LEXICOLOGICAL AND LEXICOGRAPHICAL ACTIVITIES AT THE ISTITUTO DI LINGUISTICA COMPUTAZIONALE

ANTONIO ZAMPOLLI

In 1979, the Linguistics Division (DL) of CNUCE (previously the Centro Nazionale Universitario di Calcolo Elettronico) became the Istituto di Linguistica Computazionale (ILC), an institute of the Italian National Research Council (CNR). The Linguistics Division had been constituted in 1970 to meet the demands of the increasing number of projects, for the most part of a lexicographic nature, which between 1965 and 1970 were using the resources of CNUCE in order to automatize their text processing procedures[1].

1. The three projects in course in Italy before 1965 were R. Busa's pioneer *Index Thomisticus* (R. Busa, A. Zampolli 1968), phonological, lexical and grammatical statistical studies on Italian texts in Padua (Zampolli 1960, 1968a), and the early text processing experiments for the lexical archive of the «Accademia della Crusca» (Duro 1973, Accademia della Crusca 1966, Zampolli 1973b). These three projects were transferred to CNUCE, Pisa, in 1965/6 under the direction of A. Zampolli for the computational part. The authoritative example of the Accademia della Crusca and the success of the conference organized in 1967 by the Accademia dei Lincei (1968), together with the support given to computing in the humanities by CNUCE (Faedo 1973), encouraged a rapid increase in the number of text processing projects underway in Italian universities and led to the constitution by Zampolli in 1968 of the CNUCE Linguistics Division (LD), to which S. Emmanuele and S. Rossi were associated as computer analysts. From the very beginning, Zampolli conceived the software for text analyses as a generalized procedure (Zampolli 1973b). The LD standards for text encoding were adopted by all Italian projects and the texts produced in mrf by the users of these standards were stored in the CNUCE archives. The complexity and the high costs involved in linguistic analyses of continually growing text corpora induced Zampolli to promote the construction of parsers and also of the Italian machine dictionary which began as a project of the chairs of mathematical linguistics of Pisa University (Zampolli 1968c, 1969, 1973b, 1975). G. Ferrari, G. Malatesta, G. Cappelli, D. Brogna, E. Picchi and R. Bindi were assigned by CNUCE to the LD and in 1973 Zampolli was offered 12 scholarships by a state information retrieval project interested in the use of the Italian machine dictionary. The young linguists awarded the scholarships were given permanent positions at the LD in 1975. The growth in size and the intensive research activity (partly due to the International Summer Schools organized

Although our activities have been diversified in various ways over the years, lexicological and lexicographical researches still represent one of our main centres of interest, to which different projects are linked in a closely knit design. This design reflects our firm conviction that linguistic data processing must be considered as a single disciplinary sector in which all the different applications of the computer in theoretical and applied studies on language must converge[2].

The scope of this article is to briefly describe the research activities of the ILC with particular reference to their relationship with lexicology and lexicography.

1. *METHODOLOGIES AND SOFTWARE FOR TEXT-PROCESSING*

For more than 15 years all, or almost all, of the electronic text processing work in Italy has been performed by or in collaboration with the ILC (previously the DL). This activity regards different disciplinary areas: classical and modern literature and philology, stylistic and statistic studies, psycho-linguistic researches, historical linguistics, dialectology, etc.[3], but above all lexicology and lexicography. In all cases, the aim of automatic text processing is to organize (in computer storage or in printed form) all the linguistic units of a certain level occurring in a text (words, syntagms, grammatical categories, etc.) so that they can be retrieved and manipulated efficiently, rapidly and economically.

Dictionary production is usually divided into three stages: 1) acquisition of documentation, 2) preparation and editing of lexical entries, 3) printing and publication.

Photocomposing techniques are now used routinely by most publishing houses for the third of these stages, and electronic

with the support of Pisa University in 1970 (Zampolli 1973a), 1972 (see articles in Cahiers de Lexicologie no. 38-40) and 1974 (Zampolli, 1977a)) stimulated the development of the LD as a separate organization and resulted in the constitution in 1978/79, after a difficult period of internal and external conflict and discussion, of the «Istituto di Linguistica Computazionale» of the Italian National Research Council.

2. For a discussion of the relationships between the different fields of linguistic data processing see Zampolli (1977c).

3. See Zampolli (1969, 1973b) for a first inventory of projects.

text-processing systems are used with increasing frequency by scholarly dictionaries for stage 1. Computational support is very rarely used in stage 2.

In the following section, we will briefly describe the main features of the text-processing procedures implemented at the ILC. Section 6 gives a rapid glance at experiments in progress aimed at the preparation of a procedure for computer-assisted editing. As is well known, the various text-processing procedures adopted in the different specialized centres reflect a common logical computational schema. For this reason, we will only mention here some of the options which characterize our procedures.

1.1 *Acquisition of Texts in Machine-Readable Form (MRF). The Textual Archive of the ILC.*

From the very beginning, we have tried to introduce a generalized encoding system capable of representing all the graphemes which can be found in both historical and natural language texts, plus those which may be inserted during the pre-editing stage[4]. Most Italian text-processing projects follow the standards established by our system and the texts recorded in MRF by projects using our procedures are conserved in our archives[5]. Inputting at a keyboard is still the most common method used to record texts as, for economic reasons, we have not yet been able to purchase an optical character reader, such as the Kurzweil KDEM, for example. However, we have experimented with some success the acquisition and insertion into our archive of texts recorded in MRF by publishing houses for photocomposition[6]. Particular success has also been obtained in the exchange of data with other archives such as the Trésor de la Langue Française. The economic advantages and the consequent widening of study and research facilities for the entire Italian academic community are self-evident. In particular, I am now promoting the acquisition of homogeneous

4. See Zampolli (1975) and Zampolli, Brogna (1979).

5. See Marinelli (1981) for a first inventory of machine readable texts.

6. As an example, see our agreement with the Madrid newspaper «El Pais», which gave us 30 million running words in machine readable form.

corpora, e.g. the Brown Corpus, the Lund-Bergen Corpus, the Thesaurus Linguae Graecae.

1.2 *Production of Indexes and Concordances*

Our software can be considered as generalized in the sense that it produces the results usually requested from linguistic and literary text processing[7], for texts in any language and from any period[8]. The user decides the form and structure of his results by choosing from a limited set of alternatives. As an example, let us discuss the points to consider when preparing a concordance. (References to published concordances are given between square brackets.)

1.2.1 *Nature and Organization of Entries*

Units from different linguistic levels may founction as concordance entries. We have produced concordances in which the entries are graphic forms [Busa, Zampolli, 1975], lexical forms [Tagliavini, 1965], lemmas and forms [Accademia della Crusca, 1971a], syntagms, codes representing syntactic [De Mauro, Policarpi, 1971], semantic and thematic structures [Pavese, 1979]. In some cases, the articulation in sections and subsections of the concordance entries approaches the microstructure of a dictionary entry [Legum Iustiniani Imperatoris Vocabularium, 1977].

1.2.2 *Selectivity versus Completeness of Documentation*

The decision to produce a complete documentation or to select particular items of interest concerns:

– The entries: for example, some units may be excluded by a stop list [cf. Finzi et alii, 1977], by frequency criteria, or by manual intervention during the lemmatizing stage.
– The occurrences: for example, only the global frequency or only the index locorum is given for extremely frequent words. Mixed documentation is often produced, e.g. selected contexts and/or selected references, followed by the indication of frequencies of the remaining occurrences [Legum Iustiniani Imperatoris Vocabularium, 1977].

7. Direct and reverse indexes, various types of frequency distribution, rhyme indexes (see Zampolli 1967 and Cirese 1973), index locorum, etc. See Zampolli, Bindi, Orsolini (1981) for specimina.

8. The text must, of course, be written or transliterated in alphabetical form.

1.2.3 *Construction of Contexts*

The user can specify the maximum length of the context, which can be constructed using different algorithmic rules.

– The word for which the context is being constructed (CW) is always collocated at the center of the context [Lomanto, 1983].
– The text is segmented in a sequence of contextual units, identified by delimiting factors chosen by the user: changes of references (new line, verse, paragraph, etc.), punctuation marks (colon, exclamation mark, question mark), marks which have been manually inserted during the pre-editing stage. Each contextual unit functions as a context for all the CWs it contains (e.g. the verse [Tagliavini, 1965]). The so-called contrastive concordances constitute a very interesting case. If different versions of a text are to be compared (translations or different editions), any word of any version will receive as its context not only its contextual units but also the corresponding contextual units of the other versions [Segre, Zampolli, 1976], [Bozzi, 1981].
– The context depends on the grammatical nature of the word. For example, a preposition may be given its prepositional group as context. Obviously, this algorithm presupposes previous grammatical tagging of the text [Index Thomisticus].
– The computer gives an over-abundant context, which is then manually reduced by the exclusion of interactively selected syntagms [Accademia della Crusca, 1971a].
– The position of the CW in its context is «adjusted» according to the presence on both sides of predefined, hierarchically ordered, textual phenomena (weak punctuation marks: comma, brackets, semicolon, etc.; strong punctuation: colon, question mark, exclamation mark; reference level: verse, chapter, etc.). This system is by large that most frequently chosen by our users as it seems to provide the best compromise between two oposing requirements: space saving and «significance» [Duro, 1981].

1.2.4 *Ordering of Elements*

Ordering may affect both the sequence of the entries (e.g. the forms may be listed under their lemma in alphabetical order or in a particular morphological order) and the sequence of the contexts: text order, chronological order, alphabetical order of the preceding and/or following words. This last arrangement, in fact, seems to be preferred by classical philologists, in particular when studying formulas or the influence of an author on successive authors [Lomanto, 1983]. The concordance of the *Grammatici Latini* (Keil Edition) is a typical example of this.

If the user wishes to produce lemmatized indexes and concordances, the unit of the lexical system to which each word in the text belongs must be indicated. The end product of lemmatization will be a set of records each containing the following information: a word from the text, its reference, and its lemma. Our lemmatization procedures are described in the following section.

2. *LEMMATIZATION AT THE ILC*

At the ILC we have conducted a number of experiments in order to assess the relative merits of different lemmatization systems; both completely manual and semiautomatic systems have been examined.

2.1 *Manual Lemmatization*

In our experience, lemmatization performed for lexicographical projects consists of the basic operations specified below. In our procedure, these operations are performed on the concordances[9].

2.1.1 *Lemmatization of unambiguous forms*

The lemma is entered at the side of the first occurrence of the form. The procedure then automatically assigns this lemma to all occurrences of the same form.

9. For manual lemmatization, we advise our users to lemmatize by working on the concordances of the word forms, a method which has been found to be far the most economic. Furthermore, as the lemmatizer has in front of him all the contexts of each form, a greater uniformity of treatment is ensured.

2.1.2 *Lemmatization of homographic forms*

The different lemmas which can be assigned to a homographic form in the text are written at the side of the form. Each of these lemmas is given a number. The number of the lemma which must be associated to a particular occurrence is entered at the side of its context.

2.1.3 *Composition and Decomposition of forms*

Our procedure permits us to:
- reconstruct lexical forms which are separated in current spelling, e.g. locutions (*a gambe levate*), compound noun forms (*treno merci*), compound verb forms (*ho visto*);
- decompose orthographically united forms, e.g. enclitic and verb forms, (*dirtelo*: *dire+ti+lo*), articles and prepositions (*allo* = *a+il*), etc.

2.1.4 *Assignment of words into different processing categories*

In some applications, there is a strongly felt desire to reduce the mass of data by excluding part of the lexical material from the processing work which may follow lemmatization.

Our procedure permits different types of treatment for a lemma. We can decide to print:
- only the frequency of occurrence;
- only the index locorum;
- only certain contexts, selected by the lemmatizer;
- all the contexts.

2.2 *Lemmatization using a Machine Dictionary (MD)*

2.2.1 *Lemmatization using an MD of word-forms*

This system is currently used at the ILC to lemmatize modern and contemporary Italian texts. The Italian Machine Dictionary (DMI) consists of about 1 million graphical forms which have been automatically generated by applying a morphological inflection algorithm to a set of 120,000 Italian lemmas, taken from modern Italian printed dictionaries and recorded in machine-readable form. Each form receives an «analysis», i.e. a

set of linguistic information which includes its lemma (or lemmas, if the form is homographic in the Italian lexical system), and the grammatical categories of both the form and the lemma(s).

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

The algorithm:

– «purifies» the text word-form to be lemmatized by separating enclitics from verbal forms, and «reconstructing» the final vowel when it has been «cancelled» by elision or truncation;

– searches the «purified» word-form in the DMI, i.e. identifies the form in the DMI which is «graphically identical» to the given word-form;

– associates, to each word-form «found» in the DMI, the «analysis» suggested by the DMI;

– prints new concordances in which each word-form is followed by its associated «analysis».

When examining these concordances, the lemmatizer:

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

– checks the lemmas proposed for homographical forms and selects the appropriate lemma for each single occurrence, using the encoding system already described above for manual lemmatization;

– assigns a lemma (or lemmas for homographic forms) to forms not found in the DMI. These «new» forms are then considered for insertion in the DMI by the DMI maintenance-team.

2.2.2 *Lemmatization by Morphological Segmentation: the Morphological Analyzer for Spanish*

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

We have adopted a development of these methodologies in a system constructed to treat Spanish texts. This project (Ratti et al., 1982) aims at the construction and testing of an integrated, economical and portable system for the semi-automatic lemmatization of Spanish texts, including an attempt at the disambiguation of homographic forms. At the same time, we intend to acquire experimental data on the potential of a local syntagmatic parser, conceived within the lexicographical paradigm. The feasibility of applying the results of this project to a similar system for Italian texts will also be evaluated. Particular care is thus being given to the applied and experimental aspects of the project.

The system is composed of three components which we will now describe.

2.2.2.1 *The Preprocessor*

This component aims at lemmatizing the most frequent words in an economic, straightforward way. It compares each word, inputted in text-order, with a table of about 750 words containing the most frequent «empty» words (pronouns, adverbs, articles, prepositions and conjunctions extracted from the Juilland Frequency Dictionary) and the most frequent «idioms» extracted from the Spanish corpus stored in the ILC textual archive[10]. As a result of this «look-up table», about 50% of the words in a text immediately receive an «analysis» and are submitted straightway to the third component, without entering the more expensive morphological processor. Furthermore, very simple «graphemic» rules also identify and treat digits, a subset of proper nouns, words keypunched in specially marked fields (spurious words, foreign words, quotations, etc.).

2.2.2.2 *Morphological Analyzer*

Conceived according to the traditional scheme for automatic morphological analysis, this component processes forms which have not already been identified by the preprocessor, in

10. «Idioms» are the syntagms defined as «locución», «modismo», «modo adverbial» in the Casares dictionary. Only idioms with «empty» head-words are included in our table.

alphabetical order. It includes a stem dictionary (about 6,000 stems at present, extracted from Juilland), an affix table (about 200 suffixes and 130 prefixes), a table of endings (subdivided into 82 inflectional paradigms) and a table of possible enclitics. All these elements are given morphological codes which specify their respective compatibility.

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

2.2.2.3 *Context Analyzer*

This component operates on the words recollocated in text order after analysis by the two previous components, attempting to resolve homography between forms belonging to different parts of speech. For each possible pair of parts of speech, rules have been formulated which examine the context immediately surrounding the homograph, up to a maximum of eleven contiguous words. Specific elements (words or grammatical categories) or sequences of elements are searched. The formulation of these rules, essentially based on cooccurrences, has been aided by statistical algorithms which have been applied to already lemmatized texts in the archives of the ILC, to provide the frequency distribution of words and grammatical categories in the immediate context of homographs in the sample corpus. At present, an estimated 80% of words in a text can be successfully analyzed. The remaining 20% are submitted to an interactive, manual disambiguation process. A feasibility study, which explores the possibility of constructing a similar system for Italian texts, will also examine whether this strategy for disambiguation at relatively low costs can be advantageously used as a preliminary step towards the Italian syntactic parser (see section 5).

3. *THE ITALIAN MACHINE DICTIONARY (DMI)*

3.1 *Present Contents of the DMI*

The DMI consists of three linked files:

a) *The File of Lemmas (DMIL)*
The DMIL consists of 120,000 records. Each record contains:
– The *lemma-word*, written according to its usual orthography, plus codes indicating particular phonetical features which cannot be directly computed by rules based on sequences of graphemes[11].
– *Morphosyntactic labels*: parts of speech, gender, etc., a total of 60 different subcategories.
– *Indication of foreign words*, i.e. words which do not pertain to the Italian phonological system (e.g. *walkie-talkie*).
– *Homograph codes*: distinction is made between «lexical» (*pesca*: fruit/sport) and «grammatical» (*amico*: noun/adjective) homography.
– *'Semantical' explanation*: a very brief definition (synonyms, paraphrases.....) is given to distinguish between homographic lemmas which pertain to the same parts of speech. (Homographs at the grammatical level are already distinguished by grammatical codes.)
– *Pointers*: these give the user the possibility to treat certain lemmas as part of other lemmas[12].
– *Usage Status*: archaic, dialectal, popular, literary, etc.
– *Paradigm*: grammatical codes specifying the type of morphological inflection.

11. These codes distinguish, in specific contexts, between (ó : ò), (é : è), (s : z), (ts: dz), (i : i̯), (u : u̯), and marks *gl*, *gn* in those rare cases where they represent a biphonomatic nexus. They also denote the position of the tonic stress.

12. Zampolli (1969 and 1975, chapter 2) discusses the feasibility of a DM which can be adapted to different criteria of analysis by different users. When present, pointers have the following form:
(A)—R(i)→N(L), where *N* is the numeric code in the DMI which unambiguously identifies the lemma (L) to which the current lemma (A) is related. R (i) specifies the type of relation: e.g. (A) is a graphical variant of (L); or (A) (according to certain criteria) may be considered (not as an autonomous lemma, but) as an inflected form of L (e.g. *me*, *mi* of *io*). If the user of the DMI decides to «activate» a relation, the occurrences of (A) will be lemmatized under (L).

b) *The File of Forms (DMIF)*
This file contains approximately 1 million forms, automatically produced by an inflection algorithm applied to the lemmas of the DMIL. All possible inflected forms were produced for each lemma, with the exception of alterated forms, forms with enclitics, truncated and elided forms.
This file contains:
– *The form*: written with its usual orthography. The inflection algorithm transfers the phonetic specification from the lemma to the inflected form and computes any modifications (e.g. changes in position of the tonic stress).
– *Morphosyntactic labels*: (number, gender, tense, etc.). Homographic forms of the same lemma (e.g. *dica*: 1st, 2nd, 3rd pers. sing. subjunctive/3rd pers. sing. imperative) are given all pertinent morphological labels.
– *Usage*: Particular forms are marked as archaic, literary, dialectal, etc.

c) *The File of Definitions (DMID)*
This file contains 150,000 definitions for the nouns, adjectives and verbs present in the DMIL.
– *Type of Definition*: 1) *Relational*, e.g. *abbacchiare*: *atto effetto dell'* ('fixed' part expressing a function) *abbacchiare* ('variable' part, called the 'semantic base'); 2) *Synonymical*, e.g. *abavo*: *trisavolo*, *arcavolo*; 3) *Complex*, e.g. *abaca*: *pianta bulbosa tropicale delle musacee* (tropical bulbous plant). Seven other types are used and combined in different ways with the three preceding examples.
– *Definitions*: the definitions of the reference dictionaries have been reformulated in accordance with specific criteria established for each definition type.
– *Taxonomy*: a numbering system is used which represents the classification of the different meanings of a polisemic lemma.
– *Usage*: popular, archaic, etc.
– *Semantic Procedure Code*: e.g. metaphor, metonimy, extension, etc.
– *Technical Sublanguages*: More than 100 codes are used to indicate that a specific definition belongs to a particular sublanguage or terminology, e.g. anatomy, physiology, bureaucracy, economics, law, agriculture, astronomy.

3.2 *General Objectives of the DMI*

It is our conviction that a machine dictionary must be a multifunctional tool: for semi-automatic lemmatization in lexicographic text-processing; as the lexical component in a parsing system; as a computational representation of the lexical system; etc.

The quantity of data, and the fact that they are only accessible via the alphabetical ordering of the entries, have, up until now, prevented the exploitation of the immense treasury of linguistic information contained in printed dictionaries. The transcription of printed dictionaries into machine-readable form means that these difficulties can be at least partly overcome.

Let me give a very simple example. From the very beginning, the structuralist school had affirmed that quantitative studies on the structure of the lexicon would be a necessary complement to quantitative studies on running texts. We may well ask why until now, despite their claims, statistical researches on the lexical system of a language have been very rare.

While texts may be regarded as 'given' objects, the inventory of the units and functions of a lexical system must be constructed. In a certain sense, printed dictionaries are the only comprehensive representation available of the lexical system of a given language, or at least of the way in which it has been codified within lexicographical tradition. However, their dimension makes manual exploitation extremely difficult.

The computation of the relative frequencies of the different phonological oppositions, fundamental for the study of the so-called functional load in the phonological system, requires an exhaustive inventory of all the minimal pairs. The only possible way to compile an inventory of this type is to execute a very heavy and complicated program on the phonological transcription, in MRF, of all the possible forms (including inflections, variants, etc.) of a language.

3.3 *Future DMI Research Programmes*

We shall be considering three main areas: methods of access to the DMI, lexical (sub)sets recognized by the DMI; linguistic information supplied by the DMI.

3.3.1 *Access Methods*

A general system is now being developed. The aim is to create a «dictionary server», capable of *receiving* requests from different types of users (*programs*: lemmatizing procedures, parsers, natural language processing systems; *human*: linguists, terminologists, translators, «ordinary» users of a future «telematic dictionary»), of *finding* the relevant lexical entry/ies, and of *answering* with information of the type specified by the user (a subset of the information stored in the DMI). Some modules of this system are now in operation: look-up by morphological segmentation; retrieval of relevant lexical entries using certain information categories. Some categories may be given immediately (substrings of the lemma, parts of speech, usage limits, homographic types, keywords in the definitions, etc.). Others must be «computed» by specific algorithms, which may follow, for example, chains of hierarchies (e.g. part <–> whole), relations (*act of* ...), etc., explicitly or implicitly embedded in the definitions.

The way in which these modules can be organized and integrated in a query system language with other modules, still to be developed, is now under study.

3.3.2 *Extension of the lexical (sub)set recognized by the DMI*

Some of the problems which have to be dealt with for the DMI may well be of general interest.

3.3.2.1 *«Unknown» words and DMI Updating*

In some applications, it may well be important to provide a tentative analysis of a word which is «unknown» to the DMI, to aid the successive parsing stage for instance. In other applications, such as lemmatization in a lexicographic project, it may be desirable to submit the «new» words to the attention of the lexicographer. In particular, new words must be carefully examined for (eventual) insertion in the DMI. Since the DMI can be considered as a computational representation of the Italian lexical system, and because of prospective use as a general purpose «telematic» dictionary, its updating may be of great importance.

Apart from typing errors and casual omissions, a word in a text may be «unknown» to the DMI for two main reasons:

a) The new word is a neologism.

A «page by page» comparision of the Xth (1970) and XIth (1983) editions of the *Zingarelli* Dictionary immediately reveals the extensive evolution of the Italian lexicon in little more than ten years (about 10% of the lemmas are new). The large majority of the new words are complex words[13] (in particular derived with prefixoides, suffixoides) and a few highly productive suffixes. There are also numerous loan words and compounds, which are usually orthographically separate locutions.

Different processes are used to attempt to assign information to a new word. If morphological decomposition is successful, a complex word can be fully recognized in terms of its elements. Even if decomposition does not result in a complete analysis, some elements of the word may have been identified, e.g.:

- the second word of the compound is recognized ((X)–*dipendente* as in *Hi-Fi-dipendente*): the category of this word may be tentatively assigned as a result of the analysis;
- a string SUFFIX+ENDING (e.g. *–izzare*, *-izzazione*) is recognized, but not the base: the syntactic label may be assigned with a (statistically) reasonable degree of certainty;
- only a (possible) inflectional ending (*–ebbero*, *–emmo*) is recognized. In only a very few cases is an ending represented by a univocal string. In all other cases, the information related to different possible homographic endings may be «transmitted» to the parser.

Experiments on suffix and ending decomposition without dictionary control have shown (for English) that in over 80% of cases the morphosyntactic classes of words could be determined correctly.

13. Since the very beginnings of computational linguistics, and in particular of machine translation, there have been discussions as to whether «complex words» should be inserted directly into the MD or whether they should be analyzed by morphophonological decomposition, in order to reduce the size of the MD and to permit the recognition of complex neologisms. At the present stage, complex words are autonomous DMI entries. Exhaustive indexes can thus be compiled, regularities recognized, etc. Further research is needed to decide on the final treatment(s) of complex words in the DMI. The problems, which can mask the morphological transparency of decomposition, do not seem to be much more complex than those posed by inflection. However, it seems impossible, for example, to completely specify rules defining the affixes

b) The «new» word is a member of a set intentionally omitted from the DMI.
Proper nouns, abbreviations and acronyms, foreign words and technical terms are included in this set. An ad-hoc treatment may be envisaged for the first three types. *Proper nouns* may be recognized as such (i.e. receive the label «proper noun») by playing on information such as «capital first letter, not preceded by strong punctuation». For *abbreviations* and *acronyms*, the strategy may consider situations such as: «sequence of capital letters; string ending with a period; etc.». We must also consider the advantages to be gained from inserting into the DMI the most frequent proper nouns (in particular when these are homographic with other words), the most frequent acronyms and abbreviations, and «foreign words» recorded (with increasing frequency) in «good» Italian dictionaries.

The situation for technical terms is different. They may, in many cases, be recognized by a complete decomposition process. Our problem is to find an «overall» strategy for the treatment of technical terms in our lexical database. The problem will become more pressing if our text-processing system and our parsers are to be applied to technical texts, both for lexicographical and practical NLP-type applications. Our working hypothesis is to have the DMI at the centre of a star-like structure of technical dictionaries, so that the user can select the DMI plus a domain-specific terminological subset.

«compatible» with a given base. The risk is, obviously, that of a «false homography»: i.e., of creating an inexisting derivate noun *minata* (like *fucilata*), homograph with the past participle of *minare*. The frequency of such risks can be verified in the DMI. A possible solution is to specify, for each base, the affixes actually accepted, as rules formally specifying also how to compute the morphosyntactic label, the semantic subcategorization, the selectional restrictions on the derived words imposed by the properties of the base. Different procedures will probably be studied for different applications and for different complex word types. The alterative derivation must certainly be treated by decomposition, except for «lexicalized» cases (such as *cavallone*). Only compounds not «listed» in the DMI must be (tentatively) treated by decomposition. Efforts to treat derived words by decomposition may be worth while for affixes known to be still very productive, such as *-tore*, *-zione*, *-ista*..., and in particular the so-called «prefissoidi» and «suffissoidi». (Semantic transparency is proportional to actual productivity). Specific application needs must also be considered; e.g., translation of words which exist as derived words in one language but not in another; in certain grammars, cases in which an adjective modifies only the verbal base in a derived noun.

3.3.2.2 *Analysis of Fixed Phrases*

We can distinguish at least three types of fixed phrases:
1. non-inflected fixed phrases such as «per lo più», «di volta in volta», «non ti scordar di me», «a malapena», «in cambio di», «piedi piatti», «fuggi-fuggi», «amor proprio», «tira e molla».
2. inflectable fixed phrases such as: «tavola calda», «ubriaco fradicio», «far fuori», «tener testa», «ufficio informazioni», «nave scuola», «ragazza squillo» «calcio d'angolo», «figlio della lupa».
3. There are also semi-fixed inflectable phrases whose components can be modified by the addition of further linguistic entities such as: «far (gran) conto», «fare (buon) uso», «pensarci (bene) su».

In the first case, the solution may be very simple (see the Spanish morphosyntactic preprocessor). In the second case, the situation is complicated by cases such as *tavole calde*. The problem here is that, while an adjective may be inflected in a text, its form cited in the dictionary is neither its basic form nor the form which it has in the text. The solution could be to consider the sequence of bases. This implies the identification of the phrase after morphological analysis. This fact and the need to recognize (at least in principle) the syntactic structure of the phrases quoted in point 3, could suggest integrating in a certain way the recognition of inflectable and discontinuous fixed phrases within the syntactical analysis (e.g. at the level of nominal groups, verbal groups, etc.).

3.3.3 *Completion of the DMI Linguistic Information*

The existing MDs are of different types, ranging from a simple machine-readable transcription of a printed dictionary to lexical structures used as a knowledge source by AI systems, and to multifunctional lexical databases[14].

14. Dictionaries stored in MRF can be divided into the following types on the basis of their origins.
a) *Transcribed Printed Dictionary (TD):* The text of a printed dictionary (PD) is recorded in machine readable form without any supplementary information, apart from (eventual) photocomposition codes. The number of TDs will probably increase in the near future because of the diffusion of photocomposition. (In fact, there are already 15 TDs for English).

It is our opinion that a major goal of computational linguistics, and in particular of computational lexicology, is to construct a lexicological data base. This should include:
– traditional lexicographical data;
– data resulting from processing of text corpora;

b) *Classified Transcribed Dictionary (CTD):* Codes explicitly classifying different linguistic information are inserted into the machine readable text of a TD. In other words, information on the nature and structure of the data is recorded not only implicitly via changes in type-faces (as a side-effect of photocomposition commands) but also via codes explicitly intended for future access and retrieval. Unfortunately, there are only a very few cases of this so far. The *Longman's Dictionary of Contemporary English* is probably the best known example.
c) Sometimes a TD is progressively transformed into a CTD or an MD. J. Olney's work on Webster's Seventh New Collegiate Dictionary and on the New Merriam-Webster Pocket Dictionary is well known. Optical character readers may be used to put PDs into machine readable form, for example with the aim of publishing new, and updated, editions. A recent interesting project intends to create a lexical data base recording the (out-of-print) French-Danish Blinkberg and Høybye dictionary. Experiments were conducted in order to assess whether the structure of the printed dictionary (special printed symbols, defining formulas, etc.) may be used to create, automatically or semi-automatically, an explicit linguistic classification of the dictionary information.
d) *The Lemmatizing Dictionary (LD)*: The lemmas which appear progressively in a corpus of electronically processed texts are recorded in MRF. The lemmas are supplied with information which will be useful for the lemmatization of other texts: parts of speech, inflectional paradigms, homographical relationships, brief explanations of the meanings of lexical homographs, etc. Several centres for literary, philological and lexicographical text-processing have created dictionaries of this type.
e) *The Machine Dictionary (MD):* The lemmas of a printed dictionary (PD), or of a set of dictionaries, are recorded in MRF (mainly for the lemmatization stage of text-processing) together with selected information, extracted from the PD but directly classified and encoded in a format suitable for EDP. Our present Italian Machine Dictionary is a prototype.
f) *The Parsing Machine Dictionary (PMD)*: Until now, it was all too common for each separate natural language processing project to construct its own dictionary. These dictionaries are almost never related to any PD and the number of lexical items is very limited, usually restricted to a specific technical domain. The dictionary is focused on syntactic and semantic (if not pragmatic) information, highly formalized and strongly finalized to the parsers (or the generation programs) that use it.
As we will explain in more detail in section 6, we believe that, in the near future, we will see the development of lexical databases, and that the major existing printed dictionaries will be restructured and transformed in computerized, interactively accesible form.

– linguistic information for natural language processing sistems (e.g. for parsers, etc.);
– lexicological data obtained by automatic and/or semi-automatic interactive processing of the three preceding categories of information (see 6.1.1).
Within this framework, we must complete our DMI in a number of ways.

We plan to add certain information normally contained in printed dictionaries which we have not yet recorded in MRF (e.g. etymology, phraseology, contexts)[15].

We must also add the syntactic and semantic information used by parsers, e.g. verbal valences and/or case frames, semantic subcategorization, selectional restrictions, semantic formulae, etc. I cannot discuss here the range of different choices. The preliminary problem to be answered is whether it is inevitable, when inserting linguistic information, to choose at the very beginning a precise theoretical framework in connection with a specific parser/generator, or if it is possible to conceive a «pre-theoretical» status of this information, «neutral» with respect to its utilization in different theories. We must bear in mind that the construction and maintenance of a large computerized dictionary requires an enormous effort. Up until now, it has been extremely difficult, if not impossible, to use a dictionary designed for one system in another. In other words, we must investigate the possibility of building a lexical database in such a way that it can function as a «pivot», from which dictionaries required by different NLP can be (semi)automatically built.

4. LEXICAL SYSTEM FOR LATIN

For several reasons (i.e. the rich lexicographic tradition and the fact that – as a supranational language – it is the object of computerized researches in different countries), Latin seems to

15. In any case, the DMI is not a faithful transcription of a printed dictionary. This raises a problem. The actual definitions in the DMI have been «normalized» and partly abbreviated during transcription. This certainly simplifies their processing (some examples are given in 6.1.1), but we must now evaluate whether and to what extent they are still representative of lexicographical practice.

be, at least from a certain viewpoint, particularly suitable to study and experiment the feasibility of a lexical-computational tool which:

- consents a certain degree of flexibility to the user in his choice of lemmatization criteria;
- can be applied to the lexical analysis of linguistically non-homogeneous texts (for example texts pertaining to diachronically and synchronically different strata of Latin) (Zampolli, 1975);
- improve the retrieval of lexical units, identified in different ways and classified at different linguistic levels (lemma, sublemmas, forms) by different centres and scholars using different lemmatizing criteria (for a first report see Zampolli, 1976).

A first experiment was carried out in collaboration with the members of the 'Gruppo di Lessicografia Latina' promoted by the «Lessico Intellettuale Europeo». Each of the participating centres gave us a sample of lemmatized concordances. A procedure was designed (Bozzi, Emmanuele, 1980) which aimed at merging the respective lemmas in a homogeneous inventory. With the precious collaboration of N. Marinone of Turin University, our Institute is now working on the creation of a generalized lexical system which will function as an MD for the analysis of texts, and as an aid in comparing texts lemmatized using different criteria (Bozzi, 1982). On completion, this system will include the following components:

a) An algorithm (derived from the Spanish morphosyntactic analyzer) which decomposes Latin word-forms into prefixes, bases, inflectional endings, enclitics.
b) Open inventories of these elements. The bases have been taken from the Oxford Latin Dictionary and Georges.
c) Morphographemic rewriting rules, aiming at «neutralizing» diachronic and stylistic variants.
d) The set of morphological options, which characterize the different Latin lemmatization systems, organized in a dependency tree structure. Each lexical unit on a father-node may be considered as the «arch-lexeme» of the lexical units in its daughter nodes. Each different «lemmatization theory» identifies its own lexical units by specific «proper analysis» of the tree. Following this tree, equivalences and cross references may be established among lexical units recognized in a text by different «lemmatization theories» adopted by different centres.
e) Database tools connecting the system to our Latin corpora (see section 6).

5. *SYNTACTIC-SEMANTIC PARSER FOR ITALIAN*

Computational linguists have always been greatly interested in the study of the content and organization of machine dictionaries because of their usefulness in natural language processing systems.

On the contrary, lexicographers have never shown any particular interest in parsing systems. After all, the parsers in question are not, so far, capable of exhaustively treating the variety and quantity of phenomena present in the corpora which constitute the basic documentation of the lexicographer.

In computational lexicography, parsers could be used for at least three different purposes:

- to assist in the disambiguation of homographic forms during text lemmatization;
- to study collocations, and to identify citations exemplifying particular syntactic and semantic patterns, at the editing stage;
- to analyze the definitions given in a dictionary in order to improve the 'discovery procedures' which attempt to reveal latent information in the dictionary (see paragraph 6.1.1).

We hope that the syntactic-semantic parser for Italian (ATNSYS) now being prepared at our Institute will also be considered, in the future, as a tool for these three tasks.

ATNSYS is a syntactic parser based on Wood's ATN (Augmented Transition Network) written in MAGMA-LISP (a dialect of LISP implemented in Pisa). It is oriented towards the treatment of complex sentences from narrative and descriptive texts.

5.1 *The System*

The system has a preprocessor, separated from the parser for reasons of efficiency, which can lemmatize the items of the string in input. Furthermore, it isolates, one by one, the sentences with a simplified grammar based mainly on a control of the punctuation marks.

At present, a dictionary is included which contains about 18,000 forms extracted from the DMI on the basis of Bortolini et al., 1972. Devices to modify, erase or insert new lexical items have also been implemented.

The grammar consists of four networks (50 states, 122 arcs), and, at present, analyzes complex declaratives with relative clauses, complements, and both explicit and implicit subordinate clauses.

Cappelli et al. (1978) give a description of the recognition strategies, especially of those which represent special properties of the Italian language, namely missing subjects, NP deletion in complements, classification of relative clauses, reflexive pronouns, etc.

In output, ATNSYS gives a 'deep-structure' type representation of a canonic form of the sentence (subject, verb, complements governed by the verb, other complements). Only those classical transformations which lead to this canonic order of the constituents are used. Functional labels on certain constituents enrich the structural representation. A set of labels corresponding to the classical complements (cause, locative, aim, etc.) has been chosen.

A definite set of arguments, together with an indication of their surface realization, is associated to each verb in the dictionary. This corresponds to some notions of 'verbal frame'.

Functional labels are first assigned to the arguments belonging to the verbal frame; afterwards, when possible, they are assigned to the other constituents of the sentence[16].

5.2 *Features and theoretical aims*

The ATN model has been chosen for its characteristics of perspicuity and modularity, as well as its capability to maintain the original distribution of the constituents of the input string.

16. At present, the assignment of functional labels is performed by operations of intersection between lexical information. The verb, by its own frame, activates a list of hypotheses which specify both the possible prepositions by which one argument may be realized in surface, and the requisites (selection restrictions) of the noun, head of the prepositional phrase. The hypotheses are verified by intersection between the expectations, the prepositions present in the text, and the selection restrictions of the head (Cappelli et al. 1980).

This last feature makes ATN formalism suitable for studying a perceptual model of the analysis process, given that it seems that some perceptual strategies may be formalized into a network (Kaplan, 1972). The aim of the project is not only to build a system in order to analyze the most complex utterances possible and to establish a formalized grammar to be progressively improved, but also to achieve a series of theoretical objectives, i.e. the identification of the limits of ATN from the point of view of the integration between linguistic theory, psycholinguistic experimentation and computational efficiency.

New elements have already been introduced into Wood's ATN, i.e. new categories for the tests and new actions. The goal is to treat specific phenomena of Italian more easily, to use lexical information extensively (in particular for the verbal frame), and to access information stored during analysis of already processed sentence constituents.

It should be pointed out that the fact that attention has been focused on the syntactic level does not imply that the syntactic and semantic analyses should be separate components working in a hierarchical manner.

A study has been begun which aims at investigating the possibilities of integrating knowledge of different kinds in one process of analysis, namely to integrate the level of the conceptual knowledge representation in a global framework which takes into account the perceptual model.

This hypothesis has the following features. A process, of a distributional kind, guaranteed by ATN formalism, controls and triggers lexical cognitive procedures. These procedures affect a conceptual knowledge representation based on the KLONE language, designed and implemented at our Institute (Cappelli, Moretti, 1983). Implications and possible uses of this language on lexicological and lexicographical research must certainly be explored in the near future, in particular for the study and representation of lexical knowledge and structures.

Experiments already performed have suggested changes and improvements in some crucial features of ATN on the level of integration between linguistic theory and computational efficiency, namely concerning the utilization of procedural lexical information in the control of the analysis process and the

utilization of the information already stored during the analysis, independently of the computational levels.

5.3 *Statistical Applications*

ATNSYS has already been experimented on samples of narrative and scientific texts with the aim of verifying some statistical hypotheses on syntax which are of utmost importance for the study of style, but which are neglected today[17]. Quantitative data on syntactic patterns could be a valuable help in editing a lexical entry, for citation selection, microstructure design, etc.

6. *TOWARDS A LINGUISTIC DATABASE (LDB)*

In order to increase the potential use of textual material in MRF[18] and bearing in mind the wide-spread «penetration» of

17. It is well known that ATN is a non-deterministic model where arcs exiting from each state are attempted in the order in which they have been written by the human compiler of the network. In ATNSYS (whose strategy is «search depth first») a heuristic device has been integrated which dynamically rearranges the arcs exiting from a state according to frequency data collected during previous analyses. This device is divided into two components. The first consists in the acquisition of statistical data, i.e. «the frequency, for each arc exiting from a node, of the passage across that arc, in relation to the arc of arrival» (Cappelli et al. 1980). This leads, of course, to a probabilistic distribution: experiments have been conducted whose results would seem relevant to the study of the relationship between the distribution of syntactic features in a sample and in the entire text. In my opinion, ATNSYS can be used to perform, at a syntactic level, the research program suggested by those who have tried to propose a quantitative model of stylistic phenomena within the generative transformational paradigm (Doležel 1969, Rosengren 1971, Zampolli 1975). This model requires the quantification of the choices, made by an author, from the various alternatives that the syntactic competence offers at each stage of text production. Researchers also tend to verify these ideas at the receiver level, convinced that the expectations of the reader or listener, and hence the perceptual strategies, are not only determined by the subject of the utterance but are also related to the style and register of the expression. Relationships with studies in the field of «sublanguages» are evident. The second stage of the statistical device dynamically re-orders the arcs of the network. It has been designed principally in order to optimize the computational load and the parser «penetration» (Ferrari, Stock 1980). In other words, coming from a state, the arc with the maximum of probability in relation to the specific arc of arrival is first attempted. Many measures are set up which take into account increases in system efficiency.

18. The continually increasing set of linguistic data in MRF is not

computing technologies in our «information society» (terminals, personal computers, telematic networks), which leads to a greater familiarity with computational tools, we should like to suggest working with the following objectives:

a) «Release» the researcher from the static ordering of the indexes and concordances (and, even more so, of the traditional dictionaries), profiting from data base methodologies and distributed access technologies, in favour of multiple dynamic interactive access to the data.
b) «Enrich» the archives, not only with regard to the quantity of the primary data (texts, dictionaries, etc.), but also, and above all, by the explicit representation of the units and relations of the different linguistic levels (secondary data)[19]. With this aim in mind, we must attempt to:
- reduce the cost of linguistic text analyses, using automatic or computer-assisted procedures (dictionary look-up, parsers, etc.);
- propagate minimal norms and computational tools which make it possible to compare analyses performed by different researchers (see section 4 above);
- encourage collaborations not only for data acquisition in MRF but also for analyses. Whoever uses a text from an archive must «deposit» the analyses he has made, even if only partial, so that they can be exploited by other users.

c) «Assist» the researcher in identifying and processing pertinent data. It seems necessary to:

accompanied by a growing use of these data by a proportionally wide range of users. In fact, the texts stored in MRF in the archives of specialized centres are seldom used once the researcher who first stored them has concluded his work. A great effort at both the organizational and research levels is needed to encourage greater use of the large corpora of linguistic data stored and processed in MRF, at costs which are certainly not trivial in terms of time, human resources and public funds. Furthermore, literary and philological electronic text processing has so far been mainly limited to the production of printed indexes and concordances of *graphical units*, which basically have the function of facilitating retrieval of lexical units. The researcher usually works completely manually on these printouts.

19. Secondary data, e.g., phonological patterns, marks of «things worthy of note» (proper nouns, quotations, etc.), lemmas and their conceptual and semantic categories, metric scanning, formulas, types and dependencies of sentences and clauses, syntactic and semantic microstructures, etc.

- identify the «profiles» of potential user categories, describing the operations and the procedures which constitute their research activities;
- create interrogation, processing and access functions to assist the performance of these operations;
- make available in the same «interactive environment» different «knowledge sources»: dictionaries, bibliographies, rules, parsers, statistical analyses, etc.

We use the expression *Linguistic Data Base* (LDB) to define *a set of linguistic data* of different types (not only texts and dictionaries, but also, for example, the results of socio-linguistic studies, linguistic models, bibliographic data, etc.) *finalized* for the interactive utilization by multiple categories of potential users, *stored*, *structured* and *linked* for this purpose, and *associated* to specialized software modules for access, interrogation and on-line processing.

The design of an LDB of this type is still under development and several projects of our Institute are naturally converging in this direction. Let us now mention briefly some examples of our activities, and the principal areas in which we feel our work should be concentrated in the immediate future.

6.1 *Examples from the activities of the ILC*

6.1.1 *The DMI as a Lexical Database*

When a dictionary is organized on a DB structure, accessible at many levels, it becomes evident that it contains far more information than was immediately apparent[20]. Certain research projects of our Institute on the DMI aim at designing procedures that identify and evidence this latent information. These procedures make a large-scale analysis of the definitions,

20. The simplest form of access to dictionary data consists of reproducing dictionary entries on a suitable display device in response to a query about a specific word. The next step is to reproduce (a portion of) the dictionary entries in response to a query that matches a portion of its information fields (string of characters in the lemma, word(s) in the definition, etc). There are, however, certain dictionary properties which cannot be indexed because they are not made explicit either in the printed dictionary or in its computational transcriptions.

thus enabling, via interactive consultation, the retrieval and reordering of different semantic subsets within the lexicon.

Some types of semantic subsets have been partly retrieved for further analyses: equivalence relationships in a taxonomic organization of the entries; synonymical relationships; collocational relationships; derivational relationships. A typical example is the selection from the lexicon of different natural taxonomies, in order to study hyponomic subsets[21].

A working group of our Institute now intends to experiment a possible organization of the Italian lexicon in semantic fields, and eventually to produce a dictionary organized according to these fields[22]. Some quantitative aspects of the derivative phenomenon have been investigated, e.g. the percentage of derived words in the DMI together with their distribution according to certain categories such as the part-of-speech of the base and of the derived word. Different possible treatments of derived words in the DMI have also been examined by working interactively on a subset of selected suffixes. A proposal for formalizing some of the uniform modifications which can be effected by the suffixes on the meaning of the bases has been made in order to obtain a more compact and normalized treatment of derived words[23].

21. In practice, the processing starts by interactively examining the definitions of an initial set of appropriately selected defining words. A new set of words appearing in these definitions is selected, and the definitions are then examined, and so on recursively until the members of the smallest set are found. A simple parser is now being implemented to assist the interactive selection of pertinent words by evaluating their function in the definitions. (Calzolari, 1983b).

22. The group has adopted the theory and terminology of E. Coseriu in which the lexical field is «una estructura paradigmática constituida por unidades léxicas (lexemas) que se reparten entre sí una zona de significación común y que se hallan en oposición inmediata las unas con las otras». An interactive procedure essentially permits the retrieval of lexical units whose definition contains a given word or cooccurrence of words. This helps to exhaustively retrieve groups of definitions from the DMI expressing «similar meanings». Other functions make it possible to «compare» these definitions and to organize them hierarchically in tree-like structures, which can be printed using photocomposition (Ratti et al., 1983).

23. The first stage is to recognize and extract recurrent and general patterns of definitions associated to the suffixed words: e.g. «relating to», «characteristic of», «set of», «act or effect of», etc. These could be grouped into «types of modifications» with an associated label. When a label is attached

6.1.2 *The Lexical Database as a Tool for Development and Education*

The developers' community, sometimes in collaboration and sometimes in competition with publishing houses[24], wishes to produce new «electronic» products based upon lexicographical data which until now has been marketed by the publishers of printed books. The new market open for exploitation includes office automation, personal computers, word processing, home information services, educational software. Analogous interest is shown by public organizations. It is well known that the U.S. National Institute of Education is exploring the feasibility of building pocket-book sized, portable computerized dictionaries. The study will include the type of prospective user, number of entries, type and ordering of information, simple multiple access modes aimed at providing further search capabilities enhancing the usability of the dictionary, both on the educational and entertainment level: search for synonyms, spelling games, word search (through its definition), pronunciation (via sound synthetizers), grammar and word usage correction[25]. Word processors plus an MD and associated parsers are thought to be tools which encourage the acquisition of writing skills.

6.1.3 *The LDB as a Lexicographic Workstation*

Bernard Quemada has already amply illustrated the feasibility of an LDB containing information that could be used to

to derived words it will have an effect similar to a lexical rule consisting in a semantic operation on (one of the semantic readings) of the base. A further step will be to consider the possibility and productivity of rules computing the transfer of syntactic properties (syntactic features and selectional restrictions) from the base to its derivatives (Calzolari, 1983a).

24. Of course the publisher attempts to capitalize on the «electronic rights» of their electronic lexicographical data. Permission is usually granted, through the stipulation of agreements, to use their data for scientific studies and researches aimed at enhancing the lexicographic patrimony of knowledge. Dictionaries, as we have seen, are sources of the information needed by computational linguists to construct natural language processing systems. In this case, the behaviour of the publishing houses is less clear. The problem of copyright becomes more difficult when knowledge derived from printed dictionaries is used in products of potentially commercial interest.

25. See the French project described by Ch. Muller at the ALLC Symposium, 1982.

generate dictionaries of different levels and scopes, instead of having a different DB for each dictionary. A major challenge in designing an LDB is to develop a «lexicographic workstation», at which the lexicographer can use the different sources of knowledge existing in the LDB for the task of editing the dictionary entries. According to some scholars, the publication of a dictionary in the future will represent a design decision made at an editor's workstation, in which components of an underlying lexical database are «sculpted» together in an attractive visual form, without changing any of the underlying computer data. This prospective may not be so far removed for the updates of existing traditional dictionaries. The lexicographer could effectively examine and characterize newly found citations automatically extracted from incoming text stream, modifying and creating lexical entries in an existing DB. In this way the dictionary is continually updated and remains contemporary with the use of language (Amsler, 1983).

Far more complex is the situation if we devise an integrated system for the compilation of new dictionaries in which the gap between data collection via electronic text-processing and final photocomposition is filled by computer-assisted editing of the entries (see section 1)[26]. Some experimental projects underway essentially aim at facilitating the access to textual samples and dictionary entries stored in an LDB, and the storing of preliminary versions of lexical entries for further processing.

A first component ensures quick access to the data of the LDB, thus enabling the lexicographer to use the corpora interactively via terminal. Some possible functions of this component were presented in our computer demonstration during the Workshop. For example, the lexicographer can search specific word-forms, word-forms matching (beginning,

26. In 1966, in cooperation with the Accademia Della Crusca, we decided to produce as a final result of the text processing operations for «Il Tesoro delle Origini», a mechanographical «citation slip» (scheda-contesto) containing the following printed information for each corpus occurrence not excluded during lemmatization: lemma, form, date, reference, and a context of about 700 characters. We also envisaged the possibility of automatically merging the electronically produced citation slips with citation slips xerographically or manually produced for texts submitted to a rare excerption (Accademia della Crusca, 1966).

containing or ending with) a specified string of graphemes, cooccurrences of word-forms and/or grapheme strings in a given span of text. If the texts are already lemmatized, the lexicographer may operate on both lemmas and/or word-forms. As output, the component provides the lexicographer with information on the relative and absolute frequencies in the corpus. The lexicographer may then request the contexts of selected occurrences to be displayed on the video screen or to be output in printed form[27]. The context, which is algorithmically «cut out» by the computer, may be interactively modified by the addition or exclusion of selected syntagms.

A second component enables the lexicographer to consult existing dictionaries in the LDB[28]. The lexicographer will obviously benefit from the multiple dictionary access techniques described above (cf. 3.3.1), which enable him to search different information fields in the entries of both existing dictionaries and the dictionary which is being constructed[29].

Summarizing, the lexicographer can construct the «microstructure» of a lexical entry by examining lexical entries in existing dictionaries and comparing their descriptions with contexts in the corpus.

We have demonstrated a function which permits the insertion and cyclic reordering of selected contexts in the different sections of the microstructures, thus producing a preliminary version of the new dictionary entry. All stored information can be altered, expanded and corrected at any time, and used immediately by multiple access procedures, in order to improve homogeneity and coherence[30].

27. The system developed by E. Picchi (1983) for interactive access to very large textual corpora also offers other search functions such as, for example, the recursive definition of the access keys. A key may be defined as a list of categories, each category being in turn definable as a list of categories, and so on down to categories defined as a list of lemmas and/or forms, and/or (sub)strings. This system has also been designed for philological or literary applications. Typical examples are searches in a text for cooccurrences of words belonging to specified semantic or thematic fields, or for fixed or semifixed compound words.

28. As an example, see the lexicological project of the Mannheim «Institut für Deutsche Sprache».

29. E.g., to recall entries pertaining to the same technical subfield, or to the same definition or derivation type of the entry being constructed.

30. E.g., it is possible to recall and to decide to reactivate, at any stage *n* of this processing cycle, the structure of the data at any of the preceding *n-1* stages, by a trial and error type process.

The functions we have programmed so far serve as examples, and we have deliberately not coordinated them into an integrated procedure. We feel that greater collaboration between lexicographers and computational linguists is needed if a complete experimental procedure is to be constructed[31]. In particular, the operations which the lexicographer performs when preparing a dictionary entry must be analyzed accurately. Specific computational functions must be identified and created in order to allow some of these operations to be performed interactively (thus avoiding a number of successive transcriptions of the same data). We also suggest exploring the feasibility of at least partially automatizing some searches on large quantities of data. For example, to identify syntagmatic patterns or particular collocations, or to choose citations which reflect the frequency distribution of different usages of a lexical unit in a corpus.

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.» 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» (Kiepfer, 1982). We could thus have to continue to print concordances or, even better, citation slips which can be used by the editor in the traditional manner. In order to avoid retyping selected citations, all the citations stored in the computer's memory can be univocally numbered. The editor could then key in only the microstructure (headword, etymology, grammatical information, definitions, etc.) and for each section could key in the code numbers of the citations he wants.

The first explicit and general discussion on this problem was probably held during a round-table between lexicographers from more than ten countries held in Pisa, 1972[32]. The

31. For early proposals of this kind of procedure see Quemada (1973) and Zampolli (1973c). Some demonstrations were given at our 1972 International Summer School. Some procedures (or parts of procedures) have already been implemented. See Venezky (1977) and Lentz (1981).

32. This round-table was organized during our 1972 Summer School, with

conclusions were very uncertain. The situation today is probably somewhat different due to the evolution of database methodologies and of workstation technologies, which seem to offer today the chance of «simulating» on the video the games of «solitaire» which the lexicographer has always played ordering and reordering the traditional slips[33].

6.1.4 *Statistical Linguistics*

Researches on the quantitative aspects of the DMI as a representative sample of the Italian lexical system have already been mentioned above. Our text archive must also be regarded as a source for statistical research. Some studies have already been made at the phonological level (Marioni, 1974), lexical level (Crocetti, 1976), and grammatical level (Zampolli, 1975).

As is known, the diffusion of text processing has partly falsified the models and 'laws' formerly established (Guiraud, 1954, Herdan 1956, 1960). Within both the structuralist and transformalist frameworks, some linguists have tried to draw up a research program leading to the construction of a new explicative model of frequency stability and variation in the texts. Both approaches agree with the necessity of building such models inductively by identifying, through frequency distribution analyses in sufficiently extended and stratified corpora, the relationship between subsets of texts (literary genres, different authors, diachronical distribution, etc.) and the stability or instability of the frequencies of the linguistic units.

The archive of the ILC, which contains diversified and stratified corpora, offers useful material, although we hope to

the participation of Quemada, Barr, Aitken, De Tollenaere, Bailey, Dimitrescu, Muller, Zampolli, and others. The Proceedings were published in «Cahiers de Lexicologie», no. 38-40.

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

fill up progressively the gaps existing in some fields (e.g. scientific texts, spoken texts). However, the major obstacle to the development of such research plans (both in general and at the ILC) is the scarcity of texts analyzed at different linguistic levels. We hope that the cooperative effort required for the construction of an LDB will contribute to extending the subset of analyzed texts.

6.2 *Future Research and Development*

It is only possible to outline here a few of the many paths which are likely to be taken by future research and development.

Definition of the types of data which could be merged into an LDB. Primary data: e.g. texts in MRF, dictionaries, replies to questionnaires, etc. Secondary data: processing and analyses of primary data (indexes, concordances, distribution of frequencies, lemmatization, grammatical tagging, etc.).

Formats. Each data category may require a specific structure at the computational level and a representation formalism at the linguistic level. Once again, we have the problems of data «neutrality» with respect to different theoretical positions, and of the minimal standards for data representation and analyses.

Juridical aspects. Copyright for the different parties involved: authors, publishing houses, those responsible for data recording and processing, software writers, DB managers, etc. Almost nothing has been done in this sector so far[34].

Hardware. The mainframe computer, peripheral units and workstations require certain optimal features: on-line management of a sufficient quantity of data, facilities for character changing and programmability, flexibility and size of screen. Technical compatibility must also be considered for large user networks. For a discussion of these points see the article by P. Bratley in this volume.

Software. This must be conceived as an open set of modules for different functions, some general (e.g. data acquisition and maintenance) and others problem-oriented (e.g. creation of indexes and concordances, dictionary servers, parsers, statistic-

34. Copyright and organizational problems were continually evidenced during this workshop, and particularly during the final discussion.

al packages, pattern research, etc.). In this way, the substantial convergence of different trends of linguistic data processing in a unified field of «computational linguistics» is reaffirmed.

Interaction. An LDB will mainly be used in interactive mode. According to his specific needs, the user must be able to combine query functions and interactive data access with software modules, supplied by the LDB or written ad hoc if necessary. For this purpose, the study of the professional profiles of the potential user classes is essential[35].

User information and training services. During the workshop, there was discussion on the alternatives available for training and updating lexicographers. Similar alternatives are raised for the LDB. These vary from clearing-house services, to the simple presentation of articulated «menus» for the occasional user, the in-depth knowledge of data structures and access functions, and the computational linguistics experience required to prepare user-specific software modules and personalized computational linguistic tools[36].

Exchanges with other LDBs, archives, etc. Efforts must be made to promote reciprocal exchanges of information (centralized clearing-houses, questionnaires, etc.), and technical and scientific compatibility (e.g. exchange formats, generalized data

35. It is our opinion that not only different kinds of linguists, lexicographers and literary critics but also other categories could well be interested, e.g. language teachers, terminologists and specialists in technical fields, journalists, sociologists, politicians, lawyers and, in consideration of the diachronic dimension of the LDB, historians of cultures, ideas, economics, politics and various other scientific disciplines as well. For instance, we can think of the traditional types of users for a historical dictionary. It would really be necessary to conduct a user survey in order to estimate the frequency of consultation, the psychological attitude of the users and their experience with other DBs, the potential range of applications, the elements of the LDB which are of main interest, query types, facilities required, etc. On the other hand, it is also true that the user is formed during his effective use of the DB, progressively discovering new application possibilities. We hope that a contribution will come from the publishing houses, who, we believe, are going soon to have to face these problems in order to structure and distribute the computerized versions of their dictionaries and, in general, of their lexical databases.

36. In any case, it seems clear that the rapid spread of computational tools in everyday life will necessarily involve the humanities. At the workshop it was suggested that within the academic world the construction of a database should also «count» as a scientific publication.

conversion programs, theoretical frames of reference and «metalanguages» to describe unambiguously the data and their analyses[37]).

Collaboration with users. The organization and management of an LDB requires a considerable effort at both administrative and financial levels, which must be justified by an adequate use. Collaboration is necessary at many levels: to identify fields of interest and potential users; to study user profiles; to study and establish «norms» and «conventions» for data representation and analysis; to adopt technically compatible hardware and software; to put into common use data, analyses, software modules; for cost sharing[38].

The concept of linguistic database must include that of a «bank»: a place where «goods» are not only withdrawn but also deposited and thus made available to others.

7. *INTERNATIONAL MEETINGS AND TEACHING ACTIVITIES*

The ILC regularly organizes seminars and conferences within its field of activity. Among the meetings organized in Pisa (in cooperation with the Chair of Computational Linguistics of the University), we can mention:

- International Symposium «Lexicon electronicum latinum» (1968)
- Colloque international sur l'élaboration électronique en lexicologie et lexicographie (1970) (Zampolli, 1973a)
- COLING 73: Vth International Conference on Computational Linguistics (1973) (Zampolli, Calzolari, 1977, 1980)
- European Science Foundation Workshop on «The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries» (1981)
- Round Table on «Ordenadores y Lengua Española» (1981) (Catarsi et al., 1982)

37. During the workshop discussions, it appeared clear that standardization, as such, is not a realistic solution. The most important and urgent step is to define clearly the content, the material and the analysis criteria.

38. I will promote experiments in areas where groups of researchers are already well intentioned towards cooperation, as in the case of the prospective Italian users of the «Thesaurus Linguae Graecae».

- Round Table on «Knowledge Representation in Italy» (1981) (Cappelli, 1983)
- «Computers in Literary and Linguistic Research», VII International Symposium of the ALLC (1982) (Cignoni, Peters, 1983)
- Conference and Inaugural Meeting of the European Chapter of the Association for Computational Linguistics (ACL 1983).

The ILC is involved in various didactic activities. Our researchers collaborate with colleagues at the University of Pisa (which in 1970 introduced the only official Italian university chair in computational linguistics) and also with other universities for seminars, courses, theses, etc. In reply to the invitation of the President of the Italian National Research Council, the Institute is collaborating in the new Italian «dottorato di ricerca» for the mathematical and computational linguistics curriculum. The staff of the ILC collaborates in professional training courses for teachers. Furthermore, the advice and consultation provided by the Institute to its users (in particular lexicographers) must not be forgotten. This has always been the most efficient method for the dissemination of knowledge and methodologies for linguistic data processing in Italy.

The most relevant activity is surely the International Summer School in Mathematical and Computational Linguistics, sponsored by the University of Pisa and CNR. The following editions have been held:

1970 – Introduction to Computational and Statistical Linguistics (Zampolli, 1973a)
1972 – Lexicology, Lexicography and Electronic Data Processing (Various articles in Cahiers de Lexicologie, No. 38-40)
1974 – Natural Language Understanding (Zampolli, 1977a)
1977 – Syntax, Semantics and Computational Linguistics.

We have had many requests for the organization of the next Summer School on the topic of «Lexical Databases».

REFERENCES

Accademia della Crusca, *L'Accademia della Crusca e l'Opera del Vocabolario*, Firenze, 1966.

Accademia della Crusca, *Concordanze del Canzoniere di Francesco Petrarca*, Firenze, 1971.

Accademia Nazionale dei Lincei, *Atti del Convegno «L'Automazione elettronica e le sue implicazioni scientifiche, tecniche, sociali»*, Roma, 1968.

Accademia Nazionale dei Lincei, *Atti del Convegno «Tecniche di classificazione e loro applicazione linguistica»*, Roma, 1975.

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

ARONOFF M., *Word Formation in Generative Grammar*, MIT, 1976.

BARTOLETTI COLOMBO A.M. (Ed.), *La Costituzione della Repubblica Italiana, Testo, Indici, Concordanze*, Firenze, 1971.

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

BOZZI A., *Il trattato ippocratico e la sua traduzione latina tardo-antica. Concordanze contrastive con il calcolatore elettronico e commento linguistico-filologico al lessico tecnico latino*, Pisa, 1981.

BOZZI A., Esperimento di fusione automatica di lemmari latini in machine readable form: problemi, metodi e risultati, in M. FATTORI, M. BIANCHI (Eds.), *RES-III Colloquio Internazionale del Lessico Intellettuale Europeo*, Roma, 1982.

BOZZI A., EMMANUELE S., Thesaurus Mediae et Recentioris Latinitatis, Forthcoming.

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

BUSA R, CROATTO-MARTINOLLI C., CROATTO L., TAGLIAVINI C., ZAMPOLLI A., Una ricerca statistica sulla composizione fonologica della lingua italiana parlata con un sistema IBM a schede perforate, in *International Association of Logopedics and Phoniatrics, Proceedings of the XIIth International Speech and Voice Therapy Conference*, Padova, 1962, pp. 542-562.

BUSA R. (Ed.), De lexico electronico latino, Actes du Seminaire International sur le Dictionnaire Latin de Machine, *Calcolo*, V (1968) Supplement 2.

BUSA R., ZAMPOLLI A., Centre pour l'Automation de l'Analyse Linguistique (C.A.A.L.), Gallarate, in *Les machines dans la linguistique*, Prague, 1968, pp. 25-34.

BUSA R., ZAMPOLLI A., *Concordantiae Senecanae*, Hildesheim, New York, 1975.

CALZOLARI N., An empirical approach to circularity in dictionary definitions, *Cahiers de Lexicologie*, XXXI (1977) 2, pp. 118-28.

CALZOLARI N., Polisemia e omografia nel Dizionario Macchina dell'Italiano, *Studi di Lessicografia Italiana*, II (1980), pp. 283-313.

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

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

CALZOLARI N., CECCOTTI M.L., Esperienze preliminari alla organizzazione di archivi per l'accesso diretto a dati linguistici, NT-DMI/DB 1, ILC-CNR, Pisa, 1980.

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

CAMPANILE E., ZAMPOLLI A., Problems in Computerized Historical Linguistics: the Old Cornish Lexicon, in A. ZAMPOLLI, N. CALZOLARI (Eds.) (1977), pp. 50-69.

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

CAPPELLI A., FERRARI G., MORETTI L., PRODANOF I., STOCK O., Automatic Analysis of Italian, in *Proceedings of the AISB-80 Conference on Artificial Intelligence*, Amsterdam, 1980.

CAPPELLI A., FERRARI G., MORETTI L., PRODANOF I., STOCK O., Semantica lessicale in un analizzatore sintattico automatico per l'italiano, in *Atti del Convegno SLI «Lessico e Semantica»*, Roma, 1981, pp. 363-72.

CAPPELLI A., FERRARI G., MORETTI L., PRODANOFF I., Towards an Integrated Model of Sentence Comprehension, *Linguistica Computazionale*, II (1982), pp. 45-57.

CAPPELLI G., CATARSI M.N., RATTI D., SABA A., Análisis morfológico y lematización automática de textos en lengua española, NT, ILC-CNR, Pisa, 1981.

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

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

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

CIRESE A.M., Inventaires et repertoires lexicaux, formulaires et métriques des chants populaires italiens, in A. ZAMPOLLI (Ed.) (1973a), pp. 209-239.

Concordanza dei grammatici latini, *Atti dell'Accademia delle Scienze di Torino, II – Classe di scienze morali, storiche e filologiche*, CXIII, Supplement, Torino, 1979.

CROCETTI C., Criteri di analisi del testo per una statistica del lessico dei quotidiani, Thesis, University of Pisa, 1976.

DE MAURO T., POLICARPI G., Ricerche sulla struttura del periodo italiano, in *Atti del Convegno SLI «L'insegnamento dell'Italiano in Italia e all'estero»*, Roma, 1971, pp. 583-694.

DEVOTO G., Linguistica matematica e calcolatori, in A. ZAMPOLLI (Ed.) (1973a) p. V.

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

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

DURO A., Elaborations electroniques de texts effectuées par l'Accademia della Crusca, pour la preparation du Dictionnaire Historique de la langue italienne, in A. ZAMPOLLI (Ed.) (1973a), pp. 53-75.

DURO A., *Concordanze e Indici di frequenza dei Principi di una Scienza Nuova – 1725 di G. Vico*, Roma, 1981.

DURO A., ZAMPOLLI A., Analisi lessicali mediante elaboratori elettronici, in Accademia dei Lincei (1968), pp. 121-139.

FAEDO A., Discorso di Apertura, in A. ZAMPOLLI (Ed.) (1973a), pp. VII-IX.

FATTORI M., BIANCHI L. (Eds.), *I Colloquio Internazionale del Lessico Intellettuale Europeo*, Roma, 1976.

FERRARI G., STOCK O., Strategies Selection for an ATN parser, in *18th Annual Meeting of the Association for Computational Linguistics and Parasession on Topics in Interactive Discourse, Proceedings of the Conference*, Philadelphia, 1980.

FINZI A., ROSSELLI F., ZAMPOLLI A., *Concordancias y frecuencias de uso en el léxico poético de Antonio Machado*, Pisa, 1977.

FINZI A., ROSSELLI F., ZAMPOLLI A., *Diccionario de concordancias y frecuencias de uso en el léxico poético de Cesar Vallejo*, Pisa, 1978.

GRUPPO DI PISA, Il Dizionario di Macchina dell'Italiano, in *Atti del Convegno SLI «Linguaggi e Formalizzazioni»*, Roma, 1979.

GUIRAUD P., *Les caractères statistiques du vocabulaire*, Paris, 1954.

HAYS D.G. (Ed.), *Readings in Automatic Language Processing*, New York, 1966.

HERDAN G., *Language and Choice and Chance*, Groningen, 1956.

HERDAN G., *Type-token mathematics. A textbook of mathematical linguistics*, The Hague, 1960.

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

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

KAPLAN R.M., Augmented Transition Networks as Psychological Models of Sentence Comprehension, *Artificial Intelligence*, III (1972).

KAY M., Standards for Encoding Data in a Natural Language, *Computers and the Humanities*, I (1967) 5, pp. 170-177.

KAY M., Morphological and Syntactic Analysis, in A. ZAMPOLLI (Ed.) (1977a), pp. 131-234.

KIEPFER B., Computer applications in Lexicography. Summary of the State-of-the-Art, Yale University, 1983.

Legum Iustiniani Imperatoris Vocabularium: Novellae, Pars Latina, Milano, 1977.

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

Les Machines dans la Languistique, Praga, 1968.

LOMANTO V. (Ed.), *Concordantiae in Q. Aurelii Symmachi Opera*, Hildesheim, 1983.

MARIONI B.M., Studi di statistica fonematica su un corpus di italiano parlato, Thesis, University of Pisa, 1974.

MARINELLI R., A Provisional List of Texts processed by the ILC from 1968 to 1980, NT-ARC 1, ILC-CNR, 1981.

PAVESE, C.O., *La Lirica corale greca*, Roma, 1979.

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

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

RATTI D., CATARSI M.N., SASSI M., SABA A., RUIMY N., Datos y métodos para la organización del léxico italiano en campos semánticos, Simposio de la sociedad de lingüística Española, Barcelona, 1983, Forthcoming.

RATTI D., SABA A., CATARSI M.N., CAPPELLI G., *Analizador Morfosintáctico de textos en lengua española*, Pisa, 1982.

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

ROVENTINI A., Relazione sul funzionamento e sul rendimento del DMI in lemmatizzazione, NT-DMI, ILC-CNR, Pisa, 1981.

SCALISE S., *Morfologia Lessicale*, Padova, 1983.

SEGRE C., ZAMPOLLI A., Le Concordanze diacroniche dell'Orlando Furioso, in «*Ludovico Ariosto: lingua, stile, tradizione*», Milano, 1976.

TAGLIAVINI C., Introduzione, in *Concordanze della Divina Commedia*, IBM Italia, 1965.

TAGLIAVINI C., Applicazione dei calcolatori elettronici all'analisi e alla statistica linguistica, in Accademia Nazionale dei Lincei (1968), pp. 111-118.

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

WOODS W.A., Transition Network Grammars for Natural Language Analysis, *Communications of the ACM*, XIII (1970), pp. 591-602.

WOODS W.A., Lunar Rocks in Natural English: Explorations in Natural Language Question Answering, in A. ZAMPOLLI (Ed.) (1977a), pp. 521-569.

ZAMPOLLI A., Studi di statistica linguistica eseguiti con impianti IBM, Thesis, Padova, 1960.

ZAMPOLLI A., Nota tecnica, in *Raccolta Barbi di Canti Popolari Italiani. Esperimento di Elaborazione Elettronica*, E1/RB, Pisa, 1967, pp. II-XI.

ZAMPOLLI A., Recherche statistique sur la composition phonologique de la langue italienne exécutée avec un système IBM, in *Les Machines dans la Linguistique* (1968a), pp. 25-34.

ZAMPOLLI A., L'elaboratore elettronico negli studi linguistici, in *Rivista IBM*, (1968b) 2, pp. 14-19.

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

ZAMPOLLI A., Due conversazioni sullo stato attuale della linguistica computazionale, Pisa, 1969.

ZAMPOLLI A., Nota Tecnica, in BARTOLETTI COLOMBO A.M. (Ed.) (1971), pp. XVII-XXVI.

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

ZAMPOLLI A., La Section Linguistique du CNUCE, in A. ZAMPOLLI (Ed.) (1973a), 1973b, pp. 139-199.

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

ZAMPOLLI A., Humanities Computing in Italy, *Computers and the Humanities*, VII (1973d) 6, pp. 343-360.

ZAMPOLLI A., L'elaborazione elettronica dei dati linguistici: stato delle ricerche e prospettive, in Accademia Nazionale dei Lincei (1975), pp. 23-107.

ZAMPOLLI A., Les dépouillements électroniques: quelques problèmes de méthode et d'organisation, in M. FATTORI, L. BIANCHI (Eds.) (1976), pp. 173-197.

ZAMPOLLI A. (Ed.), *Linguistic Structures Processing*, Amsterdam, 1977a.

ZAMPOLLI A., Introduction, in A. ZAMPOLLI, N. CALZOLARI (Eds.) (1977), 1977b, Firenze.

ZAMPOLLI A., BINDI R., ORSOLINI P., Problemi e metodologie per l'automazione degli spogli lessicografici, NT, ILC-CNR, Pisa, 1981.

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

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

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