Introduction

1 Corpus linguistics, the use of computers in the humanities, computational linguistics

1.1 *The tradition of corpus linguistics*

The main goal of "corpus linguistics" is the study of language through the analysis of large quantities of naturally occurring data. The use of "authentic data", excerpted from texts selected according to specific criteria, has a long tradition in various disciplines.

Lexicographic activities, for example, normally make use of large citation archives, excerpted from textual corpora. This is particularly true for "scholarly lexicography".[1] "The source data of scholarly lexicography are the examples, in actual use, of the vocabulary being documented. It is important to mention "in actual use", because information derived from lexicographer's intuition is never used as the sole basis for an entry or part of an entry" (Weiner 1994, p. 417).

Before the introduction of computers, lexical data took the form of quotations, registered on hand-written or type-written slips of paper organised into large manual files, excerpted from the "reading" of a set of texts, selected according to explicit or implicit criteria, and intended to define or to represent the language or the linguistic subject to be documented by the envisaged dictionary.

The major European lexicographical projects[2] adopted computerised methods and techniques as soon as they became available, in some cases starting in the early 1950's[3] to produce indexes, build concordances, do various types of frequency counts and produce automatically, with the computer, paper slips similar to those produced manually.

The production of indexes and concordances has a long history, dating back to the *Concordantiae Scripturae Sacrae* of the Middle Ages.[4]

Some authors (Bortolini *et al.*, 1971, p. IX) date the initial consideration of the quantitative aspects of language use even further back in time: specifically, to the period of the Alessandrini grammarians, who took into account the opposition between rare words (or, even, "apax

[1] By the term "scholarly", "academic" or "institutional" lexicography we designate the various lexicographic activities promoted by public institutions, such as research institutes, language academies, universities, and so on, as distinct from "commercial lexicography", typically supported by publishing houses. Of course, several borderline cases exist, depending on differing organizational structures in various countries. We prefer, therefore, to use the term *scholarly* lexicography, making reference to the typical product, the *scholarly dictionary*: usually a monolingual, comprehensive dictionary, founded on objective evidence that has been collected by empirical methods, which conforms to the so-called "historical principles", and whose primary audience will be the specialist: the student of language and of literature that is relatively inaccessible, the historian, etc. (See Weiner 1994, pp. 413-414).

[2] In particular, the institutional ones.

[3] The first experiments on the use of computers in text analysis are due to R. Busa S.J., who in 1951 published "a first example of word index automatically compiled and printed by IBM punched card machines" (Busa 1951).

[4] "(...) the Ministers of the Gospel (...) have been able to enhance their sermons with abundant biblical quotations, thanks to the friendly pages of one of the many, both old and new, editions of the Concordantiae Scripturae Sacrae which can be found in parish houses" (Busa 1951, p. 8). In 1907, Herman Schöne had already enumerated fifty indexes and lexicons of individual Greek authors; in 1914 Paul Rowald listed 144 of them for Latin authors.

legomena") and words of high frequency and usage.[5] In the XIX century, frequency counts of linguistic units were established to support practical applications, for example in cryptography and stenography. F.W. Kaeding's study, for instance, entitled *Haufigkeitswörterbuch der deutschen Sprache* (published in 1898) was based on a corpus of eleven million words, manually analysed to count the frequencies of graphemes, syllables, and words. Using this data, J.B. Estoup, in his *Gammes sténographiques* (published in Paris in 1907), noted the statistical regularities in the list of the word forms in a text, ranked in order of decreasing frequency, and this became the starting point of the well-known work of G.K. Zipf (1935). The period between the two wars was defined (Michea 1964) as the 'heroic era' of the frequency dictionaries. Several corpus-based projects, aimed at producing frequency dictionaries to be used mainly for language teaching, were carried out in a number of countries after World War I. The Prague linguistic school, in the same period, provided a theoretical foundation for the study of the quantitative aspects of the language, in the structural linguistic paradigm (Trubetzkoy 1939). When, as noted above, in the '50s and '60s, the use of computers spread to various sectors of the humanities, in particular for the production of indexes, concordances, and frequency counts,[6] the volume of textual data, in machine readable form, available for statistical studies increased rapidly. The first half of the '60s was characterised by an outstanding effort to clarify the methodological problems of various aspects of the statistical study of linguistic data. Some (for instance, Heilmann 1963) claimed that the probability of occurrence of a given linguistic unit, estimated on the basis of its frequency in a corpus, was one of the pertinent features of this unit, as part of a structured linguistic system. The interpretation of the frequency, in a corpus, of units of various types and linguistic levels was discussed within the framework of the relation between a sample (the corpus) and the population (the language as a whole) (Moreau 1962). Textual corpora also provided the quantitative data for a set of experiments carried out to evaluate the effectiveness of the use of statistical techniques as tools for various types of stylistic studies, as had been proposed in the '50s and the early '60s, in particular by the French school (Guiraud 1954 and Muller 1964).[7]

Some scholars in Corpus Linguistics have repeatedly emphasised that, in some sense, modern Linguistics has its roots in Corpus Analysis since "linguistic theorising received its modern impetus from historical Linguistics, and historical Linguistics is rooted in the analysis of corpus-based data concerning lexical, phonological, semantic and grammatical evolution. Wherever insight into the historical development of language may originate, the final appeal must be to the historical record - to the corpora of extant texts" (Biber & Finegan 1991, p. 205).

It has been suggested that the roots of modern Corpus Linguistics can be traced to the age of post-Bloomfieldian structural linguistics, originating in the USA. "This was when linguists

[5] "Per quanto l'esigenza di una considerazione quantitativa del linguaggio dal punto di vista scientifico sia relativamente recente e l'applicazione dei metodi statistici alla linguistica sia diversamente interpretata e valutata, bisogna riconoscere che, fin dall'epoca dei Greci e dei Romani, filologi e grammatici, più o meno consapevolmente, hanno tenuto conto del fattore quantitativo, per lo meno come opposizione fra voci rare e obsolete (o addirittura *hapax legomena*) e parole di particolare frequenza e uso. E' ben noto che la grammatica classica, pressappoco come fu formulata dagli alessandrini, rappresenta un compromesso fra i principi degli *analogisti* e quelli degli *anomalisti*.
L'analògia, cioè la regola, era formata da ciò che veniva indicato come *normale* e per ciò stesso più frequente, l'anomalia da ciò che era eccezionale e per ciò stesso più raro". (Bortolini *et al.*, 1971, p. IX).

[6] Zampolli (ed. 1973) and *Les Machines dans la Linguistique* (1968) provide a survey of these activities.

[7] Dyer (1973) provides a discussion of this issue.

(such as Harris and Hill in the 1950's) were under the influence of a positivistic and behavioristic view of the science, and regarded the 'corpus' as the primary explicandum of linguistics. For such linguists, the corpus - a sufficiently large body of naturally occurring data of the language to be investigated - was both necessary and sufficient for the task in hand, and intuitive evidence was a poor second, sometimes rejected altogether" (Leech 1991, p. 8).

The advent and the rapid success of the generative linguistics paradigm, between the '50s and the '60s, provoked heated controversy concerning the value to be attached to the role of corpora, as opposed to the role of introspection, as a source of significant data for linguistic research (Herdan 1964). The views of Chomsky and his followers on the inadequacy of corpora, as opposed to the adequacy of introspection, were criticised in vain by representatives of both the corpus linguistics and the statistical linguistics camps.

Chomsky and his disciples rapidly gained the leadership in the international scene. Activities in the field of statistical and corpus linguistics[8] continued thanks, above all, to the introduction of computers, but clearly separated and almost ignored by the prevalent linguistic school. In 1959 R. Quirk (1960) announced the start of a program for the collection of a British-English corpus of written and spoken language, the "Survey of English Usage Corpus" (SEU), and, soon after, N. Francis and H. Kucera (1964, 1971, 1979) launched the Brown Corpus project, intended as "a standardised sample of written American-English to be used with computers". It was soon followed by others, and in particular by the Lancaster-Oslo/Bergen (LOB) Corpus (Johansson *et al.*, 1978), where it was decided to adopt the same model of the Brown Corpus: five hundred samples of two thousand words each, distributed in twenty sub-corpora, for a total of one million words. The decision was deliberately taken in order to allow for a comparison between the British and American English corpora. During the same period, the idea of ensuring comparability between corpora, thanks to their homogenous composition, was adhered to in the creation of a series of corpora used as the basis for the creation of frequency dictionaries for all the Romance languages (A. Juilland *et al.*, 1965; also Bortolini *et al.*, 1971). In 1975, J. Svartvik began the conversion of SEU into machine readable form (MRF), thus creating the LONDON-LUND corpus of spoken English.

The COBUILD corpus of modern English, part of the Birmingham Collection of English Texts (Sinclair 1987), formed a valuable resource for the lexicographers of the Collins Cobuild Dictionary (Sinclair *et al.*, 1987), and of many other related publications.

Afterwards, a number of projects for building corpora of different sizes, and intended for different purposes, flourished. Thereby, a community of "corpus linguists" could develop, especially for the study of different varieties of English. They gather regularly at the ICAME meetings (a list of English corpora can be found in the appendix to the volume edited by Aijmer & Altenberg 1991; Zampolli 1990 provides a survey for other languages). However, their activities went virtually unnoticed ("little noticed by the main stream", Leech 1991, p. 8), in spite of the growing number of scientific publications based on these corpora (listed in the bibliography of Altenberg 1991).

[8] It has to be noted that, often, opposers to corpus linguistics failed to clearly distinguish between the use of corpora as a source of information on 'real' language usage, and as a statistical (representative) sample for quantitative studies.

1.2 *Textual processing for the humanities and computational linguistics*

The above mentioned criticism towards the statistics and corpus based approach, in the early '60s, had some relevant consequences also on the development of computational linguistics (CL). It is well known that, when the use of electronic data processing techniques[9] on linguistic data began, two lines of research were, quite independently, activated:

Machine translation (MT).

Lexical text analysis[10] (LTA: production of indices, concordances, frequency counts, etc.).

While MT was promoted mainly in 'hard-science' departments, LTA was developed mainly in humanities departments and, probably for this reason, the two lines had, initially, very few contacts.[11]

At the beginning of the '60s, the perception of a possible reciprocal interest was explicitly manifested, in particular through the invitation of MT researchers such as Tübingen (1960) and Besançon (1961) to LTA conferences.[12] The topics quoted were, specifically, text encoding systems for different alphabets,[13] detection of the frequency of linguistic elements in large corpora, and automated dictionaries.

The year 1966 was particularly important for both lines of research, but for opposing reasons. The Prague International Conference 'Les machines dans la linguistique' ratified the international acceptance of the LTA as an autonomous interdisciplinary field which included new dimensions of processing (for archeology, historical linguistics, dialectology, etc.) and was henceforth called Literary and Linguistic Computing (LLC), whereas the ALPAC report (1966) brought about an abrupt arrest in the majority of MT projects throughout the world and the beginning of the so-called 'dark ages' of MT. Following, *de facto*,[14] the recommendations of the ALPAC report, basic research on natural language processing (NLP) occupied the area characterised so far by

[9] In fact, the first concordances and indices were produced not with 'electronic machines', but with 'punched card electrical accounting machines' (Busa 1951, p. 22).

[10] We lack an English expression which corresponds exactly to what is a technical term in the Romance languages: "dépouillement électronique des texts", in French; "spoglio elettronico di testi", in Italian.

[11] For the history of the first years of MT, see Booth, Cleave and Brandwood (1958), pp. 1-7; Vauquois (1975), pp. 14-32; Nagao (1989).

[12] In the Introduction to the 'Actes du Colloque International sur la Mécanisation des Recherches Lexicologiques' held in 1961 in Besançon, B.Quemada says: 'Un des buts de ce Colloque sera aussi de mettre en contact des chercheurs qui sans s'ignorer tout à fait, n'échangent guère d'informations alors qu'ils travaillent sur une matière commune: la langue, et plus particulièrement, le lexique dans diverses disciplines. Nous avons la chance d'accueillir ici à côté des lexicologues et des lexicographes français et étrangers, des spécialistes de la traduction automatique (vocabulaire de base, terminologies scientifiques, speciales, dictionnaires automatiques, homographes, synonymes) de la traduction "artisanale" (...) de la documentation automatique (...) de la pedagogie des langues vivantes'. And R. Busa, in an article with a very significant title (given the period) 'L'analisi linguistica nell'evoluzione mondiale dei mezzi di informazione', in a debate on 'the two cultures: the fracture between sciences and humanities', says that 'the development of linguistic automation is triangular: lexical analysis, information retrieval, mechanical translation', (Busa 1961, p. 117).

[13] M. Kay (1964), reporting on an informal meeting on the issue of Formats for Machine Readable Text at the end of the IBM-sponsored Literary Data Processing Conference (Yorktown Heights, 1964), and in an article in the fifth issue of the *Computers and Humanities* (Kay 1967), explicitly stressed the common interest of MT and humanities researchers on this topic.

[14] But not, I think, inspired by it.

MT activities, and computational linguistics emerged as a new disciplinary activity.[15]

In spite of the ALPAC statements,[16] CL focused mainly on the development of methods for the utilisation of linguistic models - in particular formal grammars - in the analysis and generation of isolated sentences, in an almost exclusively monolingual framework. The already cited success of the generative paradigm led to an almost complete disinterest in corpora analysis and quantitative data.

The analysis of texts and the quantitative approach were attracting, instead, much attention in the LLC area at the time due, among other things, to the continuously increasing availability of texts in MRF.

But, on the other hand, the LLC delayed taking advantage of the know-how, methodologies and tools produced, from the very beginning, by MT in the field of automatic lexicons. MT not only had developed research on specialised hardware,[17] storage, access techniques, inflectional and derivational morphological analysis, but certain projects had already begun to collect large sets of monolingual and bilingual lexical and terminological data.

Very few exceptions can be reported in the LLC field, all primarily motivated by attempts to automatise the lemmatisation of texts for the production of lemmatised indices and concordances. To my knowledge, the first experiments are related to Latin, at the Centro per l'Automazione dell'Analisi Letteraria (CAAL) in Gallarate and at the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) in Liège. These two systems[18] were presented and compared at the Pisa 1968 meeting 'De lexico electronico latino', during which the first proposal for a *multifunctional lexicon* was also presented (the DMI: Italian Machine Dictionary), conceived (Zampolli 1968 and 1987) not only for lemmatisation, but also as a repository of lexical knowledge both for computer programs (parsers, generators, phonological transcription, etc.) and scholarly use (qualitative and quantitative research on the structure of the Italian lexicon).

The CL activities which came after MT, almost completely neglecting the development of large lexicons, were mainly using small toy-lexicons of a few dozen words.[19]

For several years the problem of the relationship between LLC and CL was practically ignored.

As local organiser of the Pisa COLING 1973, A. Zampolli endeavoured to include in the call for papers, and to promote in the Conference, sections explicitly dedicated to topics which could

[15] The Chairman of the Committee on Science and Public Policy, in a letter to the President of the National Academy of Science, stated: 'the support needs for computational linguistics are distinct from automatic language translation' (ALPAC 1966, p. 2). And on page 29, one reads 'work toward machine translation together with computational linguistics work that has grown out of it'.

[16] We quote from the recommendations: 'Small scale experiments and work with miniature models of language have proven seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpora' (ALPAC, p. IV).

[17] See, for example, the optical disk developed by IBM in the early '60s as a storage medium for bilingual dictionaries.

[18] The CAAL Latin machine dictionary was made up of an alphabetical list of forms, progressively accumulated from processing the texts of St. Thomas Aquinus. The *LASLA Dictionary* was based on a list of stems, extracted from the Forcellini lemmas, and an associated morphological analyser (see Busa, etc., 1968).

[19] This situation was still true until very recently. 'A recent workshop on linguistic theory and computer applications (Withelock *et al.*, 1987) reports an informal poll to establish the average size of the lexicon used by the prototypes discussed (..) the average size was about 25 (words)' (Boguraev & Briscoe 1989, p. 10).

MT activities, and computational linguistics emerged as a new disciplinary activity.[15]

In spite of the ALPAC statements,[16] CL focused mainly on the development of methods for the utilisation of linguistic models - in particular formal grammars - in the analysis and generation of isolated sentences, in an almost exclusively monolingual framework. The already cited success of the generative paradigm led to an almost complete disinterest in corpora analysis and quantitative data.

The analysis of texts and the quantitative approach were attracting, instead, much attention in the LLC area at the time due, among other things, to the continuously increasing availability of texts in MRF.

But, on the other hand, the LLC delayed taking advantage of the know-how, methodologies and tools produced, from the very beginning, by MT in the field of automatic lexicons. MT not only had developed research on specialised hardware,[17] storage, access techniques, inflectional and derivational morphological analysis, but certain projects had already begun to collect large sets of monolingual and bilingual lexical and terminological data.

Very few exceptions can be reported in the LLC field, all primarily motivated by attempts to automatise the lemmatisation of texts for the production of lemmatised indices and concordances. To my knowledge, the first experiments are related to Latin, at the Centro per l'Automazione dell'Analisi Letteraria (CAAL) in Gallarate and at the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) in Liège. These two systems[18] were presented and compared at the Pisa 1968 meeting 'De lexico electronico latino', during which the first proposal for a *multifunctional lexicon* was also presented (the DMI: Italian Machine Dictionary), conceived (Zampolli 1968 and 1987) not only for lemmatisation, but also as a repository of lexical knowledge both for computer programs (parsers, generators, phonological transcription, etc.) and scholarly use (qualitative and quantitative research on the structure of the Italian lexicon).

The CL activities which came after MT, almost completely neglecting the development of large lexicons, were mainly using small toy-lexicons of a few dozen words.[19]

For several years the problem of the relationship between LLC and CL was practically ignored.

As local organiser of the Pisa COLING 1973, A. Zampolli endeavoured to include in the call for papers, and to promote in the Conference, sections explicitly dedicated to topics which could

[15] The Chairman of the Committee on Science and Public Policy, in a letter to the President of the National Academy of Science, stated: 'the support needs for computational linguistics are distinct from automatic language translation' (ALPAC 1966, p. 2). And on page 29, one reads 'work toward machine translation together with computational linguistics work that has grown out of it'.

[16] We quote from the recommendations: 'Small scale experiments and work with miniature models of language have proven seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpora' (ALPAC, p. IV).

[17] See, for example, the optical disk developed by IBM in the early '60s as a storage medium for bilingual dictionaries.

[18] The CAAL Latin machine dictionary was made up of an alphabetical list of forms, progressively accumulated from processing the texts of St. Thomas Aquinus. The *LASLA Dictionary* was based on a list of stems, extracted from the Forcellini lemmas, and an associated morphological analyser (see Busa, etc., 1968).

[19] This situation was still true until very recently. 'A recent workshop on linguistic theory and computer applications (Withelock *et al.*, 1987) reports an informal poll to establish the average size of the lexicon used by the prototypes discussed (...) the average size was about 25 (words)' (Boguraev & Briscoe 1989, p. 10).

delineate the area of common interest.

The attempt was successful in terms of joint participation, and it was probably not just by chance that J. Smith presented there, at an international level, the newly founded Association for Literary and Linguistic Computing ALLC (Smith 1973). It was ignored, however, by the CL establishment.

In fact, in those years a (so to speak) 'puristic' approach characterised the general reflections of CL, which was searching for a definition and a disciplinary identity.[20]

In this respect it is interesting to compare the Foreword of H. Karlgreen, chairman of the COLING 1973 Scientific Committee, and the Introduction to the *Proceedings* of Zampolli (1973).

The situation has changed only in the last few years. A variety of concurrent factors have contributed to finally establishing increasing contacts between LLC and CL. The awareness of the several relevant areas of common interest and needs is gaining ground on both sides. Recently, some co-operative projects have been jointly formed, especially at the international level.[21]

This convergence is partly the result of the work of some Institutes whose activities programmatically and institutionally cover both fields,[22] but above all it is fostered by the emergence of a new framework.

LLC has been, from the very beginning, interested in processing large texts, but it has failed, in gėneral, to develop computational methods suitable for analysing the texts beyond the graphemic level.

CL, from the other side, was developing "sophisticated" linguistic models and methods for their implementation, but the results were clearly inadequate for dealing with real texts.

But the recent emergence of the so-called *language industry* paradigm has forced CL to focus (at least part of) its efforts on the processing of "real language uses", in "real-world" texts for some practical applications.

Let us briefly trace the origin and the development of this paradigm which is bringing about an awareness on the part of the LLC and CL that they should develop and share common methods, tools and linguistic descriptions in dealing with large texts at various linguistic levels.

1.3 *Language Industries (LI), Language Engineering (LE), and the need for reusable Linguistic Resources (LR)*

The term 'industries de la langue' was launched at the 1986 Tours Conference organised by the Council of Europe (Vidal-Beneyto 1986). It covers both those activities where computer assistance is being developed for supporting the traditional applied linguistics professions (such

[20] The article 'The Field and Scope of Computational Linguistics' of D. Hays in the *Proceedings of the Budapest COLING 1971* is particularly relevant. It is interesting to observe the evolution towards a 'puristic definition' of CL by the author, in respect to the more eclectic attitude of his chapter on 'computational linguistics' in the *Encyclopaedia of Linguistics, Information and Control* (1969).

[21] The TEI (Text Encoding Initiative), jointly promoted by the ACH, ALLC, ACL, is a clear example.

[22] The role of D. Walker who co-operated, for more than ten years, with our efforts to move in this direction should be recognised (Zampolli, 1994).

as, for instance, lexicography, translation, and language teaching) and those activities directed towards the development of new applications: systems for natural language interfaces, speech analysis and synthesis, automatic indexing and abstracting, office automation, machine translation and, more generally, various new forms of monolingual and multilingual services.

Computational linguists, information systems developers, funding agencies and international organisations are increasingly aware of the strategic, industrial, and cultural potential of LI, emerging as an autonomous sector within the information industries (Nagao 1989).

But it is now also recognised that we are still far from being able to fully exploit this potential, and that a major engineering effort is required to use the presently available know-how and the existing prototypes for the construction of language industry products adequate to the needs of "real" users.

The term *language engineering* is now increasingly used to stress the fact that this effort should have a central role in R&D. A major task is to produce *robust* NLP components, capable of dealing with 'real texts', to be incorporated in reliable LI products, ranging from spelling checkers, to information retrieval, to machine translation.

The availability of adequate language resources (LR) is an essential condition to achieve such robustness.

In this framework, the term LR refers to (usually large) sets of language data and descriptions in machine-readable form, to be used in building, improving or evaluating written and spoken language processing algorithms, components or systems. Examples of LR are written and spoken corpora, lexical databases, grammars, terminologies.[23]

In the field of speech processing, systems are built on technologies directly based on the use of corpora of speech data[24] representing the domain of the intended applications.

In written language processing, textual corpora are recognised as the primary source of data

[23] The term, which may be extended to include basic software tools for the preparation, collection, management or accessing of the resources, was used, as far as I know, for the first time in the discussions between A. Zampolli, V. Zue, M. Libermann and J. Carbonell at the NSF-ESPRIT workshop held in Torino in Semptember 1991. It was subsequently introduced, through Zampolli 1991b, in Danzin 1992.

[24] In this *Introduction* and in the NERC study we consider a corpus as "a collection of pieces of language that are selected and ordered. according to explicit criteria in order to be used a sample of the language" (Sinclair 1994). These corpora can consist ônly of written texts (textual corpora), or only of spoken texts (spoken corpora), or include both written and spoken sections (language corpora). The current jargon distinguishes between *spoken* and *speech* corpora, which serve the needs of two different scientific communities. The traditional work of the *corpus linguistic community*, to which the first term refers, "when spoken language is addressed, starts with deriving an orthographic transcription from a recording of large stretches of speech. This transcription is afterwards enriched using different annotation systems aiming at reflecting all the important events that take place in the process of speech production - especially when speech is spontaneously produced or an interaction takes place between two or more speakers - and that are not adequately captured by conventional spelling. Furthermore, grammatical information such as parts of speech (tagging) and syntactic structure (parsing) can be added to carry out linguistic descriptive work. The main aim is to acquire large amounts of data reflecting the natural use of language, therefore emphasis is usually put on the naturalness and spontaneity of the recording, avoiding experimentally controlled situations where the speaker is constrained to utter a number of previously prepared short sequences. (...) Within the *speech community*, the emphasis so far has been on speech databases rather than on spoken corpora in the sense described above. This is due to the need to obtain controlled speech data for basic research aimed at modelling and describing the articulatory and acoustic properties of speech or, in the field of speech technology, to derive data for speech synthesis or to build up material for training and testing speech recognition systems. (...) The central issue here is the speech signal itself, and its symbolic representation is usually made by means of a phonetic alphabet - the IPA or a computer-readable equivalent being the commonly agreed international system - allowing the phonetic modifications of words when they are spoken in context to be represented. The speech wave is first segmented into units that can be related to phonetic symbols and labelled to temporally synchronise a symbol representing a set of phonetic categories with a given part of the signal - a process known as *alignment*; the phonetic representation can be also related with the orthographic representation and thus aligned with the speech signal". (Llisterri 1994, pp. 4-5). It must be noted that the success obtained in deriving speech technology from the analysis of speech corpora, in particular in the second half of the '80s, has contributed in a significant way to the "revival" of the interest for language corpora in CL.

needed to inform the description, for computational systems, of the "real use" of languages in different communicative contexts.

It is also self-evident that, in many real-world applications, NLP systems must be able to deal with tens and hundreds of thousands of lexical items.

Experience has shown that the creation of adequate large-scale LR is a costly enterprise, impossible for a single organisation, however wealthy, to carry out alone. Duplication of efforts must, as far as possible, be avoided since financial and human resources are limited. In the past, for example, the usual practice has been for each project to construct its own ad hoc lexicon, geared to one specific application or to a specific piece of research: for each new application, and even for updating an existing application, the lexicon-building could well restart from scratch, even within the same company and research team.[25]

In this context, the *reusability of LR* has become a key concern. This expression appears more and more frequently in the definition of the objectives of national and international projects and, in particular, in designated activities planned in both the third and fourth Framework Research Programs (FRP) of the European Commission. It subsumes two major complementary issues.

The first concerns the reuse of existing partial LR, usually designed for a specific application, as a 'help' in constructing new LR: for example, various machine-readable dictionaries (MRDs) have been investigated as being potentially rich and valuable sources of lexical information to help in the construction of computational lexicons.[26]

The second issue concerns the construction of new large-scale multifunctional LR, i.e. of LR explicitly intended for multiple uses, which are capable of serving, through appropriate interfaces, a wide variety of present and future research and applications.

2 The emergence and the evolution of the concept of reusable LR

The workshop 'On Automating the Lexicon', held in Grosseto (near Pisa) in May 1986, is usually recognised as the landmark event marking the starting point of the process which has led to establishing the sector of reusable LR as it is today.

This workshop was suggested in a proposal presented by A. Zampolli (1985) to the CETIL (the first EC committee of experts on linguistic information processing), for a comprehensive programme aimed at the eventual integration of monolingual and bilingual lexicons and corpora in a linguistic knowledge base.

The objective of the Workshop[27] (Walker *et al.*, 1995) was to survey research efforts, current practices and potential developments in work on the lexicons, machine-readable

[25] See the articles by Ingria and Cumming in Walker *et al.* (1995), and Boguraev *et al.* (1989).

[26] See, for example, the work of ACQUILEX, a project funded by the European Community within the framework of the ESPRIT Basic Research Program (Boguraev *et al.*, 1988).

[27] This Workshop, sponsored by the EC and organized by N. Calzolari, L. Rolling, J. Sager, D. Walker, and A. Zampolli, was preceeded by a preparatory meeting organized by Donald Walker in Palo Alto in 1983, and followed by other Workshops on specific aspects of computational lexicology in New York (1986) and Stanford (1987).

dictionaries and lexical knowledge bases, with special consideration for the problems created by working in a multilingual environment. The brief was to recommend directions for future activities. The participants were chosen to bring together, for the first time, representatives of all those working on the lexicon: lexicologists, grammarians, semanticists, lexicographers, computational linguists, artificial intelligence specialists, cognitive scientists, publishers, lexical software producers, translators, terminologists, and representatives of funding agencies and of professional associations. The final recommendations, transmitted to the EC and widely distributed, can be summarised as follows (see Zampolli 1987 for the complete text of the recommendations):

1. To establish procedures for creating multifunctional databases from the information contained, both implicitly and explicitly, in those traditional dictionaries that exist in machine-readable form.
2. To develop computational tools for more efficient handling of lexical and lexicographical data, and to provide 'workstation' environments within which these tools may be used by lexicologists and lexicographers.
3. To explore the possibility of creating multifunctional lexical databases capable of general use, despite divergences of linguistic theories and differences in computational and applicational frameworks.
4. To study the possibility of linking lexical databases and large text files, in both monolingual and multilingual contexts, in order to determine the most effective ways of exploiting the relationships among the various lexical elements.

The Grosseto Workshop, whose aim was to further the development of the scientific, technical and organisational conditions conducive to the creation of large multifunctional LR, has been followed by an increasing number of fresh initiatives, particularly at the international level.[28]

Immediately after the Grosseto workshop, we undertook two actions

a) The day after the Workshop, A. Zampolli set up an informal working party (Hans Uszkoreit, Nicoletta Calzolari, Bob Ingria, Bran Boguraev and Antonio Zampolli) to explore the feasibility of constructing large-scale LR, explicitly designed to be multifunctional, i.e. capable of serving, through appropriate interfaces, a wide variety of present and future researches and applications. A crucial and controversial problem was to define the extent to which it was possible, as well as desirable, to make LR, at least within certain limits, "polytheoretical", i.e. usable in (applications of) different linguistic theories.[29] The initial working party was gradually extended, with the support of our Institute and of the ACL, to form the so-called 'Pisa group'. The aim of this group, which included outstanding representatives of the major schools of thought in linguistics and computational linguistics, was to investigate in detail the possibility of a

[28] See a summary of these projects in Varile, Zampolli (eds.) 1992.

[29] "Current wisdom was that the "content" of the linguistic information attached to the lexical entries is so strictly dependent on the particular requirements of the specific linguistic theory, explicitly or implicitly adopted in the analyser/generator for which a computational lexicon is built, that it cannot he reused elsewhere. During the Workshop my persuasion, that these descriptions are largely based on the identification of the same linguistic properties of the lexical entries, was reinforced, so that, I decided to set up the informal working party to explore and to prove the feasibility of exploiting this fact for the construction of polytheoretical LR" (Zampolli, 1989).

polytheoretical representation of the lexical information needed by parsers and generators, such as the major syntactic categories, subcategorisation and complementation. The common representation sought was one that could be used in any of the following theoretical frameworks: government and binding grammar, generalised phrase structure grammar, 'lexical functional grammar, relational grammar, systemic grammar, dependency grammar, and categorial grammar. This group worked on examples in various languages and began by examining in detail the way in which the foregoing theories would handle a representative sample of English and Italian verbs (Walker *et al.*, 1987).

b) A. Zampolli, in the Grosseto workshop report for the EC DG-XIII prepared in co-operation with D. Walker, suggested a large two-phase programme: first a one-year phase, to define the methods and the common specifications for a co-ordinated set of mono and multilingual lexical data bases and corpora for the European languages, and a second three-year phase for their actual construction (Zampolli & Walker 1987).[30]

This proposal was endorsed by the CGC12, the Committee of experts then advising the DG-XIII on "linguistic problems". We were hoping for rapid implementation, but it became progressively clear that such a large programme required, as a preliminary condition, the explicit inclusion of the availability of LR within the major objectives of the research actions of the Commission, and that this inclusion was dependent on the formulation of an overall policy for the field of LR.

To the numerous requests of our R&D community recommending the creation of adequate LR for the various languages, the Commission has, for ten years, limited its reaction to the launching of some, not always co-ordinated, definition studies and preparatory actions.

The first of these studies (the ET-7 project) arose from the favourable reaction of the CGC12 to a proposal by A. Zampolli (1988), supported by a technical working party, to promote a project aiming at the recommendation of a methodology for the concrete construction of shareable lexical resources, building on the encouraging results of the "Pisa-Group". Since different theories use different descriptive devices to describe the same linguistic phenomena and yield different generalisations and conclusions, ET-7 proposed the use of the "observable differences" between linguistic phenomena as a platform for the exchange of data. In particular, the study confirmed the feasibility of some basic standards for the description of lexical items at the level of orthography, phonology/phonetics, morphology, collocations, syntax, semantics and pragmatics, (Heid & McNaught 1991).

Other projects focusing on the notion of standards are, in the field of lexical data, the ESPRIT project MULTILEX, whose objective is to devise a shareable model for multilingual lexicons (Khatchadourian & Modiano 1993), and the EUREKA project GENELEX, which concentrates on a model for monolingual generic lexicons (Antoni-Lay *et al.*, 1993). In the area of textual corpora the EC sponsored the NERC study, aimed at defining the scientific, technical and organisational conditions for the creation of a Network of European Reference Corpora, and at exploring the feasibility of reaching a consensus on agreed standards for various aspects of corpus building and analysis. This project, whose results are published in this issue, is described in detail below.

[30] The two European LRE and LE PAROLE projects (1995-1998) can be viewed as the implementation of this proposal, albeit ten years later.

The European Speech community had independently organised an outstanding standardisation activity, co-ordinated mainly through the ESPRIT Project SAM (Fourcin & Gibbon 1993).

Feeling it necessary to co-ordinate their activities, the representatives of these various standardisation projects spontaneously formed an initial, preparatory group.[31] Enlarging this group, the EC established the EAGLES project in the framework of the Linguistic Research and Engineering (LRE) programme. This project aims to provide guidelines and *de facto* standards, based on the consensus of the major European projects, for the following areas: corpora, lexicons, formalisms, assessment and evaluation, and speech data (EAGLES 1993; Calzolari & McNaught 1994). The project also encompasses an international dimension which includes: support for European participation in the Text Encoding Initiative (TEI); preparation of a survey of the state-of-the-art in Natural Language and Speech Processing, jointly sponsored by the NSF and the EC (Varile & Zampolli, forthcoming); preparation of a Multilingual Corpus (MLCC) intended to support co-operation with similar ARPA sponsored initiatives; and exploration of possible strategies for international co-operation and co-ordination in the field of LR.

Another step was taken by the Commission in 1991 on the occasion of a call for proposals launched in the context of the late EUROTRA activities in the second Framework Research Program (FRP). It seems to me particularly significant that, among those submitted to the call, four of the five projects retained focused on different issues of LR: the reuse of machine-readable dictionaries,[32] the role of collocations in lexical and corpus work, the use of corpus based statistics in morphosyntactic analysis and the extraction of terminologies from corpora. In effect, the EUROTRA teams, whose beneficial role in the promotion of the awareness of the relevance of LI and in the formation of a commonly shared expertise in all the European countries is sometimes underestimated, recognised that the lack of adequate LR had been one of the major difficulties for the EUROTRA project.

A crucially important achievement has been, in my opinion, to obtain the recognition, in the relevant EC Departments, of the *infrastructural role* of LR.

In 1991, the DG XIII asked a panel of experts, chaired by A. Danzin,[33] to produce a "strategic" document, delineating the general framework, the benefits, the main objectives, the priorities, and the organisational and financial conditions for the development of LI in Europe. In his contribution to this panel A. Zampolli (1991b) expressed the opinion that this development could be based only on the facilities and conditions provided by a dedicated European infrastructure, that the establishment of this infrastructure is the responsibility of European and national authorities, and that adequate reusable LR are a central component of this infrastructure.

[31] Because the Istituto di Linguistica Computazionale was involved in many of these projects, A. Zampolli felt the danger of the contradiction embedded in the multiplication of initiatives aiming to establish - independently - potentially diverging standards for the same LR. Taking advantage of the fact that at the MULTILEX Kickoff meeting (1990) the co-ordinators of the different projects gave an overview of their programs of work, he presented an analysis of their commonalities and un-coordinated over-lapping. In a meeting of the co-ordinators called in Pisa in spring 1991, the co-ordinators decided to meet regularly, to ensure the synergy and convergence among the projects, and requested the support of various EC Departments. The initial support was granted by G. Velasco and J. Soler. After a few meetings, the group of co-ordinators representing TEI, ET7, GENELEX, MULTILEX, NERC, ACQUILEX, SAM formed the initial nucleus of the EAGLES initiative following a suggestion by the co-ordinator of the LRE programme, R. Cencioni.

[32] See Sinclair, Hoelter & Peters 1995.

[33] The panel included A. Danzin, H. Coltof, B. Oakley, A. Recoque, H. Schnelle. J. Laver, C. Rohrer, A. Zampolli, and was assisted by the EC officers R.F. de Bruine, R. Cencioni, F. Mastroddi, J. Roukens.

These statements were included in the final report of the panel (the so-called Danzin Report, 1992), which was very influential in the formation of the current strategy of the Commission.

In fact, the issue of LR is now regularly present in the current initiatives of the EC in the field of language processing. For example, the LRE program includes, in addition to the above mentioned EAGLES project, other projects dealing with different aspects of LR.

MULTEXT is a very large project aiming at providing a reusable set of basic software tools for corpus work: construction, mark-up, linguistic annotation, access, terms extraction, etc.

DELIS aims at defining and testing methods and tools to build lexical entries based on the evidence extracted from textual corpora.

RELATOR aims at defining and providing the basis for a European organisation capable of ensuring the preservation and the distribution of LR produced in international and national European projects. RELATOR has suggested that this infrastructure should take the form of an Association which will provide the basic legal and organisational framework for the different operations involved in the task of collecting and distributing existing LR: identify LR available for distribution, evaluate their suitability for potential users, validate their conformance to a set of minimal technical and linguistic requirements, negotiate the licenses or the rights for distribution with the LR providers, etc. These suggested proposals have been recently endorsed by the DG-XIII, thanks to the support of M.V. Parajon-Collada, Deputy Director of the DG-XIII, and the European Association for Linguistic Resources (ELRA) has been founded.

3 The NERC Consortium and the NERC feasibility study

It can be easily noted that all the projects mentioned so far are of a preparatory nature: they explore the feasibility, define technical prerequisites, prepare basic working methods and tools, but still do not undertake the concrete creation of large LR.

In this context the NERC study, published in this issue of *Linguistica Computazionale*, should be viewed as the major effort, so far, to prepare the ground for the creation of corpora for the European languages.

3.1 *The Expert group on corpora of the Council of Europe*

The NERC study was directly preceded by the activities of a group of experts on corpus work which Mr. Vidal-Beneyto, Director of Education, Culture and Sport of the Council of Europe, asked A. Zampolli to set-up, on the occasion of the already cited 1986 Congress of Tours.

The aim was to define priorities and possible actions in the field of corpora. The initial group was formed by representatives of Institutes with a well-known tradition of working with large textual corpora: the Pisa group (A. Zampolli), the Institut Nationale de la Langue Française (INaLF, B. Quemada), the University of Birmingham (J. Sinclair), the Institut für Deutsche Sprache (W. Teubert), the University of Malaga (M. Alvar-Ezquerra). These Institutes began collaboration by comparing relevant aspects of their activities (criteria for corpus composition, encoding formats, linguistic analysis, intended usages and users, etc.) in order to explore the possibility of harmonising and co-ordinating their work in the field of corpora.[34]

In the following years, the group extended their contacts to Eastern European Institutions through a series of workshops organised in Grosseto (1987), Dubrovnik (1988), Budapest (1988).

3.2 *The Initial NERC Consortium*

In 1988, the Directorate for Education, Culture and Sport of the Council of Europe and the DG-XIII in Luxembourg decided to exchange information on their respective activities and points of view on LI. It was agreed that top priority should be placed on the availability of adequate LR. It was also immediately apparent that, although various preparatory actions had been launched in the sector of lexicons, the same could not be said of corpora.

Acting on a suggestion of Mr. Velasco of the DG XIII-E, who had attended some of the meetings of the Council of Europe corpus group, A. Zampolli (1991) prepared and presented to the DG-XIII a proposal for a feasibility study on the possibility of setting up a "Network of European Reference Corpora".

The major aims of the proposed study were:

- to verify the need and the possibility of setting up a European infrastructure linking together Institutes having the necessary prerequisites (know-how, manpower, tools, data, etc.) to answer the needs of textual corpora of the European R&D community,
- to explore the feasibility of harmonising the relevant scientific and technical aspects (composition criteria, encoding formats, linguistic annotation, etc.), and to agree on a common workplan, so as to ensure the complete interoperability of the various national reference corpora[35] in the European multilingual context.

The Commission accepted the proposal and decided to contribute to the costs involved in the study.

It would have been desirable, from the outset, to involve all of the European countries in this study. Unfortunately, due to budget restrictions, the Commission was forced to limit the number of partners to six. The NERC Consortium was formed, adding, to the five Institutes already involved in the Council of Europe corpus group, the INL (Instituut voor Nederlandse Lexicologie, Leiden), so that the initial NERC Consortium included those six languages which, at the time,

[34] The fact that the corpus work of the majority of these Institutes was intended mainly for corpus linguistics and lexicographic purposes, more than for direct NLP use, reflects the general context of those years, described above in part 1. Some results of the activity of this group are reported in various articles published in Vidal Beneyto (1991).

[35] "A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text type that are selected. A large reference corpus may have a hierarchically ordered structure of components and subcorpora.
Questions of balance and representativeness recur in the discussion of reference corpora. (...) Reference corpora in several languages, constructed on similar principles, form a group of *comparable corpora*. (...) A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages" (Sinclair 1994, p. 10, 11).

had the highest number of native speakers in the European Union. Pisa was in charge of the co-ordination.[36]

3.3 *The NERC2 Consortium*

During the execution of the study, representatives of the EU countries not included in the initial Consortium (Belgium, Denmark, Greece, Ireland, Portugal) asked to become involved. On the suggestion of B. Maegaard (Copenhagen) and J. Roukens (EC-DG XIII), the Commission asked us to form a second Consortium including the following partners: ILSP (Institute for Language and Speech Processing, Athens), CST (Centre for Sprogteknologi, Copenhagen), St. Patrick's College - University of Dublin, BELTEXT - Université de Liège, Centro de Linguistíca da Universidade de Lisboa, and Pisa as co-ordinator.

It was called the "NERC2 Consortium" to distinguish it from the initial one, which was consequently called NERC1. The main tasks of the NERC2 Consortium were to analyse the recommendations and proposals issued in the NERC1 feasibility study, and to verify their adequacy to the needs of the respective countries and languages.

The NERC2 Consortium has endorsed the results and the recommendations of the NERC1 feasibility study. In addition, it has suggested some modifications to the estimated costs for developing the spoken component of the corpora, and has provided additional information and recommendations for the specific linguistic situations exemplified by Portugal, Belgium, Ireland (NERC2 Final Report, 1993).

The suggestions have been incorporated in the NERC1 report which, in this modified version, constitutes the *Final NERC Report* published in this issue.

3.4 *The NERC Final Report and the structure of this volume*

The NERC study workplan was organised into 11 workpackages. Each workpackage was co-ordinated by one of the partners, who was also responsible for drafting a report on the results achieved in his/her workpackage.

The reports of the 11 workpackages have been collected, reordered and edited by an Editorial Committee[37] to form the *Final NERC Report* as published in this issue of *Linguistica Computazionale*.

[36] The members of the NERC1 Consortium will be designated from now on as follows:

PSA: the Pisa group, comprising researchers of both the *Istituto di Linguistica Computazionale del CNR* (ILC: Institute of Computational Linguistics of the National Research Council of Italy) and the Linguistics Department of the University of Pisa. The contract with the CEC was managed by the *Consorzio Pisa Ricerche*, a Consortium which includes, among other industrial and academic members, the CNR and the University of Pisa.
BIR: School of English - The University of Birmingham
MAN: Institut für Deutsche Sprache, Mannheim
LEI: Instituut voor Nederlandse Lexicologie, Leiden
MAL: Departamento de Filología Española, Facultad de Filosofía y Letras - Universidad de Málaga
PAR: Institut National de la Langue Française - INALF - CNRS, France

[37] N. Calzolari (Chair, Pisa), M. Baker (Birmingham), J.G. Kruyt (Leiden).

The structure of this issue, therefore, directly reflects the organisation of the NERC study workplan. It consists of eight chapters. Five contain the final report of one workpackage. Three contain the final reports of two workpackages, which have been united into the same chapter since they deal with complementary aspects of the same issue.[38]

The table below indicates:

- the correspondence between chapters and workpakage reports
- the co-ordinating partner of each workpackage.

Ch.	WP	Title	Co-ordinating Partners
0	11	Implementation Plan	PSA (A. Zampolli)
1	1,2	User Needs	MAN (W. Teubert)
2	6	Corpus Design Criteria	LEI (J.G. Kruyt, P.G.J. van Sterkenburg)
3	3	Text Representation: Written Language	PAR (P. Lafon)
	4	Text Representation: Spoken Language	BIR (J.M. Sinclair)
4	7	Text Acquisition and Reusability	PAR (P. Lafon)
	5	Access and Management Software Tools	BIR (J.M. Sinclair)
5	8	Linguistic Annotation of Texts	PSA (N. Calzolari)
6	9	Corpus Annotation Tools	BIR (J.M. Sinclair)
7	10	Knowledge Extraction	PSA (N. Calzolari)

The *initial chapter (chapter 0)* focuses on a series of recommendations to meet the needs, for textual corpora and lexicons, of the European R&D communities.

The first step recommended is to provide a set of comparable harmonised reference corpora and generic lexicons for all the European languages. The lexicons should be gradually enriched, to extend their coverage and to progressively include new types of linguistic information for which adequate specifications will eventually be provided by advancements made in the state-of-the-art in linguistics and computational linguistics. The corpora should be regularly updated to

[38] E.g., chapter 3 deals with the representation of written (workpackage 3) and spoken (workpackage 4) texts.

reflect the evolution of the languages: the reference corpus should become a "monitor" corpus.[39]

The report argues that a permanent dedicated European infrastructure is needed to ensure the desired continuity, and suggests a two-tier organisation: a *European network* of national focal points, each of which co-ordinates, in its own country, a *national network* of corpora and lexicons providers and users.

The chapter discusses various aspects of the envisaged infrastructure, from the desirable characteristics of the focal points to a possible overall organisation and management, and suggests a workplan for the first years of activity.

Initially, the mandate of the NERC study was restricted to the field of corpora. However, during the second half of the project, the Commission requested that the NERC Consortium also prepare recommendations and a proposal for a workplan in the field of computational lexicons, parallel to the one prepared for the field of corpora. This report, prepared by Pisa (N.Calzolari), was initially intended for presentation in a workshop on the perspectives of LI, held in Luxembourg in the spring of 1992. After the discussion and the endorsement of this report by the workshop' participants, the Commission suggested that it be included in the final NERC Report. It is attached as appendix 1 to the initial chapter.

Chapter one describes the various actions undertaken by the NERC consortium to assess the needs for corpora by different types of users (workshops, interviews, questionnaires, etc.), and summarises the requests of current and prospective users, in the form of short-term and medium-term recommendations.

Chapter two surveys the various criteria adopted, explicitly or implicitly, in the design of a number of existing corpora, and suggests a prototypical composition scheme for the multifunctional comparable reference corpora to be created, during the first phase of the envisaged workplan, for each European language.

Chapter three discusses the principles that should inform the establishment of a set of guidelines for text representation, i.e. for encoding the relevant structural and typographical elements of written texts, and for transcribing and encoding spoken texts. It recommends adopting the TEI guidelines as the basis for the development of the written text representation conventions. As far as spoken texts are concerned, four different levels of transcription are identified, on the basis of the comparison of the TEI Guidelines with the experience gained in various projects, and their relevance for different tasks and different contexts is discussed. An example of guidelines for English spoken text transcription is provided in the two appendices.

Chapter four gives an overview of the main functionalities that should be provided by the software tools which should be made available for working with corpora. It starts by discussing

[39] "The first model (of a *monitor* corpus) was of a corpus of a constant size, so that the software of the day could cope with it, which would be constantly refreshed with new material, while equivalent quantities of old material would be removed to archival storage. The constitution of the corpus would also remain parallel to its previous states.
This model gave rise to the idea of *rate of flow* as the best way of managing the corpus. Instead of setting, say, 10 million words as the proper proportion of that genre, the setting could just as easily be 10 million words a year. Or a month, or a week. The language would flow through the machine, so that at any one time there would be a good sample available, comparable to its previous and future states.
Such a model opened up new prospects for those interested in natural language processing, and it added another dimension to contemporary corpora - the diachronic. New words could be identified, and movements in usage could be tracked, perhaps leading to changes in meaning. Long term norms of frequency distribution could be established, and a wide range of other types of information could be derived from such a corpus. (...) Over time the balance of components of a monitor corpus will change. New sources of data will become available and new procedures will enable scarce material to become plentiful. The rate of flow will be adjusted from time to time" (Sinclair 1994, p. 11).

the technical and operational problems concerning the acquisition of texts from various types of sources (OCR, photocomposition, etc.): methods for the conversion from the source format, effort needed, etc. Then it reviews the basic software functions required for accessing, managing, maintaining and making available very large corpora.

Chapter five discusses the feasibility of establishing commonly agreed linguistic annotation schemata. We use the term "annotated" to indicate a corpus "enriched" with a systematic encoded representation of linguistic categories occurring, at one or more levels of linguistic description, in the texts and, in some cases, of their (structural) relationships. An annotation schema has two components: 1) the set of annotation symbols (form) with a definition of their meaning (content), and 2) the guidelines for their application. The chapter discusses the situation at the phonological, morphosyntactic, syntactic and semantic levels, but focuses in particular on the morphosyntactic one. The majority of corpora, already collected or in progress, are "raw" corpora.[40] Very few annotated corpora exist, but the number is constantly increasing. This trend has been particularly strong in the last few years and is expected to continue, in particular at the morphosyntactic level. This increase is also encouraged by the spread of probabilistic taggers. This chapter compares in detail the tagsets used in 14 major corpora (4 British-English and 3 American-English, 2 French, 2 Italian, 1 Swedish, 1 German, 1 Dutch) and discusses the methodological problems involved, and possible solutions for establishing a common tagset.[41]

Chapter six argues that a new generation of tools is required to adequately analyse the continuously increasing quantity of textual data available. Lemmatisers, taggers, parsers and, in particular, various types of "detectors" of lexical patterns can assist the researchers in various tasks (for example, in identifying multi-word lexical units and collocations, and in selecting lexicographically relevant examples of word uses), and can be used to (semi)- automatically disambiguate word-meanings, etc.

"The new corpora suggest a new kind of inquiry into the nature and the structure of language - not one where the main aim is confirmation of what is already fairly well agreed, but one where the exploration is likely to uncover facts about language and languages that have not been available before."

Chapter seven further elaborates on the extraction of linguistic information directly from corpora, focusing in particular on the use of this information in a number of NLP tasks: lexicons construction, speech applications, word processing, document retrieval, machine translation.

Various classes of methods and techniques are reported for the extraction of morphological, morphosyntactic, syntactic and semantic information.

[40] We oppose annotated corpora to *raw* corpora, i.e., corpora without linguistic annotation.

[41] The results of NERC, in particular of Workpackages 4,6,8, have been taken as the basis of the EAGLES Corpus Group work, in particular for the issues of spoken text representation and morphosyntactic annotation.

3.5 *Follow-up of the NERC Study*

3.5.1 *The PAROLE project*

The NERC Report was issued at the end of 1993, and our recommendations have been endorsed by the high-level evaluators appointed by the Commission to review the NERC results. Within the LE area of the Telematics Programme in the 4th FRP (1995-1998), LR is one of the three major lines for which funding, necessary to support the large-scale actions required to begin the creation of adequate LR, has been approved. The Commission has included LR in the 1994 MLAP call, issued to support projects intended to prepare, at the organisational and technical levels, the activities of LE in the 4th FRP. Three projects have been selected in answer to this call: POINTER for terminological LR, SPEECHDAT for speech LR, PAROLE for written LR.

The MLAP project PAROLE has been proposed by a Consortium which essentially consists, with a few exceptions, in the union of the NERC1 and NERC2 partners, and is co-ordinated by SITEC (Munich) and the Consorzio Pisa Ricerche (Pisa).

The main goal of MLAP-PAROLE, which is just now coming to an end, is to create, using as its main input the NERC recommendations, the organisational conditions and to provide the technical specifications necessary to start and carry on the work requested to construct and make available, to the European R&D communities, a set of reusable reference corpora and generic computational lexicons harmonised for all the European languages.

MLAP-PAROLE has begun formation of the two layers of the European infrastructure for the production of written LR, already anticipated in the initial chapter of the NERC report. The first layer has been realised through the setting-up of the non-profit European Association PAROLE which at present includes, as funding members, 14 focal points, one for each of the following languages: Portuguese, Spanish, Irish, English, Catalan, French, Italian, Greek, German, Dutch, Belgian French, Swedish, Finnish, Danish. The focal points have been selected on the basis of the criteria listed in Chapter 0 of this report. The task of the focal points is to ensure:

- continuity in those activities relating to the creation, maintenance and updating of WLR,
- co-ordination and harmonisation, at the international level, of the above activities,
- provision of services and consultancy in the field of WLR, for both the Research and Development communities.

Each of the focal points has begun to organise a language-specific network which will constitute the second layer of the European infrastructure.

MLAP-PAROLE has issued detailed technical specifications regarding the content, the coverage, the format and the encoding conventions of the envisaged lexicons and corpora, so as to ensure their harmonisation and interoperability. The work in PAROLE is based on the results available from EAGLES. The EAGLES initiative, whose work must be viewed from a long-term perspective, aims specifically at defining standards and common specifications in preparing the ground for future standard provision. The EAGLES Working Groups have provided guidelines for the following aspects:

- Corpus and text typology: Any corpus and its constituent texts must be classified and typed to become really useful. A preliminary proposal for an agreed set of parameters for classifying and typing corpora and texts is available in *Corpus Typology* (Sinclair 1994). This work is explicitly intended as a follow-up to Chapter 2 of the NERC report.

A common encoding corpus standard: Following the recommendations of NERC, EAGLES has issued detailed recommendations regarding the adaptation of the TEI Guidelines to the specific needs of corpora intended to support LE research and applications (Ide & Veronis, *Corpus Encoding Standard,* 1996). The TEI Guidelines offer a rich choice of alternatives for encoding, thus part of the work is aimed at producing a specialised solution that will be the default for corpus encoding. Three levels of mark-up have been distinguished for text corpora, ranging from the most gross type of mark-up (situating the document within a corpus), through to an increasingly refined structural mark-up, to a mark-up for linguistic annotation. This work was carried out in close co-operation with MULTEXT.

Morphosyntactic and syntactic information in lexicon encoding and in corpora annotation: EAGLES has been working towards a general framework for annotation that will ensure the compatibility and interchangeability of concrete annotation schemata based on it. The annotation framework also allows for extension to language-specific phenomena and for variation in the degree of granularity of annotation.

For morphosyntax, the work of EAGLES directly builds on the results of NERC, in particular those reported in Chapter 8. Four degrees of constraints are proposed in the description of word categories by means of morphosyntactic tags, ranging from obligatory specification of the major parts of speech, through widely-recognised features and generic features, to language-specific features (Leech & Wilson, 1994 and Monachini & Calzolari, 1996).[42]

- Other areas of work within EAGLES are tools for corpus work, parallel corpora and spoken texts.

The last area also builds explicitly on the NERC results.

While EAGLES provides, in general, a global framework for standardisation, PAROLE translates these broad guidelines into functional and operational specifications. The PAROLE reports will have defined, for lexicons, corpora and related tools, the minimal level of standardisation which a multilingual project aiming at building very large harmonised written language resources should achieve.

MLAP-PAROLE has provided a detailed workplan for producing and making available an initial set of written LR (reference corpora and generic lexicons) for all the European languages. This workplan has been submitted as a proposal to the second LE call, issued in March 1995 and dedicated to LR. The proposal was accepted (LE-PAROLE) and the project will commence in the first half of 1996, with the goal of creating a set of comparable corpora whose minimum size will be 20 million running words including all the languages of the current PAROLE Association. In addition, a set of computational lexicons, at minimum 20,000 lexical entries, encoded at the morphological, morphosyntactic, and syntactic levels for all the languages except Irish and Belgian French, will also be included.

The LE-PAROLE Consortium will use the services of ELRA in making the corpora and lexica widely available. The ELRA members are organised into three colleges: WLR, speech resources, terminology. The PAROLE members have formed the nucleus of the WLR college,

[42] Preliminary EAGLES recommendations for syntactic annotation are also available, focussing on surface syntactic bracketting (Leech, Barnett & Kahrel, *Guidelines for the Syntactic Annotation of Corpora*, 1996).

and will actively co-operate on a study for the preparation of an ELRA manual for the validation of WLR. This is a difficult task: contrary to the field of speech, up until now no systematic attempts have been made to design a methodology for the validation of WLR. When available, the results of this study will constitute an essential complement to the NERC study.

3.5.2 *The TELRI Copernicus project*

The TELRI (Trans-European Language Resources Infrastructure) project, proposed and co-ordinated by W. Teubert of IDS, has been recently approved in the framework of COPERNICUS, a program of the EC aiming at promoting and reinforcing co-operation in the field of research and development between the countries of the EU and Eastern European countries. The main goals of the TELRI project, which includes the NERC1 partners and 14 partners from East European countries, is to complement the European infrastructure for written LR, establishing appropriate links with relevant language and language technology centres in the eastern countries.

4 Conclusion

Lets conclude this Introduction with the hope that the process started at the 1986 Grosseto workshop, and in particular the ideas brought forward by the NERC Consortia, will finally, after so many definition and preparatory actions, give way to the creation of the LR necessary for the various European languages. The actions launched by the Commission of the EU should constitute the basis for a co-ordinated series of activities at the European level. Associations like ELRA and PAROLE should contribute towards promoting the involvement of the national authorities in support of their languages, and towards ensuring the necessary continuity.

In my opinion, it is very important to promote co-ordination and co-operation also at the international level, between countries of different continents. Strong competition exists between different economic blocks for the development of applications and services based on human language technology. The availability of LR in a given language is an initial factor necessary for the successful development of technologies and products for these languages. So, to make LR for a given language available to potential economic and industrial competitors is a critical decision.

Looked at from another point of view, the development of a true multilingual information society is in the best interests not only of all economies, but also of all cultures since it helps to preserve the vehicular function and the cultural identity of each language.

The availability of LR, co-ordinated and harmonised for all the languages, is thus an issue of strategic relevance. For speech, the R&D community has already organised itself through the setting up of the COCOSDA initiative. For WLR, we have taken the initial steps along these lines. At the occasion of the international workshop on corpora, organised in Pisa in 1992 in the framework of the NERC study, to which official representatives and project' representatives of ARPA and NSF took part, we proposed the founding of LIRIC,[43] an initiative which should

[43] Linguistic Resources for International Cooperation.

constitute the equivalent of COCOSDA for written languages.

5 Acknowledgements

Let us also express our appreciation to the EU officers who, on various occasions and at different times, have supported our initiatives for the promotion of the field of LR in Europe: L. Rolling, S. Perschke, R. Cencioni, N. Brinkhoff-Button, N. Varile, J. Soler, J. Roukens of the DG-XIII in Luxembourg; their directors: V. Parajon-Collada, Deputy Director General of the DG-XIII, and F. de Bruine, Director of the DG-XIII E; S. Bensasson, D. Cornwell, M. Bosco, K. Varghese, M. Boukerche of the ESPRIT program in Brussels.

References

Aarts, J., and Meijs, W. (eds., 1986): *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam: Rodopi.

Accademia Nazionale dei Lincei (1975): *Colloquio sul tema: Tecniche di classificazione e loro applicazione linguistica (Florence, 1972)*, Rome.

Actes du Colloque International sur la Mécanisation des Recherches Lexicologiques, Besançon 1961, *Cahiers de Lexicologie*, 3, 1961.

Aijmer, K., and Altenberg, B. (1991): *English Corpus Linguistics: Studies in Honour of Svartvik*, London & New York: Longman.

Almanacco Letterario Bompiani 1961, Milano, 1961.

ALPAC Report (1966): *Language and Machine: Computers in Translation and Linguistics*, Washington, DC: National Research Council. Automatic Language Processing Advisory Committee.

Antoni-Lay, M.H., Francopoulo G., and Zaysser, L. (1994): "A Generic Model for Reusable Lexicons: The Genelex Project", *Literary and Linguistic Computing*, 9, 1, 47-54.

Atkins, B.T.S., and Zampolli, A. (1994): *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.

Atkins, B.T.S., Levin, B., and Zampolli, A. (1994): *Computational Approaches to the Lexicon: An Overview*, in Atkins, B.T.S., and Zampolli, A., (eds., 1994), 17-45.

Bazell, C., Catford, J., Halliday, M.K., and Robins, R.H. (eds., 1986): *In Memory of J.R. Firth*, London: Longman.

Biber, D., and Finegan, E. (1991): *On the exploitation of computerized corpora in variation studies*, in Aijmer, K., and Altenber, B. (eds., 1991), 204-220.

Bindi, R., Monachini, M., and Orsolini, P. (1989): *Italian reference Corpus*, Pisa: Istituto di Linguistica Computazionale.

Boguraev, B., Briscoe, T., Calzolari, N., Cater, A., Meijs, W., and Zampolli, A. (1988): *Acquisition of Lexical Knowledge for Natural Language Processing Systems*, Acquilex Technical Annex, ESPRIT Basic Research Action No. 3030 (unpublished MS).

Boguraev, B., and Briscoe, T. (1989): *Computational Lexicography for Natural Language Processing*, London & New York: Longman.

Booth, A.D., Cleave, J.P., and Brandwood, B.A. (1958): *Mechanical Resolution of Linguistic Problems*, London: Butterworths Scientific Publications.

Bortolini, U., Tagliavini, C., and Zampolli, A. (1971): *Lessico di Frequenza della Lingua Italiana Contemporanea*, IBM Italia.

Busa, R. (1951): "Sancti Thomas Aquinatis hymnorum ritualium: Varia specimina concordantiarum", in *Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milan: Bocca.

Busa, R. (1962): "L'analisi linguistica nell'evoluzione mondiale dei mezzi di informazione", *Almanacco Letterario Bompiani 1962*, 103-117.

Busa, R. (ed., 1968): *Actes du Séminaire International sur le dictionnaire latin de machine, Calcolo*. Supplemento n.2 al vol.V.

Calzolari, N., and Picchi, E. (1994): *A Lexical Workstation: from Testual Data to Structured Database*, in Atkins, B.T.S., and Zampolli, A. (eds., 1994), 439-467.

Calzolari, N., and McNaught, J. (1994): *Editors' Introduction*, EAGLES document EAG-EB-IR-2.

Cignoni, L., Peters, C., and Rossi, S. (1983): *European Science Foundation Survey of Lexicographical Projects*, Pisa: Istituto di Linguistica Computazionale.

Cumming, S. (1995): *The Lexicon in Text Generation: Progress and Prospects*, in Walker, D. et al., (1995), 171-206.

Danzin, A. (1992): Groupe de réflexion stratégique pour la Commission des Communautés Européennes (DG XIII), *Vers une infrastructure linguistique européenne*, Document available from DG XIII-E, Luxembourg.

Dyer, R.R. (1973): *The Measurement of Individual Style*, in Zampolli, A. (ed., 1973), 325-348.

"EAGLES Update", *Elsnews Bulletin*, 2, 1, 1993.

Estoup (1907): *Gammes Sténographiques*, Paris.

European Commission DG-XIII (1994): Telematics Applications Programme (1994-1998), Work Programme, Luxembourg.

Fattori, M., Bianchi, M. (eds., 1976): *I° Colloquio Internazionale del Lessico Intellettuale Europeo*, Rome: Edizioni dell'Ateneo.

Fourcin, A., and Gibbon, D. (1993): "Spoken Language Assessment in the European Context", *Literary and Linguistic Computing*, 9, 1, 79-86.

Francis, W., and Kucera, H. (1964; revised 1971 and 1979): *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers.* Providence, R.I.: Department of Linguistics, Brown University.

Guiraud, P. (1954): *Bibliographie critique de la statistique linguistique*, Utrecht-Anvers: Mouton.

Guiraud, P. (1960): *Problèmes et méthodes de la statistique linguistique*, Paris: P.U.F.

Hays, D.G. (1967): *Introduction to Computational Linguistics*, New York: American Elsevier.

Hays, D.G. (1969): *Computational Linguistics: Introduction*, in Meetham, A.R., and Hudson, R.A. (eds., 1969), 49-51.

Hays, D.G. (1976): *The field and scope of computational linguistics*, in Papp, F., and Szépe, G. (eds., 1976), 21-25.

Halliday, M.K. (1986): *Lexis as a linguistic level*, in Bazell, C., Catford, J., Halliday, M.K., and Robins, R.H. (eds., 1986).

Harkin, D. (1957): "The History of Word Counts", *Babel*, 3, 113-124.

Heid, H., and McNaught, J. (eds., 1991): "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", *Eurotra-7 Final Report*, Stuttgart.

Heilmann, L. (1963): "Considerazioni statistico-matematiche e contenuto semantico", *Quaderni dell'Istituto di Glottologia VII*, Bologna: Università di Bologna, 34-45.

Herdan, G. (1964): "Quantitative Linguistics or Generative Grammar", *Linguistics*, 4, 56-65.

IBM (1964): *Literary Data Processing Conference Proceedings, September 9, 10, 11 1964*, Washington.

Ide, N., and Veronis, J. (1996): *Corpus Encoding Standard*, EAGLES document EAG-CWG/CES.

Ingria, J.P. (1995): *Lexical Information for Parsing Systems: Points of Convergence and Divergence* in Walker, D. et al., (1995), 93-170.

Johansson, S., Leech, G.N., and Goodluck, H. (1978): *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Oslo: Department of English, University of Oslo, 25-36.

Johansson, S. (1980): "The LOB Corpus of British English Texts: Presentation and Comments", *ALLC Journal I.*

Juilland, A., Edwards, P.M.H., and Juilland, I. (1965): *Frequency dictionary of rumanian words*, The Hague: Mouton.

Juilland, A., and Traversa, V. (1973): *Frequency dictionary of italian words*, The Hague: Mouton.

Kaeding, F.W. (1898): *Häufigkeitswörterbuch der deutschen Sprache*, Berlin: Steglitz.

Kalgress, H. (1973): *Foreword*, in Zampolli, A., Calzolari, N. (1973), xiii-xiv.

Katzen, M. (ed., 1991): *Scholarship and Technology in the Humanities*, London: British Library Research.

Kay, M. (1964): *Report of Informal Meeting on Standards Formats for Machine-Readable Texts*, in IBM (1964), 327-328.

Kay, M. (1967): Standards for Encoding Data in Natural Language, *Computers and Humanities*, I, 5, 170-177.

Kay, M. (1983): *The Dictionary of the Future and the Future of the Dictionary*, in Zampolli, A., and Cappelli, A. (eds., 1983), 161-174.

Khatchadourian, H., and Modiano, N. (1994): "Use and Importance of Standards in Electronic Dictionaries: The Compilation Approach for Lexical Resources", *Literary and Linguistic Computing*, 9, 1, 55-64.

Kucera, H., and Francis, W.N. (1967): *Computational Analysis of Present-Day American English*, Providence: Brown University Press.

Leech, G. (ed., 1990): *Proceedings of a Workshop on Corpus Resources, Oxford, January, 1990,* London: DTI Speech and Language Technology Club.

Leech, G. (1991): *The state of the art in Corpus Linguistics*, in Aijmer K., Altenberg B. (eds., 1991), 8-29.

Leech, G., and Wilson, A. (1994): *Morphosyntactic Annotation*, EAGLES document EAG-CWG/Annotate.

Les Machines dans la Linguistique (1968): Prague.

Llisterri, J. (1994): *Spoken Texts*, EAGLES document EAG-CWG/Spokentx.

Locke, W.N., and Booth, A.D. (1955): *Machine Translation of Languages*, New York: The Technology Press of the MIT-J. Wiley & Sons, Inc.

Maegaard, B. (1988): "EUROTRA, The Machine Translation Project of the European Communities", *Literary and Linguistic Computing*, 3, 2, 61-65.

Malmberg, B. (1966): *Les nouvelles tendances de la Linguistique*, Paris: P.U.F.

Meetham, A.R., and Hudson, R.A. (1969): *Encyclopaedia of Linguistics, Information and Control*, Oxford: Pergamon Press.

Michea, R. (1964): "Les Vocabulaires fondamentaux", in *Recherche et techniques nouvelles au service de l'enseignement des langues vivantes*, Strasbourg: Université de Strasbourg: 21-36.

Monachini, M., and Calzolari, N. (1996): *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, EAGLES document EAG-LWG/Morphsyn.

Moreau, R. (1962): "Au sujet de l'utilisation de la notion de fréquence en linguistique", *Cahiers de Lexicologie*, 3, 140-158.

Muller, Ch. (1964): *Initiation à la Statistique Linguistique*, Paris: Larousse.

Nagao, M. (1989): *Machine Translation: How Far can it Go?,* Oxford: Oxford University Press.

NERC Final Report, (1993): Pisa: Istituto di Linguistica Computazionale.

Papp, F., and Szépe, G. (eds., 1976): "Papers in Computational Linguistics", *Proceedings of the 3rd International Meeting on Computational Linguistics, Debrecen, 1971*, Budapest: Akadémiai Kiado.

NERC2 (1993): Report, Pisa (unpublished).

Quemada, B. (1961): "Introduction", *Cahiers de Lexicologie*, 3, 1, 13-18.

Quemada, B. (1987): "Notes sur la lexicographie et la dictionnairique", *Cahiers de Lexicologie*, 51, 2, 229-242.

"RELATing to Resources" (1993): *Elsnews Bulletin*, 2, 5.

Rowald, P. (1914): *Repertorium lateinischer Wörterzeichnisse und Speziallexica*, Leipzig.

Schöne, H. (1907): *Repertorium griechisher Wörterzeichnisse und Speziallexica*, Leipzig.

Sinclair, J.M. (ed., 1987): *Looking-up: An Account of the COBUILD Project*, London: Collins.

Sinclair, J.M., et al. (1987): *Cobuild Dictionary of English Language*, London: Collins.

Sinclair, J.M. (1994): *Corpus Typology*, EAGLES document EAG-CWG/Corptyp.

Sinclair, J.M., Hoelter, M., and Peters, C. (eds., 1995): *The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing*, Studies in Machine Translation and Natural Language Processing, Brussels-Luxembourg: Office for Official Publications of the European Communities.

Smith, J. (1973): *Ideals Versus Practicalities in Linguistic Data Processing*, in Zampolli, A., and Calzolari, N. (eds., 1973), V, II. 2, 895-8.

Sperberg-McQueen, C.M., and Burnard, L. (eds., 1990), *Guidelines for the Encoding of Machine-Readable Texts for Interchange*, Chicago: ACL-ACH-ALLC Text Encoding Initiative.

Table ronde sur les grands dictionnaires historiques (1973): Florence: Olschki Editore.

The NERC 2 Final Report, (1993): Pisa (Document Presented to the EC).

Trubetzkoy, N.S. (1968): "Grundzüge der Phonologie", *Travaux du Circle Linguistique de Prague*, 7.

Varile, G.B., and Zampolli, A. (eds., 1992), *Synopses of American, European and Japanese Projects Presented at the International Projects Day at COLING 1992*, (Linguistica Computazionale VIII), Pisa: Giardini Editori.

Varile, G.B., and Zampolli, A. (forthcoming): *Survey of the State of the Art in written and Spoken Language Processing*. Sponsored by the Commission of the European Union and the National Science Foundation of the USA, (Linguistica Computazionale) (forthcoming).

Vauquois, B. (1975): *La Traduction automatique à Grenoble*, Paris: Dunod.

Vidal-Beneyto, J. (1986): "Presentation", in *Encrages*, special issue dedicated to *Les Industries de la langue enjeux pour l'Europe: Actes du colloque de Tours*, 5-7, Saint-Denis: Université Paris VIII-Vincennes, i-vii.

Vidal-Beneyto, J. (ed., 1991): *Las industrias de la lengue* , Madrid: Ediciones Piramide.

Walker, D., and Zampolli, A., "Foreword", in Boguraev, B., and Briscoe, T. (eds., 1989), xiii-xiv.

Walker, D., Zampolli, A., and Calzolari, N. (eds., 1987): *Towards a Polytheoretical Lexical Data Base*, Pisa: Istituto di Linguistica Computazionale.

Walker, D., Zampolli, A., and Calzolari, N. (eds., 1995): *On Automating the Lexicon. Research and Practice in a Multilingual Environment*, Proceedings of a Workshop held in Grosseto, Oxford: Oxford University Press.

Weiner, E. (1994): *The Lexicographic Workstation and the Scholarly Dictionary*, in Atkins, B.T.S., and Zampolli, A. (eds., 1994), 413-438.

Whitelock, P., Wood, M., Somers, H., Johnson, E., and Bennett, P. (eds., 1987): *Linguistic Theory and Computer Applications*, New York: Academic Press.

Zampolli, A. (1968): *Projet pour un lexique électronique de l'italien*, in Busa, R. (ed., 1968), 109-126.

Zampolli, A. (1968): *Projet d'un dictionnaire italien de machine-intervention*, in Busa, R. (ed., 1968), 109-126.

Zampolli, A. (1970): "Cronaca: La terza International Conference on Computational Linguistics", Stoccolma, 1969, *Archivio Glottologico Italiano*, LV, 1-2, 272-279.

Zampolli, A. (1973): "Humanities Computing in Italy", *Computers and the Humanities*, 7, 6, 343-360.

Zampolli, A. (ed., 1973): *Linguistica Matematica e Calcolatori*, Atti del Convegno e della Prima Scuola Internazionale, Pisa, 16/VIII-6/IX 1970, Florence: Olschki.

Zampolli, A. (1973): *L'Automatisation de la recherche lexicologique: Etat actuel et tendances nouvelles*, META 18, 1-2, 101-136.

Zampolli, A. (1973): *Introduction*, in Zampolli A., Calzolari, N. (eds., 1973-7), xix, xxviii.

Zampolli, A., and Calzolari, N. (eds., 1973-7): *Computational and Mathematical Linguistics*, *Proceedings of the International Conference on Computational Linguistics 1973*, Florence: Olschki.

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

Zampolli, A. (1976): *Les dépouillements électroniques quelques problemès de méthode et d'organisation*, in Fattori, M., Bianchi, M. (eds., 1976), 173-178.

Zampolli, A. (ed., 1977): *Linguistic Structures Processing*, Amsterdam: North-Holland.

Zampolli, A., and Quemada, B. (1981): *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries*, Final Report on the ESF Pisa Workshop, presented to the ESF (unpublished).

Zampolli, A. (1983): *Lexicological and Lexicographical Activities at the Istituto di Linguistica Computazionale*, in Zampolli, A., and Cappelli, A. (eds., 1983), 237-278.

Zampolli, A., and Cappelli A. (eds., 1983): *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries (Proceedings of the European Science Foundation Workshop, Pisa, 1981)*, (Linguistica Computazionale III), Pisa: Giardini Editori.

Zampolli, A., and Walker, D. (1986): *Multilingual Lexicology and Lexicography: New Directions*, paper presented at the CG12-TSC of the EC (unpublished).

Zampolli, A. (1987): "Perspectives for an Italian Multifunctional Lexical Database", in Zampolli, A., Cappelli, A., Cignoni, L., and Peters, C. (eds., 1987), 304-41.

Zampolli, A., Cappelli, A., Cignoni, L., and Peters, C. (eds., 1987), *Studies in Honour of Roberto Busa S.J.*, (Linguistica Computazionale IV-V), Pisa: Giardini Editori.

Zampolli, A., and Walker, D. (1987): *Report on the Grosseto Workshop*, Pisa (unpublished).

Zampolli, A. with consultation of Spang-Hansen, H., Perschke, S., Cerdà, R. (1988): *Reusability of lexical resources*, Document CG12/159/88/b presented to the CGC-12 (Comité de Gestion et Consultation), Luxembourg (unpublished paper).

Zampolli, A. (1989): "Introduction to the Special Section on Machine Translation", *Literary and Linguistic Computing*, 4, 3, 182-184.

Zampolli, A. (1991): *Technology and Linguistic Research*, in Katzen, M. (ed., 1991), 21-51.

Zampolli, A. (1991): *Preliminary Considerations on the Constitution of an ELTA (European Language Technology Agency)*, Pisa, Document prepared for the DG XIII.

Zampolli, A. (1991): *Corpora de Referencia*, in Vidal-Beneyto, J. (ed., 1991), 119-124.

Zampolli, A. (1991): *Bases multifuncionales de datos léxicos*, in Vidal-Beneyto, J. (ed., 1991), 127-146.

Zampolli, A. (1994): *Introduction*, in Atkins B.T.S., and Zampolli, A. (eds., 1994), 3-15.

Zampolli, A., Calzolari, N., Palmer, M. (eds., 1994): *Current Issues in Computational Linguistics: In Honour of Don Walker*, (Linguistica Computazionale IX-X), Pisa: Giardini Editori; Dordrecht: Kluwer.

Zampolli, A. (1994): *Preface*, in Zampolli, A., Calzolari, N., and Palmer, M. (eds., 1994), ix-xv.

Zipf, G.K. (1935): *The Psycho-biology of Language: An Introduction to Dynamic Biology*, Cambridge, Mass.: MIT Press.