INTRODUCTION

"Linguistica Computazionale" is very pleased to publish a selection of the papers presented at the VII International Symposium of the Association for Literary and Linguistic Computing. The majority of the presentations were concerned with computer-aided lexicography, statistical linguistics, text-processing tools and software.

As it has been impossible to publish all of the communications presented at the meeting, I should like to summarize here, in brief, the trends which I feel emerged during the Symposium as a whole.

The use of the computer in lexicology and lexicography was the subject of many of the papers and, in particular, of the invited paper by B. Quemada, Director of the Institut de la Langue Française, Nancy (*).

It seems unthinkable nowadays to start a lexicographical project without first considering the possibility of using the computer. The most recent technical developments, together with the latest advances in computational linguistics, are profoundly modifying the traditional relationships between linguistics, lexicology and lexicography. Furthermore, in addition to the traditional printed dictionary, new forms of lexicographical tools now seem desirable. Different methodologies, presented at the Symposium mainly by researchers working on large historical dictionaries, are now available to extract basic documentation, which can be used for dictionary compilation, from machine readable textual archives.

The Symposium also evidenced the gradual development of the machine dictionaries (MD).

The first MDs were produced as components of earlier natural language processing systems (automatic translation, information retrieval, etc.). They normally concerned lexical subsets restricted to the specific technical domain of these systems.

A set of independent but concomitant factors, both theoretical (e.g. the lexicalist trend of certain transformational-generative schools) and applicative (e.g. the growing popularity of data bases, telematics, etc.) have affected both the basic conception and the principle objectives of the MD. In fact, the machine dictionary nowadays is increasingly seen as a particularly convenient way to dynamically represent tendentially exhaustive sets of linguistic knowledge on which studies can be made which would otherwise be impossible.

The papers presented at the Symposium included: descriptions of existing MDs; methods used to compare different MDs available for the same language and attempts to unify their information; methods to access the information explicitly or implicitly represented in a machine dictionary; MDs as data bases for encyclopedic, lexical, grammatical and orthographical information, accessible to the general public through computing networks; MDs viewed as the

---

* The illuminating classification of activities in progress, oriented towards future prospects in this sector, presented by B. Quemada, appeared in Vol. 3 of this journal.

computational representation of the lexical system of a language (studies on derivatives, idioms, the verbal frame, collocations, etc.).

"Computer-Aided Lexicology, Lexicography and Terminology" was the topic of a working group attended by about a hundred participants. From the discussion a clear trend emerged in favour of the creation of linguistic data bases conceived as organized sets of texts, dictionaries, parsers and specialized software. These data bases can be used in different ways: for the production of indexes, concordances and other types of lexical documentation, in a variety of forms and using different publication techniques (photocomposition, microfiche, etc.); for the editing of short or long-term, general-interest or specialized dictionaries; for data query and retrieval services which hopefully will soon be available to a wider public in view of the recent rapid developments in communications and networking, etc. In this way, contacts and exchanges among researchers from different environments would be stimulated. Scholars would be able to access the data base material for lexicological, philological and linguistic studies, in return "depositing" their own results or analyses which will thus be available for other researchers.

The creation of data bases of this type raises the problem of copyright; a problem in which various factors play different roles. These include the "owners" of the original material (printed texts and dictionaries); those responsible for its electronic processing and insertion into the data base; the various categories of public and private users; exchanges with other data bases.

The conception of such data bases as a new kind of public "library", the demand for data exchange, and other reasons of a more purely scientific nature, contribute to encouraging discussions on the problem of standards. The adoption of generalized standards seems at the present moment to be impossible and maybe not even completely desirable. It would nevertheless be possible to define a methodological framework of reference, and, in the first place, a "metalanguage" to describe, clearly and univocally, the linguistic material stored in a data base, its structure, and the various analyses performed at different linguistic levels.

Many of the papers at the Symposium (almost half the total) illustrated studies in the sector of statistical linguistics. In the 50s and 60s the complexity of linguistic mechanisms and extra-linguistic factors which determine the distribution of the frequency of occurrences of the linguistic units was underestimated. It was felt that the observations made up to that moment authorized generalizations ("laws") at the theoretical level, and applications at the service of other disciplines (e.g. in philology, in studies on relative chronology and authenticity, attribution, etc.). The statistical apparatus was for the most part correct and sophisticated, but the conclusions were soon found to be contradictory and unreliable owing to the lack of an adequate linguistic model which could be used to interpret the data.

The programme of work which emerged from the communications presented at the Symposium could be outlined as follows. The quantitative behaviour (i.e. stability or variations in the distribution of frequencies) of a large number of linguistic units and structures at different levels (phonological, morphological, syntactic, ...) should be observed in texts produced in widely varying communicative contexts. A major aim is to identify, where possible, relationships between certain "production factors" ("literary genres", target audiences, early and late works, etc.) and the stability and variation in the frequency of the linguistic units. The next step could consist of the formulation of explanatory hypotheses to discover the nature and the motivations of these relations.

This programme is clearly greatly facilitated by the availability of the large text corpora already stored in the existing archives (Trésor de la Langue Française, Thesaurus Linguae Graecae, ICAME, ILC, etc.). Statistical

processing of quantitative data produced by analyses of these large corpora, illustrated during the Symposium, has already produced evidence of previously unknown phenomena. For example, the presence of surprising trends in the distribution of frequencies in diachronically stratified corpora.

A number of fundamental problems clearly emerged from the discussions. A model for the interpretation of quantitative data must be inserted into a more general model of linguistic communication, in which the communicative contexts and their relationship with the execution ("performance" in the generative sense) must be identified. On the other hand, studies both of large corpora (undertaken by important philological or lexicographical projects) or of single texts or authors (by individual researchers) are usually bound to the lexical or morphological level. Quantitative analyses at other linguistic levels, especially syntactic and semantic, are extremely rare. This situation, discussed on several occasions at the Symposium, must be attributed not only to the state of present-day linguistic theories, which are, in a certain sense, inadequate for the description of large texts and corpora, but also to the enormous amount of work necessary, in the absence of suitable parsers, for manual analyses at the syntactic and semantic level.

Some communications examined the role of the receiver (reader-listener), and a comparison could be attempted with similar trends emerging from the so-called computational-cognitive paradigm. Psycholinguistic experiments (tests, questionnaires, etc.) aim at eliciting the opinion of the receiver on the correlation between the frequency, in the text, of certain linguistic phenomena and the effect perceived at the stylistic level. These data are integrated by quantitative text analyses including frequency dictionaries, which are apparently regaining in popularity. The techniques used to evaluate the distribution of frequencies in various subsets of the sample corpus are integrated by an evaluation of the frequency of usage given by the communicative competence of native speakers.

In the extremely lively round table on "Statistical Linguistics, Stylometry, Literary Analysis", two main trends were seen. These were defined by some participants as contrasting and by others as complementary. On one hand, there was the tendency, predominantly of French and Italian scholars, to perform exhaustive analyses on large corpora in machine readable form in order to study the general quantitative structures of natural language. On the other hand, a common trend of the North-European and American school seems to be to search for author and text specific quantitative characteristics, to be used as a complementary tool in the resolution of concrete problems connected, in general, with the characterization of style.

It was clear, however, that in any case this type of research must alternate between qualitative hypotheses and quantitative verifications. Therefore, a continual "dialogue" with the data is essential. In this respect, the importance of the role played by the growing popularity of micro-computers was emphasized more than once.

Part of the debate during the round table on "Software for Text Processing" was dedicated to the impact on literary and linguistic computing of recent technological developments. It is hoped that within the next few years the principal Institutes of a number of countries will be linked by networks. These Institutes have worked in the past or are now working on the production of software for the most commonly requested types of text analyses and processing. They also create and manage extensive textual archives. The possibility exists, thanks to recent developments in different types of mass storage systems, to offer their users on-line access to these large textual corpora on-line.

The advance of telematics means that a potentially vaster public now has access to linguistic data bases. The likelihood of finding a given text already in machine readable form is increasing. The spread of personal and

micro-computers now makes the use of electronic data processing in everyday research work much more feasible.

Although to a much lesser extent than in the past, it seems that a major problem is still how to simplify the approach to computational methodologies for the linguist and philologist. A number of solutions have been proposed and discussed including university courses in computational linguistics and the organization of post-graduate schools (similar to the Pisa summer school). A major help seems to be the availability of generalized software and computational tools designed specifically for philological and linguistic research. This situation seems close to being achieved for text processing (indexes, concordances, etc.). Studies which aim at producing an inventory of available software were described.

In the same way, it can be hypothesized that in the medium-term a researcher wishing to analyse a text will have an already tried and tested machine dictionary available for languages such as Italian, German, Swedish, etc.

The situation for parsers appears different. It has been announced that generalized analyzers, which operate at the syntagmatic level to provide at least a partially automatic disambiguation of homography between different parts of speech, should soon be available for Italian, Spanish, French and German. On the contrary, there are no indications that, in the short term, semantic and syntactic parsers, aiming at assigning a structural description to sentences, will be applicable. However, the progressive acquisition of commonly adopted methodologies has been recorded. For example, ATN has been seen to be generally employed in natural language processing for man-machine communication and data base interrogation.

I must also mention the number of projects presented in the field of historical linguistics, which was to some extent unexpected. These included the collection of data on classical and non Indo-European languages, models of linguistic changes and genetic relationships, etc.

In conclusion, I should like to express my gratitude to all those who contributed to the success of the Symposium. These include of course the Programme Committee and, in particular, its co-ordinator J. Hamesse. I should also like to thank G. Nencioni (President of the Scientific Committee of the ILC and of the Accademia della Crusca), T. Bolelli (representing the Rector of the University of Pisa), J.M. Smith (ALLC Chairman) and A. Garzella (representing the Mayor and Town Council of Pisa) who welcomed the participants at the Opening Ceremony. Our appreciation also goes to all the local authorities who contributed in various ways to the organization of the Symposium.

Finally, particular thanks must go to the staff of the Istituto di Linguistica Computazionale and especially to the Local Organizing Committee of the Symposium: L. Cignoni, V. Parrinelli, C. Peters and S. Rossi.

Antonio Zampolli