A. ZAMPOLLI

INTRODUCTION

FIRENZE
LEO S. OLSCHKI EDITORE
MCMLXXVII

INTRODUCTION

The Time, Place, and Nature of the Conference

The Proceedings consist of two volumes containing the texts of all the papers accepted for the 5th International Conference on Computational Linguistics, which took place at Pisa from 27 August to 1 September 1973 and was held in the *Palazzo della Sapienza*, the center of the University. The conference was sponsored by the *Consiglio Nazionale delle Ricerche* (CNR), the *Centro Nazionale Universitario di Calcolo Elettronico* (CNUCE), the University of Pisa, and the IBM Scientific Center of Pisa.

In his preface, B. Vauquois, President of the International Committee on Computational Linguistics, has described the history of the previous conferences, and H. Karlgren, President of the Program Committee, has explained the principles which were adopted in the choice and the division of the papers. It is my task to describe briefly the organizational aspects of the conference. First of all, however, I should like to make a few observations concerning its content in relation to recent developments in the area of automated language processing (ALP). I use this term rather than *computational linguistics* as it is far more general in its implications, encompassing all studies, theoretical or applied, on the use of computers or computational techniques in the processing of natural language.[1] I consider computational linguistics to be a subset of ALP, a subset to which I shall refer by the abbreviation CL.

[1] See, for example, S. Lamb, *The Digital Computer as an Aid in Linguistics*, «Language», xxxvii (1961) 3, pp. 382-412, also *Linguistic Data Processing*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 1965, pp. 159-188; J. Gardin, *A Typology of the Computer Uses in Anthropology*, *ibid.*, pp. 103-117. P. L. Garvin, *Computer Participation in Linguistic Research*, «Language», xxxviii (1962) 4, pp.385-389; C. A. Montgomery, *The 1969 International Conference on Computational Linguistics: A Progress Report*, «Computers and the Humanities», iv (1970) 3, pp. 193-198; D. G. Hays, *Applied Computational Linguistics*, in G. E. Perren, J. L. M. Trim (eds.), *Applications of Linguistics*, Cambridge University Press, 1971, pp. 64-84, and also *The Field and Scope of Computational Linguistics*, «The Finite String», ix (1972) 9-10, pp. 1-6; R. F. Barnes, Jr, *Computational Linguistics and Linguistics*, in *Research Trends in Computational Linguistics* (Centre for Applied Linguistics, Arlington, Virginia, 1972), pp. 1-6; and, in particular, the chap-

The complementary subset of computational linguistics (CL) is given the abbreviation TP, where TP is formed from the expression *text processing* which designs the nucleus of that subset.[2] Using these abbreviations we can obtain the equation TP = ALP – CL. Obviously, it is not possible to define clearly the boundary lines between CL and TP. CL activities, which are focused on linguistic algorithms, are principally directed towards the study of linguistic models, and in general, towards the formalization, representation, and calculus of linguistic structures. TP activities are mainly concerned with the processing of collections of language data, usually large, very often for purposes of reorganization, extraction, summarization, etc. of some linguistic elements of the text, designated at the 'surface' level, i.e. distinguished by shape or code pattern.

It seems to me that one of the most valuable aspects of the 5th ICCL was that it, for the first time, intentionally, officially, and successfully brought together the experts from these two fields.[3] It could be said that the 5th ICCL underlined the complementary aspects of CL and TP.[4]

ters dedicated to Automated Language Processing in the *Annual Review of Information Science and Technology*, edited by Carlos A. Cuadra; attention is specially drawn to the article by D. Walker in Volume 8 (1973), pp. 67-119, and the forthcoming article by F. J. Damerau.

[2] I take these two terms from the above article by D. Walker which was published about the same time as the 5th ICCL (1973).

[3] See the Call for Papers of the 1973 International Conference on Computational Linguistics.

The lectures will be concerned mainly with the following topics:

a) Informatic systems for the analysis and generation of linguistic structures. These systems can consist either of only a description language or can also contain the exploitation algorithm.

b) Practical experience of automatic treatment with the use of the preceding tools (analysis, generation of natural languages, man-machine communication, question-answering systems, mechanical translation, information retrieval, etc.).

c) Informatic systems for the analysis of texts and of linguistic data.

d) Practical experience of analysis of texts and linguistic data in various fields of linguistics and humanities (historical linguistics, lexicology, lexicography, terminology, quantitative linguistics, stilometry, stylistics, philology, textual criticism, dialectology, translation, popular traditions, documentary analysis of texts, psycholinguistics, etc.).

[4] The choice of topics announced in the Call for Papers for Coling 76, the 6th International Conference on Computational Linguistics, which will take place at Ottawa in the summer of 1976, substantially confirms the timeliness of this broadening of scope.

The following themes have been chosen for sections of the program:

1. General problems and methods of computational linguistics from the linguistic, the mathematical and the computational points of view.
2. Computational semantics I: deductive logic and artificial intelligence systems.

From a theoretical point of view, it must be remembered that many research projects currently in progress in TP are aimed at extracting, from linguistic facts, data and information which constitute the primary material that must be considered in theories and models of CL. At times, information obtained on the statistical and lexical composition of specific corpora is also used in the construction of algorithms and in the choice of working strategies for systems in CL; reference can be made, for example, to the use of statistical methods in several speech understanding systems or in some projects for machine translation. From an operational point of view, typical TP procedures include some crucial operations on the texts or data which are substantially the same as some of those requested from some components of typical systems in CL. Two of the more obvious examples are morphological analysis and the distinguishing of homographs for lemmatization.

The fact that these operations in TP are still performed manually is partly because of the inadequacies of the components of the CL systems in analysing, in a satisfactory manner, the variety and complexity of the texts and data usually processed in TP, but it is also a result of the lack of exchange of information and collaboration among researchers in the two fields. Those who have worked for some time in TP, however, are well aware of the fact that the development of applications according to the 'classic' methods and techniques of the 1950s and 1960s has reached saturation point. If we continue to use current methods, according to the current rules of the game (for example: processing, at a simple graphemic level, millions of running words, in order to produce frequency counts, concordances, lexical cards, etc., without any linguistic analysis), real prospects of development do not exist. Although the speed of the computer is continually being increased and programs are becoming more sophisticated, lexicographers and linguists are not able to profit from these facts proportionally because current methodology already produces much more data than any rea-

[3] Computational semantics II: semantics of natural languages.

[4] Automatic syntactic parsing and synthesis of natural languages. Information retrieval. Man-machine communication.

[5] Computational lexicography and stylistics, including concordances and statistical studies.

[6] Speech recognition and synthesis; graphemics, including character recognition; language-graphics interfaces.

[7] Machine translation and machine-aided translation; automated terminology dictionaries.

sonably sized team of linguists could possibly analyse, working according to current procedures. If the analysing operations are left to a successive phase, this would not alleviate the problem as it is not clear how we can resolve the enormous operational difficulties which are due to the sheer quantity of the documentation and material gathered.

It is a fact that after the 2nd International Summer School at Pisa (1972) where the participants were exposed at the same time to courses on both computer-aided lexicography and parser principles and techniques, attempts were multiplied to use the methods and techniques, studied by CL, in the field of TP, obviously with all the precautions, simplifications, and the limits imposed by the extent of the materials to be processed. An example which can be cited is the system for the automatic distinguishing, in large corpora, of homographs among the various parts of speech.

It is significant that this new attitude is shown above all in the field of the analysis of texts for lexicographical, statistical, stylistic, and literary studies. It is seen that in many of the papers in the sections of these Proceedings on *text corpus editing*, *text comparison*, and *grammatical analysis*, there is evidence of developments which would have been difficult to have foreseen only two or three years ago.

This trend can easily be detected also in the papers in the section on *meaning extraction*. It is well known that the difference between automatic translation and automatic documentation is that, in the latter, simple and well known surface working methods are able to offer an acceptable rendering, even if the process of textual analysis does not produce a linguistically adequate representation of the syntactic and semantic structure of sentences and paragraphs. Nevertheless, it seems clear that the possibility of applying linguistic knowledge to extract and represent, at least locally and partially, the structure and/or meaning of clauses and sentences, makes it possible to increase the efficiency of algorithms and retrieval strategies.

There is no doubt, however, that the natural meeting point for CL and TP is that represented by the *lexicology* section. On the one hand, the linguistic schools have recently formulated new abstract lexical models, thus enriching the already outstanding tradition of lexicology, particularly strong in Europe. On the other hand, in lexicography, attempts are being made to clarify the theoretical premises with the declared intention of transforming lexicography from an 'art' or 'technique' to an 'applied science'. Projects in which great machine dictionaries are planned should be considered within such a framework. A machine dictionary is seen as an archive which is suitable for registering the body of knowledge produced by linguistic research in its dynamic development. This knowledge can be incorporated, for example, under the form of syntactic/semantic features, relationships of synonymy and antonymy, codes of semantic fields, case structures, possible constructions, rules of selection and co-occurrence, and must be able to be modified allowing both for the evolution of linguistic theories and the incorporation of additional data coming from the corpus. The frequencies of various linguistic units within the texts are recorded in the machine dictionary. These frequencies should permit an attempt to set up, on an inductive basis, the identification on the diachronic and synchronic axes of subsets of texts of quantitative homogeneous characteristics. Only in this way would it be possible to establish on a sounder basis the field of linguistic statistics which is to some extent in a state of theoretical crisis today despite the advances in technology.

With regard to *grammatical analysis*, automatic morphological analysis seems to have reached an advanced stage of development both in the optimization of algorithms and the extension of the lexicon covered in various languages.

As far as syntax is concerned, parsing systems are influenced by the general situation of transformational grammars. As is well known, a universally accepted model does not exist, and transformationalists usually limit themselves to the consideration of restricted even if highly complex groups of phenomena, each one proposing *ad hoc* modifications to the theory instead of clearly defining and completely specifying a class of grammars. The studies on augmented transition networks are probably the most relevant contribution in this field of CL, even at the theoretical level.

The situation is more favourable in a certain sense, however, for the development of algorithms which automatically generate sequences in the language of a given grammar from initial structures. These are useful both in verifying the generative power of a grammar and in the debugging of errors in the writing of a set of rules, as described in the section on *testing and simulation*.

The papers in the section on the *study of formal properties* show how these studies have been gradually widened to cover all the various levels of linguistics: phonological, morphological, syntactic, and semantic.

In the section on *discovery procedures*, the well known debate on the possibility of automating discovery procedures in linguistics was not emphasized as on other occasions. Instead, the papers give concrete examples of computational procedures which can be of great assistance at the heuristic level of research projects.

In the *translation* section, the renewed theoretical involvement of the teams which continue to work in this much discussed field is clearly shown.

As I have already stated, the meeting of those carrying out research in the fields of CL and TP was one notable feature of the conference; the other was the emphasis placed on the area of semantics which was the theme of the two invited papers and also appeared as a topic in many papers, even those not included in the section on *semantical calculus*. The interest in semantic and pragmatic studies has been characteristic of the principal schools of theoretical linguistics in the early 1970s. The discussions on the problems of constructing and modelling 'intelligent' natural language understanding systems by those working in the fields of cognitive psychology, linguistics, information science, and artificial intelligence, have led to exciting new developments in our understanding of the *faculté de langage*. The 5th ICCL has probably provided the first official forum for the description of such researches which produces a new set of tools and knowledge in the field of natural inference, semantics formalization, knowledge representation, memory models, etc. Such a conceptual framework emerging into CL from the field of artificial intelligence is beginning to have an influence on the theoretical positions of the schools of contemporary linguistics; influence which seems to be much stronger and more profound than that of CL in the past. In one sense it is possible to say that, at least as far as this field is concerned, the traditional relationship between linguistics and CL, which sees CL as applying already formed linguistic theories, or at most attempting to give a complete specification of them, is being changed: linguistics is receiving new ideas and suggestions from developments in CL and artificial intelligence.

Organization of the Conference

Participants

There were 328 scholars from 34 countries who participated in the Conference. Some of these countries were represented for the first time at an ICCL meeting. Participants were from the following countries:

Australia 1, Austria 2, Belgium 12, Bulgaria 3, Canada 16, Czechoslovakia 5, Denmark 6, Finland 3, France 43, Germany 66, Greece 1, Holland 9, Hong Kong 1, Hungary 7, India 2, Israel 2, Italy 61, Ivory Coast 1, Japan 3, Norway 6, Poland 3, Portugal 2, Republic of Zaire 1, Rumania 2, South Africa 1, Spain 6, Sweden 8, Switzerland 5, Turkey 1, Uruguay 1, United Kingdom 15, USA 25, USSR 5, and Yugoslavia 3. Attendance was nearly double that of previous meetings.

Program

The texts received in response to the Call for Papers published in February 1973 were submitted to the scrutiny of the Program Committee (see p. V) which accepted 111 papers.

In the opening ceremony, which took place on the morning of 27 August, speeches were made by G. Scaramuzzi on behalf of the University of Pisa, G. Capriz on behalf of CNR, T. Bolelli on behalf of the *Comitato Scientifico*, G. Torrigiani on behalf of CNUCE, and B. Vauquois on behalf of the International Committee on Computational Linguistics. (Transcripts of their speeches can be found on pp. XXIII-XXXVII). Following the reception offered by the University of Pisa, the two invited papers were read in a plenary session which was presided over by C. G. Cecioni who welcomed the participants on behalf of the *Comitato Nazionale per le Scienze Storiche, Filosofiche e Filologiche* of CNR. In order that all of the 110 accepted papers could be presented within the time allotted, two parallel sessions were conducted throughout the duration of the conference. Each communication was allowed 30 or 40 minutes according to the degree of generality of its content. August 29 was devoted to a sight-seeing visit of Volterra with a formal welcome by the Mayor of that city. The Mayor of Pisa received the participants in the *Sala delle Baleari* on the evening of 30 August. C. A. Mastrelli, President of the *Società Italiana di Glottologia*, welcomed the participants on behalf of Italian linguists and the *Accademia della Crusca*. An organ recital was held in the *Chiesa dei Cavalieri* for the benefit of the participants on the evening of 30 August. At the concluding banquet, on the evening of 1 September, T. Bolelli greeted the participants on behalf of the Rector of the University of Pisa. B. Vauquois, as President of the International Committee on Computational Linguistics, closed the conference.

Throughout the conference, the CNUCE computers, in particular the IBM 360/67 and the IBM 370/155, were at the disposal of the participants for demonstrations and other work. In fact, the programs described in the papers of I. Bátori, J. Courtin and G. Veillon, B. Henisz-Dostert and F. B. Thompson, G. Ferrari, M. Quézel-Ambrunaz and P. Guillaume, M. Pêcheux, C. Cipolli and A. Calabrese, were all demonstrated on the computer.

Organization of the Proceedings

The Proceedings are published in two volumes which contain the texts of the two invited papers and all the papers accepted for the conference, including those whose authors were unable to attend at the last moment.

The Program Committee grouped the presentations into eleven sessions:

I Study of Formal Properties
II Testing and Simulation
III Discovery Procedures
IV Lexicology
V Text Corpus Editing
VI Semantical Calculus
VII Quantitative Description of Language Systems
VIII Grammatical Analysis
IX Meaning Extraction
X Translation
IX Text Comparison

The first volume of the Proceedings consists of sessions I to VI; the second volume contains sessions VII to XI. The two invited papers (I. A. Mel'čuk and W. A. Woods) were included in sessions VIII and VI, respectively. Within each session the papers have been printed in alphabetical order of the authors' names (the name of the first author has been used in cases where there was more than one). Each volume has an index of the papers contained within that volume and, in addition, Volume 2 has a general index. As far the rules and criteria, which guided the Program Committee in their decisions, reference should be made to the preface by its President H. Karlgren. A brief index of themes, compiled with many suggestions from H. Eggers and A. Tomberg, can also be found at the end of the second volume. This list makes no attempt at classification and serves only to assist the reader in finding certain topics.

Acknowledgements

As General Co-ordinator of the 5th ICCL, I should like to express my deep gratitude to the *Consiglio Nazionale delle Ricerche*, CNUCE, and the IBM Scientific Center of Pisa who jointly financed the organization of the conference.

In particular, I wish to thank A. Faedo, the President of CNR and Director of CNUCE, and G. Torrigiani, the Secretary of the Board of Directors of CNUCE, who, showing a clear appreciation of the necessities, the goals, and the aspirations of our discipline, have established a Division of Linguistics at CNUCE. This Division has been responsible, to a great extent, for the rapid development of linguistic and literary computing in Italy. I should also like to thank all the members of the Scientific Committee for their valuable advice and assistance in the organization of the conference; the *Società Italiana di Glottologia* and its President, C. A. Mastrelli, and Secretary, R. Lazzeroni, and the *Società di Linguistica Italiana* and its President, P. Ramat, and past-President, T. De Mauro, for the help they gave for the Italian participation at the conference; and, of course, the Program Committee and in particular its President, H. Karlgren, for their intensive work in selecting and classifying the papers.

It is also my pleasant duty to express my gratitude to all those who have contributed to the success of the conference: the Organizing Committee and all the local authorities who helped in providing hospitality for the participants, and particularly R. Stefanini who was responsible for the co-ordination of the facilities generously placed at our disposal by the University of Pisa; all the personnel of the Linguistics and EDP sections of CNUCE, particularly P. Bronzoni and E. Picchi who organized the demonstrations using the computer, all the secretarial staff of the conference with special mention for the co-ordinator, L. Bertoni, and for B. Ghelarducci and M. Pistelli. Thanks are also due to G. Ferrari who co-ordinated the work involved in the preparation of the preprints, assisted by N. Catarsi, M. L. Ceccotti, L. Pecchia, I. Prodanof, D. Ratti, G. Stilli, and G. Turrini.

I must also thank CNUCE who gave financial support towards the publication of the Proceedings. N. Calzolari Zamorani of the Linguistics Division of CNUCE and European Secretary of the International Committee on Computational Linguistics has been responsible for the initial editing of the texts, the compilation of indexes, and the

co-ordination of the correcting of the proofs,[5] and I thus feel it only right that her name should be associated with the editorship of these volumes. I should like to express my particular gratitude to the publisher, L. S. Olschki, for the special care taken in the publication of these Proceedings.

A. Zampolli

Pisa, 1 October 1973[6]

[5] The correction of the first and second proofs was entrusted to students at the University of Pisa. The third proofs, however, were sent to the authors in order to receive their final approval.

[6] The notes have been added after this date.