Scientific Report

Tullio Gregory (Convenor)

The last few years have seen a considerable quality leap in commonly-available computer technology: memory-capacity of billions of characters, almost a thousand megahertz of processor speed, and increasingly rapid and efficient data-transfer to local and territorial networks. Huge investments in communication protocol, digital data-codification, and SGML (TEI, HTML, XML) languages have all permitted the multiplication of web-sites offering a wide variety of documents and services and a vast supply of CD-Rom and general multimedia material.

Within the humanities, only a few years ago the establishing of digital libraries – offering access via Internet to “esoteric” documents such as manuscripts, bibliographic catalogues, and ancient or rare texts translated directly into graphic resources – seemed a daunting, even impossible task. Any number of encyclopaedias and dictionaries for ancient and modern languages are today consultable on-line. In addition, large collections of texts are available on CD-Rom, which are undeniably precious to scholars, like among others the *Teubneriana* collection, the *Corpus Christianorum*, the *Thesaurus linguae Graecae*, the *Index Thomisticus*, the *Aurea latinitas*, the *Letteratura italiana*, the complete works of Pico della Mirandola, G. Bruno and B. Spinoza,

All these developments open up the possibility of a new philology for the analysis of texts in electronic version, with the gradual elimination of intermediate phases of editing, correction, etc., since the study of the classics can take benefit from the plurality of levels of information – unthinkable on the printed page – made available by the hypertextual structures that may caracterize those texts.

While, however, the application of information technology to the written text has considerable achievements to its credit, a number of issues crucial to our field of research still require a solution. The very ease of obtaining data from traditional support-material raises the deontological question of the choice of reliable editions and the application of codification methods meeting customary bibliographic and philological (and, of course, technological) standards.

The increasingly frequent appearance on the net of professional operators necessitates watertight policies covering brand-names, copyright, and royalties. Computer methodologies for critical textual analysis and the extrapolation of their "lexicological system", as the premise for any study or evaluation, can and must be further encouraged.

In effect, in more dynamic research areas, like Internet technology and multimedia supports, the results of computational linguistics are being reviewed and exploited, as in the case, for example, of associated relevancy-indexes in research on documents found on the net, or the first attempts at automatic on-line translation, and recent experiments in introducing into Internet the formalisation and standardization of the data of computer semantics. One result is that, with increasing frequency, programmes for corpora analysis allow the user lexicographic or lexicometric elaboration such as concordances, frequency-indexes, and various kinds of statistics.

Other potential instruments, however, such as the lemmatisation and automatic translation, or scanning of ancient texts and manuscripts to a reasonable degree of precision, remain only a remote possibility, according to the specialists. They should not, on the other hand, be considered a pipe-dream: in a number of fields, a more articulated approach through the synergy of different methodologies and disciplinary contributions can reasonably be expected to produce significant results.

On these subjects, several scholars were invited to submit papers (the programme of the Workshop is attached to this report). The profiles of participating persons were particularly significant.

Padre Roberto Busa, Professor Emeritus of the "Pontificia Università Gregoriana" in Rome, is a pioneer in the field of computer-based textual analysis, lexicography and bibliographic research, In the early '50 he promoted the "Index Tomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae". In 1992 he founded the "Scuola di Lessicografia ed Ermeneutica" in the Faculty of Philosophy of the "Università Gregoriana" and is now running the project "Totius Latinitatis Lemmata"

Bernard Smith moved into the European Commissions information and technology programmes in 1993 and in 1999 was nominated Head of Unit of Cultural Heritage Applications in the Information Society Directorate General, Luxembourg. His interests span new technologies, cultural and scientific heritage resources and institutions, digitisation policies and programmes, digital library research and increasingly digital preservation issues.

Stefano Rodotà, lawyer, politician, and expert in public law, is president of the "Italian Privacy Authority".

Bernard Quémada was vice-president of the "Higher Council for the French Language" and director of the "National Institute for the French Language" (INALF). A linguist and lexicologist, he was the founder of the series "Materiaux pour l'histoire du vocabulaire français".

Maria Teresa Cabré is the director of the "University Institute for Applied Linguistics" (IULA), Pompeu Fabra University (Barcellona, Spain). Linguist, lexicologist, and terminologist, she has held a number of post in various international organisations and has held courses and seminars at the major Universities in Europe and Latin America.

Susan M. Hockey joined the "School of Library, Archive and Information Studies" as Professor on January 2000. She was previously Professor and Director of the "Canadian Institute for Research Computing in Arts" at the University of Alberta (1997-99), and Director of the "Center for Electronic Texts in the Humanities" (CETH), Rutgers and Princeton Universities (1991-97). She is an Emeritus Fellow of St Cross College Oxford. She was Chair of the "Association for Literary and Linguistic Computing" from 1984-97 and a member of the "Steering Committee for the Text Encoding Initiative". She has served on many digitization and standardization projects, as well participating in the ALCTS Task Force to Define Bibliographic Access in the Electronic Environment, the original Dublin Core meeting at OCLC, the Society of Biblical Literature Seminar on Electronic Standards for Biblical Languages.

Antonio Zampolli is full professor of Computational Linguistics at the University of Pisa, and founder (1968) and director of the Linguistic Division of CNUCE, transformed in 1978 into the Institute of Computational Linguistics of the National Research Council (ILC-CNR), Pisa. His main research interests are literary and linguistic text analysis, mathematical methods in humanities, digital language resources, multimodality, standards for literary and linguistic data processing, computational lexicology and lexicography, modalities and strategies for international co-operation. Co-ordinator of several European projects for the production of language resources, including the current standardisation ISLE/EAGLES jointly funded by the NSF and the EC, he is organizing currently two Italian national projects: "National infrastructure for the linguistic resources in the field of automatic processing of written and oral natural language" and "Computational Linguistics: mono- and multilingual researches".

Heinrich Schepers is Emeritus Professor at the Westfälischen Wilhelms-Universität of Münster (Germany), editor of "Studia Leibniziana" and Director of the "Leibniz-Forschungstelle".

Lou Burnard runs the "Humanities Computing Unit" of the "Oxford University Computing Services", co-ordinating a number of activities in the humanities, most notably: the Centre for Humanities Computing providing IT training, support and consultancy for all humanities faculty within the University; the Humanities Computing Development Team developing the use of technology in humanities; the HUMBUL Humanities Hub providing access to scholarly resources in the Humanities on the net; the Oxford Text Archive archiving and distributing electronic texts for scholarly use. He joins the ALLC-ACH-ACL Text Encoding Initiative project as its European Editor in 1989.

Joseph Denooz is full professor at the University of Liège and Director of the "Laboratoire d´analyse statistique des langues anciennes" (LASLA). He is also President of the Editorial Board of "Revue. Informatique et Statistique dans les Sciences Humaines" (RISSH), an electronic journal dealing with computer science or statistics as applied to the Humanities.

Eugenio Picchi, CNR researcher, is the creator of the DBT package for the access, administration, and consultation of text archives, widely used for the distribution of texts on Internet or on CD-Rom. He is at present engaged in two main lines of research: "Mono- and Bilingual Text Analysis and Corpus Management Systems" and "Mono- and Bilingual Lexical Database (LDB) and Lexical Knowledge Base (LKB) Management Systems".

Paul Tombeur, a historian and Mediaeval philologist, was the founder and director of the "Centre de traitement électronique des documents" (CETEDOC, Catholic University of Louvain la Neuve, Belgium). He now runs the Center "Traditio Litterarum Occidentalium" and its "Library of Latin Texts".

Paolo Galluzzi was full Professor of History of Science at the "Università di Siena" until 1993-94, when joined the University of Florence. From 1982 he runs the "Istituto e Museo Nazionale di Storia della Scienza" of Florence, and from 1995 is President of the Foundation "Scienza e Tecnica". He is also Director of "Nuncius", international review of History of Science, and Head of the Committee for the "Edizione Nazionale dei manoscritti e dei disegni di Leonardo da Vinci". Chairman of the International Scientific Board of the "Nobel Foundation" at Stockholm, he is charged to organize the Nobel Museum and is member of the scientific boards of several prestigious reviews and cultural institutions in Italy and Europe. He promoted the "Storia della Scienza Einaudi", many exhibitions and several information systems or digital hypertexts on relevant subjects in the History of Science and Technology.

Pierre Lafon, University of Bordeaux-1, was a member of the Academic Council of the review "Mots", published by the "Laboratoire de lexicologie politique" of the "Ecole Normale supérieure", Fontenay/Saint-Cloud (France). A mathematician by training, specialist in algorithms and combinatorial analysis, he has published widely on linguistic statistics and lexicometry.

Andrea Bozzi, CNR researcher, is the author of the Latin machine dictionary "LemLat", and is at present working on the development of a system for the preservation of old manuscripts using digital optical tools and the implementation of a specialized workstation for their transcription and electronic processing.

Marco Veneziani, CNR researcher, has published several indexes and concordances of Italian and Latin philosophical works of the Eighteenth Century, also working on text and technology data for information retrieval.

As can be noted, many of participating persons and Institutes shared a common background and raised analogous questions. For example, the work of the Lessico Intellettuale Europeo (LIE-CNR), covering the history of ideas and electronic lexical analysis of cultural texts, has traditionally developed in parallel with other European institutes (CETEDOC, LASLA, INALF, ILC-CNR) working on lexicographic methodology, data-acquisition, electronic archives, machine dictionaries, and the processing of natural-language texts. Over the years the LIE Data-bank of philosophic texts of the seventeenth- and eighteenth centuries has grown according to definite criteria, thus offering procedures for documentation and quotation, together with the stringency of a linguistico-informatic treatment, a classical data-base together with hypertextual facilities, the traditional bibliographical *thesaurus* with the scope of a lexicographic undertaking. However, after a decade of experience, LIE consider it necessary to reconsider basic concepts such as the exhaustive treatment of linguistic data, to overhaul instruments such as the indexes and concordances of works and of authors, to update editing supports such as books and microfiches, and to acquire know-how on avant-garde information techniques, particularly those concerning content-analysis.

Likewise, many papers emphasized a noteworthy delay in the recourse to standards as Unicode, TEI and XML, failing enough supporting resources. Other relevant problems arise from the need of lemmatizing texts in view of reliable linguistic analysis. But all participating Institutes would like to overcome these difficulties and to offer itself on the net as a qualified and reliable source of documentation for the history of philosophy, of science, and of ideas, through the open verification of academic and philological choices, as well as of technical and editing criteria.

The Workshop therefore was focused on texts in machine readable form, and made a survey – as yet lacking in Europe – of the "state of the art" and of new trends of scholarly research at the level of the major academic institutions, from acquisition and codification techniques and specialised data-bases to the development of computational lexicography and more recent methods of lexical and linguistic analysis.

A comparison-confrontation with other European institutions working on the study and analysis of computer texts has proved to be indispensable at a moment, such as the present, of great technological and technical dynamism in exegesis, linguistics, and lexicography. In fact, an awareness of the various operative solutions, and a review of the multiple research prospects, are desirable objectives within the perspective of an ongoing cultural and academic exchange. What was aimed at was thus the “contamination” of humanistic research and new technologies, which will be also the main peculiarity of the “Proceedings” of the Exploratory Workshop, now going to press.

PROGRAMME

Computer texts: documentation, linguistic analysis and interpretation
Tullio GREGORY (Convenor)

Friday 14 June 2002

09h00	**Tullio GREGORY** (Rome) *Opening and Welcoming Remarks*
09h15	**Bernard SMITH** (Luxembourg) *Cultural content and Digital Heritage – Past, Present and Future*
10h00	**Stefano RODOTÁ** (Rome) *European legal protection of brand-name, copyright and royalties*
10h45	***Coffee break***
11h00	**Bernard QUÉMADA** (Paris) *De la carte perforée au cyberespace. Les nouvelles technologies et la lexicographie française*
11h45	**M. Teresa CABRÉ CASTELLVÍ** (Barcelona) *Research frontiers in applied linguistics*
12h30 - 14h30	***Lunch break***
14h30	**Susan M. HOCKEY** (London) *Reusable Electronic Texts: Towards a Digital Library for Humanities Scholarship*
15h15	**Antonio ZAMPOLLI** (Pisa) *Risorse linguistiche per il trattamento del linguaggio naturale*
16h00	**Heinrich SCHEPERS** (Münster) *"Res non verba". Accessing Leibniz texts by means of philosophical concepts*
16h45	***Coffee break***
17h00	**Lou BURNARD** (Oxford) *TEI and XML: a marriage made in Heaven?*
17h45	**Manfred THALLER** (Cologne) *Textual markup, textual models: Depth or Abyss?*

Saturday 15 June 2002

09h15	**Joseph DENOOZ** (Liège) *Méthodes et applications de banques de données grecques et latines*
10h00	**Eugenio PICCHI** (Pise) *Experience in the field of textual elaboration of large textual corpora: linguistic and software tools*
10h45	***Coffee break***
11h00	**Paul TOMBEUR** (Louvain la Neuve) *Lemmatisation and scientifix analysis of text content*
11h45	**Paolo GALLUZZI** (Florence) *Digital versus Traditional Libraries: problems of commensurability*
12h30 – 14h30	***Lunch break***
14h30	**Pierre LAFON** (Lyon) *Analyses automatiques de textes: possibilité et limites de l'automatisation*
15h15	**Marco VENEZIANI** (Rome) *Thesauri, machine dictionaries and metadata*
16h00	***Coffee break***
16h15	**Andrea BOZZI** (Pisa) **– Alberto RAGGIOLI** (Lucca) *Digital documents and computational philology: the Digital Philology System (DiPhiloS)*
17h00	**Roberto Busa SJ** (Milano) *A global communication network: a linguistic challenge to be taken up*
17h45	**Tullio GREGORY** (Rome) *Final remarks*
18h30	***End of the meeting***

Maximum time for each speaker: half an hour.
Lectures will be followed by a discussion

ESF/SCH Exploratory Workshop on:

Computer texts:

documentation, linguistic analysis and interpretation

Strasbourg, 14-15 June 2002

FINAL LIST OF PARTICIPANTS

Prof. Tullio Gregory (Convenor)
Lessico Intellettuale Europeo e Storia delle idee - CNR
Via Nomentana, 118
I - 00161 Roma
e-mail: liecnr@liecnr.let.uniroma1.it

Prof. Susan M. Hockey
School of Library, Archive and Information Studies
University College London
Gower Street
London WC1E 6BT (England)
e-mail: s.hockey@ucl.ac.uk

Prof. M. Teresa Cabré Castellví
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra
La Rambla, 30-32
E - 08002 Barcelona (Spain)
e-mail: teresa.cabre@trad.upf.es

Prof. Dr. Heinrich Schepers
Leibniz-Forschungsstelle
Rothenburg 32
D - 48143 Münster (Germany)
e-mail: schephe@uni-muenster.de

Prof. Lou Burnard
Oxford University Computing Services
13, Banbury Road
Oxford OX2 6NN (England)
e-mail: lou.burnard@oucs.ox.ac.uk

Prof. Paul Tombeur
31, rue du Haut Chemin
B - 1370 Lathuy (Belgique)
e-mail: p.tombeur@brepols.com
e-mail: tombeur@tedm.ucl.ac.be

Prof. Joseph Denooz
LASLA
Université de Liège
Quai Roosevelt, 1b
B - 4000 Liège (Belgique)
e-mail: Joseph.Denooz@ulg.ac.be

Dr. Eugenio Picchi
Istituto di Linguistica Computazionale - CNR
Area della Ricerca CNR
Via G. Moruzzi, 1
I - 56124 Pisa
e-mail: picchi@ilc.cnr.it

Dr. Andrea Bozzi
Istituto di Linguistica Computazionale - CNR
Area della Ricerca CNR
Via G. Moruzzi, 1
I - 56124 Pisa
e-mail: andrea.bozzi@ilc.cnr.it

Ing. Alberto Raggioli
META - Multimedia and WEB Applications
Viale Carlo del Prete, 347/F
I - 55100 Lucca
e-mail: alberto.raggioli@metaonline.it

Dr. Marco Veneziani
Lessico Intellettuale Europeo e Storia delle idee - CNR
Via Nomentana, 118
I - 00161 Roma
e-mail: veneziani@liecnr.let.uniroma1.it

Prof. Paolo Galuzzi
Istituto e Museo di Storia della Scienza
Piazza dei Giudici, 1
50122 Firenze
e-mail: galluzzi@imss.fi.it
galluzzi@galileo.imss.firenze.it

Mr. Roberto Busa
Aloisianum
21013 Gallarate (Varese)