

## Preface

*With this volume in honour of Don Walker, Linguistica Computazionale continues the series of special issues dedicated to outstanding personalities who have made a significant contribution to the progress of our discipline and maintained a special collaborative relationship with our Institute in Pisa. I take the liberty of quoting in this preface some of the initiatives Pisa and Don Walker have jointly promoted and developed during our collaboration, because I think that they might serve to illustrate some outstanding features of Don's personality, in particular his capacity for identifying areas of potential convergence among the different scientific communities within our field and establishing concrete forms of cooperation. These initiatives also testify to his continuous and untiring work, dedicated to putting people into contact and opening up communication between them, collecting and disseminating information, knowledge and resources, and creating shareable basic infrastructures needed for progress in our field.*

*Our collaboration began within the Linguistics in Documentation group of the FID and continued in the framework of the ICCL (International Committee for Computational Linguistics). In 1982 this collaboration was strengthened when, at COLING in Prague, I was invited by Don to join him in the organization of a series of workshops with participants of the various communities interested in the study, development, and use of computational lexica. I was unable to participate in the first workshop organized by Don at SRI in 1983, because I was involved in a CETIL meeting<sup>1</sup> at the same time in Luxembourg, where the suggestion of holding a second workshop in Europe was accepted. So, Don, Nicoletta Calzolari and I, together with Loll Rolling (CETIL promoter) and Juan Sager (CETIL President), organized the workshop "On Automating the Lexicon" in Grosseto, sponsored by the CEC, the Council of Europe, ACL, AILA, ALLC and EURALEX. The objective of this workshop was to survey research efforts, current practices, and potential developments in work on the lexicon, machine-readable dictionaries, and lexical knowledge bases, with special consideration of 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 CEC*

<sup>1</sup>CETIL was a CEC Committee of Experts in Language and Information.

and widely distributed, could be summarized as follows: <sup>2</sup>

- 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.
- 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.
- 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.
- 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.

Don dedicated an unbelievable amount of time, effort, and care in preparation of this workshop. Don, Nicoletta and I were in touch nearly every day for more than a year. I remember this, with joy, as one of the happiest and most fruitful periods of my working life. During the organization of the workshop, and as we went along establishing relationships with the participants invited to it, we had the feeling that a new research paradigm was emerging more and more clearly every day. In Pisa, we found here a confirmation of our work on texts, reusable lexica, and related tools, despite the almost complete lack of interest in the major trends in contemporary computational linguistics. Don found continual confirmations for his intuition of the multi-disciplinary centrality of the lexicon, and for the necessity of a vast organization to provide the various communities, represented in the workshop, with basic linguistic resources.

The Grosseto Workshop is now recognized as marking the starting-point of a new phase in the field of computational linguistics. This phase is characterized by an increasing number of fresh initiatives, particularly at the international level, whose aim is to further the development of the scientific, technical, and organizational conditions conducive to the creation of large multifunctional linguistic resources, such as written and spoken corpora, lexicons and grammars. An updated version of the proceedings of the Grosseto Workshop, edited by Don, Nicoletta and me, will appear soon (Walker et al, 1994). Don put a lot of effort into collecting and updating the articles, and we are extremely grateful that he was able to finish this task.

Don was particularly active in promoting actions whose purpose was to realise the strategic vision that emerged during the workshop. He organized two follow-up workshops, one in New York in 1986, the other, in 1987, in Stanford, in the framework of the Linguistic Summer School of the LSA.

<sup>2</sup>See Zampolli, 1987, for the complete text of the recommendations.

He also organized a panel on linguistic resources at the 1988 Pisa Summer School on Computational Lexicography and Lexicology, explicitly designed to contribute to the goals identified during the Grosseto Workshop. Unfortunately, the day before the panel was due to meet, Don notified me that he was not able to come, because his illness had just been discovered and he was about to have his first operation. I still remember that phone call with deep emotion. Don encouraged me to carry on with the work regardless. The way in which, during the five years of his illness, he was able to continue working with energy and enthusiasm, guiding and encouraging us in the various initiatives we were promoting together, and always planning new activities, was for me a miraculous example of what devotion to one's discipline, community and ideals can really achieve.

The day after the Grosseto Workshop, I set up a group (Hans Uszkoreit, Nicoletta Calzolari, Bob Ingria, Bran Boguraev) to explore the feasibility of constructing large-scale linguistic resources, explicitly designed to be multifunctional, i.e. capable of serving, through appropriate interfaces, a wide variety of present and future research and applications. A crucial and controversial problem was to what extent it was possible, as well as desirable, to make linguistic resources, at least within certain limits, "polytheoretical", i.e. usable in different linguistic frameworks. Don immediately intervened and secured for the group, initially supported by our Institute, the help of the ACL. He was enthusiastic about the goal of this group, (the so-called "Pisa - Group"), which was extended to include outstanding representatives of some major linguistic schools, and (I quote Don's words) "investigated in detail the possibility of a polytheoretical representation of the lexical information needed by parsers and generators, such as the major syntactic categories, subcategorization and complementation. The common representation sought was one that could be used in any of the following theoretical frameworks: government and binding grammar, generalized phrase structure grammar, lexical functional grammar, relational grammar, systemic grammar, dependency unification grammar and categorial grammar" (Walker et al, 1987).

After the Grosseto Workshop, we wrote a report together for the CEC DG-XIII, suggesting a large two-phase programme: a one-year phase, to define the methods and the common specifications for a coordinated set of lexical data bases and corpora for the European languages, and a second three-year phase for their actual construction (Zampolli, Walker, 1987). This programme, it seems, is finally being realized now, even if organizational and financial problems have made its progress slower than thought.

The first step taken by the CEC consisted in a feasibility study: the ET-7 Project, building on the encouraging results of the "Pisa-Group" work. This project was launched by the CEC with the aim of recommending a methodology for the concrete construction of shareable lexical resources. Since different theories use different descriptive devices to describe the same linguistic phenomena and yield different generalizations 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 has assessed the feasibility of some basic standards for the description of lexical items at the level of orthography, phonology/phonetics, morphology, collocation, syntax, semantics and pragmatics.

(Heid, McNaught 1991)

The second major step is now underway, and is represented by the establishment of projects focussing on the notion of standards. Two examples in the field of lexical data are the CEC ESPRIT project MULTILEX, whose objective was to devise a model for multilingual lexicons (Khatchadourian and Modiano, 1994), and the EUREKA project GENELEX, which concentrates on a model for monolingual generic lexicons (Antoni-Lay, Francopoulo and Zaysser, 1993). In the area of textual corpora the CEC sponsored the NERC study aiming at defining the scientific, technical and organizational 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 corpora building and analysis (see NERC Final Report, 1993). The European Speech community had independently organized an outstanding standardization activity, coordinated through the ESPRIT Project SAM (Fourcin, Gibbon, 1993). Feeling it necessary to coordinate their activities, the representatives of these various standardization projects formed an initial, preparatory group. Enlarging this group, the CEC established the EAGLES project in the framework of the 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, lexica, formalism, assessment and evaluation and speech data. (Elsnews Bulletin, 2(1), 1993). The project also encompasses an international dimension, influenced by Don's suggestions, which includes: the support of the European participation in the TEI; the preparation of a survey of the state-of-the-art in Natural Language and Speech Processing, jointly sponsored by the NSF and the CEC; the preparation of a Multilingual Corpus (MLCC) intended to support Cooperation with similar ARPA sponsored initiatives; and the exploration of possible strategies for international cooperation and coordination in the field of linguistic resources.

We are now on the verge of the third step needed to complete the programme suggested in our 1987 report to the CEC: to define and provide a common set of basic multifunctional reusable linguistic resources for all the European languages, available in the public domain, and to create an infrastructure for the collection, dissemination and management of new and existing resources. The second task is taken care of, at the experimental level, by the new LRE-RELATOR project (Elsnews Bulletin, 2(5), 1993). The MLAP call issued recently by the CEC DG-XIII, gives us the hope that the first task will be finally undertaken within the 4th Framework Research Programme of the CEC.

Given his interest for making linguistic resources shareable, and people and disciplines cooperate, the deep involvement of Don in standardization efforts is not surprising. His participation, from the very beginning, in the process of setting up the Text Encoding Initiative, is yet another testimony to his capacity to overcome (often artificial) disciplinary boundaries and of his determination in realizing his strategic vision of the future. Nancy Ide, President of the ACH, in November 1987 held a brainstorming workshop of representatives of the ACH and ALLC at Vassar to explore the desirability and feasibility of standards for encoding and exchanging texts in the field of humanities. Don, informed of this event, contacted Nancy and was also invited together with Bob Amsler. I remember my

emotion on hearing Don (in the evening I had gathered Don round the fire together with Susan Hockey and Nancy Ide, to discuss the possibilities of cooperation of the three Associations) announcing that he would ensure the support of the ACL to a joint initiative with the ACH and ALLC. Those of us old enough, like me, to have lived through the entire history of linguistic data processing, will probably understand my reaction. In the '50s and early '60s, activities such as machine translation, document retrieval, lexical text analysis for humanities, statistical linguistics, quantitative stylistics, which later on were separated by scientific and organizational barriers, were linked by frequent contacts and recognized each other as "poles" of a disciplinary continuum, linked together by an interest in computational processing of large "real" texts and language data. In the second half of the '60s, computational linguistics, in an effort to define its disciplinary identity, and under the influence of the generative paradigm in linguistics and of the embryonic AI, became progressively detached from the interest and work on "real" language data. The occasions for contact with the community of humanities computing became increasingly rare, despite the efforts of some institutes, which refused this separation, and continued to consider the computational processing of real language data as a unifying goal to which the various disciplines could contribute relevant know-how, methods and tools.

Reversing this trend, Don went on to play a central role within the TEI, where he contributed his capacities to mediate, to organize and balance our efforts, and to envision the future. Despite our tremendous loss, we are determined to realize his vision of the TEI as a bridge between standardization activities in various continents. He also recognized the areas of potential synergies between computational linguistics and literary and humanistic computing, and identified the benefits potentially deriving from mutual cooperation. He promoted, in particular through the friendships established in the framework of TEI with Susan Hockey (chairman of the ALLC), various events jointly sponsored by the ACH, ACL, ALLC, and above all, he actively contributed, from the initial steps to the establishment of the CETH (Center for Electronic Texts in Humanities).

Don's interest in linguistic resources (conceived as a fundamental infrastructure for natural language and speech processing) and standards (conceived as a means for allowing exchanges and cumulative efforts) naturally converged with his vision of "Language, information and knowledge....combined in human efforts after communication"<sup>3</sup> and with the profound humanism that pervaded all his activities. At the basis of his capacity to open up lines of communication between peoples and communities and to propose new collaborative initiatives, of his will to create the best working conditions for everyone, of his continuous and untiring work in the promotion of scientific development in our sector, considered by him to be inseparable from the development of people, lie, I believe, not only his desire to open new perspectives maximizing synergies and interdisciplinary fertilization, but also and above all his genuine, profound interest in people. In the various activities of his long career, he was always inspired by a sense of help, and by a deeply democratic and humanistic conception of research and its management.

I would like to relate a personal anecdote which revealed to me Don's strong

---

<sup>3</sup>See Don's article in this volume.

values. One evening, on the occasion of a reunion of the "Pisa Group", I invited Don to my house, with Annie Zaenen, to listen to some music, and at a certain moment I put on the records of "The Weavers Reunion at Carnegie Hall" of 1953 and 1963. I was surprised to see that Don became more and more moved as each song played, and I asked him why. He told me that the Weavers remained, at the beginning of his scientific career, during the worst period of the cold war, one of the most visible symbols of democracy and freedom of culture in his country, the basic values that had inspired his life.

To close, I would like to thank, in the name of our journal *Linguistica Computazionale*, all those who have contributed to the realization of this issue, moved by admiration, gratitude and love for Don and his work. The Scientific Council of the ILC, in its role as editorial board of the journal, has approved and sustained my proposal to dedicate this number to Don Walker. Nicoletta Calzolari, Susan Hockey, James Pustejovsky and Susan Armstrong have signed with me the invitation to the contributors. As co-editor of this volume, Martha Palmer expended substantial scientific and organizational effort in collecting and preparing these articles, for which our heartfelt thanks. We are grateful as well to Carolyn Elken, Tarina Ayazi and Antonietta Spanu for their precious assistance.

Antonio Zampolli

*I only wish to express my deep gratitude to Don, with very simple words, and of my many different memories of him just touch on one or two of those which perhaps mean most to me. I strongly feel the need to thank Don for what I received from him as a friend, as a fellow linguist, and as a person.*

*As a friend he treated me as I later realised he treated others whom he trusted: he helped me, he encouraged me, he gave me the right opportunities in the right moments, and I always knew that I could rely on him, at any moment. And this is a very rare gift indeed. We worked together in a wonderful way, as only friends can do.*

*As a linguist, we all know what a great contribution he made to our field and in how many different ways, both big and small: moving things forward, putting people together, encouraging newcomers, understanding in advance what would be important in the coming years, and acting concretely to make things happen. Among his many talents, he was a great builder.*

*As a person, I really must express my immense admiration for the way in which, once he knew of his illness and right up to the last days, he not only engaged in his own very private but constant battle but was constantly and courageously able to talk about the future with others, to make plans, to act towards the achievement of results that he knew that he personally would never see. And always with a smile and a kind light in his eyes. This for me was the clearest sign of his greatness. This memory will remain with me forever.*

Nicoletta Calzolari

## References

- [1] Antoni-Lay, M.H., G. Francopoulo, L. Zaysser, "A Generic Model for Reusable Lexicons: The Genelex Project", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [2] "RELATing to Resources", *Elsnews Bulletin*, 2 (5), 1993.
- [3] "EAGLES Update", *Elsnews Bulletin*, 2 (1), 1993.
- [4] Fourcin, A., D. Gibbon, "Spoken Language Assessment in the European Context", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [5] U. Heid, J. McNaught (eds.), "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", *Eurotra-7 Final Report*, Stuttgart, 1991.
- [6] Khatchadourian, H., N. Modiano, "Use and Importance of Standard in Electronic Dictionaries: The Compilation Approach for Lexical Resources", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [7] LRE, "Technical Report Document", Unpublished document available from R. Cencioni, CEE- DGXIII, Luxembourg.
- [8] NERC Final Report, Istituto di Linguistica Computazionale, Pisa, 1993.
- [9] Walker, D., A. Zampolli, N. Calzolari, "Toward a polytheoretical Lexical Database", Istituto di Linguistica Computazionale, Pisa, 1987.
- [10] Walker, D., A. Zampolli, N. Calzolari (eds.), *On Automating the Lexicon: Research and Practice in a Multilingual Environment*, Proceedings of a Workshop held in Grosseto, Oxford University Press, Oxford, 1994.
- [11] Zampolli A., Walker D., "Multilingual Lexicography and Lexicology; New Directions", Unpublished paper presented to CGXII-TSC of the CEC.
- [12] Zampolli A., "Perspectives for an Italian Multifunctional Lexical Database", *Studies in honour of Roberto Busa S.J.*, A. Zampolli, A. Cappelli, C. Peters (eds.), *Linguistica Computazionale*, Pisa, 6, 1987, 301-341.
- [13] Zampolli A., "Technology and Linguistic Resources", in M. Katzen (ed.), *Scholarship and Technology in the Humanities*, British Library Research, London, 1991.