The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

Vol. 8 n°3 & n°4 2003

Special Issue in Memory of Antonio Zampolli

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Magali Jeanmaire

Layout:
Magali Jeanmaire

Contributors:
Nicoletta Calzolari
Khalid Choukri
Nancy Ide
Martin Kay
Joeph Mariani
Bernard Quemada

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr
Web sites: www.elra.info
or www.elda.fr

Dear readers,

For more than 6 years, I have shared this page (President and CEO Letter) with Antonio Zampolli. Having endorsed my suggestion to have an independent ELRA newsletter that would report on ELRA activities, he also insisted on contributing to this letter regularly and helped to sum up important issues.

Usually the newsletter focuses on some aspects of our work related to Language Resources identification, validation, distribution and technology evaluation. In this double issue, (8.3 and 8.4), we have decided to honour the memory of Antonio. We have asked a number of friends, colleagues and personalities who shared Antonio's fights over a number of decades to contribute.

We have also thought of how we could honour and remember Antonio.

The first thing we should do all together is continue our work on LRs and evaluation, with the same enthusiasm and commitment. I am sure that a number of roadmaps and paths are well marked and we need to build upon them. One major issue Antonio initiated but did not get enough time to consolidate was the integration of spoken and written communities, and I am confident that all together we will carry this on.

We can also honour Antonio’s memory by continuing to make LREC a very successful event for the years to come. An important event during which we will continue to offer discussion forums to the whole community. We will continue to invite official representatives from funding and sponsoring agencies, and ensure that a common agenda is worked out, strongly advocating for the expectations of the R&D and the industrial communities as Antonio Zampolli used to do.

We can also honour Antonio’s memory through the acknowledgement and recognition of outstanding contributions rendered to our field by his peers. Hence, the birth of an Award bearing the name “The Antonio Zampolli Prize”, which will be awarded every second year.

I would also like to recognize Don Angel Martin Municio, the former President of the Spanish Royal Academy of Exact Sciences, Physics and Natural Sciences. Don Angel played a predominant role in the establishment of ELRA, but we will also remember him as he who encouraged, set up and supported the LREC organisation, not only from a scientific angle but also through his networking capabilities.

Antonio, Don Angel, I miss you both.

Thank you and Rest in Peace.

Khalid Choukri

A Few Words from Khalid Choukri, ELRA CEO

Dear Antonio,

I am a very privileged person. I am among the lucky few who worked with you for the last 8-9 years on a daily basis. I am, perhaps, the only contact outside your Institute. Now I realise how much I owe you technically, undoubtedly, but above all in terms of human relationships.

Having shared some of your most recent challenges in so many cities worldwide, I profess my admiration for your intuitive analysis of debates and their context. Many times I noticed lack of focus during crucial meetings only to realize in retrospect how you mastered the situation by expressing clear opinions on your longterm goals and results.

I also realised how passionate you could be and how easily you communicated such passion. Do you remember the discussion with Juan Carlos, King of Spain, during the launch of the Royal Academy dictionary? After having realized that you could not meet the King of Spain with a red and colourful tie (!), you agreed, that morning, to invest in a black one. Regardless of the tie’s color, you managed to communicate your enthusiasm and concerns naturally, in such a friendly way!! Who cared about protocols?! And now, I remember the King of Spain laughing at your Italian jokes!

This story happened in the latter part of your fight to structure what has become Human Language Resources.

Few young people will remember you as one of the pioneering persons behind Eurotra, and many other key projects (some of them are described in your testimony paper herein), which all designed and articulated the entire subject. Among the major ones, let me just quote ELRA and ENABLER.

ELRA was not an easy infrastructure to set up, but you managed to gather consensuses from major laboratories in Europe, from key industrial players, and from a large number of European Commission officials. As a follow up to the mission of ELRA, you hoped to see all European countries launching strong national initiatives in this area. With foresignt, you spotted the need to coordinate these initiatives and streamline their progress towards common goals.

A large number of initiatives are well established today (ELRA, ELSNET, EAGLES achievements, interoperability concerns in Language Resources and standards, etc.). Some may forget how difficult it was to push and initiate them when no one cared about infrastructures, basic resources, etc. You also identified the needs of our community and had suggestions in mind to respond to them.

Do you remember how LREC started? It looked like a brainstorming at the end of a meeting, but we knew you had had it in mind for a long time. I remember how convincing you could be as a scientist and as a friend. Don Angel Martin Municio, who we will also remember, offered his support and the lovely city of Granada to host the 1st LREC. I remember taking the bets, on a lovely Andalusian patio, on the number of attendees. How many did you expect? Between 50 and 70? But your twinkling eyes meant something else!

I always enjoyed your European spirit and also realised how universal your thinking was. Despite all the barriers, you encouraged true international cooperations, and how many times did you help to promote initiatives that no one would have believed in at first glance? The number of meetings about standards of metadata you attended and contributed to, e.g. OLAC, is clear proof of your determination.

You gave the impression of focusing on larger issues and not on details; working with you for the launching of LREC and its success, I realise how much importance you gave however to tiny details. From Granada to Las Palmas via Athens, I saw this: you cared about the tie-microphones as well as about the quality of the accepted papers! Not to mention the weather!!! Do you remember that we were ready to buy 600 umbrellas in Athens in order to convince you to hang the posters in the Peristylion, the open space of the impressive Zappeion Megaron conference centre?

The fact that you always insisted on considering the participants as our guests and also claimed that it is not necessary to go to bad venues to organise productive meetings goes back to 1973 Coling, the Pisa Summer Schools, the meetings that many of us missed and imagine with such an envy (Santorini, etc.). LREC proves this.

It is impressive to remember your rhetorical spirit which catalyzed a number of things, but also your perseverance. You never gave up when you knew what was at stake.

I remember that during the first years of activity at ELRA, you were very enthusiastic to see the progress of our work related to identification of existing Language Resources. You helped us, with an impressive directory of contacts from all over the world, come up with the best LRs catalogue.

I do not know if it is because you started your career working at IBM that you could easily understand concerns expressed by industry. I remember our debates on pricing policies! Though you pretended not to be good at these financial matters, I know that you saw the underlying long term issues: you defended the low price at which LRs should be offered to researchers whenever possible.

Dear Antonio, you always promised that “one day, you will see, I will write the history, or story, of computational linguistics, Language Resources, and shed light on all the fights that one had to carry out. Maybe I will write that story, once retired in my mountains… I will tell young people how difficult it was to start this, how hard it was to struggle against the funding agencies, the big established institutions, etc.” I looked at you, with amusement and scepticism. To tell you the truth, this is *the* promise I have never believed in and trusted. I am sure you knew that from my expression. I have never imagined you living a retirement of any kind. I have never seen you as someone who might withdraw from a battle before its final victory, and the number of battles ahead of us remains numerous, and you did part of the job: the path is well marked out, even if there is still a very long way to go.

I could go on with these anecdotes for pages and pages but I know you do not like us to express our recognition too much, though you appreciated “El Relicario” of Sarita Montiel!

Antonio, Thank you for everything.
Rest in Peace.

Khalid Choukri

The Opening Ceremony of LREC 2002 in Las Palmas, with Antonio 5th from the left.

A Few Words from Joseph Mariani, who succeeded Antonio Zampolli as ELRA President

I remember my first encounter with Antonio Zampolli which coincided with the launching of Elsnet, the European Language and Speech Network, back in the early 90s. The main idea was to gather the spoken and written language processing communities, which were on separate tracks at that time, mainly due to different scientific backgrounds. The fact that speech includes ambiguity right from the signal which is to be processed, induces the use of methods taking into account this ambiguity, which also necessitates training data. This was not so obvious at that time for the processing of written language. We have come a long way since then.

As I was at that time (and since 1988) the Chairman of the brand new European Speech Communication Association (ESCA - now ISCA since it has become international), I was invited by Ewan Klein, the Elsnet convenor, and Antonio to join the Elsnet Executive Board, when it was formed in 1991.

There were ups and downs in our relationships at the beginning, given the difference in the way we approached our respective scientific fields. I remember in July 1992, in a nice summer evening in a pub in Dublin, where a workshop dedicated to “Integrating Speech and Natural Language” was organized; I was having dinner in front of Antonio, and he expressed the opinion that I was not defending enough the speech community in front of the European Commission!

However, we quickly developed strong links and we became close friends.

We frequently met at the Elsnet Board meetings, and came to the opinion that there was a strong need for promoting language resources in Europe, as Antonio had been a strong supporter of data-based approaches for Natural Language Processing for years, and had had to fight for his ideas.

He organized the Relator proposal which was submitted to the European Commission, and he led the project when it was accepted. The main outcome of the project was probably the proposal of creating a European association in that field, and Antonio used his wonderful talents to convince the Relator Advisory Committee, chaired by Brian Oakley, and the EC DG XIII Deputy Director General, V. Parajon Collada, that such an association was mandatory for ensuring the future of Language RDT in Europe. He asked me to draw up a business plan, which demonstrated that the initiative was worth being tried. Many discussions took place in the meantime, as it was a major political issue. In particular, I remember those which were held in the Santorini island, beside the swimming pool, or during long lasting lunches and dinners, during the “Language Engineering and the Information Highway” Workshop, in May 1994.

It was decided to launch the European Language Resource Association (ELRA) by the end of 1994. The Association was created in February 1995. Antonio came to meet me at the 1995 Eurospeech conference, in Madrid. I ensured him of my full support, and ELRA started its activities, with Antonio as president and myself as a VP. Language Resources came, members came, success increased. Antonio was very anxious about ensuring the well being of ELRA. He took a lot upon himself, often travelling and participating in meetings and conferences to promote the Association. I remember the General Assembly in December 1996, where the first year results were to be communicated to our members. Antonio laid directly on the ground for 5 minutes to gather his forces before chairing with brio the General Assembly.

The idea of launching a scientific conference was formulated at an Elsnet Executive Board meeting in May 1997, with the aim of having an event of interest for the scientific community, on an international level. We thought that having language resources and evaluation together was mandatory. The ELRA Board reacted very positively to the idea, and the late Angel Martin-Municio, proposed enthusiastically to have the first issue in Granada. Antonio dreamed of the Alhambra, of Sarita Montiel and of “El relicario”. He also remembered with emotion a meeting organized by Don Angel in La Rioja. But he was still very anxious about the number of papers and participants that we could attract. The result was a huge success, and the banquet in the Alhambra is one of the most pleasant memories of the genius of Antonio for getting people together.

Antonio and I were often on the same wavelength. We often met at Elsnet EB meeting, where a Language Resources and Evaluation group was under his responsibility, or at ELRA meetings, sometimes in the marble room of the Duomo hotel, where Antonio liked to organize meetings, sometimes in the beautiful countryside of Tuscany. LREC went on, with increasing success, and Athens succeeded to Granada, and Las Palmas to Athens, with more papers and more participants each time.

Antonio thought that it would be important to also have an international body where the written language community could meet and discuss language resources and evaluation, similar to the Cocosda committee that we created for spoken language at the Eurospeech 1991 conference in Genoa. Discussions took place at LREC'02 in Las Palmas, and the decision to create the International Committee on Written Language Resources (ICWLR) was taken in 2003, following his initiative.

But Antonio's health was not good. He fell when riding his bicycle, in the streets of Pisa. Antonio asked me to put my candidacy as ELRA President in the spring of 2002, as he didn't feel that he could face another 2-year term, and I was elected after him, while he became Honorary President.

He had another dramatic fall from the podium at the conference centre during LREC'02, and was driven to the hospital. He gathered his last forces to participate in the Banquet on the last day, where I was so happy to offer him a record of Sarita Montiel singing "El relicario". His last request was to ask me to replace him at the ALLC/ACH conference in Tübingen, in July 2002, and to support the launching of the ICWLR. After that, we never met up again. He regularly cancelled his participation at Elsnet meetings and at the meetings of the ELRA Board. He was also much affected by the death of Angel Martin Municio in November 2002.

And in the morning of August 22, 2003, I received the sad information that I lost a friend.

A Few Words from Nicoletta Calzolari, from ILC-CNR in Pisa

It is not easy to write about Antonio for someone who was so close to him as I was. I prefer to speak to him as if he were still among us. When something relevant happens in our field I still believe he is in some way with us. Could he miss an important event? It wouldn't be like him!

I want to thank Antonio, from all of us, because if we are now speaking and writing about Language Resources, we owe this to him. And personally, I owe so much to him and I have learnt so much from him that I simply cannot express it. Now, when some difficult decision has to be made, the first thing I ask myself is: what would he have done? I may decide to do differently, but this is always my starting point.

Antonio dedicated all of his life to work, and here I will try to touch just a few of the so many significant moments of his work.

Computational Linguistics started in Italy with Antonio Zampolli. He was the first one (and for many years the only one) to hold a chair of Computational Linguistics, and he founded the Istituto di Linguistica Computazionale of the CNR in Pisa, of which he always remained the Director. This was probably his most important achievement (which leaves us with a big, but also endeavouring, inheritance), and we have now asked that the Institute be named after him.

In the early '70s he organised the famous Pisa Summer Schools, at a time when summer schools were not as trendy as they are now, gathering in Pisa the best names of the time. He always remembered that Joan Bresnan and Ron Kaplan gave birth to lexical functional grammar on a tower in San Gimignano (instead of looking at the wonderful view!). A whole generation of computational linguists all over Europe were trained in these Schools, which are still remembered as big events by many. We just found video recordings of these Schools and we'll try to digitalise them to make them available to all. They constitute a historical treasure for our field, and this will be done in homage to Antonio, such a "visionary" in our community.

I have always seen him as a man of great "visions", as few are, capable not just of anticipating but also of creating the future (the future, that others would have seen much later), always pushing forwards in new directions and fostering new initiatives with never ending energy (difficult for us sometimes to follow him), often struggling to make his intuitions become a reality (he struggled so much in his life…).

He also had the great capability of mixing people coming from different communities, thus creating the perfect mixtures to develop new ideas. An example of this is the famous Grosseto Workshop in '86: the last morning he gathered 4 or 5 of us for breakfast, and from that breakfast a series of initiatives started towards the creation of standards and best practices!

He was able to communicate to everyone his enthusiasm and his passion (this is what still lives with us!).

He was the designer and coordinator of such a large quantity of national, European, and international initiatives and projects. He founded so many associations, boards, committees, networks (ELRA and ELSNET just to mention two of them), and chaired almost all of them (however, he considered that being Director of our institute in Pisa was the task of utmost importance!).

He "invented" the field of Language Resources, and recognised their "infrastructural" role, at a time when it was almost a shame to speak about data. However, he believed in it and pushed for it and fought for it. The beginning of the era of Language Resources (although the term did not exist yet) was the famous Grosseto Workshop, which we organised along with Don Walker (another man of visions able to shape the future, a great friend of both of us, whom I would like to remember here together with Antonio). Now everyone speaks about language resources, as if it were normal, but someone had to struggle to open this direction of work.

He designed a long series of standardisation initiatives, from TEI to EAGLES, and innumerable projects to set up a European and world-wide infrastructure for language resources, culminating with the setting up of a European Network of National projects (ENABLER) and the founding of the International Coordination Committee for Written Language Resources and Evaluation (ICCWLRE), with representatives from all the continents.

He invented LREC, such a successful conference, where once again there is a real gathering of so many different communities. He told me: "There were 5 minutes left at the end of an ELRA Board meeting, and I launched this idea…". And that's how LREC was born.

He was a guide and a "maestro" for so many of us, a "living legend" as has been said. But also, in a unique combination, a man who was loved for his humour, his kindness, his enthusiasm, his vitality, his intuitions, his friendliness to all, his wonderful stories, his childish attitude sometimes, his love for the mountains and for music. He would also make us angry sometimes (in particular, those closest to him), but then he would win us over with his kindness…

The last issue (2 volumes) of the journal of our Institute, "Computational Linguistics in Pisa", with articles from all of us at ILC and an introduction with his viewpoint on the Institute (the article reprinted in this number of the ELRA Bulletin), was strongly wanted by him as if he felt that it was time to leave a testimony of all his initiatives, that is to say his life. He wanted recognition, and he deserves it, from all of us, including the younger generation.

Antonio Zampolli is no longer with us, but he will continue to live with all of us who loved him. We will continue to work along one of the many paths that he opened, to look into the future with him, to join forces for this future, to give rise to new ideas and to struggle together to bring these ideas to life.

Past & On-going Trends in Computational Linguistics: a View from the Istituto di Linguistica Computazionale

Antonio Zampolli

1 - Brief Historical Overview

1.1 The Beginnings

In the years between the end of the 50's and beginning of the 60's, the primary objective of P.R. Busa S.J., who was universally acknowledged as the pioneer of electronic text processing since 1948, was to process the entire corpus of writings either authored by or attributed to Saint Thomas Aquinas. This corpus had a total of 10 million occurrences. This work was carried out at the Centro per l'Automazione dell'Analisi Linguistica (CAAL) in Gallarate, near Milan, financed in large part by IBM Italy. In those times, the enterprise was of vast proportion and importance.

In 1960, after my University degree, with a dissertation entitled *Studi di statistica linguistica eseguiti con impianti IBM* (*Studies of linguistic statistics carried out with IBM equipment*), I started to work in Gallarate, where I was P.R.Busa S.J. assistant. I was in charge of organizing and coordinating electronic text procedures: namely, inputting texts on punched cards (an activity involving about 60 people) and processing cards with Unit Record machines (involving 30 operators). I was also responsible for the computational and linguistic aspects connected with the processing of Latin texts (methods for contextualization, lemmatization and morphological analysis, etc.) and their translation into algorithms for the computers in use at that time (IBM 7090 and IBM 1401). This work also led me to the creation of a Latin Machine Dictionary (based on Forcellini's list of lemmas - *Lexicon Totius Latinitatis*) and to the development of look-up algorithms.

The procedures and programs implemented at CAAL were soon to be adopted within the projects of other Institutions, including the processing of the *Enciclopedia Dantesca* and the lexical archive of the future *Institute for Juridical Documentation* (*Istituto di Documentazione Giuridica of CNR*).

At that point, I was engaged by IBM and started to work at the Scientific Centre (Centro Scientifico) in Pisa where I was asked to provide consultancy and assistance for projects employing electronic processors in the fields of linguistics and humanities in general. An important reference point for this type of activity was CNUCE (National University Centre for Electronic Computing) in Pisa, inaugurated in 1965 by the President of the Italian Republic, Giuseppe Saragat, who on that occasion presented the volume of *Concordances and Indexes* of Dante's *Divina Commedia*. In 1969 the number of Italian projects using methods, procedures and standards for the representation of texts and linguistic analysis had become so numerous, that the Director of CNUCE accepted, on my advice, to set up a Linguistics Division under my direction.

In 1969, I also held a course - the first in Italy - on Computational Linguistics. One of the participants, Beniamino Placido , who was at that time director of the electronic text processing office for the Chamber of Deputees, was convinced of the importance of applying the procedures of information retrieval to the laws for the members of Parliament, using a machine dictionary able to “project” query terms onto text terms in a more efficient manner. The dictionary was to contain not only the information needed for automatic morphological processing, but also information related to the semantic types: definitions, synonyms, etc. The funds made available by the Chamber of Deputees to finance this project allowed CNUCE to assign a total of 15 grants for the compilation of the Italian machine dictionary (*Dizionario Macchina dell'Italiano*) (DMI). In 1974, as CNUCE was turned into CNR, these grants became research positions. In the meantime, CNUCE had assigned 2 staff units, 5 more later on, to the Linguistics Department, which counted a total of 20 staff members, together with the new researchers. The information encoded in the DMI has been the starting point of a number of important lexical projects conducted at the Institute such as ACQUILEX and ITALWORDNET.

1.2 The Linguistics Department at CNUCE (1969-1978)

As Director of the Linguistics Department at CNUCE, I was essentially concerned that programmatic research in the sector could jointly develop the two main trends of Computational Linguistics. After a period of frequent and interesting collaborations (1955-65), the field of Computational Linguistics had progressively parted: On the one hand, with *Humanistic Text Processing*, HTP (use of computational tools for humanistic research on texts and documents, in particular through the production of indexes and concordances). On the other hand, with *Natural Language Processing*, NLP, developed in connection with the pioneering activities of Machine Translation. NLP aimed at the application of formal models largely elaborated by generative-transformational linguistic schools to the “computation” (or more precisely to the computational analysis, intended as both identification and representation) of linguistic structures underlying the texts, or vice versa, for the generation of texts starting from the representation of these structures.

In particular, if the HTP trend very quickly capitalized on the increasing number of opportunities offered by technological progress (photocomposition, video terminals, increasing availability of text characters, on-line connections, etc.), it constrained itself mainly to the analysis of graphical text units. HTP ignored NLP know-how and methods for the identification and representation of units and properties at higher levels of linguistic analysis (morphological, lexical, syntactic, etc.). Few research centres in the world could have been said to be active in both trends.

In my quality of director, I oriented the activities of the Linguistics Department (and of the Institute for Computational Linguistics, ILC) towards the development of joint research in both sectors, using common techniques, methods and knowledge.

Our research in Pisa mainly focused on the creation and adaptation of NLP tools for the improvement of HTP applications with the purpose of expanding the HTP applications themselves, through e.g. procedures for semi-automatic lemmatization and morphosyntactic tagging of texts, using lexical knowledge bases as a support for text look-up. Moreover, thanks to the experience gained in processing large amounts of classical texts, we were encouraged to increase the linguistic coverage and robustness of NLP components, to adopt sophisticated methods of quantitative analysis, and to promote awareness of the need to define representation standards. Ensuing research developments, starting from the 90's, with the current trends on digital content processing, have confirmed the efficacy of this strategic choice.

1.3 Summer Schools

The years between the 70's and 80's were characterized by a number of international activities, which aimed at promoting these strategic objectives. In 1970, I organized the first International Summer School in Computational and Mathematical Linguistics, in Pisa, entitled *L'elaborazione elettronica di dati linguistici e letterari* (*electronic processing of linguistic and literary data*). The panel of tutors included R. Dyer (University of New York), M. Gross (University of Vincennes), D.G. Hays (University of Buffalo), O. Menchi (University of Perugia), Ch. Muller (University of Strasburg) and J. Raben (Queens College New York). The experience was repeated in Pisa two years later, and in 1974 and 1977, with prominent international scholars representing the most innovative trends in the sector. The last Pisa International Summer School of Computational Linguistics (the 5th one) was organized in 1988, on behalf and at the request of the *European Science Foundation* (ESF). The choice of the topic Computational Lexicology and Lexicography was dictated by awareness of the growing interest in lexicons from many language-related disciplines. Computer usage was becoming a point of convergence for these disciplines and the “potential focus” of an innovative and inter-disciplinary collaboration.

These summer schools strongly influenced the development of computational linguistics at both the national and international level. It is widely acknowledged that they formed business executives of European CL, promoted the collaboration between the two main trends of CL (HTP and NLP) and the convergence of the methodologies based on rules with those based on quantitative data. Furthermore, institutionally speaking, the schools established a contact between European and American researchers, and contributed to the development of the state-of-the-art in crucial areas of CL, Linguistics and Artificial Intelligence. To give only some examples, in the 2nd school, comparison was made between data-driven and rule-driven approaches. During the 3rd school, a new linguistic theory (Lexical Functional Grammar) was founded in Pisa. It inspired the first dictionary, the LDOCE - *Longman Dictionary of Contemporary English*, to be designed according to the principles of CL. The LDOCE dictionary gave rise to a series of studies, which were later to develop through the researches of the IBM group in Yorktown, the ACQUILEX project and other later initiatives. The 4th Summer School witnessed the seminal activities of the group launching frame semantics in the bay-area.

For Italy, the summer schools also represented an unprecedented opportunity to import methods and techniques of analysis, which were later developed by ILC in original manner: it includes e.g. syntactic parsers in MAGMA-LISP - based on the *Augmented Transition Network* (ATN) model, integrated and optimised by statistical data, methods for knowledge representation, development of formal grammars, innovative approaches and methods in the field of computational lexicology, and lexicography.

2 - Laboratory of Computational Linguistics of CNR (1978-1980)

In the first years of the Laboratory of Computational Linguistics, research activities were systematically given priority over those of assistance and consultancy that had absorbed in the previous stages most of the human resources part [1]. The support and collaboration of the Scientific Council, and in particular of its President, Professor G. Nencioni, member of the CNR Scientific Committee for the Humanities and President of the Accademia della Crusca, were of great help during this period. The main lines of research activated in the early stages of the laboratory were the following:

- Development of the Italian Machine Dictionary, consisting in the digitisation of ca. 100,000 lemmas associated with an algorithm for phonological transcription and a morphological analyzer/generator;
- Development of methodologies to acquire linguistic information from machine-readable dictionaries;
- Feasibility studies for application to the Spanish and Latin languages of methodologies developed for Italian;
- Development of an Italian Parser based on ATN;
- Quantitative linguistics research on the archive of texts in *Machine Readable Form* (MRF) available at the Laboratory;
- Creation of an international network of textual databases - through the definition of effective exchange procedures - as an experimental phase preliminary to the design of an international standard.

3 - International Activities

This paragraph illustrates briefly the most important international initiatives and experiences, which started before the creation of ILC. Promotions of, and participation in, these activities, alongside many other international initiatives, were crucial to the creation of ILC, in particular for what concerns the different types of activities carried out therein, and the role and recognition of ILC in both the national and international context. These made it possible for the Institute to promote and disseminate methodologies, strategic principles and scientific paradigms (in Kuhn's sense), ensuring harmonious progress of our discipline with the strategic needs of the national community.

3.1 Machine Translation

In 1976, I was entrusted by the Italian Government to be part of a panel of experts intended to advice on the possible acquisition, on the part of the European Community, of the American system of Machine Translation (MT) SYSTRAN. With the collaboration of the German delegate H. Zimmermann, I persuaded the EC to promote - alongside SYSTRAN - a novel system for advanced translation based on European technology. In the same years, in my quality of President of the Scientific Committee of COLING '78 (Bergen), I emphasised the need for CL to demonstrate its application-oriented potential and to establish contacts with international Research Agencies such as the EC.

The activities, designed to prepare the grounds for what was later known as the EUROTRA European project for machine translation, required a delicate mediation between the French (Grenoble) and German (Saarbrücken) schools of MT, each trying to impose its own methods of analysis and generation. I suggested a general organization of the translation system's architecture based on the concept of *interface structure* (IS), which is a structure of semantic-syntactic representation of the text to be translated, specified according to a common representation format accepted and adopted by all participating groups (originally 7, later 9 European languages) and independent of the methods of analysis adopted. Each group could then use its own methodology to process texts in its own language (the source language), provided that IS representations were produced according to common specifications. A source language IS was then to be turned into a target language IS, which provided the basis for generating the surface text in the target language. The development of interface structures for a number of European languages represents one of the most significant achievements of EUROTRA. The key concept of IS, together with the basic architecture of the EUROTRA MT system, were maintained until the very end of the project.

EUROTRA did not produce an effective system of multilingual translation, but was decisive for the creation of a network of European computational linguistics institutes. It promoted the creation and organization of specialized teams in some countries where computational linguistics was not represented, and produced fallout know-

[1] This can be better understood if one considers that CNUCE was essentially devoted to the provision of services to the public.

how and experiences that were essential to the future development of computational linguistics in Europe. As far as ILC was concerned, participation in the Italian group of EUROTRA fostered technical competence in the field of lexical databases and of the so-called transfer grammars.

3.2 Language Resources

"Academic" lexicographers were the first to show an interest in the possibilities offered by the use of computers in lexicographic processing. This interest gradually shifted from the collection of texts in MRF and the production of indexes and concordances to the ordering of quotations and editing of lexical entries. The majority of "academic" lexicographical projects was lengthy and represented a heavy burden for the budget of the different Research Centres. In 1980, the *European Science Foundation* (ESF) asked me to conduct a survey on the European lexicological and lexicographical projects supported by public funding. The results of this survey showed that if a large part of these projects used the computer, inappropriate methods were exploited. In the light of these findings, the ESF urged me, in 1981, to organize a workshop on the *Possibilities and limits of the computers in producing and publishing dictionaries*, in Pisa. The main purpose of the workshop was to assess the state-of-the-art, evaluate the research priorities, formulate strategic recommendations and, if possible, minimise time and costs of the projects financed by public research bodies. The delegates of the European Research Centres federated in the ESF and those of the American *National Endowment for the Humanities* (NEH), managers of large lexicographical enterprises and experts in CL that I had chosen for their potential innovative contribution, participated in the discussions. Important reflections emerged, some of which are still valid today. In particular, since the various schools of thought in linguistics were unable to provide an adequate methodology for the descriptive meaning of extended sets of lexical entries, lexical semantics was to develop new theoretical approaches on more consistent and theoretically motivated grounds for the identification, definition and structuring of the meanings contained in dictionaries. Nowadays, the joint efforts of generative semantics, psycholinguistics (WordNet) and knowledge representation (ontologies) are engaged in the realization of this important task. It was also necessary to reflect on the innovative use of technology for the building of new types of dictionaries, e.g. for electronic publishing.

For the construction of tools that could help the lexicographers in their work, a study on the "modus operandi" of the academic lexicographer was to be conducted (these recommendations led, for example, to the construction of the ILC lexicographical workstation). The necessity to establish standards for the codification and exchange of textual data, linguistic analyses, and lexical descriptions emerged. Such recommendations were considered as the first concrete step towards initiatives of standardisation which were later established internationally, such as the TEI (*Text Encoding Iniziative*), and EAGLES (*Expert Advisory Group on Language Engineering Standards*).

As a result of this workshop, the ESF created a group of experts in computational lexicography (1983, Zampolli, Quemada, Zimmermann, Van Sterkenburg, Weiner). Later, in 1985, I was appointed member of the *Standing Committee for the Humanities* (SCH) of the ESF, as representative subject for linguistics and CL for the years 1985-1990.

The workshop on *Automating the Lexicon*, which I organized in May 1986 in collaboration with D. Walker, J. Sager, L. Rolling, and N. Calzolari, is universally recognized as the starting-point of the process which led to the establishment of the Language Resources (LRs) concept. The workshop examined research, current experiences and possible development of the activities on lexicons and corpora, with special regard to the multilingual environment. The final recommendations, transmitted to the EC in 1987, resulted in a series of European projects (ACQUILEX, ET-7, MULTILEX, MULTEXT, GENELEX, DELIS, etc.), organization and research activities (e.g. the 1988 *Summer School Computational Lexicology and Lexicography*).

ILC played a major role in the diffusion of a new paradigm, the so-called data-driven approach, characterizing the entire disciplinary field of Computational Linguistics. This is based on the use and study of extended collections of linguistic data and their descriptions, the so-called Language Resources (e.g. representative and tagged large-sized corpora, lexicons as complete as possible, vast language grammars).

In a meeting organized in Turin in September 1991 by ESPRIT and the *National Science Foundation* (NSF), which brought together 110 representatives of European research and 10 representatives of north-American research, spoken and written corpora and lexicons were indicated as common top-priorities for both communities. As a representative of European research, I introduced the expression Language Resources (LRs) to highlight the infrastructural role of the components, comparing it to the role of the basic resources (e.g. aqueducts, electricity, roads) needed for industrial development in a geographical area. I then suggested the expression in my report at the Danzin panel (1992). The final report of the panel validated this term (which became part of the terminology of the Commission - and consequently of the current literature in the field), as well as the definition of the LRs infrastructural nature.

The heavy costs for the creation of extended LRs require considerable organizational efforts. In the last decade, however, our community has been persuaded to engage itself in the construction and use of adequate LRs, owing to strong techno-scientific and economic-organizational pressures.

It was commonly agreed that the main activities which were to be carried out for LRs could be subdivided into four sub-sectors which I clearly identified and described for the first time in a communication held with Nicoletta Calzolari, during the workshop organized by the EC at Santorini in 1993, on the future of the Language Industry (LI) in Europe:

- Elaboration of consensual standards;
- Creation of the necessary LRs;
- Distribution and sharing of the resources;
- Creation of synergies among national projects, European and international projects, industrial initiatives.

The launch of the activities in relation to each of these 4 sub-sectors was one of the main research lines carried out by ILC, both in the national and international areas, especially through the proposal, the coordination and the realization of international projects [...]. These projects aimed to disseminate awareness of the central importance of LRs for CL in the different countries, and to respond -albeit partially- to the needs for LRs of the R&D community.

3.3 Language Industry

The great success met by some NLP systems in the mid80s highlighted the importance of carrying out automatically some linguistic operations and the maturity of existing NLP technologies. It was clear that CL was in a good position to respond to the profound and urgent needs of the emerging Information Society, in the direction of application-driven development.

I did my best to promote at ILC research activities that aimed at creating and experimenting prototypes of application systems and their components. I submitted to CNR the project *Methods and tools for Language Industries in Italian Cooperation*, which was later approved. Researchers, designers, national and international research agencies became increasingly aware of the strategic, industrial and cultural possibilities offered by the language industry, emerging as an independent sector in the information industries.

In line with this trend, the expression Language Industries (LI)[2] started to be used and covered both economical and commercial applications based on computational systems which automatically perform language operations and tasks that are an essential part of the application.

Other terms were rapidly introduced, drawing the attention of the public to particular aspects in the field. For example, in the 3rd European Framework Program of Research, the sector was called Language Research and Engineering; in the 4th one, Language Engineering, highlighting the need to develop the engineering aspects of the NLP technology so that it could be used in concrete applications. In the 5th Framework Programme, and in more recent American governmental programs, the expression Human Language Technology was used to emphasise the role that CL technologies can play for the development of a user-friendly, man-centred society.

In the preparatory phase of the 6th Framework Programme (2002-2006), now underway, which seems especially intended to expand European research and face highly innovative and priority themes for the future development of society, the treatment of language is considered within the framework of various actions, such as:

- Knowledge Management and Content Creation unit;
- Interface and Cognition Unit.

3.4 Networking

From the very beginning, my activities aimed to promote the techno-scientific and organizational collaboration between the different communities involved in CL. The following are only some of the initiatives I sparked in this direction.

3.4.1 Foundation and Management of ELSNET

In 1991, within the framework of the EC program ESPRIT, E. Klein (Edimburgh) and myself founded ELSNET, the *European Network of Excellence in Speech and Human Language Technologies*. ELSNET is a European forum dedicated to Human Language Technologies (HLT), which operates in an international context, bringing together ca. 150 nodes, both public and private. All the research areas of human communication in relation to language and discourse are considered within ELSNET, which clearly aims at improving HLT R&D in Europe, uniting the main operators of the sector and providing an open and proactive platform forum to:

- Launch innovative actions;
- Analyse the present and develop future visions;
- Encourage a common environment including resources, standards and assessments;
- Develop, share and exploit knowledge and experiences.

a) “Language Resources” Task-force of ELSNET

From its very foundation, ELSNET has set up a task force in the sector of LRs and their evaluation, whose main objectives are the following:

- Provide a platform to exchange information and know-how in the areas of language resources, evaluation and standards in Europe and with other emerging or high-technology countries;
- Disseminate methods and know-how on LRs generated within other European projects;
- Foster synergies between national projects and multinational actions;
- Promote limited, but well focussed initiatives in order to:
  - Better define the needs of the users;
  - Promote the creation of small prototypes/initial models of innovative resources involving a methodological or technical risk;
  - Explore the possibilities of organizing the activities needed to construct, maintain and update the resources presenting a certain grade of uncertainty;
  - Support other similar initiatives;
  - Stimulate the evolution of the sector.

These initiatives can then move along an independent route, but the variety and complexity of the know-how pooled by ELSNET and the comparative flexibility and adaptability in decision-making have already proved to be essential factors.

ELSNET has recently started an international extension action through contacts with other continents and the design of a roadmap for the research activities and applications that can reasonably be envisaged in this sector for the next ten years.

3.4.2 Language Resources Distribution

One of the major problems that should be solved to ensure reusability of LRs, with consequent saving of efforts, time and money, is the organization of a mechanism for the distribution of LRs. The Research Agencies of the American Government were first concerned about this issue, initially discussed in 1988 during a meeting in which I took part as European expert. On this occasion, the *Linguistic Data Consortium* (LDC) was constituted at the University of Philadelphia. LDC is supported through public funding (ARPA and NSF) and annual subscriptions by its associate members. Aware of the strategic import of the initiative, I proposed to the EC the constitution of a similar European organization.

a) RELATOR, 1993-1995

The EC published a call for a feasibility study, to which I applied with a proposal including LIMSI (CNRS-Paris, France), DFKI (Saarbrücken, Germany), the University of Edimburgh (United-Kingdom), CST (Copenhagen, Denmark) and ILC as project partners. I also acted as coordinator for the ensuing project, called RELATOR. Thanks to the scientific expertise of our Institute and its partners, the work of an American public relations journalist, the assistance of a Steering Committee (including the main European Industries of the sector and chaired by the Director General of the DGXIII), the aid of the Legal Offices of the Commission and of the CNRS Institute for International Rights, as well as the monitoring of a Committee of Learned People[3], RELATOR proposed the foundation of a European Association for Language Resources, which was registered in Luxemburg (ELRA-*European Language Resources Association*) in February 1995.

b) ELRA (*European Language Resources Association*)

ELRA is guided by an Executive Board of 12 members. ELRA has the mission of promoting LRs and of coordinating and proceeding with their validation, distribution and re-utilisation within a European framework. The ELRA statute specifies the following activities:

- Assess, select, implement the means necessary for the distribution of LRs;
- When suitable, organize and handle the acquisition of LRs from the producers and develop the technical and legal frameworks for their validation and distribution to the interested users;

[2] The expression Industries de la Langue was launched in 1986 at the International conference organised by the Council of Europe in Tours, France. It refers both to the activities where the computer may assist the human user in the Applied Linguistics traditional tasks (lexicography, translation, etc.), and to the activities for the development of new applications (man-machine interface systems, machine translation, etc.)

[3] Members of this committee are: Danzin, expert in Language Industries for the French gouvernment and for the EC; B. Quemada, vice-president of the Conseil Supérieur de la Langue Française; and B. Oakley, past-president of Logica and IST expert for the EC

- On request of the European organizations financing programs for the creation of LRs, provide advice on the distribution and validation of these resources;
- Function as source of information with regard to the contents and availability of LRs for all the parties involved in Europe;
- Identify the needs for LRs and stimulate the appropriate organizations for the creation of the LRs to meet these needs.

Unlike LDC, ELRA -whose activities were originally supported by a EC contract- has now become independent, thanks to the amount of resources distributed (ELRA holds a percentage on the income from resource selling) and to the subscription fees of member organisations. As a matter of fact, ELRA is not an Association of physical persons, but of Bodies having juridical identity[4]:

One of ELRA principal activities is the organization of the LREC conference (International Conference on Language Resources and Evaluation).

3.4.3 LREC, international Language Resources and Evaluation Conference

The LREC conference is intended to provide an overview of the state-of-the-art in the LRs area, to exchange information on both present and future activities, to discuss the language resources and their applications, to discuss methodologies and demonstrate evaluation tools, to explore opportunities and promote initiatives for international cooperation, etc. The need for a conference like LREC derives from the awareness that a large number of players working on various LRs aspects, concentrating on topics of considerable importance for their respective professions (e.g. linguists, computational linguists, publishers, software engineers, cultural operators, multimedia telecommunications and computer industries, language experts and language teachers, knowledge engineers, on-line service providers), belonging to different communities, with their own specific organizations and specific conferences, rarely had the opportunity to meet and exchange information and explore possible synergies and co-operations. The number of participants in the three editions of LREC seems to confirm that the conference meets the needs expressed by the HLT community.

3.4.4 ENABLER, European National Activities for Basic LanguagE Resources

The ENABLER network, that I had firmly been recommending since the times of NERC, and that has recently become an IST project financed by the EC, was launched in the light of the following considerations. Over the last years, the majority of national European projects in the field of HLT have identified LRs as the primary area to be supported by national funds, on the basis of consultations involving researchers, companies, service suppliers, etc. The availability of LRs is also a pressing issue, which touches directly the sphere of linguistic and cultural identity. It is a crucial precondition for the integration of a language in the Information Society, and of the citizens speaking this language. However, the LRs production costs are very high, and this can hinder (and has partially already done so) the involvement of industrial manufacturers, in particular the small and medium-sized enterprises, in the field of NLP. Furthermore, it has emerged clearly in recent discussions that the European Framework Programme as well as the American Agencies such as the NSF and DARPA are unsuitable for complete and constant support to infrastructural initiatives. It is clear that we can only produce synergies, scale economy, convergence and accumulation of the efforts, which are to provide the infrastructural LRs needed to realize the full potential of a global multilingual information society by:
- Combining the energies of different initiatives and different institutional commitments;
- Exploiting to its best the “modus operandi” of the national Authorities in different national situations;
- Meeting the needs and priorities of each industrial, R&D community, on the grounds of a clear distinction of tasks and roles for all those working in the field of HLT.

The Enabler Network aims at activating the progressive realization of this extremely urgent cooperative framework, supporting connections, establishing exchange mechanisms and encouraging the cooperation and interoperability of the results of national projects and activities. In many of the member countries of the European Union, important national authorities have financed these projects and activities in order to supply different types of LRs to the relevant languages.

Central aims of the Network are to:
- Strengthen and institutionalise - under the Commission's umbrella - the core of existent national initiatives, establishing appropriate connections with a regular, updated and structured record of the situation, publicly available for all organisational, technical and relevant information;
- Provide an official forum for discussion as well as discussion and mechanisms of general coordination and exchange of information, data, optimal practices, tools sharing, multilateral cooperation, and bilateral cooperation on specific subjects;
- Expand gradually the initial Network, identify and promote the inclusion of representatives from national complementary initiatives;
- Encourage synergies among the national activities and increase compatibility and interoperability of their results, facilitating technologies transfer among languages;
- Maintain the compatibility of the different LRs, by increasing, with national funding, the original nuclei produced following some common technical specifications of the European projects. This aims first at guaranteeing the implementation of a large infrastructural platform of standardized European LRs, which is an essential prerequisite for future large-scale multilingual LRs, secondly at exploiting the multilingual character of the European Union;
- Improve the visibility and strategic impact -with respect to the USA, Japan, etc.- of the different national activities, which result from the decisions taken by the politicians of the various countries on the grounds of similar needs and priorities;
- Provide a forum to discuss the industrial needs and formulate a common agenda of medium- and long-term industrial priorities;
- Promote the exchange of tools, specifications and validation protocols produced within national projects, to avoid duplication of efforts and dispersion or divergence of approaches;
- Contribute to the creation of a European organization for standardization of the metadata description of the different types of LRs (spoken language resources, written language resources, multimedia and multimodal resources);
- Promote the industrial exploitation of LRs;
- Encourage and contribute to the process of design and progressive implementation of a global cooperative network recognized internationally for the provision of LRs.

The participants in the network are Institutes which, in their own countries, coordinate governmental national programs for LRs, or which have been institutionally entrusted to produce them.

[4] The most prominent European, American and Asian research centres (e.g. ILSP in Greece, DFKI in Germany, LIMSI in France, CST in Denmark, SPEX in the Netherlands, Cervantes in Spain, Carnegie Mellon in the US, etc.) as well as some of the major industrial players (Microsoft, Philips, Siemens, Xerox, Sony, Panasonic, etc.). ELRA is supported by the French Ministry of Research [...]. Recently ELRA had a contract supported by the EC and by NSF for the standardisation of its catalogue and procedures with the LDC ones.

a) International Committee for Written Language Resources and Evaluation

In the framework of the ENABLER Network, we have just set up an International Committee for Written Language Resources and Evaluation, which brings together the groups and international scientific organizations involved in CL and LRs, and in particular those operating in the field of written, spoken, and multimodal LRs.
Main objectives of the Committee are to:

- Constitute a "proactive" forum for the exchange of techno-scientific information about current projects;
- Define a minimum set of LRs which should be made available for as many languages as possible, and transmit this message to the competent national and international authorities;
- Organize joint programs of research and production of common multilingual resources;
- Facilitate participation in the network of European national projects (ENABLER), of national projects of other countries (American, Asiatic, etc.) which have expressed the intention to collaborate.

4 - Institute for Computational Linguistics, ILC 1980-2001

In 1980, the Laboratory was turned into a CNR Institute.
To ensure collaboration and synergies with the major foreign research centres, to increase the financial endowment coming from CNR with funding coming from other sources, and to perform an action of mediation between Italian and foreign research as envisaged by our Statute, it was necessary to:

- Continue and strengthen the presence of ILC at the international level;
- Perform a profound action of strategic and scientific innovation also at the international level, encouraging the development of our discipline along the research lines described so far, with the view of moving out of the impasse the discipline had reached, and of better responding to the main needs of our national community.

This required the promotion of new scientific paradigms, the creation of new bodies, the launch of international and national projects for the management of which I was assisted by my collaborators, the participation in the decisions of international Research agencies, and discussions with the delegates involved in the infrastructure responsible for making fundamental strategic choices.

4.1 Main Research Trends of ILC

I shall report here the main objectives of the studies underway at ILC[5]. As a whole, according to the opinion expressed by ILC Scientific Committee, which was confirmed by CNR Scientific Committee for the Humanities to which ILC belongs, a coordinated set of activities reflecting the state-of-the-art of Computational Linguistics was formed. Its scientific merits were widely acknowledged internationally as a result of the position of leadership the Institute has reached in the most important European projects in this sector.

4.1.1 Standards and Computational Language Resources

The term *computational language resources* (LRs) designates, as we have already seen, large sets of linguistic data, accompanied or constituted by formalised annotations and representations at different levels of linguistic description, used to build, extend, make fully operational and assess models, algorithms, components and systems for automatic processing of spoken and written language. The main types of resources generally referred to are written and spoken corpora, computational lexicons, repositories of grammatical information, terminological collections, as well as software tools for the creation, acquisition, validation, access, analysis of such resources, and, more generally, basic software methods and components ensuring the fundamental functions of NLP systems.
From the second half of the '80s, the lack of adequate Language Resources was identified as one of the main obstacles to the success of CL R&D activities for the construction of appropriate models dealing with language in real use and for the transfer from prototypes to applicative systems to be used in real operational contexts.
A large part of our research activities in this sector required the involvement at different stages (direction, planning, definition, specific methodologies, implementation, assessment, demonstration) of Communitarian/international projects and/or of programs of national interest that we promoted and often coordinated[6].

a) Definition of Standards for Computational Lexicons and Written, Spoken and Multimodal Resources

This work is principally carried out in collaboration with the project EC-NSF ISLE[7]. ISLE takes advantage of the joint work of the most prominent groups of experts - both academic and industrial, European and American - operating in the sector. Results are thus based on a broad consensus for the main leading players in the field. Aim of the research is to define standards for:

- The formalization and codification of lexical information at different levels of linguistic description (in particular for semantics), for both monolingual and multilingual lexicons, with particular regard to the needs of multilingual computational systems;
- The treatment of multiwords, at the level of both lexical codification and text annotation;
- The design of ontologies for generic and specialised lexicons, and for different types of applications requiring deep text understanding;
- The annotation of written corpora at various levels of linguistic and conceptual description, related both to the evaluation of automatic parsing systems, semantic disambiguation, etc., and to the annotation required by application systems such as information extraction, machine translation, etc., and to the use of annotated corpora for machine learning;
- The annotation of dialogue at the morpho-syntactic, syntactic and pragmatic levels;
- The annotation of multimodal corpora;
- The definition of LRs validation criteria.

b) Creation of Lexical Resources: Lexical Knowledge Bank of Italian and "Dynamic" Information Acquisition

Lexical Knowledge Bank for Italian (LKB-It)

The recognition of the central role played by a computational *Lexical Knowledge Bank* (LKB) for Italian, to be used in different systems and applications, is part of a growing trend in CL (and theoretical linguistics in general), which places the lexical component at the centre of any NLP system. Our aim is to supplement the Italian lexical database available at ILC with further information at the morphological, syntactic, semantic and collocational levels, and pursue its transformation into a knowledge base (KB) containing information whereby logical inferences can be drawn. Coverage extension is envisaged both in terms of quantity, i.e. with a larger number of entries and in terms of quality, i.e. with the levels of information (addition of different types of entries, such as proper names, multiwords, terminology, further semantic information, etc., addition of bilingual and/or multilingual links). In particular, the morphological, syntactic and semantic information envisaged by the PAROLE/SIMPLE model, with

[5] The directors of research are: Nicoletta Calzolari (Standards and Computational Language Resources), Eugenio Picchi (Methods and Technologies for Multifunctional and Textual Linguistic Databases), Andrea Bozzi (Computational Philology), Irina Prodanof (Models and Methods for NLP), Giovanna Turrini (Language Technology for Didactics and Diability), Antonio Zampolli (Coordination of National and International Activities)

[6] For example the European projects ACQUILEX, DELIS, ET-7, ET-10, SPARKLE, MULTEXT, MULTILEX, NERC, EAGLES, EuroWordNet, Onomastica, LS-Gram, COLSIT, PAROLE, SIMPLE, MATE, NITE, ISLE, MUSI, POESIA, the American project NSF, XMELLT, and the national projects underway.

[7] I am the ISLE coordinator for the the European side and Nicoletta Calzolari is responsible for the working group on Computational Lexicons in Europe.

sense relations encoded following the Ital- / EuroWordNet model will be cast into a comprehensive and coherent architecture.
With a view to an application-oriented usage, we have also designed a software interface managing different lexical resources from a unified perspective, and adapting them to the specific needs of either human or system users. The lexical models adopted here play a central role in the functioning of any NLP prototype or system. This brings us to some of the most controversial and heavily debated issues concerning the interaction between lexical coverage and intended applications:
- How to settle on in a principled way the types of lexical information (especially semantic) required by a specific analysis/generation system;
- To what extent it is convenient to encode lexical knowledge without having first assessed its usability in a concrete system;
- What sense distinctions, among those reported by normal look-up dictionaries, are worth maintaining when encoding the entries of a lexical knowledge base for NLP.
In this respect, it is particularly instructive to look at the outcome of evaluation campaigns such as SENSEVAL, an international initiative co-organized by ILC for the comparative evaluation of automated sense disambiguation systems, dealing with corpora annotated on the basis of semantic networks of the WordNet type.

“Dynamic” acquisition of information from textual sources

An essential aspect of the development and richness of LKB is represented by the interaction of manual and automatic or semi-automatic modes of extension. These are linked to the complementary notions of “static” and “dynamic” lexical resources. From a theoretical point of view, “static” lexical resources can not adequately cover any possible corpus and/or meet any application requirements. Consequently, these basic “static” resources need to be combined with increasingly refined tools for the “dynamic” acquisition of lexical information (at different levels of linguistic description) starting from different types of corpora to reflect real communicative contexts, and/or for different application needs. The acquisition activity, in which our group made pioneering work, should consist in enriching or tailoring the basic resources to different types of texts and/or applications. The acquisition process itself involves a cycle of:
- Corpus analysis/annotation;
- Acquisition of information at a specific linguistic level;
- Assessment of the results;
- Iterative analysis/annotation refinement (back to step *i*) at a more complete or 'higher' level of linguistic analysis.

c) Written, spoken and multimodal corpora

Written corpora, i.e. collections of texts in electronic format, represent:
- The natural knowledge source for studying and describing the features of language use in different communication contexts, in the monolingual as well as in the multilingual contrastive context, to be used when developing the components of computational systems for automated processing of texts and contextually-embedded utterances;
- Reference standards against which the performance of language engineering methods and systems engineering should be assessed.
Following the EAGLES specifications, we are implementing:
- The continuous extension of an Italian reference corpus within the framework of a monitor corpus, which is a reference corpus periodically updated to reflect the changes in a language over the years, to be used on the market in some applications which include much larger amounts of textual data, appropriately encoded and annotated at the linguistic level;
- The creation of a first nucleus of parallel corpora (formed by texts translated into different languages), for which there is a strong request from industrial operators;
- The annotation of written and spoken corpora at different levels of linguistic description;
- The creation of software components for multi-level automatic annotation of written and spoken corpora in XML;
- The production, starting from the Italian reference corpus, of a new frequency lexicon of the Italian language;
- The collection and annotation of multimodal corpora.
In order to make an adequate use of the corpora that are being created for many languages, it is necessary to:
- Define objective criteria for the composition of different types of corpora, which can be considered as “suitable” samples illustrating a certain linguistic universe;
- Develop methods for the automatic annotation of corpora at complex linguistic and extralinguistic levels, especially with respect to classes of specific application domains and tasks;
- Implement methods and annotations allowing (semi-)automatic extraction of linguistic knowledge from the corpora;
- Increase the variety of corpora in different domains and for different application purposes (corpora made of special types of language, e.g. the collection of data recorded from children speakers to study the development of linguistic competence and which can provide information for the development of didactic tools and components for the consumers market).

d) Multilevel linguistic annotation of dialogue corpora

Large quantities of annotated spoken data are necessary to model the effects of the different variability sources (environmental and acoustic context, modes and channels of communication, emotional and social condition of the speaker, dialects, etc.) onto linguistic units such as phonemes, pronunciation and sequence of words, etc. Most of the available tools and technologies designed for automatic processing of language have been designed for application to written and monologic language. The application tools thus based on the currently available technologies cannot be used for the processing of dialogue interaction, leaving out a sector of the information society which is rapidly evolving, given the current growth rate and technological development in the field of telecommunications.
The research here aims at:
- Validating the linguistic specifications for dialogue annotation at the morpho-syntactic, syntactic and pragmatic levels on a representative dialogue sample;
- Expanding the reference corpus (200 man-machine and 200 man-man dialogues collected) within the framework of the national project TAL, in collaboration with the most important industrial operators working in the field;
- Annotating progressively a corpus of 1000 dialogues, using tools of automatic, specifications-compliant pre-annotation and an editor interface for manual correction.

e) Multimodal corpora

In order to model interactive man-man dialogues as well as man-machine ones and to play an appropriate role in the recent trend towards a human-centered use of the computer, it is necessary to avail oneself of multimodal resources (spoken, written, gestual, etc.).
We communicate not only through words, but also through intonation, glances, hands, gestures, facial expressions, etc. These modes are integrated and complemented to provide different types of information for communication purposes. Nowadays there is a growing impulse within HLT to take the various communicative channels into account. Using the results of an experiment carried out within ELSNET, alongside the recommendations of the NIMM group of ISLE, ILC started the creation of multimodal corpora and their annotation, as well as the design of software tools to handle and annotate multimodal data, in collaboration with the HTL-NITE project.

f) Modular architecture of software components for multilevel automatic annotation in XML of written and spoken corpora

Software architecture is currently being studied for the annotation of texts at different levels of linguistic analysis. Each level is designed as conceptually independent and declarative, linked to a common document containing the original text (raw or normalised), and interfaced with other levels of annotation through XML pointers. The modularity of the architecture makes as natural as possible the introduction of further annotation levels layered upon the main ones, namely morphology, morpho-syntax (POS tagging), immediate constituency syntax (chunking), functional syntax and pragmatics. Clearly, each layered annotation module can be used as an independent component and integrated in other software systems designed for specific applications.

4.2 Methods and tools for humanistic research

The study, prototyping and use of computational methods supporting research in various humanistic disciplines, with particular regard to the linguistic, literary, philological and lexicophraphic fields, has, to a certain extent, represented the platform upon which our group has developed. At the international level, the activities of ILC aim to:

- Promote and exchange know-how, methods, resources and tools with other CL trends;
- Incorporate the possibilities offered by recent technological development into the procedures studied for the humanists;
- Base the design of the components and computational tools on the study of the modus operandi of the different humanistic disciplines, so as to develop dedicated modular and flexible workstations;
- Promote awareness of the role of humanistic text processing in the Information Society, and encourage the humanists to assume their role through the use of innovative methods and the promotion of national and international projects.

It should be observed that, from a technological point of view, the number of possible uses of the computer in the human sciences is constantly increasing. In addition to the text processing possibilities, predominant at the beginning, some new media (sound, images, colour, etc.) allowed to develop new methodologies for the use of the computer in the humanistic disciplines. This opens up new perspectives of dialogue between socio-economic partners and human sciences, the latter being the main suppliers of linguistic, literary, and cultural contents for the new digital “media”[8]. However, it should be noticed that the processing of images also supports the development of applications for textual and philological analyses[9]. The possibilities offered to the researcher vary according to his specific disciplinary interests: from the support in stylometric studies to the application of hypertextual and hyperlinking mechanisms for data-mining techniques on historical data, documents, literary texts, to the study of the formal properties regarding the structure of the data and algorithms assisting the researcher in his institutional tasks, or to the use of LRs, methods and techniques of (semi)-automatic linguistic analysis, both for the identification of information, structures and linguistic features in different language corpora (ancient or contemporary, children's or adults', from different linguistic communities) and for the study of their development. These studies have progressed in two distinct but complementary directions.

4.2.1 Methods and technologies for multifunctional textual and linguistic databases

The first trend appeared within the programs used for decades by the CNUCE users for electronic text processing, and now appearing in interactive form (also remotely accessible through Internet). They have progressively been supplied with additional functions: morphosyntactic statistical taggers, maximum flexibility of the contextualization algorithm, visualization of the data and statistical calculations in real time, capacity to operate simultaneously on different alphabets and languages, interrogation of the text through group (families) of words formulated by the user or derived automatically from associated lexical knowledge bases (words that are linked by several types of semantic relations e.g. synonymy, hyponymy, hyperonymy, etc., e.g. extracted from ItalWordNet), systems for automatic alignment of parallel texts, automatic search for translation equivalents in comparable parallel corpora, etc. These functions have proven to be very useful for linguistic and literary studies, especially for those concerning the history of thought. All these modules are integrated into an effective lexicographic workstation with components aiming at assisting the lexicographer in his institutional activities (design of the entry structure, choice and classification of contexts, etc.).

4.2.2 Computational philology

The second trend concerns the joint processing of image and text transcription, mainly at the service of philologists. The general aim is to create a set of methods, software and experimental data integrated into a workstation which allows humanist researchers, especially philologists, to use easily and efficiently the different technologies borrowed from the fields of artificial intelligence and of the processing of images in text study.

With the development of digital technology for libraries and archives, new frontiers have opened up not only for philological and linguistic research, but also for the preservation and fruition of cultural heritage. The research, which integrates products implemented at the Institute[10] with specialized industrial products and others available on the market, offers a series of functions, that can be combined in various ways according to the application underway, and that can be summarized as follows:

- Database management functions: coordinated functions of information retrieval, management of the different informative annotation levels, support to collaborative work (handling of multiple access to a single document), input/output functions, standards and handling of descriptive formats conformant to the different metadata systems in use;
- Image processing function: support to image segmentation, to dynamic text-image matching, enhancement and electronic restoration, automatic scanning, automatic transcription of ancient type-setting with the aid of an NLP component, protection of the integrity and authenticity of documents, handling of multiresolution formats;
- Philological work function: support to text analysis and handling of annotation using linguistic knowledge sources (statistical indexes, thesauri, lemmatizers, etc.), to hypermedia connection of various documents, as well as handling of normalized standardized tagging (SGML, HyTime and XML);
- Support and utility functions: functions for annotation (separated from the document) at various levels, functions following indepen-

[8] This opens up new job opportunities in the area of cultural activities, where information methods may be used: as a matter of fact, requests for courses at our institute are made by people coming from this sector.

[9] If some of the research activities conducted at our institute use multimedia for teaching or learning purposes or for the creation of a software environment for multimedia semiautomatic tagging of video material (to be combined with the use of ontologies in which the concepts, when possible, are connected with images or prototype frames), other research in the philological-textual field applies text-image association both in the context of cultural services provided to the public (e.g. digital libraries), and in the context of linguistic-philological research (e.g. for aligning the image of a text with its transcription).

[10] For the processing both of the texts and the images, the commercial products available are by no means sufficient: they do not have the specialization required by the peculiarity of the applications, which originates in the structure of the formal specific data of the disciplines and in the functionality of the algorithm using the structure and which must be based on the methodological analysis of the researcher’s operating mode.

dent segmentation routes, verbal and visual indexes of words and places, etc.

In particular, in order to strengthen this trend, I have promoted the following international initiatives:

- A series of meetings between experts to draw a 'roadmap' of the sector, moving out of the present impasse characterised by constant technological refinement not always corresponding to effective methodological innovation from scientific research point of view;
- A series of round-tables held in conjunction with the ALLC-ACH Conferences, where representatives of agencies typically interested in human sciences, such as NEH, ESF, etc., representatives of the Information Society (NSF, CE, etc.), researchers, and even commercial users can discuss priorities, strategies, possibilities of collaboration and the most efficient methods to incorporate the contribution of humanistic disciplines in the Information Society;
- A meeting with computational lexicographers, “academic” lexicographers and “commercial” lexicographers to examine the state-of-the-art over 20 years after the Congress sponsored by the ESF in Pisa in 1981, and to explore and evaluate the synergies that new techno-scientific development can establish.

It should be observed that, because of the possible connections with cultural content (leading for example the EC to launch a programme called eContent outside the Framework Programme), research in this field has found increasing complementary support in various initiatives[11].

As far as electronic publishing is concerned, the number of editorial enterprises using our software and methods for textual and multimedia applications on CD, HDVD, etc., is rapidly increasing. In particular, it is worth mentioning the lexicographical projects both for the production of “traditional” dictionaries on CD-ROM and for the creation of “innovative” dictionaries, including on the one hand descriptions of various linguistic levels, on the other hand annotated reference texts.

4.2.3 Models and methods for Natural Language Processing, and mono- and multilingual application prototypes

As previously mentioned, one of the results of the first editions of the International Summer School was to disseminate methods (e.g. ATN-and/or CHART-based) for the analysis and generation of expressions in natural language. These methods based on formal rules and mechanisms of inference from knowledge bases have characterized CL in the first decades. A large part of our activities at the Institute have been performed in this direction; e.g. we have written a great number of formal Italian grammars designed for different types of parsers. Since the establishment of the paradigm of the language industry, as soon as robustness, linguistic coverage and minimum overgeneration have become the principal requisites for the components of input text analysis, it has been necessary to resort, as in many other countries, to the use of statistical models and shallow methods of parsing.

a) Statistic methods for induction of computational language models

The need to face the variety and complexity of technical and sector languages, as well as language fragmentation and context dependence of language-in-use, requires an increased use of inductive methods for automatic learning, recovery, organization and handling of linguistic knowledge (for example in purpose-oriented dialogue interaction i.e. business transaction, exchange of information, etc.). This contributes to render more robust and wider-coverage traditional rule-based NLP tools. In this respect, the Institute is actively engaged in implementing a series of methods for inducing computational models of written and spoken language starting from attested linguistic evidence, possibly annotated at one or more levels of analysis. In particular, research is subdivided into four levels of linguistic analysis:

- Morphology: methods for automatic identification of classes of morphologically complex units that have a large impact on robust processing of real texts, e.g. derivatives, compounds, multi-word units, etc;
- Syntax: mathematical models of the correlation between i) morphosyntactic tags, ii) non-recursive syntactic constituents (or chunks), and iii) functional tags, aimed at developing Markovian and analogy-based methods for the analysis of sequences of such units;
- Lexical-semantics: methods for automatic induction of lexico-semantic classes and lexical taxonomies based on specific technical domains;
- Pragmatics: methods for the acquisition of correlation measures between linguistic structures and communicative acts.

The research method takes place in two phases:

- Phase A: investigation of the statistically critical causes of failure of automatic analysis and preliminary study of the correlation among the relevant linguistic units for each of the above levels of analysis;
- Phase B: development of software components for the induction of computational models at each of the levels mentioned above.

b) Shallow parsing

Recently, we have designed and implemented a number of parsing tools. These can analyse different text types robustly and identify syntactic nuclei and relations which are conducive to further processing stages for the identification (at various levels of granularity) of proper “semantic” aspects (semantic classification, links to thesauri and ontologies, etc.). This approach to syntactic analysis is based on two tools, the *Chugger* and the *Functional Analyzer*.

The Chugger implements finite-state parsing technologies and realizes, at the same time, morpho-syntactic tagging of words in context (identification of the syntactic category of a form in a given linguistic context) and segmentation of the text in non-recursive syntactic constituents (chunks). Therefore, it combines the functionalities of a tagger with those of a chunker. The Chugger has been developed from a pre-existing module, Chunk-it, which realizes the segmentation in chunks of previously tagged texts. The functional analyzer uses as input the output of the Chugger and, based on a finite-state dependency grammar, it recognizes the principal grammar relations between the different sentence elements: e.g. identification of the subject, complements, complex nominal nuclei, etc. It represents the principal component for semantic information extraction and includes an interface with a database of lexical information, both syntactic and semantic, extracted from the computational lexicons available.

This set of components, augmented with further basic processing tools, responds well to the needs of different applications requiring a “robust” analysis of input texts[12]. They have recently met considerable interest from Web companies, which are considering including these levels of analysis in their products to improve the 'performance' of their systems. Synergy with private companies offers an important contribution to research, providing a working application environment and already active users.

Beside the work carried out on shallow and stochastic parsing, further research is under way, following the traditional, or ‘classic’ approach of the ‘70s and ‘80s. This aims at creating a cycle of theoretical analysis, design, experimentation, and methodology for the main NLP applications. For example, the growing request for programs handling linguistic and textual information, connected with the increase of distributed and on-line information services, confirms the need for advanced models and prototypes in the areas of retrieval and handling of on-line documents as well as for flexible and modular NLP interfaces. The fields of application are manifold, but the required technology boils down to precise modules and prototypes, such as the interpretation of queries in natural language, the identification of informa-

[11] For example, special (BIBLOS) and strategic (Cultural Heritage) CNR projects, MURST projects (ex 40%, CIBIT, Parnaso Programs Bibliofilo), projects of the Ministry of Labour (ADAPT: TECLA), EC projects (BAMBI, MEMORIA) and EC-NSF (CHLT).

[12] For example in due projects recently approved by the EC: MLIS-MUSI and POESIA

tion patterns in on-line documents, the presentation of the information, and the generation of sentences in natural language.
Our research aims at:

- Carrying out experimental work on methods for natural language interpretation, information retrieval and representation, and answer generation, possibly through multimedia technology;
- Producing programs for the analysis and handling of linguistic information independent of the type of application and the repository of possibly available linguistic knowledge (syntactic, lexical, semantic, etc.) programs that can be combined modularly to form application prototypes;
- Providing pre-industrial prototypes of information management systems;
- Offering methods for the analysis of application-driven problems and for the development of systems handling and presenting information.

Different models are studied and put to use, either based on formal rules, on model inference from data analysis, on knowledge representation, or on a joint usage of the three approaches mentioned above. Research, development and validation of computational models for language processing can contribute to a number of theoretical and practical issues, including a better understanding of the linguistic and cognitive basis of language usage through computer simulations and the implementation of prototypes and tools capable of addressing the needs for innovative applications based on, or including, NLP technologies. A clear example of this development is the "computation", in one form or another, of the language meaning, whose processing lies at the root of the implementation of new tools aiming at improving the processes of communication, information retrieval and text comprehension.
The application-oriented goals currently pursued are the following:

Models handling textual information for public services

Texts have different "information density": parts of them can be considered irrelevant with respect to classification of their content. The development of pre-processing methods for filtering out insignificant parts, as well as local analysis of information dense "text islands" is a basic step in tailoring NLP methods to specific applications.

Job offering and job seeking: automatic service on Internet

The aim is, on the one hand, that the offered jobs meet efficiently the need of the job seekers; on the other hand, that professional training courses can be planned. The service is envisaged for Internet. The research activity consists of testing and validating text processing methods and tools for the extraction of various linguistic levels of information (lexical, conceptual, syntactic) from texts. The research also aims at designing a model able to process the extracted information to prepare a description of the professional competence conveyed by an offer.

Querying in natural language a manualistic knowledge base

Interpretation of clarification questions does not require "full" text-understanding capabilities. It is generally sufficient to identify those significant text elements, which allow, through a mapping mechanism, to recognize the constituents of a possible reply. The main research goal is here to identify and develop methods for filtering out "useless" parts of a question and for processing the semantic content of useful parts, with a view to eliciting relevant responses from a knowledge base.

Multilingualism and on-line services

Services for textual information management are more relevant if they are developed in the framework of on-line multilingual research. The merge between multilingual processing methods and document processing methods is neither immediate nor obvious. Nonetheless, it is a necessary important step forward towards the realization of a real international service. The UNL project promoted by UNU/IAS (*United Nations University/Institute for Advanced Studies*), that I coordinate for Italy, as Director of the Italian Language Center, is carrying out a feasibility study for the creation of a permanent on-line service. This service would allow to translate the documents in a natural language (source language) into a kind of "universal" language (*Universal Networking Language*, UNL, based on a set of very general logical relations), which is then generated in the different target languages. The UNL project is developing a network of "national centres", each of which is responsible for providing the programs necessary to translate the texts written in the respective national language into UNL, and vice versa. The languages so far included in the network are Italian, German, French, Spanish, Portuguese-Brazilian, Russian, Lithuanian, Arabian, Hindu, Indonesian, Mongolian, Chinese, Japanese and Swahili.

4.2.4 Language technology for didactics and disability

a) Application of CL tools and methods to teaching and special teaching[13]

In collaboration with the Department of Computer Sciences in Turin, our Institute has implemented a hypermedia language laboratory for the study of Italian as both native and foreign language. The software is an interactive didactic environment offering a motivating context for activities generally considered rather difficult and boring, such as dictionary look-up and enrichment of the lexicon. The system brings to the surface the knowledge and experiences that are associated by a child, who then uses them to build up some new information. The language laboratory is made up of strictly connected multimedia tools: *Addizionario*, a dictionary for children, written and illustrated by the children themselves, and the *Activity Book*, a creative tool by which children can build up a kind of customised dictionary, changing and tailoring to their needs the information imported from the core dictionary, or adding new words to the existing 1,000 ones, which include information such as definitions, examples-in-use, drawings, sounds, etc.
The product is addressed to different categories of users: to children of compulsory school for activities including lexical enrichment and language reflection, to teachers for the preparation of didactic units designed to meet the needs of the students, to psychologists and therapists to help identify disturbances in development and learning and for rehabilitation activities, and to the editors of young learners' dictionaries at the same time "appealing", easy-to-use and respectful of the capacities, tastes and interests of the users.
The *Addizionario* dictionary will be experimented in different social contexts and in other countries , with both normal and impaired subjects. The analysis of a corpus of child language should provide information for an optimization of didactic tools.

4.2.5 Coordination of national and international activities

I have described above some of the goals of these activities, consisting in the design, promotion, organization and coordination of activities which aim at pushing the state-of-the-art through collaborations between the different protagonists from Italy and abroad, public and private bodies to converge towards innovation in sectors we consider crucial, from both the techno-scientific and organisational point of views, for the development of a discipline meeting the needs of the Global Society, and in particular the priority needs of our country. Summing up:

[13] The special CNR project "Use the Computer in the Teaching of Special Languages" had made it possible to realize a preliminary feasible study.

[14] Experimentations are underway for Mexican, Welsh, English.

a) At the national level

The Institute for Computational Linguistics participates in the following promotion and coordination activities:

- Establishment of a national network: this network is intended to bring together the different groups involved in NLP-related work in our country, in order to identify needs and priorities, define joint programmes of work, monitor the progresses achieved, avoid duplication of efforts, create convergence and promote sharing of language resources, facilitate transfer of technologies, integrate different types of competencies, develop new training curricula, etc.;
- Promotion of national programmes exploiting and expanding the know-how and results offered by the Institute, able to respond to needs identified as high-priority for the development of the national community at large. The majority of the themes included in these programmes correspond to research and studies carried out at ILC. The Institute can reinforce its research activity through synergies and institutionalised collaborations with highly reputed national representatives, fully accomplishing its function of national coordinator in the field of CL, which is one of the institutional tasks of its Statute. This will also make it possible for ILC researchers to rely on additional human and financial resources.

b) At the international level

At the international level, with the assistance of my collaborators at ILC, I have been engaged in complying with the tasks envisaged by the ILC statute (namely, establishing appropriate links between national and international activities). This has been achieved through initiatives which aimed first at strengthening synergies with NLP Research Centres in the most advanced countries (both European and non-European, and with public and private bodies) then at promoting international development of the state-of-the-art in strategic areas, in an attempt to include the Italian language within the multilingual framework of the global society. These initiatives can be summarized as follows:

- Bilateral agreements (at present with Institutes in Japan, the USA, Bulgaria, Spain, Cuba, Mexico, Argentina, France);
- Participation in the activities of Associations, as responsible for coordination through the presidency of these Associations (e.g. ELRA, PAROLE, ALLC, EURALEX, etc.), or through my participation in their executive committees (ICCL, AILA, ACL, ACH, SIGLEX, etc.);
- Organization of international events: for example the International Conference on Language Resources and Evaluation; different workshops on various CL issues, with European, Asian, or North-American participants; round tables during Congresses (COLING, LREC, ALLC, etc.);
- Participation in international networks: in the Network of Excellence for Natural Language and Speech (ELSNET), where Italy is represented in its Management Board; in the coordination group for the national projects of the European governments (ENABLER); in the International Committee for Language Resources, OntoWeb (Ontology-based information exchange for Knowledge Management and Electronic Commerce);
- Promotion of the participation of the Institute in international projects (more than 30 European projects over the last years);
- Promotion of the collaboration between the European Union, the United-States, and other technologically advanced countries, in the NLP sector.

5 - Programs of National Interest

Two programs of national interest[15] have been launched recently and promoted by ILC on the grounds of experiences, know-how, and perspectives acquired through the international activities previously described and the direction of ILC research. The two programs of national interest are:

- TAL (national infrastructure for language resources in the field of automatic processing of written and spoken language), costing ca. 2.6 million euro, financed by MURST for a total of ca. 1.8 million euro in compliance with law 46/82 art.10. The project, entrusted to a group of 13 private bodies[16], according to constitutive law, was brought to completion in autumn 2001.
- The *Computational Linguistics: monolingual and multilingual research* cluster, for a total cost of 4.6 million euro, financed by MURST, in compliance with law n.488 of 19/12/1992 (Cluster 18), for ca. 3.1 million euro, is divided into 8 projects, each entrusted to an executor[17], that avails itself of numerous collaborations in various juridical form.

<u>ILC in the new CNR Network</u>

On 1st January 2002, ILC became independent and autonomous within the CNR network, thanks to its worldwide-acknowledged position of Centre of Excellence, both at the national and international level.

The major role of leadership of ILC, also gained thanks to the relentless effort the Institute put into maintaining its identity and integrating different aspects of NLP, is reflected in its capacity to:

- Attract considerable external funding (national and international) in extremely competitive areas;
- Influence the strategic activities implemented by national and international bodies;
- Participate in the executive roles of the major national Institutions in the field, and recently;
- Participate in numerous Expressions of Interest for Networks of Excellence and Integrated Projects for the 6th Framework Programme (2002-2006);
- Promote an Expression of Interest grouping the major CL Centres in Europe and all over the world.

The three major thematic sectors appearing in the constitution of the Institute intend to continue and further develop the traditional lines of activities described so far, both in the framework of ILC institutionalised national vocation and in the context of international finalized and externally-funded projects.

These sectors are:

- Design of standards and development of computational linguistic resources;
- Models and methods for natural language processing, mono and multilingual application prototypes;
- Computational methods and tools for humanistic research, with particular regard to linguistic, literary, philological and lexicographical disciplines.

The “new” ILC intends to continue its major role in supporting the CL sector, by defining and stimulating a number of coordinated actions which respond to the needs of our country in this field: from the promotion of strategies and programmes of national interest financed by MIUR (Ministry of Education, of University and of Research), to proposals of independent university curricula (masters, doctorates), to the connection between national and international communities, to the proposal and coordination of initiatives and communitarian and international projects, to encouraging technological connection and transfer towards the industry.

[15] ITAL, special CNR project, had made it possible to carry out a first, although limited, feasibility study

[16] The partners are: CPR, Consorzio Pisa Ricerche; ITC, Istituto Trentino di Cultura; CSELT, Centro Srudi e Laboratory Telecomunicaioni; Synthema; CVR, Consorzio Venezia Ricerche; CERTIA, Centro per la Ricerca, Sviluppo, Formazione nelle Tecnologie e Applicazioni Informatiche; Quinary; Alceo; Computer Sharing; Delco; GST, Gruppo Soluzioni Tecnologiche; Interactive Media; NECSY, Netwrok Control Systems.

[17] The subjects involved are the following: CPR, Pisa; CIRASS, Napoli; THAMUS, Salerno; ICL-CNR, Pisa; Synthema, Pisa; Istituto Universitario Orientale, Napoli; Dipertimento di Scienze Storiche del Mondo Antico, University of Pisa; Sportello per la Cooperazione Scientifica e Tecnologica con i Paesi del Mediterraneo (SMED) of CNR, Napoli.

A Few Words from Nancy Ide

It is difficult to know what to say in tribute to Antonio Zampolli , whom I had the good fortune to count not only as a colleague, but also as a close personal friend for 20 years. A rehearsal of his enormous contributions to computational linguistics and humanities computing would be stating what all of us already know, and would almost certainly fail to do justice to his impact on both of these fields.

Antonio's great talent was as a facilitator. He had an uncanny knack for understanding even the most technical details of a new or unfamiliar problem, assessing the relative importance of each, and determining with almost miraculous speed what the key issues were and what the next step should be. He would often sit at meetings, drawing pictures (faces, usually) on his notepad for lengthy periods and appearing to be completely disinterested, only to interrupt at some point and drive right to the heart of the issue. He similarly would almost immediately assess, with almost alarming accuracy, each individual involved, their motivations and agenda and the politics of the situation. Then, he would do what was necessary to make what had to happen, happen. Some people felt he was serving his own agenda, and sometimes he was; but the top item on that agenda was always to get an area he felt was vital addressed, by whatever means; and he was almost always right about what was vital. The evidence of his visionary abilities is clear: Antonio was a champion for language resources since the 1970's. Fifteen years ago, a conference on language resources would have had relatively few attendees, while now LREC is one of the most massively attended conferences in the field. (Of course, Antonio also knew the appeal of an exotic location…)

Antonio's other great talent was his ability to put work behind every evening, and put as much energy into making the evening enjoyable-for everyone, not just himself - as he put into making the day's work productive. He seemed to take it as a personal challenge when confronted with someone who wouldn't loosen up; and more than once, I've seen him evoke some very uncharacteristic behaviors from people in this category. Conference dinners will never be the same without Antonio.

Antonio's legacy is the projects he fostered: development of language resources, standards for their representation, international collaboration, and much more. We will keep these projects alive in his memory, and raise a few glasses of wine at each conference dinner. Here's to you, Antonio; we miss you.

A Few Words from Martin Kay

Every second Saturday through the summer of 1956, I left Victoria station in London charged with escorting some three hundred holiday makers who had never left their homeland before to a variety of sunny places between Genoa and Rome. On the intervening Saturdays, I left Roma Termini to collect those that had been delivered two weeks before by somebody else, and took them back to London. On arriving in Rome, I took a long hot shower in the place that large Italian stations provided for that purpose and headed for the Piazza Navona to feast my eyes on its three Bernini fountains, to watch cheerful children who play and laugh and never cry, and to consume a cool half bottle of Orvieto. I loved the Piazza Navona and could regain my physical and mental strength there better than anywhere else.

Some forty years later, the telephone rang in the middle of the night at an hour when good news rarely arrives. It was Antonio saying that he had kept his part of our bargain, and found me an apartment on the top floor, overlooking the fountains. It remained for me to keep my part and find him a house on the steep and winding section of Lombard Street in San Francisco.

Antonio was very fond of San Francisco, and my wife and I generally made a point of taking him there for dinner when he was in the Bay Area. Naturally, we did not seek out Italian restaurants, but he had a sense that told him when there had been an Italian hand in the preparation of the meal, and he marched off into the kitchen, emerging some minutes later with a Sicilian or a couple of Tuscans and a bottle of Brunello di Montalcino. On one occasion, he found an Italian chef with a German assistant and kept everyone in the car in hysterics, including himself, all the way back to Palo Alto by declaiming loudly in Italian with a heavy German accent various things that the assistant might have said. “Kvanti ziamo a kvesto tavólo?”

Antonio was the little magician who could turn a fantasy into reality and reality into a source of marvelous merriment. But beneath the surface, there was deep and genuine concern for his friends and colleagues, for the field of endeavor he had done so much to bring into being, for Pisa, for Italy, for Europe and the world. Oh, yes, and for the Dolomites. Lombard Street was fantasy, but Cortina d'Ampezo was the center of the world, and Antonio had a house in the mountains nearby in which I believe he was able to feel a peace that he could not find in any other place. That was his Piazza Navona. My wife and I visited him there once and he took us walking among the towering chimneys and along the ledges that make that place unique in the world. When we reached a break in the path that Iris thought she might not be able to cross, Antonio said he would carry her. He was not joking. In this place, he was sure footed and knew exactly what he could do. When the ledge became so narrow that I no longer had the courage to continue, there was no macho urging forward, because he knew that you could not be safe if you did not feel safe. He was a man of the mountains: careful, and professional, and competent.

In 1978, Professor Antonio Zampolli founded what I believe to have been the second department of computational linguistics in the world (after Gothenberg the year before), though he had been leader of the linguistic division of CNUCE since 1968. But his reputation as the father of European computational linguistics goes back to the early seventies when he organized a truly remarkable series of summer schools in Pisa to which many of today's foremost research centers owe their origins. He scoured the world for anyone with experience in this fledgling field and brought them to pass on what they had learned. Before our very eyes, he turned what had been a hobby or a passing fantasy for a few people into a discipline and a profession. The hotels and parks and bars and restaurants of Pisa were awash in algorithms, and lexica, and morphemes, and parsing, and semantics. We were born too late to have been with Sartre and Hemingway in Paris, but we were with Zampolli in Pisa when the history of what matters to us was in the making. If the leaning tower has been closed to the public since those days, it is probably because the people who should have been working on keeping it from falling were seduced away from engineering and geology to language and computing.

Antonio was not simply a member of just about every organization that connected computing with the humanities or ordinary language; he was vigorous member and, in many cases, a founding member. The International Committee on Computational Linguistics had frequent cause to be grateful for his innumerable contributions to their work and, most especially, for his bringing the fifth International Conference on Computational Linguistics to Pisa in 1973. This vintage year of the period of the great Pisa summer schools was also the one in which the Association for Literary and Linguistic Computing came into being with Antonio as one of its founding members. In 1983, he became president of that organization, a position that he retained for the rest of his life.

I vividly remember an occasion on which Antonio had been assigned by the organizers of a workshop that we both attended to respond to my contribution. To my shame, I allowed his restless activity during my presentation to strengthen the conviction that my efforts were about to receive less than the treatment they deserved. They got what they deserved, and more. With charm and humor and even deference, the weaknesses I had thought might go unnoticed were revealed one by one, and remedies were suggested for errors I had not even suspected. I was put, gently but firmly, where I belonged. From the experience, I learned a lot about myself that I should have learned earlier. But I learned a respect for Antonio Zampolli which was to grow continually deeper over the years. He gave so much of his life to making good things possible for other people that one was always in danger of forgetting that he was a man of great intellectual power and creativity.

I am able to appreciate, probably better than most, one particularly arcane activity in which Antonio's creativity was manifested, namely that of causing the punched-card machinery that were the precursors of modern computers do things beyond the imagination of their designers. I have this privileged perspective because, at about the same time in the early history of our field, Antonio was working with Padre Busa in Milan, I was working with Margaret Masterman in Cambridge. He was trying to derive a phonemic representation from Italian written texts and I was trying to parse sentences with a formalism now mercifully forgotten. We were both trying to do these things by finding ingenious new ways of wiring the plug boards of various punched-card machines designed to meet the needs of accountants. He succeeded.

I have said that Antonio devoted most of his time and energy to making good things possible for others. In the latter years, he spent much of his time in Brussels and Luxemburg and Washington, not only securing the support of his institute but also tending to health of his discipline and bringing people together in whose potential interactions he saw benefits for a wider community. He was justifiably proud of his achievements as a match-maker, forging collaborations and even life-long friendships among the most unlikely partners. At every Coling conference, he organized a panel on the funding of research in computational linguistics throughout the world, mainly so that young people in our field should be exposed to as many opportunities as possible.

I have tried to limit the flow of anecdotes that flood the mind when one thinks of Antonio because, like no one else that I have known, he enriched the lives of everyone that he touched with unforgettable little personal things. I will end with just one more. After a noisy dinner with much wine and laughter during one of the summer schools, a number of us emerged from a restaurant long after its accustomed closing time into the relative cool of the outside air. Everything was illuminated by the eerie green light of a full moon. Antonio stopped as he came out of the building and suddenly fell quiet and serious. “Everybody follow me”, he said, and set off towards the grassy close that surrounds the cathedral, the baptistery and the tower. Everyone followed in silence. When we arrived, he said “This place is called Piazza dei Miracoli-The Square of Miracles. I will show you why.” And slowly he walked through the square, following a particular path that he knew well, pointing silently now at a carving on a building, now at a formation of stones in the wall, now at a silhouette on the other side of the square. Nothing was said, but we all understood the reason for the name. One more miracle had occurred there by the little miracle maker of Pisa.

A Few Words, in French, from Bernard Quemada (English version below)

La disparition d'Antonio Zampolli a consterné ses amis et ceux qui ont travaillé avec lui, mais aussi toute la communauté des chercheurs en Traitement automatique de la langue. Leur dette s'ajoute à celle de ses collaborateurs de l'ILC de Pise qu'il a accompagnés et soutenus sans relâche. Sa personnalité et ses actions ont eu un retentissement international et elles ont profondément marqué le champ multidisciplinaire de la linguistique informatique.

Ses qualités d'animateur, reconnues à travers le monde, ne sauraient occulter la part déterminante qu'il a prise dans la promotion de la philologie électronique et de la linguistique informatique dès leurs débuts en Europe. Nul autre n'a été aussi présent aux étapes clés de leur développement. Sans mesurer sa peine, il s'est placé avec ses collaborateurs à l'avant-garde des orientations de la recherche les plus novatrices, s'efforçant de répondre à toutes les sollicitations, souvent au détriment de ses travaux personnels. Les réalisations de son laboratoire et les responsabilités qu'il assuma avec succès à la tête d'associations internationales ont fait de lui un remarquable défenseur des enjeux stratégiques des Industries de la langue dans la Société de l'Information. Il fut écouté et entendu par son gouvernement et les instances européennes, sans lui, beaucoup de projets nationaux ou multilatéraux n'auraient pu voir le jour, et l'ELRA que nous connaissons n'existerait pas, du moins sous sa forme actuelle.

Après Padre Busa, Zampolli s'est placé en tête des promoteurs des traitements électroniques, s'attachant à développer et à faire partager les savoirs et les ressources. Son parcours est éloquent à cet égard. Mieux que tout autre universitaire de sa génération, il a expérimenté les mutations scientifiques et techniques d'un demi-siècle de progrès et maîtrisé les connaissances philologiques, linguistiques et informatiques correspondantes. Les analyses de la Summa à Gallarate, les traitements statistiques pour sa thèse “littéraire”, sa participation aux corpus informatisés de la Crusca, l'avaient familiarisé avec la production des données textuelles et les meilleures solutions techniques pour les élaborer. De ce fait, il a entretenu de nombreux et fructueux contacts à l'échelle mondiale, outre-Atlantique en particulier. Dès ses premières responsabilités, il s'était attaqué aux coûts élevés des ressources informatisées en préconisant la coopérations pour produire des données, et des politiques d'échanges pour leur réemploi, militant aussi pour la généralisation de normes performantes, toutes choses qui demeurent au premier plan.

Ses liens avec IBM à Pise, puis son intégration au Centre national de calcul universitaire (CNUCE) où il a pu disposer de moyens puissants, lui ont fait partager les vues concrètes et réalistes des ingénieurs. Devenu le premier enseignant de linguistique informatique dans une faculté des lettres italienne, il a toujours pris en compte les impératifs industriels et stratégiques qui échappent si souvent aux spécialistes des humanités.

Sans ménager ses forces, il a fait de Pise le principal pôle d'attraction du Vieux Monde pour ce qui touche aux applications du Traitement automatique de la langue. Les Cours d'été qu'il organisa au cours des années 70, en accueillant des stagiaires et des enseignants du monde entier, ont eu un grand retentissement. Les Colloques, où il invitait les meilleurs spécialistes, ont jalonné l'évolution du domaine ; ils sont à l'origine d'importantes innovations et ont donné un élan décisif à l'usage des corpus et des données dictionnairiques. Missionnaire infatigable, il a couru le monde autant pour s'informer que pour faire savoir. Dans les multiples rencontres internationales auxquelles il a participé, ses interventions ne passaient jamais inaperçues, et il a contribué lui-même à la création et au fonctionnement d'organismes internationaux à la tête desquels il a assumé des charges importantes avec compétence et autorité. Les rencontres du LREC sont les dernières manifestations auxquelles il a attaché son nom. Leur succès, dont il était fier à juste titre, atteste la validité des options qui ont conduit son activité pendant les dernières décennies.

A vrai dire, il était, particulièrement doué pour y réussir. En véritable Italien, il alliait l'esprit de finesse à l'esprit de géométrie que les Français opposent souvent pour définir les caractères. En ingénieur linguiste et linguiste ingénieux, il avait une aptitude rare à saisir les points essentiels, les orientations et transformations qui s'imposaient, sans oublier la promptitude avec laquelle il savait mettre à profit les occasions favorables. Tout cela avec beaucoup de témérité tempérée par davantage encore d'anxiété.

A côté de cette image du spécialiste, nous conserverons vivante l'image du solide grimpeur, nostalgique de ses chères “Montagnes” ; du commensal qui, avec autant de psychologie que d'astuce, lisait dans les lignes de la main de ses voisins de table ; du collègue qui animait les soirées d'adieux avec un talent et un tonus peu communs. Personnalité forte et attachante, aussi riche que contrastée, voire contradictoire, tantôt mûre ou enfantine, affirmée ou inquiète, sérieuse ou rieuse, mais toujours généreuse, dévouée et fidèle en amitié. A notre tour de lui garder la fidélité du souvenir.

A Few Words from Bernard Quemada (English translation)

The loss of Antonio Zampolli has dismayed all his friends and those who worked with him, as well as the whole community of researchers in Natural Language Processing. Their own debt adds up to that of his collaborators at the Institute of Computational Linguistics in Pisa, whom he accompanied and supported without respite. His personality and his actions have had international repercussions and they have deeply marked the multidisciplinary field of computational linguistics.

His qualities as a leader, acknowledged all over the world, should not hide the decisive role that he played in promoting computational philology and computational linguistics, at their early start in Europe. None but him has been as active in all the key steps of their development. Without measuring his effort, he and his collaborators have placed themselves at the avant-garde of the most innovative research directions, making every effort to respond positively to all requests, often at the expense of his own work. The achievements of his laboratory and the responsibilities that he successfully took on as the leader of several international associations, made of him a remarkable defender of the strategic stakes for Language Industries in the Information Society. He has been listened to and heard by his government and by the European authorities. Without him, many national and multilateral projects would not have been launched and ELRA, as we know it today, would not exist in its current form.

After Padre Busa, Zampolli appeared as a leader for promoting electronic text processing, taking particular care in developing and sharing knowledge and resources. In this respect, his trajectory speaks for itself. More than any other academics of his generation, he has experienced the scientific and technologic mutations that took place in half a century of progress and he mastered the corresponding knowledge in philology, linguistics and computer science. His analysis of the Summa in Gallarate, the use of statistical approaches for his PhD in the humanities, his participation to the Crusca's electronic corpora had made him familiar with the production of text data and the best technical solutions to design them. Indeed, he maintained a large number of fruitful contacts world-wide, especially across the Atlantic. As soon as he took on his first responsibilities, he grappled with the issue of the high costs of electronic resources: he recommended cooperation for data production; he promoted an exchange policy for reusing them; he also defended the wide dissemination of efficient standards. All those tracks remain today of primary importance.

His links with IBM in Pisa, and then his joining the CNUCE (National Academic Centre for Computation) where he benefited from powerful means, made him share the concrete and realistic point of view of engineers. He was the first one to teach computational linguistics in an Italian faculty of arts, while he has always considered industrial and strategic needs, which often escape the attention of specialists in humanities.

Without sparing his forces, he made of Pisa the main focus point in the Old World for what concerns applications of Natural Language Processing. The Summer Schools that he organised during the 70's, which gathered trainees and teachers from all over the world, were received with great success. The Conferences, where he used to invite the best specialists, have turned out to be milestones in the evolution of the domain; they have initiated meaningful innovations and gave a decisive momentum to the use of corpora and terminological data. Indefatigable missionary, he roamed the world as much to gather information as to disseminate it. In the numerous international meetings which he participated in, his interventions were never left unnoticed and he personally contributed to the creation and the functioning of international bodies, at the head of which he took important responsibilities with competence and authority. The series of LREC conferences are the last events to which he attached his name. Their success, for which he was rightfully proud, proves the validity of the options which guided his activities in the past decades.

As a matter of fact, he was particularly talented in succeeding. As a genuine Italian, he combined a subtle spirit and a sense of geometry, two features which French people tend to oppose when defining personalities. As an engineer in linguistics and an ingenuous linguist, he had developed a rare ability in understanding the key points, in knowing the directions to take and in identifying the transformations which had to be conducted, without forgetting to mention the swiftness with which he was able to take advantage of favourable opportunities. All of this, with a lot of rashness in his personality, moderated by even more anxiety.

Beside this image of a specialist, we will keep alive in our mind the image of the sturdy climber, nostalgic for his dear “mountains”; the table companion who would read the hands of his neighbours with as much psychology as cleverness; the colleague who was able to liven up farewell evenings as no one else, with talent and energy. A strong and engaging personality, rich and contrasted or even contradictory, sometimes mature, sometimes childlike, self-assured or apprehensive, serious or humorous, but always a generous, loyal and faithful friend... It is our turn to keep him faithfully in our memories.

Antonio Zampolli Prize for

“Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation within Human Language Technologies

Antonio Zampolli, a pioneer and visionary scientist, was internationally recognized in the field of computational linguistics and Human Language Technologies (HLT). Through the establishment of ELRA and the LREC conference, he also contributed geartly.

To reflect Antonio Zampolli's specific interest in our field, the Prize will be awarded to individuals whose work lies within the areas of Language Resources and Language Technology Evaluation with acknowledged contributions to their advancements.

The prize will be awarded for the 1st time in May 2004, at the LREC 2004 conference in Lisbon (24th - 30th May, 2004)

For more Information: www.lrec-conf.org