Introduction of the Conference Chairman: Antonio ZAMPOLLI

The first LREC, held in Granada in May 1998, registered an unexpectedly high number of submissions testifying to the perceived need of a common venue for the variety and richness of activities in the field of LRs and evaluation.

The further increase in the number of submissions received for this Second LREC2000 is an implicit recognition of the success of the First, and of the contribution of this initiative to the formation of a R&D community in a field of increasing relevance in the context of the blooming Information Society (IS).

Numerous actors are working in different sectors, on different aspects of LRs, focusing on issues of particular relevance for their professional interests: linguists, computational linguists, language engineers, publishers of multimedia products, cultural organisations, software and telecommunication industries, education and language technologists, knowledge engineers, service providers on telecommunication networks, etc. As they belong to different communities, which have their own specific organisations and conferences, they seldom have the opportunity of a common venue to exchange information and explore possible synergies and co-operation.

LREC aims to provide such a venue, promoting the awareness that all those working for LRs will benefit from considering themselves as members of a well-identified field. As stated in the Conference Announcement, the aim of this Conference is "to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding ongoing and planned activities, language resources and their applications, discuss evaluation methodologies and demonstrate evaluation tools, explore possibilities and promote initiatives for international co-operation in the areas mentioned above".

In the Introduction to the Granada Proceedings, I summarised my perception of the main challenges for the future of our fields, both from the organisational and from the technological point of view. I think that this general framework still holds. In my personal opinion, for some of those topics work is effectively taking place; some topics have not yet received the attention they deserve; strategical choices are still pending for some relevant organisational problems.

In the last two years we have witnessed a remarkable acceleration in the development of Internet, mobile and broadcasting networks, digital multimedia, etc.

Natural languages are the main vehicles for digital content products and services in business, education and culture.

The cultural treasures of different nations should be made available to everyone. Citizens of various countries wish, and should have the right, to access the digital contents and services in their own languages and using appropriate modalities.

An increasingly high volume of electronic information is becoming available in languages other than English. This requires a concerted effort in the customisation of content for the needs of different cultures, the expansion of technology to different languages, the development of a variety of enabling interfaces: spoken language, text and other modalities.

In these Proceedings, we observe, in fact, an increase in the numbers of papers on research or projects aiming at answering the following needs:

- the provision and use of multilingual and cross-lingual LRs (corpora, lexica, tools, etc.) which are necessary to develop and test multilingual and cross-lingual applications;
- the recognition, annotation, management, retrieval of content in spoken language and text: consider, for example, the increase in the number of papers devoted to semantic and conceptual annotation;
- the provision of resources and technologies for spoken dialogue understanding, and in particular for the design and collection of innovative resources for studies of and applications based on multimodality;
- the evaluation of resources, techniques, tools, systems for dialogue management and content processing.

Research reported in the Proceedings on these and on other issues, which were already mentioned as a part of the urgent research agenda for our field in my Introduction to the Granada Proceedings, confirms that LRs and evaluation are inter-linked at multiple levels and that the LR field is a clustered field, including not only production tasks but also research, directed to provide general methods and tools which should be considered as core enabling technologies, and a substantial part of human language technology (HLT).

LRs have the function of providing the linguistic specific knowledge necessary for operating on a given language.

In the previous Introduction (p. XVII) I suggested the hypothesis that in many cases it would be possible to transfer methods, technologies, algorithms from one language to another, provided that adequate compatible LRs exist for the target language.

Last year, the second HLT Call of the Fifth Framework Programme explicitly asked for submissions for this type of projects.

Recent trends and progress reinforce the observation that methods involved in annotating corpora and in acquiring and structuring knowledge from them are increasingly connected with various aspects of knowledge engineering. This long-term research challenge involves methodological and technical problems very similar to those encountered in such major HLT applications as information extraction, filtering, summarisation and retrieval, and the management of content in general. In addition, the success of methods based on empirical evidence, derived from corpus analysis, brings up a new challenge: to integrate the data-driven approach with the rule-based approach, and to generalise the results so far obtained, in a sparse and uncoordinated way, with empirical methods, to conjure an overall, possibly theoretically motivated, picture, encompassing both the approaches.

The papers in these Proceedings also confirm that progress in the state of the art in LRs and progress in spoken language, text and multimodality processing should advance hand in hand.

If we had real-size lexicons with very fine-grained semantic/conceptual information, would there be systems (non ad hoc toy systems) able to use them?

It seems that there is a loop between i) lack of suitable, large-size and knowledge intensive resources (lexicons and corpora, with many different types of syntactic and semantic information encoded), and ii) the ability of systems to use them effectively: the two paths should be pursued in parallel, they should closely interact with each other, and be gradually integrated.

It seems to me, moreover, that not much work is being done on a related issue, which was also listed in the research agenda proposed in the previous Introduction (ibid.), and which — I believe — deserves, on the contrary, serious attention because of its theoretical, practical and organisational implications, namely: which features should a given LR possess in order to be transported for reuse in another task or domain, and which are the best methods for providing an easy and cost-effective customisation of existing resources for different systems, a central requirement for

LR producers and distributors to meet the specific demands of developers[1]. This seems to me a special case of the more general well-known question of portability of components and technology, which has bedevilled the field of computational linguistics from the beginning, and clearly impinges on the possibility of boosting the state of the art on the foundation of results already achieved.

This question is also related to an as yet unsolved general policy and organisational issue.

As stated in the Granada Introduction (pp. XXI–XXII), the infrastructural role of LRs has obvious implications for policy and organisation at national and international level.

The IS and Technology are the driving forces for radical transformations in the organisation of social, economic and cultural life worldwide, and LT will play a key role in the accessibility and the usability of the IS infrastructure, in all its sectors, from information handling to human computer interaction to technology enhanced human to human communication. In addition, LT will mediate access to, and gain full benefit from, our various cultures and heritages.

Only languages for which adequate LR products and systems have been developed will be available over the IS network.

The availability of adequate LRs in a language is the key condition for the development in it of applications and services that are informed by LT. LRs are the most expensive part of any LT system.

The infrastructural role of LRs requires that LRs are made timely available for as many languages as possible. All these considerations lead to the question of who has the responsibility to make LRs available for a given language.

It has been repeatedly affirmed that each State "owns" its national language and should take the responsibility of providing the corresponding infrastructural LRs.

In fact, in Europe, several member States of the EU have decided to support, by means of projects "of national interest", the production of "generic" LRs, i.e. LRs which are not bound to be used for a specific application but should be considered multifunctional, such that they provide repositories of linguistic data and knowledge, which should drastically reduce the cost, time and effort of building LRs customised for (a class of) domain or task specific applications.

These large-scale generic LRs are often produced by extending the initial nuclei of small-scale generic LRs created within European projects according to common specifications based on internationally agreed standards/guidelines, thus offering possibilities of maintaining technical coherence, interchangeability, interoperability among the extended LRs, and in the end, of connecting them with interlingual links to build multilingual LRs.

---

[1] A major question is: does the constitution of generic LRs represent a means to efficiently and effectively reduce the cost of producing LRs directly usable for a specific application?
Probably the answer is different for different types of LR and of information provided. First of all, let me note that the very existence of ELRA is based on the concept of reusability of LRs: for example, the very successful spoken corpora produced by the SPEECHDAT projects are currently much in demand as foundations for the development of different applications. We should try to: 1) examine, for each type of resource, the classes of applications in which it is reused; 2) analyse which existing resources of a certain type are not reusable or in fact are not reused by applications, and why; 3) identify which features a given resource should possess in order to be reusable for a given class of application.
A generic corpus is normally a corpus constructed to study/represent the "general" language. Linguists have always asked the question if "general" language really exists, or if it is to be considered a juxtaposition of different special languages.
In a lexicon we certainly can find information which is true and useful for any application (e.g. part of speech) and its customisation concerns more the representation format than the content. The same could be said of syntactic and a part of the semantic information. The situation of conceptual information is however different: this information seems to be clearly domain/task dependent, and it is an open question whether it is worthwhile to represent generic conceptual information, and whether it is possible to build an increasingly growing knowledge base juxtaposing representations of conceptual knowledge related to different domains, using the same or compatible structural frameworks for the various domains.
In a recent discussion, industrial players reported that, whereas multilingual IR and MT could well require, at least partially, different information, it is very unlikely that a customer would accept to pay the cost of building two different lexicons, one for MT, and one for IR.

That, of course, would require a concerted effort, and the ENABLER initiative, jointly launched by ELRA, ELSNET, PAROLE/SIMPLE and EAGLES, is an example of the strong demand of our R&D community to the Commission to contribute to ensuring the required co-ordination by applying the principle of subsidiarity.

In other words, we are asking the Commission to take a strategic role, in addition to supporting application-oriented projects and the ad hoc LRs each of them needs.

The often-asked question of if and how reusability is dealt with in the Commission's current approach, and in particular how duplication of efforts will be avoided in the provision of LRs for projects, pertaining or not to the same cluster, can be seen as a special case of the more general question of how European projects — in particular in the field of HLT — could produce not only specific demonstrators of products, but also general technical and scientific advances, reusable by the entire R&D community, a goal that seems moreover to be at the centre of an explicit strategy of the North American Funding Agencies.

Even if national authorities would take responsibility for the provision of monolingual LRs for their own languages, in this way countering the market forces which privilege the more widely-used and economically-important languages, the problem of and responsibility for a multilingual LR policy remains.

The Commission, naturally, has to face the consequence of future extension of the Union to new countries, which will include Eastern European languages. But the globalisation of the IS is already posing the problem of interaction with language communities outside Europe. LRs will be the basis for network economy and social relationships across the continents. On the one hand, this extension will require an increasingly selective approach in deciding what technological developments can be supported. To this end, predictive evaluation could be very useful. On the other, this will call for open and well-organised international cooperation in the field of LRs[2].

It is vital, I believe, that the results achieved in the last decade, through cooperative efforts and a sometimes painful process, are not dispersed or lost, but preserved and put to use: the recognition of the infrastructural role of LRs and of the need to avoid duplication, ensuring reusability, concentrating and cumulating efforts; the progress towards the creation of an infrastructure dedicated to LRs; the promotion of consensual standards; etc.

We are still in the initial phase of the process. As exemplified above, substantial research efforts are needed to respond to a number of impending scientific and technical challenges and problems, whose solution is central for written and spoken language processing, and evaluation.

The present embryonic infrastructure must be reinforced, so that the same infrastructure is able to co-ordinate and perform, avoiding duplication, different complementary tasks: to provide and update the general repositories of linguistic data and knowledge which should be available for as many languages as possible; to produce at reasonable cost and in due time customised LRs to answer specific requests of developers; to offer services the LE community urgently needs: information, consultation, validation, etc.

---

[2] LRs are a clear case for international cooperation. In fact, standards for LRs, and the creation and distribution of LRs have been indicated as clear priorities for transatlantic cooperation in the post-LREC98 Workshop on cross-lingual information management (E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, [Eds]., "Multilingual Information Society: Current Levels and Future Abilities", to be found at http://www.cs.cmu.edu/~ref/mlim/index.html), jointly sponsored in Granada by the NSF and the European Commission.
This indication was confirmed in the special session on HLT at the Conference organised on "New Vistas in Transatlantic Scientific and Technical Cooperation" (Washington DC, June 1998), and two of the five projects jointly launched by the NSF/DARPA and the CEC in the first call of the HLT line in the Fifth Framework Programme deal with these topics.

We are very grateful for the participation of national and international Funding Agencies in LREC: the strategy they adopt will play a key role for the future of LRs and evaluation.

Acknowledgements

It is my duty and pleasure, in my capacity of Chairman of LREC, to thank all those who have contributed to the preparation of LREC and the publications of these Proceedings. LREC has been made possible due to the generous support of Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis and Gregory Stainhaouer.

The members of the ELRA Board have endorsed with enthusiasm our proposal of making this Conference biennial, and have been generous with practical advice and suggestions.

The Local Committee, George Babiniotis, Stelios Bakamidis, George Carayannis, Christophoros Charalambakis, Nikos Hatzigeorgiu, Ioannis Dologlou, Evita Fotinea, Maria Gavrilidou, Michael Kopidakis, Stella Markantonatou, George Papakonstantinou, Stelios Piperidis, Kostas Rekatsinas, Gregory Stainhaouer and Michael Strintzis, has spared neither energy nor effort in preparing LREC2000, in order to make it possible to maintain the high organisational standard set by the first LREC in Granada.

I would like to express our deepest gratitude to the eminent members of the International Advisory Committee, Sture Allen, Souguil Ann, Roberto Cencioni, Zhiwei Feng, Hiroya Fujisaki, Efstratios Galanis, Angel Martin Municio, Mark Maybury, Bernard Quemada, Gary Strong, Athanassios Tsaftaris, Piet G.J. Van Sterkenburg and Jialu Zhang, coming from Academia, Industry, International Organisations, whose prestigious co-operation clearly underlines the high cultural, scientific, social and strategic role and implications of the topics of the Conference.

Each of the about 400 submissions has been reviewed by at last 3 reviewers: they deserve particular acknowledgement for having accepted to review a number of submissions far greater than that originally foreseen. In fact, given the quantity of submissions received, the Programme Committee was forced to assign, on average, to the members of the Scientific Committee, nearly twice the number of submissions they had accepted to review, while keeping unchanged the time span allowed for the review process.

As for the first LREC, the Programme Committee has been at the heart of all the scientific, organisational and financial measures and decisions. The members have practically followed day-by-day the preparation of the Conference, dedicating an impressive quantity of time and effort to monitoring progress and solving problems. The Committee has worked along the lines already proven for Granada, collectively sharing responsibility for the overall scientific and organisational structure of the Conference, which has been established on the basis of the results of careful preparatory work, which was distributed among the Committee members for 3 major areas: NLP, speech and speech evaluation, terminology and evaluation for NLP.

Let me now express my deep and sincere gratitude to the individual members of the Programme Committee, who have been of invaluable help for me in a task which grew continuously with the increasing number of participants and the differentiation of the initiatives.

Joseph Mariani, for sharing day-by-day with me the responsibility for sometimes difficult scientific and organisational decisions, and for providing invaluable scientific competence and administrative experience.

Harald Höge and Bente Maegaard, for contributing greatly to the scientific programme of the conference, and Harald, for being instrumental in keeping the financial situation under control.

Khalid Choukri, for inspiring everyone with his enthusiasm, energy and efficiency, and the entire ELDA staff, for their support.

Nicoletta Calzolari, for her dedication and invaluable contribution in designing the structure and finalising the organisation of the scientific program.

In addition to his work as a member of the Program Committee, George Carayannis has dedicated an invaluable part of his already busy time to all the organisational aspects of LREC.

His very positive attitude has been a continuous and encouraging source of optimism for all of us during the preparation of the Conference, in particular when facing the enormous logistic problems of adapting the prestigious Zappeion Megaron to the needs and requirements of the Conference: he has taken upon himself and his team not only the risk and the burden of planning and executing this transformation, but also of finding the funds necessary to cover the cost of this operation, totally unexpected and well in excess of the initial budget.

George Carayannis has generously allowed the ILSP personnel to dedicate a large effort to LREC. I would like to express to the ILSP personnel in general the gratitude of the Programme Committee and, I am sure, of all the participants. I would like to thank, in particular, for her dedication Ms Evita Fotinea, who had handled efficiently, as events manager, many time consuming and demanding problems related to the organisation of the Conference.

The workshops, whose Proceedings are published in separate volumes, have contributed greatly to the scientific relevance of the Conference, both for the quality of the work and the choice of topics. Their organisers deserve the gratitude of all the participants at LREC.

The active participation of outstanding representatives of the major Funding Agencies has given a particular significance to LREC, and has confirmed, together with the participation of National and International Authorities who have honoured LREC with their presence, the high strategic value of LRs and Evaluation for the development of a user-friendly Information Society.

I would like to personally thank all the authors of the papers published in this volume: the number, variety and quality of the contributions will provide one of the most complete overviews and reliable definitions of the field of LRs and Evaluation, of its state of the art, of the progress accomplished in the last two years, and of the main trends and research directions.

This large participation would not have been possible without the precious co-operation of the Associations and Consortia who have joined ELRA in promoting LREC (ACL, ALLC, COCOSDA, EAFT, EAGLES, EDR, ELSNET, ESCA, EURALEX, FRANCIL, LDC, PAROLE, TELRI, etc.), and of the major national and international organisations, including the European Commission — DG XIII, ARPA, NSF, the IC/863 HTRDP Project (China), the National Natural Science Foundation of China, the ICSP Permanent Committee (Korea), the Natural Language Technical Committee of JEIDA (Japan) and the Japanese Project for International Co-ordination in Corpora, Assessment and Labelling.

In addition, I would like to thank all of the Authorities, who have given time and resources in support of LREC: General Secretariat of Research and Development of the Greek Ministry of Development; Greek Ministry of National Economy; Greek Ministry of Culture; Greek Research and Technology Network (GRNET S.A.).

Many thanks are also due to the companies and organisations which support LREC 2000, namely COMPAQ; GE Capital, Information Technologies Solutions; Eugenides Foundation; Singular SA; Baan Eastern Europe Localisation Centre SA and Consorzio Pisa Ricerche.

I would like, in particular, to express our thanks and gratitude to the staff of the ELRA distribution agency (ELDA) who contribute to the organisation of LRECs and provide the necessary support during the preparatory phases as well as during the conference: Jeff Allen, Audrey Mance, Valérie Mapelli, Valérie Raymond.

Sara Goggi and Sergio Rossi, of the Pisa Institute of Computational Linguistics of the National Research Council, have taken the full responsibility of creating and managing the database of submissions, and have been instrumental, through the timely and work-intensive updating of the information, for making it possible to keep the review process within the very short time allowed.

Antonio Zampolli
Chairman
Second International Conference on Language Resources and Evaluation
President of ELRA