Introduction of the General Chairman: Antonio ZAMPOLLI, President of ELRA

Introduction

The term "language resources" (LR), refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language and speech algorithms or systems. Examples of LR are written and spoken corpora, lexical databases, grammars, and terminologies, although the term may be extended to include basic software tools for their acquisition, preparation, collection, management, and use.

The development of robust and effective language processing systems depends crucially on the availability of various types of large scale LR. For example, in the field of speech processing, systems are built with technologies directly based on the use of corpora of speech data representing the domain of the intended applications. In written language processing, textual corpora are recognized as the primary source of data needed to inform the description, for computational systems, of the "real use" of languages in different communicative contexts. It is also self-evident that, in many real-world applications, NLP systems must be able to deal with tens and hundreds of thousands of lexical items.

Experience has shown that the creation of adequate large-scale LR is a costly enterprise, difficult for a single organization, however wealthy, to carry out alone. Duplication of efforts must, as far as possible, be avoided since financial and human resources are limited. In the past, for example, the usual practice has been for each project to construct its own ad hoc lexicon, geared to one specific application or to a specific piece of research: for each new application, and even for updating an existing application, the lexicon-building could well restart from scratch, even within the same company and research team. In this context, the *reusability of LR* has become a key concern. This expression appears more and more frequently in the definition of the objectives of national and international projects.

This terms adverts to two important aspects of LR. The first concerns the reuse of existing partial LR, usually designed for a specific application, as a 'help' in constructing new LR: for example, various machine-readable dictionaries (MRDs) have been investigated as being potentially rich and valuable sources of lexical information to help in the construction of computational lexicons.

The second issue concerns the construction of new large-scale multifunctional LR, i.e. of LR explicitly intended for multiple uses, which are capable of serving, through appropriate interfaces, a wide variety of present and future research and applications.

The choice of the term "Resources", coined rather recently[1], was intended to capture the idea that large collections of language data and descriptions play, for the development of effective NLP systems and their applications, an essential infrastructural role comparable to the role that basic resources such as highways, railways, electrical networks and energy play for the industrial and economical development of a country.

The Past

After 40 years of work in Computational Linguistics (CL), we still lack an adequate set of LR for the different languages, despite the fact that they are recognized to be "mission-critical".
This situation, in a certain sense a paradoxical one, can be explained by considering both theoretical and practical factors:

- The cost of building, maintaining and updating adequate LR, admittedly very high.
- The tendency, dominating in the 70's and 80's, to study an allegedly small amount of "critical" language data rather than the variety of linguistic phenomena occurring in the real use of the languages in different communicative contexts;

In the first half of the 80's, the mainstream Computational Linguistics (CL) community was characterized by a near complete lack of interest for the development of LR, and only few voices expressed any concern for them. However, around the middle of the 80's, a few people did pick up on the negative effects of this situation for the scientific progress of CL and the development of socially and economically relevant applications. At the same time they recognized the possibility that some current developments, if intelligently harnessed, could bring about a substantial change. These developments included the emergence of the paradigm of "Language Industries", then calling for real-life applications; the spread of every-day computer use in word-processing, which was making large amounts of language data available; and the renewed theoretical interest arising in different disciplinary fields for lexica and thesauri, which required widespread collection of structured lexical data.
To stimulate this convergence, in 1986, D. Walker, A. Zampolli, J. Sager and N. Calzolari organized a workshop in Grosseto (near Pisa) which brought together, for the first time, representatives of all those potentially interested in the study and use of lexica and corpora. The Grosseto workshop is usually recognized as the landmark event, the start of the process which has brought about today's substantial activity in reusable LR.

The Present

Even a cursory analysis of the present situation clearly shows that the LR field is evolving rapidly, both at the technical and organizational level.

---

[1] As far as we know, the term Language Resources (LR) was proposed for the first time in the Report of A. Zampolli, "Constitution of an European Language Technology Agency", written in 1991 as a contribution to the preparation of the so-called "Danzin Report" ("Vers une infrastructure linguistique européenne, 1992"), produced by a panel of experts, lead by A. Danzin, who were requested by the European Commission to design a general framework for the development of the European language industry and to identify priorities and strategies.
The Report was very influential, and from this point forward, the term LR entered the literature. It now regularly appears in the documents of the European Commission (Work Programmes, Calls for Proposals, etc.).

Current international initiatives aim at satisfying some of the major organizational needs, following the recommendations of the feasibility studies previously promoted by some of the major Funding Agencies.

Consensual de facto standards are emerging for several aspects of LR, and are widely applied in several national and international projects, already showing the benefits of their applications for the harmonization, and therefore, the inter-operability and usability of resources produced in different contexts (e.g., the EAGLES morphological tag-sets).

European Commission (hereinafter referred to as the "Commission"), after numerous efforts spent in promoting feasibility studies, now has started to support the creation of "real resources", ranging from the ones specialized for a specific domain or class of tasks (e.g. SPEECHDAT), to the initial nuclei of harmonized "generic" lexica and corpora for all the European Union Languages (e.g. PAROLE, EUROWORDNET, SIMPLE).

In several cases those initial nuclei are the starting points for nationally sponsored projects aimed at their extension (e.g. in Italy, Sweden, Denmark, Greece), or for complementary activities in other countries, so that the landscape of available LR will hopefully soon include other languages in addition to the traditional "resource- rich" ones (esp. English and French).

The need to preserve the value of LR produced in those projects through continuous extension and updating is recognized, motivating the setting up of networks which, by their very nature, should ensure continuity after the individual projects have come to an end: e.g. the PAROLE and TELRI associations link together, respectively in the European Union and in the Eastern European countries, organizations whose institutional mandate includes the provision of LR for their own languages.

The need to preserve, actively promote the use of , and effectively distribute LR, has caused the USA and EU authorities to put in place, respectively, LDC (the Linguistic Data Consortium) and ELRA (the European Language Resources Association).

These initiatives are already providing encouraging results, but the overall implementation schemata clearly demands regular updating to reflect the technical and strategical evolution of their environment. Strategic decisions are needed to ensure their continuity and extension, and to promote cross-fertilization and globalization through international cooperation.

The demands of various sectors of the Information Society for new types of applications and services, together with the first successes achieved in exploiting the data collected in the recent years, are stimulating various research efforts, directed at improving the usefulness and the reusability of existing LR and the design and construction of new types of LR: better linguistic knowledge provided by better LR allows for the research of

better methods and tools, which in turn, provide the acquisition of better knowledge, whose incorporation

improves the quality of LR.

We can consider a few examples:

The Information Society urgently requires products integrating language and speech technology: this calls for the immediate availability of LR designed to support this integration.

The need for robust components implies a facility to enrich lexica and grammars dynamically for use in a variety of broad-coverage applications. This requires the (semi)automatic capability to discover and acquire linguistic knowledge from corpus evidence. This long-term research challenge involves methodological and technical problems very similar to those encountered in such major LE applications as information extraction, summarization, text classification and information retrieval.

The success of methods based on empirical evidence, derived from corpus sources, brings up a new challenge: how to integrate the data-driven approach with the rule-based approach in a complementary manner[2].

A working hypothesis, which is currently gaining momentum, is that technologies, methods, tools, algorithms, which have been designed, applied and evaluated positively for a given language, could be usefully transported for use in another language, provided that adequate LR are available, functionally "equivalent" to the ones used for the first language. This hypothesis should still be considered a research issue, but positive evidence is provided, for example, by the proven transportability of methods and tools for speech applications systems based on statistical modeling techniques, such as Markov models and neural network, which typically learns by example from very large data sets organized in terms of many variables and many possible values.

A related, but for several aspects clearly different, research issue is to identify which features a given LR should possess in order to be transported to another specific domain or task, and which are the best methods to provide for easy and cost-effective customization, a central requirement for the LR producers to meet the demands of developers for specific LR and for developers who eventually need to adapt existing LR to their particular systems.

The methods and problems involved in acquiring and structuring linguistic knowledge are increasingly

---

[2] See, for example, in the Introduction, written by Y. Wilks, to the special issue, dedicated to Natural Language Processing, of Communications of the ACM (January 1996, 39, 1): „The field of natural language processing (NLP) has changed dramatically in recent years. Indeed, when we visited the topic five years ago, we concentrated on theoretical developments such as knowledge representation. Today, the combination of pressure from U.S. government funders - in particular, the specific goals of various ARPA programs - and the Zeitgeist itself have pushed NLP toward specific applications, systems evaluation, and above all, larger-scale language processing systems. Theoretical issues remain very important, but there is growing skepticism about the importance of small-scale, research systems and whether many of them are genuinely original, as opposed to being notational variants in a field not very aware of its own history. Rumor has it the word "hermeneutics" is now regularly heard in the corridors of Palo Alto research centers and that may be a sure sign of desperation among some of the more theoretically oriented."

connected with various aspects of "knowledge engineering", whose work can progressively contribute to the field of LR.

As exemplified in several papers presented to LREC or to the accompanying workshops, the research work required includes the development of new technologies, methods and tools[3].

Evaluation is a major feature of this landscape. Evaluation and LR are closely connected in many ways. Both play central roles of the infrastructure for natural language and speech processing: it is currently an issue whether both should be supported within the same organizational structure.

The relevance of both had, for many years, been sadly overlooked by the research community: this has changed with the emergence of a new paradigm, that of a language industry based on language engineering. This has attracted the attention of the major funding agencies to the very varied potential of language technology: strategic, social, economic and cultural. Hence component-robustness, system-coverage, market-acceleration, and user-reliability have become key issues.

All three types of evaluation which are usually distinguished (adequacy, performance, diagnostic) require, in different measures, the use of large naturally-occurring written or spoken corpora, annotated at different levels. Many surveys indicate the development of increasingly large and sophisticated annotated language corpora as one of the major contributions of evaluation to the development of speech and language processing technologies. In particular in America, a major side-effect of the comparative evaluation exercises promoted by ARPA, activities to which reference is made throughout LREC, has been to increase the support for the LR infrastructure. A clear success of the ARPA evaluation exercise has been in promoting the practice of sharing resources, methods, tools between different players, and the effective use of the services of a common infrastructure.

Currently, the Commission is considering, in preparation of the 5th FP, the possibility of introducing a comparative evaluation schema in Europe. The multilinguality of Europe clearly complicates the evaluation exercise, but poses interesting questions largely common to evaluation and to LR, as shown by the feasibility study underway (ELSE).

In addition to the organization aspects mentioned above, LR and evaluation share several research issues.

---

[3] Consider, for example, the following tasks:

- ensure the concrete reusability of LR with the provision of standardized access tools.
- acquire in a semi-automatic way additional and a "deeper" level of linguistic knowledge both to enrich generic resources, and to adapt them for domain- and task-specific use.
- make feasible the annotation of language data at a "deeper" linguistic level (e.g. semantically-tagged corpora, semantically-encoded lexica).
- collect and process new types of data required for more advanced tasks (e.g. different types of dialogue for different types of voice-based services)

link the different types of LR so that they mutually enrich each other (e.g. spoken and written corpora; corpora and lexica, etc.), offering in this way also an interpreted set of material for education tasks.

Consider the following examples:

- Localization: How a lexicon can be reused for evaluation actions with "a change of language", both a foreign language or a different type of the same language (technical language, etc.).
- Customization: What constraints are put on LR, in particular, the lexicon content, to ensure total – partial reuse for evaluating a different type of system. The customization methods and their economic viability, central issues both for LR producers and users, find in the evaluation framework a very effective and reliable test.
- Validation: The quality of lexica and corpora can strongly influence the performance of a given component or system, for its coverage, quality, linguistic content, etc. The validation of a LR both intrinsically and extrinsically (in relation to given tasks) is a central issue for the LR development and reuse.
- Date format and standards: Evaluation reinforces the adoption, the effective use, and the evolution of standards in several ways: common resources should conform to an agreed standard to be reused; interfaces between common resources and given specific systems must be implemented; vice versa, the results provided by a given system should be converted into the standard used for the test-suite; etc.
- Consistency and inter-operability of different types of LR: It is important, for example, to ensure and validate consistency between corpus data and lexicon data.
- Methods for annotating corpora: The need of designing methods and tools for annotating corpora is obviously common to LR and evaluation.
- Guidelines for distributed LR creation and use are needed to ensure consistency in the practical application of the adopted standards.

A major common feature is certainly the emphasis placed by both LR and evaluation on the use of data-driven methods. The ARPA evaluation exercise has clearly confirmed their usefulness.

The Future

Summarizing, the current landscape is very rich, complex and rapidly changing. Research and organizational activities often develop without synergy between them. The state of advancement can vary widely in different sectors and even in different countries. The risk of dispersion of efforts and delays in the results should lead to a more efficient organization and exchange of competences and information.

ELRA obviously offers a good observation point on the variety and complexity of the ongoing and planned initiatives and on the needs of the R & D communities still unsatisfied.

In '86, the perceived need was to promote the awareness of the necessity of adequate multifunctional LR as basic infrastructure for the R & D work.

Today this relevance has been widely recognized. Numerous actors are working in different sectors, on different aspects of LR, focusing on issues of particular relevance for their professional interests: linguists, computational linguists, language engineers, publishers, multimedia and culture operators, software and telecommunication industries, education and language technologists, knowledge engineers, service providers on the

telecommunication network, etc. Pertaining to different communities, which have their own specific organizations and conferences, they seldom have the opportunity of a common venue to exchange information and explore possible synergies and cooperation.

LREC aims to provide such a venue, promoting the awareness that all those working for LR 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 cooperation in the areas mentioned above".

The variety of Associations and Consortia who have joined ELRA in promoting the Conference is in itself a demonstration of the variety of the activities related to LR and of the perceived need for a common venue. The large number of countries and languages represented are an indication of how largely the awareness of the relevance of LR has spread starting from the Grosseto workshop. We hope that LREC will have, "mutatis mutandis", a comparable impact on the field.

I would like to stress in particular the significance of the participation of international and national authorities and funding agencies.

The concept of LR was introduced as the primary component of the infrastructure which is essential in supporting the development of Language Technology (LT) and its applications. The infrastructural role of LR has obvious policy implications at the national and international level.

Information Society (IS) and Technology are the driving forces for radical transformations in the organization 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 gaining full benefit from, our culture and heritage.

Only languages for which adequate LR products and systems have been developed will be available over the IS network. On the worst hypothesis, citizens who are not able to communicate in the languages implemented in the global network would be denied full participation in the IS. Authoritative sources have already warned that languages for which LT will not be adequately developed run the risk of losing their status as media of communication in the IS; because languages and cultures are inextricably linked, that will seriously threaten one of our most valuable human assets, linguistic and cultural diversity. To avoid this danger it is necessary to support multilinguality. Multilinguality has two obvious aspects – a citizen should be able to access the services of the IS in his or her own language; but should also be able to communicate and use information and services across language barriers.

The availability of adequate LR in a language is the key condition for the development in it of applications and

services that are informed by LT. LR have the function of providing the linguistic knowledge specific to a language, and the linguistic knowledge needed to ensure the multilingual links among languages. As stated before, in many cases, it will be possible to transfer methods, technology and in particular, software tools, from one language to another, provided that adequate LR exist for the second language.

LR are the most expensive part of any LT system. Today, for the major part of the languages, only embryonic nuclei of LR exists, which can not be effectively used in real systems without a substantial enlargement of their coverage. That absolutely requires that efforts are cumulated and not duplicated, reusability of LR ensured and enhanced, existing LR and specific technical knowledge exploited. The provision of LR, and, consequently, the development of the products and services required by the IS are feasible only if we are able to reach a substantial economy of scale.

The infrastructural role of LR, in addition, requires that LR are made available, in time, for as many languages as possible, in the public domain. All these considerations lead to the question of who has the responsibility to make LR available for a given language.

An easy answer would be that each State "owns" its national language, and should take responsibility for supporting the related infrastructure. A recent EUROMAP draft survey shows that the support of LT is extremely uneven across Europe at the national level. Several member states have no policy on the support of their national language within the IS, "a situation which threatens the survival of those languages in the mainstream". This problem is particularly acute for the provision of LR, which are language-specific[4].

Even if national authorities would take responsibility for the provision of the monolingual LR for their own languages, in this way countering the market forces which privilege the more widely-used and economically-important languages, the problem 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 globalization of the IS is already posing the problem of interaction with language communities outside Europe. LR 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 development can be supported. To this end predictive evaluation could be very useful. On the other, this will call for open and well-organized international cooperation in the field of LR.

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 LR and of the need to avoid duplications, ensuring reusability, concentrating and cumulating efforts; the

---

[4] In this situation, a substantial help can be provided from existing networks specifically dedicated to LR, as PAROLE. LR should also be actively promoted in general LT networks. For example, when I suggested and contributed to the design of ELSNET, I made LR a priority issue on its agenda.

progress towards the creation of an infrastructure dedicated to LR; the promotion of consensual standards; etc. We are only in the first phase of the process. As exemplified above, substantial research efforts are needed to respond to a number of incumbing scientific and technical challenges and problems, whose solution is central for natural language, speech, and evaluation. We should promote the awareness that LR are a clustered field, which includes not only "production" tasks, but also core research directed to providing general methods and tools which should be considered as core enabling technologies, and are a substantial part of the human LT.

The presently embryonic infrastructure should be reinforced, so that the same infrastructure is able to coordinate and perform, avoiding duplications, different complementary tasks: to provide and update the general repertories of linguistic data and knowledge which should be available for as many languages as possible; to produce at reasonable costs and in due time customized LR to answer specific requests of developers; to offer services the LE community urgently needs: information, consultation, validation, etc.

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

Acknowledgments

It is my duty and pleasure, in my capacity as Chairman of LREC, to thank all those who have contributed to the preparation of LREC and the publication of these Proceedings. LREC has been made possible from the generous support of the DGXIII of European Commission, the Fundacion Banco Central Hispano, the University of Granada, ELRA, the Real Academia de Ciencias Exactas, Físicas y Naturales, the Oficina de Ciencia y Tecnología (Presidencia del Gobierno), the Consorzio Pisa Ricerche, the Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche (CNR).

The members of the ELRA Management Board have endorsed from the very beginning our proposal of LREC and have been generous with practical advice and suggestions.

The members of the Scientific Committee have reviewed the submissions to LREC in the very short time allowed by the overall tight schedule constraints for the preparation of the Conference.

The Local Committee, Angel Martín Municio, Rosa Castro Prieto, Natividad Gallardo San Salvador, Antonio Rubio y Cristina Carrasco Fonseca, has spared no energy nor any effort in preparing the LREC, showing an admirable flexibility and creativity in adapting the organization to the number of participants, triple the amount of what was initially foreseen by the Program Committee.

The Program Committee has been at the heart of all the scientific, organizational and financial measures and

decisions. The members have followed practically day-by-day the preparation of LREC, dedicating an impressive quantity of time and effort to monitoring the progress and solving the problems.

Antonio Rubio has provided the facilities and organized the demonstrations, and Natividad Gallardo and Antonio Rubio, together with Rosa Castro Prieto, have taken care of the production of the Conference Proceedings as well as the nine accompanying Workshop Proceedings. They both have been marvelous in accepting to further reduce the time available in order to allow the deadline for text submissions to be extended.

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 organizers 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 honored LREC with their presence, the high strategic value of LR and Evaluation for the development of an user-friendly Information Society.

I would like to personally thank all the authors of the papers published in this volume: the number, variety, quality of the contributions, will provide one of the most complete overview and reliable definition of the field of LR and Evaluation.

This large participation would not have been possible without the precious cooperation of the Associations and Consortia who have joined ELRA in promoting LREC (ACH, ACL, ALLC, Cocosda, EAFT, Eagles, EDR, Elsnet, ESCA, Euralex, Francil, LDC, Parole, Telri, etc.), and of the major national and international organizations, including European Commission - DG XIII, ARPA, NSF, the IC/863 Project (China), the ICSP Permanent Committee (Korea) and the Japanese Project for International Coordination in Corpora, Assessment and Labeling.

In addition, I would like to thank all of the numerous companies and organizations who have given time and resources in support of LREC.

Let me now express my deep and sincere gratitude to some individuals 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.

Angel Martín Municio, with the prestige of his high office, for finding the necessary financial resources and political support to make LREC feasible, and for his experience and equilibrium, which has been instrumental in overcoming difficulties of every kind.

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

Harald Höge and Bente Maegaard, for contributing greatly to the scientific program of the conference, and to

Harald for being instrumental in keeping the financial situation under control.
Khalid Choukri, for inspiring everyone with his enthusiasm, energy and efficiency, and to the entire ELDA staff for their support.
Nicoletta Calzolari, for her dedication and invaluable contribution in designing the structure and finalizing the organization of the scientific program.
Simone Saint Laurent, who has dedicated her energies, covering diverse roles for LREC. Her cooperation with Rosa Castro Prieto and Cristina Carrasco Fonseca, of the Local Committee, has been a key factor in the preparation of LREC.
Finally, to Cristina Carrasco Fonseca and Sergio Rossi, for their generous contribution to creating and organizing the databases for LREC.

Antonio Zampolli

Chair

First International Conference on Language Resources and Evaluation