Message from ELRA Secretary General and ELDA Managing Director Khalid Choukri

Welcome to this LREC 2014, the 9th edition of one of the major events in language sciences and technologies and the most visible service of ELRA to the community.

ELRA, the **European Language Resource Association**, is very proud to organize LREC 2014 under the auspices of **UNESCO** (the United Nations Educational, Scientific and Cultural Organization), through the patronage of Her Excellency Madame Irina Bokova, UNESCO’s Director General, and of Madame Vigdís Finnbogadóttir, former President of the Republic of Iceland and UNESCO Goodwill Ambassador for Languages.

I would like to express my heartfelt thanks to Her Excellencies Madame Irina Bokova and Madame Vigdís Finnbogadóttir for their patronage and support, assuring them of the community continuous efforts to address the common concerns and the crucial challenges, we all share.

It is an important symbol and a path for ELRA that strongly advocates for the preservation of languages, all languages, as major components of our cultures and efficient instruments for boosting education, literacy, and reducing the digital divide.

Welcome to Reykjavik, where you will certainly experience a true Mediterranean atmosphere, associated now with LREC, in the very North, standing in the middle between Europe and America. After having organized LREC in areas that identify themselves with largely spoken language families (Roman, Semitic, Turkic languages), we are heading to a country where the language played a special role in particular through the medieval Icelanders' sagas but also “preserving” itself over centuries as well as preserving the Old Norse spoken by the Vikings.

Organizing LREC under the patronage of UNESCO is an important symbol for ELRA that strives to stimulate the emergence of language technologies so they contribute to better education and easy access to our common knowledge, in all languages. Since its foundation, ELRA has been an active contributor, in particular shedding light on under-resourced languages. The first LREC, in 1998, already featured a workshop on “Minority Languages of Europe”, a tradition that continues to date, going beyond the initial geographical and geopolitical coverages. Furthermore, several LRECs have seen the organization of specialized workshops and panels dedicated to educational applications. It is our credo to strive and encourage young generations to learn foreign languages, as many as one can handle. Learning foreign languages, including sign languages, is an extraordinary journey in other humans’ cultures and traditions.

HLT (Human Language Technologies) should also support such endeavor and help underprivileged communities access the tremendous and wonderful human being world heritage, in particular UNESCO referenced ones. We hope that our community, through the backing of automated translation and other multilingual tools, improves such accessibility. Contributing to the efficiency of our translation and localization experts should support the cross-cultural fertility we all promote.

Dear ELRA Members, Dear LREC participants,
It is a great and renewed pleasure to address the LREC audience for the ninth time and share with you these thoughts and remarks. On the 28th of May 2014, we will also be remembering the first LREC that took place exactly 16 years ago, on May 28th 1998, and the visionaries who felt the need for such forum.
This 9th LREC is a special milestone as it gives the opportunity to celebrate LREC's 15th anniversary and ELRA's majority after 18th year of dedicated activities and services. It is an opportunity to review the activities carried out so far, draw some conclusions, and plans for the years to come. Some of these topics have been discussed at a workshop held on 19-20 November 2013 in Paris, and attended by representatives of the most distinguished organizations active in our field.

Allow me to take you 18 years back and walk together remembering the landscape as it was, at least on the European scene, from the Language Resources and Language Technology perspectives.

Just remember that the web was only in its infancy in 1995, when ELRA was established. The first reviews and surveys of existing resources in Europe were conducted in a set of projects funded by the European Commission. The field was split over three major domains, represented by clearly three different communities, associated to three big Language Resources categories: speech processing (spoken data), written text analysis (textual corpora and general lexica), and terminological resources (specialized dictionaries). The challenge for ELRA was to try and establish bridges between these different communities (hence LREC) but also capitalize on the findings of these projects to consolidate a catalogue of Language Resources (as stated in ELRA's foundation mission).

ELRA came out with its first catalogue of resources in 1996, a simple plain list comprising 30 resources. We were proud to publish such a catalogue (hardcopy) but we realized, with great humility, the huge task in front of us, we immediately understood that listing such resources could not serve the purpose for which ELRA has been set up: to ensure that LRs are used and re-used, possibly repackaged and repurposed. Users still had to negotiate themselves with the right holders, often located in other countries and different legal systems.

Such a mission required understanding the rationales behind data production and inventing new economic models, different from the ones in use including by the other data centers. A major dimension to be understood and managed was the legal issues behind ownerships, copyright and other associated rights. We had to address such issues and clear all legal aspects so that a user could access the LRs through an easy licensing schema. The mission of ELRA shifted from an archiving house of EU-funded project outcomes to a true distribution agency. For the next decade, we consolidated our identification activity and ensured that a large number of resources were catalogued and made available by ELRA to the community at large under fair conditions and easy licensing. ELRA acted as the EU instrument in distributing all LRs that were co-funded by the EC within its R&D frameworks.
We had the feeling that we were moving from scarcity to an organized framework that would help the community access an abundance of LRs.

Acting truly for multilingualism, we had to get accustomed to negotiating and clearing rights in multiple legal systems. The role of ELRA became even more crucial when users realized they could sign a single agreement to license multiple resources provided by a large number of suppliers, from all over the world.

We were (and still are) under no illusion about how good our coverage was. Through our market analysis and surveys, we knew that less than 20% of existing resources were publicly traded, the 80% were not released and not exchanged even when the right holders were public entities funded by tax payers (a few percentages were privately sub-licensed).

On the other hand, the surveys and inquiries received by our helpdesk (that is still in operation) clearly indicated that many needs were not fulfilled at all despite our supply.

To help people disclose what they had in their archives but also get tribute and scientific recognition for the work done to produce LRs and conducted evaluations, ELRA initiated, in1998, this conference:

the Language Resources and Evaluation Conference (LREC), a forum that aimed at bringing together all interested parties. With over 1200 attendees for the last editions (including over 30% of student and young researchers), LREC became one of the major events in our field. LREC focusses on all issues related to LRs and Evaluation of HLTs. It also gives room to specialized events that run as satellite workshops/tutorials to the conference.

ELRA viewed LRECs as important channels to discover existing Language Resources on which the community works but also to help identify gaps and trends. As such, LREC helps consolidate the community while drawing a clear picture of the state of affairs. The paper about "Rediscovering 15 Years of Discoveries in Language Resources and Evaluation" by Joseph Mariani (ELRA former president and current Honorary President) et. al. reviews some of these findings through an analysis of the papers published in the LREC proceedings over the 15 past years.

ELRA also designed LREC to become one of the best places to meet friends and colleagues, to share ideas and visions, and to plan for new collaborations, proposals and projects. As such LREC also contributed to the community building, an essential part of ELRA's mission. As a supplementary contribution, ELRA endorsed the publication of the **Language Resources and Evaluation Journal**[1] by Springer, on the very same topic.

Inspired from the discussions that took place at LREC, ELRA launched its project called "Universal Catalogue" (UC), with the aim to make it an inventory of all existing LRs within our field, either identified by the ELRA team or through input by the community. The UC comprises LR descriptions, independently of whether such resources would be made available or not. The underlying idea was and is to prioritize ELRA's negotiations, taking into account the requests of our members but also help potential users discover existing material before starting heavy production processes and hopefully negotiate directly with the right holders.

While maintaining our efforts devoted to the Universal Catalogue, ELRA took advantage of LREC to establish the LRE Map (Language Resources and Evaluation Map, http://www.resourcebook.eu/): a resource book that associates scientific publications to descriptions of LR and/or tools. LRE Map, an integrated component of the LREC submission system, requires from all LREC contributors to fill in a simple description of the LRs or the Language Tools (LT) mentioned in their submissions. By doing so, ELRA initiated a community-based bottom-up process that helps describe Language Resources (over 4000 unique LRs so far), consolidating the area of language resources. We are very grateful to the other conferences that adopted the LRE Map to collect more data on the existing Language Resources. Over time such "live" inventory of resources and tools, associated with scientific publications, will constitute a very useful knowledge base for the benefit of the community.

A critical issue that we learnt from the cataloguing and distribution activities is the difficulty to associate a unique name with a given LR. We realized that, despite our efforts and those of other data centers, referencing the LR used and/or described in scientific publications is very fuzzy and we see a large variety of names used for the same resources, even by the same author. This inconsistency could not be prevented even by data centers that could and did enforce the use of their identifiers, as part of the licensing agreement (i.e. ELRA)!

It is one of my deepest regrets that the community missed out a great opportunity to set up its own persistent identification system to name the LRs we are handling. The major instrument could have been the DOI system if we did come to a consensus to have one DOI assigner. It is probably too late as many centers and LR owners became DOI assigners and each can assign a different DOI to a LR.

To overcome such issue, the major organizations behind distribution and sharing of Language Resources, decided to introduce an identifier that is independent from Internet (and hence from DOIs), independent from the right-owners as well as from distribution agencies. This was inspired from the publishing community that adopted the ISBN schema, almost half a century ago. Such identifier, referred to as ISLRN, International Standard Language Resource Number (www.islrn.org), will allow

[1] LRE Journal, http://link.springer.com/journal/10579

a unique identification of a resource, independently from where it is stored, whether it is available or not, which licenses it is associated with, etc. ELRA, LDC[2], and AFNLP[3]/O-COCOSDA[4], committed to establish, run and moderate the ISLRN server at no charge for the community. The initiative will be steered by an international committee consisting of representatives of the major players from the NLP12 group. ELRA, LDC, and AFNLP/O-COCOSDA, in partnership with the major organizations within the field, would like to ease the citation of Language Resources and hence better assess the impact factor of each resource (the NLP12 Paris declaration is available at: http://www.elra.info/NLP12-Paris-Declaration.html).

It is clear from the setting up of ISLRN that it does not prevent data centers and resource right holders from using whatever local identifiers including DOI to refer to their resources but it will be more efficient if such identifiers are used in addition to ISLRN. The ELRA Board is discussing how to enforce such an identifier, making it compulsory for all publications at LREC and LRE Journal.

As mentioned above, ELRA celebrated its 18th anniversary on November 18-19-20, 2013, through a workshop and the NLP12 meeting (http://www.elra.info/ELRA-18th-Anniversary.html). The meeting was an excellent opportunity to gather several influential representatives of the community and discuss several pending hot topics that require more coordination and harmonization. In addition to the identification of LRs, including the endorsement of ISLRN proposal, the participants felt a strong need to harmonize the organization of their conferences and later on with those of neighboring domains. We have seen recently many important events running into each other with conflicting plans such as very close deadline dates for Call For Papers, similar dates for submission of abstracts or final manuscripts, similar milestones for the review process, etc. Given that most of the work is freely carried out by the peers (review scheduling and conduct, paper selection, program design, proceeding preparation, etc.), a conflicting planning demanded more efforts to those who had to juggle with more than one event, if they had to submit a final paper, an abstract on some new research, while reviewing other authors' papers, while continuing their usual work!

To avoid this situation, the NLP12 representatives agreed to develop an internal tool that would help the organizers view their plans while visually reviewing other events' planning, and getting some warnings and alarms. It is clear that, given the number of annual events, such conflicts are impossible to resolve, but at least some of the negative effects could be better handled.
Again, this should help better consolidate the activities of the community, improve synergies, and save some efforts.

New initiatives, European Commission debates on Licensing and Copyright

Regarding the licensing activities, ELRA took part to a large stakeholder dialogue in 2013/2014 organized by the EC about "Licences for Europe". ELRA contributed to the activities of a Working Group on "Text and Data Mining". The WG participants represented most of the parties involved in Data/Text mining both from the supply side (providers of data such as publishers, broadcasters, collective management of copyright and related rights organizations, etc.) as well as the demand side (Librarians, archivists, research centers, technology developers, etc.). ELRA, as a representative of the Human Language Technology developers, both from research and industry, brought in its knowledge of the community concerns and expectations. ELRA highlighted the importance of accessing substantial amounts of data to develop and assess performances of new NLP technologies that are the basis of most of today's search and mining applications. More details: http://ec.europa.eu/licences-for-europe-dialogue/en/content/about-site.

In addition to expressing the requirements and expectations of our community, emphasizing the new trends for free and open resources, ELRA advocated for an intermediate solution based on simplifying the access to copyrighted material for research purposes. ELRA argued that the solution for a competitive Europe requires a revision of the copyright regulations, to adopt a clear rule on the fair use

[2] Linguistic Data Consortium (LDC), https://www.ldc.upenn.edu/
[3] AFNLP : Asian Federation of Natural Language Processing, http://www.afnlp.org/
[4] O-Cocosda, see the 2014 meeting announcement at http://saki.siit.tu.ac.th/ococosda2014/

for research purposes of copyrighted language resources. Further to these WG meetings, the EU invited organizations to express their views on the necessary copyright amendments, which ELRA did along these lines. We are looking forward to hearing of the next steps. The contributions are listed at: http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_en.htm.

Despite all these consolidation actions, we have also seen a fragmentation of our field. The last few years have seen an extraordinary development of the web (and more globally of the Internet). The culture of open source and free resources shifted from a fashion phenomenon to a strong and a lasting social and economic best practice. Such expansion has encouraged many institutions to establish their own repositories and offer their resources via internal infrastructures.

This trend definitely increases the availability of LRs (particular with the adoption of free/open sources spirit and licenses like Creative Commons) but renders their discoverability more tedious and their identification more complicated. In Europe, it has become affordable, from all points of view, to set up a LR repository (see details at the ELRA helpdesk at this conference) even if many institutions still rely on staff's personal pages to host resources and disseminate the corresponding information. With almost 30 different and independent entry points, META-SHARE is certainly the most sophisticated example of a distributed and networked repository set, with repositories listing as few as 5 resources and others i.e. ELRA with over a thousand. It is still a challenge to bring down the number of different applicable licenses (over 30 now) to the dozen prescribed by META-SHARE and inspired by ELRA and the Creative Commons spirits. Such a network should prevent profusion of unlinked/unrelated repositories.

Such "paradigm" shift boosted the sharing of language resources and tools while impacting the distribution mechanisms. To keep a proactive role with respect to its mission, ELRA has anticipated some of these changes and new tasks (e-commerce meta-share repository, ISLRN assigned to all its resources, e-licensing and e-signature, a LR forum, etc.) are in an advanced stage and announcements of these novelties under preparation.

To support this consolidation requirement and vital need, ELRA is involved in a new EU funded project called MLi (European Multilingual data & services Infrastructure). As a EU support action, MLi is working to deliver the strategic vision and operational specifications needed for building a comprehensive European Multilingual data & services Infrastructure, along with a multiannual plan for its development and deployment, and foster multi-stakeholders alliances ensuring its long term sustainability. We hope to share these visions with the LREC participants on the ELRA and MLi booth at the HLT Project Village that features exhibition booths for many EU projects, at this conference.

Acknowledgments

Finally, I would like to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful.

I would like to thank our Silver Sponsor Holmes Semantic Solutions, and our Bronze sponsors: EML (European Media Laboratory GmbH), IMMI, VoiceBox and K-dictionaries. I also would like to thank the HLT Village participants, we hope that such gathering will offer the projects an opportunity to foster their dissemination and hopefully discuss exploitation plans with the attendees.
I would like to thank the impressive local advisory committee. Its composition of the most distinguished personalities of Iceland denotes the importance of language and language technologies for the country. We do hope that it is a strong sign for the long term commitment of the Icelandic officials.

I would like to thank the LREC Local Committee, chaired by Professor Eiríkur Rögnvaldsson who helped us with all logistic issues, herein in Iceland and Gudrun Magnusdottir, who introduced us to Iceland and for her continuous support.

Finally I would like to warmly thank the joint team of the two institutions that devote so much effort over months and often behind curtains to make this one week memorable: ILC-CNR in Pisa and my own team, ELDA, in Paris. These are the two LREC coordinators and pillars: Sara Gogi and Hélène Mazo and the team: Victoria Arranz, Paola Baroni, Roberto Bartolini, Irene De Felice, Riccardo Del Gratta, Francesca Frontini, Ioanna Giannopoulou, Johann Gorlier, Olivier Hamon, Jérémy Leixa, Valerie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, and Priscille Schneller.

We were very happy, for this LREC, to enjoy the friendly support and efficient help of Sigrún Helgadóttir – Researcher at the Árni Magnússon Institute for Icelandic Studies, to whom I extend my warm thanks.

Now LREC 2014 is yours; we hope that each of you will achieve valuable results and accomplishments. We, ELRA and ILC-CNR staff, are at your disposal to help you get the best out of it.
Once again, welcome to Reykjavik, welcome to LREC 2014

Khalid Choukri
ELRA Secretary General and ELDA Managing Director