Introductory message of Khalid Choukri,
ELRA Secretary General
ELDA Chief Executive Officer

ELRA and ELDA are very pleased to welcome you in Miyazaki to this 11th LREC to celebrate the 20th anniversary of LREC with all of you this week.

On behalf of the ELRA/ELDA team I would like to share with you some news on the activities we conducted since the last LREC in **Portorož (Slovenia)**.

The Declaration of Granada

But first let me to share some feelings about this special LREC with you, as we are celebrating the 20th anniversary of this major forum established in 1998 in Granada (Spain), organized for its 11th edition, here in Japan.

Soon after the establishment of ELRA in 1995, its Board realised that, at that time, the language resources and the evaluation of language technologies were given very little attention at the main events. Today, we are glad that such message is spread widely and is endorsed by the major conferences in which special sessions are expressly devoted to Language Resources and Evaluation!!

Remember the first LREC, remember Granada, not only the Alhambra! With over 400 participants instead of the expected 100 attendees, we realized the importance of such forum for the community. This was confirmed over the years by a steady attendance of 1200 participants to the last editions of LREC.

I would like to take this opportunity to go back to the spirit of Granada, paying a tribute to those who were behind it, Professors Antonio Zampolli and Angel Martin Municio. I would like to bring up one of the major outcomes of that first event: "the declaration of Granada". Its recommendations are still relevant and topical, more urgent than ever to implement.

The declaration of Granada[1] comprised 10 articles. I am highlighting and commenting here some of the crucial ones that we can continue to endorse today:

- "**At this moment, language resources are one indispensable key to unlock the potential of the global information Society**"

We are still facing this issue 20 years later and if we agree that the Information Society has made tremendous progress with the emergence of social networks which have strengthened links within and between communities, social or commercial activities cross borders are still hindered by language barriers. In 2015, surveys mentioned that 24 languages are used in LinkedIn user interfaces, 48 on Twitter, 91 on Google Translate (as pairs for its translation of content and now about 103), over 150 on Facebook, just over 300 in Wikipedia. These numbers may seem impressive, but remember that this is **out of 7097 living languages or 3,909 with writing systems.** And most of these languages are used in interfaces with automatic processing of content used in Search and/or MT only. Language Resources are essential assets. Back in Granada, we stated that

[1] Granada Declaration: http://www.elra.info/media/filer_public/2013/09/06/v3n3.pdf)

"They constitute an essential infrastructure". Such infrastructure is missing for a huge number of languages. The LRE Map service provided by ELRA, inventorying the LRs reported in major conferences, continue to expose the existing gaps.

- **"All sectors of society, and all languages, have an interest in seeing these resources developed, for a variety of purposes, economic, social, industrial and cultural."**

ELRA continues to promote the concept of Basic Language Resource Kit, a Kit that would help process every language for (at least) the basic NLP functions. We stressed the importance of this approach to policy makers, emphasised the need to support small communities, and mentioned the lack of interest from private sector for non-lucrative/non-strategic languages. We also insisted that such "*core language resources should remain in the public domain*" to ensure a wide use by both research and development stakeholders. Reviewing the current situation at major data centers and repositories, we can barely count more than 100 different languages, often with scarce resources (many speech resources for the major languages, very few treebanks, very few aligned corpora, mostly aligned with English, etc.)

- **"For each language, there is a need for strategy to co-ordinate existing resources and create new ones."**

ELRA, along with LDC, their partner in the USA, did their best to offer distribution/sharing channels for Language Resources produced within publicly funded projects and some offered by private bodies. However the identified resources represent less than 15% of what exists. Coordination of the distribution but also documentation and production, have proved to be challenging. We still feel it is crucial to coordinate building roadmaps for every language and enhance the involvement of local public and private bodies. It is also essential to continue international cooperation to disseminate the know-how acquired for a given language. We are glad that a conference like LREC contributes to sharing such expertise and value the implication of governmental (regional and national) and international bodies.

We introduced the International Standard Language Resource Number (now part of the activities of the International Standardisation Organisation, ISO TC37/SC4) to assign a unique identifier with each identified Language Resource to improve the way we reference it (this is also part of the LREC submission process that distinguishes Bibliographical data from LR data). The idea is not only to provide an ID, unique and persistent, wherever the LR is stored, even for those LRs on local servers outside the Internet. This is an uphill struggle but we are convinced that it is an important step in our work to improve the identification of existing resources, the assessment of LR impact factor as well as the citation mechanism.

- **"When resources have been created, there is a continuing requirement for support and maintenance."**

This is a key part of our mission and we tried to convince data producers and funders to account for the necessary maintenance of and support for Language Resources. We introduced the validation process and the "bug" reporting mechanism, as part of ELRA procedures, to encourage sharing experiences on the use of LRs and their enhancement over time. We still face funding scenarios that provide subsidies for data production and not for other issues like IPR clearance, documentation, sharing, maintaining, etc. In Granada, we anticipated that resources would undergo some repurposing with the new uses that emerge and we insisted on the need to envisage a wide range of applications on the basis of the same resources. The community seems to be sensitive to this, but some legislators are debating the adoption of more legal constraints. We need to join forces to convince funders and decision makers about the importance of more openness and long term policies. The introduction of the Data Management Plan (DMP) by ELRA, and soon the DMP Wizard, will help each data manager to adopt up-to-date standards and best practices for data management.

- **“Understanding of the role, usefulness and optimum means of preparation for language resources is a research theme in itself.”**

Over the last decades, and especially within the last 3-4 years, we have seen an impressive breakthrough in the HLT field. The new data-intensive machine learning and the computing capabilities, are proving the crucial usefulness of LRs. Making LRs widely available is the core mission of a few organisations. ELRA is very happy to be among these organizations and is making the necessary investments to acquire more expertise to cost-effectively produce and share LRs. The setup of an internal legal team is helping to shed light on a large number of legal issues that impede the use/re-use of LRs. Working on standards is also an important aspect to help facilitate the interoperability and sharing of data. One of our mottos was that “Common evaluation requires common standards”. We still feel that common tasks in the “challenges” and evaluation campaigns are essential instruments to assess progress, share knowledge, and improve cooperation. It is a pity that many “Evaluation campaigns” are happening with very little coordination which makes them hard to find for new comers.

> ***Granada was 20 years ago and we see that some visionary recommendations are still needed today. A multilateral, concrete, and lasting cooperation remains on top of our action.***

ELRA activities since 2016

Now allow me to get back to ELRA activities carried out over the last couple of years.

We continue our actions on data sharing, through the identification, negotiation, and distribution agreements with right holders when necessary. We continue to produce resources for projects as well as for partners. Our policy remains consistent: whenever the data is offered to the community, after the shortest possible embargo period, the costs for partners are set to production costs. This position remains fundamental to our policy. We continue to invest in research and development of tools to improve and automate our production procedures. Most of our tools are shared as open source packages.

We continue also to work on our quality control methodologies so as to supply validated resources with validation procedures that guarantee the adequacy of the produced datasets with respect to the initial specifications and the state of the art.

To ensure an efficient distribution of Language Resources, ELRA has migrated its catalogue of resources to a new platform, based on e-Commerce features, redesigned with a new interface and an improved navigation. This foreshadows further developments that will incorporate e-licensing, e-payment and e-delivery of resources.

ELRA continues to support the set-up of LR repositories for data deposit by third parties. Based on its involvement in the jointly-developed META-SHARE platform, we continue the promotion of such efforts to ensure that the major data holders adhere to some common practices. A new repository was set up as part of an EU service contract to store data for MT provided by the public sector. Such initiative is now spreading across Europe, and a coordination action is establishing local repositories (known as Local Relay Stations). If we succeed to set up such stations for each country in order to collect all language datasets produced by translations services and secure these for MT training and tuning, one can anticipate good progress for these languages and domains. The repositories can accommodate any Language Resource modality.

If the establishment of such a local repository is of interest to your organization and your network, let us discuss how to work on it together.

As part of this process, we continue to work on all issues related to sustainability and preservation of data for the generations to come.

An updated ELRA Data Management Plan is made available and reviews all necessary aspects for an optimal management of resources with an easy-to-use checklist. We are working to automate the customisation of such DMP for each project. Our members will benefit from this automatic DMP Wizard, accompanied with the support of our experts, free of charge. We hope that such approach will improve sustainability and preservation of Language Resources but also make them easy to identify.

ELRA continues to be involved in the new trends in HLTs. It continues to support the new trends in MT. Many of our projects (some of which are funded under a European Program known as Connecting Europe Facility (CEF) focus on data production, including via requests for donations from translation services, but also crawling of adequate data to which we have access and re-use rights. Many resources come from organizations that belong to the Public Sector. A directive (called Public Sector Information directive, PSI) entered into application in the European Union, similar rules exist in many other countries, stating that publicly produced data should be made publicly available. This makes some of the resources needed by our community (e.g. textual corpora) available for new domains and new genres. Some geographical areas offer a multilingual environment (EU, India? South Africa, etc.), and hence more resources should be available for MT development.

Unfortunately there are still important legal restrictions on the re-use of data, even for research purposes. We continue to vilify the current legal framework, in particular in Europe, e.g. the European Union is working on a new directive on copyright in the Digital Single Market. The initial proposal for this act contained a mandatory exception for text and data mining carried out by research institutions. However, the current debates within the European decision makers seem to suggest that the exception will fall short of meeting the objective of the exception. The beneficiaries of the new exception may be limited to public research institutions, and – more importantly – 'lawful access' will be a prerequisite for data mining, which will probably result in wider implementation of digital protection measures by right holders. It is unlikely to get the exception for research that we claim since years now as a fair use doctrine for research purposes (that remains the privilege of a few countries).

The current legal framework has a strong impact on the capacity of the community to produce IPR cleared and sharable data. ELRA heavily invested in legal training and has been, for many years now, one of the few organizations that works both with in-house legal experts and a network of external practitioners/lawyers.

Another critical novelty in Europe is the new legal framework governing the processing of personal data. It goes beyond the users expectations, for more ethical behaviour on the management of their data. This may hinder the new developments of resources and technologies (e.g. Crowdsourcing activities). The new regulation (General Data Protection Regulation (GDPR)) will impose more restrictions on managing several aspects of data e.g. data protection by design and by default, privacy impact assessment, pseudonymisation and anonymization, before the data can be shared (this will of course impact also production, repackaging, repurposing of data).

To share information on these matters, a dedicated workshop on legal and ethical issues continues to be organized within LREC and will be held this week as well.

Of course, ELRA does not focus on EU issues and EU languages only (we distribute resources for more than 70 languages). In 2017, ELRA entered into an important agreement with the International Speech Communication Association (ISCA[2]) to join forces in the promotion of activities related to the Less-Resourced Languages (LRL). ELRA and ISCA agreed to merge their groups and set up a join Special Interest Group for Under-resourced Languages (SIGUL[3]). Co-chaired by a representative of ELRA and a representative of ISCA, SIGUL will continue to organize events for the LRL and encourage cooperation actions to support these languages.

As you may know, United Nation General Assembly proclaimed 2019 as the International Year of indigenous Languages. UNESCO is leading the corresponding events. ELRA proposed to organize an important international event related to HLT and Indigenous languages. We hope to draw attention to the importance of HLT and LRs for the preservation and development of local cultures and put under spotlights the role our community could play for these languages.

We continue to develop the LRE Map application. LRE Map was established to reference all LRs described by authors when submitting papers to conferences and journals. Started with LREC, it is used by other events but not as widely as we hope. In addition to identifying over 7000 instances of LRs, it helps identify existing gaps for languages lacking such modalities and ensure a minimal cooperation when planning new productions. If you are involved in the organisation of a conference, let us see how we can work together.

ELRA is also taking part in several standardisation activities. It is naturally involved in ISO/TC37/SC 4 on Language Resource management but also on ISO/IEC JTC1/SC35 about user interfaces and accessibility. ELRA brings its knowledge of the HLT field to ensure that all ICT services and products are accessible to all, in particular to users with specific needs. Some of the HLT applications are offering valuable services when converting speech into text, text into speech, sub-titling/captioning audio-visual streams, providing audio descriptions, translations (e.g. subtitles), easy-reading features (both in mono- and multilingual contexts). Such services are valuable to everyone and not only hearing or visually impaired users. Translation from text or speech to Sign languages is a big challenge that many partners are working on and ELRA will support them.

As a conclusion to my message, I would like to reiterate my statement uttered at almost all LRECs since 1998. Please remember that we can help you share your data for all types of use. We can work out a contractual framework that suits your expectations, including adopting very permissive licences and a free-of-charge policy. We can guarantee the availability as well as the sustainability of your resources. During the conference, an ELRA booth is available where we will be happy to interact with you on such topics.

About 10 years ago, we identified about 20 resources, some were on the web, others well known to the community. We keep monitoring their availability. Believe it or not, about 30% disappeared and these are not necessarily the ones that were obsolete and useless. Some right holders also disappeared and the “orphan” resources with them.

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 Sponsors: Google, Amazon, Arcadia, EML (European Media Laboratory GmbH), GSK Language Resources, Riken-AIP, Yahoo Research Japan, and the publisher *Hituzi Syobo our media sponsor: MultiLingual Computing, Inc.*

I also would like to thank the HLT Village participants, we hope that such gathering offers the projects an opportunity to foster their dissemination and hopefully to discuss exploitation plans with the participants.

[2] https://www.isca-speech.org/iscaweb/index.php/about-isca

[3] http://www.elra.info/en/sig/sigul/

I would like to thank the Local Advisory Committee. Its composition of the most distinguished personalities of Japan denotes the importance of language and language technologies for the country.

I would like to thank the LREC Local Committee, chaired by Prof. Hitoshi Isahara and the LREC Local Organizing Committee, for providing support to the organization of this LREC Edition in Japan.

Finally I would like to warmly thank the joint team of the two institutions that devoted 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 Goggi and Hélène Mazo, and the team: Roberto Bartolini, Damien Bihel, Irene De Felice, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Vladimir Popescu, , Caroline Rannaud, and Alexandre Sicard.

Now LREC 2018 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 Miyazaki and Japan, welcome to LREC 2018