LREC 2022
13th International Conference on Language Resources and Evaluation

Opening Address

Khalid Choukri

ELRA Secretary General and EDA Chief Executive Officer

Dear guests, dear friends, Dear ELRA members, Dear LREC Participants,
Je suis très heureux que nous ayons repris la vie que nous chérissons tous en nous retrouvons ici à Marseille tous ensemble pour LREC 2022, avec l'espoir que la pandémie ne vous a pas trop impacté.

It is my great pleasure to welcome you all, four years after our last face to face LREC conference in Miyazaki, Japan, in 2018.

Welcome to all those who joined us today here in Marseille but also warm greetings to those who could not join us in person but are participating remotely.

It has been a challenge to organize an LREC as a hybrid conference, LREC which spirit is to meet friends and colleagues and revitalize our networks. This new framework (in person and remote participation) requires strict organizational processes and logistics and we hope that we will succeed to make the best out of it.

Welcome to this 13th edition of LREC.

ELRA & ELDA missions

ELRA has gone trough an important restructuring to better account for the community expectations in terms of Language Resources and Evaluation of technologies. Some of this is elaborated upon by our president in his message. The major changes are related to the association membership and its governance. The technical, legal, logistics and other practical tasks will continue to be our core business while improving our connections to our members.

As most of you know, ELRA missioned ELDA to carry out its operational tasks and to put in operation its strategy. The General Assembly decides on the annual actions to conduct, and regular strategic meetings design the long-term perspectives and the associated roadmaps.

I have been proud to coordinate such activities with an excellent multidisciplinary team. Let me highlight some of the practical tasks that we managed sustainably for over a quarter of a century.

LREC to promote recognition of LRs & Eval research

I am sure that many of you still remember the first time, 1998, in Granada after that ELRA had decided to establish a conference, fully dedicated to Language Resources and Evaluation of Human Language Technologies. ELRA focused on topics the were and still are fundamentally relevant to ensure that researchers can devote the necessary efforts to produce and package or re-purpose Language Resources and get the deserved peers recognition. We all understand that such recognition is essential for the curricula, through the citations and conference indexing. From the start, the idea was to oversee these topics in the largest inclusive context possible. It was of paramount importance to promote the field beyond the main streams of the moment. The main streams were mostly a strong focus on languages such as English, Mandarin, Arabic and very few European languages and with very few technical approaches.

LREC genesis came with all languages and modalities in its DNA and hence promoted works on less-resourced languages and in particular the indigenous ones spoken by small communities. They also need to benefit from our achievements. Since that first LREC and the promotion of language resources for all languages, we can see the smart increase in the number of languages that are tackled by the research and industry communities. LREC shows this through the number of languages cited in the LREMap (`https://lremap.elra.info/`) . LREMap, as initiated by ELRA at LREC2010, help report on the language resources used/produced/used by the LREC authors. The initiative has been taken by other major conferences today. With about 8000 items and more than 100 different languages today, LREMap emphasizes the progress made so far. Nevertheless, we are still far away from the 7000 human languages out of which at least 30% would positively benefit from our field work.

The other important dimension of LREC (and LRE, Language Resources and Evaluation Journal, the associated ELRA Journal by Springer) is the promotion of evaluation of technologies and reproducibility of experiments. The number of evaluation campaigns has grown exponentially and, while this is necessary for the assessment of existing approaches, it became impossible to keep track and monitor this through an accessible inventory. It became so easy to initiate a challenge but very difficult to design this in a long-term roadmap with strong incentives that federate a large community able to conduct reproducible research. IT is very common to see a challenge established for our round without too much scientific conclusions for the future. Some initiatives started focusing on less resourced languages. This should be highly encouraged and may be made mandatory so each time a benchmarking is initiated, at least one language that is not part of the few large ones is considered if it makes sense.

Of course, the budget may not always be sufficient but many funding agencies would be ready to support such action. ELRA joined forces with other partners and is willing to continue such support if necessary.

It is crucial that the community decides to coordinate such challenges and set-up a repository for the campaigns but also ensure that developed resources, metrics, know-how is documented and made available. ELRA is prepared to play a role on this.

ELRA Catalogues & ISLRN

Our primary mission was to act as a data center for Language Resources sharing. To this end, we devoted our efforts to identifying useful data sets, reviewing the community publications, the outcome of the funded projects and the disseminated information. We worked on clearing all Intellectual Property Rights (IPR) and other legal issues with the right holders (not necessarily the data owners). Additional efforts were needed to document them using very detailed meta-data descriptors and we are happy to see these meta-data schemas re-used by others. We also designed the appropriate licenses to make the LRs easily licensable and usable for various purposes. Licensing remains a critical task as we need to balance the expectations of all parties (academic users, commercial partners, funding agencies, right holders, etc.). These different tasks, important to ensure the possibility to share LRs, are documented in our Data Management Plan (DMP, `http://www.elra.info/en/services-around-lrs/dmp/`) that we make available and customizable by third parties for their own resources with our free support when needed.

We also included, in our identification and cataloguing work of LRs, the packages produced within the evaluation campaigns I mentioned above. The package often comprises the test data, the metrics, as well as the benchmarking results and publications. These packages allow to reproduce the work done with an evaluation campaign by newcomers and consist of a useful “Exist strategy” that capitalizes on the work carried out during the campaign.

As of today, over 1500 Language resources are catalogued by ELRA and made available, covering more than 80 languages and language varieties, with a large set of modalities (audio, videos, OCR, texts, lexica, etc.), and various annotations and tagging, etc. We also continue other identification tasks, either on projects like we do for the European Commission through the ELRC contract or through other internal initiatives. For instance, within the ELRC contract (`https://www.lr-coordination.eu/`), our consortium identified over 4300 resources, out of which 3450 are available to all users under permissive licenses e.g. CC and the like. I also mentioned that ELRA continues to identify and clear the IPR aspects for resources submitted to the LREMAP that inventories all resources listed by the authors of LREC papers since LREC 2010 and may other conferences that joined us. With almost 8000 items, many are available from their right holders or data centers. In addition to this, we continue to make available the set of Language Resources that have been trusted to us, often under permissive licenses, within the Share-Your-LRs initiative, an LREC feature.

LR unique identification

A major problem that we are facing as a community is the duplicate resources in the different data centers that harvest online meta-data elements, producing their own inventories of data descriptions. In early 2000, a number of Language Resources players (inc. ELRA, LDC but also centers like Alaska Native Language Center, Langues et Civilisations à Tradition Orale, American Indian Studies Research Institute) agreed to join forces to establish OLAC, the Open Language Archives Community, that would be the repository of the meta-data to which researchers could refer.

With the emergence of a sizable number of infrastructures, many of them harvesting meta-data elements, it is essential that we assign unique identifiers to each LR to avoid that users get confused when acquiring or licensing resources.

Several major data centers have introduced the concept of ISLRN (International Standard Language Resource Number, `https://islrn.org/`), inspired from the publishing world. Since its inception, more than 200 institutions have joined this service, requesting ISLRN identifiers for their resources. We count now over 3200 identifiable resources in a unique way, wherever they are, even if they get renamed, etc. Of course ISLRN is part of the meta-data descriptors but it is more reliable than the naming and is not related to internet or to a given data-center or the internet-based Doi.
We hope to make this mandatory in the major conferences to ensure that we can keep track of existing resources that are distributed through multiple channels.

Indigenous Languages

To strengthen its mission on processing all languages, ELRA continues to encourage high-quality publications on all language resources and on endangered ones at LREC, at the many satellite workshops or the Language Resources and Evaluation Journal (LRE journal). ELRA took an active part to the various debates that emerged a few years ago on Indigenous languages. The important one was initiated by the United Nations and its agency UNESCO, through the International Year of Indigenous Languages (IYIL), in 2019.

ELRA and its partners managed to bring together different communities with their respective expertise and experience related to language technologies for all at one of the most important events, the LT4ALL conference in 2019 (https://lt4all.org/en/). Through the organization of this first conference, we hope that we have open new doors and charted an important new path. We are very proud that the initial action has considerably grown to lead to the decade of Indigenous languages (2022-2032) and hopefully with strong commitments from public, civil society, private and government agencies, aiming to design an achievable roadmap for 2030. ELRA is very confident, now that the ELRA-ISCA joint SIG on the under-resourced languages (SIGUL) is fully operational and in good hands, to continue the work on this topic. More will be discussed during this conference and at the SIGUL workshop.

LR production

ELDA has been and remains a major player in the production of language resources, both in R&D projects and for its partners. This production service covers all modalities, all languages, and all types of annotations. ELDA methodologies are designed with its partners from research and industry to ensure efficients production processes and high quality outcomes with a customized validation procedure. ELDA is proud to work with a large number of LREC participants to carry out such production activity.

ELRA and Legal /Ethics

Part of the ELDA mission is to support the use of language resources or to conduct evaluation campaigns and challenges in a clean legal context and conditions. Our experts continue to clear all IPR and other legal dimension to make the resources shareable. We continue to document these legal developments worldwide through regular reports but also through direct support to cases faced by the community (http://www.elra.info/en/dissemination/legal-issues-papers/). The main outcome of this is the very efficient partnership established with other data centers and experts to organize a useful workshop that takes place at each LREC. It is the community duty to ensure that all what is done complies with the legal frameworks but also adhere to high standards when it comes to ethics. ELRA, as a data center player, focuses its efforts on developing resources that reflect reality and accounting the ground truth that can help assess state of the art systems. But this ground truth should not compromise with the various biases that clearly impact the use of our technologies when based on data learning paradigms. The community needs to make our resources irreproachable, with transparent methodologies for the production as well as for the validation and the sharing. When the systems require a ground truth that clearly comprises bias, this should be explicitely documented, and an ideal situation contemplated.

This also applies to technology development. We need to ensure that systems' analysis and outputs are fair, leading to impeccably explainable results and compliant with our ethics.

Ethic is for many of us universal but likely, for many others, culturally based. Whatever we think, the use of our resources and technologies should aspire to a fundamental respect of humankind. This is why at each LREC, ELRA and its partners organize the specific workshop, mentioned above, to discuss ethical dimensions of our work, the novelties of legal frameworks all over the world, to brainstorm about requirements regarding respect of privacy, personal information, requirement for anonymization of our data sets, etc. LREC 2022 will continue this tradition through the dedicated legal & ethical issues workshop.

Future Plans and Plans for the future

Where are we heading today? The new ELRA structure and governance will help continue to monitor the current developments. The community is heavily investing in IA-based packages and very large language models are produced and made available for various technologies and various languages even with multilingual approaches. ELRA supports such initiatives and is working to make available such Language Models, both the ones freely available under permissive licenses and the proprietary ones that could be license with particular conditions and restrictions. It's important to seek re-usability and avoid re-training again and again similar models, given the energy consumption needed. Of course such modeling should take place for new languages and new domains.
While Language Modeling is the current state of the art, new techniques and approaches will emerge sooner or later so as a community we should not abandon fundamental research on how to develop high-quality raw resources usable beyond today's state of affairs as well as the zero-resources approaches that may help many under resourced languages.

Acknowledgments

The European Language Resources Association, ELRA, and the LREC Committees acknowledge with gratitude the support and sponsoring of the following institutions.

- Google (Diamond)
- SADILAR (Silver)
- Vocapia (Silver)
- 3M (Bronze)
- Emvista (Bronze)
- Expert.ai (Bronze)
- Grammarly (Bronze)
- Délégation générale à la langue française et aux langues de France, Ministère de la Culture, France (Institutional Support, French Gouvernment)
- Région Sud Provence-Alpes-Côte d'Azur (Institutional Support, Local Gouvernment)
- MultiLingual (Media)

I would like also thank our colleagues from the local commitee, chaired by Frédéric Béchet and Philippe Blache, with the support of Nadera Bureau. The local team was involved in the organization of two consecutive LRECs (2020 and 2022). 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. Thank you also to the ILC-CNR and ELDA teams who contributed to the scientific and organizational aspects of this LREC: Victoria Arranz, Roberto Bartolini, Lucille Blanchard, Francesca Frontini, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Mickaël Rigaud, Alexandre Sicard, Kamelia Yahiaoui.

Another word of thanks

A very particular gratefulness and gratitude on behalf of all of us, for the work carried out for more than thirty years. Knut Hofland modestly wrote *"I started the Corpora list after the 13th ICAME conference in Nijmegen in 1992 with 62 subscribers. The list has now more than 6200 subscribers"*. Knut deserves our recognition for his excellent initiative, and all the efforts he did put in maintaining and moderating the Corpora list over these 30 years. ELRA is proud and prepared to take over, with the support of other institutions.

I hope that you will enjoy this 13th edition of the LREC conference, organized as a hybride event to allow more colleagues to join us, I hope that you will also appreciate the workshops that are going on site and virtual,

Marseille, France, June 21, 2022
Khalid Choukri
ELRA Secretary General and ELDA Chief Executive Officer