# The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

January - December 2014

*Vol.17 n.1 & 4*

# LREC 2014 Special Issue

**9th International Conference on Language Resources and Evaluation**

LREC 2014
Reykjavik
May 26 - 31
Harpa Conference Centre



**Contributors:**
*Irina Bokova*
*Nicoletta Calzolari*
*Khalid Choukri*
*Vigdís Finnbogadóttir*
*Mikel L. Forcada*
*Iryna Gurevych*
*Joseph Mariani*
*Amália Mendes*
*Eiríkur Rögnvaldsson*
*Volker Steinbiss*

***ISSN: 1026-8200***

***ELRA/ELDA***
*Secretary General:*
*Khalid Choukri*
*9, rue des Cordelières*
*75013 Paris - France*
*Tel: (33) 1 43 13 33 33*
*Fax: (33) 1 43 13 33 30*
*E-mail: choukri@elda.org*
*Web sites:*
*http://www.elra.info or*
*http://www.elda.org*

# *Contents*

# *Dear Colleagues,*

Like every two years, the special issue of the ELRA newsletter is devoted to the Language Resources and Evaluation Conference (LREC). The Ninth edition of the Language Resources and Evaluation Conference took place last May in Reykjavik (Iceland) under the Patronage of the UNESCO and with the support of Madame Vigdís Finnbogadóttir, Former President of Iceland and UNESCO Goodwill Ambassador for Languages.

This edition has been again very popular: 1220 participants coming from 58 countries registered to the main conference, workshops and tutorials. This time, Germany and the United States brought an equally high number of participants. The participation from European countries remains very strong and Asia is well-represented especially if we account for the many participants coming from both Japan and India.

*Nicoletta Calzolari, Vigdís Finnbogadóttir, Irina Bokova, Khalid Choukri at Harpa on May 26, 2014*

In 2014, a new feature has been introduced: the LREC Repository of shared LRs. In addition to describing their language resources (data, tools, web-services, etc.) in the LRE Map - now a normal step in the submission procedure of many conferences - the authors had the possibility to share their LRs by uploading them in a special LREC META-SHARE repository set up by ELRA.

And the features introduced in 2010 and implemented in 2012 have been extended:

- The *LRE-Map*, a mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources during the submission process. It is now part of the LREC standard submission process to both the main conference and the workshops.

- The *HLT Village*, an initiative supported by the European Commission to encourage EC-sponsored projects to gain visibility by showing their objectives, progress and activities, (live demos and/or documentation) and to foster exchanges between the projects teams and the NLP research community present at LREC. This time, 16 projects took part in this Village.

Ten years ago, the ELRA Board created the Zampolli Prize, a prize for “Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation”, to honour the memory of its co-founder and first president, Antonio Zampolli.

In 2014, the Antonio Zampolli Prize was awarded to **Alex Waibel**, from Carnegie Mellon University (USA) and Karlsruhe Institute of Technology (Germany).

**THE NEXT EDITION OF LREC WILL BE HELD IN PORTOROŽ (SLOVENIA) ON MAY 23 TO 28, 2016.**

Now concerning the content of this ELRA newsletter dedicated to LREC 2014, a few sessions’ summaries are proposed along with the Opening Ceremony speeches. A short report on participants’ feedback is also available.

Last but not least, the new resources added to the ELRA catalogue are listed at the end of this newsletter.

Nicoletta Calzolari, President

Khalid Choukri, Secretary General

# *INTRODUCTION*
# *by Nicoletta Calzolari, LREC 2014 Conference Chair and ELRA President*

*Nicoletta Calzolari*

I wish to express to Mrs. Irina Bokova, Director-General of UNESCO, the gratitude of the Program Committee, of all LREC participants and my personal for her Distinguished Patronage of LREC 2014. Languages - mentioned in the first article of UNESCO Constitution - have been at the heart of UNESCO mission and programmes throughout its history.

I am also especially grateful to Madame Vigdís Finnbogadóttir, UNESCO's Goodwill Ambassador for languages and former President of Iceland (1980-1996), first woman in the world elected as head of state in a democratic election, for the continuous personal support she has granted to LREC since our first visit in Reykjavík in 2012. In her name the Vigdís International Centre for Multilingualism and Intercultural Understanding has been established under the auspices of UNESCO to promote multilingualism and raise awareness of the importance of language as a core element of the cultural heritage of humanity. I quote a sentence from a recent interview where she says: "The land - our nature - and language, those are our national treasures": this tells a lot of why this LREC is in Iceland!

## Some figures: all records broken!

LREC 2014, the 9th LREC, with its 1227 submissions, has set a new record! We received 21% more submissions than in 2012. We continue the tradition of breaking our own previous records: out of the 1227 submissions, after the reviewing process by well 970 colleagues, we accepted 745 papers. We also accepted 22 workshops and 9 tutorials. More than 1100 participants have already registered at the beginning of May.

These figures have a meaning. The field of Language Resources and Evaluation is continuously growing. And LREC continues to be - as many say - "the conference where you have to be and where you meet everyone".

Every time I underline the fact that a relatively high acceptance rate (60.7% this time) is for us a reasoned choice. It is important to get a pulse on the situation, to monitor the evolution of the field in the many varieties of approaches and methodologies, and in particular for many different languages. For us, a lexicon in any language is as important as a lexicon in American English. Multilingualism - and equal treatment of all languages - is a feature at the heart of LREC. Other venues promote a sense of exclusivity (also through the equation low acceptance rate and great merit); we always encourage a sense of inclusiveness. This is a typical feature of LREC that makes it a special conference. Quality is not necessarily undermined by a high acceptance rate, but also by the influence of the papers on the community: the ranking of LREC among other conferences in the same area proves this. According to Google Scholar h-index, LREC ranks 4th in Computational Linguistics at a similar level of conferences using much lower acceptance rates, just like the LRE Journal also ranks 4th in the general field of Humanities, Literature and Arts.

## LREC 2014 Trends

Language Resources (LRs) being everywhere in Language Technology (LT), LREC is a perfect observation point of the evolution of the field. Looking at all the topics, while building the program and putting all the pieces together, the most striking (even if not surprising) new trend was for me the application of sentiment/opinion discovery/analysis to social media shown by so many papers.

A very rough sketch of LREC 2014 major topics and trends, from my viewpoint, is the following:

- There is a completely new topic:
  - ♦ Linked Data, also the hot topic of this edition
- Topics that were quite new in 2012 and are now consolidated:
  - ♦ Social Media, in particular combined with subjectivity, as said above
  - ♦ Crowdsourcing and Collaborative Construction of LRs
- Other increasing (not the biggest in absolute terms) topics with respect to last LREC are:
  - ♦ Subjectivity: Sentiments, Emotions, Opinions
  - ♦ Less-resourced languages, in line with the value we give to safeguarding world's linguistic diversity
  - ♦ Extraction of Information, Knowledge discovery, Text mining: always a very hot topic
  - ♦ Computer Aided Language Learning

- Stable Big topics:
  - ♦ Infrastructural issues and Large projects, and also Standards and Metadata, receive the usual attention by the LREC authors
  - ♦ Lexicons and Corpora (i.e. the most typical "data"), of many types, modalities and for many purposes and applications: they are the prominent and most crowded topic
  - ♦ Semantics and Knowledge, in all their variations: from annotation of anaphoric information, to ontologies and WordNets, sense disambiguation, named entities recognition, information extraction, to mention just a few
  - ♦ Syntax, Grammar and Parsing continues to be a largely represented topic: not solved
  - ♦ Machine Translation and Multilingualism are areas on which a lot of work is carried out
  - ♦ Speech and Multimodality keep the same level: good but not enough
  - ♦ Dialogue and discourse, with contributions from both the Speech and Text communities
  - ♦ Evaluation is pervasive/everywhere: we are proud to give evidence to its being an essential feature in the LT landscape
  - ♦ Tools, systems for text analysis and applications are presented in many papers

A usual observation is the relevance of *infrastructural issues* and the attention that LREC - and ELRA- pay to them. They are mostly neglected in other conferences. Infrastructural issues play an important role for the field of LRs and for the LT field at large. But it is a fact that the first to recognise their importance have been people of the LR area. LRs are themselves of infrastructural nature and quite naturally call for attention to these issues. The infrastructural nature of LRs, captured by the term "Resources", was highlighted in the Introduction of Antonio Zampolli to the 1st LREC in Granada in 1998.

The fact that so many topics are represented at LREC means also that all the various LR and LT subcommunities are present at LREC: this increases the LREC impact and gives to LREC the characteristic of being a true melting pot of cultures, and an enabler of new cooperation initiatives.

## 15th LREC Anniversary

LREC was born in 1998 and on the occasion of its 15th Anniversary, Joseph Mariani has prepared an analysis of all the past LREC Proceedings, rediscovering the dynamism of the field while looking at the major contributors, topics, trends, also comparing them with an analogous survey done for the speech community on the Interspeech conference series. There will also be a Quiz for all the LREC participants and a winner! The survey paper is in the Proceedings as a special paper for the 15th Anniversary and will be presented at the Closing Session.

## ELRA and LREC: a tradition of innovations at the service of our community

I am proud to announce a number of recent initiatives of ELRA and LREC that touch topics that are at the forefront of a paradigm shift and together help advance our field and increase confidence in scientific results. As an introduction I use some words of Zampolli in 1998: "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)", observing also that their activities "demand regular updating to reflect technical and strategical evolution of their environment". We try to keep with this recommendation.

These innovations - introduced by ELRA and/or LREC - must not be seen as unrelated steps, but as part of a coherent vision, promoting a new culture in our community. We want to encourage also in the field of LT and LRs what is in use in more mature sciences and ensure reproducibility as a normal part of scientific practice. We try thus to influence how our science is organised or should be organised in the future.

I give here a quick picture of some innovations that are critical for the research process and constitute a sort of manifesto for a new kind of sustainability plan around LRs.

## LRE Map

The LRE Map (http://www.resourcebook.eu/), started in 2010, is now an established tool, consulted every day and used in other major conferences. At this LREC we have collected by the authors descriptions for more than 1000 resources in more than 150 languages! Spreading the LR documentation effort across many people, instead of leaving it only in the hands of the LR distribution centres, we also encourage awareness of the importance of metadata and proper documentation. Documenting a LR is the first step towards identifiability, which in its turn is the first step towards reproducibility.

Recognising the value of Linked Data, we just published the LRE Map in LOD (Linked Open Data).

## Share your Language Resources and Reproducibility of research results: the vision

After encouraging sharing LR metadata, the next step is sharing the actual content. ELRA has embraced in the last years the notion of "open LRs": we show this also with the "Share your Language Resources" initiative started in this LREC. With it we ask all the authors to consider making background data available with their paper. More than 300 LRs have been made available: a big success for the first experiment! Showing the community commitment to sharing.

LRE Map and Share your LRs must be seen not as isolated initiatives, but as complementary steps towards implementing a new vision of the field. On one side we encourage opening data that could be valuable to others, on the other we try to encourage a sort of cultural change in our community. Here the vision: it must become common practice also in our field that in conferences and journals when you submit a paper you are offered the opportunity to upload the LRs related to your paper. We must unlock the material that lies behind the papers: the adoption of such a policy will make the whole picture clearer. We had to fight in the '90s for concepts like "reusability", which finally led to promoting the need of developing standards in our field (this was still a hot topic in

1998 at the time of the 1st LREC). Now the need for standards is consolidated and we consider it normal, but we need to start another campaign for encouraging more resource sharing. Researchers are not yet sharing very well; they tend to hold back knowledge. I hope that this sharing trend will be more easily embraced by younger colleagues who are familiar with everyday use of social media of all sorts and free ideas sharing: we must port the same attitude in the research environment. This will fundamentally change the way of making science, in a sort of light revolution towards openness of science in all its facets. Hopefully it will diminish the unfortunate phenomenon of reinventing the wheel from time to time, instead of building on your colleagues' findings.

This vision has to do with many important aspects: shifting to a culture of sharing, re-use, reproducibility of research results. If we want to become a mature science we should make data sharing become "normal" practice. Even more important in a data-intensive discipline like LT. The small cost that each of us will pay to document, share, etc. should be paid back benefiting of others' efforts and become worthwhile. This will also lead to a greater opportunity of collaboration, encouraging bigger experiments by larger collaborative teams (something else we should learn from more mature sciences). Moreover, reproducibility encourages trust.

## ISLRN

A major achievement of ELRA has been the recent establishment of the *International Standard Language Resource Number* (ISLRN) (http://www.elra.info/Establishing-the-ISLRN.html). It is a unique identifier to be assigned to each LR. Organised and sustained by ELRA, LDC and AFNLP/Oriental-COCOSDA, the ISLRN Portal provides unique identifiers to LRs. LRs in the ELRA and LDC catalogues have been the first to get an ISLRN (just one if a LR is stored in both catalogues!).

When you publish a LR it can get an ISLRN and thus become a citable product of research. Data/LR citation must become normal scientific practice also in our field, as it is in others. To make a LR citable can then pave the way to the design of a sort of "impact factor" of LRs. This can become an important incentive for the field, so that researchers can get the credit they deserve also for the LRs they developed.

ISLRN is not only linked to the possibility of getting proper "recognition" for LR developers. It would also enhance experiment replicability, an essential feature of scientific work. It may thus become a very important advance in our field.

## META-SHARE sustainability by ELRA

Through these initiatives we try to encourage community efforts towards: documentation of LRs, possibility of identification of LRs, LR sharing, making research results reproducible. There is a lot of buzz these days around these types of topics. As I said above, all these initiatives are closely related and must become integrated with each other.

For them to become common research practices these activities must be well organised and require good mechanisms behind to become possibly a set of related services on a common platform. Pooling together data from all the research described in conference and journal papers will obviously need an infrastructure for distributing research results and such a LR platform must be sustained.

The ELRA Board has decided to support the META-SHARE platform, but META-SHARE, as sustained by ELRA, must in turn be adapted to be able to support these types of initiatives and thus become also a platform for sharing reproducible research results. We must find ways to make these practices as easy as possible and rewarding for the researcher. META-SHARE, in ELRA view, should become also the obvious repository (recognised by the community) where all these types of actions are sustained and where all research results become available, discoverable, identifiable, and citable. ELRA is taking these steps to start enabling to keep track of connected research activities like papers and supporting underlying resources, in an all-inclusive way.

## LREC Proceedings in Thomson Reuters Citation Index

A great recent achievement for ELRA and LREC has been the fact that the LREC 2010 and LREC 2012 Proceedings have been accepted for inclusion in CPCI (Thomson Reuters Conference Proceedings Citation Index). This is a significant achievement for LREC and it will provide all LREC authors with a deserved recognition. It is for us of great satisfaction, in particular for the benefit it can bring to young colleagues.

## ELRA 18th anniversary and NLP12

Coordination is an important issue when infrastructural issues are at stake. None of the actions above can or should be conducted and tackled in isolation.

For this reason we - ELRA - organised, on the occasion of ELRA majority as its 18th anniversary, the first meeting of the major associations/organisations in the field of Language Resources and Technologies, Computational Linguistics, Spoken Language Processing, Big Data and Digital Humanities, the so-called NLP12 (http://www.elra.info/NLP12-Paris-Declaration.html). We started to discuss issues of common interest to coordinate some of the activities and we adopted some common resolutions, such as the encouragement of language resources and tools sharing and promotion of best practices for language resource citation in publications.

Together we should be able to take the necessary steps to better serve the field and the respective communities and to strengthen the bridges between various communities (e.g. Language Technology and Humanities).

## ELRA for Open science

I am excited and proud that we, as ELRA and LREC, can contribute to such a (quiet) revolution towards shaping a new type of open scientific information space for the future of our field, the Language Resources and Technology future. I have always felt it is our duty to use the means that we have in our hands to try to shape the future of the field, and in this case to play a role in how to change scientific practice and have an impact on the overall scientific enterprise!

Trying to be always forward-looking and to act in a proactive way to serve the field, ELRA continues to be a community-aware association. I would like to work for it to become more also a community-driven association. We would like to discuss with all those who are interested about how to tackle the challenge of truly open research (which is more than open access!) so that we can take the necessary further steps to make this process more efficient, faster and more collaborative.

It is clear that in such a campaign for the cause of reproducibility and open science and for a proper system of attribution and citation - two closely related aspects - we must involve also funding agencies that should help in supporting the necessary policy actions. For sure we will involve in this initiative the NLP12 group. But I strongly believe that the most important change must come from the mind-set of researchers. This is where LREC can help, I hope…

The message that ELRA has for its community, the LREC community, is: We are here to help!

**Acknowledgements**

In this last part I wish to express my deepest gratitude to all those who made this LREC 2014 possible and hopefully successful.

I first thank the Program Committee members, not only for their dedication in the huge task of selecting the papers, but also for the constant involvement in the various aspects around LREC. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program. To Joseph Mariani for his always wise suggestions. And obviously to Khalid Choukri, who is in charge of so many aspects around LREC.

I thank ELRA and the ELRA Board: LREC is a major service from ELRA to all the community! A very special thanks goes to Sara Goggi and Hélène Mazo, the two Chairs of the Organising Committee, for all the work they do with so much dedication and competence, and also the capacity to tackle the many big and small problems of such a large conference (not an easy task). They are the two pillars of LREC, without whose commitment for many months LREC would not happen. So much of LREC organisation is on their shoulders, and it is visible to all participants.

A particular expression of gratitude goes to the Local Committee, and especially to Eiríkur Rögnvaldsson (its Chair) and Sigrún Helgadóttir: they have worked with great commitment and enthusiasm for many months for the success of LREC always looking at the best solutions to the many local issues.

All my appreciation goes also to the distinguished members of the Local Advisory Board for their constant support.

Among the Icelanders I wish to mention Guðrún Magnúsdóttir, for a very simple reason: the idea of having LREC in Iceland came out during a lunch that the two of us had together in Berlin!

I express my gratitude to the Sponsors that believe in the importance of our conference, and have helped with financial support. I am grateful to the authorities, and all associations, organisations, companies that have supported LREC in various ways, for their important cooperation. Furthermore, on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a wonderful job.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events.

A big thanks goes to all the LREC authors, who provide the "substance" to LREC, and give us such a broad picture of the field.

I finally thank the two institutions that have dedicated such a great effort to this LREC, as to the previous ones, i.e. ELDA in Paris and ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least, thanks are thus, in addition to Hélène Mazo and Sara Goggi, to all the others who have helped and will help during the conference: Victoria Arranz, Paola Baroni, Roberto Bartolini, Irene De Felice, Riccardo Del Gratta, Francesca Frontini, Ioanna Giannopoulou, Johann Gorlier, Olivier Hamon, Jérémy Leixa, Valérie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller. You will meet most of them during the conference.

I also hope that funding agencies will be impressed by the quality and quantity of initiatives in our sector that LREC displays, and by the fact that the field attracts all the best groups of R&D from all continents. The success of LREC for us actually means the success of the field of Language Resources and Evaluation.

And lastly, my final words of appreciation are for all the LREC 2014 participants. Now LREC is in your hands. You are the true protagonists of LREC; we have worked for you all and you will make this LREC great. I hope that you discover new paths, that you perceive the ferment and liveliness of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation … which you will show at the next LREC.

LREC is not exactly in a Mediterranean location this time, even if all the hot water around gives some Mediterranean flavour! But the tradition of holding LREC in wonderful locations continues, and Reykjavík is a perfect LREC location! I am sure you will like Reykjavík and the friendliness of Icelanders. And I hope that Reykjavík will appreciate the invasion of LRECers! With all the Programme Committee, I welcome you at LREC 2014 in such a wonderful country as Iceland and wish you a fruitful Conference.

Enjoy LREC 2014 in Reykjavík!

Nicoletta Calzolari Zamorani
Istituto di Linguistica Computazionale del CNR
Via Moruzzi 1
56124 Pisa, Italy
glottolo@ilc.cnr.it

# LREC 2014 Opening Ceremony Speeches

*Address by Mrs Irina Bokova, Director-General of UNESCO, 26 May 2014*


*Irina Bokova*

Honourable former President of Iceland, Dear Vigdís, Excellencies, Ladies and Gentlemen,

It is a great honour to address this Language Resources and Evaluation Conference, for which UNESCO is pleased to give its patronage.

We are here in Reykjavik and I believe Iceland is a living example of how a deep commitment to language is a healthy foundation for engaging in dialogue with others.

Languages are who we are: protecting this identity is a matter of human rights.

It is through language that we make sense of the world around us, that we can transform it for the better.

This starts on the benches of schools. It is why UNESCO works with governments to integrate Multilingual Education as a means to promote languages and quality learning.

Children should be taught in the language that is most natural for them to speak, not in one they cannot understand. The evidence is clear: the use of mother tongue in school can be a powerful cure against illiteracy.

In Latin America, we promote the Intercultural Bilingual Education approach, so that indigenous languages become important pedagogical resources.

In Asia-Pacific, we foster supportive polices to integrate mother tongue-based multilingual education into early learning.

The world's most pressing challenges can only be solved collaboratively.

In today's increasingly multicultural societies, speaking several languages is an asset - for working cross-culturally, for living together, for understanding each other.

Multilingualism is the foundation of global citizenship.

This is the core message of the International Decade for the Rapprochement of Cultures, led by UNESCO.

It is also the core purpose of the Vigdís International Centre for Multilingualism and Intercultural Understanding, which aims to become a leading institution to foster mutual respect. I wish to warmly commend Ms Vigdís Finnbogadóttir, our Goodwill Ambassador for Languages, who was also instrumental in the creation of UNESCO's International Mother Language Day celebrated every 21 February.

It calls for stronger public policies to promote multilingualism as the pulse of cultural diversity.

UNESCO is determined to do so, notably through the Convention on the Protection and Promotion of the Diversity of Cultural Expressions, adopted in 2005.

A proverb in Marovo, a language of the Solomon Islands, says *"Those who cannot name the good things of sea and land, cannot find them, cannot benefit from them and do not know how to look after them well."*

Artic indigenous people have developed an extremely sophisticated terminology to describe their environment and the effect of climate change. For example, there are some 120 Inupiaq terms for sea ice from Wales, Alaska, including almost 75 terms for types of ice conditions.

What a wealth of knowledge!

Each of the 6,000 languages spoken in the world contains its own wealth of knowledge: imagine what we could learn on climate change or biodiversity if we could only unlock this potential!

This is why we have built a Linguistic Diversity indicator as an integral part of the Biodiversity Indicators Partnership, and we say that biological diversity is the other side of cultural diversity: the two are closely linked.

We published the first book ever-written in Mayangna, a language spoken by the indigenous people from the Central American tropical rainforest, to document their knowledge on biodiversity.

We must develop far more of such tools, indicators and statistics, both offline and online.

It is not enough to invest in technology - we must invest in local content, in local languages, and give people not only the tools to access knowledge, but also the means to contribute to the enrichment of humankind, so that the digital divide does not deepen the language divide.

This is one key aspect of the work of the Broadband Commission for Digital Development.

Early on in my first term, I signed an agreement with ICANN, the organization that coordinates the Internet's domain names system, making it possible for these names to be created in any alphabet.

Since then we have made available several new training materials on indigenous languages on the Open Training Platform - our Atlas of the World's Languages in Danger is also accessible online.

Iceland has taken bold decisions to enrich online databases with content to improve the accuracy of online translation tools.

This is a lesson for all policy makers.

Here more than anywhere else, we understand that language is another word for culture.

In times of limits - of our resources, of our planet - this is our ultimate renewable resource: human ingenuity and this is the key to dialogue, creativity and sustainability.

This is the message of UNESCO and I believe it has never been so important.

Irina Bokova
Director-General of UNESCO

*Address by Madame Vigdís Finnbogadóttir, Former President of Iceland and UNESCO Goodwill Ambassador for Languages, 28 May 2014*

Distinctive guests, ladies and gentlemen, a warm welcome to Iceland!

**Whoever doesn't live in poetry cannot survive here on earth** said our Icelandic Nobel-prizewinning author, Halldór Laxness, in one of his outstanding novels in the late twentieth century.

It is hardly possible to commence an address here in Reykjavík, the UNESCO city of literature, without making reference to our literature - which in olden times was written on parchment, then on paper, next on typewriters, and now on computers. Innumerable tales have been told in many texts - whole sagas and epics - tales of peril at sea and long journeys between Iceland and the outside world - adventures that inspire imagination and creativity. And now the texts travel at the speed of light to distance corners of the globe - and exciting tales of faraway places are no longer as exotic as they once were. And we still see the technical possibilities expanding and developing, for transporting the very heart of the human spirit - words, the expression of the human mind, exchange of speech between people - which is on the programme of this LREC event here in Iceland... where more than 20 generations have kept our language alive almost unchanged for eleven hundred years.

The conservation of a language is a tremendous responsibility - for languages store up memories of what happened before our time - and they store up knowledge already acquired - not least knowledge of the environment, which is more vital for humanity than ever before, now when the whole world has become one region, one system, and wherever we are, we are constantly reminded to treat Mother Earth with care.

By the same token, a crucial issue is… how the technology under discussion here during these days is put to use, to invigorate every single language on earth - how it will be used to build bridges between languages, and invigorate every one of them - to minimise the barriers between languages, to facilitate translations between them - so that we will be able to live in each other's poetry - wherever we are on the planet. That is actually the core of the International Language Centre that has been established here at the University of Iceland, under the auspices of UNESCO, category 2, the only one in the world, and I have the honour is bearing my name.

As a UNESCO spokesman for the safeguarding of languages and their cultures, I imagine all those nations that emerged from the collapse of the Tower of Babel - to gain their own languages and their own cultures - as a huge multicoloured tapestry - a beautiful piece of art. Some of the threads in the tapestry are strong and dominant, while others are less noticeable, in pale and delicate colours. And in the fabric of the tapestry, it is not least those deli-

*Vigdís Finnbogadóttir*

cate shades which require our care and attention - because they are the threads that hold the tapestry together. That matters to mankind, because the question of how we survive here on earth depends on our ability to express ourselves.

The technology must be used to enlighten people about languages - to awaken their curiosity about languages, and their interest in how each language reflects the place where it is spoken. In Icelandic, for instance, we have more words than most other languages for the waves and the sea - the ocean that surrounds us - reflecting its many manifestations. And we have more weather-related words in Icelandic

than in the average language - and everyone here will understand the need for that. And we have a rich and subtle vocabulary for describing horses - probably more than in any other language - except perhaps in the plains of eastern Asia - reflecting the fact that the Icelandic horse was the Icelanders' companion and only form of transport down the centuries, as Iceland was a roadless country until the early 20th century, a traditional society of fishermen and farmers. Everything in the world is reflected in languages - and that guides the way people live in every place.

It is thus of paramount importance that technology should be used to animate the spirit - and not to repress it, not to pave the way for human solitude - living with machines.

Ladies and gentlemen: "Culture means to do things the very best way," a distinguished Icelandic philosopher has said. And that is the principle I hope to see in technological development regarding the languages of the world.

Ladies and Gentlemen, - I wish you fruitful deliberations on this very important subject in the days to come.

Vigdís Finnbogadóttir
Former President of Iceland and
UNESCO Goodwill Ambassador for Languages

*Message from Khalid Choukri, 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 Resources 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 (Romance, 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 under privileged 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 dis-

cussed 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, in 1998, 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 publica-

(1) LRE Journal, http://link.springer.com/journal/10579

tions, 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 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)] and 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[(5)]. 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-foreurope-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 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

(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

(5) NLP12 meeting: http://www.elra.info/ELRA-18th-Anniversary.html

many institutions to establish their own repositories and offer their resources via internal infrastructures.

This trend definitely increases the availability of LRs (particularly 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, here 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, Valérie 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 / ELDA
9, rue des Cordelières
75013 Paris, France
choukri@elda.org

*Message from Eiríkur Rögnvaldsson, Chair of the Local Organizing Committee*

Dear LREC 2014 Participants,

On behalf of the Local Organising Committee, the Local Advisory Board, and all the participating organisations. I would first of all like to express my profound gratitude to UNESCO and to the Director-General Madam Irina Bokova for kindly agreeing to act as patron for the conference. I would also like to thank Madam Vigdís Finnbogadóttir, UNESCO Goodwill Ambassador for languages and former president of Iceland, for her invaluable support.

Iceland is a country with a rich literary heritage and Icelandic is known for having changed less in the course of the last thousand years than most other languages having a documented history. In the conference bag that you received upon arrival at the conference venue you will find a small booklet, a gift from the local organisers. The title of the booklet is *Hávamál - the Sayings of Óðinn*. It contains an English translation of the famous poem Hávamál which is more than one thousand years old. Icelanders today can still read and understand the text of Hávamál as it is preserved in the 13th century manuscript Codex regius - there is no need for translation. We want to keep it that way.

However, we now feel that Icelandic is being threatened by globalisation and modern technology. Our precious language,

*Eiríkur Rögnvaldsson*

which the manuscripts have preserved for us for almost thousand years, might be lost in only a few decades, if it does not adapt to current and future information technology. Icelandic language technology is still in its infancy, and the META-NET survey of 30 European languages demonstrated that language technology support for Icelandic is very limited. Therefore, it is both symbolic and extremely important for us to have the LREC conference here in Iceland - the first time the conference is held outside of the Mediterranean area. It will hopefully raise awareness of the importance of language technology, among politicians, policy makers, academics, journalists, and not least the general public.

I would like to thank all who have made it possible to have the conference here in Iceland - the ELRA team in Paris and Pisa, the Programme Committee and all the reviewers, the Local Committees, our conference bureau Congress Reykjavík, the Ministry of Education, Culture and Science, and others who have supported the conference in various ways. I cannot mention all the people involved but I want to single out two names. Guðrún Magnúsdóttir first mentioned to us the idea of having LREC in Iceland, and she convinced the ELRA team that this was possible - and feasible. Sigrún Helgadóttir has been the secretary of the Local Organising Committee and has done a tremendous job in organising and coordinating innumerable things that had to be taken care of.

Let me conclude by welcoming you all to Iceland, to Reykjavík, UNESCO city of literature, and to the Harpa Concert Hall and Conference Centre, winner of the 2013 European Union Prize for Contemporary Architecture - Mies van der Rohe Award. I sincerely hope you will enjoy LREC 2014, and that the conference and your visit as a whole will be an unforgettable experience.

Eiríkur Rögnvaldsson
Chair of the Local Organizing Committee

## *LREC 15th Anniversary Celebration, Joseph Mariani, ELRA Honorary President, on behalf of the conference Program Committee*

Activities in the area of Language Resources and Evaluation greatly increased over the past 30 years, due to the importance of Language Resources to conduct research investigations in language sciences and to develop language processing systems which are based on automatic Machine Learning.

Some milestones may be identified in this area, such as the launching of the evaluation campaigns of speech recognition systems by NIST for DARPA in 1987, the creation of the Linguistic Data Consortium (LDC) and of the Coordinating Committee on Speech Databases and Speech Input/Output Systems Assessment (Cocosda) in 1991. This was followed by the launching of the European Language Resources Association (ELRA) in 1995, which organized the first Language Resources and Evaluation (LREC) conference in 1998. The oriental branch of Cocosda organized the Oriental-Cocosda conference for the first time on the same year. The Language Resources and Evaluation Journal published by Springer was initiated in 2005.

The idea of adding a scientific dimension to the Language Resources distribution activity provided by ELRA through an international conference specifically devoted to Language Resources and Evaluation was first proposed by Joseph Mariani in 1997. The first conference was held in 1998 in Granada (Spain). It was organized and chaired by Antonio Zampolli. Following its great success, the LREC conference has been organized every two years since then and is now chaired by Nicoletta Calzolari, with the continuous support of Khalid Choukri from the very beginning. Nowadays, LREC conferences gather more than 1,200 participants which form a strong interdisciplinary community.

The LREC 2014 conference gives the opportunity to warmly acknowledge and thank those who contributed to make the whole enterprise successful on the occasion of the LREC 15th anniversary.

It was felt useful to reconsider the last 15 years of research in the area of Language Resources and Evaluation through an analysis of the proceedings of the former 8 LREC conferences gathered in the LREC Anthology, with the purpose to illustrate the strong dynamism and the multiple facets of our domain and to inspire us in building up the steps for the forthcoming 15 years.

This analysis is similar to comparable exercises which were conducted in the Computational Linguistics community through an analysis of the ACL Anthology which covered 50 years of the ACL conferences and was presented within a specifically dedicated workshop at the 2012 ACL conference in Jeju (Korea), or in the Spoken Language Processing community through the analysis of the ISCA Archive which covered 25 years of the ECST, Eurospeech, ICSLP and Interspeech conference series and was presented at the 2013 Interspeech conference in Lyon (France).

A previous exercise over 15 years of the IEEE ICASSP conference was conducted by the end of the 1980s and served for deciding the launch of the ESCA Eurospeech conference.

A Quiz is also being organized on the content of the analysis during the LREC 2014 conference.

The survey will be presented during the Closing Session on Friday May 30, and the results of the Quiz with the name of the winner(s) will then be reported.

Joseph Mariani
ELRA Honorary President

*Joseph Mariani*

# *LREC 2014 Antonio Zampolli Prize*

*Speech given at the Closing Ceremony by Nicoletta Calzolari and Khalid Choukri*

*This year, the Antonio Zampolli Prize was awarded to:*

***Alex Waibel***

from Carnegie Mellon University (USA) and Karlsruhe Institute of Technology (Germany)

*From the Prize statutes:*

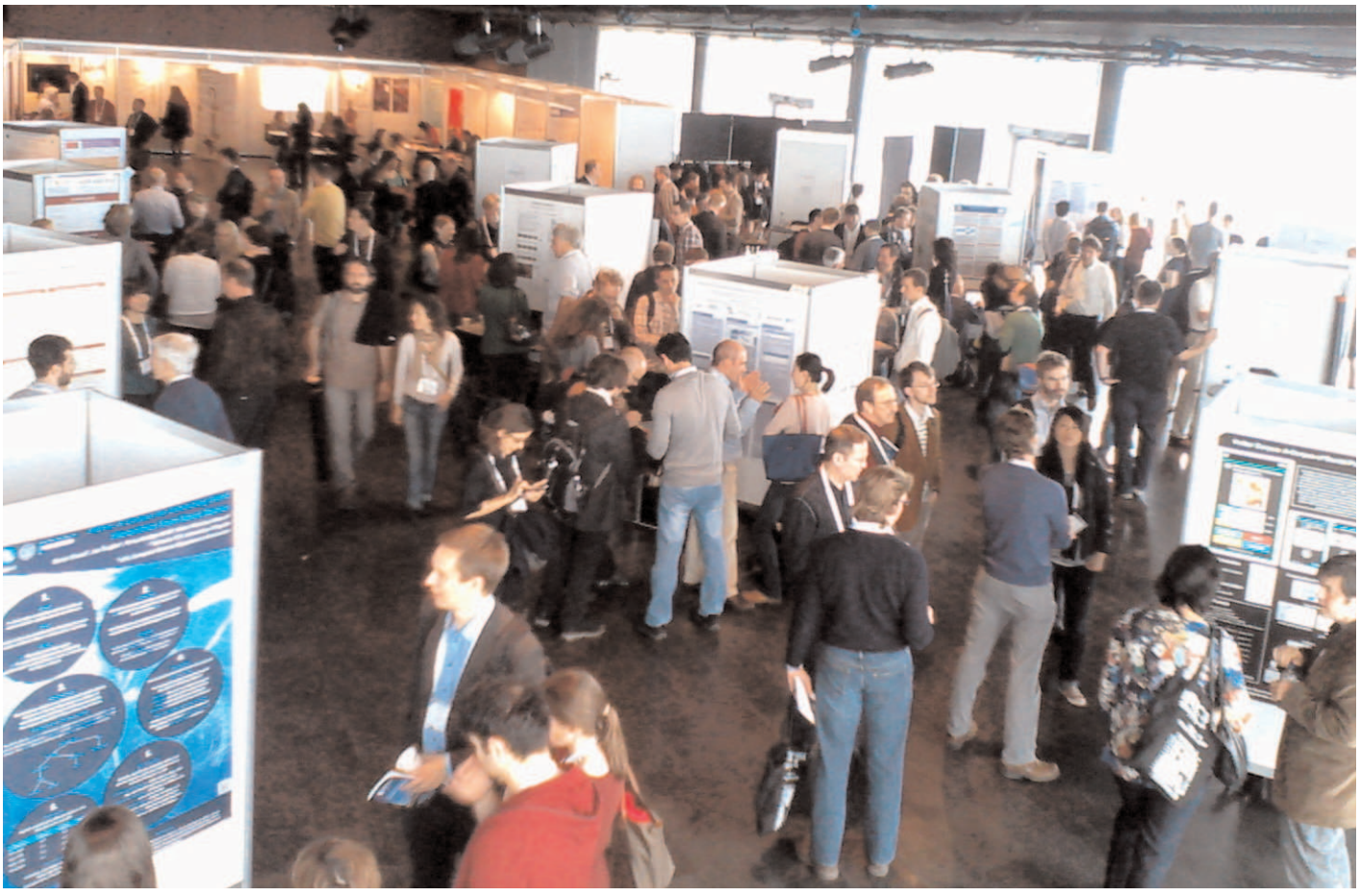
"The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation."

Just a few words about the Antonio Zampolli Prize, the prize created by the ELRA Board in order to honour our founder and first president who did so much for the field of language resources. Citing the prize articles: "The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation. In awarding the prize we are seeking to reward and encourage innovation and inventiveness in the development and use of language resources and evaluation of HLTs". At the LREC 2014 conference, the Prize was be awarded for the sixth time. The ELRA Board has been very happy to receive nominations made by outstanding people in the field, and we recognize there are several persons who are eligible for this prestigious prize.


*Khalid Choukri, Alex Waibel, Nicoletta Calzolari*

# LREC 2014 Oral and Poster Sessions Summaries

*The references given in the summaries all point to papers presented in each session. For the complete references, we invite you to refer to the LREC 2014 Proceedings, which are available online: http://www.lrec-conf.org/proceedings/lrec2014/index.html*



## O11 - Collaborative Resources

*Iryna Gurevych*

**Papers presented:**

- Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines - Marta Sabou, Kalina Bontcheva, Leon Derczynski and Arno Scharl
- Towards an Environment for the Production and the Validation of Lexical Semantic Resources - Mikaël Morardo and Eric De La Clergerie
- Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar - Dan Flickinger, Emily M. Bender and Stephan Oepen
- Mapping CPA Patterns onto OntoNotes Senses - Octavian Popescu, Martha Palmer and Patrick Hanks

LREC 2014's session on Collaborate Resources took place on Wednesday, 28 May from 16:45 to 18:05 in the Silfurberg A hall of the Harpa Conference Center. It was chaired by Iryna Gurevych of the Ubiquitous Knowledge Processing Labs at Technische Universität Darmstadt and DIPF Frankfurt. The session was well attended and featured four paper presentations.

The first talk, "Corpus Annotation through Crowdsourcing", was presented by Leon Derczynski. It provided a general overview of crowdsourcing corpus annotations, including guidelines, best practices, and caveats. The speaker explained that approaches to crowdsourcing could be categorized into three different genres according to the reward scheme (or lack thereof) employed. He then went on to describe the general four-step workflow for every crowdsourcing annotation project, along with the necessary decisions and deliverables at each step. At the end of the talk, discussion focused on how to determine the best genre and (where applicable) remuneration for a given task.

The second talk, presented by Éric de la Clergerie, was titled, "Towards an Environment for the Production and the Validation of Lexical Semantic Resources". The talk described a procedure for producing a large amount of lexical semantic information by running a parser and terminology extractor on large-scale French-language corpora, as well as an online interface for visualization and collaborative evaluation of this data. Discussion centred on possibilities for how the interface itself could be evaluated from an HCI perspective.

The third talk, "Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar", was given by Dan Flickinger. The work, which has been ongoing for some 20 years, aims at producing a rich catalogue of interoperable semantic analyses, and particulary in identifying semantic fingerprints for various phenomena. To this end they have developed a description language for logical form, applied it to a large data set, and searched for characteristic patterns in the resulting annotations. The discussion period focused on the work's relevance to collaborative resources, and specifically on the researchers' aspirations for analytical formal collaboration on the task.

Iryna Gurevych
Ubiquitous Knowledge Processing Labs at Technische Universität Darmstadt
DIPF Frankfurt, Germany
gurevych@ukp.informatik.tu-darmstadt.de

## *P15 - Lexicons*

*Amália Mendes*

**Papers presented:**

- Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing - Ismail El Maarouf, Jane Bradbury, Vít Baisa and Patrick Hanks
- GLÀFF, a Large Versatile French Lexicon - Nabil Hathout, Franck Sajous and Basilio Calderone
- Bilingual Dictionary Construction with Transliteration Filtering; John Richardson, Toshiaki Nakazawa and Sadao Kurohashi
- Bootstrapping Open-Source English-Bulgarian Computational Dictionary - Krasimir Angelov
- MotàMot Project: Conversion of a French-Khmer Published Dictionary for Building a Multilingual Lexical System - Mathieu Mangeot
- RELISH LMF: Unlocking the Full Power of the Lexical Markup Framework - Menzo Windhouwer, Justin Petro and Shakila Shayan
- Building a Dataset of Multilingual Cognates for the Romanian Lexicon - Liviu Dinu and Alina Maria Ciobanu
- LexTec - a Rich Language Resource for Technical Domains in Portuguese - Palmira Marrafa, Raquel Amaro and Sara Mendes

This poster session introduced new lexical resources, either mono or bilingual, an alternative LMF schema, and the use of a pattern dictionary for semantic parsing. One such new resource is GLÀFF, a large-scale versatile French lexicon extracted from Wiktionary, and containing for each entry inflectional features and phonemic transcriptions (Hathout et al.). Lextec is a rich computational language resource in Portuguese, encoding a representative set of terms for ten different technical domains and integrating each entry in a domain-specific wordnet (Marrafa et al.). The paper by Dinu & Ciobanu addresses the specific case of building a dataset of multilingual cognates for the Romanian Lexicon and proposes a dictionary-based approach to identifying cognates based on etymology and etymons. The MotàMot Project produced a multilingual lexical system covering French and an under-resourced language: Khmer, reusing existing dictionaries (Mangeot). Two other bilingual lexicons were introduced: a bilingual Japanese-English transliteration lexicon of technical terms in the scientific domain (including MWE), using a novel transliteration similarity measure for translation pairs identification (Richardson et al.); an open-source English-Bulgarian dictionary, based on existing and freely available resources for the two languages, which can be used as either a pair of two monolingual morphological lexicons, or as a bidirectional translation dictionary between the languages (Angelov). Standards and interoperability in the production of lexicons are addressed by Windhouwer et al.: the authors present RELISH LMF, an alternative LMF schema that uses two validation XML languages instead of a DTD and provides a set of extensible modern schema modules. The potential of lexical resources for semantic parsing is explored by El Maarouf et al.: the paper reports the results of experiments in semantic parsing using the Pattern Dictionary of English Verbs.

Amália Mendes
Centre of Linguistics of the University of Lisbon (CLUL), Portugal
amalia.mendes@clul.ul.pt

## *P49 - Multimodality poster session*

*Volker Steinbiss*

**Papers presented:**

- Representing Multimodal Linguistic Annotated Data - Brigitte Bigi, Tatsuya Watanabe and Laurent Prévot
- Single-Person and Multi-Party 3D Visualizations for Nonverbal Communication Analysis - Michael Kipp, Levin Freiherr von Hollen, Michael Christopher Hrstka and Franziska Zamponi
- The AV-LASYN Database: a Synchronous Corpus of Audio and 3D Facial Marker Data for Audio - Visual Laughter Synthesis-Huseyin Cakmak, Jerome Urbain, Thierry Dutoit and Joelle Tilmanne
- Linking Pictographs to Synsets: Sclera2Cornetto - Vincent Vandeghinste and Ineke Schuurman
- The MMASCS Multi-Modal Annotated Synchronous Corpus of Audio, Video,

Facial Motion and Tongue Motion Data of Normal, Fast and Slow Speech - Dietmar Schabus, Michael Pucher and Phil Hoole

- Mining a Multimodal Corpus for Non-Verbal Behavior Sequences Conveying Attitudes - Mathieu Chollet, Magalie Ochs and Catherine Pelachaud
- The IMAGACT Visual Ontology. an Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action - Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini and Alessandro Panunzi
- Multimodal Dialogue Segmentation with Gesture Post-Processing - Kodai Takahashi and Masashi Inoue
- The D-ANS corpus: the Dublin - Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech - Shannon Hennig, Ryad Chellali and Nick Campbell

The multimodality session presented a wide range of corpora with two or more input modalities, and some posters gave also an outlook to the related research: A corpus of biosignal and audio/video recordings of conversational speech; a multimodal corpus for dialogue segmentation with gesture post-processing (to improve segment boundary detection); a synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis; a corpus recording audio, video, facial motion and tongue motion of normal, fast and slow speech; a multimodal corpus for non-verbal behavior conveying attitude change; and a corpus with a mapping between synsets and pictograms, to help people who are not sufficiently proficient in reading. The other posters presented methodologies. An annotation scheme was described that takes into account the uncertainty of segment boundaries, such that non-matching boundaries from different streams are handled properly. Motivated by language learning, a visual ontology in particular for action verbs was presented. A set of 3D visualization methods was proposed to be used in nonverbal communication analysis.

Dr. Volker Steinbiss
RWTH Aachen University, Germany
steinbiss@informatik.rwth-aachen.de

## *P53 - Machine Translation*

*Mikel L. Forcada*

**Papers presented:**

- HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation - Ondrej Bojar, Vojtech Diatka, Pavel Rychlý, Pavel Stranak, Vit Suchomel, Aleš Tamchyna and Daniel Zeman
- Online Optimisation of Log-linear Weights in Interactive Machine Translation - Mara Chinea-Rios, Germán Sanchis Trilles, Daniel Ortiz-Martínez and Francisco Casacuberta
- An Efficient and User-friendly Tool for Machine Translation Quality Estimation - Kashif Shah, Marco Turchi and Lucia Specia
- Word Alignment-Based Reordering of Source Chunks in PB-SMT - Santanu Pal, Sudip Kumar Naskar and Sivaji Bandyopadhyay
- Comparing the Quality of Focused Crawlers and of the Translation Resources Obtained from them - Bruno Laranjeira, Viviane Moreira, Aline Villavicencio, Carlos Ramisch and Maria José Finatto
- N-gram Counts and Language Models from the Common Crawl - Christian Buck, Kenneth Heafield and Bas van Ooyen
- A Corpus of Machine Translation Errors Extracted from Translation Students Exercises - Guillaume Wisniewski, Natalie Kübler and François Yvon
- Pre-ordering of Phrase-based Machine Translation Input in Translation Workflow - Alexandru Ceausu and Sabine Hunsicker
- A Wikipedia-based Corpus for Contextualized Machine Translation - Jennifer Drexler, Pushpendre Rastogi, Jacqueline Aguilar, Benjamin Van Durme and Matt Post

This poster session on machine translation was a very well attended one. All nine posters were on display on time, but the authors for poster “Word alignment-based reordering of chunks in PB-SMT” asked a colleague to hang it, who however was not present during the session. Discussions were lively and extended well into coffee break time. Five out of nine of the posters dealt with corpora to fuel data-driven machine translation: Buck et al. crawled and automatically curated a massive monolingual corpus of English and showed BLEU score improvements when using language models based on them; Laranjeira et al.’s poster used multiword terms for focused crawling of comparable corpora in the domain of dermatology; Bojar et al. described the acquisition, filtering and correction of Hindi monolingual data and English-Hindi parallel data; Wisniewski et al. described a corpus of machine translation errors built from student postediting exercises, and Drexler et al. used articles cited in Spanish wikipedia entries to add small amounts of in-domain target-language data to improve the translation into Spanish of the corresponding English wikipedia article. Chinea-Rios et al. used a new regression-based method to tackle the difficult problem of online learning of feature weights in interactive machine translation with minor improvements, confirming the hardness of the task. Shah et al. presented two web-based demos to estimate the quality of raw machine translation. Finally, Ceausu and Hunsicker used dependency-grammar based pre-reordering before hierarchical phrase-based machine translation and found improvements of significance to translation workflows where succesful parses were obtained.

Mikel L. Forcada
Universitat d’Alacant, Spain
mlf@dlsi.ua.es

# *Miscellaneous Information*

## LREC 2014 Conference Survey Report

Following each edition of the Conference, ELRA conducts an online survey of LREC participants to collect feedback, improve the overall organization of the event and address the concerns and needs of LREC participants.

This year, 225 respondents participated in the survey which is less than in 2012 (280 respondents). The survey contained 22 questions, including some on Reykjavik as a conference destination which only a few respondents answered. The majority of the responses reflected positive feedback, in particular with regard to the conference organization, and many comments congratulated and thanked the LREC 2014 organization. Some answers and comments on the quality of the papers but also the space allocated to poster sessions, HLT Village or coffee-breaks were found very useful and will be taken into account in the next edition's organization. In 2014, the Gala Dinner was organized as a standing dinner for the first time; the many respondents' comments on the dinner show that the setting has been variously appreciated.

## Language Resources and Evaluation Journal

During the review process of abstracts submitted to LREC 2014, the reviewers are asked to assess the appropriateness of papers to be published in the LRE Journal. A number of accepted papers having met this criteria (positive review from the 3 reviewers) have been selected and their authors invited to submit an extended version of their conference paper to the LRE Journal. After a regular review process, the selected papers will be published in a special issue of the LRE Journal.

## LRE Map

The new resources collected during the LREC 2014 submission process have been added to the LRE Map. The interface of the website is being redesigned and the data normalized for an enhanced browsing and access.

*www.resourcebook.eu*

## MEDAR Grants

In 2012, the MEDAR Consortium (involved in MEDAR and NEMLAR Projects on "Mediterranean Arabic Language and Speech Technology", completed with the support by the European Commission) decided to award grants to Masters' and PhD students in order for them to attend LREC 2012. Seven grants were awarded to students who came to Istanbul and attended the conference.

In 2014, the consortium has decided to perpetuate the initiative. Although, like in 2012, each grant covered the registration fees and supported part of of the travel and accommodation expenses, only one grant was awarded to Kamal Abou Mikhael, PhD Student at the American University of Beirut (AUB) in Lebanon, to come to Reykjavik and attend LREC 2014.
Kamal Abou Mikhael produced an extensive report covering workshops, oral and poster sessions.

The full report is available at:

*http://lrec2014.lrec-conf.org/media/filer_public/2014/12/09/lrec2014-medar-report-kamal-abou-mikhael.pdf*

## LREC 2016

The next Language Resources and Evaluation Conference will take place on May 23 to 28, 2016. The Conference Centre, located in Portorož (Slovenia), will host the 10th edition of the conference.

*www.lrec-conf.org/lrec2016*

# New Resources

## *Desktop/Microphone Resources*

• ELRA-S0366 LECTRA (LECture TRAnscriptions in European Portuguese)
http://catalog.elra.info/product_info.php?products_id=1221

This corpus is composed of the audio and the manual transcriptions from seven 1-semester University courses in Portuguese. The corpus contains a total of 28 hours of audio speech that were manually transcribed by several trained annotators. The corpus is comprised of technical University lectures.

• ELRA-S0367 CORAL Corpus
http://catalog.elra.info/product_info.php?products_id=1222

The CORAL Corpus is a collection of spoken dialogues in European Portuguese. It consists of 56 dialogues about a predetermined subject: maps. One of the participants (giver) has a map with some landmarks and a route drawn between them; the other (follower) has also landmarks, but no route and consequently must reconstruct it. Only orthographic transcription was done for the whole corpus. A pilot recording was annotated in several levels.

• ELRA-S0368 Nepali Spoken Corpus
http://catalog.elra.info/product_info.php?products_id=1219

The Nepali Spoken Corpus contains audio recordings from different social activities within their natural settings as much as possible, with phonologically transcribed and annotated texts, and information about the participants. A total of 17 types of activity were recorded. The total temporal duration of the recorded material is 31 hours and 26 minutes.

• ELRA-S0369 CLIPS_MT_MANUAL
http://catalog.elra.info/product_info.php?products_id=1220

CLIPS_MT_MANUAL is a sub-corpus of the original Italian CLIPS corpus (Corpora e Lessici dell'Italiano Parlato e Scritto). This corpus contains 3228 inspected and partially repaired WAV signal files, each containing one dialogue turn (*.wav), 3228 corrected original CLIPS annotation files (*.acs, *.phn, *.std, *.wrd), 3228 BAS Partitur files containing the annotation tiers ORT, KAN and SAP (*.par), 3228 EMU database annotation files (*.vot, *.hlb) covering 30 maptask dialogues performed by 30 speakers (each speaker pair performing two different map tasks) recorded in 15 different locations in Italy in 2000-2004.

• ELRA-S0370 MoveOn Speech and Noise Corpus
http://catalog.elra.info/product_info.php?products_id=1223

The MoveOn Speech and Noise Corpus is a corpus recorded under the extreme conditions of the motorcycle environment within the MoveOn project. The speech utterances are in British English approaching the issue of command and control and template driven dialog systems with a focus on - but not limited to - the police domain. The major part of the corpus comprises noisy speech and environmental noise recorded on a motorcycle. Several clean speech recording sessions with the same recording setup (including the motorcycle helmet) in an office environment complete the corpus.

• ELRA-S0371 PortMedia French and Italian corpus
http://catalog.elra.info/product_info.php?products_id=1224

This corpus contains 700 transcribed dialogues from about 140 French speakers and 604 transcribed dialogues from about 150 Italian speakers (several dialogues per speaker). The method chosen for the corpus construction process is that of a 'Wizard of Oz' (WoZ) system. This consists of simulating a natural language man-machine dialogue. The scenario was built in the domain of touristic information and reservation. A manual transcription and semantic annotation of the corpus are provided with corresponding wave files.

## *Telephone Resources*

• ELRA-S0365 aGender
http://catalog.elra.info/product_info.php?products_id=1214

aGender contains speech sample recordings over public telephone lines with read and (semi-)spontaneous speech. Native German speakers called a voice portal from their private phone, and read text + answered some open questions. The corpus contains the voices of 945 German speakers (approx. minimum of 100 speakers per class), each delivering 18 speech items in up to six different sessions.

## *Pronunciation Dictionaries*

GlobalPhone is a multilingual speech and text database collected at Karlsruhe University, Germany. The GlobalPhone pronunciation dictionaries contain the pronunciations of all word forms found in the transcription data of the GlobalPhone speech & text database. The pronunciation dictionaries are currently available in 18 languages: Arabic (29230 entries/27059 words), Bulgarian (20193 entries), Croatian (23497 entries/20628 words), Czech (33049 entries/32942 words), French (36837 entries/20710 words), German (48979 entries/46035 words), Hausa (42662 entries/42079 words), Japanese (18094 entries), Polish (36484 entries), Portuguese (Brazilian) (54146 entries/54130 words), Russian (28818 entries/27667 words), Spanish (Latin American) (43264 entries/33960 words), Swedish (about 25000 entries), Turkish (31330 entries/31087 words), Vietnamese (38504 entries/29974 words), Chinese-Mandarin (73388 pronunciations), Korean (3500 syllables), and Thai (a small set with 12420 pronunciation entries of 12420 different words, and does not include pronunciation variants, and a larger set which contains 25570 pronunciation entries of 22462 different words units, and includes 3108 entries of up to four pronunciation variants).

Available GlobalPhone Pronuncation Dictionaries are listed below:

- ELRA-S0340 GlobalPhone French Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1197

- ELRA-S0341 GlobalPhone German Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1198

- ELRA-S0348 GlobalPhone Japanese Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1199

- ELRA-S0350 GlobalPhone Arabic Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1200

- ELRA-S0351 GlobalPhone Bulgarian Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1201

- ELRA-S0352 GlobalPhone Czech Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1202

- ELRA-S0353 GlobalPhone Hausa Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1203

- ELRA-S0354 GlobalPhone Polish Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1204

- ELRA-S0355 GlobalPhone Portuguese (Brazilian) Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1205

- ELRA-S0356 GlobalPhone Swedish Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1206

- ELRA-S0358 GlobalPhone Croatian Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1207

- ELRA-S0359 GlobalPhone Russian Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1208

- ELRA-S0360 GlobalPhone Spanish (Latin American) Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1209

- ELRA-S0361 GlobalPhone Turkish Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1210

- ELRA-S0362 GlobalPhone Vietnamese Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1211

- ELRA-S0363 GlobalPhone Chinese-Mandarin Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1212

- ELRA-S0364 GlobalPhone Korean Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1213

- ELRA-S0372 GlobalPhone Thai Pronunciation Dictionary

http://catalog.elra.info/product_info.php?products_id=1229

## Written Corpora

• ELRA-W0074 Amharic-English bilingual corpus
http://catalog.elra.info/product_info.php?products_id=1215

The Amharic-English bilingual corpus contains parallel text from legal and news domains in Amharic script, in transliterated form and in English. The size of the corpus is of 232,653 words in Amharic and 291,701 in English.

• ELRA-W0076 Nepali Monolingual written corpus
http://catalog.elra.info/product_info.php?products_id=1216

The Nepali Monolingual written corpus comprises the core corpus (core sample) and the general corpus. The core sample (CS) represents the collection of Nepali written texts from 15 different genres with 2000 words each published between 1990 and 1992. It is based on FLOB/FROWN corpora and contains 802,000 words. The general corpus (GC) consists of written texts collected opportunistically from a wide range of sources such as the internet webs, newspapers, books, publishers and authors. It contains 1,400,000 words.

• ELRA-W0077 English-Nepali Parallel Corpus
http://catalog.elra.info/product_info.php?products_id=1217

This corpus consists of a collection of national development texts in English and Nepali. A small set of data is aligned at the sentence level (27,060 English words; 21,756 Nepali words), and a larger set of texts at the document level (617,340 English words; 596,571 Nepali words). An additional set of monolingual data in Nepali is also provided (386,879 words in Nepali).

• ELRA-W0078 NE3L named entities Arabic corpus
http://catalog.elra.info/product_info.php?products_id=1226

The Arabic corpus contains 103,363 words coming from articles extracted from “Le Monde Diplomatique” newspaper, and published in 2004. 2 named entity categories were taken into account: Time and Amount.

• ELRA-W0079 NE3L named entities Chinese corpus
http://catalog.elra.info/product_info.php?products_id=1227

The Chinese corpus contains 79,302 words coming from articles extracted from “Le Monde Diplomatique” newspaper, and published in 2001. 3 named entity categories were taken into account: Person, Place and Organisation.

• ELRA-W0080 NE3L named entities Russian corpus
http://catalog.elra.info/product_info.php?products_id=1228

The Russian corpus contains 75,784 words coming from articles extracted from “Izvestia" newspaper”, and published in 1995. 2 named entity categories were taken into account: Time and Amount.

## Evaluation Packages

• ELRA-E0041 CHIL 2007+ Evaluation Package
http://catalog.elra.info/product_info.php?products_id=1196

The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker’s voice and background sounds.

The CHIL 2007+ Evaluation Package includes: 1) CHIL 2007 Evaluation Package (see ELRA-E0033) and 2) additional annotations which have been created within the scope of the Metanet4u Project (ICT PSP No 270893), sponsored by the European Commission.

• ELRA-E0042 CLEFeHealth 2013 Evaluation Package
http://catalog.elra.info/product_info.php?products_id=1218

The CLEFeHealth 2013 Task 3 Evaluation Package contains data used for the User-centred health information retrieval Shared task at the CLEFeHealth Lab conducted in 2013. Task 3 aimed at evaluating information retrieval to address questions patients may have when reading clinical reports.