

Annual Report 2016



Word by the President & Secretary General of ELRA

Many strategic activities will be carried out in 2017 and 2018 by the ELRA Team (Board and staff). The highlights are as follows :

- **Preparation of LREC 2018**

For the first time, the conference will be organized in Japan.

- **LRE-Map**

LRE-Map, a major feature of LREC, will offer a stand-alone version to help insert more resources by contributions. The main LRE-Map site will supply additional features to provide a sharper picture of LR availability and re-use.

- **Membership Policy**

The new plans and strategy for the evolution of the membership policy, to extend membership to all individuals attending LREC will be reviewed and finalized. The ELRA statutes will be changed accordingly. Plans will consider the design and the implementation of new services around LRs to better respond to members' expectations.

- **LRL Committee**

Work of the Less-resourced Languages Committee will continue. Several events are under discussion to collect and compile more facts and see how to design actions to ensure that Less-resourced Languages are considered by the research community.

- **Strategic meeting**

A strategic meeting will be convened in 2017 with representatives of major Language Technologies players in Europe so as to define a consistent and harmonized position with regard to language technologies at the European level.

"We continue to vision 2017-2020 with emphasis on a renovation of ELRA policy and activities. ELRA continues to discuss potential extension of its membership to new types of members in particular to attract individuals so it can serve them better. ELRA is also looking to more LR production activities on request."

Henk van den Heuvel



Khalid Choukri, Secretary General



ELRA Board

Board Officers

Henk van den Heuvel
President



CLST, The Netherlands

Thierry Declerck
Vice-President



DFKI, Germany

Maria Gavrilidou
Secretary



ILSP, Greece

Tatjana Gornostaja
Treasurer



Tilde, Latvia

Board Members

Gilles Adda



LIMSI-CNRS, France

Nuria Bel



UPF, Spain

Antonio Branco



Univ Lisbon, Portugal

Marko Grobelnik



JSI, Slovenia

Simonetta
Montemagni



ILC-CNR, Italy

Honorary Presidents

Nicoletta Calzolari



ILC-CNR, Italy

Joseph Mariani



LIMSI-CNRS, France

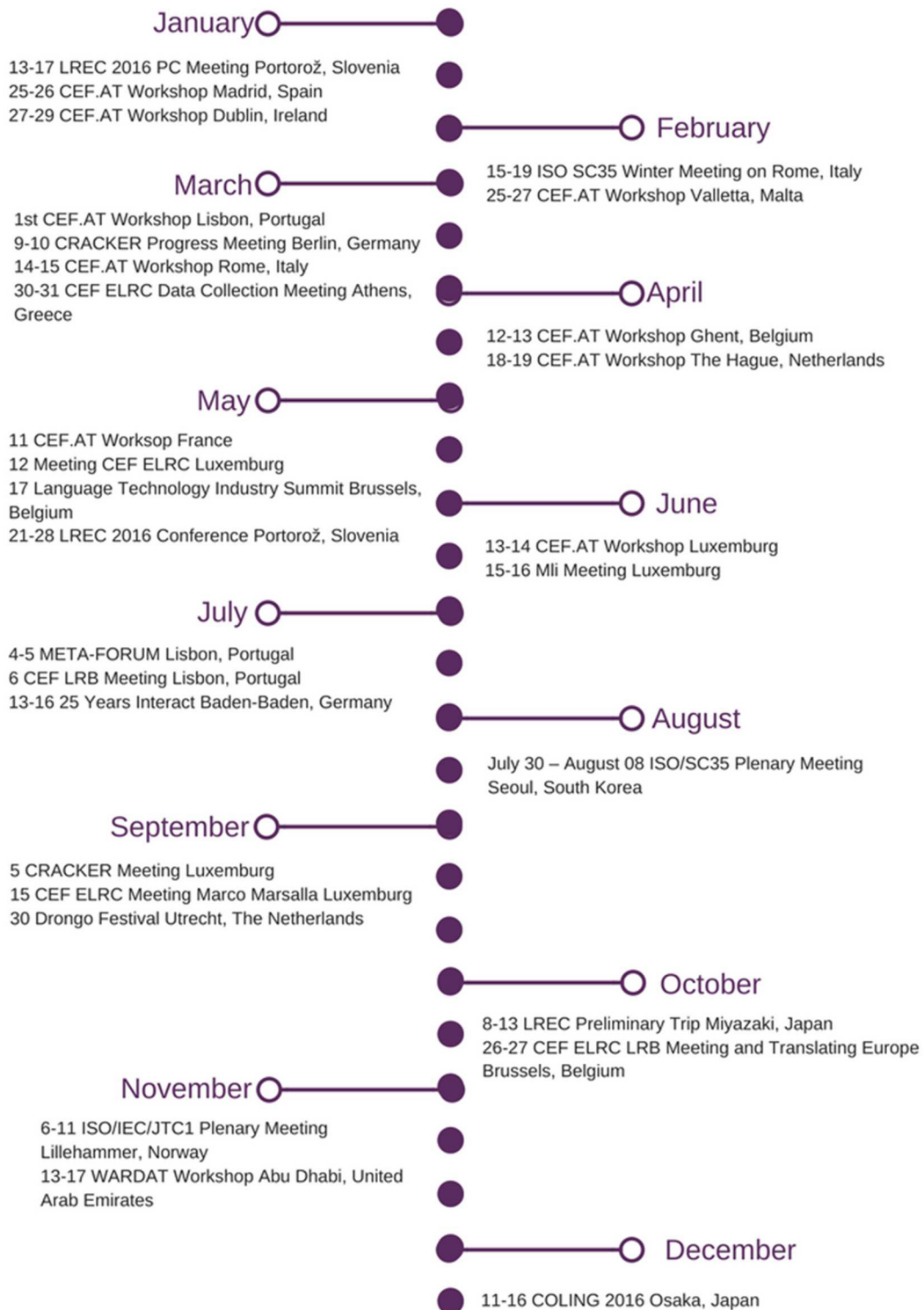
Secretary General

Khalid Choukri



ELDA, France

2016 at a Glance



Highlights on 2016 Activities

With over 50 institutional members, the representatin of the field is steady.

- **LREC 2016 in Portorož**

This 10th edition was a very successful event which gathered over 1200 participants representing all continents, Europe and North America remaining the best represented continents. 1250 abstracts have been submitted and 744 (203 Orals and 541 Posters) have been presented. Overall, the feedback survey shows the LREC 2016 participants were broadly satisfied by the organization and by the quality of the papers. For this edition, they acknowleged the Welcome Reception as their preferred social event.

- **Language Resources**

New resources have been added to the catalogue on a regular basis. All of them are allocated an ISLRN to ensure that they are unique on one hand and improve both their persistency and the computation of their impact factor on the other hand.

- **Less-Resourced Languages Committee**

The LRL committee is now in operation and a first overview (following a survey) will be conducted to draw a clear landscape of the LR gaps for some of the languages.

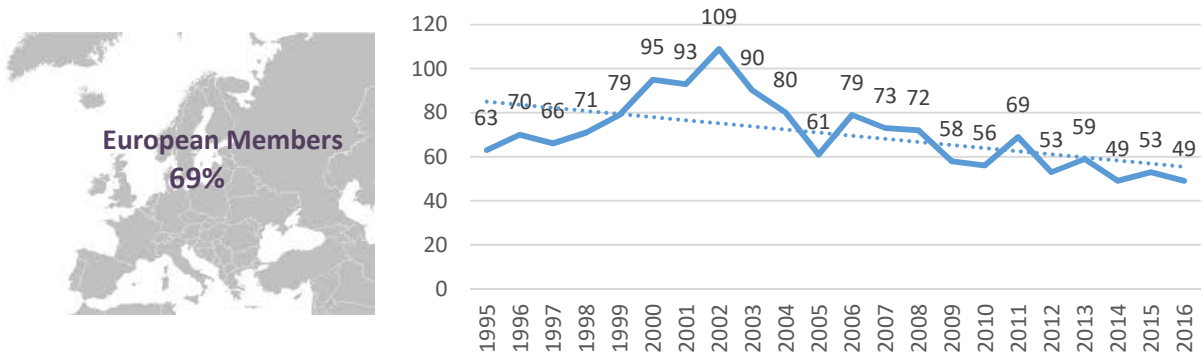
- **CEF-AT Initiative**

ELRA has continued its involvement in the European CEF (Connecting Europe Facility) and is contributing to the European Language Resource Board, targeting the compilation of resources for MT training and improvement.

ELRA Members

ELRA membership scheme is **institution-based**. Any organisation, public or private, European or non-European, can join. However, full membership, with voting rights, is available only to organisations legally established in Europe, as per the article 5 of the Statutes.

2016 ELRA Membership at a glance

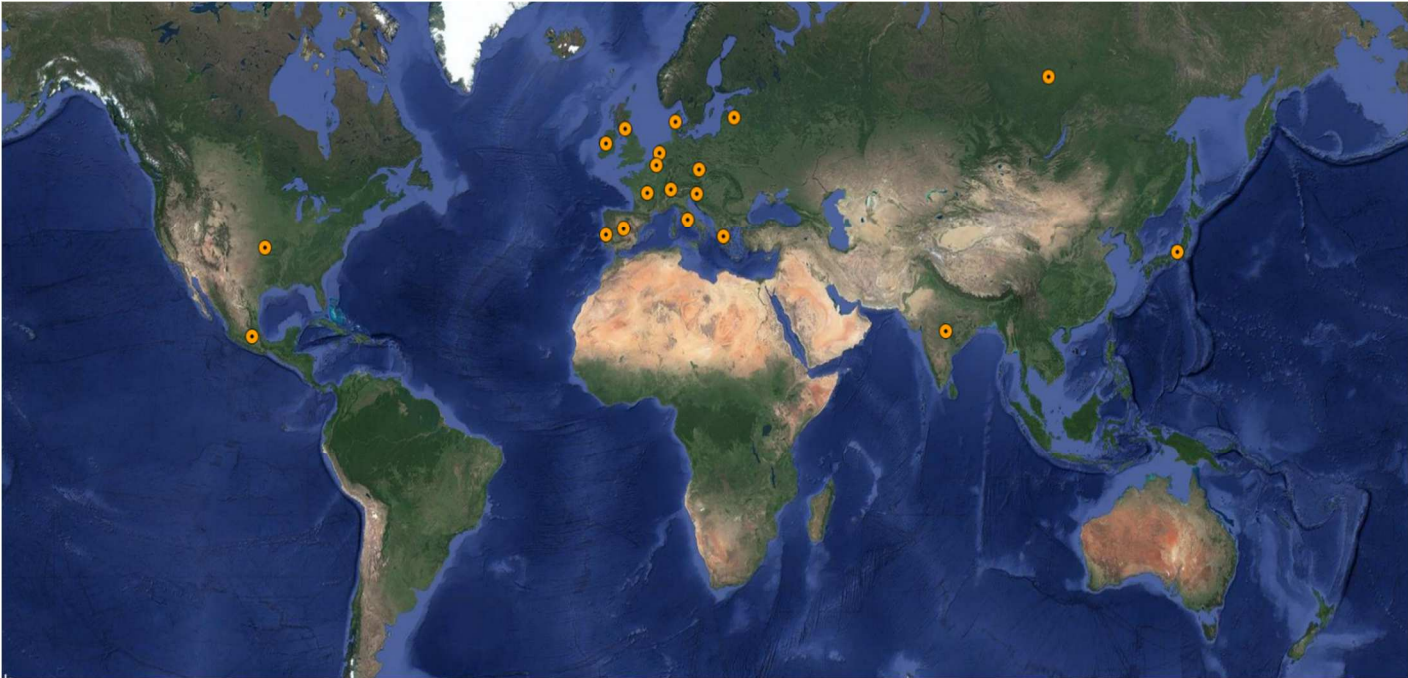


16 new members joined in 2016

33 members have renewed their membership in 2016

6 members joined in 1995 & have renewed membership continuously ever since

Where are our members?



Service to Members

- **Price discounts** on the resources, up to 70% reduction on some items, including the resources produced by ELRA.
- **Discounted fees** for members registering to the biennial conference LREC.
- **Legal and contractual assistance** are also provided to the members of the association
- **Regular information** through the monthly Members' News bulletin and the updates brought to the www.elra.info web site.
- Free online access to **Language Resources and Evaluation (LRE) Journal**, published by Springer, is provided to all ELRA members.

Legal Support HelpDesk

Using, producing, sharing or distributing Language Resources can trigger legal questions related to Intellectual Property Rights (IPR) management which are not always easy to sort out. For nearly 20 years now, ELRA has established close co-operation with legal experts to clear such IPR issues, to design licensing schemas and draft licenses, but also to provide assistance on any contractual or legal matter that may arise during the Language Resources life cycle, including acquisition, production, sharing, or distribution phases.

With this Helpdesk service fully dedicated to IPR issues, we are extending our legal support to the whole Human Language Community. This Helpdesk provides services similar to those offered by the META-SHARE network.

Please contact us by sending a message to elra-ipr-helpdesk@elra.info.

ELRA License Wizard

To allow an easy understanding of the various licenses that exist for the use of Language Resources (ELRA's, META-SHARE's, Creative Commons', etc.), ELRA has developed a License Wizard to help the right-holders share/distribute their resources under the appropriate license. It also aims to be exploited by users to better understand the legal obligations that apply in various licensing situations.

The **License Wizard** works as a web configurator that helps the user:

- ♦ select a number of legal features,
- ♦ obtain the user license adapted to their selection,
- ♦ define which user licenses they would like to select in order to distribute their Language Resources,
- ♦ integrate the user license terms into a Distribution Agreement that could be proposed to ELRA or META-SHARE for further distribution through the ELRA Catalogue of Language Resources.



Reach the License Wizard portal and find the license for your Language Resources:
<http://wizard.elra.info>

Support hlt-related events

A number of events have been sponsored by ELRA in order to support and endorse their activities within our field. In 2016, these are:

- ◆ TALN 2016
- ◆ EAMT 2016
- ◆ CLiC-it 2016
- ◆ O-COCOSDA

In addition, ELRA has renewed its subscription to EAMT as an Institutional Member.

Our members are welcome to contact us if they seek such support for events (conferences, workshops) that focus on HLT and preferably on LRs and Evaluation.

ELRA Activities and Projects

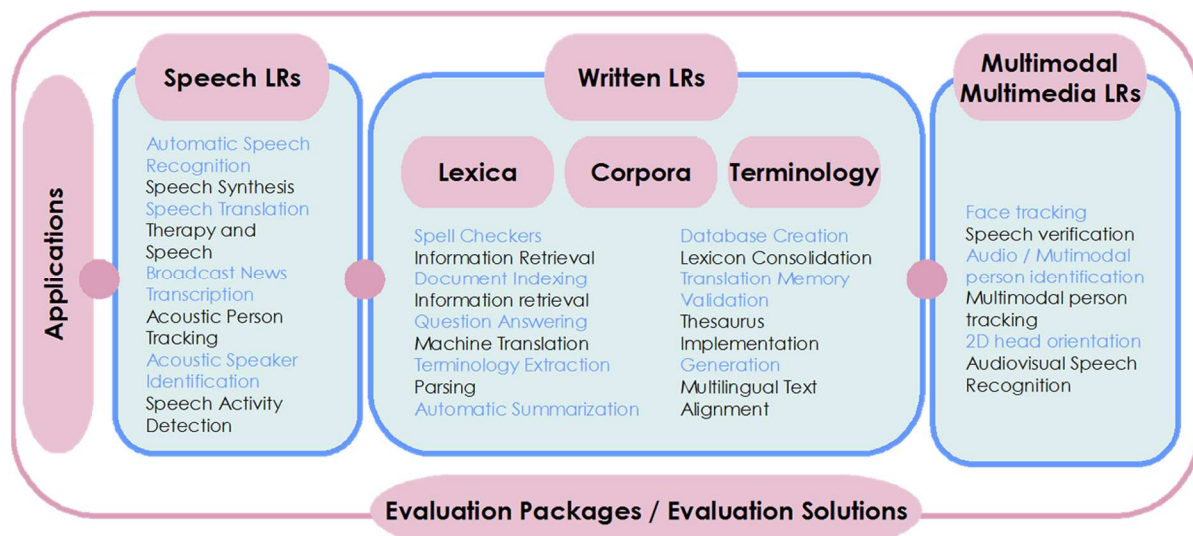
In addition to the promotion of language resources for the Human Language Technology (HLT) sector through market surveys, publication of a regular newsletter and organisation of LREC conference and other workshops, the activities of ELRA are organised around the following services:

- ◆ Identification, collection and distribution of language resources
- ◆ Production of language resources
- ◆ Validation of language resources
- ◆ Evaluation of systems, products, tools, etc., related to language resources
- ◆ Standardisation

The promotion of the language resources production also includes our support of both the infrastructure for evaluation campaigns and the development of a scientific field of language resources and evaluation, e.g. via the LREC conference.

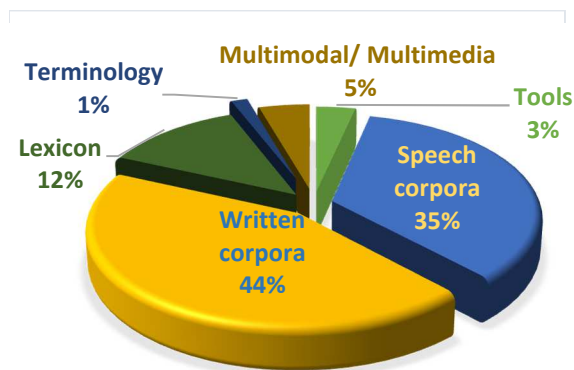
Language Resources Catalogue

The Language Resources collected by ELRA are made available to the public through the ELRA Catalogue that can be accessed online at <http://catalog.elra.info>. The distribution of resources, which cover a wide variety of languages and belong to different modalities, is shown below:



The term **Language Resources** refers to sets of language data and descriptions in machine readable form, used in many types of areas, components, systems, applications, including the creation and evaluation of natural language, speech and multimodal algorithms and systems, the software localisation and language services, the language-enabled information and communication services or the knowledge management; e-commerce, e-publishing, e-learning, e-government.

Universal Catalogue



The Universal Catalogue is an important identification feature. Information regarding Language Resources (LRs) identified all over the world are gathered in this publicly available repository. The LR's are generally located by the ELRA team, but external feedback from our members, collaborators, and web-site visitors is also included. Our aim is to provide researchers and developers information about existing LR's and spare them the effort of searching or rebuilding similar resources. The Universal Catalogue collaborative: through a simple form, anyone can add some basic information, point to an existing language resource or enrich the current description of a LR already present in the Universal Catalogue.

The Universal Catalogue is accessible through the following web site: <http://www.universal.elra.info>.

LRE Map

The initiative for LR & Tools identification, the LRE Map

Initiated by ELRA and FlareNet at LREC 2010, the LRE Map is 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. The feature has been so successful that it has been implemented also at other major conferences including COLING, IJCNLP, Interspeech, LTC, ACLHT, O-COCOSDA, RANLP, in addition to the *LRE Journal*.

The LRE Map feature is now part of the LREC standard submission processes for both the main conference and all the workshops. For LREC 2014, 1070 LR forms have been filled in (to compare with 900 LR forms in 2012). Globally, **5055 LR** resource type forms have been filled in within all the conferences, including LREC 2014, which have adopted the LRE Map.

At LREC 2014, as a new feature, the LRE Map was offering to the authors to share their resources by uploading them during the submission process.

All information on the LRE-Map can be found at <http://www.resourcebook.eu>.

Language Resources: Catalogues & Distribution

2016

Language Resources

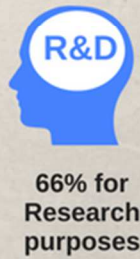
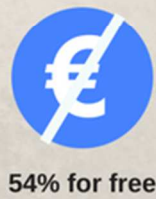
by the numbers



Catalogues Figures



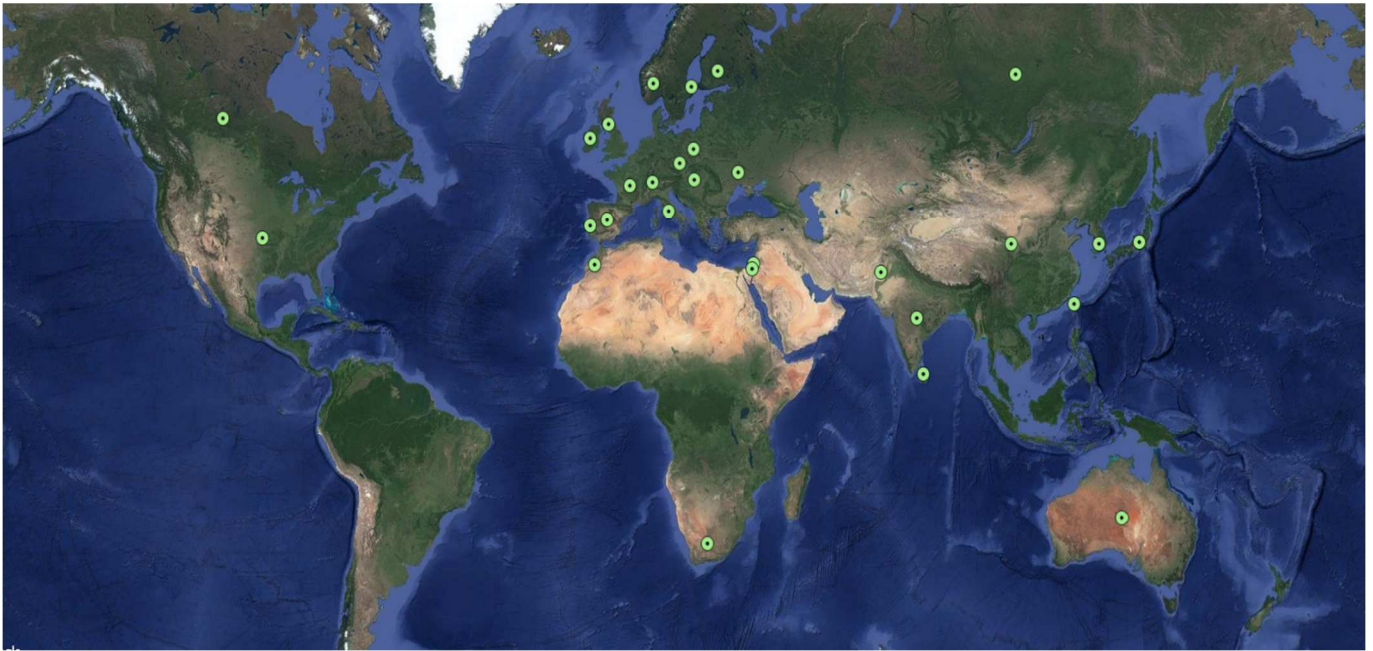
Distribution Figures



New LRs in 2016 & Distribution

Speech Corpus	
Desktop/Microphone Speech	Broadcast Speech Resources
Collins Multilingual database (MLD) – WordBank with audio files <i>ISLRN: 309-438-781-042-2, ELRA ID: ELRA-S0382</i>	TRAD Pashto Broadcast News Speech Corpus <i>ISLRN: 918-508-885-913-7, ELRA ID: ELRA-S0381</i>
Collins Multilingual database (MLD) – PhraseBank with audio files <i>ISLRN: 398-655-047-044-5, ELRA ID: ELRA-S0383</i>	FoxPersonTracks: a Benchmark for Person Re-Identification from TV Broadcast Shows <i>ISLRN: 168-132-570-218-1, ELRA ID: ELRA-S0374</i>
Arabic Speech Corpus <i>ISLRN: 866-568-447-697-8, ELRA ID: ELRA-S0384</i>	Speech related Resources
Serbian emotional speech database (GEES) <i>ISLRN: 462-780-920-598-3, ELRA ID: ELRA-S0385</i>	GlobalPhone Swahili Pronunciation Dictionary <i>ISLRN: 010-360-238-702-2, ELRA ID: ELRA-S0376</i>
GlobalPhone Swahili <i>ISLRN: 200-331-212-512-8, ELRA ID: ELRA-S0375</i>	GlobalPhone Ukrainian Pronunciation Dictionary <i>ISLRN: 022-652-862-222-7, ELRA ID: ELRA-S0378</i>
GlobalPhone Ukrainian <i>ISLRN: 456-398-378-806-1, ELRA ID: ELRA-S0377</i>	Telephone Speech
JV_TDM Corpus <i>ISLRN: 371-240-320-910-4, ELRA ID: ELRA-S0379</i>	SecuVoice <i>ISLRN: 583-080-936-563-9, ELRA ID: ELRA-S0386</i>
Large Farsdat <i>ISLRN: 067-486-870-902-0, ELRA ID: ELRA-S0380</i>	
Written Corpora	
ROMBAC - Romanian balanced corpus <i>ISLRN: 162-192-982-061-0, ELRA ID: ELRA-W0088</i>	TRAD Arabic-English Parallel corpus of transcribed Broadcast News Speech <i>ISLRN: 812-050-111-234-9, ELRA ID: ELRA-W0102</i>
NPChunks <i>ISLRN: 412-883-442-173-8, ELRA ID: ELRA-W0089</i>	TRAD Arabic-French Web domain (blogs) Parallel corpus <i>ISLRN: 138-395-895-757-7, ELRA ID: ELRA-W0103</i>
EUROPARL Corpus Parallel Corpora: Portuguese-English <i>ISLRN: 435-502-922-727-2, ELRA ID: ELRA-W0090</i>	TRAD Arabic-English Web domain (blogs) Parallel corpus <i>ISLRN: 762-161-069-435-5, ELRA ID: ELRA-W0104</i>
Linguatools Webcrawl Parallel Corpus German-English 2015 <i>ISLRN: 800-190-274-236-9, ELRA ID: ELRA-W0091</i>	TRAD Arabic-French Mailing lists Parallel corpus - Test set <i>ISLRN: 895-850-015-188-4, ELRA ID: ELRA-W0105</i>
TRAD Pashto Monolingual text Corpus <i>ISLRN: 394-903-293-388-0, ELRA ID: ELRA-W0092</i>	TRAD Arabic-English Mailing lists Parallel corpus - Test set <i>ISLRN: 858-529-510-480-2, ELRA ID: ELRA-W0106</i>
TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Training data <i>ISLRN: 802-643-297-429-4, ELRA ID: ELRA-W0093</i>	TRAD Arabic-French Mailing lists Parallel corpus - Development set <i>ISLRN: 333-026-450-858-0, ELRA ID: ELRA-W0107</i>
TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Test data <i>ISLRN: 547-897-479-723-3, ELRA ID: ELRA-W0094</i>	TRAD Arabic-English Mailing lists Parallel corpus - Development set <i>ISLRN: 213-044-240-074-6, ELRA ID: ELRA-W0108</i>
TRAD Pashto-English Parallel corpus of transcribed Broadcast News Speech - Test data <i>ISLRN: 006-102-605-738-4, ELRA ID: ELRA-W0095</i>	TRAD Chinese-French Web domain (blogs) Parallel corpus <i>ISLRN: 464-017-697-777-3, ELRA ID: ELRA-W0109</i>
TRAD Pashto-French News Articles Parallel corpus <i>ISLRN: 649-628-149-051-7, ELRA ID: ELRA-W0096</i>	TRAD Chinese-English Web domain (blogs) Parallel corpus <i>ISLRN: 982-341-079-331-4, ELRA ID: ELRA-W0110</i>
TRAD Pashto-English News Articles Parallel corpus <i>ISLRN: 612-936-517-010-2, ELRA ID: ELRA-W0097</i>	TRAD Chinese-French News Articles Parallel corpus <i>ISLRN: 153-566-144-442-2, ELRA ID: ELRA-W0111</i>
TRAD Arabic-French Newspaper Parallel corpus - Test set 1 <i>ISLRN: 922-732-502-473-8, ELRA ID: ELRA-W0098</i>	TRAD Chinese-English News Articles Parallel corpus <i>ISLRN: 626-096-751-907-7, ELRA ID: ELRA-W0112</i>
TRAD Arabic-English Newspaper Parallel corpus - Test set 1 <i>ISLRN: 764-187-795-074-0, ELRA ID: ELRA-W0099</i>	TRAD Chinese-English Email Parallel corpus – Development Set <i>ISLRN: 447-281-370-489-0, ELRA ID: ELRA-W0113</i>
TRAD Arabic-French Newspaper Parallel corpus - Test set 2 <i>ISLRN: 722-323-886-920-3, ELRA ID: ELRA-W0100</i>	TRAD Chinese-French Email Parallel corpus – Development Set <i>ISLRN: 255-358-917-604-3, ELRA ID: ELRA-W0114</i>
TRAD Arabic-French Parallel corpus of transcribed Broadcast News Speech <i>ISLRN: 862-201-329-808-4, ELRA ID: ELRA-W0101</i>	TRAD Chinese-English Email Parallel corpus – Test Set <i>ISLRN: 985-956-234-357-3, ELRA ID: ELRA-W0115</i>
Monolingual Lexica	TRAD Chinese-French Email Parallel corpus – Test Set <i>ISLRN: 239-027-077-538-0, ELRA ID: ELRA-W0116</i>
MCL - Multifunctional Computational Lexicon of Contemporary Portuguese <i>ISLRN: 489-956-642-755-8, ELRA ID: ELRA-L0096</i>	Terminological Resources
LEX-MWE-PT - Word Combination in Portuguese <i>ISLRN: 353-430-176-260-6, ELRA ID: ELRA-L0097</i>	Collins Multilingual database (MLD) – WordBank <i>ISLRN: 990-814-402-335-7, ELRA ID: ELRA-T0376</i>
	Collins Multilingual database (MLD) – PhraseBank <i>ISLRN: 452-383-219-228-0, ELRA ID: ELRA-T0377</i>
LRs now also available for commercial purposes	
Khresmoi manually annotated reference corpus <i>ISLRN: 764-036-829-417-7, ELRA ID: ELRA-W0081</i>	Arabic Speech Corpus <i>ISLRN: 866-568-447-697-8, ELRA ID: ELRA-S0384</i>

Where are our customers?

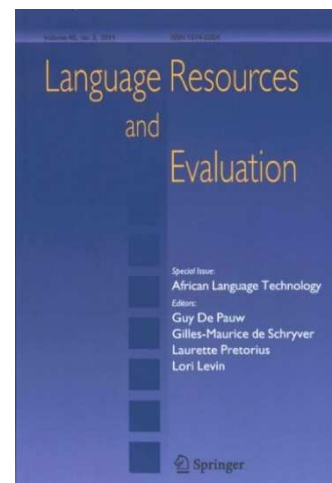


Language Resources and Evaluation Journal

The LRE Journal, published by Springer, is the first publication devoted to the acquisition, creation, annotation, and use of LRs, together with methods for evaluation of resources, technologies, and applications. The ELRA members are granted complimentary access to the journal through the society.

In 2016 the following regular issues, edited by Nicoletta Calzolari and Nancy Ide, have been published:

- Volume 50, n° 1. Special Issue: Computational Semantic Analysis of Language: SemEval-2014 and Beyond
- Volume 50, n° 2. Special Issue: LREC2014: State of the Art in Language Resources and Evaluation
- Volume 50, n° 3. Regular Issue
- Volume 50, n° 4. Regular Issue



For LREC editions, the LREC organisers and the JLRE Editors have agreed with Springer on some special conditions for subscription to the JLRE to be offered to LREC participants.

International Standard Language Resource Number

The **International Standard Language Resource Number (ISLRN)**, a unique and universal identification schema for Language Resources which provides Language Resources with unique identifier using a standardised nomenclature, ensures that Language Resources are correctly identified, and consequently, recognised with proper references for their usage in applications in R&D projects, products evaluation and benchmark as well as in documents and scientific papers.

The ISLRN Portal has been officially opened on April 3rd, 2014. Both ELRA and LDC have started to assign ISLRN to their resources. The moderation is handled by ELRA and LDC. Since the official opening of the ISLRN Portal to December 2016, 2352 numbers have been assigned.

Reach the ISLRN web site and request ISLRN for your own Language Resources @ <http://www.islrn.org>

Some Key Projects



CRACKER is a Coordination and Support Action under the H2020 Programme from the European Commission. Launched beginning 2015, it is coordinated by DFKI.

CRACKER aims at providing planned coordination and support to the European Machine Translation research community, in the context of the the Digital Single Market setup. The principle behind this support action is that "sharing data, results and evaluation instruments" leads to a successful evolution and this collaborative work needs to go through the definition of priorities, planning, organisation, organisation of evaluations and then sharing of the outcome. For that purpose, CRACKER counts on a Consortium of experts covering all these needs and it builds upon established networks, infrastructures and evaluation campaigns. ELDA/ELRA participates in this project providing its expertise in all relevant aspects of the language resource sharing landscape.

More information can be found on the following website: <http://cracker-project.eu/>



The European Language Resource Coordination (ELRC) was launched to collect language resources for building machine translation systems for public service administrations across all EU Member States and Iceland and Norway which will better meet the everyday needs of public services across Europe.

ELRC is an unprecedented public-sector data collection effort aiming to provide CEF.AT with language and translation data (mono- and bi-lingual data) relevant to the daily needs of European national administrations. It aims not only to close the gap between the capabilities of the current MT systems offered by the EC to the national administrations and the actual, day-to-day requirements of national public services across Europe, but also to support Europe's national languages at the grass-roots level.

More information can be found on the following website: <http://lr-coordination.eu/>



The **MLi Support Action** is working to deliver the strategic vision and operational specifications needed for building a comprehensive European MultiLingual data & services Infrastructure, along with a multi-annual plan for its development and deployment, and foster multi-stakeholders alliances ensuring its long term sustainability.

More information can be found on the following website: <http://mli-project.eu>.

Production Projects

The following production projects were achieved in 2016 by the team at ELDA.

- Morphological and Named Entity Annotation in French, English and Arabic
- Sentiment Annotation in French Tweets
- Spoken Language Production in American English

LREC, the Language Resources and Evaluation Conference

The Language Resources and Evaluation Conference is an international scientific event which aims at providing an overview of the state of the art, exploring new R&D directions and emerging trends, exchanging information regarding Language Resources and their applications, carrying out the evaluation of methodologies and tools, on-going and planned activities, industrial uses and needs, requirements coming from the e-society, with respect to both policy issues and technological and organisational ones.

LREC provides a unique forum for researchers, industrial players and funding agencies from across a wide spectrum of areas to discuss problems and opportunities, find new synergies and promote initiatives for international cooperation in support of investigations in language sciences, progress in language technologies and development of corresponding products, services, applications, and standards.

In 16 years LREC which is organized every other year has become the major event on Language Resources and Evaluation for Human Language Technologies. The conference programme is organised around parallel oral and poster sessions during the main conference (3 days) and workshops and tutorials (2 days before and 1 day after the main conference). Since 1998, the LREC conference has been organized every other year with an increasing success.



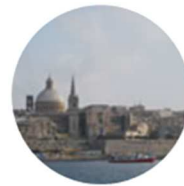
LREC 2016 Portorož
1200+ participants



LREC 2014 Reykjavik
1200+ participants



LREC 2012 Istanbul
1200+ participants



LREC 2010 Malta
1200+ participants



LREC 2008 Marrakech
1100 participants



LREC 2006 Genoa
800 participants



LREC 2004 Lisbon
950 participants



LREC 2002 Las Palmas
730 participants



LREC 2000 Athens
600 participants



LREC 1998 Granada
510 participants

LREC 2016, 11th edition



MAIN CONFERENCE: 9-10-11 May 2018 **WORKSHOPS & TUTORIALS:** 7-8-12 May 2018

<http://www.lrec-conf.org/lrec2016/lrec2018.htm>



@LREC2018

The format of the 11th edition of LREC is the one adopted at the Reykjavik edition: two days for the pre-conference workshops and tutorials and one day for the post-conference.

The format of the Main Conference remains unchanged (on 3 days)

PROGRAMME COMMITTEE

- **Nicoletta Calzolari** – CNR, Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa - Italy (Conference chair)
- **Khalid Choukri** – ELRA, Paris - France
- **Christopher Cieri** – LDC, Philadelphia - USA
- **Thierry Declerck** – DFKI GmbH, Saarbrücken - Germany
- **Koiti Hasida** – The University of Tokyo, Tokyo - Japan
- **Hitoshi Isahara** – Toyohashi University of Technology, Toyohashi - Japan
- **Bente Maegaard** – CST, University of Copenhagen - Denmark
- **Joseph Mariani** – LIMSI-CNRS & IMMI, Orsay - France
- **Asuncion Moreno** – Universitat Politècnica de Catalunya, Barcelona - Spain
- **Jan Odijk** – UIL-OTS, Utrecht - The Netherlands
- **Stelios Piperidis** – Athena Research Center/ILSP, Athens - Greece
- **Takenobu Tokunaga** – Tokyo Institute of Technology, Tokyo - Japan

PROGRAMME

The Scientific Programme will include invited talks, oral presentations, poster and demo presentations, and panels, in addition to a keynote address by the winner of the Antonio Zampolli Prize.

Join ELRA

The annual membership fees are shown in the Fees column.

A Fidelity Programme has been implemented to reward the faithful members. The Miles column displays the number of miles (1 mile=1€) which is allocated to each member joining the association and/or remaining an ELRA member. Miles can be used to purchase resources, pay the membership fees or the registration to LREC.

Important: As part of the revision of the association's membership policy, the ELRA Board has decided that the fidelity program will be abandoned in September 2017. All the remaining miles can be cashed until 1st May 2018.

	Fees (€)
Non-profit making organisations	750
European small/medium-sized companies (< 50 employees)	1000
European profit making organisations (>= 50 employees)	1500
Non-European profit making organisations	5000

<http://www.elra.info/en/join-elra>



ELRA • 9, rue des Cordelières, 75013 Paris • France
Telephone: +33 1 43 13 33 33 • Fax: +33 1 43 13 33 30
info@elda.org • www.elra.info • www.elda.org