2012 annual activity report

Content

Word by the President & Secretary General of ELRA

The Board has been discussing how to draw a new strategy for ELRA for the 5 years to come in order to consolidate ELRA as a key player in the HLT community, in particular in its commitment to anticipate the community's new expectations e.g. free resources or distributed repositories, but also considering its relationship with the European Commission. A new landscape and new trends are perceived through many initiatives popping up within the community, in particular in the framework of the EC funded R&D programs (FP7, CIP, Horizon2020, CEF, etc.). As a consequence, a new structure is being put into operation:

- A new vision of data sharing and distribution, based on META-NET outcomes (repository software e.g. META-SHARE, new metadata schema, new standardised licenses, etc.).
- An ELRA infrastructure committee to renovate the ELRA activities with a strong focus on the ELRA membership structure (i.e. opening up the ELRA membership to new types of members or friends);
- The LRE Map introduced at LREC 2010 and used by other conferences, that would, jointly with the Universal Catalogue, help sharing knowledge on LRs and Evaluations;
- The role we play within the HLT community through LREC, and how to enrich this through new services.
- Last but not least, the set up of task forces to monitor ELRA activities related to the ISLRN, Crowdsourcing, etc.

> **"We continue to vision 2013-2015 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."**

Part of this will continue to be backed up by FlaReNet and the main challenge of FLaReNet follow-up is how to ensure that ELRA can help sustain the FlaReNet assets including the language matrices derived from the LRE Map (to identify gaps) and the contact points (the to-become LRE-Friends), in addition to the setting up and promotion of a sustainability analytic model for LR with the corresponding web portal.

Nicoletta Calzolari, President   **Khalid Choukri, Secretary General**

ELRA Board

Board Officers

Nicoletta Calzolari (ILC-CNR, Italy), President

Nick Campbell (Trinity College, Ireland), Vice-President

Isabel Trancoso (INESC, Portugal), Vice-President

Maria Gavrilidou, (ILSP, Greece), Secretary

Andrejs Vasiljevs, (Tilde, Latvia), Treasurer

Honorary President
Joseph Mariani (IMMI & LIMSI, France)

Board Members

Thierry Declerck, (Language Technology Lab DFKI, Germany)

Marko Grobelnik, (J. Stefan Institute, Slovenia)

Henk van den Heuvel, (CLST, Radboud Universiteit, The Netherlands)

Frédérique Segond, (Objet Direct, France)

Secretary General
**Khalid Choukri**, (ELDA, France)

2012 at a glance

PROMISE Winter School is organized in Italy on 4-8 February.

ELRA holds the 8th edition of **LREC** in Istanbul on 21-27 May.
1225 participants attend.

ELRA releases the 1st set of **Free Language Resources**.

META-SHARE V3 Stable Release

Nicoletta Calzolari is elected **President of ELRA**.

nuary	February	March	April	May	June	July	August	September	October	November	December

*ELRA sponsors **EACL 2012** in Avignon (France).*

*ELRA sponsors **JEP TALN 2012** in Grenoble (France).*

ELRA attends **META-FORUM 2012** in Brussels, on June 20-21.

*ELRA sponsors the 13th edition of **Interspeech** in Portland (Oregon) and attends the conference.*

ELDA attends the **CLEF 2012** Conference in Rome.

ELRA organizes the **Annual General Assembly** in Paris.

*ELRA sponsors **IWSDS 2012** in Ulm (Germany).*

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.

2012 Membership at a glance

Who are our members?

50	total number of members
15	new members joined in 2012
Members having renewed their membership	**35**
6	members joined in 1995
European members	**82%**
50% European non profit organisations	
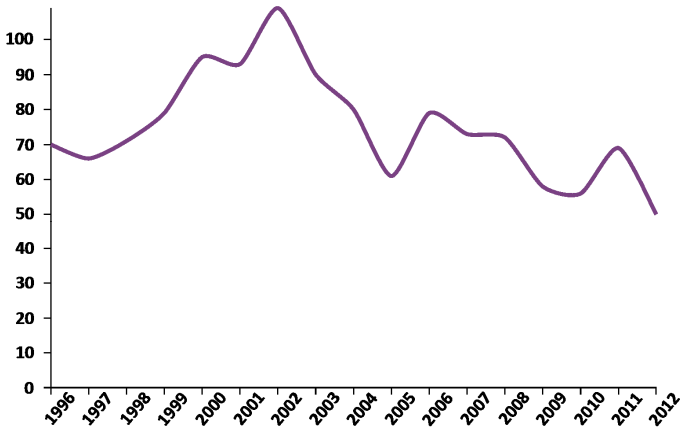
Membership Over the Years

Where are our 2012 members?

Services to members

- **Price discounts** on the resources, up to 70% reduction on some items, including the resources produced by ELRA. Discounted rates are also offered to 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.

- Special prices are proposed to the members for the **Language Resources and Evaluation (LRE) Journal**, published by Springer.

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 over 16 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.

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

Support HLT-related Events

A number of events have been sponsored by ELRA in order to show our commitment to the field. In 2012, these are:

- EACL 2012
- Interspeech 2012
- IWSDS 2012
- JEP-TALN 2012.

In addition, ELRA continues to support the Linguist List.

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.

LRE Friends

The Language Resources and Evaluation Friends (LRE-F) were established at the last LREC in Istanbul. The LRE Friends are individuals interested in Language Resources and Evaluation. They can be scientists, students or professors, involved in HLT research activities in universities, small and medium companies or international groups; decision-makers or project managers in large public institutions, etc.

ELRA Activities

Apart from the promotion of language resources for the Human Language Technology (HLT) sector through market surveys, publication of a quarterly 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 production of language resources also includes our support of the infrastructure for evaluation campaigns and our support in developing a scientific field of language resources and evaluation, e.g. via the LREC conference.
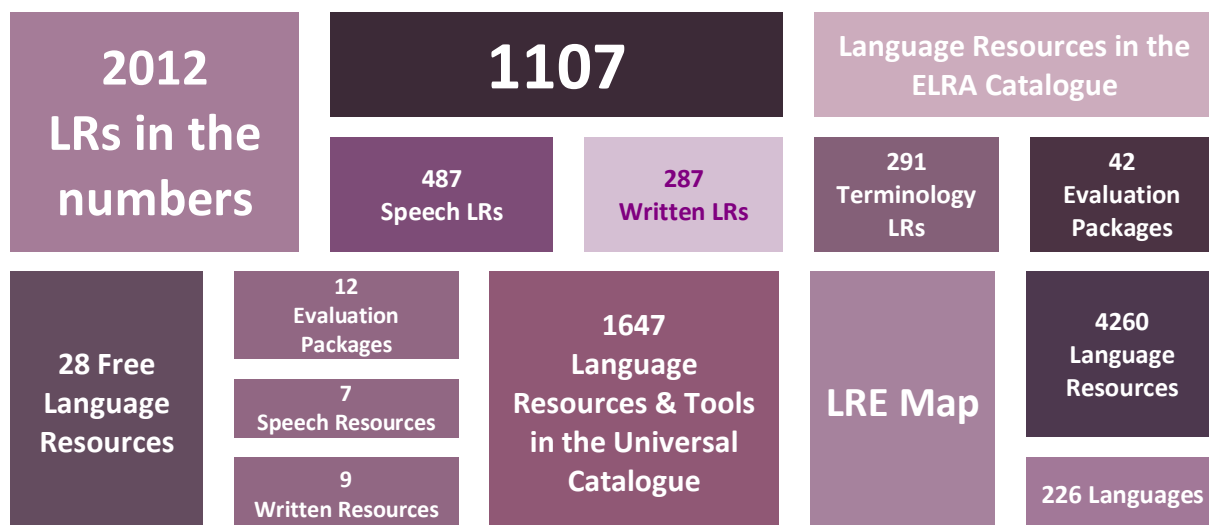
Language Resources Catalogue

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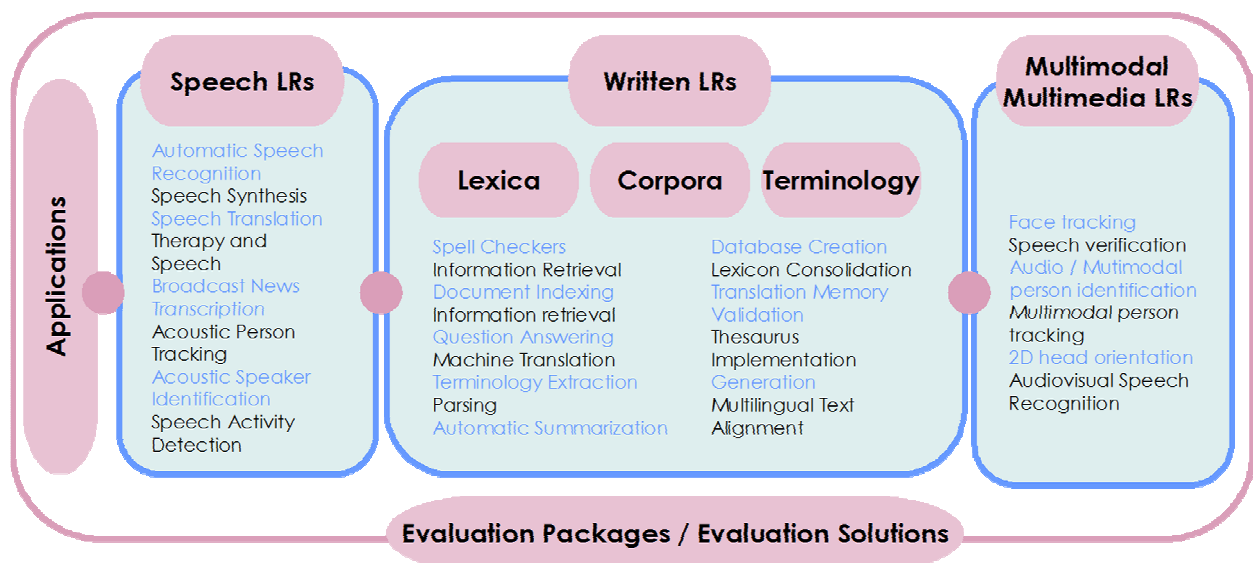
The term **Language Resources** refers to sets of language data and descriptions in machine readable form, used in many types of areas, components, systems, applications, including the creation and evaluation of natural language, speech and multimodal algorithms and systems, the software localisation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

ELRA Catalogues in the numbers

2012 LRs in the numbers

1107

Language Resources in the ELRA Catalogue

487 Speech LRs

287 Written LRs

291 Terminology LRs

42 Evaluation Packages

28 Free Language Resources

12 Evaluation Packages

7 Speech Resources

9 Written Resources

1647 Language Resources & Tools in the Universal Catalogue

LRE Map

4260 Language Resources

226 Languages

New LRs in 2012 & Distribution

The Language Resources collected by ELRA are made available to the public through the ELRA Catalogue and in 2012, a number of new agreements have been signed with providers. They are shown below:

Speech LRs

Speech Desktop/Microphone
- ELRA-S0339 Acoustic database for Polish unit selection speech synthesis
- ELRA-S0342 Acoustic database for Polish concatenative speech synthesis
- ELRA-S0345 Spoken Portuguese Corpus
- ELRA-S0346 Fundamental Portuguese Corpus
- ELRA-S0347 GlobalPhone Hausa

Speech Telephone Resources
- ELRA-S0343 VERIF1DE
- ELRA-S0344 LILA Hindi Belt database

Broadcast Resources
- ELRA-S0338 ESTER 2 Corpus

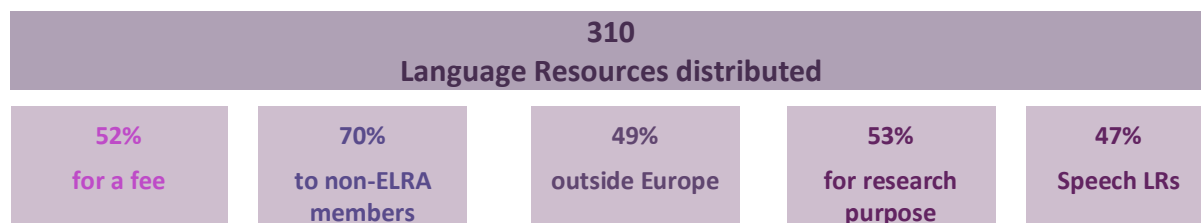### Written LRs & Evaluation Packages

Written Corpora
- ELRA-W0055 CINTIL-TreeBank
- ELRA-W0056 CINTIL-PropBank
- ELRA-W0059 LT Corpus
- ELRA-W0060 PTPARL Corpus
- ELRA-W0061 CINTIL-DependencyBank
- ELRA-W0062 CINTIL-DeepBank

Evaluation Packages
- ELRA-E0035 DEFT'08 Evaluation Package
- ELRA-E0039 CLEF QAST (2007-2009) – Evaluation Package
- ELRA-E0040 MEDAR Evaluation Package

Monolingual Lexicons
- ELRA-L0088 Arabic Morphological Dictionary

310 Language Resources distributed

52% for a fee	70% to non-ELRA members	49% outside Europe	53% for research purpose	47% Speech LRs

Free Resources in 2012

Anticipating users' expectations, ELRA has decided to offer a large number of resources for free for Academic research use. Such an offer will consist of several sets of speech, text and multimodal resources that will be regularly released, for free, as soon as legal aspects are cleared.

Whenever this is permitted by our licences, please feel free to use these resources for deriving new resources and depositing them with the ELRA catalogue for community re-use.

The first set of free Language Resources for the Academic research community is composed of the following :

- **AURORA-CD0002 AURORA Project Database 2.0 - Evaluation Package**
- **AURORA-CD0004-01 AURORA Project Database - Aurora 4a - Evaluation Package**
- **AURORA-CD0004-02 AURORA Project Database - Aurora 4b - Evaluation Package**
- **AURORA-CD0005 AURORA-5**
- **ELRA-E0002 TC-STAR 2005 Evaluation Package - ASR English**
- **ELRA-E0011 TC-STAR 2006 Evaluation Package - ASR English**
- **ELRA-E0025 TC-STAR 2007 Evaluation Package - ASR English**
- **ELRA-S0001 ACCOR – English**
- **ELRA-S0009 COST232**
- **ELRA-S0239 N4 (NATO Native and Non Native) database**
- **ELRA-L0010 MULTEXT Lexicons**
- **ELRA-W0003 CRATER corpus**
- **ELRA-W0033 CRATER 2 Corpus**
- **ELRA-W0023 MLCC Multilingual and Parallel Corpora**
- **ELRA-W0048 TUNA Corpus**

In line with this strategy and spirit, ELRA will continue to negotiate and enter into agreements to offer more free resources. Follow our announcements for this purpose.

**To obtain the free LRs, you should
download the End-User agreement from www.elra.info
and follow the instructions given in the agreement**

Where are our clients in 2012?

Language Resource & Evaluation Journal
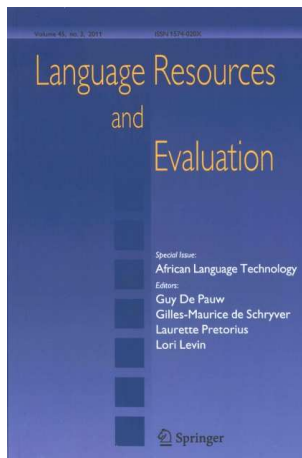
The LRE Journal, published by Springer, is the first publication devoted to the acquisition, creation, annotation, and use of LRs, together with methods for evaluation of resources, technologies, and applications. The ELRA members are granted complimentary access to the journal through the society on the condition that they subscribe to the publisher's table-of-contents alert service.

In 2012, the following regular issues, edited by Nicoletta Calzolari and Nancy Ide, have been published:

- Volume 46, n° 1, Spring 2012
- Volume 46, n° 2, Summer 2012
- Volume 46, n°3, Fall 2012
- Volume 46, n°4, Winter 2012

As for previous 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 2012 participants.

Some Key Projects

Focusing on Medical Information analysis and retrieval, the project KHRESMOI which stands for *Knowledge Helper for Medical and Other Information users* addresses both challenges of trustworthiness and complexity levels in online health information. The project is funded under the EU 7th Framework Programme ICT theme. The consortium, consisting of 12 partners from 9 European countries, will develop a web-based multi-lingual multi-modal search and access system for biomedical information and documents over the next four years. The system will allow querying in several languages, in combination with visual image queries. It will return translated document summaries linked to the original documents.

META-SHARE is one of the objectives within META-NET, it aims to be "a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources." (http://www.meta-net.eu/meta-share). In its aim to achieve such a network, work is being carried out for the specification of a metadata schema which builds upon currently available schema, knowledge and expertise and provides a unified schema capable of handling the requirements of the community. These requirements comprise both the description of Language Resources and that of tools or technologies.

PANACEA, which stands for *Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies* has held its kick-off meeting in Barcelona, Spain, on January 21-22, 2010. The PANACEA consortium consists of 7 partners: Universitat Pompeu Fabra, Spain, Consiglio Nazionale delle Ricerche - Instituto de Linguistica Computazionale, Italy, Institute for Language and Speech Processing, Greece, University of Cambridge, UK, Linguatec, Germany, Dublin City University, Ireland, and ELDA, France. The project is coordinated by Núria Bel, Universitat Pompeu Fabra, Spain.

PROMISE, which stands for *Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation* is a Network of Excellence funded in part by the 7th Framework Program of the European Commission with 10 partners from 7 countries running for 3 years, starting in September 2010.
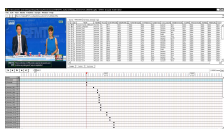
REPERE consists in an evaluation campaign for multimedia information processing systems. More specifically, it deals with person recognition within TV programs. On the organization side, two partners are funded by the DGA (Direction Générale de l'Armement): the LNE (Laboratoire national de métrologie et d'essais) organizes the campaign while ELDA is producing the evaluation data : acquisition of broadcasted programs, audio & text transcription, Named Entities annotation, head localization, head description & identification...

FLaReNet Follow-up

FLaReNet is a project funded by the European Commission under the eContentplus programme, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable. The project is now finished but ELRA agreed to take over the management of the national contact points. In addition, ELRA agreed to work on the idea to host and maintain a repository of best practices and standards

LRE Map

The new initiative for LR & Tools identification, the LRE Map

Initiated by ELRA and FlareNet at LREC 2010, the LRE Map is a new 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. Nearly 2000 language resource forms have been filled in. Apart from providing a portrait of the resources behind the community, of their uses and usability, the LRE Map intends to be a measuring instrument for monitoring the field of language resources. The feature has been so successful that it has been implemented also at COLING 2010 and EMNLP 2010, while other major conferences are in the pipeline, in addition to the *LRE Journal.*

The LRE Map is now part of the LREC submission process which has been slightly updated for the LREC 2012 submission to consider the changes brought to the 2012 version of the Map. Further to the very high number of submissions to LREC 2012 (1027 for both oral and poster papers), the number of forms in the map has risen accordingly to reach 4260, with more than 3000 resource types covering 226 languages. More data will be compiled, cleaned, and reported on in 2012, better incorporating LREC 2012 expected input.

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

LREC, the Language Resources and Evaluation Conference
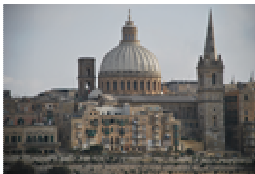
The Language Resources and Evaluation Conference is an international scientific event which aims at providing an overview of the state of the art, exploring new R&D directions and emerging trends, exchanging information regarding Language Resources and their applications, carrying out the evaluation 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 14 years LREC which is organized every other year has become the major event on Language Resources and Evaluation for Human Language Technologies.

The conference covers a full week and LREC 2012's programme will be organised around parallel oral and poster sessions during the main conference; 4 days before and after the conference will be dedicated to workshops and tutorials.

Since 1998, the LREC conference has been organized every other year with an increasing success. The following is a brief overview of all the places that have welcomed LREC:

8th	**LREC 2012** Istanbul 1200+ participants www.lrec-conf.org/lrec2012		4th	**LREC 2004** Lisbon 950 participants www.lrec-conf.org/lrec2004	
7th	**LREC 2010** Malta 1200+ participants www.lrec-conf.org/lrec2010.		3rd	**LREC 2002** Las Palmas 730 participants www.lrec-conf.org/lrec2002	
6th	**LREC 2008** Marrakech 1100 participants www.lrec-conf.org/lrec2008.		2nd	**LREC 1998** Granada 510 attendees www.lrec-conf.org/lrec1998	
5th	**LREC 2006** Genoa 800 participants www.lrec-conf.org/lrec2006.		1st	**LREC 2000** Athens 600 participants www.lrec-conf.org/lrec2000	

LREC 2012, 8th edition

The LREC conference took place in Istanbul, Turkey from May 21 to 27, 2012 under the Patronage of Ms Neelie Kroes, Vice-President of the European Commission and Digital Agenda Commissioner, and Mr Nihat Ergün, Minister of Science, Industry and Technology of the Republic of Turkey. It was a very successful edition that brought together more than 1200 participants from 60 countries, all continents, registered to the conference workshops and tutorials.

- A new feature has been introduced this year: the **Language Library** an experiment of a community-built repository allowing sharing of multidimensional and multi-level processed/annotated resources.

A number of important changes brought to the conference in previous editions have been perpetuated:

- The **LRE Map** feature is now part of the LREC submission process and for the 2012 edition around 900 LR forms have been filled in.
- The **EU Village**, set up to promote EC-funded projects with the support of the European Commission, has attracted a high number of projects including ACCURAT, META-NORD, Bologna Translation Service, CESAR, DIRHA, EASTIN-CL, EU-BRIDGE, FAUST, Khresmoi, LISE, Let's MT, META-NET, METANET4U, MONNET, MultilingualWeb-LT, PANACEA, PortDial, PROMISE, SUMAT, transLectures, TrendMiner, TTC (see http://www.lrec-conf.org/lrec2012/?Projects-EU-Village-2012)..
- The LREC 2012 proceedings were published online before the conference in order for participants to download them from the web site www.lrec-conf.org/lrec2012 before the beginning of the conference.

The **Antonio Zampolli Prize** was awarded to 1) Charles Fillmore and Collin Baker, from the International Computer Science Institute (ICSI), University of California Berkeley (USA), and 2) Oriental Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (Oriental COCOSDA).

The conference received the support of the following companies and institutions CELI, Nuance, EML, IMMI META-NET, Quaero and KDICTIONARIES which have committed to support the conference. In addition, we had arranged displays for a number of publishers, including Springer which offered a special JLRE subscription rate to LREC participants.

Access to the conference information remains available at www.lrec-conf.org/lrec2012.

LREC 2014
Reykjavik

Next edition of LREC will take place in Reykjavik (Iceland) on May 26-31, 2014.
All information can be found at http://lrec2014.lrec-conf.org.

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.

	Fees (€)	Miles
Non-profit making organisations	750	200
European small/medium-sized companies (< 50 employees)	1000	250
European profit making organisations (>= 50 employees)	1500	375
Non-European profit making organisations	5000	1250