The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

January - June 2010

Vol.15 n.1 & 2

LREC 2010 Special Issue

7th International Conference on Language Resources and Evaluation

lrec2010 malta May 17-23

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

ISSN: 1026-8200

ELRA/ELDA
Secretary General: Khalid Choukri
55-57, rue Brillat Savarin
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, this is the special issue of the ELRA newsletter devoted to the Language Resources and Evaluation Conference (LREC). The Seventh edition of the Language Resources and Evaluation Conference took place last May in Valletta (Malta) under the Distinguished Patronage of Dr George Abela, President of Malta and with the support of Herman Van Rompuy, President of the European Council.

This edition has been again very popular: 1246 participants coming from 61 countries registered to the main conference, workshops and tutorials. This time, Germany brought the highest number of participants, but the participation from America and Asia was also significant.

For LREC 2010, three major changes were introduced:

• The *LRE-Map*, 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 existing or under-development language resources and tools, of their uses and usability, the LRE Map intends to be a measuring instrument for monitoring the field of Human Language Technologies.

• The *EC Village*, an initiative to encourage EC-sponsored projects to gain visibility by showing their objectives, progress and activities, either through demos, or through brochures or posters for projects still at their early stages. 15 projects took part in this Village. Both ELRA and LDC were present and took this opportunity to highlight their role within the community.

• The *Special sessions*, a new experiment of oral sessions on "hot topics", with dedicated time slots left for more exchange and discussions between the authors, chairpersons and audience.

A joint COCOSDA/WRITE workshop on *Language Technology Issues for International Cooperation* was held on Saturday, May 22nd 2010. This meeting was organised jointly by COCOSDA, the International Committee for Co-ordination and Standardisation of Speech Databases, WRITE International Committee for Written Resources Infrastructure, Technology, and Evaluation, and FLaReNet. It aimed at reporting on existing initiatives (worldwide) and international cooperation among various initiatives and communities around the world, within and around the field of Language Resources and Technologies.

Six 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 2010, the Antonio Zampolli Prize was awarded to Mark Liberman, from the University of Pennsylvania (USA). We made Mark Liberman's presentation available on-line, from the LREC home page:

www.lrec-conf.org/lrec2010

Now concerning the content of this ELRA newsletter dedicated to LREC 2010, 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.

Stelios Piperidi, President
Khalid Choukri, Secretary General

INTRODUCTION
by Nicoletta Calzolari, LREC 2010 Conference Chair

Let me first express to His Excellency Dr George Abela, President of Malta, the gratitude of the Program Committee and of all LREC participants for his Distinguished Patronage of LREC 2010. We thank him also for having kindly offered us the Presidential summer residence, the beautiful Verdala Palace, for our welcome reception.

This is the 7th LREC. LREC is still young in comparison to other conferences of the field, but after only 12 years we have established a new record: we received 930 submissions and have accepted 662 papers. We have 22 Workshops and 9 Tutorials. These figures are a confirmation that the field of Language Resources and Evaluation is flourishing more than ever. In support of this statement is the fact that also ACL and COLING this year decided to have a specific Track called "Language Resources".

So far more than 1100 people have already registered. LREC thus continues to be - as many say - "the conference where you have to be and where you meet everyone". The high acceptance rate is an important characteristic that gives LREC its special flavour, even more important considering the flaws of the reviewing system of main conferences. But mostly we don't want just the "peaks" (if we can speak about peaks), we want to provide a global vision of the field (the "lay of the land"), with its many features (and for many languages): LREC is therefore the event where we gather a broad representation of the current trends of the field. The preparation of this year program was as usual a challenge, like a big puzzle, but also quite an interesting task giving the perception of the huge range of topics in which our community is engaged. Some of this year main trends:

- Semantics and Knowledge, in all its variations, from annotation of anaphoric, temporal, spatial information, to ontologies and lexicons, to disambiguation, named entities recognition, information extraction, and so on.
- Subjectivity, declined in various nuances: emotions, opinions, sentiments. It becomes almost a workshop track inside the main conference.

Nicoletta Calzolari

- Machine Translation and multilingualism, almost another track.
- Infrastructural initiatives, strategies, national and international projects are of major interest, as usual inside the LREC community.
- Both Lexicons and Corpora keep their strong position.
- Tools and systems for text analysis at many levels.
- Dialogue and discourse, with contributions from both the Speech and Text communities.
- Speech and Multimodal databases, tools, systems.
- And finally evaluation and validation methodologies, as a topic per se and as an important part of quite many papers.

A problem we still have to face is the relatively small number of submissions from the Speech and Multimodal communities. We have always thought that it is important for the field to integrate work from the Written, Spoken and Multimodal areas, and to have a conference where these communities may come together. But the tradition of separate conferences makes this objective difficult to meet. We need to do more in this respect.

It is interesting to notice how, while all the other topics occur in all major conferences (maybe in different proportions), a characteristic of LREC is the relevance of what we can call "meta-research" issues. I think this reflects a peculiarity of the Language Resources and Evaluation community, which has recognised since some time the importance for the field of promoting infrastructural activities. These are considered not so rewarding from the typical researcher viewpoint, but are of extreme importance for the field as a whole. Using an ecological metaphor, the majority of researchers tend to look after their individual tree (individual benefits), but some must take care of the forest (the best for the field)! I think this is a great merit of our community, which is now starting to be recognised also outside. Why this? Probably because language resources and evaluation themselves have an infrastructural nature, they are in need of coopera-

tion, and this reinforces the awareness of the importance of joint efforts and coordinating actions. It is a fact that initiatives towards standardisation, distribution, reusing, integration and now sharing mostly come from this community.

LREC has also introduced innovations in the style of our conferences: posters (previously used mostly in Speech conferences and used at LREC since its first edition, thanks to the insistence of Joseph Mariani) are now used in every conference. Many colleagues even like poster sessions more than oral sessions.

This year we introduce other innovations. We make an experiment of two "special sessions" on hot topics - Temporal and spatial annotation, and Emotion and sentiments (this is even a special track) - where the last slot is left for discussion, to be introduced, stimulated and moderated by the chair. If successful we may have more special sessions in the next edition.

The LREC Map

This year's biggest innovation is the LREC Map, which finds its proper place inside the infrastructural actions. I was thinking about this already two years ago, but it was too late to introduce it. Somehow it is even better to start now because of the coincidence of new important initiatives around Language Resources in the last years and of new windows of opportunities. What is the Map? As all the authors know, we have asked to provide some basic information about all the resources (in a broad sense, i.e. also tools, standards, evaluation packages, etc.) - either used or created - described in their papers. Without all the authors' efforts the Map wouldn't exist, but now they gain much more: the whole Map will be available to everyone!

Lack of information and documentation about resources is, in the e-science paradigm, a very critical issue. With the Map we aim not only at starting to fill this gap, but also at encouraging the needed "change in culture". We need in fact to make each and every researcher aware of the importance of his/her personal engagement in documenting resources. It is a task as important as creating new resources and is not an accessory to be disregarded. It is the necessary service to the whole community that we all have to provide (again, the forest and not only the tree!).

Even if it is a simple idea, it has a great potential for many uses, in addition to being an information gathering tool:

- It may become a great instrument for monitoring the evolution of the field (useful also for funders), if applied in different contexts and times.
- It can be seen as a huge joint effort, the beginning of an even larger cooperative action not just among a few leaders but among all the researchers.
- It can thus become also an "educational" means towards the broad acknowledgment of the need of meta-research activities with the active involvement of many.
- It is also instrumental in introducing the new notion of "citation of resources" that could provide an award and a means of scholarly recognition for researchers engaged in resource creation.

I am proud of the fact that the Map is already being adopted in other contexts. The *Language Resources and Evaluation* Journal will also make use of it, as announced in the new Special Issue *LREC 2008: Selected papers* (freely accessible to all LREC participants). And quite recently also COLING 2010 has decided to make use of the Map. The Map is taking its first steps outside home!

Acknowledgments

The final section is dedicated, as usual, to the acknowledgments to all those who made this LREC possible and hopefully successful.

I thank first the Program Committee, and recognise their dedication in the huge task of selecting the papers, but also in the various aspects around LREC. Among them a particular thanks goes to Jan Odijk and Joseph Mariani, who have been very helpful in the preparation of the program. And to Khalid Choukri, who is in charge of so many organisational aspects around LREC.

I thank ELRA, which is the promoter of LREC, its own conference.

I am particularly grateful to Mr. Herman Van Rompuy, President of the European Council, for his support to LREC2010 and his interest in multilingualism, a great topic in LREC.

I thank our impressively large Scientific Committee, composed of 537 colleagues. They did a wonderful job, succeeding to complete their reviews on time for so many papers. Also on behalf of the Program Committee, I warmly thank all the various Committees that have provided support to this LREC, the International Advisory Board and the Local Advisory Board.

A particular thanks goes to the Local Organising Committee, and especially to Mike Rosner and Claudia Borg, who have worked very hard for many months to find the best solutions to many local issues.

I am grateful to authorities, associations, organisations, committees, agencies and companies that have supported LREC in various ways, for their important cooperation.

I express my gratitude to all the sponsors that have believed in the importance of our conference, and have helped with financial support.

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

A big thanks goes to all the 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 provided economic support and dedicated so much effort, in terms of manpower, to this LREC, as to the previous ones, i.e. ELDA in Paris and my group at ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least thanks, are thus devoted, with all my sympathy, to the people of these institutions who have worked so intensely to make this LREC possible in all its details: Roberto Bartolini, Olivier Hamon, Vincenzo Parrinelli, Valeria Quochi, Sergio Rossi, and in particular Irene Russo (substituting this time Sara Goggi in maternity leave) and Hélène Mazo, a pillar of LREC. I can say for sure that without their daily work and real commitment for many months, LREC would not have happened. We have solved together the many big and small problems of such a large conference. They, together with other young researchers of these two institutions and

with a number of volunteers, will assist you during the conference.

Now LREC is in your hands, the participants. You are the protagonist of LREC, you will make this LREC great. So, at the very end, my greatest thanks go to you all. I may not be able to speak with each one of you during the Conference (I'll try). I hope that you learn something, that you perceive and touch the excitement, fervour 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.

I particularly hope that funding agencies all over the world will be impressed by the quality and quantity of the initiatives in our sector that LREC displays, and by the fact that the field attracts practically all the best groups of R&D from all continents. This is a sign they must take into account in their programmes and funding strategies, and I recognise that they have started to do so. The success of LREC for us actually means the success of the field of Language Resources and Evaluation.

The tradition of holding LREC in wonderful locations with a Mediterranean flavour continues, and Malta is a prototype of the typical LREC location! I'm sure you will enjoy Malta during the LREC week. And I hope that Malta will enjoy the invasion of LRECers!

With all the Programme Committee, I welcome you at LREC 2008 in such a wonderful country as Malta and wish you a fruitful Conference.

Enjoy LREC 2010 in Malta!

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

LREC 2010 Opening Ceremony Speeches

Message from Stelios Piperidis, the ELRA President

Stelios Piperidis

Let me first express, on behalf of the ELRA Board and Members, our profound gratitude to His Excellency Dr George Abela, President of Malta, for his Distinguished Patronage of LREC 2010 and for honoring us with his presence. Our gratitude is also extended to His Excellency M. Herman Van Rompuy, President of the European Council for the importance that the EC confer to multilingualism, one of the cornerstones of our conference.

This 7th edition of LREC coincides with ELRA's fifteenth anniversary. ELRA was established in 1995, in a setting where data-driven techniques had just started rising, numerical and learning methods had not prevailed in the language technology field. It was the work and strategic thought of a few visionary people to set up an association that would take care of identifying, archiving, and distributing language data, as well as cater for language technology evaluation. Initial goals were later on extended into production and validation of language data, production of technology evaluation packages, as well as into offerings of specific services (upon demand) and organisation of specific events, LREC being the major such event since 1998. And every LREC edition seems to outperform the previous one in a range of dimensions, most notably in numbers of submissions and participants.

Today, we live in a totally different setting. No need to stress the importance of data any longer. Data Intensive Science and Research is the new paradigm and Digital Data Deluge a new term that challenges the minds of many of us. Infrastructures, detailed machine-readable metadata and documentation of language data and tools, data and metadata exchange and harvesting, processing tools in the form of web services, static or dynamic workflows were just unknown terms 10 years ago in our field. Yet, today, they are part of our daily vocabulary and appear high in our research and development agendas, as you will notice in the next three days.

Throughout this period, from 1995 to 2010, ELRA has achieved to establish itself as a high-quality data centre, as a main player in the field of language resources and evaluation, while it also enjoys a steady base of membership. In times changing at a dazzling speed, ELRA is constantly responding to new challenges and adapting to new cultures and trends. ELRA has recast its definition of resources by extending it to encompass tools that are, by contemporary methods, used in the LR production, validation, and evaluation processes. As announced at a previous LREC, besides its ever evolving catalogue of language data with clear distribution rights and licences, ELRA has set up the Universal Catalogue, a bottom-up, com-

munity-built resource radar, a catalogue of resources, data and tools, irrespective of whether these can be acquired through ELDA, ELRA's Distribution Agency, or not. In a similar vein, through its catalogues, ELRA reinforces its cooperation links with major data centres around the world, notably LDC and NICT, trying to unify all component catalogues into a Global Inventory of LRs. Likewise, ELRA's resources are now harvestable by OLAC.

To achieve these goals, ELRA is restructuring its activities along three axes : a) *infrastructural*, taking care of identification, cataloguing, updating metadata-based LR description and documentation, and new simpler distribution mechanisms, b) *scientific*, taking care of LRs production & validation, and evaluation, and c) *promotion*, taking care of marketing and promotion activities like LREC and information aggregation services.

Along the infrastructural line, LREC and ELRA are proud of having initiated, implemented and supported the LREC 2010 Map. As will be discussed in various other opportunities during the conference, this new concept is yet another concerted effort towards a new culture, it constitutes a very rich source of information, that I am pretty certain will be re-used many times, and functions as a unique tool for monitoring progress, identifying specific gaps, highlighting trends. The LREC 2010 Map is a truly community-built information base, and we are deeply indebted to all of you for embracing the idea and making it happen. The Map is the common achievement of all of us, our first shared meta-resource.

ELRA, through its operational body ELDA, is also proud of being a founding member of the newly launched META-NET, a Network of Excellence dedicated to fostering the technological foundations of the European multilingual information society. In the framework of META-NET, by its nature and objectives, ELRA will play an important role in META-SHARE, an open, integrated, secured and interoperable exchange facility for language data and tools for the Human Language Technologies domain and all other applicative domains (e.g., digital libraries, cognitive systems, robotics, etc), where language plays a critical role.

[…]

I would like to take the opportunity to thank all those who have worked so hard to make this conference a big success; the LREC Programme Committee, chaired by Nicoletta Calzolari, the Scientific Committee, the International Advisory Committee, Nicoletta's group in Pisa, Khalid Choukri and the ELDA staff in Paris. Particular thanks go to Mike Rosner and his Local Organising Committee for taking care of the myriad of issues that pop up when undertaking the organisation of such a complex and demanding conference as LREC, a Herculean task.

Dear LREC Participants,

The Conference is now in your hands. With your active participation in the oral sessions, your lively discussions with the presenters at the poster sessions, your visits to the EC Village to discuss with the investigators of the research projects funded by the European Commission, LREC 2010 will be yet another success.

I welcome you all to LREC 2010 in Malta and wish you a fruitful conference.

Stelios Piperidis
ILSP / R.C. "Athena"
Epidavrou & Artemidos 6
Marousi 151 25 Athens, Greece
spip@ilsp.gr

Message from Khalid Choukri, ELRA Secretary General and ELDA Managing Director

Welcome to Malta and LREC'2010; Merhba lil Malta u LREC'2010;

Khalid Choukri

Welcome to the celebration of ELRA 15th anniversary; your presence here is the best gift that friends can bring for such a wonderful event; thank you all for your contribution to make this happen;

But let me first express, on behalf of the ELDA team, our profound gratitude to His Excellency Dr George Abela, President of Malta, for his Distinguished Patronage of LREC2010 and for honoring us with his presence. Our gratitude is also extended to His Excellency M. Herman Van Rompuy, President of the European Council for the importance that the EC confer to multilingualism, one of the cornerstones of our conference.

The patronage of His Excellency Dr George Abela, President of Malta highlights the importance his Excellence attributes to this field and the encouragements of his Excellence to the diversity of our topics and I know that this is also a true commitment to see the languages reinforce the foundations of intercultural comprehensions and diversities in Europe and beyond. So in the name of ELRA and of the LREC Program Committee, and I am sure I can speak on their behalf, let me express to his Excellence our gratitude for his support and extend this gratitude to his Excellency, president Van Rompuy, assuring him with our dedication to help with the huge challenge of Multilingualism on his agenda.

I am particularly proud to welcome you in Malta where many of our visions are shared by the country highest authorities but also where our main concerns are explicitly imprinted everywhere. Maltese (Malti) is the only Semitic language that is an official language of the European Union, it has all the roots and grammar paradigms of a Semitic language while it borrowed from

Italian, Sicilian, and more recently from English a large amount of its vocabulary and uses Latin alphabet for its writing. The country population lives the multilingualism at a very high degree, mostly bilingual (Maltese, English), the population has a good command of Italian, a nice scenario to mimic at other corners of the European Union and beyond.

We are happy that this 7th LREC is taking place in that context; ELRA and its partners are very proud to continue the organisation of such a major event where one can meet all active players from both academia and industry. Going back to the 1998 vision initiated by Joseph Mariani and Antonio Zampolli for the need of a scientific event dedicated to LRs and Evaluation, I remember that we welcomed less than 500 participants in Granada (1998) followed by Athens (2000), Las Palmas (2002), Lisbon (2004), Genoa (2006) and Marrakech (2008) with increasing interest from the HLT community. We, the organizing team, felt that we reached the highest target in terms of number of participants, papers and submissions, workshops but we have broken a new record with more than 1100 registered participants a month before the Conference, about 9 tutorials and 24 workshops.

This stresses how dynamic our field is but also how vital it is in particular for a continent like Europe attempting to establish a Multilingual Union that fundamentally requires HLT to help overcoming the language barriers and bringing a crucial supporting tool for the Challenge of Multilingualism. ELRA has committed to such a vision from that start and will continue being supportive to all activities and initiatives that help making LRs available to the R&D communities and also providing the necessary resources and the infrastructures for the evaluation of technologies.

As you know ELRA is driven by its members and enjoyed over the last 15 years the membership of more than 70 active players in average, each year. Gathering such members, that enjoy all ELRA membership benefits, helped also ELRA to design and tune its strategy and orientation. I am particularly happy to welcome the 150 participants to this edition and coming from institutions, members of ELRA.

To offer useful services to its members meant for ELRA to devote human resources to the identification of useful LRs and making them available through its LRs catalogue. From a "centralised" archiving house, as was initially established, ELRA shifted its operations towards brokerage of LRs with strong focus on distribution and licensing. This required, in addition to the negotiation of distribution rights with providers, the drafting of reliable and understandable licences and licensing schemas, ensuring that use (and re-use) of LRs is not hindered by legal barriers. With the support of legal experts, the ELRA team acquired the necessary legal knowledge to handle the basic legal issues involved.

Over the 15 years of activities, more than 1000 resources have been catalogued and made available, thanks to over 250 distribution agreements (with providers from all over the world); ELRA distributed over 3500 resources for HLT development (respectively 48% for research purposes by academia, 37% for Research and technology development by industry, and 16% for evaluation of technologies), not to mention an additional 1500 copies distributed within evaluation campaigns. The availability of such resources within easy, trustable, and fair legal and exchange frameworks boosted the development of Human Language Technologies and the deployment of applications.

In addition to the over 1000 items available off the shelf from ELRA catalogue, ELRA has also continued its work on identifying existing LRs and collected useful information on over 1700 resources that constitute the ELRA Universal Catalogue, an antechamber to the ELRA catalogue and the ancestor of the LREC Map.

A number of HLT evaluation campaigns have been initiated, others have been strongly supported ensuring that the community get a clear picture of the state of the art but also ensuring that all used resources are assembled afterwards and offered to the community as "Evaluation Packages", with all required datasets, tools, methodologies, to conduct similar experiments. More than 20 technologies have been evaluated with that contexts and over 40 Packages are available within the ELRA Catalogue. In order to ensure that the community has access to all available information on Evaluation initiatives, ELRA has set up a portal collecting and compiling all sorts of information related to evaluation with quick and easy reference about protocols, metrics, tasks, resources, projects, campaigns, etc. (http://www.hlt-evaluation.org).

ELRA has also conducted the production of LRs, either on its internal funds or commissioned by partners. Many of these productions took place in the framework of European and international projects. So far ELRA compiled LRs in more than 25 languages, with processes that ensure high quality. Such LRs target different technologies. Some of the recent achievements covered audio/speech data for a variety of languages (e.g. Hindi, Korean, Colloquial Arabic(s), Canadian French, US Spanish, etc.), Broadcast News Speech Corpus for Arabic, French, Spanish, etc. Corpora for languages such as Turkish, Romanian, Kazak, Catalan, etc. aligned textual corpora for Machine Translation (several languages), video annotations with audio transcriptions, etc. etc.

Most of these resources are intended to be part of the ELRA Catalogue, even if, when commissioned, all the rights including ownership remain with the commissioning party.

Last but not least, ELRA continues its work on providing useful forums for the HLT community to debate on current situations, new trends, and shared visions for the future. In addition to LREC, established as a major milestone within our community, ELRA continues to support the organization of the Langtech events (the European exhibition on language technologies, previous one in Rome 2008, next one very likely in 2011), the organization of what is now known as the "European Language Resources and Technologies Forum" (held in the framework of the EC funded project FLaReNet in Vienna, 2009 and Barcelona, 2010), the organization of specialized events (MEDAR conferences on Arabic language, workshops on Less-Resourced Languages (last one in Poznan, in conjunction with LTC'09), etc. Such events allow the community to gather and discuss the hot topics and ensure a useful stream of information dissemination and awareness.

Now that many of the original core goals have been achieved, others expanded, and many activities are running smoothly, ELRA is moving to a new stage. After an

establishment and consolidation period, ELRA is now well equipped to tackle the new challenges emerging from the new trends within our community.

ELRA is reshaping its visions and missions throughout a number of infrastructural changes that will offer our community new services centred on LRs and Evaluation, taking into account new scenarios of LRs "consumption behaviours", while keeping in mind the rationales behind its establishment.

The core activity centred on identification of Language Resources, negotiation of distribution rights and cataloguing such LRs on our online catalogue will remain as our backbone. The modus operandi will strongly take into consideration new mechanisms of sharing and distributing LRs, combining all Web 2.0 features and inspiring both from "open source" principles and professional Business-to-Business approaches.

Part of this vision will be driven by the strong implication of ELRA, through ELDA, in some major European and international initiatives. An international initiative is ongoing to secure the largest Global catalogue of LRs, partnering with LDC, NICT, and others. Others will be driven by a number of EC newly funded projects that have just started e.g. META-NET and PANACEA to name the crucial ones; META-NET, a network of excellence that will design "an open, integrated, secured and interoperable exchange facility for language data and tools for the Human Language Technologies domain" called META-SHARE. Our expectations within META-SHARE is to make more resources sharable through open and integrated Repositories that share similar principles of community servicing and which would benefit from ELRA's 15 years of expertise. The other important project, PANACEA, is addressing issues related to cost-effectiveness of the production of LRs and its automation through adequate web-service based "factories".

Another reason to offer a forum for discussion like LREC is to ensure that needs, trends, expectations, plans, partnerships, co-operations can be discussed and brought one step forward. Given the number of initiatives, projects, committees, that address some of the issues of paramount importance to our community, it is becoming essential to harmonize and coordinate the activities of international organizations. I know that many of us share this view and I would like to take this opportunity to issue a public invitation to coordinate our efforts, through an "informal" umbrella organization, to all interested organizations (ACL, COLING, IAMT, ISCA, EAFT, LDC, GSK, AFNLP, COCOSDA/ORIENTAL-COCOSDA, etc.). LREC'2010 could be the right place to start such an initiative.

If you would like to learn more about ELRA and ELDA, the ELRA/ELDA staff is available during the conference. You will also find more information on our web sites, at www.elra.info and www.elda.org.

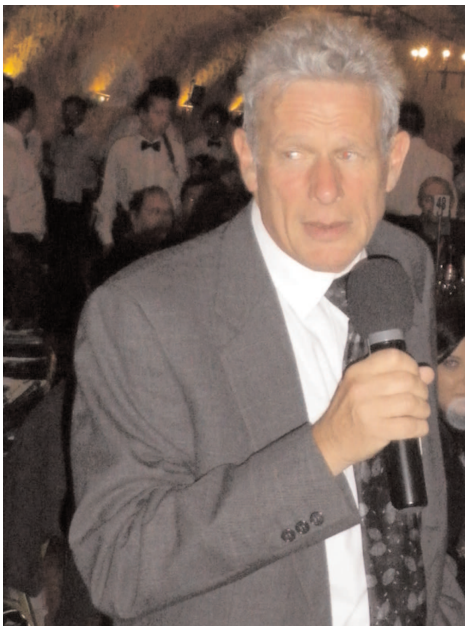
On behalf of ELRA and ELDA, and on your behalf, I would like to warmly thank the local team in Malta as well as the Maltese advisory committee.

ELRA and the Organizing Committee have made every effort to ensure the success of LREC 2010, making, both the scientific part and the social events fruitful and enjoyable. Suggestions to improve any aspects of the conference are welcome, and if you need any assistance to make of this event a more memorable one, please do not hesitate to contact our staff, who will be very pleased to help.

Once again, welcome to Malta, welcome to LREC 2010.

Khalid Choukri
ELRA / ELDA
55/57, rue Brillat Savarin
75013 Paris, France
choukri@elda.org

Message from Mike Rosner, Chair of the Local Organizing Committee



Mike Rosner

Dear LREC 2010 Participants,

On behalf of the Local Advisory Committee, the Local Organising Committee, and all the other participating organisations I would first of all like to express my profound gratitude to His Distinguished Excellency Dr. George Abela, President of Malta, for so kindly agreeing to honour the LREC 2010 by His Patronage and for his offer to assist with the training of a sign-language specialist. I also thank the President of the European Union, Mr. Herman Van Rompuy, for the support he has offered.

About ten years ago, the late Antonio Zampolli had asked me if I would be interested in exploring the possibility of organising LREC 2002 in Malta. I explored for about ten seconds. "I am sorry Antonio", I said, "It's impossible. There would be no support for that kind of conference over here. Come back in ten years and we'll see". With an answer like that I thought I was pretty safe. But it was not to be.

As Chairman of the Local Organising Committee, I am now very proud to acknowledge the help we have received from local organisations and individuals that has made this event possible. In particular, I would like to mention generous financial and material support that has been forthcoming from the Ministry of Finance, the University of Malta, the Malta Tourist Authority, and Air Malta.

In addition to these institutions there is a long list of local individuals to whom we

are indebted for the hard work that has turned LREC 2010 from an idea into the reality that you now see unfurling before you. Let us hope it does not unravel completely! I cannot mention them all, but I would like to single out a few in particular: Prof. Juanito Camilleri, Rector of the University, colleagues from the Faculty of Information and Communication Technology, from my own Department of Intelligent Computer Systems, and from the Institute of Linguistics; Rosette Micallef and her team, who run this magnificent Mediterranean Conference Centre, Rachel Borg Cardona and Aleksandra Radulovic, of Gollcher Travel, our local destination management company, Nadine Brincat, of the Malta Tourist Authority, Claudia Borg, my personal assistant throughout, who has had to face the hot fires of disagreement coming from all directions simultaneously on several occasions, and last but not least, the student volunteers who are assisting throughout.

I am extremely pleased to have inaugurated a Local Advisory Committee whose members reflect a broad cross section of those having interests that intersect with those of the LREC community. It is my sincere hope that the activities of this Committee will endure beyond the short week of this Conference and will help to establish a permanent home here for language resources, particularly for the Maltese Language.

Of course, all this local support has been received in the context of a truly European collaboration that has provided us with constant guidance and includes all the other members of the Programme Committee, Khalid Choukri, Helene Mazo and ELRA staff in Paris, and the Pisa team headed by Nicoletta Calzolari.

Let me conclude by welcoming you all to Malta. I sincerely hope you all enjoy your visit, and that the arrangements we have made will serve to make LREC 2010 a pleasant and memorable experience both scientifically and socially.

Michael Rosner
Head Department Intelligent Computer Systems
University of Malta
Msida MSD 2080, Malta
mike.rosner@um.edu.mt

LREC 2010 Antonio Zampolli Prize

Speech given at the Opening Ceremony by Stelios Piperidis

This year, the Antonio Zampolli Prize was awarded to:

Mark Liberman

Trustee Professor of Phonetics in the Department of Linguistics and Professor in the Department of Computer and Information Sciences, University of Pennsylvania.

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 LREC2010 conference, the Prize will be awarded for the fourth 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.

The LREC2010 prize is awarded to a very important linguist. A linguist that has been involved and has been fighting with a range of research areas from phonology and phonetics to gestural, prosodic morphological and syntactic ways of marking focus and their use in discourse, to formal models of linguistic annotation, to information retrieval and information extraction from text. Recently he has been working on the organization of spoken communication in the human brain, especially in relation to the evolutionary substrates for speech and language, and to analogous systems in other animals; he has also worked on agent-based models of language evolution and learning. His research is frequently conducted through computational analyses of linguistic corpora. Recognising the need for centres offering services around language data, he founded the first data centre. To terminate the agony, ladies and gentlemen please welcome this LREC's Antonio Zampolli Prize winner, Professor Mark Liberman.

The presentation given by Mark Liberman, entitled

"The Future of Computational Linguistics: or, What Would Antonio Zampolli Do?",

can be viewed from the LREC 2010 web site:
www.lrec-conf.org/lrec2010

LREC 2010 Oral Session 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 2010 Proceedings, which are available online: http://www.lrec-conf.org/proceedings/lrec2010/index.html

O3 - Dialogue and Evaluation

Sophie Rosset

This session on "Dialogue and Evaluation" was very interesting and offered a large panel of work conducted in this domain. Speaking of evaluation in the dialog systems field leads to different points of view. One can consider evaluating specific modules of a dialog system or evaluating and predicting the user satisfaction, or overall dialogue.

Gordon and Passonneau in their paper "An Evaluation Framework for Natural Language Understanding in Spoken Dialogue Systems" describe experiments for evaluating different approaches for natural language understanding in two different dialog systems. Their proposed evaluation framework addresses different questions such as "is the NLU approach robust to recognizer error?" or "what is the maximum level of noise an NLU can robustly accommodate?" which are important questions when developing NLU modules for spoken dialog system. The paper compares, given different word error rates, the results obtained by two completely different approaches (CFG-based grammar and SVM-based classifiers) in two completely different domains. The results are discussed given the task difficulty and the word error rate.

User satisfaction is one of the most important metrics for measuring the performance of spoken dialogue systems. Different methods have been proposed and most of them rely on objective features among them the speech recognition accuracy. Hara, Kitaoka and Takeda in their paper "Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System" propose to estimate user satisfaction through a N-gram model based on dialog acts sequences (user, system or user+system). The best results concern the simpler classification task (complete or incomplete task). It seems obvious that the history context is useful even for the more complex task that is the dialogue classification between satisfaction level classes (5 levels).

The need to evaluate overall dialogues and not only user satisfaction or specific modules is considered more and more important, specifically when the task or the domain is not well defined. Two papers address this question. The first one ("Dialogues in Context: An Objective User-Oriented Evaluation Approach for Virtual Human Dialogue" by Robinson, Roque and Traum) proposes a descriptive coding scheme with as objective to allow qualitative evaluations of the dialogue quality itself, for example is the kind of answer, silence or utterance, appropriate given the situation. What is interesting in this paper is that the evaluation framework allows dialogue evaluation from a dialogue perspective itself and not from a system perspective.

"Evaluating Human-Machine Conversation for Appropriateness" (Webb, Benyon, Hansen and Mival) is concerned, as the previous one, by the notion of appropriateness and propose to use this notion as a measure of conversation quality. A coding scheme is proposed regarding the appropriateness of the system (7 classes) or user (5 classes) utterances. Using this scheme, the scores offer objective and subjective performance measures which allow to show improvements over the different version of the systems.

The work described in "Constructing the CODA Corpus: A Parallel Corpus of Monologues and Expository Dialogues" by Stoyanchev and Piwek is slightly more different and concern the notion of dialogue itself as opposed to a monologue. The paper presents the construction of a parallel corpus of monologues and expository dialogues where the monologues are paraphrases of the dialogues. The dialogues are annotated with dialogue acts and the monologues with rhetorical structures. Guidelines and a tool for annotation and translation are presented.

Sophie Rosset
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
rosset@limsi.fr

O8 - Sign Language

Eleni Efthimiou

There has been a tradition of successful Sign Language workshops (this year's workshop was the 4th of a series) organised in the framework of LREC, which indicate the steadily raising interest of the HLT community for Sign Language. However, it was Session O8 of LREC 2010, which introduced Sign Language for the first time, among the subjects covered within the main conference. The session included 5 oral presentations, which gave light to a number of Sign Language (SL) related issues. A major aspect in Sign Language research is the existence of language resources including vocabularies, electronic lexicons and appropriately created and annotated corpora. Since SLs are natural languages articulated in the 3D space, simultaneously using multiple channels in order to express the linguistic message, SL corpora are collections of video. SL video-corpora comprise the necessary resource, which allows for the development of language

models of SLs. But the importance of SL corpora becomes evident also in respect to the development of the so-called SL Technologies, from Sign Language Recognition and Generation to Sign Language Machine Translation systems.

The Sign Language session provided the opportunity to a number of national and international research teams and a SL Network to present their work and highlight the points of focus in their research. Discussion included reference to methodologies and practice in creation of SL corpora, annotation procedures and exploitation perspectives of annotated data, the state-of-the-art in SL technologies and how the various SL Technology oriented projects address the currently open issues of Sign Language Processing, Sign Language Recognition, Sign Language Generation and Sign Language Machine Translation, but also various organisational and policy issues relating to corpus maintenance practices, standards for annotation and metadata, informant consensus protocols and IPR issues.

The list of presentations provided a representative picture of work done in Europe towards creation, maintenance and standardisation of SL resources, as well as development of SL technologies. Reference was made to the FP7 SignSpeak and DICTA-SIGN projects, the national projects Creagest and SignCom, and the scope and objectives of the Sign Linguistics Corpora Network (SLCN). Finally, the session brought to light the necessity of provision for interpreting from and into SL, in the case a public event addresses Deaf as well as hearing audience.

Eleni Efthemiou
ILSP / R.C. "Athena"
Epidavrou & Artemidos 6
Marousi 151 25 Athens, Greece
eleni_e@ilsp.gr

O13 - Ontologies

Thierry Declerck

The session O13 on ontologies had a focus on the relation between natural language and semantic resources. How to combine them in applications, how to use one type of resource in order to enrich the other one? Those were broadly the topics discussed by four presentations, in a well filled auditorium. The four papers were:

Inducing Ontologies from Folksonomies using Natural Language Understanding, Marta Tatu and Dan Moldovan

WikiNet: A Very Large Scale Multi-Lingual Concept Network, Vivi Nastase, Michael Strube, Benjamin Boerschinger, Caecilia Zirn and Anas Elghafari

Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia, Gosse Bouma

A Semi-supervised Type-based Classification of Adjectives: Distinguishing Properties and Relations, Matthias Hartung and Anette Frank

All papers have been presented in a good setting, and all the speakers respected the time for presentation, so that enough place for discussion was given. We noticed during this session the increasing importance of partially structured (multilingual) sources like Wikipedia. Advantages and problems of such a "monopolistic" approach to semi-structured sources have also been discussed.

Thierry Declerck
DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
declerck@dfki.de

O23 - Broadcast News

Carmen García-Mateo

In this session four papers were presented. Each paper was presented by one of the authors and get comments and questions by the audience. The papers cover different languages and techniques but all of them try to cope with the problem about how to get high-quality speech resources in the area of Broadcast News. Next, a reduced summary of each paper is presented.

KALAKA: A TV Broadcast Speech Database for the Evaluation of Language Recognition Systems, Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, Amparo Varona and Mireia Díez.

The process of designing, collecting data and building the training, development and evaluation datasets of KALAKA is described. This speech database was created to support the *Albayzin 2008 Evaluation of Language Recognition Systems*, organized by the Spanish Network on Speech Technologies from May to November 2008. This evaluation involved four target languages: Basque, Catalan, Galician and Spanish (official languages in Spain), and included speech signals in other (unknown) languages to allow open-set verification trials. Also, results attained in the Albayzin 2008 LRE are presented as a means of evaluating the database.

The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet and Jérôme Farinas

This paper presents the EPAC corpus which is composed by a set of 100 hours of conversational speech manually transcribed and by the outputs of automatic tools (automatic segmentation, transcription, POS tagging, etc.) applied on the entire French ESTER 1 audio corpus: this concerns about 1700 hours of audio recordings from radiophonic shows. The EPAC corpus will be useful to researchers who want to work on such data without having to develop and deal with such tools. These

automatic annotations are various: segmentation and speaker diarization, one-best hypotheses from the LIUM automatic speech recognition system with confidence measures, but also word-lattices and confusion networks, named entities, part-of-speech tags, chunks, etc.

Online Temporal Language Model Adaptation for a Thai Broadcast News Transcription System, Kwanchiva Saykham, Ananlada Chotimongkol and Chai Wutiwiwatchai

This paper investigates the effectiveness of online temporal language model adaptation when applied to a Thai broadcast news transcription task. The adaptation scheme works as follows: first an initial language model is trained with broadcast news transcription available during the development period. Then the language model is adapted over time with more recent broadcast news transcription and online news articles available during deployment especially the data from the same time period as the broadcast news speech being recognized. It is found that the data that are closer in time are more similar in terms of perplexity and are more suitable for language model adaptation.

DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain, Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler and Joachim Köhler

This paper reports on the design, construction, and experimental analysis of DiSCo, a German corpus for the evaluation of speech and speaker recognition on challenging material from the broadcast domain. One of the key requirements for the design of this corpus was a good coverage of different types of programmes beyond clean speech and planned speech broadcast news. Corpus annotation encompasses manual segmentation, an orthographic transcription, and labelling with speech mode, dialect, and noise type. Results from ASR, speech search, and speaker recognition on the new corpus are also presented.

Carmen García Mateo
ETSI Telecomunicacion
University of Vigo, Campus Universitario
36310, Vigo, Spain
carmen@gts.tsc.uvigo.es

O25 - Emotion, Sentiment - Special Session

Laurence Devillers

The Special session 025 "Emotion and Sentiments" has consisted of the presentation of 4 papers and of 20 minutes of discussion at the end of the session. This special session has enclosed three sessions declining in various nuances Emotions from speech and Opinions and Sentiments from text:

017 Opinion Mining and Emotions - Paolo Rosso

021 Emotion, sentiment - Inma Hernaez Rioja

025 Emotion, sentiment - Laurence Devillers

During these three sessions, the number of papers was balanced between both speech and text approaches.

- 7 papers were related to Text: Resources for opinion mining (news, movies, product reviews), Sentiment /opinion classifier (negative, positive, neutral).
- 6 papers dealt with Speech (using also sometimes the lexical content): Persuasive power (political dialogs), Spotting problematic dialog situations, Integration of emotional strategies into spoken dialogue management, Cross-corpora study, Cross-cultural difference.

There are some obvious links between emotion, sentiment and opinion.

In opinion/sentiment mining, words can be ambiguous without an emotional context analysis. For example:

"I am afraid in the dark" -> emotion

"I am afraid I don't understand what you have said" -> politeness

When it comes to emotion detection, sentiment and opinion analysis combined with emotion detection is very useful in conversation analysis to detect for example:

- the satisfaction of the customers (Voxfactory project (LIMSI)),
- the probability of the task being completed (WITcHCRafT "a workbench for intelligent exploration of human-computer conversations").

The discussion of the special session focused on the best practices for improving research in emotion, sentiment and opinion, such as:

- sharing the annotations schemes and annotated data
- sharing multi-level annotated resources with emotion, sentiment and opinion annotations
- creating open-source tools/libraries: praat, openEAR, lexical resources
- finding measures for qualifying resources (stat & meta-data)
- using multiple corpora collected in different contexts to train models, combine linguistic and paralinguistic information
- collecting multilingual and multicultural data
- building further benchmarks and evaluations: larger databases, real-life emotional data, challenge including emotion, sentiment and opinion

One of the conclusions drawn from this discussion was the proposal for a common workshop on Corpora for research on Emotion, Sentiment and Opinions at LREC 2012.

Laurence Devillers
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
devil@limsi.fr

O31 - Multimodal Annotation

Jean-Claude Martin

The session on "Multimodal Annotation" included 4 presentations covering several burning issues of the domain: standards for dialog act annotation, dialogue and their multimodal expression, annotation of combined text and image, and polysemy / function of head nods.

The first paper by Harry Bunt et al. was presented by David Traum: Towards an ISO Standard for Dialogue Act Annotation. This paper describes an ISO project which aims at developing a standard for annotating spoken and multimodal dialogue with semantic information concerning the communicative functions of utterances, the kind of semantic content they address, and their relations with what was said and done earlier in the dialogue. The proposed annotation schema distinguishes 9 orthogonal dimensions, allowing each functional segment in dialogue to have a function in each of these dimensions, thus accounting for the multifunctionality that utterances in dialogue often have. A number of core communicative functions is defined in the form of ISO data categories; they are divided into "dimension-specific" functions, which can be used only in a particular dimension, such as Turn Accept in the Turn Management dimension, and "general-purpose" functions, which can be used in any dimension, such as Informant Request. An XML-based annotation language, "DiAML" is defined. More information can be found at: http://semantic-annotation.uvt.nl/.

The second paper was presented by Volha Petukhova and aims at An Integrated Scheme for Semantic Annotation of Multimodal Dialogue Data. It shares some research goals with the first talk. The authors investigate the applicability of existing dialogue act annotation schemes to the semantic annotation of multimodal data, and the way a dialogue act annotation scheme can be extended to cover dialogue phenomena from multiple modalities. The general conclusion of their explorative study is that a multidimensional dialogue act taxonomy is usable for this purpose when some adjustments are made. They proposed a solution for adding these aspects to a dialogue act annotation scheme without changing its set of communicative functions, in the form of qualifiers that can be attached to communicativefunction tags.

The third paper was presented by Pierre Tirrilly about News Image Annotation on a Large Parallel Text-Image Corpus. He presented a multimodal parallel text-image corpus, and proposed an image annotation method that exploits the textual information associated with images. Their corpus contains news articles composed of a text, images and image captions, and is significantly larger than the other news corpora proposed in image annotation papers (27,041 articles and 42,568 captionned images). They use the text of the articles as a textual information source to annotate images, and image captions as a ground truth to evaluate their annotation algorithm. Their annotation method identifies relevant named entities in the texts, and associates them with high-level visual concepts detected in the images (in this paper, faces and logos).

The fourth paper was presented by Francesca D'Errico about Types of Nods: the Polysemy of a Social Signal. The authors analyse the head nod, a down-up movement of the head, as a polysemic-social signal, that is, a signal with a number of different meanings which all share some common semantic element. Based on the analysis of 100 nods drawn from the SSPNet corpus of TV political debates, a typology of nods is presented that distinguishes Speaker's, Interlocutor's and Third Listener's nods, with their subtypes (confirmation, agreement, approval, submission and permission, greeting and thanks, backchannel giving and backchannel request, emphasis, ironic agreement, literal and rhetoric question, and others).

Jean-Claude Martin
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
martin@limsi.fr

O39 - Information Extraction

Martine Adda-Decker

The Session 039 on Information Extraction, which took place the third day in the morning, attracted a relatively large audience (the room was almost filled at the beginning of the session). It was a very interesting session with a nice mix of papers addressing both general and more specific issues: Event Extraction (R. Grishman), Geographical Relation Extraction (A. Blessing et al.) Slovene Definition mining (D. Fi¨er et al.), Entity Mention Detection (S. Biggio et al.). A last paper dealt with English L2 proficiency to evaluate Dutch-speaking users on PubMed searches (K. Vanopstal et al.). The focus of this paper was felt a bit marginal with respect to the overall session scope. The quality of the presentations was globally very high.

The paper by R. Grishman aimed at studying the influence of corpus and task on event extraction systems (their design and evaluation). The author discussed major event extraction evaluations in English (ACE, MUC...) based on various news sources. His claim is that the characteristics of the corpora need to be understood to better understand the event extraction task. The paper of D. Fišer addressed the problem of Definition learning, and the authors argued in favor of starting with structured resources (here Slovene Wikipedia corpus) to acquire knowledge, which can then be used for less structured data. The third paper by Italian colleagues (S. Biggio et al.) reported on research on Entity Mention Detection EMD, which is an extension of Named Entity Detection, to nominal and pronominal mentions. The system was applied among others to annotate the Italian Wikipedia and has been evaluated in the framework of the Italian Evalita 2009. The work presented by colleagues

from Stuttgart University aimed at enhancing context-aware systems by extracting geographical relations on a fine-grained level. The application made use of structured and unstructured data from the German Wikipedia (German towns/ditrict locations). The work by researchers from the Flemish Gent University addressed the problem of information access in scientific English databases (PubMed) by non-native English (Flemish nursing and midwifery students). The double problem of language skills and scientific competence was investigated across two test groups with different levels of education.

Martine Adda-Decker
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
madda@limsi.fr

LREC 2010 Poster Session 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 2010 Proceedings which are available online: http://www.lrec-conf.org/proceedings/lrec2010/index.html

P14 - Word Sense Disambiguation and Evaluation

Olivier Ferret

From a global viewpoint, work presented at the Word Sense Disambiguation (WSD) and Evaluation poster session can be classified into three main categories. The first one according to the number of its articles concerns the problem of building training and evaluation corpora for WSD, which is known to be a difficult and expensive task. A classical way to reduce the cost of building an evaluation corpus in this field is to use the pseudo-word paradigm. This is the issue addressed by (Otrusina & Smrz) and Scheible: the first one focuses on the use of semantic similarity measures to choose words to conflate into pseudo-words for having the most possible realistic evaluation; the second one applies this paradigm to the evaluation of verb clustering. The two other articles in this first category were dedicated to the manual building of evaluation corpora: (Rehbein & Ruppenhofer) explores the use of Active Learning for selecting the most useful new examples for a word sense, claiming that more data is not always useful if these data bring nothing new; (Görög and Vossen) focuses in its case on bootstrapping techniques for annotating a large Dutch corpus for word senses. A second set of articles tackled more directly the WSD issue. (Tsutsumida et al.) and (Okamoto & Ishizaki) both rely on the same resource, the Associative Concept Dictionary (ACD), a dictionary of associations, for performing WSD: the first one uses it for enriching the context vectors associated to each word to disambiguate, which is a to favor the transposition to other domains; the second one turns the ACD into a lexical network and presents an activation propagation procedure on this network for performing WSD. A third article in this set, (Rakko & Constant), investigates a large set of linguistic features to determine their impact on WSD results. Finally, only one article concerned the application of WSD: (Laparra & Rigau) aim at connecting WordNet and FrameNet by disambiguating the Lexical Units of the FrameNet's frames with the synsets of WordNet as reference senses.

Olivier Ferret
CEA - LIST
Centre de Fontenay-aux-Roses
18, rue du Panorama
BP 6 - 92265 Fontenay aux Roses
Cedex - France
olivier.ferret@cea.fr

P27 - Evaluation of Speech Recognition and Speech Synthesis

Olivier Galibert

The "Evaluation of Speech Recognition and Speech Synthesis" session was comprised of a set of three posters around the topic of speech recognition, synthesis and spoken interaction.

Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon by Réveil, Martens and van den Heuvel, deals with the task of large vocabulary proper name recognition. In order to accommodate a wide diversity of possible name pronunciations (due to non-native name origins or speaker tongues) a multilingual acoustic model is combined with a lexicon comprising 3 grapheme-to-phoneme transcriptions built for 3 different languages and 4 variant generators (phoneme-to-phoneme transducers) keyed on the speaker tongue and the linguistic origin of the proper name. The experimental results show that the generated variants can be employed to improve name recognition, and that the obtained accuracy is comparable to what is achieved with transcriptions made by a human expert. They also show that the knowledge of the linguistic origins of both the proper name and the speaker are critical clues to obtain good results.

TTS evaluation campaign with a common Spanish database by Sainz, Navas, Hernáez,

Bonafonte and Campillo, describes the first TTS evaluation campaign designed to compare approaches for Spanish synthesis. Seven research institutions took part in the evaluation campaign and developed a voice from a common speech database provided by the organisation. A common set of sentences were generated and some of the synthesised test audio files were subjectively evaluated via an online test according to the following criteria: similarity to the original voice, naturalness and intelligibility. While the results show that a considerable margin for improvement still exists, two approaches obtain significantly better results, allowing to conclude that some kind of spectral control is needed when building voices with a medium size database for unrestricted domains.

DICIT: Evaluation of a Distant-talking Speech Interface for Television by Sowa, Arisio and Cristoforetti, presents an evaluation of the final prototype of a voice and remote-controlled management interface for an interactive TV. Voice, in that case, includes both natural language and command-and-control-style speech input. The task-oriented evaluation involved naive test persons and consisted of a subjective part with a usability questionnaire and an objective part covering speech component performance, interface design and user awareness, and finally task-based effectiveness and usability. The evaluation revealed a quite positive subjective assessment of the system and reasonable objective results. In addition to pointing specific design-related issues, a usability advantage of voice over remote control has been shown for certain types of tasks.

As a whole, the session presented complementary studies on varied points of the interaction loop. Proper name recognition is essential for most interactive applications, being primary designators of the objects we want the system to manipulate. Speech intelligibility and naturalness is a requirement from the user to continue with a spontaneous interaction. And finally, evaluating in which cases a well-designed speech interface can be more efficient than a remote control helps focus the research on what will be considered useful.

Olivier Galibert
Laboratoire National de métrologie et d'Essais
29 rue Roger Hennequin
78190 Trappes, France
olivier.galibert@lne.fr

Opening Session with Khalid Choukri, Nicoletta Calzolari, Joseph Mariani, Roberto Cencioni, Stelios Piperidis

Republic Hall

LREC 2010 Workshop 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 2010 Proceedings which are available online: http://www.lrec-conf.org/proceedings/lrec2010/index.html

Legal Issues for Sharing Language Resources: Constraints and Best Practices

Valérie Mapelli

http://workshops.elda.org/lislr2010

This half-day workshop took place on Monday 17th May, 2010. It aimed at enlightening the often fuzzy knowledge around legal issues that have to be dealt with at each step of the production and dissemination of a Language Resource. It also aimed at showing new lines of work in that field, as well as new possible cooperation topics. The workshop was a good opportunity for all interested parties to air their views and have an open discussion.

The workshop was organised by Khalid Choukri (ELDA, France), Denise DiPersio (Linguistic Data Consortium, USA), Marc Kupietz (Institut für Deutsche Sprache, Germany) and Valérie Mapelli (ELDA, France) and was supported by FLaReNet (Fostering Language Resources Network - http://www.flarenet.eu).

Before introducing the workshop, Khalid Choukri first presented an overview of the vision of ELRA about Legal Issues with respect to Language Resources throughout ELRA's 15 years of activity.

The workshop was divided into 3 parts, the first one being open to 3 invited talks, the second one to short talks and the last one to a panel discussion on the subject.

- Isabelle Gavanon (FIDAL, working as ELDA consultant, France) gave the first talk, entitled "*Sharing Or Not Sharing? That Is the Legal Question*". She focused on the restrictions of the current legal system in particular in the French law vis-à-vis the specific field of Language Resources, and showed the variety of existing licenses such as Creative Commons or GNU...

- Erik Ketzan (Attorney at Law, USA)'s presentation, entitled "*Translation licensing and copyright issues under United States law*", aimed to raise questions on author right limitations with respect to the production/exchange/distribution of automatically translated works. He particularly focussed on the US "fair use" and "implied license" practices for copyrighted works and the adaptation of this practice with respect to Language Technology. He also briefly showed the difference between US and European laws.

- Prodromos Tsiavos (The London School of Economics and Political Science, United Kingdom) gave a talk entitled "*Extracting value from open licensing arrangements*". This talk aimed at presenting the Creative Commons licenses, showing the various questions that needed to be dealt with when creating the various Creative Commons licensing models.

Then, four short talks were given during the second session:

- John Hendrik Weitzmann (EEAR - European Academy of Law and Computing, Germany), "*Licensing and Sharing Language Resources: An Approach Inspired by Creative Commons and Open Science Data Movements*": He introduced the legal work being carried out within the META-SHARE project which aims at becoming a universal exchange platform for Language Resources and Language Technology. He focussed on the use and adaptation of the Creative Commons license to this field of activity.

- Denise DiPersio (Linguistic Data Consortium - LDC, USA), "*Some Implications of US Initiatives for 'Fair Research*' and Open Access on the Development and Distribution of Language Resources": She presented historical facts concerning recent US initiatives with respect to legal questions linked to the use of data for Research. In particular, she presented the legal framework of the latest Fair Research in Copyright Act and Federal Research Public Access Act.

- Gilles Adda (LIMSI), "*Language resources and Amazon Mechanical Turk: legal, ethical and other issues*": He presented the legal and ethical questions that can be raised from a researcher's point of view with respect to the production of Language Resources through a crowdsourcing system.

- Christina Bankhardt (Institut für Deutsche Sprache, Germany), "*Processing Language Resources on the basis of the Consent - European point of view*": She elaborated on an IPR criterion called "Consent" and which is linked to the processing of Language Resources that include personal data.

At last, Khalid Choukri and Marc Kupietz launched a panel discussion with the invited speakers.

Valérie Mapelli
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
mapelli@elda.org

Emotion 2010 - On Recent Corpora for Research on Emotion and Affect

Björn Schuller and Laurence Devillers

http://emotion-research.net/sigs/speech-sig/emotion-workshop

The 2010 International Workshop on Emotion - Corpora for Research on Emotion and Affect - organized by L. Devillers, B. Schuller, R. Cowie, E. Douglas Cowie and A. Batliner, was the third in a series started in 2006 in association with LREC 2010. Its previous two workshops have helped to consolidate the field, and there generally now is growing experience of not only building databases but also using them to build systems. By that, this workshop aimed to continue the process, and laid emphasis on databases for practical usage.

In the 15 papers accepted for and presented at the workshop, a total of 21 databases were presented covering 7 languages (distributed over English (32%), French (26%), German (22%), Hebrew, Hungarian, Italian and Russian (5%, each)) and two non-speech sets. Considering the type of the spoken content, in 24% of the corpora a fixed such was selected as opposed to 76% of the sets featuring non-constrained speech. The nature of the contained emotion was acted in only 23%, induced in 32%, and natural in the majority of the sets at 45% clearly reflecting the current preference of more realistic display of emotion in resources. However, studio recording still prevails looking at 87% of the sets discussed being recorded in such an environment and only 13% in real-life surroundings (call centre and medical operation room). A prevailing problem also seems to be the accessibility of data looking at 47% being of proprietary nature, 24% available under a license, and 29% being freely available. Looking at the model of emotion, the ratio was 3:2 in favour of categorical versus dimensional, whereby the minimum of categories was 5, and that of dimensions 2. In addition, a 3:1 ratio was given for utilizing one of these schemas exclusively versus using both. The trend behind these figures illustrates the increasing popularity of dimensional or more complex modelling. The according number of annotators in these sets varied from only 1 (in 7% of the sets) to 17 leading to a mean of 4 annotators (2 and 3 to 6 in 37% of cases, each, and more than 10 in 19% of the cases). At the same time, the number of contained individuals reached from 4 to 100, with a total of 530 and a mean of 38, whereby the female to male ratio was at 4:3. In terms of size, the recorded speech time varied between 47 minutes and 22 hours, leading to a total of 69 hours and a mean of 6 hours per corpus resembling a minimum of 120 to 13,731 instances per set and 3,641 on average.

If one judged exclusively from these databases, the 2010 average emotion corpus would be English spoken, verbally non-restricted, natural in terms of emotion, though studio recorded, labelled in 15 classes or 4 dimensions by 4 annotators, contain 38 speakers with more of them being female producing 6h of speech that would result in 3,6k instances. Yet, it would be proprietary and not feature a defined partitioning. Overall, the obvious positive trends by that are that more and more languages are covered; and more natural databases labelled in more complex models seem to be expectable as available to the community. At the same time, increasingly more multimodal resources are to be expected.

Finally, as an agreement of the concluding panel session, further synergies from the presently partly co-existing fields of emotion, opinion, and sentiment research are emerging. Follow-up initiatives could further help foster combined community efforts for merging and common labelling of resources and make such desperately needed larger amounts of resources accessible.

Björn Schuller & Laurence Devillers
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
lrec-emotion@limsi.fr

LR and HLT for Semitic Languages

Khalid Choukri and Bente Maegaard

http://workshops.elda.org/sl2010

The workshop aimed at addressing all issues related to language technologies handling the Semitic language family. The Semitic family includes languages and dialects spoken by a large number of native speakers (around 300 million). Prominent members of this family are Arabic (and its varieties), Hebrew, Amharic, Tigrinya, Aramaic, Maltese and Syriac. Their shared ancestry is apparent through pervasive cognate sharing, a rich and productive pattern-based morphology, and similar syntactic constructions. In addition, several languages which are used in the same geographic area such as Amazigh or Coptic, which, while not Semitic, have common features with Semitic languages, such as borrowed vocabulary.

The Workshop intended to follow on topics of paramount importance for Semitic-languages NLP that were discussed at previous events (LREC, MEDAR/NEMLAR Conferences, the workshops of the ACL Special Interest Group for Semitic languages, etc.) and which are worth revisiting.

The workshop was organized as a one-day workshop consisting of two oral sessions, three poster sessions, and a general discussion on a "Cooperation Roadmap", prepared within the MEDAR project, for building a sustainable Human Language Technologies for the Arabic language within and outside the Arabic world. A large number of papers were assigned to the poster sessions to allow for more interactions and discussions.

The workshop presentations covered many aspects of Arabic Language Processing (from low morphology levels to semantics/wordnet aspects), a few issues regarding Amharic (in particular a dependency grammar), Semitic verbal morphology, an approach to a light morphology of Amazigh (though not a semitic language), an approach to enrichment of the Named Entities in Arabic WordNet, etc.

The first oral session focused on corpus aspects for Arabic including syntax and parsing, semantics/Wordnet, etc. It also featured a paper on issues regarding Amharic (in par-

ticular a dependency grammar). The second one addressed issues related to MT, elaborating on approaches to create parallel and aligned corpora for MT as well as use of such alignment tools to construct bilingual lexica.

The three poster sessions were customised to address specific areas of NLP. The first and second ones focused on topics related to basic levels of language processing e.g. morphology, tagging, language annotation, and their use in applications (QA, search tool, etc.). A paper elaborated on an approach to enrich Wordnet using an ontology. An interesting paper addressed the use of Mechanical Turk to develop resources for Arabic (a corpus of summaries in this case). A session was devoted to speech and related resources and described several collections of Arabic dialects, e.g. Algerian, Maltese, etc.

Two papers focused on Amazighe language spoken in Morocco and the adaptation of existing annotation and tagging tools from semitic languages.

The General Discussion aimed at describing the MEDAR proposal for a Cooperation Roadmap for building sustainable Human Language Technologies for the Arabic language within and outside the Arabic world. The purpose of the discussion was to determine some of the common interests among the participants which could serve as a basis for future collaboration. MEDAR has suggested a cooperation roadmap for Arabic LR and HLT - a short presentation was made - and other roadmaps were mentioned as well.

The discussion was structured along the following themes: tools for reading support, fighting illiteracy through LT, teaching dialects (and focussing on dialects in LRT), standards, repositories for free resources, sharing and collaboration, evaluation, networks for the field. An extensive report on the discussion can be found at the MEDAR website www.medar.info.

Bente Maegaard
Centre for Language Technology
University of Copenhagen
Njalsgade 80
2300 Copenhagen S, Denmark
bente@hum.ku.dk

LREC 2010 Conference Survey Report

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

This year's survey received 302 respondents (over 1246 registered participants) as opposed to 202 in 2008. The survey contained 13 questions related to the Conference Organization and the Conference Content.

The first part of the survey covered the various aspects of the conference organization. The overall results show that a great majority of respondents were satisfied or very satisfied by the conference organization. The online registration is acknowledged as a "good" (52%) even "excellent" (33%) process and the on-site registration process' score is even higher. The conference web site with a 15% "fair" opinion rate has certainly suffered from the late publishing of the conference programme which, according to some participants, have triggered practical issues in their personal organization. Finally, the conference support staff has carried the day with a 90% "good" and "excellent" opinion score. A number of very nice comments on the organization have also been expressed, stressing the smooth sequence of the sessions and within the sessions.

The second part of the survey dealt with the content and structure of the conference. Here again, the impression given by the overall results confirm the satisfaction of most participants. LREC remains the place where the HLT community can gather, allowing contacts networking, meetings and discussions. Nevertheless, if the quality of the LREC 2010 programme is generally acknowledged by the respondents (nearly 70% state that the papers met their expectations), comments show that there is still room for improving each paper's quality and the overall quality of the programme by reaching a homogeneous level for all accepted papers.

This year, for the second time, Poster Sessions have been organized in parallel with Oral Sessions. From the respondents' point of view, it seems that this structure should be kept "as is" for future editions. The current LREC format which was also questioned in the survey appears to be also the preferred one (51%): that is maintaining the satellite workshops and tutorials (2 days) before and after the main conference (3 days). From a content perspective, some respondents have found overlaps in topics between the conference and workshops and are calling for a better selection/distribution of the topics. Other suggested that "workshop hopping" is proposed by implementing a new feeing system that would allow participants to switch between different workshops instead of registering to individual events.

Finally, more or less half of the respondents found that the main conference programme was well-balanced as opposed to nearly one third who considered the programme as too heavy.

For this edition, 3 new initiatives have been introduced and the survey was questioning the participants on those initiatives.

The LREC Map, the new mechanism for monitoring the use and creation of language resources, was found useful by 53% of the respondents and they were almost 60% to assess that they were ready to enter such information in the LREC Map in the future. The implementation of the Map in other conferences was approved by nearly 50% of the respondents.

For the EC-Village as well as the Special Sessions, there was a majority of "No answer" whether "no opinion" or "did not attend" which made difficult to draw clear conclusions. Those respondents who actually answered considered that the EC Village was a "good idea" and that this initiative could become a global (rather than European) HLT projects village in future editions. Those who attended the Special Sessions found them interesting.

Many respondents wrote personalized comments, some very detailed, most of them reflecting positive spirits, congratulating the organizers for the work done and their dedication. These responses are very useful as they are more specific than the survey questions and often convey interesting suggestions.

New Resources

Monolingual Lexicons from the general domain

ELRA-L0086 Persian Multext-East framework lexicon

This is a Persian (Farsi) morphosyntactic lexicon derived from the Persian 1984 corpus (Multext-East framework) (see ELRA-W0054). It contains the full inflectional paradigms of a superset of lemmas that appear in the Persian 1984 corpus. Each entry gives the word-form, its lemma and morphosyntactic description. The lexicon contains 13,247 entries.

	ELRA members	Non-members
For research use	45 Euro	45 Euro
For commercial use	500 Euro	2,000 Euro

ELRA-L0087 Persian lexicon

This is a Persian (Farsi) lexicon of more than 40,000 entries of non-inflected forms of words. Each word is transliterated based on the proposed framework from MBROLA (Text-To-Speech synthesizer). The database includes a large variety of descriptors for each entry (plural, homograph, ...).

This lexicon has been made out from a corpus of newspaper publications collected during a period of six months from the Shargh Newspaper, a publication containing articles from diverse topics: art, culture, policy, social, sport, etc. Due to its coverage, this lexicon can be in particular interesting for Persian TTS systems, as the pronunciation of Persian words cannot be derived directly from their transcription due to the omission of short vowels in Persian writing systems.

The number of records is distributed as follows: 11,955 adjectives, 2,047 adverbs, 164 classifiers, 129 conjunctions, 85 indexes, 36,651 names, 88 numbers, 455 verbs with past stem, 435 verbs with present stem, 223 prepositions, 141 pronouns, 352 semi-sentences.

The lexicon is provided in a MS Access database.

	ELRA members	Non-members
For research use	500 Euro	700 Euro
For commercial use	5,000 Euro	7,000 Euro

Written Corpora from the general and specific domains

ELRA-W0053 Catalan-Spanish Parallel Corpus

This corpus contains more than 100 million words and it contains 10 years of bilingual articles from “El Periódico de Catalunya”. Both language data are rather close as the Catalan text is a translation of the Spanish one, partly achieved by means of Machine translation and then post-edited.

The data are aligned at sentence level and stored in text files, in a one sentence per line basis. The data are provided in plain text, with no encoding whatsoever.

	ELRA members	Non-members
For research use	2,000 Euro	3,000 Euro
For commercial use	20,000 Euro	24,000 Euro

ELRA-W0054 Persian 1984 corpus (Multext-East framework)

This corpus contains the Persian (Farsi) translation of a part of the novel “1984” (G. Orwell) annotated in the Multext-East framework (Multilingual Text Tools and Corpora for Eastern and Central European Languages). The aim of the Multext-East project was to develop standardized language resources.

The package comprises:

(i) the specifications for morphosyntactic encoding of Persian Language, based on the EAGLES/MULTEXT model and specific resources of MULTEXT-East,

(ii) the annotated Persian version of Orwell’s 1984 corpus.

The corpus contains extensive headers and markup for document structure, sentences, and various sub-sentence annotations in the XML-format following the TEI guidelines. Annotation includes POS (part-of-speech) and lemmas. The corpus contains approximately 100,000 words (6,604 sentences, 13,247 lemmas) and can easily be aligned with other corpora in the MULTEXT-East framework.

	ELRA members	Non-members
For research use	45 Euro	100 Euro
For commercial use	2,000 Euro	5,000 Euro

Terminology Resource in Natural Sciences

ELRA-T0374 Terminology database of natural sciences

This dictionary covers the three kingdoms: Animal, Vegetal, Mineral. It contains 50,000 species with numerous synonyms in French, English and Latin and many breeds and varieties. Minerals are given with their chemical formula. About 7,900 definitions in French are included.

This dictionary gathers many disciplines and topics such as: Mammals, Fishes, Birds, Insects, Reptiles, Shellfishes, Trees, Plants, Flowers, Fruits, Vegetables, Minerals, Rocks, Gems, etc. It also includes synonyms and linguistic variants.

Languages : French - English (GB, US) - Latin

Number of entries: 133,500

Number of terms per language: between -20% and -25% approx. with respect to the number of entries (i.e. ca. 50,000 terms)

Disciplines: about 105

Format: .DBF files, sorted alphabetically in French and English

A viewer is also available upon demand for an additional cost of 2676 euros. This software enables a spontaneous search French => English and English => French in the database according to different criteria:
- by beginning of term,
- by included word,
- by discipline,
- by class,
- by kingdom,
- through a free search: French, English, Latin words and synonyms in the 3 languages, chemical formula, i.e. 435,000 access points. For each term obtained, the database corpus is instantly displayed with the total of terminological data available for that term.

Viewing format: .FIC (Windev)

Please note that the prices indicated here are dependent from the number of entries available which is growing constantly. Please contact us for further details.

	ELRA members	Non-members
For research use	53,400 Euro	66,750 Euro
For commercial use	66,750 Euro	80,100 Euro

Desktop/Microphone Resources

ELRA-S0307 BABEL Polish database

The BABEL Polish Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS programme (COPERNICUS Project 1304). The project began in March 1995 and was completed in December 1998. The objective was to create a database of languages of Central and Eastern Europe in parallel to the EUROM1 databases produced by the SAM Project (funded by the ESPRIT programme).

The BABEL consortium included six partners from Central and Eastern Europe (who had the major responsibility of planning and carrying out the recording and labelling) and six from Western Europe (whose role was mainly to advise and in some cases to act as host to BABEL researchers). The five databases collected within the project concern the Bulgarian, Estonian, Hungarian, Polish, and Romanian languages.
The Polish database consists of the basic "common" set which is:
• The Many Talker Set: 30 males, 30 females; each to read 100 numbers, 3 connected passages and 5 “filler” sentences (or 4 passages if no fillers needed).
• The Few Talker Set: 5 males, 5 females, normally selected from the above group: each to read 5 blocks of 100 numbers, 15 passages and 25 filler sentences (or 20 passages if fillers not needed), and 5 lists of syllables.
• The Very Few Talker Set: 1 male, 1 female, selected from many-talker set: 5 blocks of syllables, with and without carrier sentences.

	ELRA members	Non-members
For research use	300 Euro	600 Euro
For commercial use	4,000 Euro	6,000 Euro

Broadcast Resource

ELRA-S0308 Egyptian Arabic Speecon database

The Egyptian Arabic Speecon database is divided into 2 sets:
1) The first set comprises the recordings of 550 adult Egyptian speakers of Modern Standard Arabic as spoken in Egypt (273 males, 277 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2) The second set comprises the recordings of 50 child Egyptian speakers of Modern Standard Arabic as spoken in Egypt (24 boys, 26 girls), recorded over 4 microphone channels in 1 recording environment (children room).

This database is partitioned into 25 DVDs (first set) and 4 DVDs (second set).

The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications. Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items (over 290 items for adults and over 210 items for children):

Calibration data:
- 6 noise recordings
- The "silence word" recording

Free spontaneous items (adults only):
5 minutes (session time) of free spontaneous, rich context items (story telling) (an open number of spontaneous topics out of a set of 30 topics)

17 Elicited spontaneous items (adults only):
3 dates, 2 times, 3 proper names, 2 city names, 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language

Read speech:
- 30 phonetically rich sentences uttered by adults and 60 uttered by children
- 5 phonetically rich words (adults only)
- 4 isolated digits
- 1 isolated digit sequence
- 4 connected digit sequences
- 1 telephone number
- 3 natural numbers
- 1 money amount
- 2 time phrases (T1 : analogue, T2 : digital)
- 3 dates (D1 : analogue, D2 : relative and general date, D3 : digital)
- 3 letter sequences
- 1 proper name
- 2 city or street names
- 2 questions
- 2 special keyboard characters
- 1 Web address
- 1 email address
- 204 application specific words and phrases per session (adults)
- 74 toy commands, 14 phone commands and 34 general commands (children)

The following age distribution has been obtained:
Adults: 290 speakers are between 15 and 30, 166 speakers are between 31 and 45, 94 speakers are over 46.
Children: 24 speakers are between 8 and 10, and 26 speakers are between 11 and 14.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	300 Euro	600 Euro
For commercial use	4,000 Euro	6,000 Euro

Evaluation Packages from CLEF

ELRA-E0036 CLEF AdHoc-News Test Suites (2004-2008) – Evaluation Package

The Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access (MLIA) by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

The CLEF AdHoc-News Test Suites (2004-2008) contain the data used for the main AdHoc track of the CLEF campaigns carried out from 2004 to 2008. This track tested the performance of monolingual, bilingual and multilingual Information Retrieval (IR) systems on multilingual news collections.

The CLEF Test Suite is composed of:
- News Data Collections
- Topics
- Guidelines
- Relevance assessments
- Official campaign results
- Working notes papers

The News Data Collections consist of the following datasets:

- Bulgarian
  - Sega 2002 (33,356 documents, 120 Mb)
  - Standart 2002 (35,839 documents, 93 Mb)
- Czech
  - Mladna frontaDnes (68,842 documents, 143 Mb)
  - Lidove Noviny (12,893 documents, 35 Mb)
- Dutch
  - NRC Handelsblad 1994/95 (84,121 documents, 299 Mb)
  - Algemeen Dagblad 1994/95 (106,483 documents, 241 Mb)
- English
  - Glasgow Herald 1995 (56,472 documents, 154 Mb)
  - Los Angeles Times 1994 (113,005 documents, 425 Mb)
  - Los Angeles Times 2002 (135,153 documents, 434 Mb)
- Finnish
  - Aamulehti late-1994/95 (55,344 documents, 137 Mb)
- French
  - Le Monde 1994 (44,013 documents, 157 Mb)
  - Le Monde 1995 (47,646 documents, 156 Mb)
  - SDA French 1994 (43,178 documents, 86 Mb)
  - SDA French 1995 (42,615 documents, 88 Mb)
- German
  - Frankfurter Rundschau 1994 (139,715 documents, 320 Mb)
  - Der Spiegel 1994/95 (13,979 documents, 63 Mb)
  - SDA German 1994 (71,677 documents, 144 Mb)
  - SDA German 1995 (69,438 documents, 141 Mb)
- Hungarian
  - Magyar Hirlap 2002 (49,530 documents, 105 Mb)
- Italian
  - La Stampa 1994 (58,051 documents, 193 Mb)
  - SDA Italian 1994 (50,527 documents, 85 Mb)
  - SDA Italian 1995 (48,980 documents, 85 Mb)
- Persian
  - Hamshahri 1996-2002 (166,774 documents, 611 Mb)
- Portuguese
  - Público 1994 (51,751 documents, 164 Mb)
  - Público 1995 (55,070 documents, 176 Mb)
  - Folha de São Paulo 1994 (51,875 documents, 108 Mb)
  - Folha de São Paulo 1995 (52,038 documents, 116 Mb)
- Russian
  - Izvestia 1995 (16,716 documents, 68 Mb)
- Spanish
  - EFE 1994 (215,738 documents, 509 Mb)
  - EFE 1995 (238,307 documents, 577 Mb)
- Swedish
  - Tidningarnas Telegrambyrå 1994/95 (142,819 documents, 352 Mb)

The full package consists of 2.43 Gb and is stored on 1 DVD.

For evaluation use	ELRA members	Non-members
By an academic org.	150 Euro	300 Euro
By a commercial org.	500 Euro	1,000 Euro

ELRA-E0037 CLEF Domain Specific Test Suites (2004-2008) – Evaluation Package

The Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access (MLIA) by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

The CLEF Domain SpecificTest Suites (2004-2008) contain the data used for the Domain Specific track of the CLEF campaigns carried out from 2004 to 2008. This track tested the performance of monolingual, bilingual and multilingual Information Retrieval (IR) systems on multilingual collections of scientific articles.

The CLEF Test Suite is composed of:

- Data Collections
- Topics
- Guidelines
- Relevance assessments
- Official campaign results
- Working notes papers

The Data Collections consist of the following datasets:

- German Indexing and Retrieval Test database (302,638 documents, 524 Mb):

Data collection (social sciences) including a German corpus (151,319 documents) and a pseudo-English corpus which is in fact a translation of the German corpus into English (does not contain as much textual information as the German version).

- Cambridge Scientific Abstracts - Sociological Abstracts (20,000 documents, 38.5 Mb):

Database of Sociological Abstracts from Cambridge Scientific Abstracts.

• Russian Social Science Corpus (94,581 documents, 65 Mb):

Russian sociology database data from the Russian Social Science Corpus.

• Institute of Scientific Information for Social Sciences (Russian Academy of Science) (145,802 documents, 12 Mb):

The INION-ISISS corpus consists of bibliographical data from the ISISS database (03.02.2006) covering economics (~99,000 documents) and social sciences (46,000 documents).

The full package consists of 617 Mb and is stored on 1 CD.

For evaluation use	ELRA members	Non-members
By an academic org.	150 Euro	300 Euro
By a commercial org.	500 Euro	1,000 Euro

ELRA-E0038 CLEF Question Answering Test Suites (2003-2008) – Evaluation Package

The Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access (MLIA) by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

The CLEF Question Answering Suites (2003-2008) contain the data used for the Question Answering (QA) track of the CLEF campaigns carried out from 2003 to 2008. This track tested the performance of monolingual, bilingual and multilingual Question Answering systems on multilingual collections of news documents.

The CLEF Test Suite is composed of:
- News Data Collections
- Questions
- Guidelines
- Relevance assessments
- Official campaign results
- Working notes papers

The News Data Collections consist of the following datasets:

- Basque
  - Egunkaria 2000-2002 (119,982 documents, 212 Mb)
- Bulgarian
  - Sega 2002 (33,356 documents, 120 Mb)
  - Standart 2002 (35,839 documents, 93 Mb)
  - Novinar 2002 (18,086 documents, 48 Mb)
- Dutch
  - NRC Handelsblad 1994/95 (84,121 documents, 299 Mb)
  - Algemeen Dagblad 1994/95 (106,483 documents, 241 Mb)
- English
  - Glasgow Herald 1995 (56,472 documents, 154 Mb)
  - Los Angeles Times 1994 (113,005 documents, 425 Mb)
- Finnish
  - Aamulehti late-1994/95 (55,344 documents, 137 Mb)
- French
  - Le Monde 1994 (44,013 documents, 157 Mb)
  - Le Monde 1995 (47,646 documents, 156 Mb)
  - SDA French 1994 (43,178 documents, 86 Mb)
  - SDA French 1995 (42,615 documents, 88 Mb)
- German
  - Frankfurter Rundschau 1994 (139,715 documents, 320 Mb)
  - Der Spiegel 1994/95 (13,979 documents, 63 Mb)
  - SDA German 1994 (71,677 documents, 144 Mb)
  - SDA German 1995 (69,438 documents, 141 Mb)
- Italian
  - La Stampa 1994 (58,051 documents, 193 Mb)
  - SDA Italian 1994 (50,527 documents, 85 Mb)
  - SDA Italian 1995 (48,980 documents, 85 Mb)
- Portuguese
  - Público 1994 (51,751 documents, 164 Mb)
  - Público 1995 (55,070 documents, 176 Mb)
  - Folha de São Paulo 1994 (51,875 documents, 108 Mb)
  - Folha de São Paulo 1995 (52,038 documents, 116 Mb)
- Romanian
  - Wikipedia
- Spanish
  - EFE 1994 (215,738 documents, 509 Mb)
  - EFE 1995 (238,307 documents, 577 Mb)

The full package consists of 2.16 Gb and is stored on 1 DVD.

For evaluation use	ELRA members	Non-members
By an academic org.	150 Euro	300 Euro
By a commercial org.	500 Euro	1,000 Euro