The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

January - June 2008

Vol.13 n.1 & 2

Under the High Patronage of
His Majesty King Mohammed VI

lrec 2008 marrakech

LREC 2008 Special Issue

6th International Conference on Language Resources and Evaluation

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57 rue Brillat Savarin
75013 Paris - France
Tel.: +33 (0)1 43 13 33 33
Fax: +33 (0)1 43 13 33 30
Email: choukri@elda.org
Web sites:
http://www.elra.info
http://www.elda.org

Contents

Dear Colleagues,

This special issue of the ELRA newsletter is devoted, like every two years, to the Language Resources and Evaluation Conference (LREC). The 6th edition of the Language Resources and Evaluation Conference took place last May at the Palais des Congrès Mansour Eddahbi in Marrakech, Morocco under the High Patronage of His Majesty King Mohammed VI.

This edition has been very popular: more than 1,100 participants coming from 57 countries registered to the main conference, workshops and tutorials. A majority of participants came from all over Europe, but this time, the participation from America and Asia was significant. The United States of America brought the highest number of participants.

For LREC 2008, three major changes were introduced:

- The conference structure was modified with Poster sessions being held in parallel with Oral sessions
- LREC 2008 Proceedings, provided to all participants on CD-ROM, were made available from the conference web site to all immediately after the conference.
- 25 workshops and 6 tutorials, with up to 7 sessions in parallel each day, took place during the pre- and post-conference days, covering a wide range of subjects including HLT Evaluation, LR standardization, multimodality, sentiment analysis or sign language.

A joint COCOSDA/WRITE workshop on *Common priorities and recommendations for the future of Language and Speech Resources* was held on Sunday, June 1st, 2008. As in 2004 in Lisbon and in 2006 in Genoa, this meeting was organised jointly by COCOSDA, the International Committee for Co-ordination and Standardisation of Speech Databases, and WRITE International Committee for Written Resources Infrastructure, Technology, and Evaluation. The COCOSDA session of the workshop dealt with both the current situation and the future of speech and related multimodal/multimedia resources. Then, the WRITE session focused on forward-looking views and strategies for the LRs, and introduced the new FLaReNet project, funded by the European Commission.

Four 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 2008, the Antonio Zampolli Prize has been awarded to Yorick Wilks, from the Oxford Internet Institute and the Computer Science Department of the University of Sheffield (UK). Yorick Wilks' presentation has been made available on-line, from the LREC home page:

http://www.lrec-conf.org/lrec2008

Regarding the content of this ELRA newsletter dedicated to LREC 2008, we decided to have a special double issue, due to the high number of contributions from authors and presenters at LREC 2008. We have received numerous session summaries and we are happy to offer an overview of this LREC in the ELRA newsletter. In addition, Opening Ceremony speeches and a conference report are also included.

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

Bente Maegaard, President Khalid Choukri, CEO

INTRODUCTION
by Nicoletta Calzolari, LREC 2008 Conference Chair

Nicoletta Calzolari

Let me first express to His Majesty Mohammed VI, King of Morocco, the Program Committee gratitude for his Royal Patronage of this 6th edition of LREC.

This year LREC celebrates its 10th birthday! It started in 1998 in Granada, from a great and visionary idea of Antonio Zampolli, and it was then really an adventure and a challenge. I still remember Antonio saying "we won't even have 50 participants" … and we had about 500. When LREC was established, Language Resources (LRs) - and with them Evaluation - were only just starting to receive by larger sections of the HLT (Human Language Technology) community the attention that for many years was given to other aspects of language technology (LT). After only 10 years, LREC has become one of the big and established conferences in the sector.

This year LREC has established a new record: we received more than 900 submissions, we have accepted 645 papers, and we may have more than 1,000 participants. These figures are quite impressive to me, and are a sign of the great vitality of the broad area of LRs and Evaluation.

LREC is judged - I believe - as a great observation post for feeling the pulse of today's initiatives in the field, with the possibility not just of listening to the "best or most innovative" method or approach, but of examining the large variety of approaches, of analysing the many angles, of exploring the variety of resources for many languages, the new emerging trends, the large projects and initiatives, the infrastructures. I believe that this broadness of themes, topics, perspectives, is an essential contribution to form a better global vision of our field and thus to stimulate new ideas.

We do not have to defend or promote the 'data-driven' approaches any longer: they are pervasive and have a well-deserved and ample recognition. Large-scale LRs are unanimously recognised as the necessary infrastructure underlying LT. It is the merit (mostly) of LRs that LT acquires the maturity and attains the robustness needed to become truly usable in real world applications. This is probably the major result of LRs, with an impact on the transformation of LT from 'just' an R&D sector to a technology with a great impact in society.

We can say today that a 'LR community' exists: we simply have to observe the large LREC attendance. The achievement of a worldwide linguistic infrastructure, however, requires the coverage not only of a range of technical aspects, but also - and maybe most critically - of a number of organisational and coordination aspects. The stable growth of the field brings in itself some sort of revolution, and a need now to converge. An essential element for ensuring an integrated basis is to enhance the cooperation among many communities now acting separately, such as LR and LT developers, written, speech and multimodality specialists, terminology, semantic web and ontology experts, content providers, linguists and so on. This is one of the challenges for the next years, for a usable and useful 'language' scenario in the global and multilingual network. I believe that LREC may play an important role in the integration of these various communities.

This does not mean that there are no infrastructural issues still to be discussed and solved. Let us only think about the problem of maintenance of LRs, or the big issues of interoperability and sharing. The big growth of the field should be complemented by a reflection - within the same community - on priorities and future strategies. It is a big achievement, and a great opportunity for our field, that recently a number of strategic-infrastructural initiatives have started, or are going to start, in all the continents. This is also a sign that funding agencies recognise the strategic value of our work and the importance of helping a coherent growth also through a number of coordinated actions. LREC, together with its workshops, is the place where these - and future - initiatives will be presented, discussed, and promoted.

LREC, both the main conference and its workshops/tutorials, deals with the topics of LRs and evaluation within all the modalities: written and spoken language, and multimodality. Which are the main themes and issues at LREC 2008? Some very sketchy observations delineating some trends:

- There is a lot of semantics, knowledge, ontologies, content, analysed from every possible angle, from representation and annotation to acquisition, from spatio-temporal information, coreference and discourse, to emotions, affect, opinions.
- Complementary to this, even if with less emphasis, also syntactically annotated resources and related tools continue to receive attention.
- Among the LR-related tools and systems, information extraction is still a major issue, not yet solved, and more and more ambitious; well represented are also named entity recognition, document classification, question answering and summarisation.
- In speech: broadcast news processing, speaker identification, pronunciation data, speech synthesis, recognition, dialogue, large varieties of speech corpora for many languages, but also the topics of affect and emotion.
- Multimodality and multimedia span from TV and video processing to sign language, communication and dialogue in multimodal environment, emotion and subjective content again.
- As usual, large infrastructures, architectures, big initiatives and large projects, together with the issues of interoperability and standardisation, receive wide attention at LREC.
- Many papers, as expected, on lexicons, corpora and many different types of related tools.
- Evaluation of systems and validation of resources, both for written and spoken language, are broadly represented, as well as evaluation methodologies and evaluation campaigns.
- Many papers are on multilingual resources and systems and machine translation applications, but also dialects and language varieties.
- Terminology is an important issue, and the most represented terminological domain is the biomedical one.

As another sign of the great significance and the success of the field, LREC is now somehow complemented by the *Language Resources and Evaluation* journal, endorsed by ELRA and co-edited by Nancy Ide and myself. Springer offers LREC 2008 participants a subscription to the journal at a special rate. Moreover, we will ask the authors of the papers mostly recommended by reviewers as appropriate for the journal if they want to submit a longer version to the journal.

I am also proud to announce that we have decided to make all the LREC Proceedings, together with the proceedings of accompanying workshops, available on the web as a service to the community. It will happen soon after this conference.

A few last notes, on which we would like to get your opinions. You see in the program that we experiment this year a new conference structure, with Posters in parallel with Oral sessions. This also allowed us to accommodate more posters. We have also accepted more Tutorials and Workshops than usual, because of the large number of good submissions. We invite your feedback on these changes. We are also asking ourselves, given the so high - and ever increasing - number of submissions, if we should experiment with a 4-day conference: we will ask you about this too.

Acknowledgments

First of all my thanks go to the Program Committee, that at every LREC has a harder job, having to deal with an ever increasing number of submissions, of a better quality.

Also on behalf of all the members of the Program Committee, I warmly thank all the various Committees that have made this LREC possible, and hopefully successful.

I thank ELRA and the ELRA Board, for their continuous commitment to LREC.

I thank our impressively large Scientific Committee, composed of about 500 colleagues from all over the world. They did a wonderful job, succeeding to complete their reviews in time for so many papers.

We are indebted to the International Advisory Board and the Local Advisory Board, that have provided moral support to our Conference, and to the Local Organising Committee, and in particular Abdelhak Mouradi, for helping in finding solutions to local issues.

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

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

I thank the workshop, tutorial, and panel organisers, who surround 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 wish finally to 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 institute, ILC-CNR in Pisa. Without their commitment LREC would not have been possible.

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. Despite the distance (Paris-Pisa) they have worked together as a unique team, with enthusiasm and dedication. My biggest thanks go to them: Roberto Bartolini, Olivier Hamon, Vincenzo Parrinelli, Valeria Quochi, Mathieu Robin-Vinet, Sergio Rossi, and in particular Sara Goggi and Hélène Mazo who have become over the years one of the pillars of LREC. I cannot list the many tasks they carried out, but I can say for sure that without their daily work and real commitment since many months, LREC would not happen. We have solved together many big and small problems of such a large conference. They, together with other young researchers of these two institutions, will assist you during the days of the tutorials/workshops and the conference.

Now LREC is in your hands, the participants. You are the protagonist of LREC, you will make this LREC great (I am sure). So at the very end my greatest thanks go to you all. I may not be able to speak with each 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), most of all that you profit of so many contacts to organise new exciting work and programmes in the field of LRs and evaluation, which you will show at the next LREC.

I particularly hope that funding agencies all over the world are impressed by the quality and quantity of 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. The success of LREC means to us in reality the success of the field of LRs and Evaluation.

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

Enjoy LREC in Marrakech!

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

Palais des Congrès Mansour Eddahbi

LREC 2008 Opening Ceremony Speeches

Message from the ELRA President, Bente Maegaard, University of Copenhagen

Bente Maegaard

Let me first express, on behalf of the ELRA Board and members, our profound gratitude to His Majesty Mohammed VI, King of Morocco, for his Royal Patronage of this 6th edition of LREC.

When ELRA was established only 13 years ago, in 1995, the main purpose of the association was of course the identification and distribution of language resources. But very soon the idea emerged of organising a conference covering the same fields as ELRA, with the addition of Evaluation. Such a conference should promote the field, and create a meeting place, and the first conference was organised by ELRA in 1998. And this idea proved to be very good, - the LREC conference has established itself as the main meeting point of those who believe that language resources and evaluation are main building blocks for language technology both for written and spoken language. If you want to meet somebody from the field, just go to LREC and he/she will be there.

Over time, ELRA has further developed its mission from LR distribution to also cover production, validation and support for evaluation of language technologies. And language resources have developed from relatively simple speech or written resources to more advanced resources and to multimodal resources. The ELRA Board, and the distribution agency, ELDA, are watching the development, and welcome any request for specific types of resources and even for specific resources. We may be able to find them for you, or to encourage their production.

ELRA has a number of strategic activities. We have been further developing the main activity, namely identification and distribution of LRs. The ELRA catalogue now contains around 1000 resources (1012 to be specific), i.e. resources for which we have obtained distribution rights. In addition, ELRA has asked ELDA to make the assembled list of existing LRs available to our members. We call this the Universal Catalogue, in contrast to the ELRA catalogue which contains only the LRs. We believe the Universal Catalogue is very interesting and very useful, and we hope to be able to have an online service for those who want to provide information about their own resources, so that we can help spread information about what is available.

I would also like to mention the Antonio Zampolli 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. A citation from 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. The prize covers the field of Language Resources and Language Technology Evaluation in the areas of spoken language, written language and terminology". At the LREC2008 conference, the Prize will be awarded for the third time. The ELRA Board has been very happy to receive the nominations made by outstanding people in the field, and we recognize there are several persons who are eligible for this prestigious prize.

At LREC 2008 you will have the chance to discuss strategic issues concerning language resources and evaluation and the contribution of these two fields to the further development of language technology for both spoken and written language. You will also see a multitude of language resources and tools for very many different languages that may be useful for your own work or you may get or provide new ideas for the further evolution of the field. This year you will also see several new networks and projects focussing on language resources. This is a very important development, and it is a profound pleasure to see that funding authorities are supporting the development, the distribution and the sharing of language resources.

Please take advantage of all this, and enjoy your participation!

Finally, I would like to take the opportunity to thank all those who contributed so hard to making this conference a success. This year we are for the first time outside Europe, we are in Morocco - which is not too far away from Europe! The fact that we are in Morocco, means that a very large part of the local organization work has been taken care of by Khalid Ckoukri. At the same time Khalid Choukri and the ELDA team have taken a large part of the responsibility for the practical organization of the conference, so this has been a gigantic task and we thank Khalid Choukri and all of his team.

The content part of the conference has been taken care of by the Programme Committee, as always chaired by Nicoletta Calzolari, from the Istituto di Linguistica Computazionale, CNR, Pisa. Nicoletta Calzolari and her team in Pisa have been managing more than 900 submissions, their review, their selection, final papers, and finally the creation of the programme that you will be enjoying these days.

We are deeply grateful to Nicoletta Calzolari and her team, as well as to the Programme Committee. Finally, we would like to thank the Scientific Committee who did all the reviewing and the International Advisory Committee for their valuable advice.

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

Message from Khalid Choukri, ELRA CEO and ELDA Managing Director

Khalid Choukri

Dear Mr. Taib Debbag, representing Mr. Ahmed Reda Chami, Minister of Trade, Industry and New Technologies,

Mr. Jochen Richer, representing Mr. Leonard Orban, the European Commissioner for Multilingualism,

Dear Colleagues, Dear Friends,

مرحبا بكم جميعا في مراكش - مرحببكم كلكم فمراكش

ⵎⵔⴰⵃⴱⴰⴱⵉⴽⵓⵎ ⵢⴰⵔⵏⴰⵢ ⴷⵉ ⵎⴰⵔⴰⴽⵛ

Marhaban becom jamias fi Marrakech,
mrahbabicml fmarrakech,
mrahbabikoum ran di Marrakech,

These three very short sentences stated in Modern Standard Arabic, Moroccan colloquial Arabic and Amazigh from the Rif region, North of Morocco, summarize in a simple but exceptional way both the LREC philosophy and the Moroccan attitude towards language issues, multilingualism, multiculturality, cultural diversity, variety of civilizations and their dialogues, and also toward new information technologies.

Let me first express my profound gratitude to King Mohammed VI for honoring our conference with his Royal Patronage. The Royal Patronage highlights the importance his majesty attributes to this field and the encouragement his Majesty has devoted to the different national and international initiatives within our sector. The establishment of a Royal Institute for the Amazigh language (IRCAM), the native language of Morocco, in complement to the strong support to existing institutions like the Institute for Arabization (IERA), shows the Royal Commitment to see the languages reinforce the foundations of inter-cultural comprehension and diversity in Morocco and beyond.

So in the name of ELRA, LREC Program Committee, and I am sure I can speak on their behalf, let me express to his Majesty our gratitude for His support and read a few sentences of the letter we will send his Majesty after our meeting.

To His Majesty King Mohammed VI,

His Majesty,

On behalf of the European Language Resources Association and the participants of LREC 2008 we would like to express to His Majesty our gratitude for the honor to hold our conference under His Majesty's High Patronage. We would like to express our gratefulness for the examples set up by His Majesty in boosting the multilingual nature of the Moroccan society and assure His Majesty of our full support to such royal actions and the commitment of our community to enrich such experiences through our own efforts in joint projects that build on His Majesty's Language diversity initiatives.

His Majesty,

Our activities in language technologies as well as in language resources (collection and preservation), and in evaluation of language technologies are key instruments to remove language, culture and civilization barriers. These activities are also key instruments for better educated young generations, to give all citizens access to the information society and to give elders better living conditions through speech and other related ambient technologies.

It is important for our community to inspire from His Majesty's initiatives that stress the dual aspects related to the preservation of our cultural heritage while investing on new technologies and on the training of human resources.

With our deepest respect.

Dr. Khalid CHOUKRI, on behalf of the ELRA Board and the LREC committees.

Dear LREC participants,

We are proud to welcome you in Marrakech, where we are pleased to contribute to the organization of this sixth edition of the Language Resources and Evaluation Conference, LREC 2008. We are very happy to continue the organization of such a major conference for the Human Language Technology community in such a wonderful city.

Let me take this opportunity to thank Mr. A. Reda Chami, Minister of Trade, Industry and New technologies for his support and ask Dr. Taïb Debbagh to convey our best wishes. I would like to thank Mr. Leonard Orban, the European Commissioner for multilingualism, and his representative today Mr. Jochen Richter for his participation. I would like to thank Mr. Ahmed Lakhrif the Moroccan Secretary of State for the Foreign Affairs and Cooperation for his strong support, and Mr. Mounir Chraïbi, Wali of Marrakech, although he cannot be with us today, for his continuous support and friendly help. I would like to express particular thanks to his excellence Fathallah Sejilmassi, the Moroccan Ambassador in France and his team (in particular Khalid Afkir) for his strong commitment and dedication.

I would also like to extend our warm thanks to our Moroccan colleagues from the Local Advisory Committee; you may have seen that several ministries, most of the Moroccan universities and institutions are contributing to the success of LREC 2008. The list is in our booklet to give you some contacts for potential joint projects and cooperation but let me just quickly name the institutions: the University Al Akhawayn (Ifrane), the University Mohammed V Souissi (Rabat), the University Cadi Ayyad (Marrakech), the CNRST (Centre National pour la Recherche Scientifique et Technique) (Rabat), the Royal Institute of the Amazigh Culture, (Rabat), the Institut d'Etudes et de Recherches pour l'Arabisation (Rabat), INPT (Institut National des Postes et Télécommunications) (Rabat) and ENSIAS (École Nationale Supérieure d'Informatique et d'Analyse des Systèmes) (Rabat).

I cannot forget that this year we celebrate the 10th anniversary of LREC, some of

you may remember how things started ten years ago in Granada (LREC 1998) with Antonio Zampolli and Angel Martin-Municio, I would like to say how much we owe them today.

The emergence of the Web has largely contributed to promote the sharing of language resources but also, like everything else around the web, it has made it critical to excogitate about the nature and the quality of what one gets on the web and the various implications related to such availability (quality, rights and licensing, reuse and exploitation,…). Let me mention just one issue that will be largely debated in this conference: the legal aspect. A survey of the existing licensing schemes recently conducted by ELDA showed that there are more than ten license models, several of them backed up by large institutions, and which handle similar, or even identical, legal matters. Several of them are derived from the software industry traditions and simply transposed to Language Resources. Such abundance and proliferation, instead of helping users to better understand their rights, duties, responsibilities, and liabilities, are confusing most of them who are not prepared to invest time and money in legal advises. Other aspects will also be discussed that may have a substantial impact on the next decade of Language Resources and Evaluation. The proliferation of resources will also be of major importance in our field. It is becoming natural to search the web for the latest dictionary, corpus, list of words, broadcast recordings, etc, free of charge and copyright free. One may identify a large number of such items for a given language. In many cases, there are serious doubts about the quality but also about the origin as well as the actual right owners. It is a challenge for Data Centers like ELRA, LDC, GSK, Chinese-LDC, and similar agencies, to ensure that such a trend is fully supported and streamlined for the benefits of all.

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

Suggestions to improve any aspects of the conference are welcome, and if you need any assistance to make this event a more memorable one, please do not hesitate to contact our staff, who will be very pleased to help.

Let me now give you some practical information about the next few days:

Social events: We are very pleased to invite you to the Welcome reception of tonight. This will take place in a very mysterious place that we will reach after a 20mn ride. Buses will leave the conference center at 20:15 so be on time and make sure you have your invitation with you. May I remind you that evenings in Marrakech are very chilly so dress accordingly.

Next Friday we will be happy to invite you to the ELRA Gala dinner, we always feel happy that after 3 hard days we get a chance to get together and relax in a lovely atmosphere. Again shuttles are planned for all of us and we will be leaving the conference center around 20:15.

Regarding security, please keep your badge visible; I am sure you understand the security issues behind that with more than 1000 attendees.

For Lunch: nothing is organized but I am sure you will enjoy some of the close-by restaurants (there are many in the avenue Mohammed VI just in front of the center but this may take some time). There are also lunch boxes that you can buy from the travel agent as well as tickets for the Mansour Hotel buffet.

The WIFI login is LREC2008. Please make sure that you do not use all the bandwidth and share it with your colleagues.

For any assistance please contact our staff who you can recognise by the RED lanyards they wear.

We have appointed a travel agent to handle the extra conference aspects so please talk to them to make your stay in Marrakech more enjoyable.

I cannot finish this official statement without a very personal thought to one of my team members, Victoria Arranz. I am sure that most of you have been in touch with Victoria during the last LREC and during the last few weeks and months. I am sorry to say that Victoria had to leave urgently for family reasons and I would like to convey our warm friendship.

Once again, welcome to Marrakech, welcome to LREC 2008.

Welcome reception: Fantasia El Borj Bladi

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

Message from Taieb Debbagh, Vice-Minister of Trade, Industry and New Technologies, Morocco

Taieb Debbagh

First of all I would like to welcome all participants coming from all over the world for attending the 6th International Language Resources and Evaluation Conference (LREC 2008) organised by the European Language Resources Association (ELRA) in Marrakech from May 26th to June 1st, 2008, held for the first time outside Europe.

I would also like to convey the apologies of Mr. Ahmed CHAMI, Minister of Trade, Industry and New Technologies who would have wished to be with us during this important event, had he not been called to another mission.

This Marrakech edition provides a unique occasion for Moroccan participants in particular, and for all participants from Africa and the Middle East, to join the discussion on the issues relating to the promotion of multilingualism through the production of linguistic resources in numerous languages.

Ladies and Gentlemen,

It is my pleasure to take the opportunity of this important conference to present the strategy of the Government in the development of Communication and Information Technologies.

Conscious of the strategic nature of information and telecommunication technologies, Morocco has taken numerous initiatives recently: legislative and regulatory frameworks have been reviewed, liberalisation has been initiated and considerable effort has been put on infrastructure. At the same time, several measures have been adopted in favour of small and medium sized companies.

The telecommunication field is of utmost importance in our economic and social development and it has undoubtedly benefited from the liberalisation policies conducted in recent years. Mobile telephony, in particular, has witnessed an unprecedented growth rate which has reached 66.85% at the end of Q1 2008. This means, in other words, over twenty million subscribers.

This success in the telecommunication sector must inspire our efforts for a more democratic access to the Internet in our country. Concerted actions, involving both equipment and broadband access and content, will enable us to make progress with this incredible information and knowledge tool. We will thus create one of the levers which will contribute to improve Moroccan companies' productivity and competitive position.

Ladies and Gentlemen,

The Ministry of Industry, Trade and New Technologies, has created a short-term priority action plan aiming essentially at reinforcing the foundations and the competitive position of Morocco in the fields of information and communication technologies.

This plan, which incorporates seven priority targeted actions for the period 2008-2009, will contribute to the improvement of the Moroccan position on the world scale and reduce the digital gap which separates the country from its major competitors.

The seven priorities in this plan are social transformation, adoption of communication and information technologies by companies, a user-oriented public service, modernisation of the administration, competitiveness of the sector, reinforcement of research and development (R&D), innovation in the field of information technology and improvement of the telecom infrastructure.

Ladies and Gentlemen,

The action plan for 2008-2009 will be achieved within the framework of a public/private partnership, which is an essential component in the success of this programme, the objective being to reinforce relationships with private groups, both Moroccan, through the Association of professionals of the information technologies sector, as well as international players.

The Government will implement the piloting structures and the coordination of this programme, which will be launched this July. In order to achieve this, a national council for the information society will be created, chaired by the Minister for Industry, Trade and New Technologies.

This plan, which objectives are both realistic and achievable insofar as they do not require a major financial effort is only the beginning. The Government is currently working on the development of a strategic plan for the period 2009-2013 which aims to literally transform Morocco into an information society.

This strategy is such as to provide serious momentum as regards the development of the information society both from an industrial point of view and from the point of view of the users.

Ladies and Gentlemen,

Offshore activity, particularly in relation to call-centres, provides employment today for more than 20,000 people. And this sector, which did not exist only 8 years ago, generates today 2 billion-dirham annual revenues. It has consolidated its position by rightly adopting a strategy of high quality, added value and with the maximal optimisation of human resources, which constitute the major trump card of Morocco in all fields from telephone to the Internet.

In order to reinforce its position as an offshore destination which can attract more investors in the most innovating sectors, Morocco must also promote linguistic engineering which nowadays enables extraction of textual information, translation or automatic abstracting, production of multilingual information, localisation of software and also voice recognition.

Ladies and Gentlemen,

I am sure that this conference will be a fruitful event in terms of debates and exchanges and that it will lead to results in your field.

As a conclusion, I would like to thank the European Language Resources Association for having chosen Morocco to host this important event. I thank all participants present here and wish you the best of success for your work.

Taieb Debbagh
Vice-Minister of Trade, Industry and New Technologies
1, avenue Moulay El Hassan
Rabat, Morocco

Message from Jochen Richter, Deputy Head of Cabinet with Leonard Orban, Commissioner for Multilingualism, European Commission

Cher représentant du Ministère de l'industrie, dear President and Members of the Organising Committee, dear participants,

Thank you very much for having the honour of being invited to this conference. For somebody that started his career in the early 80ies in informatics it's a real pleasure being here today.

Congratulations to the authorities of Morocco and the city of Marrakech for the excellent organisation of this event.

I am sending you the best regards of my Commissioner, Leonard Orban, who for other professional obligations cannot be here.

Ladies and gentlemen,

Although much of the discussion at this conference is going to be technical in nature, and we shall probably be hearing more about interfaces than dialogue, everyone here is in some way involved in trying to make it easier for people to communicate across linguistic and cultural boundaries.

The five previous editions of this conference have all been held in Europe. Today, migration and globalization are making our communities increasingly multicultural, and we need to communicate far beyond the borders of the European Union.

We can see this being reflected in some of the contributions to this conference, which are by no means confined to the European Union.

2008 is the European Year of Intercultural Dialogue. So it is particularly fitting that this year's conference should be taking place here in Marrakech. The Commission attaches much importance to the issue of intercultural dialogue. Languages play an important role. It is probably the best way to dive into a different culture by mastering the other one's language, being able to communicate, understand and reading literature as history of a country.

Therefore Commissioner Orban welcomed the idea born by the advisory group chaired by the French Lebanese writer Amin Malouf, that we in Europe should take account of the added-value of the languages of the migrants - obviously not an uncontroversial idea to which I will return in a moment.

Coming back to this conference, the European Commission has been actively involved from the outset in this initiative, giving support to these meetings, and also to several of the projects that are going to be discussed.

But this is only one part of a much bigger picture. I would like to briefly explain the other aspects of the Multilingualism portfolio.

What is the reason that the European Union subscribes to multilingualism. It has to do with the equality of Member States, their languages and thereby the different identities. The functioning of multilingualism is set down in the Union's ever first adopted regulation and constitutes a democratic right to understand and to be understood in your mother tongue. But is also important for legal certainty and as it is inevitable for citizens. For companies it secures a level playing field, both inside and outside the EU. And it is another symbol for the EU's foundation on cooperation between MS.

There are 5 dimensions to this portfolio.

Firstly education, where we are aiming to strengthen the understanding that language knowledge is an asset. That probably is easier when one starts to learn early. But we also need modern and attractive ways of language learning. There comes technology into play that can support these processes. That has to go hand in hand with alternative ways of motivation. There is scientific evidence that enhanced use of language families facilitates language learning. And let's not forget, also teachers need more support. All these aspects are covered by the Lifelong Learning programme that also includes support to regional and minority languages. But let me stress that the main aim is to create bridges rather than divisions. I would like also to underline that education is first of all a responsibility of Member States but the EU can provide a good platform to learn from each other.

The second aspect concerns the intercultural dialogue. This is not because of the 2008 European Year of intercultural dialogue but because of the fact that globalisation and migration are shaping our societies including vigilant discussions about what our identities constitute. This much needed dialogue is a challenge in fighting any form of what was called to be a possible clash of civilisation. This issue concerns the EU itself but also beyond. Again one can easily see the bridge-building function of languages. For this reason the Commissioner welcomed the idea of the Amin Malouf lead advisory group proposing that migrants should learn the language of the host country while society should take benefit from the other languages and cultures. In line with that principle the group proposed what they called a personally adopted language, one where your interest goes beyond learning the language as such but is close to your heart.

The third angle is the economic dimension, where once more globalisation is a determining factor. It is true to say that whole areas of life such as technology, science or financial markets seem to rely on English only. But a study from last year demonstrated that still 44% of our population is

Jochen Richter

monolingual, namely their mother tongue. And let's be honest with each other. When we are holding our exchanges in the corridors, we are often saying what we can and not necessarily what we want to say. Linguists call that form of English meanwhile Globish. To stress the economic value of language skills a business forum chaired by Vicomte Davignon is seeking to identify best practices of companies in making use of the language skills of their employees and their strategies for a language policy.

And then, fourthly, there is technology. I know what translators and interpreters will say. But no, it's not about replacing the human brain by a machine. It is about efficiency gains and support to deliver better quality. It is necessary to broaden the scope by looking at issues such as court interpretation, multilingual health services - including the 112 emergency number - and also local governmental services. Finally it is a question of long-term sustainability of a system that is likely to see a further increase of the number of official languages rather than the other way around.

Finally, there is also an external dimension. As other parts in the world already do, we should engage to teach and train people in languages beyond the EU languages. This is a question of openness and cooperation. Others have already shown interest in our system of multilingualism. We are in closer contact with China, India, South Africa and Russia.

This autumn the Commission will be gathering together all the various policy strands into a Communication on Multilingualism. This will set out our policies to safeguard and promote linguistic diversity in line with the idea of unity in diversity.

Allow me a side remark as I can imagine that this issue is of interest to some of the attending participants. The Commission adopted on 20 May its proposal for the structure and functioning of the Mediterranean Union. As the responsible Commissioner, Mrs. Ferrero-Waldner, pointed out this initiative is meant to foster the multilateral relations with the Mediterranean countries. In addition it will complement the bilateral relations of the European Neighbourhood Policy.

The new impetus this gives to the already since 12 years existing so-called Barcelona process is the recognition that despite some encouraging results there remains much to be done. To conclude on this point the priorities of this proposal are:

1. stepping up the cooperation at a political level,
2. finding the right balance of responsibilities and
3. giving the projects a profile that citizens can see the added value.

Ladies and gentlemen,

Coming back to the issue of this conference, we have, as I have indicated, our own internal need for translation and interpretation. Coping with massive information flows in a multiplicity of languages is a mammoth task in itself. Within the institutions, a series of enlargements have taken the number of official languages we have to deal with to 23.

Fortunately, a number of technical advances which have been worked on for years are at last bearing fruit, making it possible to save both time and money. Gone are the days where all a translator needed was a pen or a Dictaphone and a well-thumbed dictionary.

The European Union has been something of a pioneer in this field. Since the problem is a very urgent one for us, there has been a great incentive to look for solutions. The Translation Service has been very pro-active in the use of computers, and today they are equipped with state-of-the-art tools to assist them in their work.

Many of the texts they translate are based on previous texts or existing legislation. This has enabled us to create a very effective translation memory device. Today when our translators write a sentence, similar phrases from previous translations pop up in the form of suggestions.

This technique of re-using previously translated words or passages saves a considerable amount of time. It also ensures that terminology is used consistently - which is obviously vital in legislative texts.

We also use machine translation in the Commission - our ECMT system processes nearly a million pages in a year. It is partly used by translators to help in their work. Obviously, the raw output has to be edited to a greater or lesser extent.

Of course these are only two of the many tools available to translators, and experimentation is not confined to translation: interpretation is covered, too. In spite of years of work on speech recognition, nobody has so far perfected a speech-to-speech device which can actually replace interpreters. However, the Interpretation services have developed a number of other tools for multilingual communication, Multilingual web- streaming, chats, and videoconferencing are already commonplace, and other tools, such as a multilingual speech repository, provide aids for interpreter training.

So you can see that in looking for answers to our own internal language needs, we have found ourselves leading the way in the implementation of new language technologies. The Commission has been keen to share these new technologies with others: our policy is to re-use our information resources to get the maximum use from them. First we made our terminology and translation tools available to all staff across the institutions. And more recently, some of our instruments and resources have been made available to the public at large.

In June of last year, the vast IATE terminology database was opened up to the public. This is an in-house store of wisdom containing over 8 million terms and covers all the official languages of the EU. The content is constantly fed in by the language departments, after going through a rigorous validation procedure.

And at the beginning of this year the Commission's translators and in-house scientists released another treasure-trove to the public: our huge collections of sentences gleaned from legal documents. These cover technical, political and social issues in 22 languages (there is no Irish as yet). In this translation repository it is possible to find sentences with their equivalent in all other official languages. Elsewhere such resources are scarce for languages such as Latvian or Romanian, and they are practically nonexistent for the less common language-pairs.

These developments demonstrate how our internal policy of multilingualism can reach beyond the immediate needs of the institutions, and provide resources to researchers such as you. This is a clear

case of mutual benefit, since we in turn profit from the work that you do.

Ladies and gentlemen, in a few moments we are also here to celebrate a man who was a true pioneer of the discipline of Computational Linguistics.

This will be the third time that this prize will be awarded: it is a fitting memorial to a man who loved both literature and mathematics, and who scorned artificial boundaries between subject areas. Antonio Zampolli was one of the first to yoke together linguistics and computer science into an entirely new discipline. From his university department in Pisa he devoted his career to spreading the word about his passion.

He was also eager to establish an enduring basis for international co-operation in this field. He played a major role in masterminding the creation of the European Language Resources Association (ELRA), and, of course, these extremely successful LREC conferences.

In the work that he inspired, a shared interest in translation always brought people from different backgrounds closer together. This, of course, is the whole essence of translation, and I very much hope that that it will be the case today.

I should like to end by wishing this conference much success.

Thank you for your kind attention.

Jochen Richter
Deputy Head of Cabinet with Leonard Orban
Commissioner for Multilingualism
European Commission
BERL 08/327
1049 Brussels, Belgium
jochen.richter@ec.europa.eu

Message from Abdelhak Mouradi, Chair of the Local Organizing Committee

Abdelhak Mouradi

Dear LREC 2008 Participants, on behalf of the local advisory committee, the local organizing committee, and all participating organizations we would like to express our profound gratitude to His Majesty King Mohammed VI, King of Morocco, for so kindly agreeing to honour LREC 2008 with His Royal Patronage.

It is our pleasure to welcome you in the Imperial city, Marrakech, for the Sixth International Conference on Language Resources and Evaluation, LREC 2008.

The Conference takes place at the "Palais des Congrès Mansour Eddahbi", Marrakech, from May 26th to June 1st 2008 and it is organized in cooperation with a wide range of national and international associations and organizations.

Morocco has the privilege to be the first country out of Europe and in Africa to host the ELRA conference and to celebrate its 10th anniversary.

The LREC conference welcomes all scientists in the world working on basic research and applications in the field of Computational Linguistics, Human Language Technology and Information Society Technology.

The conference offers an excellent opportunity to highlight recent trends as well as general aspects of the field and gives to the researchers a chance to share experiences, points of view and results.

With more than 650 accepted papers and over 1000 participants, LREC 2008 represents the vitality and growth of Computational Linguistics. Therefore, we hope that the conference will provide a stimulating scientific environment for exchange of results and ideas for future research.

We would like to thank all the members of the different committees involved in the organization of this Conference.

We hope you will enjoy your visit to Marrakech, and remember both the scientific and social aspects of LREC 2008 as a pleasant and worthwhile experience. Please let us know if we can be of any assistance to make the event and Marrakech more memorable.

Abdelhak Mouradi
ENSIAS
B.P. 713, Agdal
Rabat, Morocco
mouradi@ensias.ma

Opening Session with Bente Maegaard, Taieb Debbagh, Nicoletta Calzolari, Khalid Choukri, Jochen Richter, Abdelhak Mouradi

LREC 2008 Antonio Zampolli Prize

Speech given by Bente Maegaard

This year, the Antonio Zampolli Prize was awarded to:

Professor Yorick Wilks

Professor of Computer Science, University of Sheffield, and Senior Research Associate, Oxford Internet Institute

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."

Yorick Wilks

Motivation:

Professor Wilks' research using corpora goes back to his Cambridge thesis from 1968: it was one of the first pieces of work to apply computation (LISP 1.6 on an IBM 360) to actual texts, rather than examples; the work considered 10 passages of paragraph length from philosophical literature, as well as 10 paragraphs as controls from Times editorials. The results were connected semantic structures, with the content words resolved to particular (unique) sense representations from a lexicon of 500 items containing more than one sense per word. The work was re-implemented, using longer passages of text. The underlying semantic coherence approach entered the literature as Preference Semantics, whose original goal was to determine the emergence of new senses of words in text not present in the lexicon.

In 1976 the preference semantics analysis method was extended to the analysis of case, as case was not really implemented in the first version. Yorick Wilks developed stacks of ordered patterns called "paraplates" that were applied in turn to semantic patterns drawn from a main clause and a prepositional phrase until one matched, thus determining the case and word-senses at the same time. The method was later simplified and was found to be one of the most effective determiners of correct attachment when applied to corpora.

Professor Wilks ran an NSF-supported activity at New Mexico (1986-90) to parse the whole LDOCE dictionary, extract semantic content from the entries and build ontology/thesaurus trees for the whole dictionary as a resource for large scale semantics research.

Professor Wilks' Consortium for Lexical Research at New Mexico (1990-1995) was a one of the first pre-web sites to distribute lexical and text corpora and software tools automatically via the Internet---it still gives free downloads. It was managed with Louise Guthrie and was a model for later efforts; it arose from a transatlantic series of seminars on resources organised with Antonio Zampolli.

Professor Wilks was the initial principal investigator for the software platform for NLP called GATE at Sheffield. The GATE system is now in its fourth implementation and is one of the largest and most downloaded systems of tools for NLP/CL and the automatic management, distribution and use of language resources. It has been the platform for a wide range of implementations, competitions and software developments in the EU and USA.

Professor Wilks was the Principal Investigator of a UK project for large-scale word-sense disambiguation of (adapted) LDOCE senses against corpora, using machine learning to combine the output of modules based on differing principles and resources; at the time of publication its results were the best world-wide. It was tested over a large corpus, derived, originally, by converting a large corpus annotated in one sense annotation system (Wordnet) to another (LDOCE).

Currently professor Wilks is working on a range of projects designed to reveal anomalies in large-scale web corpora, his own interest being the emergence and discovery of new senses in text, an interest going back to his starting point.

So, I am sure you understand why we have chosen Professor Wilks as this year's Antonio Zampolli Prize winner. With great success throughout a career, his work covers all the areas addressed at LREC: production of language resources, use of language resources, reuse of language resources, and distribution of language resources and tools.

Finally I want to thank the nominators, whose nomination is largely underlying this talk. I also confess that I have not mentioned the names of Yorick Wilks' collaborators - I apologize for that!

The presentation given by Yorick Wilks, entitled

"In my beginning is my end: reflections on 45 years of NLP and corpora.",

can be viewed from the LREC 2008 web site:

www.lrec-conf.org/lrec2008

LREC 2008 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 2008 Proceedings, which are available online: www.lrec-conf.org/proceedings/lrec2008

O2 - LRs: Infrastructure, Projects, Centres

Chiu-yu Tseng

The oral session on LRs consisted of 4 paper presentations on the ACL Anthology Reference Corpus, the LDC, a Geographic Information System within language sites, and CLARIN. The presentation of *ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research Computational Linguistics* by Steven Bird, Robert Date, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Wee, Brett Powley, Dragomir Radev and Yee Fan Tan discusses a digital archive of conference and journal papers in natural language processing that aims at serving as a reference repository of research results and other related usage, and that will be made public.

The LDC paper entitled *The Linguistic Data Consortium Member Survey: Purpose, Execution and Results* by Marian Reed, Denise DiPersio and Christopher Cieri reports about a member survey of the well-known and long-standing consortium to solicit user feedback, in particular the data needs of the LDC users for future reference. The survey brought forth an important feature of sustained service of publicly available corpora, be them free of charge or at a fee, after service and future development.

The presentation of *Language-Sites: Accessing and Presenting Language Resources via Geographic Information Systems* by Dieter van Uytvanck, Alex Dukers, Jacquelijn Ringersma and Paul Trilsbeek discusses about linking digital language corpora and resources via Google Earth's 3-D interface, thus incorporating GIS with language resources, as well as video information with text information through the Internet.

The last presentation of *CLARIN: Common Language Resources and Technology Infrastructures* by Tamas Varadi, Steven Krauwer, Peter Wittenburg, Martin Wynne and Kimmo Koskenniemi introduces the CLARIN project, a large scale 32-partner 23-country European research infrastructure designed to establish an integrated and interoperable infrastructure of language resources and technologies focusing on humanities and social sciences.

The four presentations collectively covered a wide dimension of research advancements and language resources, and active discussions from the participants were not limited to presentation contents only. The lively exchange brought forth lasted throughout the days during LREC 2008.

Chiu-yu Tseng
Institute of Linguistics, Academia Sinica
128, Section 2, Academia Road 115
Taipei, Taiwan
cytling@sinica.edu.tw

O3 - Corpus, Lexicon and Evaluation

Núria Bel

The session on "Corpus, Lexicon and Evaluation" gave a good overview on how the evaluation of language resources has gained in scope, as it is being used for validation purposes. Thus, the four presentations in this session are a confirmation of the existence of a continued change towards a quantitative approach that promotes the abstraction into general statements that help in decision making. Formal evaluation metrics (kappa coefficient and accuracy measures) are used as a tool to validate a tag set and annotation scheme in *Evaluating Dialogue Act Tagging with Naive and Expert Annotators* by Jeroen Geertzen, Volha Petukhova and Harry Bunt. These metrics help them in identifying what concepts of the tag set are difficult to understand and need reformulation, for instance.

Frahomíra Johanka Spoustová, Pavel Pecina, Jan Haji and Miroslav Spousta also used kappa coefficient to validate the quality of part-of-speech tagged corpus in *Validating the Quality of Full Morphological Annotation*. The metrics are used to identify the samples to be re-annotated in order to improve the quality of the corpus.

In *Evaluating Complement-Modifier Distinctions in a Semantically Annotated Corpus*, by Mark McConville and Myroslava O. Dzikovska, we saw another example of a validation strategy using kappa coefficient to evaluate the results of the merging of different annotation schemas for dictionary building.

And recall and precision metrics were used to validate the effect of grammar restrictions on the accuracy of annotation tools in *Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case* by Kremena Ivanova, Ulrich Heid, Sabine Schulte im Walde, Adam Kilgarriff and Jan Pomikalek. In this article, the authors validate the accuracy of a German Sketch Grammar for the identification of grammatical functions in German supplying the readers with very interesting facts such as the relation between precision and restrictions in the grammar rules, and the impact of additional linguistic information in the general accuracy.

Núria Bel
IULA, Universitat Pompeu Fabra
La Rambla, 30-32
08002 Barcelona, Spain
nuria.bel@upf.edu

O8 - Multimodal Annotation Tools

Jean-Claude Martin

The session on "Multimodal Annotation Tools" included 4 presentations covering several burning issues of the domain: exchange format, category registries as well as sample corpora of various multimodal behaviors.

The first paper was presented by Schmidt et al. about *An Exchange Format for Multimodal Annotations*. It described the results of a joint effort of a group of multimodality researchers and tool developers to improve the interoperability between several tools used for the annotation of multimodality. The authors proposed a multimodal annotation exchange format, based on the annotation graph formalism, which is supported by import and export routines in the respective tools.

The second paper was presented by Stoia et al. about *SCARE: a Situated Corpus with Annotated Referring Expressions*. The presented corpus enables to answer research questions related to situated language that should connect world information to the human language. The release of a corpus of English spontaneous instruction giving situated dialogs was reported. The corpus was collected using the Quake environment, a first-person virtual reality game, and consists of pairs of participants completing a direction giver-direction follower scenario.

The third paper was presented by Han Sloetjes and Peter Wittenburg about *Annotation by Category: ELAN and ISO DCR*. The Data Category Registry is one of the ISO initiatives towards the establishment of standards for Language Resource management, creation and coding. This presentation described the first steps that have been taken to provide users of the multimedia annotation tool ELAN, with the means to create references from tiers and annotations to data categories defined in the ISO Data Category Registry.

The fourth paper was presented by Hennie Brugman et al. about *A Common Multimedia Annotation Framework for Cross Linking Cultural Heritage Digital Collections*. The authors designed an Annotation Meta Model: a RDF/OWL model mainly addressing the anchoring of annotations to segments of the many different media types used in the collections of the archives, museums and libraries involved. The model includes support for the annotation of annotations themselves, and of segments of annotation values, to enable layer annotations.

The last paper was presented by Philippe Blache about *Creating and Exploiting Multimodal Annotated Corpora*. The paper presented a project which aims at collecting, annotating and exploiting a corpus of spoken French in a multimodal perspective. The author presented the annotation schemes used in phonetics, morphology and syntax, prosody, gestuality together with the type of linguistic description made from the annotations seen in two examples.

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

O16 - Biomedical Resources

Su Jian

The *Biomedical Resources* session has 5 papers addressing the latest progress on various topics with biomedical language resource development and evaluation.

An exploitation of term definitions for term matching in need for terminology and ontology construction is reported. A text comparing and clustering tool is used for this task which demonstrates the capability of grouping many related terms using their definitions. Determining the species is crucial to term grounding systems for specific database identifiers as the same terms are often used for different entities across different model organisms. A machine-learning species disambiguation system is compared with a rule-based system using "species indicating words", such as human or marine, where the machine-learning system yielded better overall results on gold-standard datasets.

A corpus for parser evaluation in the biomedical domain is constructed with a 50-abstract subset of the GENIA corpus with labelled head-dependent relations using the grammatical relations (GR) evaluation scheme for comparing parsers that use different formalisms, which has been used for parser evaluation in the newswire domain. A machine learning-based pronoun resolution system is built and evaluated on MUC, ACE, and MEDCO (GENIA coreference corpus) corpus with comparative statistics of those corpora which reveal the noticeable issues in constructing an effective pronoun resolution system for a new domain, but also provide a comprehensive view of those corpora for pronoun resolution. The problem of utilizing multiply annotated data in training biomedical information extraction systems for named entity recognition and relation extraction systems. Several methods of automatically combining the multiple annotations to produce a single annotation are compared, but none produces better results than simply picking one of the annotated versions at random. It is also shown that adding extra singly annotated documents produces faster performance gains than adding extra multiply annotated documents. The above efforts although focusing on the biomedical domain further extend general purpose natural language processing studies which are usually conducted using news articles.

Su Jian
Institute for Infocomm Research
1 Fusionopolis Way, #21 - 01 Connexis
138632 Singapore
sujian@i2r.a-star.edu.sg

O24 - Machine Translation and Multilinguality

Gregor Thurmair

In Machine Translation, the focus of research slowly turns into experiments with hybrid system architectures. The contribution of Carl, in the tradition of the METIS projects, aims at using language models as generation component following a rule-based analysis and transfer phase, and describes efficient search techniques for this purpose.

Itagaki and Aikawa describe a "Term Swapper" which is an approach to extend an SMT output such that existing user glossaries can be integrated by replacing SMT translations by user terminology translations.

The use of multilingual corpora for MT translation extraction is continued to be researched; Babych et al. aim at using corpora for the identification of translation equivalents, extending the scope from parallel to comparable corpora, and exploring issues in overlapping corpora.

Gregor Thurmair
Linguatec Sprachtechnologien GmbH
Gottfried-Keller-Straße 12
81245 Munich, Germany
g.thurmair@linguatec.de

O25 - Evaluation

Robert Frederking

This session on Evaluation concerned primarily the evaluation of parsers and annotations, with one talk on evaluating language identification.

The session began with Jennifer Foster presenting *Parser Evaluation and the BNC: Evaluating 4 constituency parsers with 3 metrics*. The parsers are four versions of the Charniak parser. They are evaluated using a new English test set (hand-corrected BNC parse trees) and three metrics: the Parseval metric, the Leaf Ancestor metric, and a metric based on conversion to labelled dependency trees and counting label and attachment scores. They report that re-ranking gives a modest performance improvement on the new test set; also, self-training with BNC data helps, but not self-training from the North American News Corpus.

Anne Vilnat then presented *EASY, Evaluation of Parsers of French: what are the Results?*. EASY has evaluated syntactic parsers on syntactic phenomena and dependency relations in French. She presented their annotation scheme and newly-developed annotated corpus, and the results of analyzing 15 parsers from 12 teams. A ROVER procedure was also described that automatically combined the outputs of the parsers.

Xavier Tannier presented *Evaluation Metrics for Automatic Temporal Annotation of Texts*. He and his co-author argue that local measures are not sufficient for evaluating the annotation of temporal relations, since a coherent picture only emerges at a global level. They seek to have fair comparisons between automatic systems, based on paying attention to these global issues. Their technique uses recall and precision on a kernel of central relations, and presents "evaluation of the evaluation" measures to support its superiority.

In the outlying data point of this session, Lena Grothe presented *A Comparative Study on Language Identification Methods*. They conducted two experiments with different evaluations comparing language identification algorithms using short words, frequent words, and n-gram-based approaches, as well as two combinations of these. They argue that their work shows the importance of using dynamic values with the "out-of-place" measure.

Finally, Eric Villemonte de la Clergerie presented *PASSAGE: from French Parser Evaluation to Large Sized Treebank*. This project aims at automatically building a large French Treebank by combining the output of several parsers, using the EASY annotation scheme presented earlier in this session. More results of the first evaluation campaign were presented, as well as additional preliminary results obtained from the ROVER procedure that was also mentioned above.

Robert Frederking
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
PA 15213 Pittsburgh, USA
ref@cs.cmu.edu

O28 - Machine Translation

Dan Tufis

The Fifth International Conference on Language Resources and Evaluation (LREC 2008) was definitely an enjoyable event, both scientifically and socially. The papers, presented in the conference as well in the satellite workshops, were very relevant for the state of the art and the main trends in HLT.

Given the large topic coverage of LREC 2008, the hot topic of machine translation has been very well represented both in the oral and poster sessions. This note refers to one of the oral sessions (O28) which contained five interesting papers.

In their contribution *Building Bilingual Lexicons using Lexical Translation Probabilities via Pivot Languages*, Takashi Tsunakawa, Naoaki Okazaki and Jun'ichi Tsujii describe a statistical approach to deriving from two bilingual lexicons, sharing a common language (called the pivot language) a third one among the other two

languages. Their method exploits the recent advances in word and phrase alignment technology and has been used to derive a merged Japanese-Chinese dictionary out of large Japanese-English and Chinese-English technical dictionaries. The resulted dictionary is evaluated in terms of Precision and Mean Reciprocal Rank and the results bring evidence of the effectiveness of the authors' approach over other methods (e.g. *exact matching*).

The second paper presented in this section *Improving Statistical Machine Translation Efficiency by Triangulation* due to Yu Chen, Andreas Eisele and Martin Kay, describes an attempt to reduce the model size by filtering out the less probable entries based on testing correlation using additional training data in an intermediate third language. The central idea behind the approach is triangulation, the process of incorporating multilingual knowledge in a single system, which eventually utilizes parallel corpora available in more than two languages. The experiments conducted by the authors using the Europarl corpus showed that the reduction of the model size for a phrase-based MT system can be up to 70% while the translation quality is being preserved.

Caroline Lavecchia, David Langlois and Kamel Smaïli bring evidence on the advantages of phrase-based models in machine translation and their paper *Phrase-Based Machine Translation based on Simulated Annealing* introduces a new algorithm exploiting the concept of interlingual triggers. These are pairs of source words/phrases (triggering events) and multiple target words/phrases (triggered events) which are best correlated in terms of mutual information with the source triggers. Most often, the correct translation of a source phrase detected as a triggering event is among the target triggered events. The Simulated Annealing algorithm proposed in the paper is used to extend the 1-1 interlingual triggers to N-M triggers, avoiding the phrase alignment step in most Phrase-Based MT systems. The authors report a 7% improvement over a reference state-of-the-art phrase-based approach that required word alignment.

The paper *Evaluation of Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation* by Marine Carpuat and Dekai Wu reports recent results in phrase-based machine translation experiments that demonstrate significant improvements when word sense disambiguation is adequately incorporated in the underlying statistical models. They present a generalisation of the WSD approach called Phrase Sense Disambiguation, underlying the automatic construction of phrasal translation lexicons. Unlike the conventional static phrasal translation lexicons which ignore all contextual information, the authors argue that the dynamically-built context-dependent phrasal translation lexicons are more useful resources for the lexical choice step in phrase-based statistical machine translation. The evaluations based on Chinese-English NIST2004 data showed significant benefits of the described methodology not only on lexical choice, but also on phrasal segmentation.

The last paper of the session, *A multi-genre SMT system for Arabic to French* authored by Sasa Hasan and Hermann Ney, discusses the most recent improvements of a SMT system developed within the TRAMES project, bringing evidence on significant performance improvements (both in quality and speed) over an earlier system. Essentially, the significant progress has been achieved by a more accurate training data preparation, using combined genre-specific models. This result emphasises very well the idea that investing efforts in preparing accurate training data is highly rewarding.

Dan Tufis
Romanian Academy, Research Institute for AI
13, Calea 13 Septembrie
050711 Bucharest 5, Romania
tufis@racai.ro

O42 - Multimodal Session

Kristiina Jokinen

Multimodality has recently gained more attention in Language Technology, and the reason is clear: to achieve a wider understanding of the use of natural language in human communication, it is also important to study all those non-verbal aspects of communication that accompany our verbal expressions when we speak. Also various human-computer applications use different input and output devices, aiming at more natural and robust interaction, and questions concern how to best integrate different multimodal components into dialogue systems, what are suitable architectures and representations, and how the users evaluate such systems.

Multimodal system development is usually a large-scale task which requires cooperative efforts of various expertise. One of the main tasks is to collect data that exemplifies the intended system functions, but also represents typical human communication patterns in a given situation. Recent research has thus brought in the collection of large video corpora of natural conversational speech, their examination through multi-level analysis and annotation, as well as model-building for human-human and human-computer interaction systems.

The increase in interest and research concerning multimodality can also be seen in LREC: workshops and tutorials were organised on related topics, and in the main conference, there were four parallel paper sessions devoted especially to multimodal corpora, annotation tools, and interaction. The oral presentations in the "O-42 Multimodal Dialogue" session dealt with multimodal system architecture, dialogue applications and data collection in the projects such as Companions and FruitCart, as well as corpus collection to study commonalities between computer-mediated communication genres and traditional ones.

The types of corpora discussed ranged from human-human discussions on predefined topics to WoZ-based interactions on photos, and to subjects talking to a computer and manipulating objects on the screen. The initial application prototypes could then be used to generate more conversations. Data annotation was also discussed, from the point of view of

studying language production and comprehension, variation between different genres of communication such as emails, blogs and face-to-face chatting, as well as fusion of speech and pointing input in open domain dialogues with senior people accompanied with photos. The oral presentations were followed by several interested questions and the discussions apparently continued during the break that followed the session.

The technical organisation of the session went fine, and the presentations could proceed smoothly.

Kristiina Jokinen
Department of General Linguistics,
University of Helsinki
PO BOX 9
FIN-00014 Helsinki, Finland
Kristiina.Jokinen@helsinki.fi

O48 - TV and Video Processing

Michael Kipp

The first talk by Einav Itamar was on *Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora*. It was about creating a corpus from a database subtitles using a variant of Gale and Church's (1993) sentence alignment algorithm. This version includes the time information of subtitle occurrence (more specifically: the normalized duration) and linearly combines it with the Gale and Church formula (weight approx. 60%).

The second talk, given by Ron van Son, introduced *The IFADV Corpus: a Free Dialog Video Corpus*, a free corpus of informal, dyadic conversations (Dutch), already transcribed with orthography, POS tags and automatically generated phoneme transcriptions and word boundaries. Some preliminary results on the impact of speaker gaze behaviour on the listener's reaction were presented. The videos do not allow to code gestures (audience question) because the hands are not always fully visible.

The last speaker, Luca Cristoforetti, presented a *WOZ Acoustic Data Collection for Interactive TV*. Under the European DICIT project, researchers recorded sessions (English, German, Italian) where 3 subjects controlled a TV set by voice which is to be used in developing acoustic pre-processing algorithms. Sessions were manually transcribed in the Transcriber tool at the word level, including some acoustic events. Questions addressed whether the setting realistically reflected a living room with acoustic and psychological surroundings. The speaker pointed out that separating walls were set up to make the room dimensions comparable to a standard living room.

Michael Kipp
DFKI, Embodied Agents Research Group
Campus D3.2, Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
kipp@dfki.de

Oral Session

LREC 2008 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 2008 Proceedings, which are available online: www.lrec-conf.org/proceedings/lrec2008

P3 - Syntactically Annotated Resources and Related Tools

Toma Erjavec

This session consisted of 11 poster presentations concentrating on methods or results in the domain of syntactically annotated language resources. The posters discussed a wide array of subjects and languages, from subcategorization to reported speech, and from Bengali to French. Below we briefly summarise the presentations.

Dino Ienco et al. presented experiments on *Automatic extraction of subcategorization frames for Italian*, where statistical subcategorization extraction methods were applied to an Italian treebank annotated with dependency relations, and the results evaluated in a cross-linguistic perspective.

Jerid Francom and Mans Hulden in *Parallel Multi-Theory Annotations of Syntactic Structure* presented an XML encoding, illustrated on the Penn Treebank additionally annotated with Functional Dependency Grammar and Government and Binding style notations, which supports multiple syntactic annotations.

Meni Adler et al. presented *Tagging a Hebrew Corpus: the Case of Participles*, where they argue for a new part-of-speech tagset for Hebrew, and illustrate its application on the *beinoni* forms in Hebrew.

Joydeep Nath et al. in *Unsupervised Parts-of-Speech Induction for Bengali* approached the issue of part-of-speech tag set design by studying the topological properties of word interaction networks, where clustering techniques lead to the induction of natural word classes. They evaluate the approach and argue that it can be easily extended to any language.

Guadalupe Aguado de Cea et al. presented *Tagging Spanish Texts: the Problem of SE*, where, concentrating on the highly ambiguous Spanish particle "se", they took a free annotation tool and modified it in various ways to improve its behaviour.

Jirí Mírovský in *Does Netgraph Fit Prague Dependency Treebank?* presented the query language of Netgraph, which is a graphical tool for searching in the Prague Dependency Treebank (PDT), and demonstrated that it is capable of searching in the most complex layer of PDT, namely the tectogrammatical layer.

Tomas By in *The Kalshnikov 691 Dependency Bank* presented a re-encoding of the PARC 700 Treebank, arguing that the original format of this treebank, meant for parser evaluation, has a number of problems, avoided in the Kalshnikov 691 format.

Natalie Schluter and Josef van Genabith in *Treebank-Based Acquisition of LFG Parsing Resources for French* showed that, with modest changes to the established parsing architectures, encouraging results can be obtained in inductive dependency parsing for French.

Svetla Koeva et al. presented *Chooser: a Multi-Task Annotation Tool*, a tool to assist manual annotation of linguistic data, while Pavlina Fragkou et al. presented *BOEMIE Ontology-Based Text Annotation Tool*, a tool which is able to locate blocks of text that correspond to specific types of named entities, fill tables corresponding to ontology concepts with those named entities and link the filled tables based on relations defined in the domain ontology.

Finally, Ralf Krestel et al. in *Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles* detailed the basic processing steps for reported speech analysis and reported on performance of an implementation in form of a GATE resource.

Poster Session

Toma Erjavec
Dept. of Knowlegde Technologies
Jozef Stefan Institute
Jamova39
SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

P7 - Term Identification / Extraction and Terminological Databases

Maria Gavrilidou

This session aggregated papers on the issues of "Term Identification / Extraction and Terminological Databases". The papers touched upon issues such as re-purposing of tools and resources, evaluation of algorithms and the impact of the types of corpora used for this task and enrichment of terminological resources with various types of information or structures.

The paper presented by Jorge Vivaldi, Anna Joan and Merce Lorente, entitled *Turning a Term Extractor into a new Domain: first Experiences*, discussed the adaptation of a semantic-based term extractor (YATE, supported by EWN) developed for one domain (medical) to a new one (economic).

The paper by Peter Anick, Vijay Murthi and Shaji Sebastian, *Similar Term Discovery using Web Search* presented an approach for discovering semantically similar terms that utilizes a web search engine both as a source for generating related terms and a tool for estimating the semantic similarity of terms.

In the domain of term identification / recognition falls the paper of Junko Kubo, Keita Tsuji and Shigeo Sugimoto *Temporal Aspects of Terminology for Automatic Term Recognition: Case Study on Women's Studies Terms*, which discussed the impact the appearance of a term in a dictionary has on its termhood status, as attested in texts written before and after the dictionary publication; the calculation was based on five automatic term recognition (ATR) measures.

Also in the same domain, the paper *A Comparative Evaluation of Term Recognition Algorithms* by Ziqi Zhang, Jose Iria, Christopher Brewster and Fabio Ciravegna, presented the comparison of five ATR algorithms and proposed a combined approach using a voting mechanism. The 6 different approaches were evaluated using two different corpora.

Another paper that belongs to the same domain is the one by Véronique Hoste, Els Lefever, Klaar Vanopstal and Isabelle Delaere, *Learning-based Detection of Scientific Terms in Patient Information*, which presented the first step towards the automatic replacement of a scientific term by its general language counterpart, with beneficial effects on readability. In this first step, the authors investigated the use of a machine-learning based approach for scientific term detection.

Two papers of session P7 were related to terminological databases. *WNTERM: Enriching the MCR with a Terminological Dictionary* by Eli Pociello, Antton Gurrutxaga, Eneko Agirre, Izaskun Aldezabal and German Rigau presented the construction of WNTERM, a multilingual (Basque and English) light-weight domain ontology (the pilot domain being ecology), which resulted from the merger of the EuroWordNet-based Multilingual Central Repository (MCR) and the Basic Encyclopaedic Dictionary of Science and Technology (BDST).

Finally, Rita Marinelli, Melissa Tiberi and Remo Bindi on their paper *Encoding Terms from a Scientific Domain in a Terminological Database: Methodology and Criteria* report on a project which aims at enhancing a maritime terminological database by means of a set of terms belonging to meteorology.

Maria Gavrilidou
Electronic Lexicography Dept.
Institute for Languages and Speech Processing
Artemidos 6 & Epidavrou
GR - 151 25 Marousi, Greece
maria@ilsp.gr

P13 - Evaluation

Maghi King

Trying to summarize a poster session which included twenty four separate and mostly unrelated presentations is an exercise doomed to failure. All I can do is give the reader a flavour of the topics treated and encourage him to go and look at the proceedings to find out more.

But before I do so, I would like to say that this poster session convinced me for the first time of the real value of heavily populated poster sessions running in parallel with more formal presentations. The atmosphere in the area set aside for the poster session was exciting: there was a constant hub-hub of discussion and argument, and above all, both those presenting posters and those coming to see them seemed to be heavily involved. At the end of the session it was difficult to persuade participants that the allotted time was indeed over: that has to be a mark of success.

About the only general statement one might make about the topics treated is that they were very many and very various. I have tried to group them together in what follows, but am only too aware that my groupings are somewhat arbitrary, and that some presentations could have been put in other groups. There were a couple of presenters who were prevented for reasons beyond their control from being at the session. I have nonetheless included their posters in the summary, basing what I say on the written version.

Relatively few presentations dealt with spoken language. One was concerned with the ability of speech recognition systems to improve their performance by adapting to the language concerned, and with evaluating different techniques for bringing the adaptation about. A second dealt with designing an evaluation for distant talking noisy speech recognition under hands free conditions, and presented an evaluation framework which is already being used in Japan. The final contribution in this group

was concerned with speech recognition and speech synthesis, but as components of a larger system. The authors were describing the end to end evaluation of a speech to speech translation system.

A second group of presentations dealt with particular aspects of the analysis of written language: posters in this group talked of the evaluation of word segmenters (for Vietnamese), part of speech taggers for French, parsers for Italian and named entity recognition where the entities to be recognized were diseases mentioned in clinical notes. One presentation dealt with the evaluation of text summarization applications and the interplay between term definition and the scoring a sentence would receive. On the computer support side, one poster evaluated different virtual keyboards used for the input of West-African languages.

A number of papers described evaluation campaigns, or, on a more modest scale, exercises. Thus one poster described EVALITA, a campaign devoted to the evaluation of NLP tools for Italian, and another described the initial steps in setting an exercise for evaluating anaphora resolution systems. A rather novel competition presented was a competition to clean up web pages, with the idea of using the cleaned up content as a corpus.

Quite a large group of presentations was concerned, in one way or another, with the development and evaluation of linguistic resources. Thus, one poster was concerned with assessing the cost of annotating a corpus, whilst another discussed the evaluation across different types of texts of a specific information structure annotation scheme. In a somewhat similar vein, one presentation compared the vocalized and the unvocalized versions of the LCD Arabic annotated tree bank. Several presentations were concerned with the creation of test and/or training data. The first described the creation of a large collection of aligned Japanese-English sentences for use in training and evaluating machine translation systems working in the domain of patents. Word-alignment was the topic of a presentation dealing with various facets of Portuguese to English word alignment. A fourth poster concerned the creation of a gold standard data set for use in the evaluation of a named entity recognition system working over the recognition of diseases in clinical notes, whilst a fifth described the creation of a ground truth data set of culturally diverse Romanized names destined for use in the evaluation of name matching systems. A fourth presentation described the development of a test suite intended for use in the evaluation of systems studying inference problems shown by English adjectives. Finally, a poster described the evaluation of inter-sentential co-reference annotation in the context of manual construction of a semantic network.

A final group of presentations dealt with lexical and semantic aspects of NLP, always within the perspective of evaluation or the creation of resources for use in evaluation. Thus, one poster talked about creating and using an artificially constructed corpus of low frequency words and their associations in order to evaluate whether use of the associations helped in finding the low frequency target word. Another offered a comparative study of the semantic and lexical relations defined and adapted for WordNet and for EuroWordNet. A third presentation described using a model of speech acts present in e-mails in order to categorize e-mail content into a set of speech acts, each with a set of associated expectations.

Maghi King
University of Geneva
28 rue des Bossons
CH-1213 Geneva, Switzerland
maghi.king@gmail.com

P14 - Evaluation: Resources, Tools, Systems, Methodologies

Diana Santos

The P14 session, on *Evaluation: Resources, Tools, Systems, Methodologies*, comprised 30 papers and illustrates well the ingenuity required for evaluation and the different directions and studies that are being actively produced in our community.

The session encompassed a wide range of contributions, from those aiming at general discussion and philosophical questions to those dealing with very specific practical problems. General application areas covered were as diverse as dialog systems, machine translation, sentiment analysis, question answering, text generation and summarization.

A fair amount of user-transparent tasks (Gaizauskas, 1998[(1)]) were also dealt with, such as word alignment, multicultural name matching, and multiword detection, while a few papers concentrated on eliciting user data or compiling evaluation resources. What one could call meta-evaluation (discussing and improving evaluation methods or resources) was well represented, too, with papers dedicated to improvements to measures or description of systems, frameworks or methodologies for evaluation. Information from the last developments and evaluation setup in ACE was also reported.

As to languages covered, however, it appears that the main bulk of evaluation is still done on English, or related to English (as the interesting task of "English inclusion detection" shows, or when translation or alignment is at stake), although a fair amount of papers were general enough (or too general) to even deal with any specific language. There were four papers concerning French, three Spanish and one paper each dealing with Arabic, Turkish and Basque. Interestingly, these "other-languages" papers were almost only in the areas of (speech) dialogue systems, often QA systems.

Diana Santos
SINTEF ICT
Box 124 Blindern
N-0314 Oslo, Norway
Diana.Santos@sintef.no

(1) *Gaizauskas, Robert. "Evaluation in language and speech technology".* Computer Speech and Language, *12 (4) (1998), pp.249-62.*

Internet Area

P19 - Morphology, Syntax and Tools

Patrick Paroubek

This poster session on "Morphology, Syntax and Tools" took place in the hall of the conference center, where 16 posters were presented. The high ceiling of this relatively open space contributed to bring down the noise generated by the large crowd that was attending the session to a reasonable level and lively discussions took place around the posters. They displayed a wide range of works from morphosyntax up to semantics, with a poster addressing the construction of a prop bank for Arabic, giving thus a good picture of the field.

One could mainly distinguish three groups of presentations, among which the first was centered on linguistic information acquisition and had 4 posters addressing: (1) the unsupervised lexical acquisition of POS tags for unknown words through generation and morphology applied to Romanian, (2) a hybrid approach (statistics/linguistics) for extracting verb-noun constructions from a German corpus, (3) the semi-automatic categorization properties of nominal compounds and finally (4) the learning of noun phrases properties from data to functions by means of self organization maps, applied to Italian.

The second group focused on treebanks with six posters for various languages, like Latin, Slovene, German, Italian, Croatian and Arabic.

The third dealt with text analysis with: (1) a presentation of a uniform formalism for partial parsing and morpho-syntactic disambiguation applied to Polish, (2) the annotation of superlatives with a corpus from Wikipedia, a useful contribution for the emerging field of opinion mining, (3) unsupervised POS tagging for unsupervised parsing especially good for noun phrases, (4) an underspecified representation of syntactic annotations design to handle ambiguity in a human friendly way and (5) a study of parenthetical in discourse corpora with examples from the Penn Treebank.

Last but not least, the support provided to the presenters by the organization was faultless and the Moroccan cookies at the coffee break were delicious.

Patrick Paroubek
LIR Group
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
pap@limsi.fr

P21 - Tools and data for Speech Systems Developments

Ryszard Gubrynowicz

The posters presented during this session were referring to three main subjects: (a) large corpora design and recording, (b) tools for speech systems, and (c) development of dialog systems.

(a) Different kinds of corpora were designed and recorded, most of them built for speech synthesis (TTS). Some of them referred to a given language (like European Portuguese or Czech), the other had a multi-language option. Two were designed for automatic speech recognition engines (ASR). There were also presentations of some specific databases such as phonetic databases (pronunciation databases) like for Japanese Dutch, Welsh, or French), and lexical databases (composed of proper names, like geographical, person, place, organization names, etc.)

(b) Various tools for ASR, speech synthesis - TTS, WOZ and machine translation were the subject of 5 papers, among them the important communication of the ECESS consortium which is developing an important framework allowing remote evaluation of software modules and tools applied in TTS, via the Web.

(c) Several communications took up problems related to designing and developing dialog systems, such as annotation, tagging, extraction of concepts, application of morphologic knowledge, etc., with particular focus on spontaneous telephonic speech (three of them done within the LUNA project supported by the EU).

Ryszard Gubrynowicz
Department of Multimedia
Polish-Japanese Institute of
Information Technology
Koszykowa St. 86
02-008 Warsaw, Poland
Ryszard.Gubrynowicz@ippt.gov.pl

P23 - Speech Corpus in Various Languages

Briony Williams

There were seven posters in session P23 on the morning of Friday 30th May. They covered a range of European languages, as follows:

- Official languages: Standard German, Austrian German, Polish, European Portuguese, Estonian
- Less-resourced languages: Catalan, Luxemburgish

The projects were innovative; in some cases demonstrating how much can be done even with few resources and little previous work, as is the case for Estonian. The Catalan speech corpus was specifically aimed at speech synthesis, showing that Catalan is one of the stronger less-resourced languages. The work on Luxemburgish, on the other hand, was the very first attempt at documenting and developing language and speech resources for this language.

The corpus of spontaneous speech for European Portuguese (Corp-ORAL), which is still under development, formed a test-bed for an interesting general-purpose tool, “Spock”, which was presented in the neighbouring poster session. This tool makes possible web-based access to an “audio concordancer”, which allows the user to search an annotated speech corpus, view the keyword-in-context search results, and then listen to the audio recording fragments in the search results. The Corp-ORAL annotated recordings are the first recordings to be made publicly searchable using this method.

Overall, the projects presented made it clear that the focus of speech corpus work has now shifted from read speech to spontaneous speech, and from general linguistic corpora to special-purpose corpora (such as for speech synthesis or dialogue research).

Briony Williams
Canolfan Bedwyr
University of Wales, Bangor
Bryn Haul, Victoria Drive
LL57 2EN Bangor, Gwynedd, United Kingdom
b.williams@bangor.ac.uk

P26 - Semantics, Semantic Resources and Semantic Annotation

Costanza Navarretta

Most of the papers in session P26 dealt with the creation, extension, exploitation and/or evaluation of semantic resources. More specifically the following sub themes could be recognised:

- the (semi-)automatic creation or extension of semantic resources reusing information extracted from monolingual and multilingual linguistic resources, such as annotated corpora, parallel corpora, word nets, computational lexica and dictionaries;
- the construction and use of a corpus annotated with information about argumentative reasoning and a corpus annotated with events and relations between events to be used in co-reference resolution;
- the evaluation of semantic resources via other resources;
- the analysis of texts using semantic resources and the application of probabilistic models of context to discover unexpected uses of words;
- the construction of an ontology to be used in multilingual search of domain texts, and the extension of an ontology describing the Chinese character system to account for the Japanese Kanji variation of the same system.

Costanza Navarretta
Center for Sprogteknologi
University of Copenhagen
Njalsgade 140-142, bygn. 25
DK-2300 Copenhagen S, Denmark
costanza@hum.ku.dk

P27 - Temporal Annotation

Dan Cristea

Five papers were contributed in the poster session P37 on "Temporal Annotation". The main interest in these papers was the automatic detection of temporal information in texts, a direction which seems to have more and more applications in information retrieval, question-answering and deep reasoning based on grasping the texts' content.

The way temporal information could be detected in Web pages and why Web pages are different from plain texts in this respect is approached by Stéphanie Weiser, Philippe Laublet and Jean-Luc Minel. They have used a symbolic approach in which patterns and rules are described to detect the temporal information. The temporal expressions to be extracted are classified into two kinds: concerning one particular event and concerning repetitive temporal information.

Sebastian Gottwald, Matthias Richter, Gerhard Heyer and Gerik Scheuermann report on WCTAnalyze, a tool for storing, accessing and visually analyzing collections of temporally indexed data, a research shown to be motivated by applications in media analysis and business intelligence, where higher level analysis should be performed on top of linguistically and statistically processed unstructured textual data. WCTAnalyze enables an efficient and effective way to explore chronological text patterns of word forms, their co-occurrence sets and set intersections, as well as to discover temporally related entities.

MiniSTEx, a spatiotemporal annotation system able to handle temporal and geospatial information as expressed in texts is presented by Ineke Schuurman. Under development at the moment, when ready the system will be able to fill in a database with time and space information to be used for automatic temporal and geospatial reasoning. Although originally developed for Dutch, MiniSTEx incorporates principles that will make it adequate for multilingual applications as well as to times and locations that are not compatible with the present day maps and calendars.

An empirical approach to the identification and resolution of temporal expressions in Spanish news corpora is reported by María Teresa Vicente-Díez, Doaa Samy and Paloma Martínez. This work presents a typology of time expressions based on an empirical inductive approach, both from a structural perspective and from the point of view of their resolution.

Georgiana Puscasu and Verginica Barbu Mititelu report on an experiment aimed to annotate all the 11,306 English verbs included in WordNet 2.0 with event classes. Two annotators assigned each verb present in WordNet the most relevant TimeML event class, as captured by the most frequent verb's meanings. It is clear that this valuable work opens new perspectives to the problem of classification of events into time classes in normal texts, but interesting results will appear mainly after fine-tuning temporal classes on verb senses.

Dan Cristea
Faculty of Computer Science
"Alexandru Ioan Cuza"
University of Iasi
16, Gen. Berthelot St.
700483 Iasi, Romania
dcristea@info.uaic.ro

Oral Session

P28 - Multilinguality and Machine Translation

Elliott Macklovitch

The poster session that I had the honour of chairing at LREC 2008 included no fewer than 27 papers and was very well attended. Unfortunately, with so many papers being presented simultaneously, there was no way I could possibly listen in to every speaker, and there's no room here for me to do justice to every paper. What I will attempt to do instead is draw out certain dominant themes which seem to re-occur in the papers and which may be indicative of general trends in our field.

Not surprisingly for an LREC conference, many of the presentations in this session focussed on the construction, or exploitation of linguistic resources for multilingual applications; and a recurrent theme was the need to minimize the amount of human effort required to construct those resources. The value of parallel corpora as a knowledge source for multilingual applications is now widely acknowledged. Indeed, the term 'parallel text' appears in the title of more than half a dozen papers in this session, with some proposing novel sources for these texts, e.g. movie subtitles (Tiedemann) or revisions to draft translations (Abekawa & Kageura). Other papers describe various machine learning techniques that now permit the desired linguistic knowledge to be directly inferred from the parallel data, ideally with minimal supervision, e.g. (Claveau) and (Spreyer et al.). Automatic alignment, which renders translational correspondences explicit across the two parallel texts, is generally a prerequisite to such higher-level processing. At the sentence level, this now seems to be a largely resolved problem (Maeda et al.); alas, the same cannot be said for finer grained word alignments, particularly across languages that are typologically different (Megyesi et al.).

One issue that appears to be receiving increased attention is the development of linguistic resources for medium- or low-density languages - roughly, all those other than the major European languages plus Chinese, Japanese and Korean. At least five of the papers presented in this session explicitly address this question. For anyone interested in the preservation of linguistic diversity, this is certainly a welcome development, since these low- and medium-density languages are spoken by more 50% of the world's population, something I learned from (Halácsy et al.). Actually, I found this five-page article to be a real eye-opener and would heartily recommend it to anyone. Using the examples of Uzbek and Kurdish, the authors clearly lay out all the difficulties routinely encountered in trying to assemble minimal linguistic resources - a modest sized monolingual corpus, a dictionary of at least 10k stems, a parallel corpus, along with basic supporting software like a segmenter and POS tagger - for so many languages in the non-industrial world. The picture that emerges is certainly a sobering one, although not entirely hopeless. What is called for is a certain amount of ingenuity, of the sort we find manifested in other papers presented in this session. One strategy which I found particularly intriguing[1] is to induce a bilingual lexicon for a new pair of languages via transitivity, using an intervening and better resourced language as a bridge; cf. (Kaji et al.) and (Nerima & Wehrli).

All in all, this was a very stimulating session!

Elliott Macklovich
Laboratoire RALI, Dept. d'informatique et de recherche opérationnelle
Université de Montréal, C.P. 6128, succursale Centre-ville
H3C 3J7 Montréal, Quebec, Canada
macklovi@iro.umontreal.ca

(1) *First proposed, as far as I know, by Mann & Yarowsky in their NAACL 2001 paper "Multipath Translation Lexicon Induction via Bridge Languages".*

P30 - Sentiment and Opinion Analysis

Diana Inkpen

This session presented research on resources and methods for various opinion analysis and generation tasks. Resources included affect lexicons built from domain-specific corpora, lexicons of dirty words, and an opinion-annotated corpus for Chinese product reviews. Analysis tasks included extraction of concepts such as products and their attributes, and their corresponding polarities. Generation of opinionated text was proposed in a system that generates jokes, and in a system that changes the valence of existing texts.

Diana Inkpen
University of Ottawa
School of Information Technology and Engineering
800 King Edward
K1N 6N5 Ottawa, ON, Canada
diana@site.uottawa.ca

Publisher

Area

LREC 2008 Conference Report

Speech given by LREC Programme Committee

Closing Session with LREC Programme Committee:
Jan Odijk, Stelios Piperidis, Bente Maegaard, Nicoletta Calzolari, Khalid Choukri, Daniel Tapias, Joseph Mariani

This year, the Programme Committee has decided to shorten the Conference Reports on each field that used to be presented during the Closing Session and make it one where each member of the Committee has chosen to highlight some trends in the field, in a subjective way.

On the **General Issues**, many papers have been submitted on the infrastructures and also on initiatives that are starting. A lot of papers were dedicated to metadata and standards, which was also the case for articles on Linguistic Web Services of numerous types. Finally, we have seen a significant number of articles dealing with sustainability which is important for our field. All this leads us to think that we are finally starting to converge and open all our resources. We hope that this trend will be confirmed at the next LREC.

With regards to **Written Resources**, the 6th edition of LREC has shown an evident consolidation of know-how and best practices in multilevel corpus processing and annotation, and also in building corpora for various purposes. This seems to extend to new national and regional languages not sufficiently covered in the past. Several other trends can be highlighted:

- The semantic resources, annotation and ontological engineering are now almost at cruising speed with 60 papers accepted at this LREC.

- Multilingual information processing and machine translation have both undergone a considerable progress with 70 papers.

- Out of the new areas, two trends emerge and receive considerable attention: spatiotemporal annotation and sentiment analysis and emotion.

In the written area, the future challenges are mostly focused on the robustness of language technology with regards to user-generated content in the Web2.0 (this will need to be investigated), and the intelligent integration/interfacing of language with other media (image, audio, video, …) and modalities in content processing and man-machine interfaces.

We should also carefully look into the language resources and technology in cognition and perception analysis. To this extent, **Multiword Expressions** is an interesting topic in itself, but not only since proper identification and processing are essential for Machine Translation, Question Answering or Information Retrieval. Around 20 papers were presented at LREC in addition to a workshop dedicated to this topic. This is promising, but more work in this area is desirable and needed.

The evolution of the **Spoken and Multimodal Resources**, with the specialization of databases as a new trend is clearly shown in the tables next page.

Concerning the **Tools, Platforms and Methodologies**, an important effort in the development of annotation methodologies, platforms and tools for speech and multimodal resources has been done. However, some thinking and some actions should be taken to standardize annotation schemes, information structure or methodologies to easily reuse and improve the already available and future resources and to guarantee the quality of the resources.

Spoken Language Evaluation is a very active field and many papers addressed evaluation principles, evaluation metrics and even evaluation campaigns on evaluation metrics. We should stress the increasing number of languages, especially if we refer to Broadcast News that now covers a variety of new languages such as Czech, Khmer, Thaï, Norwegian, or Arabic. In this field, it is important that data are provided in large quantity and good quality in order to develop acceptable systems. For Norwegian, for instance, 1,000 working hours are required to transcribe 1 hour of speech.

Increasingly complex evaluations are run on increasingly complex systems. This is the case for Speech-to-Speech translation (in dialogs), with Automatic Speech Recognition, Machine Translation and Text to Speech synthesis, even if no use is being made of prosody neither for analysis, not for generation. We have also noted an interest in comparing automated evaluation to human assessment. This complexity can also be found in Question

Traditional Databases

	2008	2006
Dialogue Corpora	30%	24%
Emotional Speech	10%	6%
Noisy Conditions	10%	4%
Broadcast News	8%	14%
Telephone Speech	6%	7%
Text to Speech	6%	20%
Non-native	4%	14%

Specialization of Databases

	% papers
Age Groups	30%
Verbal and non Verbal	10%
Medical/Legal/Lectures	10%
Sign Languages	8%
Biometric DBs for Speaker Identification/Verification	6%

Evolution of the Spoken and Multimodal Resources

Answering on speech in meetings, including conversational Speech Recognition and Q&A. Multi-party dialogs, with spontaneous Speech Recognition and Spoken Language Dialogs and evaluation of the detection of language disfluencies, are also concerned.

Finally, the Historic speech collection retrieval, including Automatic Speech Recognition, Information Retrieval and User Interface Evaluation should be quoted. Here, from a system developed for Broadcast News to Cultural Heritage application, the recognition rate (WER) has increased from 30% to 70 % and a new metric RIA (Rank Index Accuracy) has been proposed.

The Evaluation topic also addresses **Evaluation campaigns**, with several national or international campaigns such as Transtac, GALE, CLEF, Technolangue, STEVIN, CENSREC, and **Evaluation projects** such as TC-Star, CHIL or Quaero where evaluation is really the core of the project.

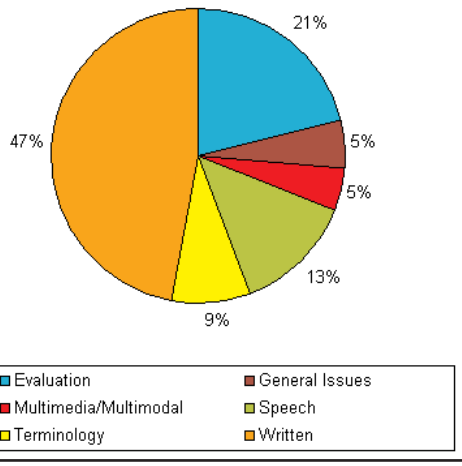
The significant number of submissions in the **Evaluation** track shows that it has become a strong component of the LREC conference, probably the only place where the topic is discussed as such. Evaluation mirrors the field because everything is being evaluated, from simple components to applications. We are evaluating semantics and various aspects of semantics. Machine Translation evaluation is developing and other multilingual applications such as Information Extraction, Information Retrieval or Q&A are strong and evolving.

This field is clearly maturing and we are beginning to agree more on metrics. But more solid work on the metrics and methodologies needs to be done and to this extent, the new project started by NIST on Metrics and Evaluation Methodologies for Machine Translation (NIST-MATR08) will be followed carefully.

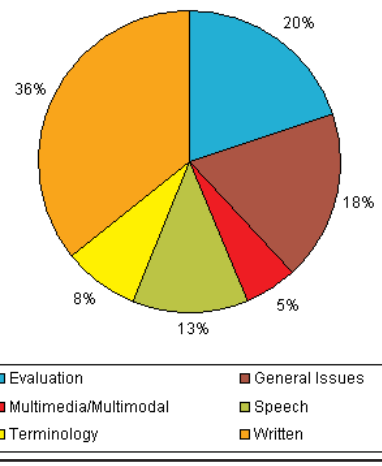
To conclude, the **Conference Statistics** (see graphs below) show that the distribution over the various tracks Evaluation, General Issues, Multimodal/Multimedia, Speech, Terminology and Written have remained stable in 2008 compared to 2006. The only difference this time is that many papers were erroneously submitted under the General Issues track whereas they actually belonged to the Written track.

We would have liked more submissions on Speech and hope to see more papers in this track for LREC 2010.

Submissions in 2006



Submissions in 2008



NEW RESOURCES

Monolingual Lexicon from the general domain

ELRA-L0085 euLEX (Lexical Database for Basque)

euLEX is a general lexicon which contains 115,000 entries, divided into 94,000 dictionary entries or lemmas, 12,000 allomorphs, 7,500 verb forms and about 1,200 dependent morphemes. All entries include linguistic information such as morphology and usage. The lexicon is in XML.

	ELRA members	Non-members
For research use by academic organisations	3,000 Euro	6,000 Euro
For research use by commercial organisations	6,000 Euro	10,000 Euro
For commercial use	15,000 Euro	20,000 Euro

Bilingual Lexicons from the general domain

ELRA-M0043 Russian => English MT optimized lexicon in OLIF XML

This lexicon is provided in structured XML of OLIF (Open Lexicon Interchange Format) format. It comprises 99,211 entries in its source language (Russian) and 134,828 entries in its target language (English). The source entries are distributed as follows: 64,487 nouns, 11,470 adjectives, 19,724 verbs, 1,762 adverbs, and 1,768 closed-class elements (interjections, special prefixes, suffixes, etc.). Nouns contain gender and number information and verbs provide details on aspect and reflexivity. The entries contain semantic information in terms of domain specification or style information (e.g., colloquial, regional use, etc.). Moreover, definitions are available for 59,775 entries, as well as collocational information for 39,148 entries.

	ELRA members	Non-members
For research use	1,000 Euro	2,000 Euro
For commercial use	6,100 Euro	8,550 Euro

ELRA-M0044 English => Swahili Bilingual Lexicon

This lexicon is provided in structured XML of OLIF (Open Lexicon Interchange Format) format. It comprises 58,247 entries in English and 58,300 in Swahili. The source entries are distributed as follows: 36,046 nouns, 3,013 adjectives, 18,308 verbs and 880 closed-class entries. The entries contain semantic information in terms of domain specification or style information (e.g., colloquial, regional use, etc.). Collocational information is also available for 17,570 entries.

	ELRA members	Non-members
For research use	45 Euro	45 Euro
For commercial use	45 Euro	45 Euro

ELRA-M0045 Cebuano => English Bilingual Lexicon

This lexicon is provided in structured XML of OLIF (Open Lexicon Interchange Format) format. It comprises 1,988 entries in Cebuano and 1,990 in English. The source entries are distributed as follows: 1,052 nouns, 462 adjectives, 405 verbs and 69 closed-class entries. The entries contain semantic information in terms of domain specification or style information (e.g., colloquial, regional use, etc.). Collocational information is also available for 500 entries.

	ELRA members	Non-members
For research use	45 Euro	45 Euro
For commercial use	45 Euro	45 Euro

ELRA-M0046 English => Czech Bilingual Lexicon

This lexicon is provided in structured XML of OLIF (Open Lexicon Interchange Format) format. It comprises 31,718 entries in English and 32,125 in Czech. The source entries are distributed as follows: 17,797 nouns, 7,748 adjectives, 6,039 verbs and 134 closed-class entries. The entries contain semantic information in terms of domain specification or style information (e.g., colloquial, regional use, etc.). Collocational information is also available for 3,065 entries.

	ELRA members	Non-members
For research use	45 Euro	45 Euro
For commercial use	45 Euro	45 Euro

Speech Telephone Databases

ELRA-S0242 SALA II US English database

The SALA II US English database comprises 4,090 US English speakers (2,017 males, 2,073 females, including some speakers with Hispanic accents) recorded over the United States mobile telephone network.

	ELRA members	Non-members
For research use	55,000 Euro	60,000 Euro
For commercial use	60,000 Euro	75,000 Euro

ELRA-S0272 MEDIA speech database for French

The MEDIA speech database for French was produced by ELDA within the French national project MEDIA (Automatic evaluation of man-machine dialogue systems), as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT). It contains 1,258 transcribed dialogues from 250 adult speakers. The method chosen for the corpus construction process is that of a 'Wizard of Oz' (WoZ) system. This consists of simulating a natural language man-machine dialogue. The scenario was built in the domain of tourism and hotel reservation.
The semantic annotation of the corpus is available in the catalogue and referenced ELRA-E0024 (MEDIA Evaluation Package).

	ELRA members	Non-members
For research use	1,000 Euro	2,000 Euro
For commercial use	5,000 Euro	10,000 Euro

ELRA-S0277 SpeechDat Galician Database for the Fixed Telephone Network

The SpeechDat Galician Database for the Fixed Telephone Network contains the recordings of 653 speakers of Galician recorded over the fixed telephone network. Each speaker uttered around 44 read and spontaneous items.

	ELRA members	Non-members
For research use	7,000 Euro	15,000 Euro
For commercial use	12,000 Euro	18,000 Euro

ELRA-S0281 LILA Hindi-L1 database

The LILA Hindi-L1 database comprises 2,030 Hindi speakers (1,012 males and 1,018 females, all speakers with Hindi as first language) recorded over the Indian mobile telephone network. Each speaker uttered around 60 read and spontaneous items.

	ELRA members	Non-members
For research use	40,000 Euro	50,000 Euro
For commercial use	50,000 Euro	65,000 Euro

Speech Microphone Databases

ELRA-S0276 Swedish EUROM1 (EUROM1_S)

EUROM1 is the first really multilingual speech database produced in Europe. Over 60 speakers per language pronounced numbers, sentences, isolated words using close talking microphone.

	ELRA members	Non-members
For research use	800 Euro	1,600 Euro
For commercial use	800 Euro	1,600 Euro

ELRA-S0278 SmartWeb Handheld Corpus (SHC)

This corpus contains recordings spoken by 156 speakers in a human-machine query situation. Users were asked to solve several tasks with a spoken query system to the WWW using a smart phone as portable device in natural environments (office, hall, restaurant, street). Recorded channels are the Bluetooth headset over UMTS (telephone quality), the Bluetooth headset and an additional collar microphone in high quality.

See also ELRA-S0279 and ELRA-S0280.

	ELRA members	Non-members
For research use	1,912.50 Euro	3,825 Euro
For commercial use	2,912.50 Euro	4,825 Euro

ELRA-S0279 SmartWeb Motorbike Corpus (SMC)

This corpus contains recordings spoken by 36 speakers in a human-machine query situation on a running motor cycle (BMW). Bikers were asked to solve several tasks with a spoken query system to the WWW using an integrated system connected to a speech server via an UMTS connection. Recorded channels are the Bluetooth helmet microphone over UMTS (telephone quality), and - partly - the Bluetooth helmet microphone and an additional neck microphone in high quality.

See also ELRA-S0278 and ELRA-S0280.

	ELRA members	Non-members
For research use	382.50 Euro	765 Euro
For commercial use	382.50 Euro	765 Euro

ELRA-S0282-01 BAS PHATT 1.0.X (sub-set)

The Ph@ttSessionz speech database contains recordings of 864 adolescent speakers of German (age range 12-20). The recordings were performed via the WWW in public schools (Gymnasium) in 41 locations in Germany. Recordings were done with SpeechRecorder in selected schools in the years 2005-2007. Both channels, the headset and the desktop microphone, were recorded in high quality.

The BAS PHATT corpus is available in two versions: BAS PHATT 1.0.X (sub-set, ELRA-S0282-01) and BAS PHATT 1.1.X (complete corpus, ELRA-S0282-02).

BAS PHATT 1.0.X contains 41 items.

See also ELRA-S0082-02.

	ELRA members	Non-members
For research use	512 Euro	512 Euro
For commercial use	2,512 Euro	3,512 Euro

ELRA-S0282-02 BAS PHATT 1.1.X (complete corpus)

The Ph@ttSessionz speech database contains recordings of 864 adolescent speakers of German (age range 12-20). The recordings were performed via the WWW in public schools (Gymnasium) in 41 locations in Germany. Recordings were done with SpeechRecorder in selected schools in the years 2005-2007. Both channels, the headset and the desktop microphone, were recorded in high quality.

The BAS PHATT corpus is available in two versions: BAS PHATT 1.0.X (sub-set, ELRA-S0282-01) and BAS PHATT 1.1.X (complete corpus, ELRA-S0282-02).

BAS PHATT 1.1.X contains 138 items.

See also ELRA-S0082-01.

	ELRA members	Non-members
For research use	1,917.30 Euro	3,834.75 Euro
For commercial use	3,917.30 Euro	6,834.75 Euro

Speech Broadcast News

ELRA-S0275 Slovenian BNSI Broadcast News Speech Corpus

This speech database consists of TV news shows (both evening news, “TV Dnevnik” and late night news, “Odmevi”), from the archive of a Slovenian national broadcaster RTV Slovenia. The recordings took place between June 1999 and May 2003. The database comprises a total of 36 hours of recordings, transcribed and manually checked using the Transcriber tool. 1,565 speakers were recorded (1,069 males, 477 females, 19 unspecified).

	ELRA members	Non-members
For research use	6,000 Euro	10,000 Euro
For commercial use	19,000 Euro	33,000 Euro

Speech Related Databases

ELRA-S0268 UPC-TALP database of isolated meeting-room acoustic events

This database has been produced within the CHIL Project (Computers in the Human Interaction Loop), in the framework of an Integrated Project (IP 506909) under the European Commission’s Sixth Framework Programme. It contains a set of isolated acoustic events that occur in a meeting room environment and that were recorded for the CHIL Acoustic Event Detection (AED) task. The database can be used as training material for AED technologies as well as for testing AED algorithms in quiet environments without temporal sound overlapping. Approximately 60 sounds per sound class were recorded. Ten people (5 men and 5 women) participated in three sessions. During each session a person had to produce a complete set of sounds twice.

	ELRA members	Non-members
For research use	500 Euro	600 Euro
For commercial use	1,000 Euro	1,200 Euro

ELRA-S0269 LC-STAR Greek Phonetic lexicon

The LC-STAR Greek Phonetic lexicon comprises 110,708 entries, including a set of 57,519 common words, a set of 45,162 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 8,027 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	21,250 Euro	27,625 Euro
For commercial use	28,000 Euro	36,400 Euro

ELRA-S0270 LC-STAR Italian Phonetic lexicon

The LC-STAR Italian Phonetic lexicon comprises 109,712 entries, including a set of 56,420 common words, a set of 45,253 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 8,039 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	14,250 Euro	21,500 Euro
For commercial use	22,000 Euro	29,250 Euro

ELRA-S0271 LC-STAR English-Italian Bilingual Aligned Phrasal lexicon

The LC-STAR English- Italian Bilingual Aligned Phrasal lexicon comprises 10,466 phrases from the tourist domain. It is based on a list of short sentences obtained by translation from a US-English 10,524 phrase corpus. The lexicon is provided in XML format.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

ELRA-S0273 LC-STAR Slovenian Phonetic lexicon

The LC-STAR Slovenian Phonetic lexicon comprises 110,900 entries, including a set of 64,521 common words, a set of 45,012 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 5,491 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	15,250 Euro	23,250 Euro
For commercial use	23,000 Euro	31,500 Euro

ELRA-S0274 LC-STAR English-Slovenian Bilingual Aligned Phrasal lexicon

The LC-STAR English-Slovenian Bilingual Aligned Phrasal lexicon comprises 12,722 phrases from the tourist domain. It is based on a list of short sentences obtained by translation from a US-English 10,522 phrase corpus. The lexicon is provided in XML format.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

Multimodal Database

ELRA-S0280 SmartWeb Video Corpus (SVC)

This multimodal corpus contains 99 recordings each containing a human-human-machine dialogue: one speaker (which is being recorded) interacts with a human partner as well with a dialogue system via a smart phone (SmartWeb system).

See also ELRA-S0278 and ELRA-S0279.

	ELRA members	Non-members
For research use	635 Euro	1,275 Euro
For commercial use	1,635 Euro	2,275 Euro