The ELRA Newsletter

April - September 2006

Vol.11 n.2 & 3

LREC 2006

Special Issue

5th 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

Introduction

Nicoletta Calzolari, Conference Chair *Page 4*

Opening Ceremony Speeches

Bente Maegaard, ELRA President *Page 7*
Khalid Choukri, ELRA CEO *Page 8*

Antonio Zampolli Prize Award Ceremony

Bente Maegaard *Page 10*

Oral and Poster Session Summaries

Semantics & Infrastructural Issues
Dan Tufis *Page 11*

Multilingual Corpora
Steven Krauwer *Page 12*

Question-Answering
Bernardo Magnini *Page 12*

Tools and Evaluation
Mary P. Harper *Page 12*

Lexicon and Pronunciation
Ute Ziegenhain *Page 13*

Morphology & Tagging
Kiril Simov *Page 14*

Anaphora & Coreference
Véronique Moriceau *Page 15*

Corpora and LR Infrastrucures
Sonja Bosch *Page 15*

Corpora: Creation, Annotation
Costanza Navaretta *Page 15*

Web Services and Digital Archives and Libraries
Zygmunt Vetulani *Page 16*

Translation
Nelleke Oostdijk *Page 16*

Workshop & Panel Reviews

Workshop on "SALTMIL SIG: Speech and Language Technology for Minority Languages"
Briony Williams *Page 17*

Workshop on "Crossing Media for Improved Information Access"
Stelios Piperidis *Page 18*

Panel on "Resources for the Processing of Affect in Interaction"
Nick Campbell *Page 18*

LREC 2006 Reports

Report on Papers on Evaluation for Spoken and Multimodal Communication
Joseph Mariani Page 19

Report on Spoken Language Resources and Multimodality
Daniel Tapias and Khalid Choukri *Page 20*

Report on Written Language Resources
Jan Odijk *Page 21*

Report on Semantic-Related Papers
Aldo Gangemi *Page 23*

New Resources *Page 24*

Dear Colleagues,

The 5th edition of the Language Resources and Evaluation Conference took place last May in Genoa, Italy. 800 participants from over 44 countries attended this fruitful and milestone event in HLT, with its rich and varied conference programme.

More than 800 submissions for poster and oral presentations were reviewed by the Scientific Committee and 512 were actually presented in Genoa. A majority (around 45%) of the papers were dedicated to the Written area. 20% of the presentations were dedicated to Spoken and Multimodal issues and 20% to Evaluation. Less than 10% of the articles dealt with the terminological issues.

In addition, a total of 18 satellite workshops and tutorials covering various fields were organised before and after the main conference. These workshops covered topics as diverse as minority languages processing, annotation science, corpora for research on emotion, terminology design, semantic web technologies, speech corpus production and validation, and for the second time after LREC 2004, the representation and processing of sign languages.

The last workshop held at LREC 2006, *COCOSDA/WRITE Roadmap for Language Resources and Evaluation in a Multilingual Environment*, was a joint meeting between COCOSDA, the International Committee for Co-ordination and Standardisation of Speech Databases, and WRITE, the International Committee for Written Resources Infrastructure, Technology, and Evaluation. As a follow-up of the successful Joint COCOSDA and WRITE Meeting, held at LREC 2004 in Lisbon on *Building the Language Resources and Evaluation Roadmap*, the aim of the workshop was to discuss the challenges for language resources in a multilingual environment, including emerging trends and priorities, and to compile a report of recommendations to be presented to funding agencies and global partners.

Two 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.

This year, the Antonio Zampolli Prize was awarded to Christiane Fellbaum and George A. Miller, from Princeton University, Princeton, New Jersey, USA, for their work on WordNet. Christiane Fellbaum's presentation, entitled "Whither WordNet?", and given at the closing ceremony, was attended by a wide audience. We made it available on-line, from the LREC home page: www.lrec-conf.org/lrec2006/.

The LREC conference is a biennial event: it was announced at the end of this edition that, in 2008, LREC will most probably be organised in Marrakech, Morocco.

Now concerning the content of this ELRA newsletter dedicated to LREC 2006, we decided to have a double special issue, due to the high number of contributions from authors and presenters at LREC 2006.

We received session summaries, as well as workshop reviews, and we are happy to offer in the ELRA newsletter an overview of this LREC. Apart from these, Opening Ceremony speeches and conference reports are also available.

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 2006 Conference Chair

This is the fifth edition of LREC, which means that LREC is only 8 years old, not even a decade, but many things have changed in these few years.

In 1998 Antonio Zampolli understood that a new community was forming, around the topic of Language Resources (LRs) and Evaluation, a community whose interests were not served completely by the major conferences of the area of Computational Linguistics. His intuition, like many others before, proved to be absolutely right, as confirmed by the ever growing number of submissions to LREC and by its extremely large attendance. When LREC was established in 1998, LRs - and with them Evaluation - were 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. LREC has already become, after just 8 years, a 'traditional' and very big conference in the sector of Computational Linguistics.

What does it mean? It is a confirmation that LRs constitute indeed the necessary infrastructure for any Language Technology (LT) and Evaluation project. This was the great intuition of Antonio and of some of us (the oldest here) back in the late '80s. Among these I'd like to mention also Don Walker, who played an important role in making the role of LRs recognised within the Computational Linguistics community.

The “data-driven” approach is no longer something for which to fight, as it was many years ago for colleagues like Geoffrey Leech: some of us still remember how his corpus analyses were badly received at a European ACL of the '80s. This era seems so far today, and the youngest may consider it absurd.

Statistical methodologies are now by far the major trend in computational linguistics, even too much, sometimes at the expense of serious linguistic analyses. In the same direction, robustness is of major relevance for the production of effective applicative systems. And data, i.e. LRs, are behind these trends. We have to pay attention to avoiding that innovative and valuable trends do not become just 'fashions'.

At the same time the recognition of the need for good quality, for comparing results, for measuring progress, and so on, has given more and more importance to evaluation methodologies, as we all know.

LREC remains the best observatory for an examination of the evolution of the field of LRs and Evaluation, and by consequence of LT. Looking retrospectively at the various LRECs, and at this LREC now, we can - and maybe must - ask ourselves a few questions:

i) whether, how and how much LRs have influenced the evolution of LT,

ii) how the field of LRs itself is changing, and based on these, but more critically for our future, questions such as:

iii) how the achievements of the last years must influence our future directions of research,

iv) if completely new trends are in front of us,

v) what will be the role of LRs in the future of LT,

vi) which infrastructural, strategic, cooperation or coordination initiatives are needed in the next years for a better deve lopment of the field.

Just a few words on the first two complementary points.

It is the merit of LRs (or at least a big part of the merit) if LT is changing so much, is acquiring maturity, and is gradually attaining the robustness needed to become truly useful in real world applications. This is probably the biggest effect of LRs, causing also a big transformation of LT, from 'just' a R&D sector to a technology with a great impact in the society.

But also the field of LRs is changing in many ways, and consequently the needs of our community are different. It is more mature, which is a trivial observation, but this may have not trivial consequences. At the beginning of the '90s three major areas were perceived - and described probably for the first time by Antonio and me at a workshop in Santorini in 1993 - as critical for the development of the field:

i) standardisation of LRs,

ii) creation of basic LRs and their annotation,

iii) distribution of LRs.

Major projects and initiatives of the '90s had objectives related to the implementation and satisfaction of these needs: i) standards were defined, accepted and used; ii) many large LRs were created. iii) ELRA came out from this vision. Where we are with respect to these needs? Work on the three tracks is still going on, as we see from LREC papers. A lot has been achieved, but a lot still has to be done, both in quantity (more standards, more LRs for more languages, the web considered an invaluable source, what was considered large 10 years ago is no longer large, more distribution) and in new ways of approaching the problems. A fourth area has clearly acquired an increasingly larger relevance for LRs and LT, i.e.: iv) methods for automatic acquisition of linguistic (or other) information.

But what is of interest to me is:

- are the three/four areas still valid, or the most critical, today?

- in which ways these areas are changing?

- which are the new needs of the field?

I think LREC is helpful in answering these and similar questions.

Preparing the programme of this Conference what have I noticed? Let me quickly touch just a few of the issues, and of the questions raised by the set of submissions.

If we consider the traditional levels of linguistic analysis, morphology seems no longer a central issue. It is probably almost solved, and also for syntax there is a lot of consolidation of achieved results. While semantics is still a topic for research and development, it is really at the centre of the scene, and this happens looking both at works on corpora and on lexicons. Ontologies are becoming central.

Systems maintain their importance also in a conference for LRs and evaluation, in particular for information extraction, information retrieval, machine translation, question answering,… Many papers, more and more, focus on evaluation, either as evaluation of tools, systems or also of LRs themselves (validation in this case), many also on evaluation methodologies per se and on usability and user satisfaction.

Moreover, new topics are emerging, linked to subjectivity more than to the 'objective' aspects of meaning, and interestingly this happens both for spoken and written research. I mean topics such as discovery, analysis, representation of sentiments, affect, opinions. This is a new area of research with potentially enormous applicative impact, in areas such as business, marketing, intelligence. The interest for these new topics does not exclude that more 'objective' areas do not present challenges, on the contrary. Despite the progress in the ability to semantically annotate texts, we are far from having 'solved' the problem of 'meaning' or of semantic interpretation of texts. To grasp, manipulate, and effectively use content, both objective and subjective aspects of it, remains the big challenge of our field. Intelligent access to content is thus a goal, maybe a revival - hopefully more successful - of the old Artificial Intelligence with new and more powerful means, i.e. new batteries of tools and resources.

Another hot topic is multilinguality. This has been sometime neglected, while it will be a major unifying factor for future R&D.

The same is true for multimodality, which is more and more important. This emerges not only from the quantity and quality of submissions, but also from two of the satellite workshops.

And general topics such as LR infrastructures and architectures, large projects, organisational and policy issues see a big growth, receiving more and more attention.

Do we have theoretical issues? Or ours is just a practical empirical field? We have both. The 'data-driven' approach is by nature empirical, and statistical methods are certainly pervasive, but theoretical reflections on language are imperative also in this area.

Do we have revolutions? Probably not. Even if the stable growth of the field brings in itself some sort of revolution. After a proliferation of LRs and tools, we need now to converge. We need more processing power, more integration of modalities, more standards and interoperability, more sharing (in addition to distribution), more cooperative work (and tools enabling this), which means also more infrastructures and more coordination.

Where are we going?

The set of LREC papers, of workshops, tutorials, are together delineating some trends. It's up to all of us to draw the consequences. In a workshop, the last day of the Conference week, we will try to see together what are the emerging trends, the challenges, the consolidated achievements, the promising new directions, the necessary synergies, the breakthroughs - if any.

This is one of the important roles of LREC, to help the community to reflect on itself to have a better vision of the future. I do not want to draw conclusions here. I leave it to the group of us together at the Roadmap workshop to try to do that.

But I would like to have some reflections on these issues at the next LREC, which may be appropriate after the first decade of its life.

Acknowledgments

First of all my thanks go to the Programme Committee, that at every LREC has a harder work, having to deal with an ever increasing number of submissions, of a better quality, while our structure is still the same. I am sure we have made mistakes, and we ask for your forgiveness.

Also on behalf of all the members of the Programme 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 more than 300 colleagues from all over the world. They did a wonderful job, succeeding to complete their reviews in time for so many papers.

We are also indebted to the International Advisory Board and the Local Advisory Board, that have provided moral support to our Conference.

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

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

I specially thank the Local Organising Committee, Lucia Marconi, Paola Cutugno and Daniela Ratti, who had succeeded in finding solutions to local problems, despite having often to face delays or changes in decisions of relevance to local matters. We have solved together the many big and small problems of a large Conference like this. They will assist you during the days of the tutorials/workshops and the Conference.

I thank the workshop, tutorial, and panel organisers, who surround LREC of so many interesting events. A big thank to all the authors, who provide the content to LREC, and give us such a broad picture of the field.

This time I wish also to thank two institutions which have provided economic support and dedicated so much effort, in terms of manpower, to this LREC, as to the previous LRECs, i.e. ELDA in Paris and my institute, ILC-CNR in Pisa and Genoa. Without their dedication LREC would not have been possible.

So I arrive to the last, but not least thanks, dedicated, 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 and wonderful team, with enthusiasm and dedication. My biggest thanks go to Hélène Mazo and Mathieu Robin-Vinet at ELDA, while at ILC to the new LREC entries Paola Baroni and Alessandro Enea, to the old LREC staff Vincenzo Parrinelli, Sergio Rossi and in particular Sara Goggi, 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.

As I said last time, 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 biggest thanks go to all of you. 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 ebullience, exuberance 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, and with the Genoa, Paris and Pisa teams, I welcome you at LREC 2006 in Genoa and wish you a wonderful Conference.

Nicoletta Calzolari Zamorani
Istituto di Linguistica Computazionale del CNR
Via Moruzzi 1
56124 Pisa, Italy

Tel.: +39 050 315 2836 (secr.)
Fax: +39 050 315 2834
glottolo@ilc.cnr.it
www.ilc.cnr.it

LREC 2006 Opening Ceremony Speeches

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

When ELRA was established 11 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 of organising a conference covering the same fields as ELRA, with the addition of Evaluation, came up, 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 lately 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. In 2006 we are investing in the production of one language resource, we are continuing the validation activities for both written and spoken resources, and developing a methodology for the validation of multimodal resources. For evaluation, we are focussing on the creation of the HLT Evaluation Portal. We have also been further developing the main activity, namely identification and distribution of LRs. The ELRA catalogue now contains around 800 resources, and as a new activity, 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 for which we have obtained distribution rights. We believe the Universal Catalogue is very interesting and very useful, also for people outside ELDA; this is the reason for making it available to our members.

After our first president, Professor Antonio Zampolli, Pisa, Italy, so tragically passed away in 2003, the ELRA Board created the Antonio Zampolli Prize. 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 LREC2006 conference, the Prize will be awarded for the second 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 2006 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.

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 the team of Nicoletta Calzolari at the Istituto di Linguistica Computazionale, CNR, Pisa and Genova, has had the main responsibility for the practical organization of the conference, supported by the ELDA team under the management of Khalid Choukri. This is a tremendous job and we thank all of them. At the same time Nicoletta Calzolari has been the programme chair, - not a small job with more than 800 submitted contributions! We are deeply grateful to Nicoletta and 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.

Message from Khalid Choukri, ELRA CEO and ELDA Managing Director

Dear LREC participants,

Welcome to LREC 2006, welcome to Genoa!

ELDA, the operational body and distribution agency of ELRA , ELDA (Evaluations and Language resources Distribution Agency), is proud to welcome you in Genoa, where we are pleased to contribute to the organisation of this fifth edition of the Language Resources and Evaluation Conference, LREC 2006. We are very pleased to continue the organisation of such an important event in such an attractive city.

For ELDA, the strong involvement in the organization of LREC is part of the core mission of the task force set up by ELRA to conduct its strategy and actions. Since the beginning, ELDA considered that it was of paramount importance to join forces with other partners to organize LREC instead of delegating its organization to an "event" organizer that would not be acquainted with the field of Language Resources and Evaluation from the inside.

LREC 2006 is the fifth biennial conference on Language Resources and Evaluation, the fifth in a very successful series of events since ELRA initiated it with the strong involvement of the founders of ELRA and the continuous support of a large number of active organizations in the field. There is no doubt that LREC has become an essential milestone in the field of Human Language Resources and Evaluation, both for academic and industrial players. With more than 800 participants, LREC proves to be an unquestionable success and a fruitful forum for all of us. One of the challenges of such an organization is to attract participants from academia and industry. With almost two-thirds of the participants from academia and one third from industry, LREC achieves one of its goals: paving the way towards a rich cooperation between all sectors in this field.

Since the very beginning, LREC 1998, ELRA has also made sure that it meets with representatives of its members through its special offer to its members that benefit from favourable registration conditions. With over 100 participants per conference who attend from organizations members of ELRA, such a goal has been achieved.

With the maintenance of a permanent web site (at www.lrec-conf.org) and the possibility to have access to all LREC proceedings, including the proceedings of workshops, ELDA is playing a role in making such treasure available over time.

ELRA and ELDA have also learnt a lot from the science and technology that are presented at LREC. Some parts of these have to do with ELDA daily activity. The infrastructure set up about 11 years ago is being challenged at various levels today.

The core activity of ELDA is identification of Language Resources, negotiation of distribution rights and cataloguing such LRs in our online catalogue. Over the years we have revised our catalogue to account for new types of resources (e.g. multimodal), new metadata sets, addition of evaluation packages, etc. Our distribution work is permanently reviewed in the light of the new trends and new distribution mechanisms in particular today's licensing schemes encourage us to investigate new modes inspired from "open sources" or GNU-like principles. This is yet another reason to offer a forum for discussion on the latest developments of Research in the field. Since 1998, LREC has also boosted international cooperation, thanks to exchanges between researchers and industrial partners who could meet in these attractive LREC locations.

The large number of satellite workshops is also a good sign of the vitality of the field. Despite the difficulty in organizing all these events together, we maintain our objective of offering the best forum to all.

To better understand the LREC conference, it is necessary to elaborate a little bit on ELRA, the European Language Resources Association.

ELRA was founded in 1995, with the support of the European Commission. The main mission of the Association was to provide a clearing house for language resources, while promoting HLT more generally. In parallel, ELDA, the Evaluations and Language resources Distribution Agency, its operational body and distribution agency, was created to handle every activity in relation to the identification, collection, production, marketing and distribution of language resources, along with the participation in HLT evaluation campaigns and other HLT projects, at the European and international levels.

The collection and distribution of language resources are major activities for ELRA and ELDA and highlight the central role played by both bodies for the advances in the field. However, other crucial services related to language resources and language technologies are also offered. These include the validation of language resources, thus ensuring the best quality of the language resources presented in the catalogue, with the support of ELRA's network of validation centres; the production of language resources, mainly SLRs within projects ELDA participates in; and the evaluation of speech and language technologies, with involvements in evaluation campaigns to ensure that evaluation resources (data test suites, protocols, methodologies, results, etc.) are packaged and made available to the HLT community, on the model of LR distribution.

In order to serve better the HLT community, ELDA is always engaged in discussions with data providers to obtain the best distribution conditions for the R&D community. We hope that the debates taking place during LREC will convince some potential providers of the extreme importance of sharing data and tools.

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

ELRA and the Organising Committee have made every effort to ensure the success of LREC 2006, making it a fruitful and enjoyable event. We hope that you will enjoy LREC 2006, benefiting the best from the conference programme, the satellite workshops, and the social programme. Hopefully, you will also get the opportunity to do some sightseeing, and enjoy Genoa and Italy.

On behalf of ELRA and ELDA, as well as on your behalf, I would like to warmly thank the local team in Genoa responsible for the practical aspects of this event. As you can imagine, organising such an important event is not an easy task to carry out.

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

Once again, welcome to Genoa, welcome to LREC 2006.

LREC 2006 Antonio Zampolli Prize

Speech given by Bente Maegaard

This year, the Antonio Zampolli Prize was awarded to:

George A. Miller,
James S. McDonnell Distinguished University Professor of Psychology, Emeritus, and
Christiane Fellbaum,
Senior Research Psychologist
Princeton University, Department of Psychology, 1-S-5 Green Hall, Princeton, NJ 08544, USA

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."

Motivation:

WordNet is a unique language resource in many ways. English WordNet provides an extraordinarily comprehensive mapping of lexical items to other semantically related lexical items. Its architecture was driven by psycholinguistics principles, yet it has been widely and effectively used in computational and empirical linguistics as well. This sense-based structure as well as its free accessibility to the research community helped to make it the language resource that is most adaptable to multi-linguality. In this fast shrinking world where preservation of language diversity is as urgent as the overcoming of language barriers, WordNet has become the double-bladed sword of choice. There are now Wordnets in at least 38 languages in the world.

George and Christiane deserve this recognition primarily for their pioneering work that created a truly meaning-driven, sharable architecture as an English language resource. However, in a broader context they have also significantly contributed to greater global communication. With their insistence on making the original Princeton WordNet freely available, and their generosity in supporting in all possible ways WordNet efforts for other languages, they have almost single-handedly created a genuinely cross-lingual shared resource platform that can encompass all the languages of the world, irregardless of financial backing or number of speakers. They have set a standard for worldwide scientific cooperation that we would all do well to emulate.

George Miller has made many crucial contributions to the understanding of human cognition. The most popular and best known contribution of his is of course the magic number seven, plus or minus two. Perhaps he was looking for a magical word when he observed that dictionaries require the human users to bring a lot of background knowledge. He often tells the following story to underline the motivation of a lexical knowledgebase like WordNet: A school child was asked by his teacher to use dictionaries to make his essay really good. Hence, instead of writing "My parents are aging", he went and looked up the dictionary for alternative, fancier word. The sentence that he came up with was "My parents are eroding". Thanks to George and WordNet, parents in the world can now age gracefully without eroding quality of life.

Christiane Fellbaum, like George Miller, has her primary academic training in psychology. She is also well-exposed to different languages in the world. In addition to her native German, and to English, she also studied Japanese. Her recent research grants include a Wolfgang-Paul Prize of the Alexander-von-Humboldt Foundation to work on German Collocation, as well as a project to work on Arabic Wordnet, funded by the REFLEX program. Her own range of multilingual work reflected her belief that linguistic knowledge is the accumulated legacy of all human beings and should be shared.

Both George and Christiane are dedicated researchers. George comes to the lab regularly even though he is now retired. Christiane, on the other hand, works so hard that she rarely finds any time for leisure but instead finds enjoyments and fulfillment in her work.

I, Bente Maegaard, gave this speech at LREC2006, but I received the input from the nominators, whom I thank very much.

The presentation given by Christiane Fellbaum, entitled ***"Whither WordNet?"****, can be viewed from the LREC 2006 web site:*

www.lrec-conf.org/lrec2006

LREC 2006 Session Summaries

Summary of the Oral Session "Semantics & Infrastructural Issues"

Dan Tufis

The session "O1-W: Lexicons, Semantics & Infrastructural Issues" included seven interesting papers addressing quite a large spectrum of issues within the thematic area.

The first paper *A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons* by Yoshihiko Hayashi, Toru Ishida introduces a new concept, "The Language Grid", as a language infrastructure for intercultural communication, available on the Internet, aiming at solving the problems of accessibility and usability in the currently available language services. The paper focuses on the dictionary access services and proposes an abstract dictionary model for the accurate meta-description of this service. The authors exemplify the ability of the model to integrate different types of dictionaries (MRD and lexical ontologies) by commenting on the navigation links between different mono and bi-lingual representations of the senses of a given lexical item (bank).

Somehow related, the paper *Moving to dynamic computational lexicons with LeXFlow* by Claudia Soria, Maurizio Tesconi, Francesca Bertagna, Nicoletta Calzolari, Andrea Marchetti, and Monica Monachini, describes a web application framework where lexicons, represented in a standardized format (here, MILE lexical model), may semi-automatically interact and thus reciprocally enrich themselves. The authors propose a workflow architecture, using an agent-based approach, where each human or software agent can participate to the workflow with one or more roles, as prescribed by a hierarchical role chart (described by XPath expressions). The lexical entries from different lexicons (here, SIMPLE/CLIPS and the ItalWordNet lexicons) become active data structures looking for mutual mapping and merging. This flow results in a dynamically updated lexical entry (with the validation of two human agents) subsuming the original ones.

Since computational lexicons usually encode normalized lexical items, most of the time they are accompanied by analysis and/or generation engines that cope with the un-normalized word-forms to be found in running texts. For languages with a very productive morphology, such engines become mandatory in any serious NLP application (if storage limits is not a roadblock anymore, manual description of all the variants of the lemmas in a lexicon might be unfeasible). This issue is addressed by Violetta Cavalli-Sforza and Abdelhadi Soudi in their paper *IMORPHE: An Inheritance and Equivalence Based Morphology Description Compiler*. IMORPHE builds on a previous morphology description compiler and extends it in a significant way, embedding the new system in an inheritance-based framework. Motivated mainly by the needs of describing the challenging morphology of the Modern Standard Arabic, the new system ensures more conciseness and modularity to the linguistic descriptions while the runtime efficiency of the morphological generation is improved.

Alon Itai, Shuly Wintner and Shlomo Yona, in their paper *A Computational Lexicon of Contemporary Hebrew,* deal with a similarly hard language from the morphological point of view. The reported work presents the Haifa Lexicon of Contemporary Hebrew, the broadest-coverage publicly available lexicon of Modern Hebrew, currently consisting of over 20,000 entries. The lexicon, developed for NLP applications, is accompanied by morphological processors (analyzer and generator) and can be used as a research tool in Hebrew lexicography and lexical semantics. It is open for browsing on the web and several search tools and interfaces were developed which facilitate on-line access to its information.

The next two papers presented in the session described ongoing WordNet projects, their methodologies and the current status of development for Basque and Modern Standard Arabic languages. The paper *A methodology for the joint development of the Basque WordNet and Semcor* authored by Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello and Mikel Quintian give a detailed description of the task (edition, tagging and refereeing), including the main criteria for solving difficult cases in the edition of the senses and the hand semantic annotation of the corpus (about 300,000 words), with special mention to multiword entries. A detailed account on the quantitative data, as well as an analysis of the agreement rates provide the reader with a clear view on the reported work and on the future development plans.

Sabri Elkateb, William Black and Piek Vossen in their contribution *Building a WordNet for Arabic* describe an incremental approach, along the lines of EuroWordNet and BalkaNet previous projects, to a Modern Standard Arabic WordNet aligned to the Princeton WordNet. In addition to the standard wordnet representation of senses, word meanings are also available in a in first order logic representation. The basis for this semantics is the Suggested Upper Merged Ontology and its associated domain ontologies. Tools to be developed as part of this effort include a lexicographer's interface modeled on that used for EuroWordNet, with added facilities for Arabic script.

Most recent NLP applications are incorporating various types of ontology-like knowledge bases (or even proper ontologies) and this trend appears to generalize. While there are many ontology editing tools aimed at expert users, there are very few which are accessible to users wishing to create simple structures without delving into the intricacies of knowledge representation languages. The paper *User-friendly ontology authoring using a controlled language* by Valentin Tablan, Tamara Polajnar, Hamish Cunningham, Kalina Bontcheva describes a system which allows the user to create and edit the taxonomical structures of an ontology (the hierarchy of classes, instances of classes, properties and their values) through statements in a restricted version of the English language. The controlled language described within is based on an open vocabulary and a restricted set of grammatical constructs. The system can "verbalize" (in the same controlled language) an existing ontology and thus, a newcomer becomes familiar with the machine "dialect" and can interact with the ontology building system without preliminary training. Sentences written in this language unambiguously map into a number of knowledge representation formats including OWL and RDF-S to allow round-trip ontology management.

Dan Tufis
Romanian Academy, Research Institute for AI
13, Calea 13 Septembrie
050711, Bucharest 5, Romania
Tel: +(40 21) 3188103
Fax: +(40 21) 3188142
tufis@racai.ro

Summary of the Oral Session "Multilingual Corpora "

Steven Krauwer

This session comprised six talks, all dealing with Multilingual Corpora, but targeting six different audiences.
The first paper addressed the problem that translation researchers and students need large general or domain specific corpora. The solution offered by Sharoff consists of a methodology to compose Internet-derived corpora and a uniform access to these corpora.
The audience for the second paper are adults who want to learn a third language that is related to a second language they already know to some extent. Ciobanu et al offer an intuitive and user-friendly environment based on existing trilingual corpora and other easy-to-assemble material.
The main beneficiaries of the third paper are minority language researchers. The paper by Feldman et al offers a method to exploit existing language resources for rapid, low-cost development of resources for new languages. The paper shows how this process behaves within and across language families.
The fourth paper may be of interest to statistical MT researchers. Kuhn et al address the problem of phrase-level alignment of parallel corpora. They present a representation format for syntactic correspondence and work on tools to automatically labelling a corpus on the basis of manually annotated seed data.
The fifth paper is interesting for those who work on NLP projects with semantic components. Rambow et al aim at an incrementally deepening interlingua notation through deep-syntactic annotation. Results for six languages show that many syntactic differences disappear.
The last paper is again about language learning, but this time the target audience are the teachers. Granfeldt et al present a corpus based method for grammatical profiling of language learners. Machine-learning methods are used to detect the learner profiles.

Steven Krauwer
ELSNET, University of Utrecht
Faculty of Arts, Utrecht University
Trans 10
3512 JK Utrecht
The Netherlands
Tel: +31 30 253 6050
Fax +31 30 253 6000
Steven.Krauwer@ELSNET.org

Summary of the Oral Session "Question-Answering"

Bernardo Magnini

Question Answering (QA) is a recent hot topic in Computational Linguistics. While a number of different techniques and approaches have been proposed the last years, issues related to evaluation and resources for QA have been added only recently to the research agenda on QA. The six papers presented at the oral QA session at LREC-2006 represent a good selection of current issues in the area: the presentations have been followed by a room crowded of people, showing the high interest for Question Answering.
The first paper, *Toward Natural Interactive Question Answering*, presented by Gerhard Fliender, considers how to move from current isolated questions to more natural dialog-based QA, addressing both anaphora resolution and ellipsis.
The second talk, *Mining Knowledge from Wikipedia for the Question Answering Task,* by Davide Buscaldi and Paolo Rosso, proposes to use Wikipedia as a source for answer validation.
Language Challenge for Data Fusion in Question Answering, by Véronique Moriceau, addresses the problem of providing the user with better answers extracted from different sources.
Summarizing Answers for Complicated Questions, by Liang Zhou, Chin-Yew Lin and Eduard Hovy, proposes a view on QA from the perspective of Automatic Summarization, including an interesting report on evaluation measures already successfully adopted for automatic evaluation of summaries.
An Answer Bank for Temporal Inference, by Sanda Harabagiu and Adrian Bejan, describes a resource, AnswerTime Bank, where a large amount of temporal questions and their respective answers are stored and made available to QA systems.
The last presentation of the session *Using Semantic Overlap Scoring in Answering TREC Relationship Questions*, by Gregory Marton and Boris Katz, addresses the evaluation of complex QA systems proposing a methodology where the performance of each component of the system is scored by means of a specifically designed tool.

Bernardo Magnini
ITC-IRST
Via Sommarive, 18 I
38050 Povo-Trento
Italy
Tel: +39 0461 314528
Fax: +39 0461 302040
magnini@itc.it

Summary of the Oral Session "Tools and Evaluation"

Mary P. Harper

This session spanned a variety of topics related to automatic speech recognition (ASR) tools and evaluation. Two of the presentations involved methods for enhancing the performance of the acoustic models in ASR systems. *Transcription Cost Reduction for Constructing Acoustic Models Using Acoustic Likelihood Selection Criteria*, by T. Kato, T. Toda, H. Saruwatari, and K. Shikano, describes a selective sampling method to reduce transcription cost for constructing task-adapted ASR acoustic models. This paper focuses on two important issues: how to select informative

samples and, given this data, how to train the task-adapted model.

Automatic Detection of Well Recognized Words in Automatic Speech Transcriptions, by J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise, discusses the development and evaluation of confidence measures for identifying words with very low error rates from automatically transcribed speech segments from French broadcast news speech using the CMU Sphinx 3.3 decoder. By using high confidence segments as additional training materials, the authors were able to significantly reduce system word error rate.

Two presentations were related to the evaluation of ASR systems. *Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech*, by J. Fiscus, J. Ajot, N. Radde, and C. Laprun, describes a multi-dimensional extension of Levenshtein edit distance calculations for evaluating ASR systems over regions of overlapping speech. As the speech community begins to evaluate on meeting data which contains a significant amount of overlapping speech, this method will support evaluation of system performance on these regions of simultaneous speech. The second paper, *Competitive Evaluation of Commercially Available Speech Recognizers in Multiple Languages*, by S. Burger, Z. Sloane, and J. Yang, describes the evaluation of the accuracy of three commercially available desktop speech recognition engines over eight languages using word error rate. The authors found that two of the systems performed comparably, while the third obtained greater error. Also, read speech was recognized more accurately than conversational speech, and system performance was affected by the language recognized.

The remaining two presentations involved higher levels of processing. *REGULUS: A Generic Multilingual Open Source Platform for Grammar-Based Speech Applications*, by M. Rayner, P. Bouillon, B. Hockey, and N. Chatzichrisafis, describes Regulus, an open source platform that provides a variety of resources to derive domain-specific speech recognizers from unification grammars. Regulus resources, now available on SourceForge, are briefly described followed by the presentation of a series of experiments that investigated the impact of various factors (e.g., vocabulary size, linguistic coverage, features, generality, and use of probabilities) on speech understanding performance. *Discourse functions of duration in Mandarin: resource design and implementation*, by D. Gibbon and S. Tseng, concerns the development of a resource consisting of annotated speech data, tools, and workflow design to support the investigation of discourse phenomena (i.e., discourse markers, discourse particles, and fillers) in Taiwan Mandarin. The resulting annotated corpus, Mandarin Conversational Dialogue Corpus, and toolkit are slated for future release. Measurement studies based on this corpus suggest that fillerS tend to occur utterance-initially; whereas, discourse particles and markers tend to occur utterance-medially, with discourse markers seldom occurring in utterance-final positions.

Mary P. Harper
School of Electrical and Computer Engineering, Schools of Engineering
1285 Electrical Engineering Building
West Lafayette, Indiana 47907-1285
Purdue University
Tel: +1 301 226-8881
Fax: +1 301 226-8811
harper@purdue.edu,
mharper@casl.umd.edu

Summary of the Oral Session “Lexicon and Pronunciation”

Ute Ziegenhain

In the session the following four papers were presented.

In the first paper *SI-PRON: a Pronunciation Lexicon for Slovenian*, by J. Gros, V. Cvetko-Orešnik, P. Jakopin, A. Mihelic, the authors describe the design and development of SI-PRON, a machine-readable pronunciation lexicon for Slovenian containing over 1.4 million lexical entries. The lexicon contains orthography, corresponding pronunciations, lemmas and morpho-syntactic information of lexical entries in a format defined by the W3C Voice Browser Activity. The lexicon is already being used in a Slovenian text-to-speech synthesis system and for generating audio samples.

In the second paper, *"Casselberveetovallarga" and Other Unpronounceable Places: The CrossTowns Corpus*, by S. Schaden and U. Jekosch, the development of a very small corpus of non-native speech that contains pronunciation variants of European city names was presented . The names were chosen from five countries spoken by speakers of four native languages. The authors describe the contents and technical specifications of the corpus as well as strategies to develop a non-native speech database.

In the third paper, *Lexicon Development for Varieties of Spoken Colloquial Arabic*, by D. Graff, T. Buckwalter, M. Maamouri, H. Jin, the authors from LDC described the development of a very interesting toolbox for the development of Colloquial Arabic lexicons from recorded speech databases. The diglossia between Modern Standard Arabic (mostly written/formal speech) and Colloquial Arabic (spoken dialects) poses special problems to orthographic and grammatical annotation. The authors described the different stages on annotation (orthography, pronunciation, morphology, POS and English translation), the user interfaces and the relational database for storing the data. It is planned to make the toolbox available by LDC.

In the last paper, *Experimental detection of vowel pronunciation variants in Amharic*, by T. Pellegrini and L. Lamel, the authors describe the selection of pronunciation variants of vowels in Amharic (Ethiopia) using a speech recognizer. The authors use different methods to create various pronunciation lexica starting on syllable level. Frequent variants for each syllable were then used to build a word-based lexicon. Results show that the inclusion of pronunciation variants during forced alignment on a radio broadcast speech database improved both the quality of the alignments and the likelihood of the acoustic models.

Ute Ziegenhain
Siemens AG
CT IC 5
Professional Speech Processing
Otto-Hahn-Ring 6
81730 Muenchen
Germany
Tel: +49-89-636-40439
Fax: +49-89-636-49802
ute.ziegenhain@siemens.com

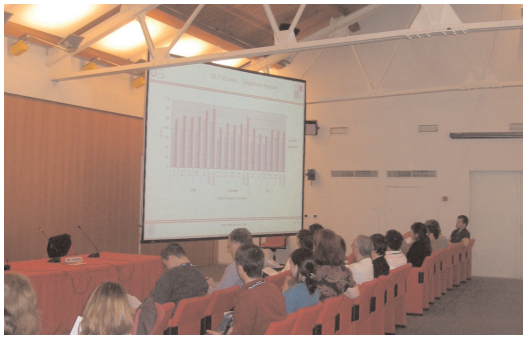
Summary of the Oral Session "Morphology & Tagging"

Kiril Simov

In Session O34-WE Morphology & Tagging four papers were presented: A. Novák: *Morphological Tools for Six Small Uralic Languages*, J. Vaneyghen, G. Pauw, D. Compernolle, W. Daelemans: *A mixed word / morphological approach for extending CELEX for high coverage on contemporary large corpora*, M. Mieskes, M. Strube: *Part-of-Speech Tagging of Transcribed Speech*, B. Hughes, D. Gibbon, T. Trippel: *Feature-Based Encoding and Querying Language Resources with Character Semantics*. The papers discuss two types of problems: how to model morphological knowledge for endangered languages and how to process speech on morphological level. Although the two topics seem to be quite different and to have different aims, they share a lot of common problems. Both areas of research require a careful treatment of problems on the boundary between phonology and morphology. In the first case it is necessary in order to model correctly the pronunciation and the writing system (which very often needs to be created by the researchers themselves) and then to model the morphological knowledge of the language. In the processing speech the problem is to recognize the boundaries of words and to analyze them morphologically. Novák presented an application of two tools for morphology processing of six Uralic languages. First he discussed the morphology, phonology and orthography of the languages. All of them are agglunative languages. The tools are "High speed Unification MORphology (HUMOR)", developed at the Hungarian company MorphoLogic and xfsf - the Xerox Finite State Tool. The selection of the tools was motivated by the fact that both of them were used for modeling other agglunative languages - Finnish and Hungarian. A comparison between the two tools was presented with respect to speed and memory requirements, the adequacy of the grammar formalisms, and applicability to other tasks additionally to analysis like lemmatization and generation. Hughes, Gibbon and Trippel emphasized the importance of the description not only of given phenomena in a language, but also the formal definition of the descriptor inventory. They defined an explicit representation of character features. The required features are represented as feature structures according to the Feature Structure Standard ISO-DIS-24610-1. An XML representation for these feature structure was developed. Standard mechanisms for processing XML are used for querying the resulted database. The main application of the developed system is to document minority and endangered languages. Vaneyghen, Pauw, Compernolle and Daelemans introduced the idea to incorporate constraints over morphological rules as a preprocessing step in language models. The module is set as dynamic. It is easily incorporated in the finite state pipe. It can operate over various units: morphemes, lemmas or wordforms. Promising experiments were reported on Dutch, which is a highly inflected language. The authors concluded that the method helps in reducing the overgeneration of new words, but also reduces the coverage. However, the constraints might be loosened or strengthened with respect to the task. This fact sets the question to the balance between overgeneration and coverage. Mieskes and Strube investigated the balance between manual annotation necessary for training of POS taggers applied to annotation of transcribed multiparty dialogs and the quality of the result of the annotation. They have used four taggers to do the task: TnT tagger (http://www.coli.uni-saarland.de/~thorsten), left3words and bidirectional taggers (http://www-nlp.stanford.edu/software/tagger.shtml), and Brill tagger (http://www.cs.jhu.edu/~brill). The authors first trained the taggers on Wall Street Journal portion of the Penn Treebank, and then they used them to annotate the transcribed speech data, which then was manually repaired. The next step was to retrain and test the four taggers over the created gold standard corpus. The author checked the result of each tagger, but also their combination was checked on the basis of majority voting of the taggers. The procedure was applied gradually up to the moment when there was not a major improvement on the result. The experiments demonstrated that between 197K and 221K tokens of manually checked data gives good tagging results with reasonable effort of manual work. The session showed that advances in morphological processing are still necessary in order to improve the documentation of new languages and to support reliably the next steps of natural language processing.

Kiril Simov
Linguistic Modelling Laboratory
Central Laboratory for Parallel Processing
Bulgarian Academy of Sciences
25A Acad. G.Bonchev Str., 1113 Sofia, Bulgaria
Tel: (+359 2) 979 2825
Fax: (+359 2) 707 273
kivs@bultreebank.org

Oral session



Oral session

Summary of the Poster Session "Anaphora & Coreference"

Véronique Moriceau

During this session, 8 posters were presented. The projects presented during that poster session covered diverse languages from different families, among which: English, German, Norwegian, Italian and Korean.
Globally 4 major points were outlined during the session:
(1) The **construction of annotated corpora** in different domains (terrorism, newspapers, etc.) for three main goals: corpus analysis, training and evaluation.
(2) The **methodologie**s for manual or automatic annotation of anaphora/references. These methods propose annotation guidelines or use for example memory-based learning, machine learning techniques, Centering Theory or logical text structure. They are applied for personal/possessive pronoun, NP or event coreference annotation and/or resolution.
(3) The **evaluation** of those techniques and performances of systems on particular languages and data: comparison of automatic annotation to manual annotation, comparison of annotation from different annotators, etc.
(4) The **applications**: in particular the viewing/editing of coreferences from manually or automatically annotated data, the anonymization of proper names, etc.

Véronique Moriceau
Université Paul Sabatier
IRIT - Equipe ILPL
118, route de Narbonne
31062 Toulouse Cedex 9
France
Tel: +33 5 61 55 74 03
moriceau@irit.fr

Summary of the Poster Session "Corpora & LR Infrastructures"

Sonja Bosch

Six very well presented posters were displayed in this session. The paper by Dimitrios Kokkinakis, *Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus - The MEDLEX Experience*, dealt with the collection, encoding and linguistic processing of a Swedish medical corpus. The significance of the paper was that in contrast to the predominantly English medical corpora that are normally available, this one is in Swedish.
Nelleke Oostdijk and Lou Boves in *User requirements analysis for the design of a reference corpus of written Dutch* analysed the user requirements study conducted for putting the tools and procedures in place that are needed to design a 500-million-word reference corpus of Dutch.
The paper by C. Onelli, D. Proietti, C. Seidenari and F. Tamburini, *The DiaCORIS project: a diachronic corpus of written Italian* described the design processes of a diachronic corpus of written Italian, including the document annotation schema and technological infrastructure designed to manage the corpus.
The annotation of the largest post-edited parallel corpus including Portuguese, was described in the paper by Diana Santos and Susana Inácio, *Annotating COMPARA, a Grammar-aware Parallel Corpus*. Challenges regarding syntactic ambiguity were also addressed.
David Guthrie, Ben Allison, Wei Liu, Louise Guthrie and Yorick Wilks in *A Closer Look at Skip-gram Modelling* investigated the use of skip-grams to address data sparsity in NLP. Skip-gram modelling of various skips with various amounts of training data was examined.
Aimilios Chalamandaris, Athanassios Protopapas, Pirros Tsiakoulis and Spyros Raptis in their paper *All Greek to me! An automatic Greeklish to Greek transliteration system*, which drew a lot of attention, presented research on the transliteration of Greek using the Latin alphabet, a phenomenon of Greek e-mail communication, and known as Greeklish. This research led to the development of the first automatic transliteration system of any type of Greeklish. Various aspects of the system were evaluated.

Sonja Bosch
Department of African Languages
University of South Africa
P O Box 392
0003 UNISA
South Africa
Tel: +27 12- 429 8253
Fax: +27 12- 429 3355
boschse@unisa.ac.za

Summary of the Poster Session "Corpora: Creation, Annotation"

Costanza Navarretta

The presentations in section P6-W dealt with various aspects of the creation and annotation of written corpora. The type of corpora accounted for in the session included general language corpora, domain specific corpora and collections of chat texts which represent a particular type of written material. The languages involved comprise Basque, Chinese, Dutch, Italian, Japanese, Portuguese, Russian, Slovenian and Swedish.
The presentations about annotational issues comprised methodological discussions, descriptions of existing annotated corpora and presentations of specific annotations. The annotation types discussed in the papers cover morfologic (POS), syntactic and semantic information. The different semantic aspects covered i.e. verb relations, interaction events in the biological domain and valency of nominalizations.
A group of presentations dealt with the creation of annotated corpora while another group dealt with tools and resources supporting the creation and annotation of corpora. Techniques for converting existing syntactic annotated resources to other syntactic annotation formats were also addressed.

Costanza Navarretta
Center for Sprogteknologi
University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S
Denmark
costanza@cst.dk

Summary of the Poster Session "Web Services and Digital Archives and Libraries"

Zygmunt Vetulani

At the poster session on Web Services, Digital Archives and Libraries six contributions were presented to the LREC participants. Five of them present on-going projects in their advanced phase and one describes a language resources center (a Web services provider). Four out of the six papers focus on Web accessible digital archives or libraries, one is about a Web accessible multilingual dictionary of sign languages.

Multilingual Search in Libraries. The case-study of the Free University of Bozen-Bolzano, by R. Bernardi, D. Calvanese, L. Dini, V. Tomaso, E. Frasnelli, U. Kugler, B. Plank presented an on-going research aiming at enhancing the Online Public Access Catalogue of the Free University of Bozen-Bolzano. The multilingual access system (Multilingual Search In Libraries / MUSIL/) has been proposed to enable access in Italian, German and English. MUSIL provides automatic translation of the query terms into these three languages. It is based on linguistic knowledge like stemming, grammars, dictionaries and thesauri combined with statistical methods for data retrieval. System architecture, interface and evaluation of search results are presented.

The following three contributions form a series focussing on digital archives and related problems.

The paper by D. Broeder, F. Offenga, P. Wittenburg, P. Kamp, D. Nathan, S. Strömqvist, *Technologies for a Federation of Language Resource Archives*, presents a Grid technology based project of distributed access to language resources (DAM-LR project). The architecture of a federation of archives is described in this paper. The authors declare having started testing the architecture on the basis of "a few well known components" within the consortium grouping several institutions (cf. affiliations of the Authors). The next two papers address two such components.

The contribution presented by P. Berck, H. Bibiko, M. Kemps-Snijders, A. Russel, P. Wittenburg, *Ontology-based Language Archive Utilization*, aims at contribution in bridging the gap which is due to differences in encoding linguistic phenomena. In particular, the paper addresses the issue of interoperability at the level of linguistic encoding and discusses a solution based on bottom-up driven ontologies (created by users) with concepts possibly related to central ontologies (as e.g. ISO DCR).

In the third one of the series, M. Kemps-Snijders, J. Ducret, L. Romary, P. Wittenburg, *An API for accessing the Data Category Registry*, present an API to the ISO DCR, i.e. a flat list of concepts used in linguistics (language engineering) whose main role is to achieve interoperability of linguistic encoding. The main DCR API functions are described. This DCR is operational and the proposed API has already been tested from a lexicon application.

M. Boekestein, G. Depoorter, R. Veenendaal *(Functioning of the Centre for Dutch Language and Speech Technology)* present the TST Centre (Centre for Dutch Language and Speech Technology). The paper describes its organisation, tasks and services consisting mainly in management of a broad collection of Dutch digital language resources, such as audio recordings, digitalized texts annotated corpora, computational lexica, POS taggers, parsers, etc. TST available products and services are briefly presented, as well as licence policy of the TST Centre. The reader will find a substantial description of functioning of the Centre, in particular as a Web provider of language and speech technologies.

We conclude this presentation with a paper by E. Suzuki, T. Suzuki, K. Kakihana, *On the Web Trilingual Sign Language Dictionary to Learn the foreign Sign Language without Learning a Target Spoken Language*. It introduces the foreign sign language teaching problems the Deaf community has to deal with. As there is no universal/international sign language and the national sign languages are not merely mimed forms of the corresponding natural languages, an additional barrier exists for the deaf community to learn foreign sign language. The paper presents an on-going project of development of a trilingual sign language dictionary (for English, Japanese, Korean) as a solution to help students to learn a foreign sign language without necessity to learn the target natural language. The main methodological choices and the progress made so far are presented, as well as the plans for future research.

Zygmunt Vetulani
Department of Computer Linguistics and Artificial Intelligence
Adam Mickiewicz University
ul. Umultowska 87
PL-61614 Poznan, Poland
Tel: +48-61 8295 380
Fax: +48-61 8295 315
vetulani@amu.edu.pl

Summary of the Poster Session "Translation"

Nelleke Oostdijk

The main focus of the poster session on speech-to-speech (s2s) translation was on resources. It included four presentations:

This poster *Are you ready for a call? - Spontaneous Conversations in Tourism for Speech-to-Speech Translation Systems*, by Darinka Verdonik and Matej Rojc was about the Turdis database of spontaneous conversations in the tourism domain which is being developed for the Slovenian language for use in developing speech-to-speech translation components. Data comprises recordings from telephone conversations that were conducted by professional tourist agents. With the recordings orthographic transcriptions are available. At the time the poster was presented, the database constituted around 43,000 words.

Through *GAIA: Common Framework for the Development of Speech Translation Technologies*, by Javier Pérez, Antonio Bonafonte an open-source software platform for the integration of speech translation components was presented. The software has already shown its usefulness in the European LC-STAR project and the Spanish ALIADO project, e.g. in obtaining text and speech corpora for speech translation.

The presentation of *Collection of Simultaneous Interpreting Patterns by Using Bilingual Spoken Monologue Corpus*, by Hitomi Tohyama and Shigeki Matsuhara was about the results of a manual investigation of simultaneous interpreting patterns on the basis of a bilingual (English-Japanese) aligned corpus of monologues. The patterns that were identified were classified into a number of types that will be used as interpreting rules in machine interpretation of simultaneous speech.

The poster *TC-STAR: New Language Resources for ASR and SLT Purposes*, by Henk van den Heuvel, Khalid Choukri, Christian Gollan, Asunción Moreno and Djamel Mostefa, presented an overview of the different resources that have been developed in the TC-STAR project, which include resources for training, development and evaluation, giving details about their properties, validation and availability.

Nelleke Oostdijk
Department of Language and Speech
University of Nijmegen
P.O. Bos 9103
6500 HD Nijmegen, Netherlands
Tel: +31 24 36 12765
Fax: +31 24 36 12907
N.Oostdjik@let.kun.nl

Poster session

LREC 2006 Workshop and Panel Reviews

Workshop on "SALTMIL SIG: Speech and Language Technology for Minority Languages"

Briony Williams

On May 23rd 2006, SALTMIL held a morning workshop on "Strategies for developing machine translation for minority languages". This was a satellite workshop preceding the biennial LREC (Language Resources and Evaluation Conference) in Genoa, Italy, and was chaired by Briony Williams. It was the latest in the series of SALTMIL workshops held as satellites to the biennial LREC conference.

The program was very full, beginning with seven invited talks, each of which stimulated many questions and discussions. This was then followed by a poster session, with sixteen contributed poster papers. About fifty people were present in total, from a wide range of countries, and representing work on a variety of minority languages.

One of the highlights was a talk by Steven Krauwer (University of Utrecht, Netherlands), the originator of the BLARK concept (Basic Language Resource Kit).

His talk, entitled "ENABLER, BLARK, what's next?", focussed on recent developments such as the CLARIN initiative (see http://www.mpi.nl/clarin/) - Common Language Resources and Technology Infrastructure.

This is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable. It is not a project proposal, but rather a proposal for a research infrastructure to be included in the European Roadmap for research infrastructures. He also introduced the concept of the "Blarkette" (a mini-BLARK) for languages with few digital resources.

The other speakers gave talks on the following topics:

- "The BLARK matrix and its relation to the language resources situation for the Celtic languages".
- "Building NLP systems for two resource-scarce indigenous languages: Mapudungun and Quechua".
- "Open source machine translation: an opportunity for minor languages".
- "Approaching a new language in machine translation".
- "Unicode Development for Under-Resourced Languages".
- "Statistical Machine Translation with and without a bilingual training corpus".

One of the participants recorded her impressions of the event as follows:

- There is an optimistic outlook for lesser-resourced languages.
- BLARK is very useful for explaining our needs to policy-makers and funders.
- There is a need to invest in basic language resources before end-user applications can be developed.
- Perhaps a Blarkette is appropriate for languages starting from nothing.
- We need to monitor developments in CLARIN.
- Great savings in time and money can be gained by piggy-backing off a closely-related language which is better resourced, and also by leveraging the work which has gone into a major language which is commonly paired with the lesser-resourced language.
- A half day was sufficient for the workshop.
- The speakers were very high-quality in terms of content and delivery.
- The workshop was well-organised.
- I found it very beneficial!

Briony Williams
Bryn Haul Heol Victoria
University of Wales
Bangor Gwynedd LL57 2EN
United Kingdom
Tel: +44 1506 200862
Fax: +44 1506 842599
b.williams@bangor.ac.uk

Workshop on “Crossing Media for Improved Information Access”

Stelios Piperidis

The development of methods and tools for content-based organization and filtering of multimedia information has become crucial in view of the convergence of technical media platforms. Advances in medium-specific (audio, image, text) processing have facilitated the development of tools for indexing multimedia content. ?ext-based indexing methods of such content still prevail; text processing has reached a level of maturity that enables shallow semantic analysis for identifying keywords, terms and names as indexing terms, with considerable progress being made in the extraction of events and facts. Experiments are ongoing on applying this type of indexing on speech recognition output as such or/and on associating web text to such output (for recovering from ASR mistakes) and then performing text-based indexing (cf. work within the PRESTOSPACE project, www.prestospace.org). Speech processing can provide automatic speech transcriptions of good quality (in certain acoustic conditions), as well as speaker turn and identification information. On the other hand, image-based indexing methods for multimedia content rely on basic image processing and in particular on the extraction of keyframes, shortcuts and low-level image features, while progress in developing face detection, face identification and object recognition technologies contribute to a more promising future for such approaches. Furthermore, there is evidence (cf. TREC-VID and Image-CLEF) that some benefits in performance can be gained through the fusion of the results of visual and linguistic analyses of multimedia content. Research on the automatic association of images with corresponding textual data go beyond fusion of medium-specific results to multimedia integration for indexing and retrieval applications (cf. UP-TV, ACEMEDIA and BUSMAN projects), while a more general notion of "crossing media" within or/and across documents seems to emerge too (cf. the REVEAL THIS project, www.reveal-this.org).

The "Crossing Media for Improved Information Access" workshop explored these new tendencies in accessing multimedia content by bringing together researchers working on the development of indexing technologies for archived and contemporary multimedia content.

Tablan, Cunningham & Ursu presented a method of automatic semantic analysis in the process of creating analytical metadata for digitized audiovisual archives in the PrestoSpace project. Tomadaki & Salway dealt with the resolution of cross-document coreference in an attempt to generate representations of film content out of various texts, such as screenplays, audio descriptions and plot summaries, in order to improve video indexing. Yakici & Crestani presented the cross-media indexing component of the REVEAL THIS project, a component that leverages the individual potential of every indexing information generated by the analyzers of diverse modalities such as speech, text and image. The initial prototype utilises the multiple evidence approach by establishing links among the modality specific descriptions in order to depict topical similarity in the textual space. Koehler described the multimedia indexing system iFinder, a development of the Fraunhofer IMK, and its usage in several research and development projects and applications. The main idea of iFinder is to integrate different multimedia extraction methods for the automatic generation of metadata of audio-visual content and to support international metadata standards, like MPEG-7.

Rehatschek et al discussed cross media tools and multi-modal analysis and their role in automating media monitoring and advancing content production, by presenting relevant results from the DIRECT-INFO and NM2 projects. Kosmopoulos et al proposed an approach to knowledge acquisition, which uses multimedia ontologies for fused extraction of semantics from multimedia content, and uses the extracted information to evolve the ontologies. Ciravegna & Staab presented the X-Media project which addresses the issue of knowledge management in complex distributed environments, by implementing large scale methodologies and techniques able to support sharing and reuse of knowledge that is distributed across different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.). Georgantopoulos et al described the cross-media summarization component of the REVEAL-THIS project. They report different ways of synthesizing the most salient elements of the constituent parts of a cross-media object, visual, auditory or textual, and adapting the way in which these salient parts are fused in accordance with the users' interests, digital equipment and the typology and semantic characteristics of the original information. Last, DeJong reviewed how the concept of media crossing has contributed to the advancement of the application domain of information access and explored directions for a future research agenda. She discussed ways to incorporate the concept of medium-crossing in a more general approach that not only uses combinations of medium-specific processing, but that also exploits more abstract medium-independent representations, partly based on the foundational work on statistical language models for information retrieval.

Stelios Piperidis
Institute for Language and Speech Processing (ILSP)
Department of Language Technology Applications in Office Systems
Artemidos 6 & Epidavrou
GR-151 25 Maroyssi, Greece
Tel: +30 210 6875300
Fax: +30 210 6854270, 6856794
spip@ilsp.gr

Panel on “Resources for the Processing of Affect in Interaction”

Nick Campbell

In line with the opening remarks on the increase of Subjectivity in Language Processing from the Conference Chair, the panel session on Day 3 of the conference was entitled "Resources for the Processing of Affect in Interactions". It was chaired by Nick Campbell and Jianhua Tao, and presentations were made by Véronique Aubergé, Anton Batliner, Ellen Douglas-Cowie, and Laurence Devillers. The latter 3 members represented the EU's Humaine Network of Excellence and presented views summarising discussions held in the half-day workshop on "Corpora for Research on Emotion And Affect" that preceded the main conference.

The panel session examined the frameworks under which resources are being collected for analysis and modelling of Expressive Speech and the display of Interpersonal or Affective Information. There has been considerable increase recently in use of the terms "emotion" and "affect" with respect to speech and multimedia information processing but problems arise when the terms are used interchangeably. The goal of the panel discussion was first to define and diffe-

rentiate the two terms, as they relate to speech processing, and then to specify the different needs and requirements of research and technology development for each. All panelists showed considerable experience in these developing fields of speech and language technology.

Batliner succinctly summarised the problem by pointing out parallels with Language & Gender discussions wherein the well-known phenomenon of Parasitic Reference results in wide use of a gender-specific term that refers to a high-status subset of the whole class in place of a neutral generic term, linking this to the common disclaimer wherein authors state that "In this paper, we use the term "emotion" in a very broad sense, not confined to the big-six, full-blown emotions. etc., etc." and then continue to limit their observations to extreme examples and fail to examine the more everyday types of speaking styles that include e.g., tiredness, motherese/reprimanding, interest, boredom, etc., and that from an application point of view are undoubtedly more relevant to interactive speech processing technologies.

Devillers explained that the "affective states" include emotions and feelings, but also signal attitudes and the interpersonal stances in a discourse, pointing out that there is a significant gap between the affective states observed with artificial data (acted data or induced data) and those observed with real-life spontaneous data; this difference being mainly due to the context. Since her goal is to build a "non-caricatural" Human-Machine Interaction system, she presented examples that showed how to collect real-life databases illustrating natural interactions instead of using biased data from artificial or contrived events.

Aubergé pointed out the difficulties in labelling affective states in a natural discourse, since these involve subjective impressionistic labelling of pragmatic intentions and inferred speaker states, but showed that empathy as a human characteristic can be employed within a scientific framework to produce verifiable, albeit multi-faceted, descriptors of speaking styles and discourse strategies. She proposed methods for the validation of labels and descriptors thus obtained, linking such research to the ecological sciences.

Douglas-Cowie produced further examples of real-world data and described techniques for time-aligning annotations to the multi-media corpora. She proposed categories and dimensions for labelling different tiers of affective information and showed how these can be directly mapped to the observable physical characteristics of the speech signal.

The panel was notable in that all participants referred to collections of real-world data and described practical annotation techniques that have been developed to face the challenges of working with multi-faceted and subjective data within an objective technological framework. In concluding remarks, the chairs pointed out that if machines are to be made sensitive to this type of interpersonal communicative information, then we will need more such natural corpora upon which to base our future research. If these corpora are acted or contrived, then the resulting technology will be of little use in the real world. The data described here display unwieldy and complex interactions of factors, but the more natural the data that we can collect, and the more complex the factors they illustrate, the closer we can come to an understanding of the mechanisms of human social communication and can perhaps model them for general use in the ubiquitous computing environment that is becoming so much a part of our everyday lives.

Nick Campbell
Spoken Language Communication Research Laboratory
Advanced Telecommunications Research Institute International
Keihanna Science City
Kyoto 619-0288, Japan
nick@atr.jp

LREC 2006 Reports

Report on Papers on Evaluation for Spoken and Multimodal Communication

Joseph Mariani

The number of papers on speech and multimodality in general is decreasing this year, in volume and in ratio (it used to be 30% of the total number of papers, and it goes down this year to 22%). After decreasing from 30% in 1998 to 25% in 2000, the ratio of papers in the general area of evaluation is now stabilized at about 20% since 2002, but evaluation is now used in all areas of Language Technologies: the ratio of evaluation papers on written language is increasing from 50% in 2004 to 70% this year, 20% are on speech (compared with 30% in 2004), 5% are on multimodality and the same on terminology.

We find evaluation activities presented at the conference for various purposes: for technology assessment (many), for usability assessment (few) and for product assessment (few). Various technologies are addressed: speech recognition, speech synthesis…, and some are dealing with both speech and language: oral dialog, speech-to-speech translation… Many different application areas are targeted: meeting transcription, closed caption TV, telephone services.

Evaluation in speech and multimodality is conducted within large programs, such as GALE supported by Darpa in the US, TC-Star, CHIL, AMI or CLEF in Europe, Techno-Langue in France (with several evaluation campaigns: Evasy, ESTER, MEDIA…), STEVIN in The Netherlands.

More and more languages are addressed in the systems which are developed and assessed. In 2004, papers concerning American English, French, German, Japanese, Portuguese, Dutch, Russian, Czech, Slovenian, Arabic, Spanish, Basque and Cypriot were presented at the LREC conference. This year, other languages have been added: Catalan, Danish, Persian, Somali, Amharic, Dutch regional variants.

There is still need for more coordination:

- In order to compare performances across languages: how to compare the quality of a system in a given language with another system assessed in another language?
- In order to use the same data for various tasks at various levels, for analyzing the influence of performances at lower levels on the overall system performances, such as the influence of POS

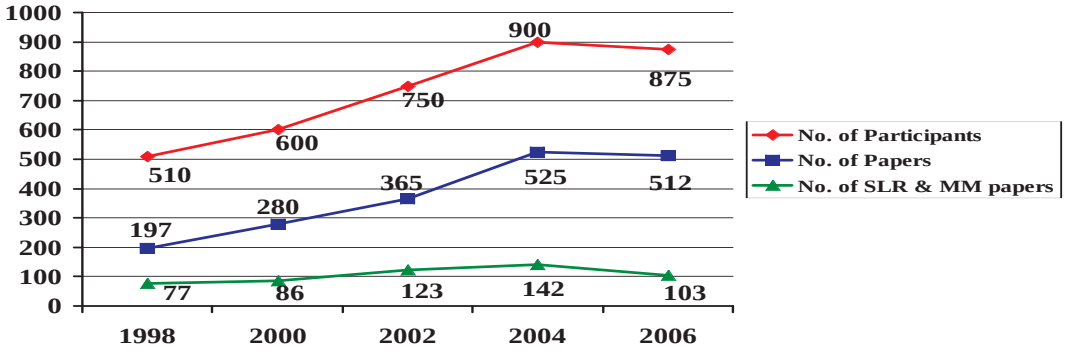
tagging, syntactic parsing and Named Entity extraction components on the quality of Broadcast News retrieval, for example. This also means that the outputs of lower level processing should be made available to upper level processing and that this goes across domains: NLP, speech processing and multimodal communication overall.
Evaluation is now mandatory in the Language Technology R&D activities, in order to know where we are, and how fast we make progress.

Joseph Mariani
LIMSI-CNRS
BP 133
91403 Orsay Cedex
France
joseph.mariani@limsi.fr

Report on Spoken Language Resources and Multimodality

Daniel Tapias and Khalid Choukri

To begin with, we would like to stress the fact that the number of papers/submissions did not increase as was expected. On the contrary (shown in the figure below), the number of papers on Spoken Language Resources (SLR) and Multimodality (MM) has decreased substantially (by 27%, from 142 in 2004 to 103 in 2006), whereas the total number of accepted papers has seen only a slight decrease (2,48%, from 525 in 2004 to 512 in 2006).



LREC 1998 to LREC 2006 - Evolution in participation and paper submission

In 2006, 140 papers dealing with Spoken Language Resources and Multimodality were submitted (out of 789) as opposed to 193 (out of 719) in 2004. A thorough analysis is required to understand better the reasons behind these figures.
In spite of all our efforts, the proportion of papers on Spoken Language Resources and Multimodality represents only 20% of the total number of papers, as opposed to 27% in 2006. The analysis may check the impact of conferences that took place just before LREC, for instance ICASSP in Toulouse the week before the conference, or the ones scheduled after it, such as ICSLP/Interspeech in September 2006.
Let us elaborate more on this area of SLR/MM resources within LREC. We do our best to keep the same structure as the report from LREC 2004 on Speech Corpus and related resources.
An impressive number of Language Resources (over 176) was presented at LREC 2006 out of which some are spoken resources. In addition to resources, several papers introduced various annotation and recording tools, processing platforms (including open-source platforms) and a number of initiatives to conduct collaborative activities in this area.
Among the speech resources, we could mention the large number of "broadcast news" collections, as within the French project ESTER, the Japanese monologue data, SINOD, etc. For the ESTER project , a French database was designed, collected and transcribed while speech and speaker recognition were being evaluated. The ProGmatica database is also worth mentioning as it focuses on a Portuguese Collection that is being used for other research activities such as prosodic analysis. An original resource is the SINOD database which is a Slovenian broadcast news collection of non-native Slovene speakers.
A number of papers addressed the needs of Text-to-Speech Synthesis (TTS) and it is good to see that a large number of languages are involved. For instance, the TC-STAR project introduced its resources for English, Castilian Spanish and Mandarin, annotated at various levels: text corpus for grapheme-to-phoneme conversion, audio data for inter-lingual speech conversion, etc. Other languages like Polish, Norwegian but also Basque and Catalan were presented. In total, about 20% of the papers on Spoken Language Resources were devoted to TTS.
Dialog Collections have also been described in detail in the papers on SLR. Some focused on evaluation and resources for evaluation (the French project Media), others focused on new resources such as (1) the Dihane Collection for Spontaneous telephone conversations in Spanish, (2) a replicate of the HCRC Map Task in Danish, (3) a Slovenian dialog database, etc. Non-native issues are also handled through the collection of specific databases: the SINOD data mentioned earlier, the European city names, French tourism dialogs, a Polish "learners" of English database, etc.
Resources for the innovative area of Speech-to-Speech translation were also introduced through several presentations, in particular the ones made by the TC-STAR Consortium (both on resources and evaluation) and some others like the presentation of parallel corpora for Valencian and Spanish. It is worth mentioning that for this topic both speech and text resources are used and constitute a joint theme for speech and written communities. Mobile "phones" and PDAs are also media that were used extensively in several collections e.g. within the German smartweb handheld corpus (SHC) and the smartweb motorbike corpus. Some of these collections refer to UMTS as well.
Several papers elaborated on multimodal resources. Very often, but not always, speech is combined with other modalities.

The most important part focuses on meetings, seminars, human-human Multimodal dialogues as well as the Wizard-of-Oz based collections with a focus on audio, video, images and associated textual inputs. Some resources are collected under "real" Human-Machine Interactions conditions, for example, the H.C. Andersen Conversations, the Bielefeld Topic Tracking (BITT) corpus, etc.
A few papers focused on "emotion" annotations on the basis of new or existing multimodal/multimedia resources e.g. the annotations of EmoTV Corpus (for French), the Belfast-Naturalistic database and the Castaway database (for English), emotions in meetings with the AMI project. There is also the Safe Corpus and the "fear type" emotions annotation and detection, as well as the use of emotion-based speech collections for TTS applications as was described for a Basque database.
An interesting document in this category is a paper on sign language and its association with "linguistic" annotations including the use of existing tools such as Anvil.
In addition to the large number of Spoken Language Resources described, a number of papers focused on Annotations of resources: both on methodologies and on tools. Some examples are: a grammar-based ASR Platform, tools for Speech-to-Speech translation components, etc.
It is important to note here the growing trend in the use/development of open-source and freeware softwares for Speech technologies (including Speech-to-speech translation components). Written and terminology language resources types follow a similar trend.
Annotation and recording tools remain an important category in our conference. Hopefully the re-use of existing tools is widely advocated for and important extensions, customizations are mentioned by authors.
We may mention the work on the multimodal Woz tool to extend the well-known approach to data acquisition. Transcription of speech/multimodal resources is very expensive and various automatic tools have been created to help cut costs and improve efficiency (for French, Japanese and also Somali!).
Resources and tools for Phonetic lexica and pronunciation "dictionaries" are also well represented for many languages including for less-studied languages (so far) e.g. Amharic, Arabic but also for non-native pronunciations. We can also mention work carried out to exploit an existing TTS system to derive a TTS system for a language that does have required resources (use of German TTS to "fake" Somali TTS).
As in the past, LREC is also the place to report on important projects, national and international initiatives and to get an update of the activities of major data centers (e.g. ELRA and LDC). A special session, O27-G : Large Programs, Data Centers and International Operations, was devoted to this aspect and helped update the audience with the activities carried out in Asia (report from Oriental-Cocosda new and past convenors), reports from several national programs in Europe (Stevin in the Netherlands, KUNSTI in Norway, Technolangue in France) and the newly established "European Center of Excellence for Speech Synthesis" (ECESS) that targets the exchange of TTS modules between its members and that is open to any TTS lab.
Many interesting topics have been addressed within LREC 2006 and it is almost impossible to highlight just a few topics. If one needs to keep in mind a few trends these could be open-source for building blocks of ASR, speech-to-speech translation, multimodal/emotion resources, language resources for some under-represented languages, customization of existing and well-known tools, tools for "semi"-automatic transcriptions.
Despite all this, there is still room for more work. Cooperation between Speech and Written language communities is taking off but needs more efforts. A large number of important languages (in terms of number of speakers) lack basic resources: coding schemes are not standardized yet and the automatic/semi-automatic tools to speed up transcriptions and to reduce costs are not widely available. For Europe a large number of initiatives are coming to an end without any clear vision on what will happen next and how the invested funds can lead to sustainable resources, not to mention the large number of countries without any coordinated program or without any program at all.
We hope that LREC 2006 was the right forum to discuss all these issues and let us wish an LREC 2008 with more speech and multimodal resources and with more reports from large programs and initiatives that could help bring this area one step ahead.

Daniel Tapias
Telefónica Móviles España
C/ Serrano Galvache, 56
28033 Madrid, Spain
daniel.tapiasmerino@telefonica.es

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

Report on Written Language Resources

Jan Odijk

In this short summary, I present a sketchy characterization of the Written Area at LREC 2006. I follow the schema of a similar report prepared by me for the previous LREC in Lisbon, which itself was based on a schema made by Nicoletta Calzolari. This makes it easier to comparatively assess the main tendencies in the field. But because the previous reports also covered the General Area and the Terminological Area, these comparisons are not perfect.

Parameters for Classification
(see table on page 22)

As in previous years, we received an impressive amount of papers for the Written Linguistic Resources (WLR) area, such that multiple parallel sessions on WLR were necessary, and a huge amount of posters had to be accommodated.
In the previous reports there were four parameters to broadly classify WLR papers: i) research vs. development, ii) type of resource/tool/etc. described, iii) linguistic description level, iv) language(s). Each has sub-classifications for which the relative order - in terms of number of WLR papers (both Oral and Poster) - is given. This provides a global quantitative, even though sketchy, overview of the distribution of interest among LREC authors, and a rough idea of the relative weight - as of today - of different

aspects related to WLR. The following table summarizes the findings (grey cells denote areas with interesting increase, while purple ones denote decrease with regards to the previous LREC):

Levels of Linguistic Description

The major trend at LREC-2006 with regard to linguistic levels of description is that there is very high coverage of contributions on Semantics, clearly beating all other levels of descriptions. Within semantics two topics were very prominent: automatic acquisition of semantic properties and the creation of semantically annotated corpora (See Aldo Gangemi's report on Papers on semantic-related topics).
Morphology, though least well represented of the levels of the description, as in previous years, actually showed a small increase in coverage. The other levels of description remained at the same rank, though we did not measure this time for terminology.

Innovation vs. Consolidation

The philosophy behind the LREC conference is that it is a conference where it is important to report not only on what is methodologically new, but also on existing linguistic resources (LR), to describe for which languages LR have been or are being developed, in which state of development they are, and evaluate what is usable in applications. That constitutes LREC's strong industrial relevance, which makes it different from other conferences, e.g. Coling and ACL.
Several trends which had set in earlier showed consolidation and further growth this time.
In particular, automatic and semi-automatic acquisition techniques and machine learning, especially for lexicons, as well as annotation of corpora remained an important topic. One particularly prominent new topic concentrated on research on how to get good results for one's research or technology even if the data on which the research or technology is based are scarce. Related to that, we also saw research on automatically or semi-automatically extending and enriching small resources and resources with no or limited annotation into larger and/or richer resources increasingly represented at LREC-2006.

Resources and Systems

As to the types of resources and systems described at this conference, we see that little changes have occurred at LREC-2006 in comparison to earlier LRECs. In the modern digital world, one sees that new types of communication increase in importance. We already had e-mail, but communications via internet messenger systems ("chatting") and via SMS on mobile phones has started to play an important role in daily life. The text types associated with such new means of communications are of a particular nature, generally quite informal and deviating strongly from other text types. It is becoming increasingly important for natural

Level of Linguistic Description	Genoa	Lisbon	Las Palmas	Athens	Granada
Morphology	5	5	3	2	2
Syntax	**2**	**1**	**1**	**3**	**1**
Semantics	**1**	2	2	1	2
Ontology/Conceptual	3	3	4	5	5
Terminology	not measured	4	5	5	4
Other	4	6	6	4	6
Research vs Development					
(Innovative) Research	1	2	4	3	4
Large Projects	3	4	3	2	1
Tool/System Development	2	1	1	1	3
Policy Issues	4	3	2	4	2
Type of Resource/Tool/other described					
Lexicon	1	1	2	2	2
Corpus	2	2	1	1	1
Methods	5	5	6	6	3
Task/Component	3	3	3	3	5
System	6	4	4	4	4
Infrastructural Aspects	4	6	5	5	5
Language(s)					
One Language	1	1	1	1	1
Many Languages	3	3	3	3	3
Bi-Lingual	2	2	2	2	2

language processing and speech technology systems to be able to deal with such text types. But that requires that LRs for such text types are available. I am therefore happy to see that LREC 2006 had a few contributions in this area, in particular in the area of SMS text corpora. The relevance of such contributions increases because the collection of text corpora for SMS and messenger systems is a non-trivial issue, which involves a proper arrangement of privacy issues, and the development of methodologies to ensure that the collected data are natural and representative. I hope that the few contributions present at LREC 2006 will stimulate and help others to collect similar data, also for other languages.

Languages

As in previous years, most papers deal with a single language. But LREC 2006 showed an increase in the number of papers dealing with bilingual and multilingual resources. And this often took a new form in that methodologies and tools were proposed to create a resource for a new language given similar already existing resources for another language. This can be seen as another instantiation of the trend mentioned earlier to develop methods to extend existing small resources and enhance resources with limited annotation into larger and more richly annotated resources.

Policy Issues and Infrastructural Initiatives

One topic was present at LREC in a very prominent way, not only in the main conference but also in surrounding workshops: proposals and discussions on the creation of a large infrastructure for language resources. Such an infrastructure is intended to improve identification, accessibility, and availability of language resources as well as collaboration on language resource production and enhancement. And with regard to accessibility, it is clearly the intention to make the language resources accessible not only to researchers and developers of natural language processing and speech technologies, but also to a wider group (e.g. researchers in the humanities faculties) that may be benefit from such resources or services around it. This reflects an important trend, in my view, that we surely will hear more about in the near future.

Jan Odijk
Linguistic Resources Division, Speech and Language Technologies
Nuance Communications International (formerly known as ScanSoft Belgium)
Guldensporenpark 32
9820 Merelbeke, Belgium
jan.odijk@nuance.com

Report on Semantic-Related Papers

Aldo Gangemi

Distribution and acceptance rate of papers on semantic-related topics

I assume here "semantics" broadly, with a meaning ranging from linguistics to logic and to cross-disciplinary applications.
The statistics on semantic-related papers submitted to LREC 2006 has been based on the submitted keywords.
The acceptance rate for semantic-related papers has been 66%, a little bit less than the overall acceptance rate (68%). Higher submission number and acceptance rates go to papers marked with Ontologies and Semantics keywords. They mark over 90% of the semantic-related papers. Papers marked with Metadata are just a few.

Keyword overlap

The analysis of keyword overlap shows very little commonality between Metadata papers, Semantics papers, and either Ontologies or Semantic Web papers; these two last keywords show a good overlap:
- 1 overlap between Semantic Web and Semantics
- 9 overlaps between Semantic Web and Ontologies
- 3 overlaps between Ontologies and Semantics
No overlap with Metadata

Analysis of overlap data

We might understand the use of keywords in terms of different research traditions. In the context of computational and corpus linguistics, Semantics has a linguistic meaning, while Metadata has a computer science meaning, and Semantic Web and Ontologies have a web- and/or logic-related meaning.
Two observations can be made on the small overlap. The first one is contingent: there has been a total temporal superposition between LREC and WWW conferences, so that some possible submissions to LREC have been skipped. The second one, probably more reasonable, is that the gap between linguistic semantics and logic-related semantics assumed for ontologies is not yet filled.

Tensions

Following that line of reasoning, I could remark some tensions in this area of lexical resources research.
- Formal vs. lexical semantics. A lot of work has been done in the last months on the relation between formal (logic-related) and lexical semantics, as a reaction to the gap that I've mentioned above. The formal and lexical communities have not yet understood each other at a satisfactory degree. The discussions during LREC 2006 workshop "OntoLex" follow this disappointing pattern
- Web- vs. monolithic resources. The web-oriented work on resources has not yet many relationships to the corpus-linguistics world and traditional lexical resources with semantic annotations. For example, the open attitude of the web world to create large repositories of folksonomies and their direct application to socially relevant resources such as Flickr and de.licio.us follows a totally approach from the traditional way of building linguistic resources and using them for semantic annotations. The actual mess of claims related to so-called metadata supports that impression.
Bottom-line: I highly recommend a more substantial interaction between the linguistic world, which may benefit from the open approaches of the web world and the formal semantic methods of ontologies, and the semantic web world, which could benefit from the reuse of large lexical resources that can be used to match the incredibly large amount of information that is made available on the web. LREC 2008 should take a clearer take towards this recommendation, e.g. by launching open challenges and calls related to the use of lexical resources over the semantic web.

Aldo Gangemi
Laboratory for Applied Ontology
Institute for Cognitive Sciences and Technology
National Research Council (ISTC-CNR)
Via Nomentana 56, 00161, Roma, Italy
Tel: +390644161535
Fax: +390644161513
aldo.gangemi@istc.cnr.it

NEW RESOURCES

ELRA-S0215 UK English Speecon database

The UK English Speecon database comprises the recordings of 606 adult UK English speakers and 51 child UK English speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0216 German Speecon database

The German Speecon database comprises the recordings of 562 adult German speakers and 50 child German speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0217 BITS Logatome Synthesis Corpus – BITS-LG

This corpus contains 11,036 recordings of logatomes spoken by 4 professional German speakers covering all German diphone combinations as well as the most prominent combination German - French – English. Each logatome was recorded in three channels: close microphone, large membrane microphone and laryngographic signal. All diphones are segmented and labelled into phonemic units.

	ELRA members	Non-members
For research use	627.17 Euro	754.35 Euro
For commercial use	4,627.17 Euro	9,000.00 Euro

ELRA-S0218 Speecon manually pitch-marked reference database for Spanish

This database is intended for the development and the evaluation of noise robust pitch marking (PMA) and/or pitch determination (PDA) algorithms. The recordings of 60 speakers were selected from the Speecon Spanish database (ELRA-S0160). The reference database comprises 60 minutes of pitch-marked speech signal.

	ELRA members	Non-members
For research use	150 Euro	900 Euro
For commercial use	300 Euro	2,000 Euro

ELRA-S0219 NEMLAR Broadcast News Speech Corpus

The Nemlar Broadcast News Speech Corpus consists of about 40 hours of Standard Arabic news broadcasts. The broadcasts were recorded from four different radio stations: Medi1, Radio Orient, RMC – Radio Monte Carlo, RTM – Radio Television Maroc. All files were recorded in linear PCM format, 16 kHz, 16 bit.

	ELRA members	Non-members
For research use by academic organisations	150 Euro	300 Euro
For research use by commercial organisations	500 Euro	1,000 Euro
For commercial use	2,000 Euro	4,000 Euro

ELRA-S0220 NEMLAR Speech Synthesis Corpus

The NEMLAR Speech Synthesis Corpus contains the recordings of 2 native Egyptian Arabic speakers (male and female, 35 and 27 years old respectively) recorded in a studio over 2 channels (voice + laryngograph). The recordings comprise more than 10 hours of data with transcriptions.

	ELRA members	Non-members
For research use by academic organisations	500 Euro	1,000 Euro
For research use by commercial organisations	1,250 Euro	2,500 Euro
For commercial use	5,000 Euro	10,000 Euro

ELRA-S0221 OrienTel Egypt MCA (Modern Colloquial Arabic) database

This speech database contains the recordings of 750 Egyptian speakers recorded over the Egyptian fixed and mobile telephone network. Each speaker uttered around 49 read and spontaneous items.

	ELRA members	Non-members
For research use	18,000 Euro	22,500 Euro
For commercial use	24,000 Euro	30,000 Euro

ELRA-S0222 OrienTel Egypt MSA (Modern Standard Arabic) database

This speech database contains the recordings of 500 Egyptian speakers recorded over the Egyptian fixed and mobile telephone network. Each speaker uttered around 49 read and spontaneous items.

	ELRA members	Non-members
For research use	12,000 Euro	15,000 Euro
For commercial use	16,000 Euro	20,000 Euro

ELRA-S0223 OrienTel English as spoken in Egypt database

This speech database contains the recordings of 500 Egyptian speakers of English recorded over the Egyptian fixed and mobile telephone network. Each speaker uttered around 47 read and spontaneous items.

	ELRA members	Non-members
For research use	12,000 Euro	15,000 Euro
For commercial use	16,000 Euro	20,000 Euro

ELRA-S0224 BITS Unit Selection Synthesis Corpus (BITS-US)

This corpus contains 6,732 recordings spoken by 4 professional German speakers covering all German diphone combinations in different prosodic contexts. Each sentence was recorded in three channels: close microphone, large membrane microphone and laryngographic signal. All recordings are segmented and labelled into phonemic units as well as annotated prosodically.

	ELRA members	Non-members
For research use	627.17 Euro	754.35 Euro
For commercial use	4,627.17 Euro	9,000.00 Euro

ELRA-L0065 KORLEX – Croatian Lexicon

The KORLEX - Croatian Lexicon provides a list of 118,252 Croatian lemmas (including 52,450 nouns, 8,985 adverbs, 14,937 verbs and 41,161 adjectives, as well as pronouns, determiners, prepositions/postpositions, conjunctions and numerals), i.e., words in canonical form, annotated with part-of-speech (POS) tag and lexical features. The lexicon data is compiled with the objective of covering the majority of text circulating in everyday use, such as in the news, in business, technological documentation, legal documentation, and politics. The resource is a flat textual file in which each textual line contains information about one lemma. The resource is encoded using ISO-8859-2 encoding, and sorted according to the standard Croatian lexicographic order.

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

ELRA-L0066 KORLEX – Serbian Lexicon

The KORLEX - Serbian Lexicon provides a list of 108,491 Serbian lemmas (including 52,027 nouns, 9,153 adverbs, 15,522 verbs and 31,052 adjectives, as well as pronouns, determiners, prepositions/postpositions, conjunctions and numerals), i.e., words in canonical form, annotated with part-of-speech (POS) tag and lexical features. The lexicon data is compiled with the objective of covering the majority of text circulating in everyday use, such as in the news, in business, technological documentation, legal documentation, and politics. The resource is a flat textual file in which each textual line contains information about one lemma. The resource is encoded using ISO-8859-2 encoding, and sorted according to the standard Serbian lexicographic order.

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

ELRA-L0067 English lexicon with morphological information

This English lexicon is made up of 174,000 inflected forms corresponding to 68,000 simple word lemmas (including 31,900 nouns, 11,800 verbs, 19,900 adjectives, 4,100 adverbs, 300 pronouns, articles, prepositions/postpositions and conjunctions). Each line in the resource file shows an inflected form, its part of speech, its related lemma and its morphological information.

	ELRA members	Non-members
For research use by academic organisations	3,500 Euro	4,500 Euro
For research use by commercial organisations	5,000 Euro	7,000 Euro
For commercial use	6,000 Euro	8,500 Euro

ELRA-L0068 French lexicon with morphological information

This French lexicon is made up of 424,000 inflected forms corresponding to 55,000 simple word lemmas (including 34,400 nouns, 7,300 verbs, 11,700 adjectives, 1,400 adverbs, 200 pronouns, articles, prepositions/postpositions and conjunctions). Each line in the resource file shows an inflected form, its part of speech, its related lemma and its morphological information.

	ELRA members	Non-members
For research use by academic organisations	3,500 Euro	4,500 Euro
For research use by commercial organisations	5,000 Euro	7,000 Euro
For commercial use	6,000 Euro	8,500 Euro

ELRA-L0069 Italian lexicon with morphological information

This Italian lexicon is made up of 862,500 inflected forms corresponding to 112,000 simple word lemmas (including 66,340 nouns, 12,030 verbs, 28,080 adjectives, 4,890 adverbs, 660 pronouns, articles, prepositions/postpositions and conjunctions). Each line in the resource file shows an inflected form, its part of speech, its related lemma and its morphological information.

	ELRA members	Non-members
For research use by academic organisations	5,500 Euro	7,000 Euro
For research use by commercial organisations	6,500 Euro	8,500 Euro
For commercial use	8,000 Euro	10,000 Euro

ELRA-L0070 Italian lexicon with morphological information and clitic verbs

This Italian lexicon is the same as the one described in ELRA-L0069, but with the addition of clitic verbs, which increases the number of inflected forms to 1,800,000 (still corresponding to 112,000 simple words lemmas). It contains 66,340 nouns, 12,030 verbs, 28,080 adjectives, 4,890 adverbs, 660 pronouns, articles, prepositions/postpositions and conjunctions. Each line in the resource file shows an inflected form, its part of speech, its related lemma and its morphological information.

	ELRA members	Non-members
For research use by academic organisations	6,500 Euro	8,500 Euro
For research use by commercial organisations	8,000 Euro	10,000 Euro
For commercial use	10,000 Euro	12,500 Euro

ELRA-L0071 Spanish lexicon with morphological information

This Spanish lexicon is made up of 816,000 inflected forms corresponding to 104,000 simple word lemmas (including 52,000 nouns, 9,800 verbs, 21,200 adjectives, 20,500 adverbs, 500 pronouns, articles, prepositions/postpositions and conjunctions). Each line in the resource file shows an inflected form, its part of speech, its related lemma and its morphological information.

	ELRA members	Non-members
For research use by academic organisations	5,500 Euro	7,000 Euro
For research use by commercial organisations	6,500 Euro	8,500 Euro
For commercial use	8,000 Euro	10,000 Euro

ELRA-L0072-01 PAROLE-SIMPLE-CLIPS PISA Italian Lexicon – Full lexicon

PAROLE-SIMPLE-CLIPS is a four-level, general purpose lexicon that has been elaborated over three different projects. The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon comprises a total of 387,267 phonetic units, 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon was encoded at the semantic level, in full accordance with the international standards set out in the PAROLE-SIMPLE model and based on EAGLES. Syntactic and semantic encoding were performed jointly with Thamus (Consortium for Multilingual Documentary Engineering), which is responsible for 25,000 extra entries (to be released soon).
This lexicon is subdivided into five different subsets:
L0072-01 Full lexicon
L0072-02 Phonetic layer
L0072-03 Morphological layer
L0072-04 Syntactic layer
L0072-05 Semantic layer

	ELRA members	Non-members
For research use	1,500 Euro	2,000 Euro
For commercial use	12,000 Euro	15,600 Euro

ELRA-L0072-02 PAROLE-SIMPLE-CLIPS PISA Italian Lexicon – Phonetic layer

PAROLE-SIMPLE-CLIPS is a four-level, general purpose lexicon that has been elaborated over three different projects. The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon comprises a total of 387,267 phonetic units, 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon was encoded at the semantic level, in full accordance with the international standards set out in the PAROLE-SIMPLE model and based on EAGLES. Syntactic and semantic encoding were performed jointly with Thamus (Consortium for Multilingual Documentary Engineering), which is responsible for 25,000 extra entries (to be released soon).
This lexicon is subdivided into five different subsets:
L0072-01 Full lexicon
L0072-02 Phonetic layer
L0072-03 Morphological layer
L0072-04 Syntactic layer
L0072-05 Semantic layer

	ELRA members	Non-members
For research use	600 Euro	2,000 Euro
For commercial use	4,800 Euro	15,600 Euro

ELRA-L0072-03 PAROLE-SIMPLE-CLIPS PISA Italian Lexicon – Morphological layer

PAROLE-SIMPLE-CLIPS is a four-level, general purpose lexicon that has been elaborated over three different projects. The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon comprises a total of 387,267 phonetic units, 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon was encoded at the semantic level, in full accordance with the international standards set out in the PAROLE-SIMPLE model and based on EAGLES. Syntactic and semantic encoding were performed jointly with Thamus (Consortium for Multilingual Documentary Engineering), which is responsible for 25,000 extra entries (to be released soon).
This lexicon is subdivided into five different subsets:
L0072-01 Full lexicon
L0072-02 Phonetic layer
L0072-03 Morphological layer
L0072-04 Syntactic layer
L0072-05 Semantic layer

	ELRA members	Non-members
For research use	375 Euro	500 Euro
For commercial use	3,000 Euro	3,900 Euro

ELRA-L0072-04 PAROLE-SIMPLE-CLIPS PISA Italian Lexicon – Syntactic layer

PAROLE-SIMPLE-CLIPS is a four-level, general purpose lexicon that has been elaborated over three different projects. The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon comprises a total of 387,267 phonetic units, 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon was encoded at the semantic level, in full accordance with the international standards set out in the PAROLE-SIMPLE model and based on EAGLES. Syntactic and semantic encoding were performed jointly with Thamus (Consortium for Multilingual Documentary Engineering), which is responsible for 25,000 extra entries (to be released soon).
This lexicon is subdivided into five different subsets:
L0072-01 Full lexicon
L0072-02 Phonetic layer
L0072-03 Morphological layer
L0072-04 Syntactic layer
L0072-05 Semantic layer

	ELRA members	Non-members
For research use	375 Euro	500 Euro
For commercial use	3,000 Euro	3,900 Euro

ELRA-L0072-05 PAROLE-SIMPLE-CLIPS PISA Italian Lexicon – Semantic layer

PAROLE-SIMPLE-CLIPS is a four-level, general purpose lexicon that has been elaborated over three different projects. The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon comprises a total of 387,267 phonetic units, 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). The PAROLE-SIMPLE-CLIPS Pisa Italian Lexicon was encoded at the semantic level, in full accordance with the international standards set out in the PAROLE-SIMPLE model and based on EAGLES. Syntactic and semantic encoding were performed jointly with Thamus (Consortium for Multilingual Documentary Engineering), which is responsible for 25,000 extra entries (to be released soon).
This lexicon is subdivided into five different subsets:
L0072-01 Full lexicon
L0072-02 Phonetic layer
L0072-03 Morphological layer
L0072-04 Syntactic layer
L0072-05 Semantic layer

	ELRA members	Non-members
For research use	150 Euro	200 Euro
For commercial use	1,200 Euro	1,600 Euro

ELRA-W0042 NEMLAR Written Corpus

The NEMLAR Written Corpus consists of about 500,000 words of Arabic text from 13 different categories. The corpus is provided in 4 different versions: raw text, fully vowelized text, text with Arabic lexical analysis, text with Arabic POS-tags.

	ELRA members	Non-members
For research use by academic organisations	150 Euro	300 Euro
For research use by commercial organisations	250 Euro	500 Euro
For commercial use	1,000 Euro	2,000 Euro

ELRA-W0043 PAROLE Italian Corpus

The PAROLE Italian Corpus comprises 3,135,651 words collected from four different domains: newspapers (2,179,800 words), periodicals (143,810 words), books (564,964 words), miscellaneous (247,077 words). Data are morphosyntactically annotated and lemmatized.

	ELRA members	Non-members
For research use	100 Euro	150 Euro
For commercial use	1,500 Euro	2,500 Euro

ELRA-W0044 Italian Syntactic-Semantic Treebank (ISST)

ISST comprises 89,941 tokens for the financial-domain part and 215,606 tokens for the general part. It is formatted in XML. This Treebank has a five-level structure covering orthographic, morpho-syntactic, syntactic; semantic and lexico-semantic levels of linguistic description. Syntactic annotation is distributed over two different levels: the constituent structure level and the functional relations level. The fifth level deals with lexico-semantic annotation, which is carried out in terms of sense tagging of lexical heads (nouns, verbs and adjectives) augmented with other types of semantic information: ItalWordNet (see ELRA-M0018) is the reference lexical resource used for the sense tagging task . Both syntactic and lexico-semantic annotations refer to the morpho-syntactically annotated text, which in turn is linked to the orthographic file with the text and mark-up of macrotextual organisation (e.g. titles, subtitles, summary, body of article, paragraphs).

	ELRA members	Non-members
For research use	100 Euro	150 Euro
For commercial use	1,500 Euro	2,500 Euro

ELRA-E0008 The CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package

The CLEF Test Suite contains the data used for the main tracks of the CLEF campaigns carried out from 2000 to 2003: Multilingual text retrieval, Bilingual text retrieval, Monolingual text retrieval, and Domain-specific text retrieval. It contains multilingual corpora in English, French, German, Italian, Spanish, Dutch, Swedish, Finnish, Russian, and Portuguese.

	ELRA members	Non-members
For evaluation use by academic organisations	150 Euro	300 Euro
For evaluation use by commercial organisations	500 Euro	1,000 Euro

PRESS RELEASE - PARIS, FRANCE

Distribution Agreement

ELRA today signed a major Language Resources distribution agreement with Beijing Haitian Ruisheng Science Technology Ltd.

ELRA and Beijing Haitian Ruisheng Science Technology Ltd today signed a major Language Resources distribution agreement. On behalf of ELRA, ELDA will act as the distribution agency for Beijing Haitian Ruisheng Science Technology Ltd and will incorporate to the ELRA Language Resources catalogue a large number of Speech resources designed and collected to boost Speech Synthesis and Speech Recognition. The resources cover mainly Mandarin Chinese with some coverage of Korean and Japanese languages.
With over 60 new resources, ELDA is strengthening its position as the leading worldwide distribution centre. With this agreement Beijing Haitian Ruisheng Science Technology Ltd will get more visibility in particular on the European market.

About ELRA

The European Language Resources Association (ELRA) is a non-profit making organisation founded by the European Commission in 1995, with the mission of providing a clearing house for language resources and promoting Human Language Technologies (HLT).
To find out more about ELRA, please visit our web site: www.elra.info

About ELDA

The Evaluation and Language resources Distribution Agency (ELDA) is ELRA operational body. ELDA identifies, collects, markets, and distributes language resources, along with the dissemination of general information in the field of HLT. ELDA also participates in some evaluation projects and campaigns, has considerable knowledge and skills in HLT applications and has participated in many French, European and international projects.
To find out more about ELDA, please visit our web site: www.elda.org

About Beijing Haitian Ruisheng Science Technology Ltd / Kingline Data Center

With rich experience in speech technology, the Kingline Data Center has concentrated our time on speech data processing since 1998. Till now, 500 hours of high quality speech synthesis corpora (read by professional speakers in Chinese, Japanese, English, Spanish, etc.) and 4,000 hours of speech recognition corpora (recorded with various microphones, desktop phones, mobile phones, and in-Vehicle phones) have been collected and processed.
To find out more about Beijing Haitian Ruisheng Science Technology Ltd / Kingline Data Center, please visit our web site: www.speechocean.com/indexe.asp

More information on the ELRA catalogue, please contact:
Valérie Mapelli
mapelli@elda.org

More information on ELRA & ELDA , please contact:
Khalid Choukri
choukri@elda.org
Hélène Mazo
mazo@elda.org

ELDA
55-57, rue Brillat Savarin
75013 Paris (France)
Tel.: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30