LREC 2004

IV INTERNATIONAL CONFERENCE

On Language Resources and Evaluation

The ELRA Newsletter

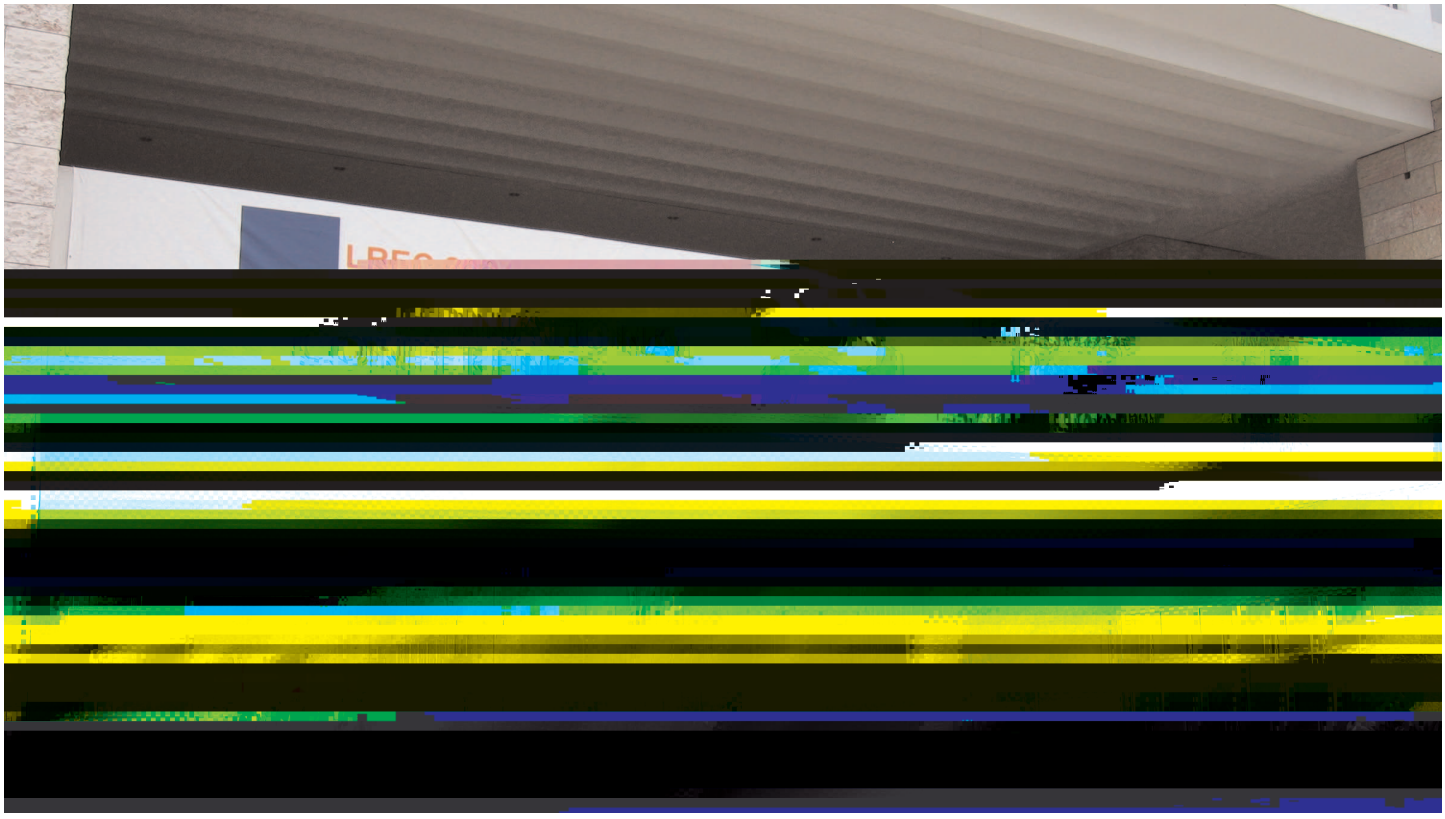
LREC 2004 Special Issue

Vol. 9 n° 2 & 3

ISSN: 1026-8200

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

Contents



Dear Colleagues,

The 4th edition of the Language Resources and Evaluation Conference took place last May in Lisbon, Portugal. It was held in memory of two dear friends we lost recently, Angel Martin Municio and Antonio Zampolli.

Close to 800 submissions for poster and oral presentations were submitted and were reviewed by the Scientific Committee: 519 were actually presented in Lisbon, where LREC 2004 was an indubitable success.

About 900 participants from 50 countries enjoyed this fruitful event in HLT, with its rich and varied conference programme. Out of 519 papers, a majority was dedicated to written resources (260), 116 dealt with spoken resources, 40 with terminological issues, 57 with evaluation, 29 on multimodal-multimedia issues and 17 were on general ones.

In addition, 18 satellite workshops covering various fields were organised before and after the main conference. These workshops covered topics as diverse as minority languages processing, parallel and comparable corpora, XML-based richly annotated corpora, speech corpus production and validation, or the representation and processing of sign languages.

One of the workshops held at LREC 2004 was a joint event between COCOSDA, the International Committee for Co-ordination and Standardisation of Speech Databases, and ICCWLRE, the International Co-ordination Committee for Written Language Resources and Evaluation.

The ICCWLRE was recently launched and aims to support international cooperation and coordination in the field of Written Language Resources and Evaluation, to set research priorities and discuss the needs in the field for the future. This new Committee for WLR and Evaluation is based on the same model as Cocosda for SLR; one of the main objectives is to share the knowledge and experiences learnt from both SLR and WLR areas and better coordinate the activities conducted in these areas.

LREC 2004 was the occasion to honour Antonio Zampolli. In addition to dedicating the conference to his memory, it was the right place to announce the awarding of the 1st Antonio Zampolli Prize. This Prize was established by the ELRA Board to honour Antonio Zampolli who was ELRA co-founder and first President, from 1995 to 2002.

The Antonio Zampolli Prize was awarded to Fredrick Jelinek, from John Hopkins University, in Baltimore, USA. At the Closing Ceremony, he gave an impressive talk, entitled “My Best Friends were Linguists”, attended by a large audience. The slides of his presentation are downloadable from the LREC 2004 web site: www.lrec-conf.org/lrec2004/

The LREC conference is a biennial event: it was decided at the end of this edition that in 2006, LREC will be organised in Genoa, Italy. We hope to meet you there!

ELRA took the opportunity of LREC 2004 to organise its Annual General Members Assembly.

The ELRA Board was renewed, with 4 new members: Gregor Thurmair, from Linguatec (Germany), Jimmy Kunzmann, from IBM (Germany), Asuncion Moreno, from UPC (Spain) and Martine Garnier-Rizet, from Vecsys (France) joined the Board.

In addition, the new President of the association was elected: Bente Maegaard (CST, Denmark) has replaced Joseph Mariani, whose term was finished.

On behalf of all our members, we would like to thank him as well as the other Board members who left for their contribution to the success and advances at ELRA.

Now concerning the content of this ELRA newsletter dedicated to LREC 2004, we decided to have a double special issue, due to the high number of contributions from authors and presenters at LREC 2004.

We received many sessions' summaries, as well as workshops' reviews, and we are happy to offer in the ELRA newsletter an overview of this LREC conference thanks to these numerous contributions, dealing for example with corpus annotation and evaluation, multimodal corpora, corpus and lexicon tools, question-answering, evaluation of speech annotation and systems, machine translation, or computational lexicons. Apart from these, Opening Ceremony speeches and conference reports are also included.

Last but not least, the new resources added to the ELRA catalogue are listed at the end of this newsletter: three new speech databases and three new written corpora.

If you would like to offer your resources to the HLT community and distribute them via ELRA and ELDA, you are kindky invited to contact us (contact details provided on the front cover).

Khalid Choukri, CEO

INTRODUCTION
by Nicoletta Calzolari, LREC 2004 Conference Chair

Antonio Zampolli launched the idea of a Language Resources and Evaluation Conference (LREC) during an ELRA Board meeting. And Angel Martin Municio proposed Granada for the first LREC. It was the perfect combination for a new adventure. Which continued with Athens, Las Palmas and Lisbon… where Antonio and Angel were with us only in spirit. But they were with us. LREC is a creature of Antonio Zampolli, one of the many initiatives put in motion by him, one he loved so much. He understood that, despite the many conferences, there was not only the space but the need of an event that could gather all those working in the areas of LRs and Evaluation.

LRs and Evaluation have, as Antonio understood very well, an infrastructural role for Language Technology (LT) and traverse horizontally every applicative area of HLT, as LREC 2004 clearly testifies. If we have a look at the programme, we see such a broad spectrum of tools, components, systems, applications represented, that we may ask "is this LREC?", but it is rightly so. Summarisation, question answering, machine translation, speech-to-speech translation, cross-lingual information retrieval, information extraction, document classification, automatic indexing of broadcast news, topic detection, text mining, e-learning, to mention just a few, need data, often lots of data. And need evaluation, and good methods for evaluation.

LRs occupy more and more space in our everyday work and are inevitably intermingled with algorithms, tools, systems, applications, etc. But I notice an important shift of focus in the various editions of LREC, from papers on 'data almost per se' to 'what we use the data for' and 'how we use the data'. This is an important and interesting change of perspective.

LRs is also a 'sensitive' issue, charged of political, social, cultural, economic, commercial, and -more and more recently- strategic implications (defense, security, etc.). As pointed out very well in the final Euromap Report, it is crucial that LTs for all languages are promoted (and that coordination initiatives are put in place), if we want to avoid a two-speed situation between languages which are interesting commercially, and today also politically, and those which are not (unfortunately the vast majority). That is why we, at LREC, recognise the importance of giving visibility and providing a large forum for discussion to policies for LR creation in different countries and for different languages, and to infrastructural issues such as distribution, cooperation, standardisation, etc.

LREC has always targeted all the communities of spoken, written and multimodal LRs, and in this LREC in particular, in addition to the usual LREC tracks -Evaluation, Multimodality, Speech, Terminology, Written- we decided to start having some 'mixed' session, and see how well this is accepted by the participants. The goal is to favour integration among different communities. We strongly believe that integration of the different, until recently rather separate, communities is an essential step for a comprehensive approach to communication, which is made up by different modalities and their complex interactions. LREC is also special in this respect, because it is one of the few conferences that really targets all these communities at the same time and at the same level of importance. This is a must for our field to contribute to the big challenges of the 'knowledge-based society'.

We think it is important to have a conference providing an overview of "what exists", not only of what is new. To report not only on what is methodologically new, but also on which LRs exist, for which languages, in which state of development, and evaluate what is usable in applications. Consolidation -which goes together with "robustness"- is therefore at least as relevant as innovation, to get hold of the situation of LRs (particularly important for industrial exploitation). With this characterisation, I think LREC allows an assessment of the level of maturity not only of the field of LRs, but of HLT in general, because of the clear interaction between LRs and HLT.

If we compare the content of the four editions of LREC, and try to make even a very superficial and cursory analysis of the prominent areas covered in the four conferences, we cannot avoid noticing a number of trends. The field is evolving, and these trends reflect very clearly the evolution of the field and the emerging needs, and provide us with a picture of where our field moves, and how it changes. Just a few quick remarks:

1. The focus of the attention is moving - on the continuum of the LR space - from one edition to the other: from issues of morphology and tagging, to grammars and treebanks (many in '02), then terminology and knowledge, semantics, semantic web and ontologies, pragmatics, multimodal dialogue, and how to model emotions (there was no paper on emotions in '98).

2. An impressive amount of papers this time are on 'how to acquire data', i.e. about methodologies and techniques for machine learning, automatic acquisition and/or classification of information. Acquisition techniques aim at creating LRs, and at the

same time rely on LRs, at some stage either of implementation or of evaluation, creating a virtuous loop. They are the real trend and the challenge of the last years, and one of the most promising research areas for the next years.

3. In this edition we had not only so many tools, components, systems, applications, but it emerged, and it is very recent, the recognition of the strategic importance, both in political and economic terms, of being able to build a new system for a given language in a very short time, or to adapt or tune an existing one very quickly, and this crucially depends on the availability of large quantities of data and on the ability to process them.

4. A quite new paradigm is also emerging, in a few papers, involving initiatives aiming at open and distributed infrastructures, for cooperative and controlled creation and maintenance of LRs. This is only feasible when the field as a whole has reached a level of stability and maturity. This may become the new 'vision' for LRs in the years to come.

5. The other pillar is Evaluation, without which no technology is credible. Many evaluation resources and many evaluation methodologies are presented at this LREC: evaluation in many cases of resources, tools or systems where semantics is at stake, from evaluation of disambiguation systems to ontology platforms, from machine translation to summarisation. Both American and European large evaluation campaigns are well represented. Also validation of LRs themselves acquires more and more importance, as a fundamental step to accompany any distribution activity. Validation is closely linked to standards.

We received an incredible number of submissions. However the success has brought with it also practical and organisational concerns. We were faced with the dilemma: should we maintain the size of the last LREC and reject many submissions, or we remain faithful to the policy of providing the broadest picture of the field of LRs and evaluation, obviously preserving quality? We have decided for the second option. This meant accepting an incredibly high number of papers, between orals and posters. This decision has also forced us to decide to reduce the length of the papers to 4 pages, to avoid ending up with Proceedings of 10 or 12 volumes! And we had incredibly large poster sessions (about 100 each day)! We certainly need to think about these issues for 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.

The figures of submissions, papers, the fact that participants were so numerous in Lisbon (almost 1000) proves that Antonio was right. Antonio would be proud of this, I believe. We have dedicated this 4th edition of LREC to Antonio Zampolli. In particular we had a special plenary session with three of the 'oldest' friends of Antonio speaking to him and for him: Bernard Quemada, Martin Kay, Makoto Nagao. I think he would have liked that.

But the true protagonist of LREC were the participants, who have made this LREC great. With all the Programme Committee, all the other committees, and somehow together with Antonio, I thank all of them and … wait for them as numerous and enthusiastic as this time at LREC 2006!

Acknowledgments

And now it is time for thanking all those who have made this LREC possible.

First of all I deeply thank the Programme Committee (PC), a very special PC which is more a group of old friends. Then I thank with sympathy the groups in Paris and Pisa, in particular: Magali Jeanmaire, Louis-Gabriel Pouillot, Sara Goggi, Sergio Rossi and Vincenzo Parrinelli. I thank our impressively large Scientific Committee, and our Advisory Board, for their important cooperation. We are also indebted to the ELRA Board, and to authorities, associations, organisations, committees, agencies, companies that have supported LREC in various ways. We particularly thank Microsoft, IBM, Priberam Informática, Porto Editora for their sponsorship to the Conference. I thank the workshop organisers, and obviously all the authors, who provided the content to LREC, giving us such a broad picture of the field. I am specially grateful to Martin Kay, Makoto Nagao and Bernard Quemada, for speaking, representing all of us, in the session in memory of Antonio. Finally I thank the fantastic Lisbon team, headed by Teresa Lino, with their enthusiasm and dedication. And at the very end my biggest thank goes to all the participants, hoping that they could profit of so many contacts to organise new exciting work in the field of LRs and evaluation, to be shown at the next LREC.

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
Email: glottolo@ilc.cnr.it
Website: www.ilc.cnr.it/

LREC 2004 Opening Ceremony Speeches

Joseph Mariani, ELRA President

LREC, the Language Resources and Evaluation Conference has now become the regular rendez-vous of those who believe language resources and evaluation are of crucial importance for the development of written and spoken language science and technology.

But this fourth issue of LREC is different from the three previous ones, as we are deeply missing our friend Antonio Zampolli, the first president of ELRA, and general chairman of the first three LREC conferences, who died in August 2003. And we also miss Angel Martin Municio, ELRA Vice-President and a major actor in the decisions of having LREC in Spain - Granada in 1998 and Las Palmas in 2002 - who died in November 2002.

Since its creation in 1995, ELRA, the European Language Resources Association, has developed a lot its activity, in strong relationship with ELDA, its Evaluation and Language Distribution Agency. With close to 100 members, and more than 700 resources in its catalogue, ELRA now appears as a major actor in the field of language technologies worldwide. Its initial activity was only related to Language Resources distribution. Since then, it was extended very naturally to Language Resources validation, and, more recently, to Language Resource production and Language Technology evaluation.

The activities of the association depend deeply on the participation and initiatives of the Board members. I take this opportunity to thank all those who participated in ELRA, since its very beginning. I would like to mention especially those who are quitting the Board, in agreement with the statutes of the association, which limit the number of consecutive terms to three: Daniel Tapias, ELRA secretary, Harald Höge, ELRA Treasurer, and Volker Steinbiss, ELRA Vice-President. I will also quit the Board for the same reason, with the satisfaction of having participated in the founding of a successful initiative, starting from the European Commission Relator project, where the idea of a European association on Language Resources was worked out, and with the pleasure to have now an healthy entity, benefiting from the support of ELDA, which has now close to 20 employees.

I would also like to thank the actors who participated in the definition and creation of ELRA, with a special mention to the representatives of the European Commission who helped the association in its early days, and especially Vicente Parajon-Collada, Roberto Cencioni and Nino Varile, and to the Relator group of high level consultants, comprising Brian Oakley, André Danzin and Bernard Quémada.

The ELRA General Assembly took place yesterday, and Bente Maegaard has been elected as ELRA new president. The new office will be in place starting tomorrow.

The idea of creating a scientific conference in the field of Language Resources and Evaluation, in order to meet the needs of the Language Technology community, both scientific and industrial, was expressed at an ELSNET Advisory Board meeting, and was immediately submitted at the next ELRA board meeting. The initiative has been a success from the very first conference in Granada, and each time since then, in Athens, Las Palmas and now Lisbon, the number of papers submitted and presented, and the number of participants increased. Antonio and Angel would have been proud to announce more than 500 papers and more than 800 attendees from 50 different countries this time. I dedicate this success to their memory.

LREC 2004 will also be the place where the Antonio Zampolli prize will be awarded for the first time, in order to recognize outstanding contributions to the advancement of Language Resources and Language Technology Evaluation.

Language Technology appears as a very active field of research and development. More and more actions are gathering the specialists of spoken language processing and written language processing altogether, with links to other communication media, such as vision and gesture. The coverage of that field by the European Commission went from a specific program on Human Language Technologies in FP5, to larger programs on Multisensorial interfaces and Knowledge management in FP6. Now we are preparing FP7, the 7th European Framework Program.

Even if the efforts devoted to that field have been large for many years, even if they resulted in many applications and products which have been put on the market, even if Language Technologies are used everyday, embedded in various devices and services, it clearly appears that the arcane of processing language is still unsolved and needs further and larger efforts, both at the basic scientific level, and at the system development one.

In Europe, the integration of 10 new member states in the European Union has enlarged the number of languages, and combinations of languages, that have to be considered in order to address both the need to preserve the individual language and culture of all Member States and regions, and the need to communicate within a large community of countries respectfully of all its constituents.

The European Commission has not enough means and forces to cover the effort aiming at providing the various Language Technologies for all those different languages. Our proposal is therefore to join forces in order to better coordinate each national effort, addressing its language, or languages, and taking care of the availability of all the Language Resources which are necessary to develop those various technologies, and the activity of the European Commission, of generic and organizational nature. This is in full agreement with the concept of subsidiarity, and with the spirit of the European Research Area.

But this question is in fact wider and can be placed at the international level. The need for better exchanges on spoken language resources and evaluation was expressed very early by the speech community, with the creation of Cocosda, the Coordinating Committee on Speech Databases and speech I/O systems Assessment, back in 1991. I'm glad to see that a comparable initiative has been developed for the written language community, with the recent creation of an International Coordinating Committee on Written Language Resources and Evaluation, which will meet for the first time together with Cocosda during this LREC conference.

Finally, I would like to thank the general chair of the conference, Nicoletta Calzolari, and the CNR team in Pisa, Khalid Choukri and the ELDA team in Paris, Teresa Lino and her team and friends in Lisbon and in Portugal, the International Advisory Committee, the Program Committee and the Scientific Committee for the tremendous work they achieved to make this conference a wonderful and renewed success.

Enjoy !

Khalid Choukri, ELRA CEO

Dear LREC Participants,

Welcome to LREC 2004, welcome to Lisbon!

ELRA (European Language Resources Association), its operational body and distribution agency ELDA (Evaluations and Language resources Distribution Agency) and the Universidade Nova de Lisboa are proud to welcome you in Lisbon, where we are pleased to organise the fourth edition of the Language Resources and Evaluation Conference, LREC 2004. We are very pleased to continue the organisation of such an important event in such an attractive city.

LREC 2004 is the fourth biennial conference on Language Resources and Evaluation, the fourth in a very successful series of events since ELRA initiated it with the strong involvement of Antonio Zampolli and Angel Martin Municio.

This event is held in Memory of Antonio Zampolli. Antonio initiated LREC in 1998, as founder and first President of ELRA, and has largely contributed over the past years and with the 3 previous LRECs to the success of the event. It is also an opportunity for us to remember Angel Martin Municio, whom we lost in November 2002.

To honour the memory of Antonio Zampolli and acknowledge his contribution to the set up of ELRA and LREC, the ELRA Board decided to create a Prize to award individuals whose work lies within the areas of language resources and language technology evaluation, with acknowledged contributions to their advancements: The Antonio Zampolli Prize will be awarded for the first time here in Lisbon.

Before giving you some practical details about the next few days, let me say a few words about ELRA and LREC: I think that to better understand LREC, it is necessary to elaborate a little bit on ELRA.

ELRA was founded in 1995, with the strong dedication of Antonio Zampolli, and 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, ELRA's 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 related projects, at the French, European and international levels.

ELRA now counts around 100 members, who belong to academic and industrial organisations involved in the use and exploitation of language resources for research and/or language technologies development or evaluation. ELRA members are offered several advantages, in particular reduced prices on the language resources available in the catalogue: at the end of 2003, ELRA's catalogue counted around 750 language resources, distributed in three colleges, namely Spoken Language Resources (SLR), Written Language Resources (WLR) and Terminological resources.

Teresa Lino, Ramoa Ribeiro, Leopoldo Guimarães, Nicoletta Calzolari, Joseph Mariani, Khalid Choukri

The language resources can be purchased by members and non-members: the whole HLT community is thus offered the possibility to access the catalogue on-line, and to buy any needed resources. In 2003, over 360 language resources were distributed.

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, but other crucial services related to language resources and language technologies are also offered. These include the validation of language resources, carried out with the support of ELRA's network of validation centres thus ensuring the best quality of the language resources presented in the catalogue, and the production of language resources, mainly SLR within projects ELRA and ELDA participate in; the evaluation of speech and language technologies is another major activity, 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 language resources distribution. More recently, it was agreed to strengthen our position in the standardisation area, getting further involved in related initiatives.

If you would like to learn more about ELRA and ELDA, you are invited to visit our web sites, at www.elra.info and www.elda.fr, and to get in touch with us. The ELRA/ELDA staff is at your disposal here in Lisbon during this week.

On behalf of ELRA, and on your behalf, I would like to warmly thank the local team in Lisbon responsible for the practical aspects of this event. As you can imagine, organising such an important event, in particular once our expectations have been revised upwards, from 700 attendees to about 1000, is not an easy task to carry in addition to the daily commitment of a university staff. I would like to thank very much Teresa Lino for having managed such organisation and extend this to her team.

I would like to thank all members of the Scientific Committee for their valuable help to review the 790 submitted papers, as well as workshops' organisers for contributing to the success of LREC.

Let me take this opportunity to thank a number of organisations which have helped or contributed to the organisation of LREC 2004: the official LREC sponsors, IBM, Microsoft, Porto Editora, Priberam; the supporters, ILC CNR , the Portuguese Fundation of Sciences and Technologies, Fundation Camoes, Institut Franco-portugais, Institut Cervantes, Instituto Italiano de Cultura, and Fundation Calouste Gulbenkian.

LREC is organised by ELRA with the support of a very large number of organisations, including ACL, AFNLP, ALLC, ALTA, COCOSDA and Oriental COCOSDA, EAFT, EAMT, ELSNET, ENABLER, EURALEX, GKS, GWA, IAMT, ICWLR, ISCA, LDC, ONTOWEB, TEI, and with major national and international organisations, including the Commission of the EU - Information Society DG, Unit E1 “Interfaces and Cognition”. We are very grateful to all of them.

I wish you all a very fruitful and successful LREC!

LREC 2004 Antonio Zampolli Prize

Speech given by Joseph Marinai

In order to honor the memory of Antonio Zampolli, the ELRA Board has decided to create the Antonio Zampolli prize, which will be awarded every two years at the LREC conference to an individual, in recognition of outstanding contributions to the advancement of Language Resources and Language Technology evaluation, for the progress of human language science and technology.

The prize consists of a medal, a certificate, and its amount is 10,000 Euros.

Nominees should be proposed by at least three individuals from three different institutions. We received for this first attribution of the prize the proposals of eight nominees in due time.

The ELRA Board, during its meeting of April 3rd, selected the winner, and I'm glad to announce that the 2004 Antonio Zampolli prize is awarded to Fredrick Jelinek.

Fredrick Jelinek started his career as a teaching assistant at MIT, where he got his PhD.

He then taught at Harvard, before rejoining Cornell University as assistant professor, then professor.

In 1972, he was appointed to a position of senior manager at the IBM T.J. Watson Research Center, where he managed the very well known speech group during 20 years.

He moved back to academia in 1993, and rejoined the Johns Hopkins university, as a professor and Director of the Center for Language and Speech Processing.

He received IEEE awards from the Signal Processing Society and from the Information Theory Society, and he is the recipient of the 1999 ESCA Medal.

Fred Jelinek is a pioneer in the statistical processing of speech and language in various areas: speech recognition, machine translation, text parsing and understanding.

The famous expression “There is no better data than more data”, that we very much like at ELRA, comes from a member of his team at IBM, Bob Mercer.

His centre organizes every year a summer school, where students and researchers develop a language processing system based on the use of Language Resources and on evaluation.

For all those reasons, the ELRA Board decided to award him the 2004 Antonio Zampolli Prize.

*The presentation given by Fredrick Jelinek, entitled **“Some of my Best Friends are Linguists”,** can be viewed from the LREC 2004 web site:*

www.lrec-conf.org/lrec2004

LREC 2004 Sessions' Summaries

Summary of the Oral Session "Corpus Annotation and Evaluation"

Nelleke Oostdijk

In the session on Corpus Annotation and Evaluation, the following three papers were presented: *A Labelled Corpus for Prepositional Phrase Attachment* (by Brain Mitchell and Robert Gaizauskas, presented by Louise Guthrie), *Annotators' Agreement: The Case of Topic-Focus Articulation* (by Katerina Veselá, Jiri Havelka and Eva Hajicová, presented by Eva Hajicová) and *A Word Alignment System Based on a Translation Equivalence Extractor* (by Ana-Maria Barbu).

The first paper describes the development of a resource that can be used for training machine learning algorithms directed at the automatic attachment of prepositional phrases. In their approach the authors investigate the five most common patterns of PP-attachment and investigate what are potentially useful data features, including features that have not been used previously. Novel data features are lexical and phrasal distances from a preposition to its attachment point and phrase function tags as they appear in the Penn Treebank II.

The second paper reports the results obtained in evaluating the annotation of topic-focus articulation in the Prague Dependency Treebank, while it also describes the measures that have been developed in order to increase interannotator consistency. The findings lead the authors to conclude that the annotation of this kind is indeed feasible, provided that the annotation has been adequately elaborated theoretically and annotators can refer to a comprehensive manual for guidance.

The last paper describes a new version of TREQ-AL, a word alignment system that uses a lexicon extracted from a training corpus by means of a translation equivalence extractor. The new version has been improved significantly by including linguistic information. Especially, the use of language-specific rules appears to play an important role here. Information referring to cognates, precedence constraints and pair assignments (alignment of pairs of consecutive parts of speech) is also shown to improve the results, although to a lesser extent.

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

Summary of the Oral Session "Annotation of Multimodal Corpora"

Wolfgang Minker

In order to create reusable and sustainable multimodal resources, a transcription model for hand and arm gestures in conversation is required. In their presentation, Thorsten Trippel, Dafydd Gibbon and colleagues argued that state-of-the-art systems for sign language transcription and psychological analysis were not suitable for the linguistic analysis of conversational gesture. They developed CoGesT, a feature-based Conversational Gesture Transcription system for the linguistic analysis as well as automatic processing of arm gestures.

Harry Bunt and Laurent Romary discussed some basic methodological issues of the activities undertaken in the ACL-SIGSEM Working Group on the Representation of Multimodal Semantic Information. Rather than proposing particular formats, the working group aims at developing methodological principles for identifying and characterising representational concepts for multimodal content. A particular focus is placed on the interoperability and reuse of multimodal and language resources.

In the last presentation of this session, Ajay S Bhaskarabhatla and Sriganesh Madhvanath gave an insight into research carried out at Hewlett-Packard Labs, Bangalore, in online handwriting recognition of Indic scripts. The authors described the ongoing process of the data collection procedure, tools for collection and subsequent annotation, user-interface issues, the annotation scheme, and the organization of the dataset.

Wolfgang Minker
Department of Information Technology, University of Ulm
Albert-Einstein-Allee 43
89081 Ulm/Donau, Germany
Tel.: +49 731 5026254
Email: wolfgang.minker@e-technik.uni-ulm.de

Summary of the Oral Session "Corpus and Lexicon Tools",

Truus Kruyt

Consistent with other poster sessions on tools, the session "Corpus & Lexicon Tools" included a large number of presentations (20 in total). The tools were developed for a variety of purposes. They concerned a large number of languages, among which Turkish, Bulgarian, Polish, Portuguese, Spanish, Catalan, Basque, Japanese and Greek. Several tools, although implemented in a specific language, were designed to be language-independent. General tendencies were the application of XML, the adherence to general availability, and the relationship between language technology and the web.

Three tools had a rather generic purpose. In their poster, *A Public Reference Implementation of the RAP Anaphora Resolution Algorithm*, Long Qiu et al. presented the publicly available tool JavaRAP, a reference implementation to be used for the comparative evaluation of the many different anaphora resolution

approaches. The tool is a Java-based implementation of the seminal Resolution of Anaphora Procedure (RAP). *FreeLing: An Open-Source Suite of Language Analyzers*, presented by Xavier Carreras et al., is a suite of basic language analysis tools (tokenisers, morphological analysers, PoS taggers, etc.), based on a client-server architecture which enables the quick and easy integration of the tools into any NLP application. The software is distributed under Lesser General Public License (LGPL). Kiril Simov et al. also reported on a sequence of basic tools, but specifically for processing XML documents in the process of XML-based corpora creation and as a platform for rapid prototyping: *The CLaRK System: XML-based Corpora Development System for Rapid Prototyping*. Relatively many tools concerned workbenches for the development of language resources. Umut Özge and Bilge Say presented the *Development of a Corpus Workbench for the METU Turkish Corpus*, a workbench that is basically usable with any TEI- and XML- compliant corpus. *Abar-Hitz: An Annotation Tool for the Basque Dependency Treebank* was presented by Arantza Díaz de Ilarraza et al. "Abar-Hitz" is a graphical, language-independent tool, which accelerates the annotation process and avoids possible mistakes made by linguists. It was designed and built in close cooperation with linguists. *Creating multi-purpose linguistic resources for Modern Greek: a deep Modern Greek Grammar*, presented by Valia Kordoni and Julia Neu, concerned the development of a re-usable deep computational Modern Greek Grammar, with the practical support of "Grammar Matrix", an open-source tool designed for rapid development of multilingual, broad-coverage grammars couched in HPSG en MRS semantics. Catarina Ribeiro et al. showed in *Semi-automatic UNL Dictionary Generation using WordNet.PT* how they semi-automatically develop a PT-UNL dictionary, by porting information from the Portuguese WordNet database to the Portuguese UNL Dictionary. UNL is a meta-language developed for conveying linguistic expressions in order to encode website information into a standard representation. The dictionary is needed to integrate the Portuguese language into this platform. *Dynamic Lexicographic Data Modelling. A Diachronic Dictionary Development Report*, presented by Paul Gévaudan and Dirk Wiebel, focussed on a lexicographical model of diachronic filiation, covering many languages. The model has the capacity to analyse highly complex cases of lexical evolution. For the task of diachronic dictionary compilation, the model is represented as an entity-relationship model and integrated into a powerful DBMS workbench. Two tools offered alternatives for common methods of corpus building. Zygmunt Vetulani addressed the problem of the absence of an easy and inexpensive way to collect naturally generated dialogue recordings. He presented *An Environment for Dialogue Corpora Collection (ENDIACC)*, an easily accessible, language-independent software platform, to provide an experimental setting for text-mode written (keyboard) dialogue corpora collection. The tool will be freely accessible for research purposes. *Using Paradigm Tables to Generate New Utterances Similar to those Existing in Linguistic Resources*, presented by Yves Lepage and Guilhem Peralta, described a method of automatic sentence generation on the basis of an existing corpus, to enlarge that corpus and to make it more domain-specific than is feasible with common corpus building. Several tools supported the development of training data, two of them concerned with handwriting recognition. *An XML Representation for Annotated Handwriting Datasets for Online Handwriting Recognition*, presented by Ajay S Bhaskarabhatla and Sriganesh Madhvanath, provided an XML representation for annotation of online handwriting data to support the development and evaluation of handwriting recognition algorithms. The representation uses Digital Ink Markup Language (InkML), a draft standard from W3C. *The SPARTACUS-Database: a Spanish Sentence Database for Offline Handwriting Recognition*, presented by Salvador España et al., is a freely available database that consists of offline handwritten Spanish sentences from four different subtasks and that is expected to be especially useful for recognition systems that may benefit from language models of restricted semantic tasks. The files are in XML. Vincent Vandeghinste and Erik Tjong Kim Sang presented *Using a Parallel Transcript/Subtitle Corpus for Sentence Compression*, a training corpus for the automatic conversion of transcripts of Dutch television programs into compressed subtitles targeted at hearing-impaired people. In *Annotation of Anaphoric Expressions in an Aligned Bilingual Corpus*, Agnès Tutin et al. reported on the development of a 25,000 words French-English corpus annotated and aligned at anaphoric level. The annotation scheme is encoded in XML; the alignment follows the EAGLES CES recommendation. The paper contains little information on tools.

Some tools concerned search engines. In *Linguistic Corpus Search*, Christian Biemann et al. described a prototype of a modular and (almost) language-independent linguistic search engine for exploring plain as well as PoS-tagged monolingual corpora in an easy and intuitive way. A 'minimalist' query language nevertheless allows powerful searches without the cognitive load of a complex formal search language. In *Concept-based queries: Combining and Reusing Linguistic Corpus Formats and Query Languages*, Felix Sasaki et al., arguing that current query languages are strongly connected to corpus formats, proposed a methodology for querying heterogeneous linguistic data represented in different corpus formats. The methodology includes an abstract, conceptual level of "Linguistic Concept Descriptions" (in RDF format) on top of existing formats and query languages. Carlos Amaral et al. presented *Design and Implementation of a Semantic Search Engine for Portuguese*. The task of this search engine is to find a sentence in a set of texts (on local hard disk or on the web) that answers questions in natural language. The result, presented as a list of the best sentences in descendent order of their scores, is crucially influenced by the quality of the language resources used by the system.

Three tools could not be categorised. In *Applying a Part-of-Speech Tagger to Postal Address Detection on the Web*, Nuno Cavalheiro Marques and Sérgio Gonçalves reported on the adaptation of a neural-network PoS tagger to a real-world information retrieval system that is capable of extracting postal addresses from internet web pages. For this system, a particular tag set was developed. Luciana Bordoni et al. presented *CHeM: A System for the Automatic Analysis of e-mails in the*

Restoration and Conservation Domain. A "Cultural Heritage e-mail Manager" automatically analyses e-mails of the Restoration and Conservation newsgroup and clusters them into content classes; the subject field of the e-mail does not suffice for this complex domain. The system automatically generates a mailing list of all the users interested in a particular content cluster. To conclude with a language-specific problem: *Bypassing Greeklish!*, presented by A. Chalamandaris et al. Greeklish is a set of transliteration patterns of Greek using the Latin alphabet. It is widely used, because e-mail and other computer devices do not support the Greek alphabet. Greeklish is extremely inconsistent, and reading it is over 40% more time-consuming reading plain Greek. The system transliterates Greeklish into Greek and also detects non-Greek, with high succes rates.

Truus Kruyt
Institute for Dutch Lexicology INL, PO Box 9515
NL-2300 RA Leiden,
The Netherlands
Tel. +31 71 5272270
Fax +31 71 5272115
Email: kruyt@inl.nl

Summary of the Oral Session "Morphosyntactic Corpora and Tools"

Zygmunt Vetulani

It is always a pleasure for me to chair a session at LREC because of the high level of contributions, the (time) discipline of presenters and the reactive public ready for questions and discussion. This time my task was to chair the LREC session 033, focusing on morphosyntactic corpora and tools. Though this four paper session is only one out of 47 oral and 27 poster sessions, morphosyntax related issues are far from being marginal at the conference: one poster session of 15 presentations addressed morphosyntactic data and tools (P14) and the term morphosyntactic appeared as keyword in a number of other contributions.

The four papers presented during this session were substantially different among themselves. It seems that the intention of Organisers when gathering them into one session was to emphasise the productivity of the domain characterised by the intersection of these three keywords: morphosyntactic, corpora and tool, and their importance for various areas of Language Technologies.

The papers presented are: *The verb in the Terminological Collocations, Contribution to the Development of a Morphological Analyser: MorphoComp*, by Rute Costa and Raquel Silva from Portugal, *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*, by Toma Erjavec from Slovenia, *The Statistical Analysis of Morphosyntactic Distributions*, by Stefan Evert from Germany, and *Utilization of Multiple Language Resources for Robust Grammar-Based Tense and Aspect Classification*, by Alexis Palmer, Jonas Kuhn and Carlota Smith from the USA.

The first of these papers, by Rute Costa and Raquel Silva, aims to contribute to the MorphoComp project, whose purpose is to develop computational morphology tools, in particular a Morphological Analyser for extracting terminological collocations from specialised corpora. The focus in the paper is on morphological tools. The second contribution, by Toma Erjavec, is more focused on morphosyntactic corpora. The recent developments of multilingual MULTEXT-East resources are presented, in particular those developed for Central- and East-European languages under the Copernicus project CONCEDE. The LREC audience is already familiar with this project (cf. LREC 2000) but it is interesting to follow this initiative in the domain of multilingual resources (morphosyntactically annotated corpus "1984" based on translations of the famous novel by Orvell). Its language coverage is already important, and it would be wonderful to see all European languages included in the project (this was one of the questions from the audience). The next paper, presented by Stefan Evert, was about a method of statistical analysis of quantitative data on the distribution of morphosyntactic features in corpora. This issue is important for highly inflected languages (as e.g. German and all Slavonic languages) where morphological analysis, essential for parsing, is hard because of syncretism. The author proposes to use a fine statistical method to help solving morphological ambiguities in corpus data. The last of the four contributions, by Alexis Palmer, Jonas Kuhn and Carlota Smith, is a contribution in discourse semantics. "Situation entity" class labels are to be assigned to predicators in written English texts. This objective is attained using multiple language resources and tools, including a parsing system for predicate-argument analysis that involves the association of morphosyntactic features. An interesting, and non-trivial, empirical observation about situation entity classification task was that inclusion of lexical information improved recall and decreased precision.

Dr Zygmunt Vetulani
Head of Department of Computer Linguistics and Artificial Intelligence
Adam Mickiewicz University of Poznan
Poland
Email: vetulani@amu.edu.pl
Web site: www.amu.edu.pl/~vetulani

Summary of the Oral Session "Question-Answering"

Carol Peters

This session gave a good panorama of many of the issues currently being investigated in the Question-Answering (QA) area, plus some that are rather on the outskirts of the main interests of this sector. The five papers presented covered a wide variety of topics ranging from both written and spoken question generation and inferential rule creation from language resources right through to QA system building and architecture. Unfortunately, there was no-one available to present the paper on *Evaluation for an End-to-End Interactive Question Answering System*, and evaluation was thus one important aspect largely missing from the session although touched on to some extent in the presentation by the University of Geneva.

The first two papers described experiments aimed at exploiting existing language

resources to build useful components for QA system development. The presentation by Karin Müller, from Amsterdam University, reported a method for the semi-automatic construction of a question treebank. Linguistic knowledge encoded in the Penn Treebank is being used in the generation of a large treebank of questions. The aim is to create resources that can be employed to develop improved question-processing modules. Francesca Bertagna, from ILC-CNR, Pisa, described experiments aimed at verifying whether the semantic information encoded in two Italian language resources, ItalWordNet and CLIPS (a computational lexicon for Italian tagged at phonological, morphological, syntactic and semantic levels), could be combined to derive primitive inferential rules that could then be used in QA systems. Unfortunately, the results were rather disappointing. Bertagna states that only rarely did these resources provide the relations needed to support complex inferences. She attributes this to both quantitative and qualitative problems with the LRs under examination but claims that experiments of this type contribute to the ongoing discussion on the ways of conceiving and representing word meaning. The third presentation by a group from Carnegie Mellon University described *JAVELIN*, an open-domain QA system, focusing on the design and implementation of the core module of the system - the information repository. The JAVELIN repository implements a consistent relational model for all the information associated with a QA scenario. It addresses two crucial requirements for advanced, scalable QA systems: module traceability and answer validation; consistency and reuse of information. The paper from the University of Geneva, presented by Agnes Lisowska, was an outlier with respect to the main focus of the session. Lisowska described an experiment in user query elicitation aimed at deriving input for the design of a multimodal meeting processing and retrieval system. The elicited queries are also used as a benchmark against which to evaluate the system implemented. The final presentation in this session, by Nelleke Oostdijk, from University of Nijmegen, described an attempt to derive a taxonomy of wh-questions from a spoken Dutch corpus and to produce a model of the way in which questions are integrated into spoken discourse. The aim was to develop an NLP system that can support natural interaction with a spoken QA system. The results obtained were compared with a collection of Internet FAQs. One of the findings was that restricting studies to wh-questions was too severe a limitation and further research is needed into alternative ways of asking for information.

QA is very much a multidisciplinary area, traditionally involving a combination of tools and methodologies from both the Information Retrieval and the Natural Language Processing domains. However, with one exception, the papers in this session were very much oriented towards the NLP side of the topic. Overall, the session can be regarded as a success. It was nice to see that not all experiments regarded English; work on both Italian and Dutch was also reported. There was a good-sized audience, averaging around sixty people, and the speakers were subjected to a reasonable number of questions from the floor, although no general discussion or debate emerged.

Carol Peters
ISTI-CNR
Area di Ricerca CNR
Via Moruzzi, 1
56124 Pisa, Italy
Tel: +39 050 3152897
Fax: +39 050 3152810
Email: carol.peters@isti.cnr.it

Summary of the Oral Session "Evaluation of Speech Annotation and Systems"

Jean-Claude MARTIN

Four papers were presented on the topic of evaluation of speech annotation and systems.

Danieli et al. presented an evaluation of consensus on the annotation of prosodic breaks in the romance corpus of spontaneous speech "C-ORAL-ROM". Their results showed that the annotation of the utterances identified in terms of their prosodic breaks is able to capture relevant perceptual facts.

Duchateau et al. gave a talk on the use and evaluation of prosodic annotations in the CGN database (Spoken Dutch Corpus). Their conclusion is that annotations for the remainder of the CGN database can be generated automatically with the same quality as the manual annotations.

Trutnev et al. compared evaluations in the domain of Automatic Speech Recognition. The main obtained results are that (1) the Hidden Markov Model HMM-based technology performs better than the hybrid approach in the case of unconstrained continuous speech, and (2) the academic systems perform better in the case of continuous speech in French, while the commercial systems show better recognition accuracy for continuous speech in German.

Finally, Veiga et al. described a method to perform word confidence measures in an automatic speech recognition system. The confidence measure is computed during the decoding phase and is based on likelihood ratios between the top hypotheses that reach a word node. Experiments were carried out on a digit database with a connected-digit recognizer. The results showed that this method outperforms word-graph confidence measure with a special grammar and is worse with a word loop grammar. The audience of this session held in the late afternoon was composed of fifteen people who asked a few questions for each presentation.

Jean-Claude Martin
Assistant Professor in Computer Science
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
Email: Jean-Claude.Martin@limsi.fr
Website:www.limsi.fr/Individu/martin/

Summary of the Poster Session "Evaluation of Language Technologies"

Andrei Popescu-Belis

Poster session P25-EW was one of the many sessions at LREC 2004 dedicated to the evaluation of language technologies, in particular for written language processing - as opposed for instance to spoken language tools, or to dialogue systems. The session had good thematic homogeneity, since three main research areas were represented: (1) the evaluation of machine translation, (2) the evaluation of parsers and grammatical resources, and (3) reports of evaluation campaigns.

The first theme had been discussed the previous day (Thursday, May 27th) in an

interesting and lively session on the "Evaluation of Machine Translation and Multilingual Systems", chaired by Maghi King. During the present session, several posters examined the reliability of metrics for the evaluation of MT, often trying to improve the BLEU metric (Papineni et al. 2002). For instance, Andrew Finch (et al.) showed that the correlation between human judgments of quality and automated MT evaluation metrics is stable when four or more reference translations are used - so using only four is enough. Stephan Vogel (et al.) studied the BLEU scores based on confidence intervals obtained from various samples of a test corpus, and showed which relative rankings of the DARPA/NIST 2003 campaign were the most certain. Bogdan Babych (et al.) introduced an alternative approach to ranking, based on usability, and compared it to two automated evaluation metrics, BLEU and LTV. Two direct applications of MT evaluation were also presented, one to the evaluation of human translation capability (Yasuhiro Akiba et al.), and the other to the comparison of a statistical and a rule-based MT system on a novel domain with a limited amount of resources (Per Wejnitz et al.): here, a concordant variation of all the scores showed that the second system performed better than the first one.

The second theme, evaluation of parsers and grammars, was somewhat closer to studies presented also in other sessions, such as "Evaluation of LR and Tools", and "Evaluation of Systems and Tools" - note however that the latter featured an application of a recent MT evaluation metric to the evaluation of answers in an e-learning environment. Jennifer Foster presented a method to evaluate the performance of parsers on ungrammatical sentences, of which a sample was collected from various sources including academic papers. Timothy Baldwin (et al.) analyzed the coverage of a grammar used for "deep processing" of English, applied to previously unseen data, and offered some suggestions to extent its lexical coverage. Gabriel Infante-Lopez (et al.) described an approach to comparing probabilistic context-free grammars based on their capacity to reduce parsing ambiguity for each sentence.
Two poster presentations summarized ongoing evaluation campaigns. Patrick Paroubek (et al.) described the evaluation protocol and the main challenges of the EASY campaign for syntactic parsers of French - a topic thus related to the second theme above. The EASY campaign is one the components of the French EVALDA multi-evaluation initiative sponsored by the Technolangue program, other components of which have been also presented at this conference. The poster summarizing NIST's recent evaluation campaigns (Alvin F. Martin et al.) was an excellent synthesis of these actions, with an attempt to outline a common, generic approach to component evaluation, in various clearly identified stages. The NIST poster identified the following phases, in a presentation that was visually clearer than the paper published in the proceedings: (a) task definition; (b) metrics, scoring software, and data; (c) rules and schedule; (d) description of participating systems; (e) evaluation, and post-evaluation workshop.
On the whole, this poster session witnessed a lot of interest from the conference participants, often accompanied by lively debate. The feedback received by the authors was probably more significant than in oral presentations, an argument in favour of posters - provided enough time and space are allowed for discussion.

Andrei Popescu-Belis
ISSCO/TIM/ETI, Université de Genève
40, bd. du Pont d'Arve
1211 Geneva, Switzerland
Tel.: +41 22 379 8681
Fax: +41 22 379 8689
Email: andrei.popescu-belis@issco.unige.ch

Summary of the Poster Session "Machine Translation"

Anna Sågvall Hein

The poster session on machine translation includes six highly relevant presentations. Three of them focus on the enhanced use of translation memories in various settings, one on the improvement of statistical translation quality by adapted language modelling, one on capturing structured feed-back from post-editors for improving a transfer grammar, and one, finally, is devoted to the relation between text difficulty and MT performance.
Kranias & Samiotou present a method for enhanced use of a translation memory in the translation process. The basic idea is to post-edit fuzzy matches automatically, making use of a dictionary of words and phrases generated by means of word alignment. As a result of the automatic post-editing, fuzzy match scores increase and low-score matches can be utilised. Data on cost reductions thus achieved are presented in the paper. The method is commercially implemented.
The enhancement of the use of translation memories is also the main goal of the study presented by Nevado, Casacuberta, and Landa. The core issue is the automatic generation of sub-sentence bisegments, typically multi-word-units, and their integration in a translation memory. Two statistical alignment strategies are investigated and applied to Basque and Spanish. For the evaluation, an intuitively generated reference alignment was used. A basic problem is the low precision of the automatic alignment, and proposals for the modification of the alignment strategies are made. It seems, that additional inspiration for handling some of the problems that are encountered in the study, among them the evaluation of partial linking, may be found in previous works on word alignment not cited in the paper.
In the paper by Gröbler, Hodász, and Kis, linguistic annotation is proposed for enhancing the usefulness of the translation memory. In particular, the automatic assignment of linguistic structure is addressed. Structure is assigned at three levels by means of POS tagging, NP chunking, and the identification of sentence skeletons. Sentence skeletons are patterns of NP slots, tags of words, and punctuation marks. Search in the memory for the linguistically annotated bisegments is handled by the translation memory system. It composes translations from constituents found in the memory. Human post-editing is part of the translation process. The result of the post-editing is automatically analysed, source as well as target segment, and fed into the memory. No formal evaluation has, so far, been carried out.
Eck, Vogel, and Waibel demonstrate that the translation quality of a statistical machine translation system may be improved if the language model of the target language is adapted to the domain of the sour-

ce text. Domain texts are extracted from large target language corpora and identified by means of information retrieval techniques, where the original translation, based on a general language model, serves as the search key. Domains in terms of documents and sentences are investigated, and sentence domains are found to give the best results. Furthermore, for the strategy to be successful the quality of the original translation has to meet certain demands, one of several issues brought up in the paper that will be further investigated.
Font Llitjós, and Carbonell present a tool for capturing structured feed-back from non-expert post-editors. The feed-back consists in a corrected version of the translation with a log of the corrections, and the specification of error categories. The error classification used for the purpose has nine categories. The data thus provided are to be used for the automatic improvement of a transfer-grammar. A user study has been carried out showing that the users are good at detecting errors but less good at determining error types. In order to deal with this problem, the MT error classification will be further developed.
Clifford, Granoien, Jones, Shen, Weinstein bring up the relation between machine translation quality and language difficulty. Initial experiments on several languages indicate some relations between MT performance and difficulty levels as they defined by the Interagency Language Roundtable standard. Primarily, this is found for MT output whose quality is good enough to be readable by human readers.

Anna Sågvall Hein
Professor in Computational Linguistic
Dean of the Faculty of Languages
Department of Linguistics and Philology
Uppsala University
Box 635
751 26 Uppsala, Sweden
Tel.+46 (0)18-471 1412
Fax +46 (0)18-471 1094
Email: anna.sagvall_hein@lingfil.uu.se

Summary of the Poster Session "Computational Lexicons"

Farah Benamara

Globally three major points were outlined during the session: (1) the description of dedicated lexical units from existing resources such as WordNet or EuroWordNet, (2) the methodologies for building computational lexicons based on various paradigms such as the corpus based approach, or Frame Net and (3) the applications that make use of these lexical descriptions.
The projects presented during that poster session covered a large number of very diverse languages from different families, among which: Japanese, Slovene, Serbo-Croatian, Portuguese, Danish, Spanish, Korean, Chinese, French and English.
Besides the classical, but of much interest, WordNet or EuroWordNet extensions, uses of FrameNet, and works around Core lexicons and the management of lexical consistency, a number of projects were devoted to less frequently encountered topics such as the lexical description of adverbs and adjectives. Let us also note some interesting projects around idioms and the introduction of implicit information into WordNets. Of interest is also a new trend in the development of conceptual relatedness and consistency measures between lexical entries in a hierarchy.
The session gathered people from at least 20 different countries, exchanging ideas and experiences on the coding of the properties of their own languages, and the difficulties encountered, technical as well as institutional. Besides authors, who were in general quite numerous for each poster, a large number of LREC participants came, got information, references and links. Due to time restrictions and to the large number of presentations, participants felt they had valuable but too short exchanges.

Farah Benamara
Institut de Recherches en Informatique de Toulouse, IRIT,
118 route de Narbonne,
31062, Toulouse, France
Email: benamara@irit.fr

Illustration of posters layout at Belem conference centre

Session room at Belem conference centre

LREC 2004 Workshops' Reviews

Workshop on "Multimodal Corpora"

Organisers: Jean-Claude Martin, Elisabeth Den Os, Peter Kühnlein, Lou Boves, Patrizia Paggio, Roberta Catizone

The full title of this one day workshop was "Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces". Around 40 people attended this workshop held on Tuesday 25th May 2004. It was the only workshop related to multimodality among the 18 LREC 2004 satellite workshops (following the 1st and 2nd LREC workshops on multimodal corpora in 2000 and 2002).
The primary purpose of this one day workshop was to share information and engage in the collective planning for the future

creation of usable multidisciplinary multimodal resources. Existing annotation of multimodal corpora until now has been done mostly on an individual basis, each researcher or team focusing on their own needs and knowledge about modality specific coding schemes or application examples. Thus, there is a lack of real common knowledge and understanding of how to proceed from annotations to usable models of human multimodal behaviour and how to use such knowledge for the design and evaluation of multimodal input and embodied conversational agent interfaces. Furthermore, the evaluation of multimodal interaction poses different (and very complex) problems than the evaluation of monomodal speech interfaces or WYSIWYG direct interaction interfaces. There are a number of recently finished and ongoing projects in the field of multimodal interaction in which attempts have been made to evaluate the quality of the interfaces in all meanings that can be attached to the term 'quality'. There is a widely felt need in the field for exchanging information on multimodal interaction evaluation with researchers in other projects. One of the major outcomes of this workshop should be better understanding of the extent to which evaluation procedures developed in one project generalise to other, somewhat related projects.

Out of 15 submitted papers, 10 papers were accepted for long presentation. They enabled the workshop to cover several dimensions of multimodal corpora:

- Multimodal phenomena: verbal and gestural feedback, visual correlates of emotional speech, facial animation, human movement notation.
- Multimodal corpora collection and analysis: guidelines, annotation schemes.
- Multimodal system design and evaluation: wizard of oz prototyping, animated agent systems and multimodal spoken dialogue systems, evaluation metrics.
- Application areas: edutainment systems (computer games, children), multi-participant meetings.

The presentations were grouped in 3 sessions:

- Recommendations for Multimodal Annotation Tools and Schemes,
- Multimodal Systems Design and Evaluation,
- Coding Schemes and Multimodal Communication.

The workshop was very successful in the sense that it really brought people from different disciplines together. For example, lively discussions took place on coding of facial and body expressions.

There was an invited talk on Corpora for Sign Language Studies and a panel discussion closing the workshop. Discussions continued in the evening at the oldest cervejaria in Lisbon, a Portuguese restaurant with walls covered with old tiles.

Workshop URL: http://lubitsch.lili.uni-bielefeld.de/MMCORPORA/

Jean-Claude Martin
Assistant Professor in Computer Science
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
Email: Jean-Claude.Martin@limsi.fr
Website:www.limsi.fr/Individu/martin/

Workshop on "Compiling & Processing Spoken Language Corpora"

Organisers: Nelleke Oostdijk, Gjert Kristoffersen, Geoffrey Sampson

Over the past few years there have been many initiatives directed at the development of spoken language corpora. At present, corpora are being compiled for many different languages from all over Europe, including various smaller languages and minority languages. Some projects are about to start (e.g. Norwegian), while others have just been completed (French, Spanish, Italian, Portuguese, Dutch). It is against this background that the workshop "Compiling and Processing Spoken Language Corpora" was organized.

The aim of the workshop was to bring together people working on the development (compilation and processing) of spoken language corpora. The workshop gave participants the opportunity to exchange views and share experiences. Moreover, the workshop was instrumental in taking stock of and evaluating the present state-of-the-art.

The workshop attracted some 45 participants. The programme offered a selection of papers that were accommodated in three sessions: (1) Corpus compilation and (orthographic) transcription, (2) Corpus annotation, and (3) Extending corpus parameters.

The first session opened with a paper by Shlomo Izre'el and Giora Rahav who reported on the progress made with respect to the compilation of a corpus of spoken Israeli Hebrew, a project that is still in a very early stage but in which various issues relating to the design of a spoken corpus have been addressed extensively. The two other papers that were presented in this session (one by Ana González Ledesma and others, the other by Sarah Creer and Paul Thompson) were concerned with the orthographic transcription and mark up of spoken language corpora, more specifically the C-ORAL-ROM corpus and the BASE corpus, and the problems encountered there.

Although each of the papers presented in the second session dealt with corpus annotation, they varied as regards the type of annotation they addressed. The paper by José Guirrao and Antonio Moreno Sandoval described a toolbox for tagging the Spanish C-ORAL-ROM Corpus. Claudio Bendazolli, Cristina Monti and others introduced their project that is aimed towards the creation of an electronic parallel corpus for the study of simultaneous interpretation from and into different languages. The last paper in this session was by Tiit Hennoste and others, and reported the experiences obtained in the development of a dialogue act coding scheme and its application to the Estonian Dialogue Corpus.

In the last session, Philippe Martin presented WinPitch Corpus, a text-to-speech alignment and analysis tool for use with large multimodal corpora (including both audio and video). Next, Fabio Tamburini and Carlo Caini described the method they have developed for the automatic detection of prosodic prominence in continuous speech. Finally, Daan Broeder and others introduced and elaborated upon the idea of a 'dynamic corpus environment' which should make it possible to maintain corpora while allowing the addition of further data and/or new types of information.

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

Belem conference centre (front)

Belem conference centre (back)

Workshop on "LRs Integration and Development in e-Learning and in Teaching Computational Linguistics"

Organisers: Paola Monachesi, Cristina Vertan, Walter v. Hahn, Susanne Jekat

The workshop on "Language Resources: Integration and Development in e-Learning and in Teaching Computational Linguistics", held on 24th May 2004, focused on the integration of LRs in the educational process and the cooperation among LRs and e-learning. Additionally, it discussed the use of LRs in the curriculum of computational linguistics. It was organised by Paola monachesi (University of Utrecht), Cristina Vertan and Walther v. Hahn (University of Hamburg) and Susanne Jekat (Zurich University of Applied Sciences Winterthur). The organisers come from different areas of research with interest in language resources (linguistics, natural language processing, translation).

The 8 presented papers covered the following topics:

1. Case studies on the use of LRs in linguistics and computational linguistics,
2. Additional skills acquired by the students when using or developing LRs (e.g. how to acquire standards),
3. Usage of LRs in the development of e-learning materials,
4. Adaptation of existing LRs for CALL environments,
5. Development of e-Content localization resources.

The workshop was organised in two sessions, preceeded by an invited lecture of Hans Uszkoreit about Ontology-based knowledge management and transfer in computational linguistics. The topic of the morning session was "Language Resources in Teaching Computational Linguistics" and contained three contributions.

Veit Reur and Petra Ludewig described the use of LRs in two group projects for students at master level. In one project, LRs are used for collocation extraction; in the other, for the construction of a vocabulary trainer. The presentation of Claudia Kunze and Lothar Lemnitzer focused on the use of existing lexical resources, in particular GermaNet, for case studies and explorative learning in virtual courses of Computational Linguistics and Language Engineering. Dan Cristea, Horia-Nicolai Teodorescu and Dan-Ioan Tufis reported on LRs used for student projects both in language and speech technology.

The afternoon session consisted of 4 talks addressing the relationship between LRs and e-learning.

Dragos Ciobanu, Karl-Heinz Freigang, Anthony Hartley, Uwe Reinke and Martin Thomas presented a rationale for the development of a multilingual resource designed to support the training of translators in their use of translation memories. The following two talks focused on a special aspect of e-learning: computer-aided language learning. In both presentations, the accent was on vocabulary learning. Galia Angelova, Albena Struchanska, Ognian Kalaydjev, Svetla Boytcheva and Irena Vitanova described LRs used in a CALL-project for learning English financial terminology. The paper of Sandro Pedrazzini, Alexandro Trivilini and Judith Knapp showed how an existing LR can be adapted for e-learning purposes, i.e., language learning. The creation of an environment for dynamic teaching materials for ESSLI (European summer School on Logic, Language and Computation) was discussed by Rafaella Bernardi, I.Dahn, G. Mishne, M. Moortgat, M. de Rijke and H. Uzkoreit.

The afternoon session was followed by a summing-up session where the organisers stressed the essential topics revealed by the talks. At the end, a one hour discussion concentrated on take-up actions, and future collaboration plans.

The workshop was attended by a quite big number of participants, all taking actively part to discussions, after each talk as well as in the panel. Several take-up actions (set-up of some working group, mailing list) will be brought to life in the coming weeks. The workshop showed once more that a deeper cooperation between specialists working in different areas (in particular, in education), with language resources, is highly desirable.

Dr. Cristina Vertan
Natural Language Systems Division
Computer Science Department
University of Hamburg
Vogt-Koelln-Str. 30
22527 Hamburg, Germany
Tel.: +40 428 83 2519
Fax: +40 428 83 2515
Email: cri@nats.informatik.uni-hamburg.de
Web site: http://nats-www.informatik.uni-hamburg.de/~~cri

Workshop on "XML-based Richly-annotated Corpora",

Organisers: Andreas Witt, Ulrich Heid, Jean Carletta, Henry S. Thompson, Peter Wittenburg

The Workshop on "XML-based Richly Annotated Corpora", on Saturday, May 29th 2004, full-day, and with 30-40 participants, was structured into 3 major sections, ranging from theory of XML corpus representation over applications to software. The part on applications was divided into a block with more linguistically-oriented ones and a block with more tool-related ones.

The workshop covered all aspects of the use of XML in the annotation of corpora, from concurrent analyses (Cristea/Butnariu) and the handling of discontinuous multiword items in a stand-off model (Pianta/Bentivogli) over questions related with text classification (Langer et al.) and the structure, annotation and modelling of a diachronic corpus (Dipper et al.) to the creation, use and maintenance of XML based language archives, with an opening towards international and global infrastructures for XML-based corpora (Wittenburg et al.).

This showed impressively that corpus linguistics has entered its XML era, and that almost all questions of corpus design, corpus annotation and corpus manipulation are now being discussed in the framework of XML-based richly annotated corpora.

Similarly, the software section also covered most aspects of practical work with XML-based corpora: Freese presented possibilities for integrating an existing format and tool box with the linguistic annotation framework, LAF, which is currently being proposed by ISO TC 37 SC 4; other presentations focused more on user interfaces for the creation of richly annotated corpora (Artola et al.) as well as on tools for the transcription and annotation of spoken language (Schmidt) and the annotation of richly annotated written language corpora (comparison of existing tools, by Dipper/Goetze/Stede).

All presentations were discussed in quite some detail, and it became clear that, for complex and richly structured corpora, representation models based on ordered directed acyclic graphs, possibly within the stand-off model, are a promising modelling device (Dipper et al., Pianta/Bentivogli). Other approaches (Cristea/Butnariu, Schmidt) deviate from this with good reasons, for example because of needs of applications, such as the annotation of spoken language and the conversion of heritage data.

In the presentations, most aspects of the manipulation of richly annotated corpora were dealt with, with the exception, perhaps, of tools for search and retrieval, which were only mentioned punctually. For very large corpora, a mapping towards a performant database (and, for example, query via SQL, Dipper et al.) were proposed, or, alternatively, there are custom-made tools for browsing and interrogation of the corpora (Artola et al., Wittenburg et al.).

The workshop clearly showed the potential of XML-based corpus technology.

Ulrich Heid
Universitaet Stuttgart
IMS-CL, Institut fuer maschinelle Sprachverarbeitung -- Computerlinguistik
Azenbergstrasse 12
D - 70 174 Stuttgar t, Germany
Tel.: + 49 711 121 1373
Fax: + 49 711 121 1366
Email: uli@ims.uni-stuttgart.de

Workshop on "Representation and Processing of Sign Languages"

Organisers: Oliver Streiter & Chiara Vettori

This year, for the first time, there has been at LREC a workshop dedicated to sign languages. For those who stumbled into the workshop, the great variety of topics and approaches might have been surprising. Since this field is considerably younger than the processing of spoken and written languages, a vast number of fundamental questions still have to be settled.

Trivially speaking, spoken languages are spoken and heard. Sign languages are signed and seen. Spoken languages have been written as ideograms, in syllabic and phonemic transcriptions. But as for sign languages? How can they be written for love letters, poems, verdicts and recipes?

One possible answer is *SignWriting*. *SignWriting* does not decompose a sign into phonemes, syllables or morphemes but body-parts, movements and face expressions. Each of them is assigned a representation. Given such an alphabet for potentially all sign languages - how may a keyboard, the input system, look like? How are the simple elements (body-parts, movements and face expressions) to be encoded and how the composed signs? As pictures, in Unicode or XML? How will this influence the input of signs, the layout and formatting, the possibilities to perform exact and fuzzy matches? A couple of presentations have been dedicated to these problems.

SignWriting, however, is not the only possible way of writing signs. Thomas Hanke in his invited talk introduced *HamNoSys*, the Hamburg Notation System for Sign Languages. The purpose of *HamNoSys* has never been the everyday communication. Instead, it complies with research requirements for corpus annotation, sign generation and dictionary construction. It thus differs from *SignWriting* in its scope and granularity.

Once fundamental questions regarding the writing of signs will be settled, derived notions such as word n-grams and character n-grams, may be used for applications such as language recognition, document classification and information retrieval. Spelling checking, syntax checking and parsing obviously will be further developments. In the current workshop, these topics still did not play a role.

Whether *SignWriting* should be used for writing recipes and poetry or the national spoken language, is still emotionally discussed. In addition, most deaf signers have not been trained in reading or writing *SignWriting*. What is known as "text-to-speech" in the processing of spoken languages would come as possible solution: a front-end to web-pages, mail boxes, etc., would sign out the written text. As shown in various presentations, avatars, i.e. virtual signers, may be constructed which translate a written form of signs into signs, just like translating "d" into the corresponding sound wave.

A front-end on the input side of the system might translate signs into a written representation. Speech Recognition becomes Sign Recognition. Two different tech-

niques have been proposed. The recognition with the help of a data glove precedes from the signer's perspective and his/her articulations. The recognition of signs with the help of cameras, the second alternative, leads to the description of signs from the observer's point of view.
A number of presentations have been concerned with the design and creation of electronic sign language dictionaries. If there is a common line in all these proposals, it might be the attempt to give the sign language an as large as possible autonomy with respect to the spoken national language. Still the list of topics does not end here...
The workshop was held in an atmosphere of collaboration and mutual respect, although no good solution could be found to assure the interpretation for deaf workshop participants. Thanks to LREC for hosting this workshop and surely, there will be a second one, hopefully 2006 in Genoa.

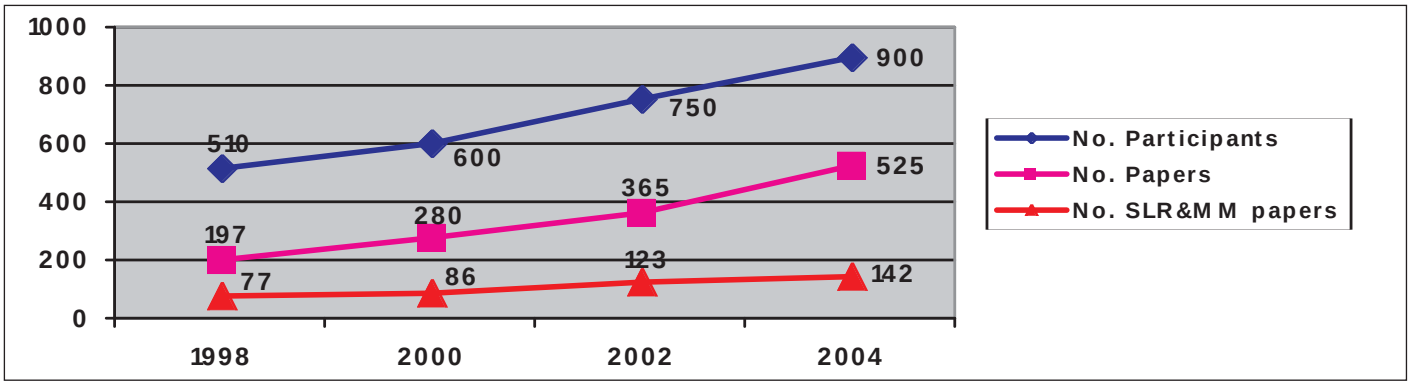
Oliver Streiter
Language and Law
EURAC research
Viale Druso/Drususallee 1
39100 Bolzano/Bozen, Italy
Tel.: +39 0471 055115
Fax: +39 0471 055099
Email: ostreiter at eurac dot edu

LREC 2004 Reports

Report on Spoken Language Resources and Multimodality

Daniel Tapias

First of all, I want to say that LREC 2004 has been a very special conference to me. On the one hand, because it was dedicated to the memory of two dear friends and great scientists (Antonio Zampolli and Ángel Martín Municio). On the other hand, because once more, the number of papers and participants has increased with respect to the previous LREC, which shows the growing interest in the area of Human Language Technologies and the consolidation of LREC as the International Conference on Language Resources and Evaluation.
The figure illustrates this by showing the evolution of the number of participants, the total number of papers and the papers on spoken language resources (SLR) and multimodality (MM), that represent about 30% of the total.



If we go beyond the figures and enter into more detail, we see that the area of Spoken Language Resources can be classified into four main topics:
1. *Speech Corpora*: in this section, I would mention the important effort made in the creation of broadcast news, emotion and expression, telephone speech and phonetically and prosodically oriented databases. For example, in the case of broadcast news (BN), it is worth mentioning projects and initiatives like Net-DC for Arabic, the COST-278 European project in which 7 European languages were recorded, the parallel corpora for Spanish and Basque from EITB, the ESTER campaign for ASR evaluation in BN in French or the Transcrigal-DB for Galicien.
Concerning the emotion and expression speech databases, it is important to talk about the work that has been carried out to produce spontaneous or semi-spontaneous emotional speech databases together with the more traditional approach consisting on recording acted emotional speech. Some of the presented papers showed the continuation of the work presented at LREC 2002, like the one corresponding to the JST Expressive Speech Processing Corpus produced for Japanese. It was also interesting the work done in this area for children's speech in German and English and the corpora produced to work in emotional Text-to-Speech conversion in Greek and Basque.
As for the telephone speech databases, it is interesting to say that many of the presented databases follow the SpeechDat design and methodology (SALA-II cellular speech in America, LILA in the Asian-Pacific area and ORIENTEL in the mediterranean area). In addition to this, important initiatives like the Fisher Corpus of human-human conversations (DARPA EARS Program), the South African languages database, the children's speech database in French, the Cypriot speech database and the Speaker Verification Database used in the NIST evaluations were presented.
In the area of phonetically and prosodically oriented databases, the Spoken Africaans Language Resource (SALAR) produced to research in pronunciation variants, the phonetically balanced Mexican Spanish VOXMEX database or the C-ORAL-ROM multilingual spontaneous speech database are good examples of the work done in this topic.
Finally, there were papers presenting many other types of corpora like dialogue, car environment and translation speech databases. In the machine translation field,

there were interesting examples like the well known NESPOLE! Corpus, a parallel corpora for Spanish and Basque and the DARPA CASTE program, which is oriented to speech to speech translation for narrow semantic domains.
2. *Annotation*: in this area, it is worth mentioning the recommendations on annotation, that were based on previous experiences by SPEX and in the collection of the Dutch Speech Corpus. There was also a discussion on metadata issues through the presentation of initiatives like IMDI and OLAC as well as methods for making collaborative annotation possible, so that the annotation is enriched when the resources are used by third parties. Finally, an important issue is the fact that most of the annotation schemes presented at the conference were based on XML representation.
3. *Tools, Platforms and Procedures*: several tools for annotation were presented, like the NITE XML Toolkit, that was developed for annotating dialogue and multimodal language corpora; the MAUS tool, that allows the production of automatic segmentation and labeling, the MDE annotation tool developed in the DARPA EARS Program and tools for collaborative annotation and for producing automatic phonemic labeling and segmentation.
Also, different platforms and procedures for recording LRs were presented. In particular, I would mention E-WIZ, SpeechRecorder and the Fisher protocol. The first allows the implementation of emotion scenarios and then record voice and video of emotional speech based on Wizard of Oz applications; the second is a platform independent audio recording software that supports speech recordings via more than two channels, and the third was developed to collect conversational telephone speech in the DARPA EARS Program.
Finally, there were several papers presenting different procedures for automatic transcription and segmentation. In this sense, there were methods based on pronunciation variants, on ASR adaptation, on taking advantage of already existing transcripts, etc.
4. *Programs and National and International Activities*: LREC 2004 was also very fruitful from the point of view of the number of programs and initiatives presented. The DARPA EARS program, the NSF TalkBank, the Dutch-Flemish HLT Program, the Technolangue program in France, the ELRA network of validation units, that was created to check and improve the quality of the language resources of the ELRA catalogue, the ELRA initiative to create an Universal Catalogue of language resources, the WALA initiative (West African Language Archive) and the ENABLER European project for adopting de facto standards, best practices, specifications and validation protocols, for promoting the industrial exploitation of language resources, etc. It is also important to mention that ENABLER supports BLARK and ELARK which goal is to define an updated set of language resources that should be minimally available for as many languages as possible.
Concerning the area of Multimodality, papers could be grouped into two main categories:
1. *Multimodal resources*: There were important contributions in the area of multimodal resources as well. In particular, I would mention the effort carried out in the creation of corpora that combines speech and gestures, like the corpus composed of conversations about blood pressure (containing speech and gestures) or the corpus composed of utterances and pointing expressions. In all these papers, the need of multimodal corpora for constructing computer models of multimodal human communication and the problems associated to the annotation and synchronization of speech and gestures are addressed, so that some proposals for improving the annotation process were presented. It is also worth mentioning the work done to use audio-visual information for improving the word accuracy of automatic speech recognizers in car environment (the AV@CAR corpus) and the NSF Talkbank project, that has audiovisual recordings of human and animal communication.
2. *Annotation and tools*: In this area, there was an interesting discussion on annotation schemes and recommendations for linking coreference relations between linguistic expressions and images, on codings, on metamodels like MMIL for representing semantic content in multimodal context (linguistic, gesture, graphical events, dialogue acts, etc.) and on challenges in the development of annotation tools to easily entry different coding schemes, to allow unlimited cross-level and cross-multimodality encoding, to facilitate the presentation of data coded with different coding schemes and to automatically or semi-automatically perform the annotation.
In addition to this, there were presentations about annotation tools like ELAN, that allows collaborative annotation through Internet, tools for annotating videos of Sign Language and tools for multimodal alignment of text and speech.
Therefore, we can conclude that there is a growing interest in LRs, annotation and coding schemes, tools and quality, which mirrors how important for creating, developing and testing new technologies, products and services, LRs are.
In LREC 2004 we have seen the important effort that has been made in the area of LRs. There are still languages for which there are no available or enough SLRs, there is a lack of a common annotation standard and coding scheme and there is a need for better tools capable to speed up the annotation process. However, there are many projects, initiatives and discussions working in these directions as we have seen during the conference, which makes next LREC even more interesting.
See you at LREC 2006!

Daniel Tapias
Telefónica Móviles España
C/ Serrano Galvache, 56
28033 - Madrid, Spain
Email: tapias_d@tsm.es

Report on Papers on Evaluation for Spoken and Written Language

Joseph Mariani

After decreasing from 30% in 1998 to 25% in 2000, and 20% in 2002, the ratio of papers in the area of evaluation is now stabilized at about 20% this year, but evaluation is now used in all areas of Language Technologies: 50% of the evaluation papers are on written language, 30% on speech, 5% across spoken and written language, 10% on multimodality and 5% on terminology.
We find evaluation activities presented at the conference in various domains: for the written language, on POS segmentation and tagging, syntactic and semantic parsing, content extraction, spelling checkers, sense distinction, coreference resolution, summarization, (crosslingual) information retrieval, Question & Answers

(Q&A) , Machine Translation, human authentification… For the spoken language, on speech recognition, oral dialog, speech synthesis, speaker recognition, speech-to-speech translation… On terminology extraction. And, for multimodal communication, both on understanding and generation, and on human communication in Virtual Reality environments.
Most papers are on technology assessment, few are on usability assessment, and some address the issue of data quality in various areas, going from Language Resources to textbooks. It appears that evaluation is used for many more languages than initially: not only in American English, but also in French, German, Japanese, Portuguese, Dutch, Russian, Czech, Slovenian, Arabic, Spanish, Basque, Cypriot… Evaluation is used in very different application areas (medical, patent retrieval, home appliances, car, meeting transcriptions…), and Language Resources of various kinds are more and more involved in those activities (Wordnets, Treebanks, (semi-)transcribed corpus…).
More and more on-going programs include evaluation at the international level (Senseval, Aurora -front end of ASR), in the United-States (TREC, EARS, ACE, TDT… conducted by NIST, which presented a "LT Evaluation cookbook" at this conference, with support from DARPA), in the European Union, within the European Commission (CLEF, TC-STAR, CHIL…), or in various European countries, such as the TechnoLangue program in France with 8 evaluation campaigns (EASY, ESTER, EQUER, MEDIA…), or the activities in Switzerland, with the evaluation of both academic and commercial speech recognition systems.
It now appears that there is a 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 in another language? Is it acceptable that a paper be rejected because it was assessed on a language which is not used for international reference comparison ?
- 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 text retrieval, for example.
Evaluation is now mandatory in the Language Technology R&D activities, in order to know where we are, and how we make progress.

Joseph Mariani
Directeur, Département "Technologies de l'Information et de la Communication"
Direction de la Technologie, Ministère Délégué à la Recherche
1 Rue Descartes
75231 PARIS cedex 05, France
Tel.: +33 (0)1 55 55 89 86
Fax: +33 (0)1 55 55 98 73
Email:
Joseph.Mariani@technologie.gouv.fr
Website: www.recherche.gouv.fr/technologie/

Report on Written and Terminological Language Resources

Jan Odijk

I present a sketchy characterization of the Written and Terminological Area at LREC 2004.
I have chosen to follow the schema of the similar reports prepared for the previous LRECs (made by Nicoletta Calzolari), which makes it easier to comparatively assess the main tendencies in the field. But because the previous reports also covered the General Area, these comparisons are not perfect. I will point out where this is the case.

Parameters for Classification
(see table on page 22)

This year we received an even more impressive amount of papers for the Written and Terminological Linguistic Resources (WTLR) area than in earlier years, such that often three (sometimes even four) parallel sessions on WTLR 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 WTLR. The following table summarizes the findings: (purple cells denote areas with interesting increase, while grey ones denote decrease with regards to previous LRECs).

Levels of Linguistic Description

There were a considerable number of papers dedicated to Morphology, though it followed the trend already set in in earlier years with a decrease partly due to the fact that it is a more or less consolidated area where many practical tools/systems exist for many languages. The exception to this is the issue of compounds, which continues to pose challenges both for language and speech technologies. This is probably due to its productive, but partially capricious, nature.
The major interest remains, as before, in Syntax and Semantics. The Syntax area thus consolidates the trend set in earlier in becoming an ever more robust field to build large resources for many languages. Semantics, on the other side, remains a hot topic: the major topics in this area, semantically annotated corpora and tools (for annotating them), work building upon-, extending- and enhancing-(Euro)WordNet, automatic acquisition of semantic properties for lexicons, and Word sense disambiguation.
As to Terminology, we see a clear increase in the number of papers, not only absolutely but also relatively, as seen in the table. The most represented topic was Automatic identification and extraction of terms, followed by papers on ontological- and knowledge-based approaches. Finally, there was a substantial number of papers on tools 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 LR, for which languages, in which state of development, 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; the issue of annotation of corpora is also getting more and more in important. The techniques used here are to a large extent statistical in nature, but we often see interes-

ting combinations of statistical and linguistic approaches.
The Web as a source for language data continues to be explored and utilised, and grows in importance. Finally, metadata remains a quite hot topic; this is positive, since we hope it will provide the means to obtain a better, quicker, and easier access to existing resources, contributing to optimise their use.
Important areas in which the techniques mentioned are applied are multiword and terminology extraction as well as identification of named entities. Other important topics covered at LREC 2004 are coreference and anaphora.
Many papers covered technologies and tools for creating language resources more efficiently, not only semi-automatically but also in a fully automatic manner; methodologies for creating dynamic, self-adaptive and continuously improving language resources were also mentioned.
Policy Issues and Large Programs may appear to be less well represented in Lisbon: instead, they were classified under the General Area and the tables give a distorted picture. I will come back to Policy Issues later.

Resources and Systems

As to the types of resources and systems described at this conference, we see that work on lexicons has taken over the first position. Papers dealing with work on corpora held the first rank at LREC 2002 in Las Palmas, at least for the Written and Terminological Area. However, this does not mean that corpora have become less important: there is still a significant number of papers on corpora and in addition most work on lexicons is corpus-based. So I believe it rather shows that focus has shifted somewhat from the creation of corpora to their use in developing other types of resources and technologies.
There was an impressive number of papers describing systems, tools, components, and related resources. The main applicative areas are:
- Machine Translation and Translation Memories
- Question Answering
- Document Classification
- Information retrieval, mainly monolingual but also Cross-Language
- Information Extraction
- Summarization
- Proofing Tools

Languages

Most papers deal with a single language. There are also papers that deal with two or a few languages, but very few papers that deal with multiple languages.

Policy Issues and Infrastructural Initiatives

The comparison with the past LRECs is not really possible, since the General area was covered in the earlier tables while the current one restricts itself to the written and terminological area. Therefore the policy issues may appear to be underrepresented. However, I would like to point out one issue.
There is an increase in researchers promoting freely accessible, open resources and collaborative creation of language resources. At the same time, we see a contrast on this issue with most industrial representatives, who generally have a mixed attitude on this matter: they want to keep the resources they created for themselves for reasons of competitive advantage; then, they are prepared to share data only because they are not able to carry the costs for the wide range of resources they need. This issue has been around for some time, and is also popping up clearly in the new Dutch HLT programme. The Dutch Language Union is preparing activities to investigate in this matter, but I believe, as Steven Krauwer suggested, that it would be good if ELRA could play a role at the international level in attempting to design a clear model to deal with this in a way that is satisfactory both for industry and for academic researchers.

Parameters for Classification	Lisbon	Las Palmas	Athens	Granada
Research vs Development				
(Innovative) Research	2°	4°	3°	4°
Large Projects	4°	3°	2°	1°
Tool/system Development	1°	1°	1°	3°
Policy Issues	3°	2°	4°	2°
Type of Resource/tool described				
Lexicon	1°	2°	2°	2°
Corpus	2°	1°	1°	1°
Methods	5°	6°	6°	3°
Task/component	3°	3°	3°	5°
System	4°	4°	4°	4°
Infrastructural Aspects	6°	5°	5°	5°
Level of Linguistic Description				
Morphology	5°	3°	2°	2°
Syntax	1°	1°	3°	1°
Semantics	2°	2°	1°	2°
Ontology/conceptual	4°	4°	5°	5°
Terminology	3°	5°	5°	4°
Other	6°	6°	4°	6°
Language(s)				
One Language	1°	1°	1°	1°
Many Languages	3°	3°	3°	3°
Bi-/MultiLingual	2°	2°	2°	2°

New Resources

ELRA-S0163 ILPho phonetic lexicon

The ILPho database is a phonetic lexicon which contains 39,000 lemmas (319,318 entries). It is distributed in two formats. The first format is compact and corresponds to an easy extension of the text format in which the Multext lexicons (réf. ELRA-L0010) (Ide et Veronis, 1994) are distributed, by adding a column where phonetic transcriptions are stored. The second format is instantiated in XML (see www.xml.org), corresponding to a set of mark-ups specifically designed within this project for lexicons representation.

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

ELRA-S0164 BAS GEO1

BAS GEO1 is a simple database about the most important location names of Germany, Austria and Switzerland together with their canonical pronunciation coded in SAMPA.
BAS GEO1 may be used as a basis for automatic speech recognition of German postal addresses or to feed a speech synthesis algorithm. Future updates will be distributed to all users automatically (if a valid email address is provided).
BAS GEO1 contains 3 data sets:
1) List of all locations with the following fields: Location ID, Gemeinde name, Gemeinde name pronunciation, Postal code, Location name, Location name pronunciation, Kreis name, Kreis name pronunciation, State name, State name pronunciation, Car code, Phone area code, Population (in 2003)
2) A list of all street names: Street ID, Street name, Street name pronunciation
3) A mapping of Locations to Streets: Location ID, Street ID

	ELRA members	Non-members
For research use	172.82 Euro	255.65 Euro
For commercial use	1,400 Euro	2,800 Euro

ELRA-S0165 MICROAES

The ATLAS Spanish Microphone Database (MICROAES) has been collected in Spain by Applied Technologies on Language and Speech, S.L. (ATLAS). This database comprises microphone recordings from 300 different speakers, who have been selected from five different dialectal areas. Sex and age distribution was also considered for speaker selection.
The corpus has 30 sets of 15 paragraphs giving a total of 450 paragraphs. Each 15 paragraph set contains at least two allophones from the extended SAMPA symbols. For this purpose, coarticulation effect between words was considered.
The recording platform is based on a laptop using a PCMCIA slot as interface to the audio equipment. Up to four microphones are recorded simultaneously:
- Sennheiser ME 104 (close distance)
- Nokia Lavalier HDC-6D (close distance)
- Sennheiser ME 64 (medium distance)
- Haun MBNM-550 E-L (far distance)

In this database all recordings have been done in an office with no discussion or meeting during the recordings. The signals are stored in a raw file format, i.e. without headers in the signal file. Each of the four speech channels is recorded at 16 kHz with 16 bit quantization.
A description of the sample rate, the quantization, and byte order used is held in the SAM label file that corresponds to each speech file. This label file also contains information about the signal quality value of the speech file.
The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. Transcription includes segment markers dividing the paragraph in portions of less than 10 seconds using speaker pauses.
The lexicon file included in this database has more that 7400 words with the corresponding pronunciation information using the SAMPA phonemic notation.
The database contains 30 hours of speech and is distributed in 30 ISO 9660 CD-ROM volumes or 5 ISO 9660 DVD-ROM volumes.

	ELRA members	Non-members
For research use	18,000 Euro	22,000 Euro
For commercial use	28,000 Euro	32,000 Euro

ELRA-W0037 The EMILLE/CIIL Corpus

The EMILLE/CIIL Corpus consists of three components: monolingual, parallel and annotated corpora.
There are fourteen monolingual corpora, including both written and (for some languages) spoken data for fourteen South Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu, Urdu.
The EMILLE monolingual corpora contain approximately 92,799,000 words (including 2,627,000 words of transcribed spoken data

for Bengali, Gujarati, Hindi, Punjabi and Urdu).
The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu.
The annotated component includes the Urdu monolingual and parallel corpora annotated for parts-of-speech, together with twenty written Hindi corpus files annotated to show the nature of demonstrative use. All other components are annotated at the sentence level. The corpus is marked up using CES-compliant SGML and encoded using Unicode.
References: Xiao, Z, McEnery, A., Baker, P. and Hardie, A. 2004. 'Developing Asian language corpora: standards and practice' in Sornlertlamvanich, V., Tokunaga, T. and Huang, C. (eds.) Proceedings of the Fourth Workshop on Asian Language Resources, pp. 1-8. March 25, Sanya.

For more information on the Emille project: **http://bowland-files.lancs.ac.uk/corplang/emille/**

	ELRA members	& Non-members
For research use by academic organisations		Free
For commercial use, see below (W0038)		

ELRA-W0038 The EMILLE Lancaster Corpus

The EMILLE Lancaster Corpus consists of three components: monolingual, parallel and annotated corpora.
There are monolingual corpora for seven South Asian languages: Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil, Urdu.
The EMILLE monolingual corpora contain approximately 58,880,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu).
The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu.
The annotated component includes the Urdu monolingual and parallel corpora annotated for parts-of-speech, together with twenty written Hindi corpus files annotated to show the nature of demonstrative use. All other components are annotated at the sentence level. The corpus is marked up using CES-compliant SGML and encoded using Unicode.
References: Xiao, Z, McEnery, A., Baker, P. and Hardie, A. 2004. 'Developing Asian language corpora: standards and practice' in Sornlertlamvanich, V., Tokunaga, T. and Huang, C. (eds.) Proceedings of the Fourth Workshop on Asian Language Resources, pp. 1-8. March 25, Sanya.

For more information on the Emille project: **http://bowland-files.lancs.ac.uk/corplang/emille/**

	ELRA members	Non-members
For research use, see above (W0037)		
For commercial use	7,500 Euro	12,000 Euro

ELRA-W0039 The Lancaster Corpus of Mandarin Chinese (LCMC)

The Lancaster Corpus of Mandarin Chinese (LCMC) is designed as a Chinese match for the FLOB and FROWN corpora for modern British and American English.
The corpus is suitable for use in both monolingual research into modern Mandarin Chinese and cross-linguistic contrast of Chinese and British/American English. The corpus sampled 15 written text categories including news, literary texts, academic prose and official documents etc published in P. R. China in the earlier 1990s for a total of approximately 1 million words. The same sampling frame and period as FLOB/FROWN were used in LCMC.
The corpus is marked up for text categories, sample file numbers, paragraphs, sentences and tokens. Linguistic annotations undertaken on the corpus include tokenization and part-of-speech tagging. The whole corpus is annotated at the word level and includes orthographic and morphological annotations. The tagging system used was produced by the Institute of Computing Science Chinese Lexical Analysis System (ICTCLAS), the Chinese Academy of Sciences. The corpus is encoded in Unicode (UTF-8) and marked up in XML
The corpus comes with a User Manual detailing corpus design specifications and part-of-speech tags. The XML structure of the corpus was validated using the parser built in Xaira. Part-of-speech tagging of all aspect markers was manually checked.
References: McEnery, A., Xiao, Z. and Mo, L. 2003. 'Aspect marking in English and Chinese: using the Lancaster Corpus of Mandarin Chinese for contrastive language study'. Literary and Linguistic Computing 18/4: 361-378. Xiao, Z, McEnery, A., Baker, P. and Hardie, A. 2004. 'Developing Asian language corpora: standards and practice' in Sornlertlamvanich, V., Tokunaga, T. and Huang, C. (eds.) Proceedings of the Fourth Workshop on Asian Language Resources, pp. 1-8. March 25, Sanya. McEnery, A and Xiao, Z. 2004. 'The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study'. Paper presented at LREC 2004. May 2004, Lisbon.

For more information on the LCMC:
www.ling.lancs.ac.uk/corplang/lcmc

	ELRA members	Non-members
For research use by academic organisations	Free	Free
For commercial use	7,500 Euro	12,000 Euro