The ELRA Newsletter

April - June 2002

*Special Issue **LREC 2002*** *Vol.7 n.2*

Contents

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr or
WWW: http://www.elda.fr

Dear Colleagues,

During the ELRA Annual General Assembly which took place on 26th April, Antonio Zampolli, president of ELRA since its creation 7 years ago, announced that in compliance with ELRA statutes, he was completing his last term as President of ELRA.

The members, the ELRA board and the ELDA staff warmly thank him for his involvement in ELRA activities and his outstanding contribution to its success.

A new ELRA board was elected: Joseph Mariani, from LIMSI-CNRS (France), became President of ELRA, and three new members joined the board: Bente Maegaard, from the Center for Sprogteknologi (Denmark), Teresa Lino, from the Universidade Nova de Lisboa (Portugal), and Nicoletta Calzolari, from the Istituto di Linguistica Computazionale del CNR (Italy). You will find in the next page a résumé of the three new board members and of the new President.

Antonio Zampolli was nominated Honorary President by the board, and Khalid Choukri remains ELRA CEO.

The summer issue of our newsletter reports on the third edition of the international Language Resources and Evaluation Conference, LREC 2002, which took place from 27th May to 2nd June 2002 in Las Palmas, Canary Islands (Spain). It was organised by ELRA, with the support of many international organisations involved in the field of HLT. With over 700 participants, and 39 countries represented, the LREC conference has once again, after the first two editions in 1998 and 2000, proven that it constitutes a milestone in the field of HLT, for both industrial and academic communities.

The success of the LREC 2002 conference is illustrated with several figures: for the main conference, 365 papers were selected, out of the 460 papers which had been submitted and reviewed. The submissions covered most of the areas in the field of HLT: written resources (280 submissions), spoken resources (100), multimodal and multimedia (25), evaluation (50), and terminology (16). As for the workshops, 18 took place, before and after the main conference, addressing a large variety of topics, such as resources and tools in field linguistics, use of semantics in various areas, language resources for Arabic language processing, machine translation evaluation, evaluation of multimodal systems and multimodal resources, etc.

This newsletter is divided into three sections, which aim at depicting an overview of what happened in Las Palmas. The first section includes the speeches that were given during the LREC 2002 Opening Ceremony by A. Martin Municio, A. Zampolli, J. Mariani, K. Choukri and H. Höge. In order to give a more concrete illustration of the event, we present in the second section a few summaries of some sessions and workshops, written by their chairpersons. Mark Maybury, who gave a keynote speech, reports on multimodal systems, resources and evaluation. Another keynote speaker, Kishore Papineni introduces a new method for machine translation evaluation. The speeches of the closing session from N. Calzolari, B. Maegaard, A. Martin Municio, D. Tapias, J. Mariani and K. Choukri can be found in the third section of the newsletter.

A few words should be added on ELRA activities. In April and July, the members of the validation committee, VCom, met in Paris to discuss the issue of the new bug report service offered by ELRA in co-operation with SPEX, for the validation of the SLR available in our catalogue, aiming at distributing data of even better quality. The bug report service is available on our web site. You are kindly invited to contribute to the success of this new service by reporting "true" and valuable bugs. It will benefit for both developers and users of language resources, and prizes will be awarded to the best contributors (the first was awarded to Tony Robinson (CUED) at the LREC 2002 Opening Ceremony). Validation centers for WLR are also presently being set up.

ELRA and ELDA are involved in many on-going activities: the OrienTel project, which aims at creating 26 databases for the Mediterranean and Middle East countries; Euromap LT, which aims at promoting HLT across Europe and where ELDA acts as the contact point for France. In the framework of the Speecon project, which aims at collecting speech data in order to promote the development of voice-controlled applications, the French recordings are now over, and the transcriptions are going on. About 100 hours of broadcast news have also been recorded on an Arabic radio in Paris (Radio-Orient), and are being transcribed, in collaboration with LDC in the framework of the Network-DC project.

In France, the Technolangue action, initiated by the Ministry of Research, has issued a call for proposals in April along four sections: development and reinforcement of language resources, creation of an infrastructure for the evaluation of language technologies, better accessibility to norms and standards and setting up a technological survey in HLT.

We are involved in the preparation of the LangTech 2002 conference, which aims at being an international forum for people and organisations involved in the development, deployment and exploitation of spoken and written language technologies in real world applications, where we are responsible for the technical exhibition. Please contact us at exhibition@langtech.org, or visit the web site at www.lang-tech.org.

New resources are described at the end of the newsletter: a Turkish speech database, SpeechDat-Car in German, a Basque spoken corpus and MultiWordNet.

Joseph Mariani, President

Khalid Choukri, CEO

Bente Maegaard

Born in Copenhagen in 1945, Bente Maegaard studied Mathematics and French at the University of Copenhagen. She joined the Department of Applied and Mathematical Linguistics from 1971 to 1990 first as lecturer then as a Research professor. Bente was appointed a visiting professor at the University of Geneva (ISSCO) in 1981. She headed the Eurotra-Denmark for two years, 1989-90. She was a research fellow at the university of Salford (UK) in 1990. From 1993 to 2001, she was a member of the Board of Directors for Munksgaard Publishers A/S. In 1994-95, Bente participated in the Executive Board of ACL (Association for Computational Linguistics). From 1995 to 2000, Bente acted as a Member of the 'Comité de suivi' for the French language technology programme AUPELF/UREF. Since 1985, she is a member of the Danish Academy for the Technical Sciences (ATV), and for 4 years (1991-95), she was a member of its Board and chairman of its Fundamental and Ancillary Sciences Group. Since 1995, she is a member of the Committee (Board) of EAMT (European Association for Machine Translation). Bente was also the Danish delegate to the Language Engineering Working Party, at the European Commission (Luxembourg). Since 2001, she is a member of Nordisk Forskningspolitisk Råd, and she holds the chair of the Danish research councils' Progamme Committee for IT research, as well as the chair of the Programme committee for the Norwegian research programme for language technology (KUNSTI). Bente was a member of the ELRA Board in 1999-2000, and was reelected last April as vice-president of the association. Bente became very recently, in May 2002 a member of the French “Comité de coordination des sciences et technologies de l'information et de la communication”.

Bente's main research interests and expertise lie in computational linguistics, machine translation, evaluation methodology, dictionaries, and corpora. She was awarded the Levison Prisen in 1991, and the Hartmann Prisen in 1997.

Maria Teresa Lino

Born in Lisbon in 1947, Maria Teresa Rijo F. Lino graduated in Romance philology from Universidade Nova de Lisboa, Faculdade de Letras in 1973, and pursued studies at Sorbonne Nouvelle Paris III, where in 1987 she obtained her PhD in lexicology. In 1996, she was honoured by the French government with the decoration of “Chevalier dans l'Ordre des Arts et des Lettres”.

Since 1977, Maria Teresa Lino works as a professor at the Linguistics Department of from Universidade Nova de Lisboa, where she launched the following subjects at different levels (masters and PhD): lexicology, lexicography; terminology and computer linguistics.

She is also responsible for the organisation of several seminars on the subjects previously mentioned, not only in Portuguese universities, but also in foreign universities.

She heads the University's research teams on lexicology, lexicography and terminology since 1992, as well as the Linguistics centre's Research Unit.

Maia Teresa Lino is responsible for the creation of a terminological data bank linked to scientific corpora and a network system on Portuguese neology and terminology with the cooperation of Brazil and other Portuguese speaking-countries.

She also leads other projects in progress like e.g. TERMEDICA (medical dictionary), terminological dictionary on senology, PHARMATERM (computationnal lexicography of pharmacology).

Maria Teresa Lino is the author and co-author of several publications in Portuguese and on international specialised magazines. She founded the Portuguese Association of Terminology (TERMIP) in 1989, and is an active member of the European Association of Terminology.

Maria Teresa Lino was elected member of the ELRA Board in April 2002.

Nicoletta Calzolari

Nicoletta Calzolari, graduated in Philosophy at the University of Bologna, was first researcher at CNUCE (Centro Nazionale Universitario di Calcolo Elettronico), then researcher at Pisa University, Department of Linguistics, and is now Director of Research (equivalent to Full Professor) at the Istituto di Linguistica Computazionale of the CNR (ILC-CNR) in Pisa, Italy. She works in the field of Computational Linguistics since 1972. Main fields of interest are: computational lexicology and lexicography; text corpora; standardisation and evaluation of language resources; lexical semantics; knowledge acquisition from multiple (lexical and textual) sources; integration and representation. She has co-ordinated and/or technically managed a very large number of international and national projects. From the beginning, Nicoletta was a chief co-editor of the EAGLES Project, and is now European responsible for the Computational Lexicon Working Group of the EU-US ISLE project. She is a member and general secretary of ICCL, and member of many international committees and advisory boards. She was invited speaker, member of program committee or organiser for quite numerous international scientific conferences, workshops, etc.

Joseph Mariani

Born in 1950, Joseph Mariani is a senior researcher at CNRS. He is now director of the "Information and Communication Technologies" department at the French Ministry in charge of Research (Division of Technology), where he manages various activities, including national R&D Networks on Telecommunications, Micro and Nano-Technologies, Software Engineering and Audiovisual & Multimedia. He was the general director of LIMSI, a CNRS laboratory in Orsay (France), from 1989 to 2000, while being also responsible of its Human-Machine Communication department, which develops research activities in spoken and written language processing, non-verbal communication (computer vision, computer graphics, gestural communication) as well as in multimodal communication, human perception, cognitive psychology and socio-economics of interactive communication. He was president of the European Speech Communication Association (ESCA, now ISCA), and was vice-president of the European Language Resources Association (ELRA) from the very beginning. He is a member of the Executive Board of the European Language and Speech Network (Elsnet), and of the advisory council of the Cocosda international committee. He was a member of the CNRS Scientific Council and Engineering Sciences department council, member of the Evaluation Committee of the French Information Science Institute (INRIA), coordinator of Francil, the Language Engineering Network of the Francophone Universities Agency (AUF), and coordinator of the "Human-Machine Interaction, Ergonomics and Acceptability of Services" committee of the French National Network on Telecommunications (RNRT). He served as a member of advisory committees for the US TIDES program and for the French programs on Audiovisual and Multimedia Industries (PRIAMM), Language Resources (DGLF-LF) and Computational Language Processing (CSLF). He is the author or co-author of more than 300 papers.

LREC 2002 Opening Ceremony Speeches

Angel Martin Municio
Real Academia de Ciencias Exactas - Spain

To the Spanish authorities, presidents of ELRA, members of the local committee, dear friends,
In the short life of the association, it is the second time that our international conference takes place in Spain.
In fact, the third international congress organised by ELRA with the support of other international institutions is now open in one of the most admired and loved regions of the country, the Canary Islands; and particularly in the very singular and beautiful location of Las Palmas as you could have the opportunity to realise. For the two previous editions, the cities of Granada and Athens offered to our conference their ancient cultural heritage for reinforcing the ideas and the enthusiasm for the European unity. The geographical parameters of the Canary Islands not only represent one of the most southern and western corners of Europe, and as such facing the American continent; the culture and the traditions, the musicality of their languages, and the modern life of these Islands will also serve as the best symbol for both: the firm links with the American nations and the encouragement to further advance of our work in the field of human language technology, a strong prerequisite for the unity of Europe.
The mayor of Las Palmas as well as the Spanish telephone company, Telefónica, have contributed to make possible this conference. The presence in this opening ceremony of the mayor of the city and that of the regional director of Telefónica gives us the first opportunity to offer them our warmest thanks.
I would like first, after the statutory cease of Antonio Zampolli as President of the Board of ELRA, to thank him publicly for his friendship and all the physical efforts, experience and technical knowledge he has dedicated during the past ten years to the birth and development of the association.
All of you also know that the organisation of such an event implies a lot of efforts, many silent activities, carried out in the local organisation of this conference by some people headed by Manuel González, Dean of the Faculty of Informatics of the Politechnical University of Las Palmas, Daniel Tapias and Nicoletta Calzolari, members of the ELRA Board. To them our gratitude.
Welcome finally to the LREC conference and a lot of thanks to the Ambassador Tomás Solís, who represents the Spanish Ministry of Foreign Affairs, for being with us at this opening ceremony.
On behalf of the Local Organising Committee, welcome again to Spain to all the conference participants.
Thank you all!

Angel Martin Municio
Real Academia de Ciencias Exactas
Fisicas y Naturales
Calle Valverde 22
28004 Madrid (Spain)
Tel.: +34 91 701 42 30
Email: presidente.racefyn@insde.es

Antonio Zampolli
Chairman of the conference, Honorary President of ELRA
Istituto di Linguistica Computazionale del CNR - Italy

First of all, let me express my warmest gratitude to the Authorities who have honoured our opening ceremony, witnessing in this way the relevance of our field for the harmonised development of our society.
It is a pleasure for me to welcome all of you to this third edition of the International Conference on Language Resources and Evaluation (LREC).
The first edition of the conference, four years ago in Granada, and the second one, two years ago in Athens, were truly successful, as the number of submissions to the present one clearly indicates.
I hope that this conference here in Las Palmas will equally contribute to establish LREC as a permanent initiative, strongly contributing to the progress of our field. At present, I am not informed about the existence of another international conference that programmatically promotes, at the same level, the interaction between research and development, speech and language, empirical and rule-based methods, multimodality and international co-operation.
Many papers presented here - both oral and poster - clearly show that our field is a very composite one: on the one hand, LRs and evaluation are central components of the linguistic infrastructure, which is an essential pre-condition for the full development of the potentiality of HLT and its applications for the benefit of our global information society.
As clearly emerged in the discussions in Granada and in Athens, a number of organisational and policy problems remain, for a large part, yet unsolved; on the other hand, the provision of adequate LRs and evaluation methods is not only a practical task which demands a labour-intensive production work, but also presents challenging research issues, at the forefront of research in HLT, such as the integration of different modalities, semi-automatic knowledge extraction from corpora, standardisation of linguistic description, methods for annotating large LRs.
Let me express my warmest gratitude to all those who have contributed to the preparation of the conference: from the ELRA Managing Board to the Programme

Committee, with particular reference to Doctor Daniel Tapias; from the Local Organising Committee to the International Advisory Board; from the ELDA staff to the sponsors which have generously contributed to the financial efforts, and to the various Organisations which have accepted our invitation to co-sponsor the conference.
In particular, I wish to thank the persons working for my Institute and for the University of Pisa who have contributed to the organisation of the conference and, first of all, Doctor Nicoletta Calzolari for substituting myself in a lot of tasks during the months of my illness.
The choice of the Canary Islands as the venue of this conference, suggested and generously supported by Professor Angel Martin Municio, has certainly contributed to the increase of the number of the participants but, at the same time, has made the solution of many organisational problems difficult. I apologize for the consequences of these difficulties, which have been increased furthermore by my present health condition. The scientific success of the conference depends on your participation: I am sure that the results of the conference will be very influential from the scientific, application-oriented and organisational standpoint.
In particular, I am sure that the conference wil1 facilitate the creation and the consolidation of a de facto community, to which researchers and developers of different thematic and geographical areas - who seldom or never have the occasion to meet - will feel to belong, sharing problems, mutually benefiting from resources, joining knowledge and efforts to search for solutions.
I wish all of you a successful conference and a pleasant stay in Las Palmas.
I hope you wil1 accept with benevolence any inconvenience or problem our organisation might cause to you.

Antonio Zampolli
Consorzio Pisa Ricerche
Via della Faggiola 32
I-56100 Pisa (Italy)
Tel.: +39 050 3 15 28 37
Fax: +39 050 55 50 13/62 85
Email: pisa@ilc.pi.cnr.it

Joseph Mariani

President of ELRA
Director, department "Information Technologies and Communication", Technologie direction, French Ministry of Research

I will say a few words, as the new ELRA president, following the elections which took place at the last General Assembly and at the first new board meeting, back on April 26th in Paris.
Taking over from Antonio Zampolli in this duty is both a pleasure and an honor.
The pleasure to share with him the creation of the European Language Resources Association, ELRA. I participated in the Relator project, coordinated by him and his institute in Pisa and supported by the European Commission in its 4th Framework Program, where the concept of ELRA was worked out, resulting in the launching of the association in 1995. As the chairman of the Relator Advisory Committee, Brian Oakley made a major contribution to make it real, and we also received much support from Vincente Parajon-Collada, as Deputy Director of DG 13 at that time. A board was elected and, since then, was renewed several times, allowing the association to benefit from the ideas and contributions of those board members over the years. I take this opportunity to thank all of them.
It was decided to organize ourselves in three colleges: spoken, written and terminology. The two first colleges have flourished since then, but we have some concern with the third one, which will need careful attention.
One of the first task we had to carry out was the nomination of a Chief Executive Officer for the association, and we had a very difficult task to make such a selection among excellent candidates, which resulted in the choice of Khalid Choukri. It was quickly followed by the creation of ELDA, the ELRA Language Distribution Agency, which allowed to gather around Khalid the task force which was necessary to first identify language resources of interest, and then attract members.
For the first time in Europe, there were people spending 100% of their time thinking about language resources, investigating the various aspects of resource identification, validation, distribution and maintenance, and addressing the legal and commercial questions related to their distribution worldwide.
Several Working Groups were set up to help us on that duty, and I would also like to thank all their members for their contribution during the critical period of the launching of the association.
When I participated in the Relator project, I was asked to draw the ELRA business plan. This was a difficult exercise, and I was anxious to see if the future was in agreement with my guesses. I am very happy to report that, with more than 90 institutional members from 19 countries, and a catalog comprising about 700 resources (200 in the speech domain, 200 in the written area and almost 300 in the field of terminology), the best targeted numbers have been achieved. So I now feel much more relaxed.
Distributing language resources was the major aim, but we quickly figured out that it was also our duty to help the international community gathering to discuss and exchange on that topic, which goes very naturally together with the evaluation topic. In close connection with Elsnet, the European Language and Speech Network, we decided to launch a biennial conference, which resulted in LREC. We met success from the very first issue of the conference in Granada in June 1998, and consolidated this success in Athens two years later and this year in Las Palmas de Gran Canaria, with more than 700 participants.
We have more initiatives and more activities going on for the future.
Validation of language resources is one of our main concerns. A Validation Committee chaired by Harald Höge is considering this aspect, both for spoken and written language, and a network of validation centers is being installed, with SPEX in The Netherlands as the very first node in this network. Harald Höge will say some words, later on, on this validation activity, which includes a “bug report” award.
Evaluation is another topic of great importance for us, which needs more attention and more efforts. ELDA has changed its

name to Evaluation and Language resources Distribution Agency. We are already involved in several projects dealing with evaluation, such as Aurora or CLEF, where we provide the development and test data appropriate to conduct the evaluation campaigns. But we will consolidate and extend our activity in that field.
Language resources and evaluation appear as the building blocks on which one should construct the language technology edifice. Two more items have to be added: Standards and here we support the initiatives which are presently being taken, such as the one in ISO, in order to ensure interoperability of systems and sharability of resources. And Survey of the technological transfer, and here ELRA participates in the organization of the LangTech conference which will take place in Berlin in September 2002, and is jointly organized with Euromap, under the auspices of Bente Maegaard, and Elsnet.
Finally, all this has to take place in an international framework.
Language is a major issue for the European Union, and for the new countries which will join the Union in the near future. We should find a way to ensure that all European languages benefit from the tools they deserve, to facilitate their use over the information and communication means. But the European Commission itself may have difficulties in finding the budget that has to be placed in front of an effort of that size.
It is my belief that Language Technologies would be a particularly good example to experiment the construction of the European Research Area, and to see how European national efforts can meet and reinforce the European Commission ones.
On a larger scale, it appears that research is now international, industry is getting international, activities in the area of language resources, evaluation and standards would benefit from more international collaborations, and ELRA and LDC already paved the way with a joint Net-DC transatlantic project. Harmonizing the various national or transnational programs in order to join forces for solving the language processing problem in a cooperative way is probably the next challenge that we should address, together with all our friends around the world.
All those activities, past and present, as I said, included the pleasure to work with Antonio Zampolli.
It was also an honor to work with a man who devoted his huge talent and energy to make it exist, to make it work, to make it a success.
In recognition for his outstanding contribution to the success of the association, the board, during its meeting of April 26th, elected Professor Antonio Zampolli as Honorary President of ELRA.
There was a great vision, and there was a long and difficult way to achieve that goal. We've made it! You've made it! Congratulations, Antonio! Congratulations, Mr President !

Joseph Mariani
Direction de la Technologie, Ministère Délégué à la Recherche et aux Nouvelles Technologies
1 Rue Descartes
75231 Paris cedex 05 (France)
Tel.: 01 55 55 89 86
Fax :01 55 55 98 73
Email: Joseph.Mariani@technologie.gouv.fr
Web site: www.recherche.gouv.fr/technologie/

Khalid Choukri

ELRA CEO

Let me first express my deepest thanks and gratitude to Antonio Zampolli for these splendid years we have all together devoted to ELRA. I would like to say in my name, on behalf of the ELDA team and ELRA members, how enjoyable these years have been, and that we expect to continue to benefit from your guidance and support for the years to come, as you [Antonio Zampolli] have been nominated Honorary President of ELRA.
We would like to ensure that, as a duty, you will think of writing the history of ELRA, closely linked with the history of language resources and NLP, for the new generation. For you to do so, we have thought of a special gift…Now you have the tool! [Antonio Zampolli was offered a fountain pen by Khalid Choukri, on behalf of the whole ELRA]
Now, ladies and gentlemen, let me tell you a few words about ELRA and ELDA to allow you to better understand the whole picture.
ELRA is a European non-for-profit organisation which was created to promote language resources and the whole field of Human Language Technologies (HLT). The association is headed by an elected board, which consists of 12 members, and which defines the policy and the strategies. These are implemented by ELDA, the Evaluation and Language resources Distribution Agency. ELDA acts as ELRA's operational body.
Membership to ELRA is open to all institutions, and affordable. Our members are distributed according to the college they belong to: speech, written or terminology: 2/3 of our members belong to the speech college, 1/3 to the written field, and very few work in the field of terminology. They are sorted into academic institutions or industry.
We have currently at ELRA about 120 members. Over the years, there has been an average of 25-30 new members each year, who belong mainly to the speech college.
The catalogue of language resources available at ELRA includes over 200 speech resources, 210 written resources, and 275 terminology resources.
It is as good as you want it to be. You are the suppliers of the resources that we catalogue, it is you who decide on what you want to share with others. We are the middlemen, the intermediary, but ready to play an active role in licensing, logistics, etc.
The LREC conference is a tremendous opportunity to talk to us about what you may have. A lot of things that you feel are useless may be of interest to someone else, a researcher, a PhD student, an industrial who develops technologies, etc.
Based on our experience, we can help you assessing the value of your resources, for free.

Some key resources in our catalogue are e.g. the databases from the SpeechDat family, the EuroWordNet databases, MHATLex, Farsdat, Logotypografia, etc.
Concerning the distribution activity at ELDA, over 200 contracts were signed in 2001, the majority of which belongs to the speech domain. ELRA's resources are mostly bought by ELRA members (3 times more than by non-members). Over the years, the distribution for research use vs the distribution for commercial use has been very stable. For example, for the last 3 years, exactly 58 % of the resources were sold for research purposes, whereas 42 % were sold for commercial use.
ELRA is getting more and more involved in the set up of its validation network. The quality of the resources available in its catalogue is a very important issue for ELRA, and it benefits both to the users and to the providers. We are implementing validation criteria, with the support of our partners, and launching a Quick Quality Check procedure, currently only available for the speech resources, to be able to give these resources a quality flag. We also offer a new service: the possibility for the users of some language resources to report the bugs and imperfections they may find. This service is available from the ELRA web site, but Harald Höge will tell you a bit more about this (see below).
Producing and commissioning the production of language resources is one of ELRA's tasks. We participate in the cross-Atlantic Network-DC project, to collect and transcribe broadcast news in Spanish, Arabic, etc., in the European Speecon project, which aims at creating speech databases for the development of consumer products, in the OrienTel project, to collect speech in Mediterranean languages - to quote a few.
This means that whenever you need specific resources, you can ask us and we will do our best to help you (through partnerships, etc.).
The evaluation activity is a major issue for ELRA & ELDA. We are setting up a team dedicated to this activity.
Initially, we had provided the language resources to be used in the evaluation process (e.g. Aurora), but we have decided recently to get involved in the evaluation campaigns themselves, e.g. CLEF, Cross-Language Evaluation Forum, and more are still to come!
For the promotion of language resources, ELRA is active in projects at European or international levels, like ENABLER, COCOSDA, Euromap LT. We have also set up a close partnership with LDC, our American counterpart, for the distribution of some resources and in the framework of the Network-DC project.
Other projects we are involved in aim at drawing roadmaps for language resources, standardisation, evaluation, etc.: TC-STAR, Intera, ISLE, etc.
As you can notice, ELRA is active and acts as a driving force for many different aspects related to language resources and evaluation. You should visit our web site to have a clear view of our activities. I will now provide you with a few practical information about LREC.
[Khalid Choukri then gave a few practical information that participants may need, for the conference organisation (bus service, ELRA desk, poster areas, commercial centre, etc.)]

Khalid Choukri
ELRA CEO
55-57 rue Brillat Savarin
75013 Paris (France)
Tel.: +33 1 43 13 33 33
Fax: +33 43 13 33 30
Email: choukri@elda.fr
Web site: www.elra.info or www.elda.fr

Harald Höge

Chairman of the Validation Committee (VCom)
Siemens AG

First Prize for the best Bug report

We will seize the opportunity during this LREC 2002 opening ceremony to award the first prize to the people who have best participated in the bug report service recently set up by ELRA.
The quality of language resources is a very important issue in the field of language engineering, and it is one of ELRA's main objective to ensure maximal quality and so to keep the “effort for use” of the resources as low as possible for the users.
In order to achieve this goal, ELRA set up a validation committee which is in charge of investigating the quality of a language resource (validation) and improving the quality of resources. For improving language resources, ELRA launched a bug report service this year. Users of language resources, who find “bugs” in the resource, have the possibility to report these bugs via the ELRA webpage. Currently the bug service is restricted to spoken language resources (SLR) and is performed in co-operation with the SLR-validation center of ELRA - the Dutch institute SPEX.
ELRA is just in the process to set up a validation center for written language resources (WLR). As soon as this validation center, is established the bug service will be extended to WLR.
In order to stimulate the users to report the bugs, the validation committee of ELRA announced bug report prizes. This year, prizes consist of PDAs.
As the head of the ELRA validation committee, I am happy to hand over the first prize for the best bug report, here in Las Palmas. The winner is Tony Robinson, from Sheffield University.

Harald Hoege
Siemens AG; CT IC 5
D-81730 München (Germany)
Tel.: +49 89 636 53374
Fax: +49 89 636 49802
Email: harald.hoege@mchp.siemens.de

LREC 2002 Keynote Speeches

Multimodal Systems, Resources, and Evaluation

Mark Maybury

1. Introduction

Mark Maybury, Executive Director of the Information Technology Division at the MITRE Corporation, gave an invited talk on multimodal systems, resources and evaluation. Mark's talk included a vision of multimodal question answering and an example of content based access to broadcast news video. He described intelligent multimodal interfaces, defined terminology, and summarized a range of applications, required corpora, and associated media. He introduced a jointly created roadmap for multimodality and illustrated an example of an open source multimodal spoken dialogue toolkit. Next he described requirements for, and an abstract architecture of multimodal systems. He concluded discussing multimodal collaboration, multimodal instrumentation, and multilevel evaluation.

Figure 1: Ask Multimodal Questions, Get Multimodal Answers

2. Multimodal Question Answering

A long range vision of Mark's is to create software that will support natural, multimodal information access. As implied by Figure 1, this suggests transforming the conventional information retrieval strategy of keyword-based document/web page retrieval into one in which multimodal questions spawn multimodal information discovery, multimodal extraction, and personalized multimodal presentation planning. In Figure 1 the user of the future is able to naturally employ a combination of spoken language, gesture, and perhaps even drawing or humming to articulate their information need which is satisfied using an appropriate coordinated integration of media and modalities, extracted from source media.

3. Broadcast News Access

As a step toward multimodal question answering, we have been exploring tools to help individuals access vast quantities of non-text multimedia (e.g. imagery, audio, video). Applications that promises on-demand access to multimedia information such as radio and broadcast news on a broad range of computing platforms (e.g. kiosk, mobile phone, PDA) offer new engineering challenges. Synergistic processing of speech, language and image/gesture promises both enhanced interaction at the interface and enhanced understanding of artifacts such as web, radio, and television sources (Maybury 2000). Coupled with user and discourse modeling, new services such as delivery of intelligent instruction and individually tailored personalcasts become possible.

Figure 2 (shown next page) illustrates one such system, the Broadcast News Navigator (BNN) (Merlino et al. 1997). The web-based BNN gives the user the ability to browse, query (using free text or named entities), and view stories or their multimedia summaries. For example, Figure 2 displays all stories about the Russian nuclear submarine disaster from multiple North American broadcasts from 14-18 August 2000. This format is called a Story Skim. For each story, the user can view story details, including a closed caption text transcription, extracted named entities (i.e. people, places, organizations, time, and money), a generated multimedia summary, or the full original video.

In empirical studies, Merlino and Maybury (1999) demonstrated that users enhanced their retrieval performance (a weighted combination of precision and recall) when BNN's mixed media presentations instead of mono-media presentations (e.g. text, key frames, video). In addition to performance enhancement, users reported increased satisfaction (8.2 on a scale of 1 (dislike) to 10 (like)) for mixed media display (e.g. story skim, story details).

4. Applications, Corpora, and Media

Table 1 (shown next page) illustrates a range of multimodal applications and associated corpora and media. What's different about these corpora from traditional linguistic corpora? Notably, the applications and associated multimodal corpora incorporate temporal and/or spatial dimensions. Consider the following examples:

- *Multimodal question answering*. The ability of users to articulate queries by typing, speaking, drawing, or singing and the ability to receive results in a range of integrated but heterogeneous media.
- *Intelligent multimodal interfaces*[1] that support more sophisticated and natural input and output, enable users to perform complex tasks more quickly, with greater accuracy, and improve user satisfaction. Intelligent multimodal interfaces are becoming more important as users face increasing information overload, system complexity, and mobility as well as an increasing need for systems that are locally adaptive and tailorable to heterogeneous

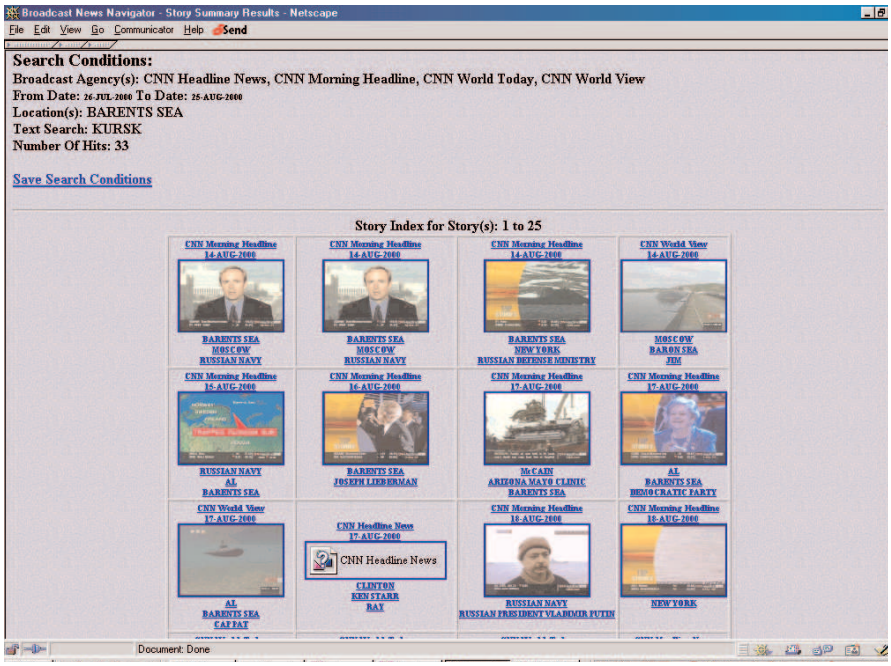

Figure 2: Tailored Multimedia News

user populations. Intelligent multimodal interfaces are typically characterized by intelligent multimodal dialogue (Maybury and Wahlster 1998, Maybury 1999).

Meeting transcription. Video tapings of human behavior that include not only (written or spoken) language discourse and visual events, but also capture the physical location of participants (in space but also in the video frames), changes in their properties over time (e.g. position to one another, attention, emotional state), and so on.

Multimodal authentication in which multiple biometric signatures of users (e.g. voice, face, eyes, gestures) are utilized to determine the identity of an individual in order to provide access control and behavior monitoring.

Each of these situations might imply audio, visual, and/or tactile modalities. Associated media have temporal extent and implied sequencing. They frequently contain information with spatial extent, coming in the form of user input, information accessed, or properties of the environment. For the user, spatial information can come from gaze or gestures (facial, hand, body) articulated by the user or system, the location (absolute or relative) of the user or the retrieved information or object (e.g. GPS coordinates of a car on a road) or simply a characteristic or property of the information retrieved (e.g. a map, blueprint, CAD/CAM diagram).

Collection and annotation of multimedia corpora is challenging. Application requirements differ in needs, such as fidelity (e.g. degree of geoposition specificity), accuracy/error rate, and timeliness. There are no standard mark up languages much less common ontologies for such phenomena as time and location, although there are several ongoing international initiatives (Cunningham et al. 2000). Evaluation of these applications is also challenging for a number of reasons, not the least of which is they are often interactive and thus it is almost impossible to replicate exact human behavior across sessions. A recent international workshop (Bunt et al. 2001) addressed future directions in multimodal systems.

Application area	Corpora (and models)	Media
Multimodal question answering	Question and answer corpora	Text, speech, graphics, video
Intelligent multimodal interfaces	Human-machine interaction corpora	Text, speech, non-speech audio (e.g. sounds, music), gaze,gesture, video
Lifelike interface agents and/or robotic interfaces	Interaction corpora (human physiology models)	Speech, gaze, gesture (facial, hand, body)
Meeting transcriptions (and human behavior analysis)	Human human communication corpora, meeting corpora	Video analysis of speech, gaze, gesture, drawings)
Authentication	Multimodal biometric corpora	Text, speech, face, iris, gesture

Table 1: Applications, Corpora, and Media

References

- Bunt, H., Maybury, M. and Wahlster, W. Dagstuhl *Seminar on Coordination and Fusion in Multimodal Interaction*. Oct. 28-Nov.2,2001. www.dfki.de/~wahlster/Dagstuhl_Multi_Modality
- Cunningham, H., Roy, D. and Wittenburg, P. 2000. *First EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora.* LREC 2000, Athens, Greece, 29/30 May.
- Merlino, A., Morey, D. and Maybury, M. 1997. *Broadcast News Navigation using Story Segments*, ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391. www.acm.org/sigmm/MM97/papers/morey
- Merlino, A. and Maybury, M. 1999. *An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News.* Mani, I. and Maybury, M. (eds.) Automated Text Summarization, MIT Press.
- Maybury, M. February 2000. *News on demand: Introduction*. Communications of the ACM. Vol 43(2): 32-34.
- Maybury, M. December 2001. *Collaborative Virtual Environments for Analysis and Decision Support*. Communications of the ACM 14(12): 51-54. www.acm.org/cacm/1201/1201toc.html.
- Maybury, M. T. and Wahlster, W. editors. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Press.

Acknowledgements

Appreciation for leadership of MITRE activities described herein goes to Laurie Damianos, Bea Oshika, Sam Bayer, Lynette Hirschman, Stanley Boykin, Warren Greif, Scott Mardis, Anita King, Rod Holland, Jay Carlson and Michael Krutsch.

[1] *See www.mitre.org/resources/centers/it/maybury/iui99 for an on-line tutorial on intelligent interfaces*

Mark T. Maybury
Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730 (USA)
Email: maybury@mitre.org
Web site:
www.mitre.org/resources/centers/it

Approximating Human Judgment of Translation Quality Automatically

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu

The advent of large parallel text collections, increased computing power, and reliable automatic evaluation metrics heralds an exciting new era for high-quality machine translation. Demand for machine translation technology is taking off as global information exchange proliferates on the internet. This has spurred a worldwide resurgence of machine translation research centered around data-driven techniques. Today's computing power enables us to try many algorithms in a short time on vast amounts of data. However, this is of limited use without automatic methods to evaluate translation quality. Researchers and developers will benefit from reliable automatic evaluation. Automatic metrics will accelerate the development cycle. One such metric, BLEU (Papineni, 2000), has proven effective for judging quality of translation into English from three language families.

Evaluating translation quality is considered difficult because there is no single gold standard or ground truth for translation. There are many possible correct translations of a given source text, differing in word choice and word order. These differences must be accounted for when judging the quality of a translation. Human judges of translation quality take these and many more subtle aspects into consideration. Collective human judgment of translation quality is therefore the gold standard of evaluation itself. However, such human evaluations are very expensive, and they take a long time to finish. Nor do we benefit from the past human effort when a new system must be evaluated. For MT system developers there is a constant need to evaluate MT quality so that they can weed out bad ideas from good ones. They need automatic evaluation of translation quality that is cheap, fast, and good.

How to measure the goodness of an automatic metric? The grand objective of any automatic metric is to approximate collective human judgment. Then we can view automatic metrics as predictors of human judgment. Prediction error of a metric is then a natural measure of goodness of a metric. Prediction error is related to correlation: the higher its correlation with human judgment, the better the metric is. *Bilingual Evaluation Understudy* (BLEU) is a new method for automatic evaluation of translation quality that correlates highly with human judgment across language pairs from different language families.

The central thesis of BLEU is that the closer a machine translation is to professional human translations, the better it is. The closeness measure, to be described later, is inspired by the precision and recall concepts from information retrieval and the word error rate in speech recognition that has driven the progress in speech technology for over a decade. However, these concepts are modified to take the multiplicity of gold standards into account. If there were a single gold standard for translation, then the traditional word error rate would be sufficient to judge the quality of a translation. BLEU indeed turns the apparent adversity of multiplicity of reference translations into an advantage. The more professional reference translations, the better it is for BLEU.

BLEU does not eliminate human effort altogether. Rather, it shifts the effort from expert judges to professional translators in that it requires one or more high quality reference translations. This up-front one-time cost is shared across all system evaluations. The marginal cost of evaluating a new system is negligible. The evaluation itself takes only seconds.

BLEU has two component scores. One is a precision score derived by counting the number of n-gram matches between the candidate translation and the reference translations. We typically count n-gram matches for *n* ranging from 1 up to 4. Shorter n-gram matches account for adequacy of the translation while longer n-gram matches account for fluency. The n-gram match counts are first turned into modified precision numbers and then geometrically averaged to get the precision score. Precision is commonly defined as the fraction of candidate items that are correct. For example, *1-gram* precision is the fraction of words in candidate translation that are also in the reference translations. According to this definition, the unigram precision of a silly translation such as “the the the the” is 1.0 if any reference translation uses the word “the”. This is not the precision number that BLEU uses. The problem with this example is clear: a reference word should be considered exhausted once a matching candidate word is identified. BLEU assigns a *modified* precision of 1/4 if “the” appears only once in any of the references and 2/4 if “the” appears twice in any reference and so on.

The second component of BLEU is a brevity penalty that penalizes unreasonably short translations. Translations that are brief compared to the reference translations incur a penalty that depends on the comparative brevity. So, in order to score high, a translation must match the reference translations in length as closely as possible. Once the length is approximately the same as the references, a translation must produce the same words in roughly the same order as the references to get high precision score. BLEU score is the product of the brevity penalty and the precision score. It is normalized to give a score of 1 to a translation that is identical to any of the reference translations.

Clearly, target sentences that do not share words with reference translations get a BLEU score of 0 - no matter how fluent or grammatical they are. Those that get high scores will match many long n-grams with references and tend to fluently splice reference translation snippets together. The n-gram matching simultaneously accounts for fluency as well as fidelity, assuming that the reference translations are fluent and faithful. In summary, to score high on BLEU, a translation must match references in length, in word choice, and in word order.

Automatic metrics derive their strength from quantity - averaging over individual errors. We view automatic metrics as statistical predictors of human judgment. So, they make prediction errors. Prediction errors have two components: bias and variance. Bias is the difference between human judgment and the metric on a given

test corpus. Variance measures the variability of the metric across different test corpora. If the test set size is too small, variance will be high and becomes smaller and smaller as the test set size increases. Human judges can assess the quality of translation by looking at just a few sentences, but automatic metrics cannot - they need more data to average over. BLEU is no exception. If there is only one sentence to test and there is only one reference translation, BLEU may assign a very low score to a perfect candidate translation if the candidate translation happens to paraphrase the reference translation using synonyms. With many reference translations, this effect disappears. Also, as the test size increases the variance of BLEU score gets closer and closer to zero. Fortunately, increasing the test size is very easy and hence variance is not a real issue. Simply by increasing the test set and the reference translations, we derive a high-quality automatic score: *quantity leads to quality*!

In practice, we do simple text normalization before matching n-grams. Case-folding is the main normalization used currently. But other sophisticated components could be used in the BLEU framework. For instance, the matching can be done after morphological reduction. Another possibility is to weigh n-gram matches differently based on the type of n-grams matched. For instance, named-entity n-grams can be given higher weights than other n-grams. The baseline implementation of the BLEU method treats all words equally after case-folding. Since BLEU considers variable-length n-gram precisions, there is flexibility in choosing the maximum-length of the n-grams. When the translation quality is higher, fluency becomes a better differentiator than adequacy. Therefore higher translation quality warrants matching on longer n-grams. When the translation quality is poor, adequacy is better differentiator than fluency. Lower translation quality warrants the use of shorter n-grams. Similarly, when the word-order is not important in the target languages, shorter n-grams are more important.

To assess BLEU's correlation with human judgment, we obtained judgments of translation quality by a pool of judges. An automatic metric ideally predicts human judgment robustly across the spectrum of translation quality and across language families. To assess the robustness across the quality spectrum, we mixed human and machine translations in the set of translations that the humans judged. The hope is that the metric will be useful in future when the MT quality approaches that of human translation if the metric can assess now the difference between the quality of human translations. To test the robustness across several language families, we considered translations from Arabic, Chinese, French, and Spanish into English. The BLEU score correlates highly with human judgments. On Chinese-English, it attains a correlation (R) of 0.99. That is, the prediction error is about 2%. On Arabic-English, the correlation is 0.98. On French-English (DARPA-94 evaluation data), the correlation with Adequacy judgment is 0.94 and with Fluency is 0.99. On Spanish-English (DARPA-94 evaluation data), the corresponding numbers are 0.98 and 0.96.

In summary, human judgment of translation quality is the gold standard of evaluation. Automatic metrics attempt to approximate human judgment. By simple counting of n-gram matches with a corpus of good-quality reference translations, we can automatically approximate human judgment remarkably well.

References

Papineni, K. Roukos, S. Ward, R.T. and Zhu W-J. (2002) *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of ACL-2002, Philadelphia, USA.

Kishore Papineni
IBM T.J. Watson Research Center
P.O. Box 218, Room 23-126D
Yorktown Heights, NY 10598 (USA)
Tel.: 914-945-2939
Fax: 914-945-4490
Email: papineni@us.ibm.com

LREC 2002 Sessions summaries

Review on the session "Large Project Initiatives for Speech Corpora"

Justus Roux

This session presented an excellent overview of activities related to the development of large speech corpora in Europe, Japan and in the Americas. The following four presentations were made:

The C-ORAL ROM project: New methods for spoken language archives in a multilingual romance corpus

Emanuela Cresti, Massimo Moneglia, Fernanda Bacelar do Nascimento, Antonio Moreno Sandoval, Jean Veronis, Philippe Martin, Khalid Choukri, Valerie Mapelli, Daniele Falavigna, Antonio Cid, Claude Blum

This well-prepared paper on a large European consortium project coordinated by the University of Florence was presented by Dr. Moneglia. C-ORAL-ROM is a multilingual corpus of spontaneous speech of around 1,200,000 words representing the four main Romance languages: French, Italian, Portuguese and Spanish. The resource will be delivered in standard textual format, aligned to the audio source in a multimedia edition. C-ORAL-ROM aims to ensure at the same time a sufficient representation of spontaneous speech variation in each language resource and the comparability among the four resources with respect to a definite set of variation parameters. The multimedia conception of C-ORAL-ROM allows simultaneously alignment and full appreciation of the acoustic information through the speech software WINPITCHCORPUS. The storage of spoken language resources is based on the identification of utterances in the four corpora through perceptively relevant prosodic properties. In C-ORAL-ROM, all the textual information is tagged simultaneously with respect to prosodic parsing and utterance limits. Each prosodic unit corresponding to an utterance is easily and directly aligned to its acoustic counterpart, thus ensuring a natural text - sound correspondence and the definition of a database of possible speech act in the four romance languages.

Detail on the project may also be found on the official web site: http://lablita.dit.unifi.it/coralrom

The present status of speech database in Japan: Development, management and application to speech research

Hisao Kuwabara, Shuich Itahashi, Mikio Yamamoto, Toshiyuki Takezawa, Satoshi Nakamura, Kazuya Takeda

Professor Kuwabara presented a very interesting paper describing the present status of Japanese speech databases. The database project in Japan started in the early 1980s initiated by JEIDA (*Japan Electronic Industry Development Association*). This database initiative aimed at creating a speech database that could evaluate performance of the then existing speech input/output machines and systems. Since then, several database projects have been undertaken, including one initiated by the ATR institute (*Advanced Telecommunication Research*). A survey was conducted on the usage of the presently existing speech databases among industry and university institutions in Japan where speech research is conducted. A short description was presented of four large corpora and sub-corpora. It has been revealed that the ATR's continuous speech database is the most frequently used, followed by the equivalent version of the Acoustical Society of Japan.

SpeechDat across all America: SALA II

Asunción Moreno, Oren Gedge, Henk van den Heuvel, Harald Höge, Sabine Horbach, Patricia Martin, Elisabeth Pinto, Antonio Rincón, Franco Senia, Rafid Sukkar

Prof. Moreno presented this paper which describes a major project following the initial SALA project. SALA II is co-sponsored by several companies that focus on collecting linguistic data dedicated for training speaker independent speech recognizers for mobile/cellular network telephone applications. The goal of the project is to produce SpeechDat-like databases in all the significant languages and dialects spoken across Latin America, US and Canada. Utterances will be recorded directly from calls made from cellular telephones and are composed of read text and answers to specific questions. The goal of the project should be reached within year 2003.

Three new corpora at the Bavarian Archive for Speech Signals - and a step towards distributed web-based recording

Christoph Draxler, Florian Schiel

Dr Schiel reported on some recent activities at the *Bavarian Archive for Speech Signals* (BAS) in Munich. BAS has released three new speech corpora for both industrial and academic use:

a) Hempels Sofa contains recordings of up to 60 seconds of non-scripted telephone speech;

b) ZipTel is a corpus with telephone speech covering postal addresses and telephone numbers from a real world application;

c) RVG-J, an extension of the original Regional Variants of German corpus with juvenile speakers.

All three corpora were transcribed orthographically according to the SpeechDat annotation guidelines using the WWWTranscribe annotation software. Recently, BAS has begun to investigate performing large-scale audio recordings via the web, and RVG-J has become the testbed for this type of recording.

All of these presentations were well accepted and generated lively discussions.

Prof JC Roux
Department of African Languages
University of Stellenbosch
South Africa
Tel.: +27 21 808 3215
Fax: +27 21 808 3975
Email: jcr@sun.ac.za

Review on the session “Speech Variabilities and Multilingual ASR”

Rainer Siemund

All three papers in the session were dealing with acoustic conditions of one sort or another, two of them within the framework of *Automated Speech Recognition* (ASR). In this summary, I shall deviate from the original order in the session and move from pure ASR matters towards user-aspects.

Database adaptation for speech recognition in cross-environmental conditions

Oren Gedge, Shaunie Shammass, Ami Moyal (all NSC - Natural Speech Communication), *Christophe Couvreur* (ScanSoft), *Klaus Linhard* (DaimlerChrysler AG)

The first paper, presented not by any of the authors but by Yaron Himmelhoch of NSC, dealt with methods of adaptation between acoustic environments typical of consumer applications as diverse as mobile phones, handheld computers or television sets. The aim of the study was to find out whether expensive speech data collections could be avoided by adapting the source data to various environmental conditions. A software tool developed in the framework of the EU-funded SPEECON project (http://www.speecon.com) performed two tasks, namely convolution of a clean speech signal with a given room Impulse Response and addition of noise to the convoluted speech signal.

It turned out that adaptation methods involving the addition of noise had a positive effect on recognition rates, reinforced by convolution particularly if far- and medium-distance microphones were used. While the data used for the presented findings was a rather small speaker-dependent sample of speech, further investigations involving speech data from several hundred speakers are under way in SPEECON. For more infos please contact *oreng@nsc.co.il*.

Diagnostic assessment of telephone transmission impact on ASR performance and human-to-human speech quality

Sebastian Möller (Institute of Communication Acoustics, University of Bochum), *Ergina Kavallieratou* (Wire Communications Lab, University of Patras)

The second paper on ASR, presented by Sebastian Möller of Bochum University, addressed the transmission channel impact on human-to-human speech communication quality as well as on ASR performance via landline, cellular and IP-based networks. The dilemma in which transmission network planners find themselves, it was argued, is to find a balance between the subjective human perception of sound quality and the rather objective measurements derived from ASR performance. In general, the findings of the presented study tentatively suggested, codecs operating at low

bit-rates as in mobile telephony appeared to have a lower impact on ASR performance than on human-to-human speech quality. Networks planned to meet human-to-human requirements will therefore usually also satisfy the requirements set by ASR. The performance prediction models tested in the paper will allow network designers to assess a system's usability already rather early in the design phase. Please contact *moeller@.ruhr-uni-bochum.de* for further details of the study.

Does the content of speech influence its perceived sound quality?

Alexander Raake (Institute of Communication Acoustics, University of Bochum)

The third paper, finally, presented by Alexander Raake, also of Bochum University, took a user's perspective on speech quality. Starting out from the assumption that different bandwidths have an effect on the perceived sound quality, the researcher presented a set of French speech data both to listeners who are French native speakers and to listeners without any knowledge of French. The text material presented to the two groups via various auditory channels consisted of *Semantically Unpredictable Sentences* (SUS) and everyday speech. Listeners were then asked to rate the sound quality of the transmitted voice on a one-dimensional rating scale. The French listeners' ratings were found to be lower for SUS, while those of the non-French listeners did not show any major dependency on text material. The reason, it was argued, is that if a given speech sign is understood by the listeners, they are unable to separate form from function and reflect content in their ratings of sound - rather irrespective of the auditory channel. More details and information on new work in the area can be obtained from *raake@ruhr-uni-bochum.de.*

Rainer Siemund
Philips Speech Processing,
Kackertstr. 10
D-52072 Aachen (Germany)
Tel.: +49-(0)241-8871-392
Fax: +49-(0)241-8871-149
Email: rainer.siemund@philips.com

Review on the session "Acquisition of Lexical Information"

Gregory Grefenstette

There is a growing interest in using the World Wide Web as a source for language models. This session at LREC mixed traditional approaches that mine specific corpora for lexical relations with newer techniques that involved using the web as an element of lexical resource building.

The session "Acquisition of lexical information" at LREC 2002 included the following papers:

"Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method"

Pierrette Bouillon, Vincent Claveau, Cécile Fabre, and Pascale Sébillot.

The authors attempt to find patterns revealing semantic relations between words. If found over a training set, these patterns could be used to extract these relations over new text. The authors here restrict themselves to qualia relations between nouns and verbs (e.g. the telic [purpose] qualia relation between "read" and "book", the agentive relation between "write" and "book"). The authors further lessen ambitions by not typing the qualia relation (i.e. as telic, agentive, etc.) they find, but just looking for any qualia relation between nouns and verbs. Unfortunately, removing the type reduces the problem to finding significant noun-verb pairs, a problem already attacked by Hindle and others in the early 1990s by techniques with less theoretical baggage. The authors implement a technique for learning the patterns between identified noun-verb pairs which involves lexical patterns such as those exploited by Marti Hearst in 1992.

"Building Concept Frames based on Text Corpora"

Birte Lönneker

This work describes an interface for manually storing abstract frames along the lines of Minsky (1975) and Mel'cuk (1984), and is more interesting for the effort put into the ergonomics of the input system than for any theoretical insights. The system is designed to be multilingual from the get-go.

"A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping"

Feiyu Xu, Daniela Kurz, Jakub Piskorski, Sven Schmeier.

This work is a mixture of Hearst's lexical-syntactic pattern discovery and information retrieval, all applied to german text. Information retrieval scoring is used to identify the top words in a domain, and the patterns are used to extract relations between these words.

"A Method for Automatically Building and Evaluating Dictionary Resources"

Smaranda Muresan and Judith Klavans.

This ambitious project finds definitions from raw text using a finite-state grammar composed of cue phrases (is called) and text markers (mostly punctuation). It was first developed for formal and layman medical texts. An online demo of this system can be found at *http://www.cs.columbia.edu/~smara/definition_extraction/def_extraction.html* This should be compared to Google's new definition finder *http://labs.google.com/glossary* [which seems to only find text in URLs containing the string "glossary"]. This is a promising area for lexical resource mining.

"Improving an Ontology Refinement Method with Hyponymy Patterns"

Enrique Alfonseca and Suresh Manandhar.

This very interesting work aims at finding the right position in an ontology to place a new word. They first relate the new item to known items using Hearst-like lexical-syntactic patterns (1992). They then use collocations features to describe the potential nodes in the entire ontology and then traverse this hierarchy using the collocation features of the new word to place. This seems to work very well and seems useful for extending ontologies automatically.

"Using Parallel Corpora to enrich Multilingual Lexical Resources"

Dominic Widdows, Beate Dorow and Chiu-Ki Chan.

This is an experiment on using parallel documents to fill a common term-document matrix containing terms from both languages. Then as David Evans and Susan Dumais have done, they reduce this matrix to single space in which bilingual terms which are probable translations are near each.

Gregory Grefenstette
Principal Research Scientist
Clairvoyance Corporation
Pittsburgh, PA, 15232 (USA)
Tel.: 412-621-0570 x137
Fax: 011-33-476-59-3911
Email: g.grefenstette@Clairvoyancecorp.com

Review on the session "Semantic Lexicons"

Anna Braasch and Bolette Pedersen

Presentations in the first session on Semantic Lexicons dealt with three different languages: English, Italian and Danish. All the talks were concerned with encoding and exploitation of semantic information in NLP-oriented lexicons, although based on two different models (viz. FrameNet and SIMPLE). A recurring feature in the presentations was the different use of the encoded information.

The first talk of the session (*Seeing Arguments through Transparent Structures* by *Charles J. Fillmore, Collin F. Baker and Hiroaki Sato*, read by the first author) presented research work exploiting the information available in the FrameNet database with the aim of disambiguating word senses in English. Various processes identifying the lexical heads of phrases that express the core semantic roles of argument-bearing verbs, nouns and adjectives were discussed. Corpus-based generalisations about frame structure and grammatical organisation are derived automatically in order to acquire information about lexical selection and collocation structures. An example of an application is the extraction of KDGs (*Kernel Dependency Graphs*) from a large body of annotated sentences using information in the FrameNet database which facilitates the recognition of selectional and collocational relations between lexical heads, and also the identification of some idiomatic expression types. Predications deeply embedded in a clause or intervening structures pose a barrier to easy access the proper semantic core. In this connection several examples were presented showing a discrepancy between syntactic head and semantic core of the structure, such as in case of transparent nouns ("*eat that kind of fish*") or support verb constructions ("*make a decision*"). The potential for using KDG's in automated abstracting and other NLP applications were sketched out - associating the core arguments with the semantic roles of the frame. The availability of the FrameNet data was also presented, being interesting and useful not only for research into the English language but also as a source of inspiration for NLP-related research into other languages.

The second talk dealt with CLIPS, a project on a multi-level lexicon for Italian (*Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, Antonio Zampolli*, presented by the first author). Firstly, the main characteristics of the underlying PAROLE/SIMPLE model were outlined. The extension of the PAROLE/SIMPLE lexicon is carried out within the framework of the ongoing CLIPS project.

The extension concerns both the quantity of encoded entries and the quality of linguistic information, especially with regard to syntax and semantics. The presentation focused on the correlation between syntax and semantics and stated that in the CLIPS project, a semantic-driven approach to syntactic encoding has proved useful, as syntactic encoding based on (rough) semantic classification showed improved consistency. Another topic was the linking of syntax and semantics, an aspect of the underlying model which needed a thorough revision in order to treat relations between the argument structure of a semantic unit and alternating structures of the syntactic unit. The information on the semantic level was discussed more in detail - the SIMPLE approach based on the extended qualia structure (an idea originating from Pustejovsky's Generative Lexicon). The pros and cons of the implemented multidimensional system were discussed - capturing the complexity of semantic features for the definition of word senses was discussed. On the one hand, qualia roles proved to be very effective in cases of concrete nouns and clearly specified events, on the other hand, they are less appropriate for abstract nouns and underspecified events - precisely because of the vague semantics of these word sense types.

On the same lines as the first talk of this session, the second part of the CLIPS presentation presented the current state of the lexical data and the possibilities for information retrieval from various applicational points of view. The basic assumption is that e.g. nouns, clustered under the same ontological type share a common semantic predicate, which often - although not always - also share the same syntactic realisation type. Once both the semantic and syntactic information is encoded on a large number of entries, queries can be formulated to the database in order to select and retrieve information combinations appropriate for various applications. Like FrameNet, the CLIPS data also allow for retrieval of lexical context of the entry word, showing typical lexical collocates. A few interesting search results were discussed, such as the selection of nouns belonging to the same semantic (sub-)type (e.g. semiotic_artifact) and the typical activities of their production (agentive role: created_by). This way of grouping information together facilitates the creation of semantic networks and also the extraction of domain specific information. The last mentioned process is based on the orthogonal relationships between word senses throughout the entire lexicon, an interesting possible application based on combinatory search. Other relevant perspectives of data querying were mentioned, such as the disambiguation of complex nominals on the basis of qualia relations. In the CLIPS lexicon, the extended use of semantic features for marking predicates' arguments allows the identification and capture of lexical units across the type hierarchy. These examples illustrated the exploitation of the detailed semantic information encoded in the entries from different perspectives, encouraging the work on semantic encoding of multifunctional lexicons.

The third and last talk of the first session (given by *Sanni Nimb*) discussed the treatment of adverbs in semantic lexicons for NLP, the project presented has the aim to extend the Danish SIMPLE lexicon with the semantic description of adverbs. In Danish, lexical semantic information on adverbs is especially important in lexicons for NLP applications (both in analysis and generation), because of their combinatory properties as regards the word order in the sentence and the verb selection (wrt. aspect and tense). Initially, a semantic classification of Danish time adverbs was presented and followed by a method of

investigating their distributional behaviour and interaction with Aktionsart and tense. The semantic types of “point in time” and “duration” were subjects of closer investigation, each of them being placed in the same set of systematically elaborated test sentences. The test showed that there is a strong relationship between the actual sense and the position of the adverb in the sentence, furthermore restrictions on tense/Aktionsart were also observed. In Danish, adverbs carry aspectual features that are relevant in machine translation into languages expressing aspect through the verb form, e.g. French. Consequently, the information on these features has to be provided in the lexical entry of the adverb. The structure of the SIMPLE model lends itself to an extension with a subontology e.g. on time adverbs - and the last part of the talk discussed the exploitation of the inheritance mechanism provided in the ontology for time adverbs. In this connection, a possible expression of synonymy/antonymy relations was discussed. Finally, some illustrative examples were chosen to show the adaptation of the SIMPLE encoding method to the extension of the set of features describing temporal meaning components. The most important observations on selectional restrictions, qualia roles and word order were implemented in the examples showing the first results of the project.

The elaboration and exploitation of wordnets also proved to have a great impact in the field of semantic lexicons for computational use. More than half of the talks in the second session on semantic lexicons were related to wordnets. One of the talks dealt with the elaboration of the German wordnet, GermaNet, whereas two others were related to the exploitation and/or further developments of already existing wordnets for English (Princetown WordNet) and Italian (ItalWordNet). Two of the talks were also related to the now completed SIMPLE project.

The talk on GermaNet (*Claudia Kunze, Lothar Lemnitzer,* presented by the latter) was mostly concerned with the discussion of the representation and standardisation of lexical databases - and wordnets in particular - with the aim of facilitating compatibility and interoperability. It was stated that a current prerequisite for interoperability - both on a monolingual as well as on a multilingual basis - is that they adhere to the XML standard. In addition, interesting perspectives regarding the application of wordnets in “Semantic Web” environments were sketched out in this talk. Since wordnets serve well as an interface between natural languages and ontologies, there is a potential for semantic web designers to apply wordnets as a basic ontological structuring to be eventually expressed in RDF (*Resource Description Framework*).

The talk on ItalWordNet (*Adriana Roventini, Marisa Ulivieri, Nicoletta Calzolari,* presented by the latter) also dealt with interoperability, but from another perspective. Here the challenge is to integrate two semantic lexical resources, ItalWordNet and the Italian SIMPLE Lexicon. A SIMPLE lexicon differs from a wordnet in the sense that, apart from semantic relations, it also includes information types such as argument structure, selectional restrictions as well as links to syntax and morphology (via the PAROLE lexicons). On the other hand, the SIMPLE lexicons are - in their current stage - much smaller than the wordnet resources and thus need to be expanded in order to be practically useful. The experiment of merging the two resources is in its initial phase; however, some considerations regarding differences were reported on: for example, the sense definition strategy in the two projects differs: ItalWordNet is a very fine-grained semantic source whereas SIMPLE establish senses rather on the basis of the formal model and thus results in a more coarse-grained sense distinction strategy. However, since the two lexical resources supplement each other in many respects, a merging of the two is considered fruitful.

The Swedish SIMPLE lexicon was in focus in a talk presented by *Dimitrios Kokkinakis* on behalf of *Jerker Järborg, Maria Toporowska Gronostaj* and himself. This lexical resource and the *Gothenburg Lexical Database* (GLDB), as well as a sense-tagged corpus for Swedish constitute the basis for a semi-automatic construction of new lexical entries with ontological information. Considering the expensive enterprise of establishing semantic lexical resources, this work presents some promising possibilities. Two approaches were presented: an approach where productive compounds - which are typically not in the lexicon - are automatically labelled with the same ontological type as the head of the compound; for instance *kryssningfartyg* is assigned the same ontological type as *fartyg*, namely *Vehicle*. The other approach relies on noun phrases with enumerative nouns, where the unknown noun is tentatively assigned the same ontological type as its sisters in the phrase.

Nabil Hathout presented a paper that also deals with semi-automatic establishment of new lexical resources on the basis of existing ones, in this case on the basis of dictionaries of synonyms (i.e. wordnets). He proposes a language-independent technique to acquire morphological constructional relations from dictionaries of synonyms. Consider the words abandon and abandonment as well as *desert and desertion*, examples of words which pairwise share a graphemic pattern. If *abandon* and *desert* are furthermore encoded as synonyms in a wordnet, then we have two individual factors indicating that *abandonment* most presumably means to *abandon* and *desertion* to *desert*. In other words, the method combines constructional links and synonymy relations in order to make more accurate predictions on the semantics of derived words.

In conclusion, in these two sessions on semantic lexicons, the audience learned a lot about the necessity of systematic and detailed semantic information in the lexicon - in order to be able to turn the material to practical account in NLP applications and language technology products.

Anna Braasch
Bolette Pedersen
Center for Sprogteknologi
Njalsgade 80
DK-2300, Kbhn (Denmark)
Tel.: +45 35 32 90 78
Fax: +45 35 32 90 89
Emails: anna@cst.ku.dk
bolette@cst.ku.dk

Review on the session "Semantic Tagging"

Kiril Simov

All the presented papers discussed a number of more general or more specific semantically oriented tasks, such as word-sense disambiguation problem, animacy recognition, metonymy resolution, flexibility of named-entity determination. Different approaches were proposed for the best performance of the particular tasks: learning methods of different kinds, transfer ideology, expert human intervention. As a whole, all the papers address the issues of real-world text applications, domain-independence, bringing into existence of large-scale language resources and, last but not least - the minimisation of human work. Some of the tasks in hand require substantial and systematic description of real world data instead of artificial theoretical examples. One of the fillings that one received after the session is that more work on the standardisation of semantic tagging is necessary. It is not clear which semantic phenomena could be represented in a corpus, what levels of such annotations are acceptable and what are the relationships between these levels. A very interesting trend of development is the transfer of successful language resources in one language to some other language by using parallel corpora or by building correspondences. The interest in the topics of the session was very high and more than one hundred people attended it.

The first paper "*Learning of word sense disambiguation rules by Co-training, checking co-occurrence of features*" by *Hiroyuki Shinnou* suggests more flexible improvement techniques over the unsupervised learning method, called co-training. Then a promising application to word-sense disambiguation problems is outlined. The experiments show that after overcoming successfully the accuracy limits over the learned rules, the proposed method becomes reliable enough, and the experiment results get much better.

The following four papers deal in different ways with the sparseness problem of semantically annotated data, reusing the existing ones as SemCor and WordNet.

The paper "*Towards a Corpus Annotated for Metonymies: the Case of Location Names*" by *Katja Markert and Malvina Nissim* concentrates on metonymy resolution and describes the treatment of location names in particular. After considering the information insufficiency (especially concerning the problematic cases) of the existing semantic knowledge sources, the authors rely on a data-driven annotation scheme (including golden standard), XML technology, hierarchical organisation of the classifiers, relevant underspecification of complex structures and evaluation refinement.

The paper by *Constantin Orasan and Richard Evans, "Assessing the difficulty of finding people in texts"*, compares several methods (WordNet-based approach and machine learning one) for adequate animacy recognition as a subtask of the anaphora resolution. Although formulated as a language specific survey (for English only), the paper discusses some general for the NLP issues as: plausibility of pure learning methods vs modular approaches with more knowledge resources added (for example, the Named-entity module) and necessity to evaluate a certain subtask with respect to other related tasks, such as anaphora resolution in this particular case. The authors conclude that the separate knowledge modules can be viewed not only as information-holders, but as noise-containers as well, because of the potential errors.

The paper "*Opportunistic Semantic Tagging*" by *Luisa Bentivogli and Emanuele Pianta* proposes an opportunistic way of handling with the sense sparseness problem. The authors suggest reusing an already sense annotated corpus of one language for the semantic tagging of data in another language (in this case - English-Italian). Hence, with awareness of the related problems, a cross-lingual annotation transfer system is pursued. It relies on word level semantic annotation, word alignment strategies and human translation expertise.

Rada F. Mihalcea's paper "*Bootstrapping Large Sense Tagged Corpora*" proposes an algorithm for automatic generation of large semantically tagged corpora, which would repair the sparseness problem by creating them in a fast and reliable way. For the starting point the author relies on SemCor corpus and WordNet, and then a bootstrapping technique is used. The results show that the generated corpus competes the hand-tagged corpora in many respects and, in addition, it minimises the human labour. Unfortunately, this paper was not presented.

The last paper "*How feasible is the reuse of grammars for Named Entity Recognition?*" by *Katerina Pastra, Diana Maynard, Oana Hamza, Hamish Cunningham, and Yorick Wilks* puts forward the question about the transfer/reuse of already existing tools (as some of the others papers do). The authors exemplify this fact by describing how a named-entity grammar can be adapted to a new domain or task. From the three obstacles: rule formalism, application and language, the last proves out to be the most difficult one. Nevertheless, a conclusion is made that the reuse operation is better than creation from scratch, at least when the cost of the human labour is concerned.

Kiril Simov
The BulTreeBank Project
Linguistic Modelling Laboratory -
CLPPI, Bulgarian Academy of Sciences
Acad. G.Bonchev Str. 25A
1113 Sofia (Bulgaria)
Tel.: (+3592) 979 28 25
Fax: (+3592) 70 72 73
Email: kivs@bultreebank.org

Review on the session "Treebanks"

Dan Tufis

The Third International Conference on Language Resources and Evaluation (LREC 2002) was definitely an enjoyable event, both scientifically and socially. The papers, presented in the conference as well in the satellite workshops, were very relevant for the state of the art and the main trends in HLT.

Obviously, given the large topic coverage of LREC 2002, the issues concerned with building, augmenting and use of treebanks could not be absent from the program. In fact, there were two sessions on "TreeBanks" and this note refers to the second one.

The session I chaired included three papers presenting on-going work and recent results in exploiting one of the most used language resource: Penn Treebank. The common thread of the papers included in this section, besides the common treebank, is the aim of making explicit various kind of information both syntactic and semantic in nature.

The first paper, "*Acquiring Compact Lexicalised Grammars from a Cleaner Treebank*" authored by *Julia Hockenmaier* and *Mark Steedman,* from the Division of Informatics of the University of Edinburgh, discusses an algorithm which translates the Penn Treebank into a corpus of *Combinatory Categorial Grammar* (CCG) normal-form derivations. In order to achieve the desired translation, they relied on a preprocessing phase, the side-effect of which was the discovery of a series of inconsistencies and annotation errors. As a result of this preprocessing phase, a cleaner version of the original Penn Treebank was obtained. Although the translation algorithm discussed in this paper does not cover the full range of syntactic phenomena encoded in the Penn Treebank, its variant of binary CCG derivations offers a solid basis for further work towards extending the current annotations with semantic information.

The second paper, "*Identifying Verb Arguments and their Syntactic Function in the Penn Treebank*", by *Alexandra Kinyon* and *Carlos A. Prolo,* from the Department of Computer and Information Science of the University of Pennsylvania, discusses problems related to automatic extraction of a verb lexicon with explicit argument structures and syntactic function of each argument. The new version of the Penn Treebank, known as "release 2" (PTB2), includes additional annotation (the function tags) to expose the sub-categorisation information. However, in order to remain impartial with respect to different syntactic approaches, the encoded linguistic decisions in PTB2 are rather non-committal. Therefore, it is not straightforward to map the PTB syntactic tags to the syntactic functions of a specific syntactic model. The authors argue in favour of their tool that allows for implementing finer-grained rules by which one is able to distinguish verb arguments from verb adjuncts and to differentiate among obligatory and optional arguments. Thus, both the correct identification of the verb frames and the reliable assignment of syntactic functions to the verb arguments are strongly supported. In the context of a grammar extraction task the reported work is expected to be refined and extended with new rules for syntactic function assignment and more importantly with means to deal with unseen sequences of tags.

The third paper of the Treebanks II session, "*From TreeBank to PropBank*" by *Paul Kingsbury* and *Martha Palmer,* from the University of Pennsylvania, addressed the issue of adding semantic information to the Penn Treebank 2. A PropBank (*Proposition Bank*) is a semantically annotated corpus making explicit the predicate-argument structure for verbs, participial modifiers and nominalizations. This paper presented the current status of the Penn PropBank which took into account about one-quarter of PTB2 and concentrated only on the verbal predicates. For each recognized predicate of a clause, depending on the contextual sense, its arguments are labelled in a neutral way (*Arg0* to *Arg5*). The labelling strategy used in the annotation does not attempt to keep the same interpretation for argument names across various senses of a word (as they are potentially described by different predicate structures). For instance, a label *Arg1* in the predicate-argument structures of two semantically different occurrences of the same verb is by no means supposed to have the same interpretation. However, predicates belonging to the same semantic class are supposed to have their argument labels interpreted the same way. To exemplify the adopted methodology, the authors provide several examples of sentences and associated predicate-argument structures. The annotation procedure is supported by detailed and comprehensive examples for different verb's syntactic realisations and the corresponding argument labels. Additionally, based on a frequency analysis, a series of frames are drawn up to describe the expected arguments. The arguments' labels (*Arg0* to *Arg5*) are also given mnemonic names. These names are in general verb-specific, but where the arguments are characteristic to a verb class, they are labelled according to established naming conventions (such as theta-role theory). As one might expect, a verb-frame can be easily extended to cover most part of the verbs in the same semantic class. The authors report on such an experiment with the verbs in class 44 of Levin's classifications and the results are very encouraging. According to their estimation, the 850 verb frames (as of beginning of April) could be easily extended to cover over 1,500 verbs.

Dan Tufis
Director
Institute for Artificial Intelligence,
Romanian Academy
13, "13 Septembrie"
74311, sector 5, Bucuresti (Roumania)
Tel.: +4021 411 29 53
Fax: +4021 410 39 16
Email: tufis@racai.ro

LREC 2002 Closing Session Speeches

Written Language Evaluation and Terminology

Bente Maegaard

In the area of evaluation of written language, 29 presentations were given, compared with 30 at LREC2000, so even if this conference had more presentations in total, this was not the case in the field of evaluation.
The papers showed some trends, of which the most important are highlighted below.
For research, evaluation is becoming an integral and more visible part of any research project. Theories have to be proven, and you need to do this statistically showing the success of your own theory and that the results are superior to other theories. This is a very sound development.
The fact that at almost any computational linguistics conference, all presentations will end by discussing the evaluation methodology and the performance, means that evaluation is no longer a specialised field, but integrated in all fields, and an important part of any researcher's daily thinking.
Similarly, it is of course of vital importance for system developers to be able to follow the progress made in the lab, and to be able to compare with competitors. There is no golden standard yet for evaluation, but the community certainly will be able to tell what counts as a good evaluation methodology, and what does not.
Even resources have to be evaluated: for research as well as for commercial development, the quality, the coverage and the validity of basic language resources, such as dictionaries, grammars, corpora, have to be evaluated.
Machine translation was the first NLP area in which evaluation was applied and where methods were developed. Despite its long history, still no generally accepted methods for the evaluation of MT exists. In his excellent keynote presentation Kishore Papineni, IBM T.J. Watson Research Center, USA, made a new suggestion for the automatic evaluation of MT. Mr. Papineni's point of departure is that evaluation has to be cheap, fast and good. He presented a method to obtain this, and he also compared with human evaluation results (see his article in this issue). Apart from this keynote, LREC 2002 had 4 more presentations on MT evaluation.
The two most popular fields in written language evaluation were the cluster Information Extraction, Information Retrieval and Question Answering (8 papers) and Lexica (7). I believe it is the first time evaluation of resources scores so high. Other areas were evaluation of parsers, grammar checkers, summarisation tools. Finally, we had one paper dealing with evaluation methodology in general, in which the current ISLE results were presented.
The trends that could be seen, apart from what is already mentioned above, follow the message of the keynote: evaluation has to be cheap, fast and objective, - and hence automated. The additional question, taken up by several, was the correlation with human evaluation - similarity to human evaluation being the target.
Terminology is one of the fields of language resources which has a very long tradition. Terminology as a science is of course discussed in separate conferences, such as TKE, but terminology remains an interesting topic for LREC. First of all, a very large part of the vocabulary in business, industry and administration language, is terminology, and consequently terminology is important for all business applications of language technology. This concerns both terminology as a resource (i.e. the terms themselves, their automatic extraction, etc.) and terminology as a part of the vocabulary (NLP treatment in grammars etc.). Out of the 12 terminology presentations (compared with 13 at LREC 2000) 4 concerned term extraction and 8 concerned terminologies and ontologies, i.e. the structure and relationships in terminology.
Terminology is still a "small" field at LREC. But it is important for the fields HLT and terminology to make progress together and to cross-fertilise each other, e.g. concerning methods for acquisition, management and evaluation. At this LREC, the programme committee still felt it was beneficial to treat terminology as a separate field; but maybe at the next conference we will rather be focussing on commonalities between the treatment of general vocabulary and treatment of terminology. If you work in the field of terminology, you may contact me with your opinion about how best to integrate terminology in LREC and how to get more high-quality presentations in this field, - 12 is very low!

Bente Maegaard
Director, professor
Center for Sprogteknologi, Njalsgade 80,
2300 Copenhagen S (The Netherlands)
Tel.: +45 35 32 90 74
Fax: +45 35 32 90 89
Email: bente@cst.dk
Web site: www.cst.dk

Written Language Resources at LREC 2002 in Las Palmas

Nicoletta Calzolari

I have chosen to follow, in this short report, the schema of the corresponding reports for the previous two LRECs, which makes it easier to comparatively assess the main tendencies in the field.

Parameters for Classification

Also this time we received an impressive amount of papers for the *Written Language Resources* (WLR) area, such that often three (sometimes even four) parallel sessions on WLR were necessary. As for Granada and Athens, I use four parameters to broadly classify WLR papers: i) research vs. development, ii) type of resource/tool/etc. described, iii) linguistic description level, iv) language(s). Each has sub-classifications for which the relative order - in terms of number of WLR papers (both oral and poster) - is given. This provides a global quantitative, even though sketchy, overview of the distribution of interest among LREC authors, and a rough idea of the relative weight - as of today - of different aspects related to WLR (yellow cells denote areas with interesting increase, while pink cells denote decrease wrt previous LREC).

Levels of Linguistic Description

Morphology is less and less an interesting topic: it is a consolidated area, where many practical tools/systems exist for many languages. The real interest is in *Syntax* and *Semantics*, with an explosion of papers on *Treebanks*. The advance of *Syntax* means that, after years of theoretical and applied work, it is finally becoming robust enough to build large resources for many languages, almost in a widespread way as morphology. *Semantics* on the other side is still the hot and relatively new - at least with large coverage - topic, crucial for all HLT applications.

Innovation vs. Consolidation

There are quite a number of relatively innovative trends - even though not completely new approaches -, in many cases continuing trends of the previous LREC:

Parameters for Classification	Las Palmas	Athens	Granada
Research vs. Development			
(Innovative) Research	4	3	4
Large Projects	3	2	1
Tool/System Development	1	1	3
Policy Issues	2	4	2
Type of Resource/Tool/etc. described			
Lexicon	2	2	2
Corpus	1	1	1
Methods	6	6	3
Task/Component	3	3	5
System	4	4	4
Infrastructural Aspects	5	5	5
Level of Linguistic Description			
Morphology	3	2	2
Syntax	1	3	1
Semantics	2	1	2
Ontology/Conceptual	4	5	5
Terminology	5	5	4
Other	6	4	6
Languages			
One Language	1	1	1
Many LAnguages	3	3	3
Bi-Multi-Lingual	2	2	2

- *Acquisition techniques* and *machine learning*, also for semantic and multilingual information;
- *Annotation*, also for *Information Extraction*, dealing with coreference, conceptual annotation, named entity recognition, etc.;
- *Semantics with wide coverage*, in lexicons, corpora, tools, systems, mono- and multilingual environment, dealing much more than in the past with multi-word expressions and ontologies;
- *Multilingual aspects*, for resources, tools, applications;
- *Web-based resources* and *tools*;
- *Metadata*, a quite hot topic.

Novelty often lies in moving towards robustness and large-scale, which is crucial in LR and critically involves research aspects. A strong research effort is also given to get new types of LR - self-adaptive, flexible, "dynamic" - to be added to core "static" and manually created LR. This will be the only way to get LR which are adequate, and with good coverage, for HLT applications.

I stress again here that LREC is a conference where it is important to report not only on what is methodologically new, but also on which LR exist, for which languages, in which state of development, and evaluate what is usable in applications. That constitutes its strong industrial relevance, which makes it different from e.g. Coling and ACL.

Consolidation - which goes together with "robustness" - is therefore at least as relevant as innovation. Mature aspects emerged in Las Palmas, in addition to the obvious POS tagging. These are:

- *Standards*, and *open architectures*, more and more felt as a priority;
- *Treebank* and *parsers*, today a must for every language;
- *Semantic lexicons*, finally also with large coverage;
- *Large scale resources*, i.e. lexicons, variously annotated corpora, grammars, for so many languages, but never enough.

Also *integration of lexicon and corpus* is at the basis of many papers, as in previous LRECs, as are descriptions of *large WLR projects*. In this respect the crucial role played by the EC, complemented by national initiatives, in the WLR field, must be again underlined. Without EC or national support many initiatives could not have happened.

Resources and Systems

There was an impressive number of papers describing systems, tools, components, and related resources. The main applicative areas - where again multilingual issues and semantics and "contents" are at stake - are:

- *Question answering*;
- *Summarisation*;
- (*Cross-Lingual*) *information retrieval*;
- *Information extraction*;
- *Machine translation*, with renewed interest;
- *Word sense disambiguation*, important component technology in various applications.

Policy Issues and Infrastructural Initiatives

The importance of infrastructural issues has been clearly recognised in this LREC as critical for a real advancement in HLT. Main topics are:

- *Standards* - either consensually agreed in initiatives such as EAGLES/ISLE or ISO, or de-facto standards, such as (Euro)WordNet, PAROLE/SIMPLE -, with emphasis on *metadata*, and an ISLE panel on standardisation for multilingual lexicons;
- *Multilinguality*, with important aspects of organisational, strategic, political nature;
- *Open architectures and platforms* for LRs, a strategic move towards a new paradigm of co-operative creation of LRs;
- *Minority languages*, with also a panel dedicated to this topic;
- *Large-scale resources*, with challenging organisational issues for international/national co-operation ;

- *Technology transfer*, important for industrial development;
- *Distribution of LRs*, with ELRA and LDC playing a major role;
- *Roadmaps for LRs*, also for specific applications such as QA.

These are obviously the more important issues for international co-operation, which is already going on between Europeans and Americans on a few issues, but should be enlarged to cover e.g. Asian languages. These are also areas for public support, given the infrastructural nature.

Overall Assessment: the field is in a good state

LREC, which is very well consolidated, 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 NLP techniques. Main mature areas are those where:
- *Technology transfer* among languages is possible;
- A *common basic platform* is reached, i.e. a level of uniformity, even repetitions. This happens also through *technology transfer* among languages, very important for the LRs field (e.g. for minority languages);
- *Products* start to emerge.

This is why it is important to have a conference providing an overview of "what exists", not only of what is new. This has always been an important parameter for evaluation of papers for LREC. LREC gives however also a clear feeling of new trends and emerging needs in the R&D community, such as:
- *Acquisition systems*, to overcome the inadequacy of "static" resources;
- *Multilingual resources*, critical for globalisation and world-wide communication;
- *Semantics and conceptual/ontological issues*, to tackle the problems of content interoperability and knowledge management;
- *Semantic-web related aspects*, such as *metadata*;
- *Use of LRs in applications*, where the gap between availability of large-scale and knowledge intensive LRs and systems ability to use them is finally decreasing;
- Importance of *being practical*, even at the expense of theoretical elegance, which shows e.g. in the need for *integration of robust components*.

A final remark is on the importance, emerged a number of times in papers and panels, of a quite new paradigm involving initiatives aiming at *open and distributed infrastructures* for cooperative and controlled creation and maintenance of LRs. This is only possible when the field as a whole has reached a level of stability and maturity. This may become the new "vision" for LRs in the next years.
At last, I want to mention one desiderata for the next LREC, i.e. having *less separation between Written and Spoken sessions*, to start encouraging and pushing towards more interaction and integration between the two big areas and communities. This will be a must for our field to contribute, effectively and globally, to the big challenges of the "knowledge-based society".

Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR
Via Moruzzi 1
56124 Pisa (Italy)
Tel.: +39 050 315 2870 (direct)
Fax: +39 050 315 2834
Email: glottolo@ilc.cnr.it
Web site: www.ilc.cnr.it/

Spoken Language Resources and Tools

Daniel Tapias

Once more, LREC has shown that the area of SLR (*spoken language resources*) is very active the whole world over, not only because of the number of papers included in the conference, 71, but also due to their quality. This fact, could be seen from the different speech sessions, since we saw:
- Papers ranging from reports about initiatives and projects orientated to the development of SLRs for minority languages like Galician and Basque, to industrial consortia focussing on the production of SLRs that cover a large number of languages like SALA II, Orientel, SPEECON and C-ORAL-ROM.
- From speech databases for improving already existing text to speech (TTS) converters, to others which will allow the creation of TTS converters for languages not yet covered by university or industry developments. The databases for Czech and Slovenian and the emotional speech databases collected in the project "Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments" for English, French, Slovenian and Spanish are some examples of the activities in this field.

We also checked the effort that is being carried out in the area of speech to speech translation in projects like NESPOLE, TC-STAR and Tongues, and on emotional and non-native speech databases (for example: Japanese English and European city names), which should allow the development of new and promising technologies and products.
The papers in the area of tools were also very interesting, showing annotation tools like the Multi-Tier annotation proposed in Verbmobil, the paralinguistic annotation for TTS conversion, the annotation of emotional states and several different and interesting proposals for dialogue annotation and modeling. However, despite the important number of initiatives in this area, annotation standards are still an open issue that, from my point of view, should be addressed at an international level.
There were important contributions in the area of automatic speech segmentation as well. In particular, I would mention the one based on statistical correction of context dependent boundary marks and another based on the Forward-Backward algorithm.
It is worth mentioning the effort that is being made by the European Commission (EC) and by the national programs and initiatives for developing new SLRs. In particular, in the Language Resources and Evaluation Panel, organized by Mark Maybury and Antonio Zampolli, the panelists (representing the EC, France, Germany, Italy, Spain and the USA) showed the status and plans for LRs development in their countries. In this sense, there were also several papers describing the status of the speech databases in Japan, the spoken Dutch corpus, the Bavarian Archive for Speech Signals, the large vocabulary speech database for Thai, etc., which shows the importance of this area in many countries. Finally, in the opening ceremony, ELRA announced the availability of a bug report service for reporting bugs found on the SLRs distributed through ELRA.
Therefore, we can conclude that there is a growing interest in LRs and their quality, which mirrors how essential for creating, developing and testing new technologies and products, high quality SLRs are.
There are still many languages for which there are no available SLRs as well as environments, recording conditions, speaking styles, etc., that need to be properly understood to improve the quality of both automatic speech recognisers and text to speech converters. Consequently, new and, in some cases, complex SLRs will have to be collected and annotated in the short and medium term future.
Additionally, as the Telecommunications and the Information Technologies come closer together, the products on offer become more complex and feature-rich, so the need for easy-to-use human-machine interfaces becomes more and more important.

In this scenario, Human Language Technologies will play an increasingly important role since they will be the key actors in facilitating the access to the benefits of the Information Society to everyone, independently of language, education, culture or special needs. Therefore, language resources, which are the foundation for building good quality Human Language Technologies and products, will continue to be a strategic component for addressing the current and coming challenges in the years to come.

Daniel Tapias
Telefonica Moviles
C/Labastida, 11
28034 - Madrid (Spain)
Tel.: +34 680 013 286
Email: tapias_d@tsm.es

Spoken Language Evaluation and Multimodal Communication

Joseph Mariani

In the domain of spoken language system evaluation and multimodal communication, several statistics may be mentioned. The number of papers on spoken language processing and multimodality has increased through the years from 77 at LREC'98 to 86 at LREC'00 and 123 this year, at LREC'02, while the ratio of the papers in these categories compared with the total number of papers is stable at 30%.
The ratio of papers on evaluation has decreased, from 30% in 1998 to 25% in 2000 and 20% in 2002.
Finally, there has been a large increase of the number of papers on multimodality, from 2 (1%) in 1998 to 6 (5%) in 2000 and 40 (15%° in 2002). This shows the growing interest of the language resource and evaluation community for this topic.
Generally speaking, the use of evaluation has been reinforced in the USA, within programs such as TIDES, EARS or Babylon. There has been presentations devoted to specific issues of interest on Rich Transcription Evaluation (RTE), conducted by NIST, and on MT evaluation based on N-grams (BLEU) proposed by IBM and also conducted by NIST. DARPA has proposed international cooperation on those two topics. The activity in this area is therefore very large at NIST, and, accordingly, at LDC which has the task of providing the language resources.
More and more efforts are also going in that direction in Japan, especially within a broadcast news transcription evaluation program.
Meanwhile, the activity is still limited and on a non-permanent basis in the European Union. However, new initiatives may be reported, such as the European Commission supported project TCSTAR-P, which aims at preparing within the last FP5 Call the coming FP6 program. Technology evaluation now appears as a specific component of the large Integrated Projects, which constitute with the Networks of Excellence the new "instruments" within FP6. In France, the TechnoLangue program has been launched which includes a large part of activity on language resources and evaluation. ELRA and ELDA decided to increase their activity in evaluation, and ELDA changed its name to "Evaluation and Language resources Distribution Agency". The Evaling association has also been launched in France, which is typically devoted to language systems evaluation.
In the area of speech technology evaluation, activities and results have been reported at the conference on pronunciation evaluation, especially for proper nouns, and on speaker verification evaluation, especially over telephone. A large activity is devoted to spoken dialog evaluation, aiming at providing methods for measuring performances in understanding and in dialog handling. Results have been reported on various applications, including usability measures from field tests.
The multimodal communication area is a very active field of investigations. It includes multimedia information processing, natural interactivity and multimodal communication systems. There are still very few evaluations conducted in that field, but many tools are now proposed for the acquisition, transcription and annotation of multimodal data. The need to have them being made largely available in the near future has been expressed by many researchers.

LREC 2002 Closing Session

Angel Martin Municio

I am a little worried because of the advice of the Major of the city of Las Palmas during the opening ceremony. You remember he recommended us to make shorter the periodicity of these Congreses. I think he said "up to a meeting each six months". I don't know wether such a kind of recommendation would be accepted or not by the leaders of our association. Nevertheless, I am quite sure we are an association of institutions, organizations and companies belonging to each of the European countries and to different scientific and technological communities. That is the reason why we have special and particular aims, on which we have to talk each day more and more urgently.
In the first and second Congresses, we explored the possibilities and promoted some initiatives for international cooperation concerning what Prof. Zampolli, in the Proceedings of the 1st Conference in Granada, called the reusability of language resources; in those days recently coined to express the idea of large collections of language data as the essential infrastructure for all languages.
Now, from the perspective of all the communications we have had during LREC 2002, we could realize that the present situation is clearly different from that we contemplated six years ago, both at the technical and organizational levels. And also the present situation of ELRA is able to promote our efforts in favour of the so-called European Research Area. I am quite sure that the framework programs of the Commission in all the fields they contain are not enough for the building of Europe, and must be either changed or complemented for some kind of excellence networks in the same way that the economical and monetary areas in the Union are treated.
You know that to improve this situation, it would be necessary to influence on both the political decisions and the scientific culture of our societies. And I wonder if ELRA could take these goals as own, and in which way ELRA could carry out these aims. I think that the political and social diffusion of the French Government Project we have known yesterday could be, perhaps, one of the many aids of ELRA in this way.
Finally, on behalf of the Local Committee of Las Palmas, I would like to thank you for coming to Spain, and have a good return home. And hasta la vista!

General Report on LREC 2002

Khalid Choukri (Khalid Choukri could not be present at LREC 2002 due to personnal reasons)

I would like to apologise for not being with you today. I missed the most useful event of 2002 to which we at ELRA & ELDA devoted most of our efforts in the last few months. I am missing this for a serious but enjoyable personal reason.

As everyone stated during the last few days, LREC has become a major event in Human Language Technologies (HLTs), tackling the most critical issues of LRs and Evaluation. ELRA is very proud to play a role in that.

The challenge of gathering during three days and even 7 with the workshops, the key players in this area every two years, turned out to be a huge contribution of ELRA to the promotion of our field. I will elaborate quickly through some data and facts about LREC.

Some raw data about the participation deserve to be mentioned to give you a more concrete idea: from 500 in Granada, 600 in Athens, now we are/were very proud to welcome over 700 participants (739 registered participants).

Figures on the registered participants show that we had this year about 100 participants from the industry sector, compared to 72 in 2000. This highlights our efforts to attract industrial organisations in addition to purely academic ones.

I would also like to stress how glad we are to offer special packages to our members allowing them to attend LREC at special conditions. This is part of our mission to serve our members and to attract new ones. So in addition to the substantial discount our members get when they purchase the resources, ELRA members can benefit from reduced registration fees.

To further illustrate the undeniable success of the LREC conference, we should also mention the increasing number of submissions, both for the papers and the workshops.

Out of the 460 submitted papers for LREC 2002, 365 have been selected - about 100 more than for LREC 2000. These papers cover many different fields of HLT, and address issues related to e.g. written and spoken resources, multimodal and multimedia, evaluation, and terminology.

The number of workshops which have actually taken place at LREC 2002 is 18, out of the 20 which had been accepted. To compare, 9 workshops had been organised in 2000. The table below illustrates the variety of the topics handled during these workshops and the number of participants for each workshop.

The program committee had a very hard time to select the right papers and workshops.

Our event is really international as well as our activities. We are an association with a European flavor and an international scope and coverage. We enjoy the backing and partnership with a large number of representative organizations which I would like to thank for their involvement in LREC. They should be proud to see the outcome of this involvement.

All continents have been represented for this edition, with 39 countries.

At the opening ceremony, Bente Maegaard told you about LangTech, this new European forum for language technology. This year (and probably the coming ones as well) we will have no exhibition at LREC, because we think that the LangTech event will constitute the right forum for a more market-oriented, “product-commercial”-oriented exhibition. LangTech is taking place in Berlin on 26-27 September , please contact us if you wish to be part of it. Details are available at: www.lang-tech.org.

See you soon and hopefully at LREC 2004 in Lisbon!

	Workshop Title	Participants
W1	International Workshop on Resources and Tools in Field Linguistics	54
W2	OntoLex 2002: Ontologies and Lexical Knowledge Bases	80
W3	Machine Translation Evaluation: Human Evaluators Meet Automated Metrics	39
W4	Annotation Standards for Temporal Information in Natural Language	28
W6	Customizing knowledge in NLP applications	32
W7	Question Answering: Strategy and Resources	54
W8	Language Resources in Translation Work and Research	48
W9	International Standards of Terminology and Language Resources Management	63
W10	Workshop on Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation	57
W12	First International Workshop on UNL, other Interlinguas and their Applications	22
W13	Arabic Language Resources (LR) and Evaluation : Status and Prospects	39
W14	Multimodal Resources and Multimodal Systems Evaluations	52
W15	Portability Issues in Human Language Technologies (HLT)	24
W16	Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data	79
W17	Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research	61
W18	Towards a Roadmap for Multimodal Language Resources and Evaluation	30
W19	Event Modelling for Multilingual Document Linking	18
W20	Beyond PARSEVAL: Towards Improved Evaluation Measures for Parsing Systems	42

New Resources

ELRA-S0121 Turkish Continuous and Isolated Speech Database

This Turkish speech database was produced by the department of Théorie des Circuits et Traitement de Signal at the Faculté Polytechnique de Mons. The corpus was designed to provide read speech data for speech recognition purposes. The database contains 14 hours of speech (1618 words) from 43 Turkish speakers (adults over 18; 22 males, 21 females) from Belgium, Germany and Turkey (Istanbul, Ankara, Malatya), recorded at 32 kHz on DAT by Sennheiser MD-441-U microphone. The speech signal was sampled at 16 kHz and digitised with 16 bits. Each speaker read a predetermined text of 215 sentences and 100 isolated words, in quiet conditions. Parts of the corpus were labelled and segmented phonemically. Phonetic and orthographic transcriptions of sentences and isolated words are provided.

	ELRA Members	Non Members
Price for research use	400 Euro	800 Euro
Price for commercial use	3,000 Euro	6,000 Euro

ELRA-S0122 German SpeechDat-Car

The German SpeechDat-Car database comprises 338 German speakers recorded over the mobile telephone network. The German SpeechDat-Car database was collected and annotated by the Department of Phonetics and Speech Communication of the University of Munich, under a subcontract of Robert Bosch GmbH, Stuttgart. This database is partitioned into 17 DVDs and 1 CD. The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. The signal data format for the in-car mobile platform recordings is 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order); the channels are multiplexed in a single file, with the channel sequence being 0-1-2-3. The format of the fixed platform audio files is 8 kHz, 8 bit alaw encoding. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items:
- 2 voice activation keywords
- 1 sequence of 10 isolated digits
- 7 connected digits : 1 sheet number (4+ digits), 1 spontaneous telephone number (9-11 digits), 3 read telephone numbers, 1 credit card number (16 digits), 1 PIN code (6 digits)
- 3 dates : 1 spontaneous date (e.g. birthday), 1 prompted date, 1 relative or general date expression
- 2 word spotting phrases using an application word (embedded)
- German data phrases
- 4 isolated digits
- 7 spelled words : 1 spontaneous (own forename or surname), 1 spelling of directory city name, 4 real word/name, 1 artificial name for coverage
- 1 money amount
- 1 natural number
- 7 directory assistance names : 1 spontaneous (own forename or surname), 1 city of birth / growing up (spontaneous), 2 most frequent cities, 2 most frequent company/agency, 1 "forename surname"
- 9 phonetically rich sentences
- 2 time phrases : 1 time of day (spontaneous), 1 time phrase (word style)
- 4 phonetically rich words
- 69 application words: 13 mobile phone application words, 22 IVR function keywords, 32 car products keywords, 2 additional common application words
- 2 additional language dependent keywords
- spontaneous sentences

The following age distribution has been obtained: 187 speakers are between 16 and 30, 72 speakers are between 31 and 45, 70 speakers are between 46 and 60, and 9 speakers are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	90,000 Euro	120,000 Euro
Price for commercial use	90,000 Euro	120,000 Euro

AURORA Databases

The AURORA SpeechDat-Car databases are now available at a lower price for academic organisations:
1/ AURORA/CD0003-01: AURORA project database - Subset of SpeechDat-Car - Finnish Databse
2/ AURORA/CD0003-02: AURORA Project Database - Subset of SpeechDat-Car - Spanish database
3/ AURORA/CD0003-03: AURORA Project Database - Subset of SpeechDat-Car - German database
4/ AURORA/CD0003-04: AURORA Project Database - Subset of SpeechDat-Car - Danish database

For research use by academic organisations	200 Euro
For research use by commercial organisations	1,000 Euro

ELRA-S0123 Basque Spoken Corpus, by Jon Aske (Department of Foreign Languages, Salem State College - Salem, Massachusetts, USA)

This is a collection of forty two narratives in the Basque language (Euskara) by native speakers. It includes sound files (MP3 format) and full detailed transcripts. Each of the narratives is a recounting of a short, silent movie that the speaker has just watched to a friend or acquaintance who has not seen the movie (no other person was present in the room, just the recording equipment). Two short silent movies were used to elicit the narratives: Twenty one of the narratives correspond to the 7-minute silent movie The Pear Story (Chafe, ed., 1980) and the other 21 are about a 12 minute collage from Charlie Chaplin's Modern Times. The recordings were made as a part of a study on Basque word order in 1993 (Aske 1997). The transcriptions are made following a modified version of the guidelines given in Edwards and Lampert 1993. The speakers were from different age groups, different dialects, and had differing language abilities. Profiles of the speakers are also included. In addition to the 42 narratives with transcripts, 53 additional sound tracks of extemporaneous speech and description of still images are also included.

	ELRA Members	Non Members
Price for research use	45 Euro	45 Euro
Price for commercial use	45 Euro	45 Euro

ELRA-M0026 MultiWordNet

MultiWordNet is a multilingual lexical database including information about English and Italian words. It is an extension of WordNet 1.6, a lexical database for English developed at the Princeton University. MultiWordNet contains information about the following aspects of the English and Italian lexical:

- Lexical relations between words
- Semantic relations between lexical concepts
- Correspondences between Italian and English lexical concepts
- Semantic fields

The basic lexical relationship in MultiWordNet is synonymy. Groups of synonyms are used to identify lexical concepts, which are also called synsets. Synsets are the most important unit in MultiWordNet. A lot of interesting information is attached to them, such as semantic fields and semantic relationships.

MultiWordNet can be used for a variety of NLP tasks including:

- Information Retrieval: synonymy relations are used for query expansion to improve the recall of IR; cross language correspondences between Italian and English synsets are used for Cross Language Information Retrieval.
- Semantic tagging: MultiWordNet constitutes a large coverage sense inventory which is the basis for semantic tagging, i.e. texts are tagged with synset identifiers.
- Disambiguation: Semantic relationships are used to measure the semantic distance between words, which can be used to disambiguate the meaning of words in texts. Also semantic fields have proved to be very useful for the disambiguation task.
- Ontologies: MultiWordNet can be seen as an ontology to be used for a variety of knowledge-based NLP tasks.
- Terminologies: MultiWordNet constitutes a robust framework supporting the development of specific structured terminologies.

The release 1.1 of MultiWordNet is currently available. It includes information about 51,000 Italian word meanings and 28,000 synsets (in correspondence with the English equivalents). It also includes a labelling of most WordNet 1.6 synsets with semantic field labels.

Work on MultiWordNet is going on. The next release will contain at least 10,000 new word meanings.

Data are contained in a specialized database server, which can be accessed by clients through a socket connection. The database server has been implemented in Lisp under the Unix and Windows environments. An application program interface and graphical browsing interface are provided with the database. A Java implementation of the database is planned for the next release.

	MultiWordNet Database (including semantic fields)		Labelling of WordNet 1.6 with semantic fields	
	ELRA Members	Non Members	ELRA Members	Non Members
Price for research use by an academic institution	350 Euro	500 Euro	Free	Free
Price for evaluation use (3 month license)	500 Euro	1,000 Euro	50 Euro	100 Euro
Price for internal use by a commercial organisation	6,000 Euro	12,000 Euro	600 Euro	1,200 Euro
Price for commercial use	10,000 Euro	20,000 Euro	1,000 Euro	2,000 Euro