The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

July - September 2000

Vol.5 n.3

Special Issue 2nd LREC

CONTENTS

Letter from the President and CEO ____________________________ Page 2

LREC Closing Session Summaries ____________________________ Page 3

LREC Panels Summaries ____________________________________ Page 9

LREC Technical Sessions Summaries __________________________ Page 16

LREC Workshops Summaries _________________________________ Page 18

LREC Opening Session Speeches ______________________________ Page 22

New resources ____ Page 30

Founding of the Global WordNet Association __________________ Page 32

Editor in Chief:
Khalid Choukri

Editor:
Khalid Choukri &
Valérie Mapelli

Layout:
Valérie Mapelli

ISSN: 1026-8200

ELRA/ELDA
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
WWW: http://www.elda.fr

Dear ELRA Members,

Following the tradition we established after the LREC'98 in Granada, this issue of our newsletter reports on the second LREC (International Conference on Language Resources and Evaluation), held in Athens from 31 May to 2 June 2000.

The second issue of the LREC series was organized by ELRA with the support of the major organizations involved in Human Language Technologies. The local organization was trusted to ILSP (Institute for Language & Speech Processing) and the National Technical University of Athens.

According to comments from participants, LREC conferences are now a major event in our field. With over 500 participants in 1998 and over 600 in 2000, we hope that LREC constitutes a significant milestone in the life of Human Language Technologies.

At this second issue we offered the opportunity to some key players (both industry and academic research centers) to show their latest products and services, through an exhibition that lasted four days. Standing by the tradition established with LREC98, we were able to accommodate ten satellite workshops. Due to time constraints, most of these workshops lasted half a day but gave the participants fruitful forums for discussions. We hope to continue such action in the future as part of our contribution to the development of the field.

Some general statistics to illustrate the LREC'2000 success : over 280 papers (129 oral versus 152 posters), over 600 participants from 47 countries and all the continents.

In order to give you (remind you) an idea of what happened in Athens, this newsletter is structured into three parts; We start with the closing session as the first part of the newsletter. It consists of general overviews drawn up by the program committee at the closing ceremony. As was done at the 1st LREC, these overviews focus on Spoken Language Resources (H. Höge), Written Language Resources (N. Calzolari), Evaluation in the spoken area (J. Mariani), Evaluation on the written area (B. Maegaard), the general aspects of LREC (K. Choukri), and some concluding remarks from the chairman of the conference (A. Zampolli) and the chairman of the local organizing committee (G. Carayannis).

The second part includes short summaries of some technical sessions and panels, reported by the chairpersons or the panels organizers. This is an attempt to highlight some major topics of the conference. This part also includes summaries of some of the satellite workshops.

The last part is devoted to the important speeches given by some key political guests and supporters during the opening session. Their messages addressed the crucial issues of Language Resources, Multilinguality and Human Language Technologies as key elements for the growth of today's economy. Their contributions and statements reflect the importance of the HLT field that goes beyond the simple economy and business competitiveness, with important social and cultural impacts.

In his speech and referring to his introductory speech at LREC’98, A. Zampolli, President of ELRA and chairman of the conference, drew a picture of the Human Language Technologies field stating that although the general framework still hold, efforts are devoted to some key topics identified in Granada but there is still a large number of areas which did not receive the attention they deserve.

We would like to express our warm thanks to all members of local organizing committee for the valuable work they did. Special thanks to G. Carayannis and E. Fotinea.

Last but not least, at ELRA/ELDA we continued to carry out our regular activities during the last quarter. We entered into agreements with a number of Language Resource providers and we added new resources to our catalogue. These are described in this issue.

ELRA/ELDA have been extending their activities and new positions are open. See below for more details or on http://www.elda.fr.

Antonio Zampolli, President　　　　Khalid Choukri, CEO

Open positions at ELDA

ELRA is expanding its activities in the Human Language Technology area. Positions for technical engineers with background in speech processing, terminology management, marketing and market analysis are available.

The ideal candidates would have:

- Experience in the field of language engineering/ computational linguistics/ speech processing or a related field;
- Experience in developing market studies and writing technical reports in the field of information technologies or Human Language Technologies;
- Knowledge of /Experience in integrating and implementing language resources in new applications;
- Experience with technology transfer projects, industrial projects, collaborative projects within the European Commission or other international frameworks;
- Experience /motivations in establishing and supervising external contracts;
- Citizenship of (or residency papers) a European Union country;
- The ability to work in at least two European languages (English being essential).

Positions are based in Paris.

Applicants should E-mail, Fax, or post a cover letter addressing the points listed above, together with a Curriculum Vitae, to:

Khalid CHOUKRI, ELRA / ELDA, 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.

LREC Closing Session Summaries

Summary on Spoken Language Resources

Harald Höge, Siemens AG, Germany

1. SLR for Commercial Use

The Databases of the SpeechDat-Family is a success story:

- The SpeechDat databases will cover soon all languages in East & West Europe and South America.

- SpeechDat has been extended to a new continent: Australia.

- Extensions to 'small' languages and dialectal areas have been presented: Welsh, Hebrew, Slovenian, Catalan, Austrian.

- Extensions to new application areas have been presented: Car (SpeechDat_Car Project), Consumer Devices (SPEECON Project).

- Open issue: availability of databases with 'accent' speech.

Providers of speech driven services record 'tons' of speech data within running applications. Providers showed willingness to share these data.

These data are characterized by:

- Describe real behavior of users
- Data is not annotated
- **Open issue:** what to do with these data

2. Basic SLR and Tools

Basic speech databases have been presented for:

- Studying dialog phenomena
- Studying multimodal issues
- Making research on content processing

For these issues a new family of SLR evolves: The Broad Cast News (BCN) databases will become a new SLR for many languages. An open issue are standards.

Within some national project SLR is produced in a large scale. These actions are uncoordinated (open issue).

Tools are further developed to:

- Record SLR
- Annotate SLR
- Validate SLR

Open issue are standards.

Production & Research on pronunciation lexica is an ongoing activity. Open issue are standards and validation procedures for pronunciation lexica.

3. New Type of SLR

SLR for speech synthesis (TTS): corpus based speech synthesis needs a new type of annotated and segmented speech databases. First databases were presented. Automated speech segmentation tools are still not good enough.

SLR for speech dialogues: these SLR are dialog-atoms (also called speech objects, dialogue modules) which describe a semantic concept with the properties:

- Embedded in a dialog act
- Comprises grammars, prompts, dialog strategies and language models

Examples of such Dialog-Atoms are dialog acts to ask for a money amount, for names for time information. Open issues are:

- Theoretic background
- Transfer to other languages
- Standards

4. Production of 'Good' SLR

There are two basic approaches to produce good SLR:

- A quality stamp is attached on a SLR. This means the SLR is checked against a list of specification criteria. For this approach a first proposal has been made by SPEX. This approach has to be applied.

- Know how has to be increased in order to improve the design (specification) of SLR optimal suited for building speech systems. This approach has started in the Cost action 249 where recognition results from different SpeechDat databases are derived. Open issue: experiments on existing databases should lead to conclusion for optimal design of SLR.

5. Conclusion

Following conclusions can be made:

- SLR is a fast growing field with global dimension
- New types of SLR are coming
- Theoretic background for optimal design of SLR has to be developed
- A new effort in setting standards has to be made

Harald HÖGE
Siemens AG, ZT IK 5,
81730 München, Germany
Email: harald.h.hoege@mchp.siemens.de

Spoken Language Evaluation at LREC-2000

Joseph Mariani, LIMSI-CNRS, France

A general overlook on the conference shows that the number of papers on speech has increased from 77 at LREC'98 to 93, following the general increase of papers at LREC. Overall, the ratio of speech papers remains the same (about 1/3), and the ratio of speech papers addressing evaluation also remains the same (about 1/3 of speech papers).

The evaluation paradigm has been used by Darpa in the US since 1984 to monitor their program. It implies an infrastructure which is mostly provided by NIST, for protocols, and LDC, for data. The participation of non-US laboratories in speech related evaluation campaigns started in 1992. Speech systems evaluation is now extended more generally to dialog or language systems evaluation (including speech), in programs such as Communicator or TIDES, where the corpus-based evaluation approach slightly moves to the evaluation of modules using the same architecture, among contractors and affiliates. Compared with this long term effort, the European activities on that topic are more limited and of a non-permanent nature, at the EC level (such as Sqale), or at the level of a country (Verbmobil in Germany) or of a group of countries (such as the AUF Francophone actions). There is no permanent infrastructure yet, but possible cross-project evaluations will be considered, among other clustering activities, within the IST-HLT Class project. Activities are now also developing in Japan, on a Broadcast News type of task.

Many discussions are being pursued on the relationship between technology evaluation and usage evaluation. It should be stressed that a good technology is necessary, although not sufficient, to develop a good application. Therefore, the link between technology evaluation, usage evaluation and basic research should be established and maintained. It should also be stressed that evaluation of language processing systems is of course of interest for customers, but even more of interest as a feedback for the technology developers. Usage evaluation results were especially depicted in two papers: one comparing two speech dictation systems, considering criteria such as functionality, usability and maintainability, and another one on evaluation of speech input in car applications, studying the recognition rates in relation with the mental workload, and learnability of the system.

In their paper, NIST discussed design issues in speaker recognition evaluation, considering one or two speaker detection, and applying speaker recognition to speaker tracking and speaker segmentation. Another paper addressed the problem of cross-lingual interpolation of speech recognition models. The evaluation of text-to-speech systems is still a very active area, where evaluation at the perceptual level of speech compression or of voice quality (for synthesizing the enclitic stress), of methods, such as the Analysis-Modification-Synthesis one, or for encoding units, such as dissylables, are studied. Prosody is a topic of special attention, and guidelines for end-to-end TTS system evaluation in Japanese have been proposed.

But spoken dialog evaluation is probably the most exciting research topic nowadays. It is a very difficult problem, which makes it still an open research issue, especially if we consider comparative evaluation, as it deals with evaluation of interactive systems, and requires the availability of large transcribed dialog corpus.

The EC DISC project has issued Best Practice methodologies to design a spoken language dialog system. Prosody in on-line evaluation of spoken language dialog systems has also been studied.

In the Paradise paradigm, M. Walker and coll. at AT&T Labs propose dialog systems evaluation methods which aim at predicting user satisfaction as a weighted integration of objective measures in dialog performances, such as the recognition rate, the understanding rate, the number of turns etc. It is shown that this measure seems to be relatively independent from the task and from the subjects, and this approach is used in the US Darpa Communicator program. Other researchers propose a different approach, the DCR (Declaration-Control-Reference) paradigm, to assess the ability of the system to deal with various dialog phenomena.

Several dialog systems evaluation results were reported, such as the influence of dialog prompting on the dialog performances, and various criteria and metrics are proposed for the evaluation of dialog components such as the number of turns, the turn duration, the dialog duration, the correction rate, the recognition time, the word accuracy, the implicit recovery, the sentence understanding measure, but also the user frustration or the information bit rate.

As previously mentioned, speech goes more and more together with natural language processing, and complete systems are evaluated, as in speech-to-speech translation systems evaluation. Two papers report experiments in this area, for English-to-German and Japanese-to-German speech translation. A Graphical Evaluation Tool is described, and ways of evaluating the systems, either accuracy-based or task-based, are mentioned, including quality-fidelity measures, or goal completion. For building translation systems, the use of comparable corpora instead of parallel corpora, which may be obtained in much larger quantities, is an alternative which is presently considered. Another area where speech and language are considered together is multilingual Topic Detection and Tracking (TDT), where results are reported.

But speech is also used within multimodal communication systems, and the way to evaluate multimodal groupware systems, or multimodal meeting recognition and tracking, involving speech, NL and vision, starts being considered.

The importance of the availability of IPR-clear language resources for language system evaluation is often mentioned. The production of resources which will be used for evaluation often induces that the corresponding data is of quality, as it should be delivered in due time, and it should be in agreement with the specifications. The corresponding data can be made available after the campaign for other laboratories to compare their system with the state of the art, or to measure progress since the evaluation campaign. Operational systems, such as the ones which are used for telephone applications, allow for producing huge amount of data, even larger than what a laboratory can deal with.

The evaluation of the speech resource quality is itself a topic of interest. Spex conducted validation and improvement of spoken language resources for the Speechdat project, or for ELRA. The quality of an Italian Broadcast News corpus has been assessed, and, generally speaking, the quality of the corpus used for evaluation is ensured by the evaluation participants themselves, as they won't easily accept that the test data contain errors which may be considered as caused by their system.

The need for corpus annotation and design tools appears clearly, including annotation convention commonly admitted and shared. The availability of transcription tools, such as the Transcriber software jointly developed and distributed by the LDC and the DGA in France, or the MATE workbench, creates de facto standards. The need for tools helping in the design of Wizard of Oz environments, and for speech recognition tools for processing speech data, is also clearly identified. A dictation free software is made available by the Information Technology Promotion Agency (IPA) in Japan, and the COST 249 in Europe distributes a reference recognizer to provide reference recognition results for multilingual applications.

Finally, the need for international cooperation in Human Language systems evaluation has been stressed at many occasions. Science and core technology are international, and should therefore be evaluated at an international level, while HLT evaluation has to be multilingual, and therefore implies a very large effort which should preferably be shared. Having an international multilingual evaluation action would optimize the participation of laboratories. A good candidate for a common task would be to address Broadcast News on Demand, which includes speech transcription in various conditions, speaker recognition, Named Entity extraction, Topic detection and tracking, crosslingual multimedia information retrieval and translation, and dispose of very large amounts of data.

In conclusion, the evaluation paradigm appears to be of most importance to accompany research in Language Technology development. Speech and natural language processing methods are now intimately merged in actual systems, and both technology and usage evaluation should be considered. Installing an international evaluation infrastructure will be a challenge for the coming years.

Joseph Mariani
LIMSI-CNRS,
BP 133, 91403 Orsay Cedex (France),
Email: mariani@limsi.fr

Written Language Resources at LREC 2000

Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR, Italy

Parameters for Classification

The first remark for the Written Linguistic Resources (WLR) area is the impressive amount of papers (almost always two parallel sessions on WLR were necessary) and the variety of topics.
As for Granada, 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.

Levels of Linguistic Description

The real surprise is the explosion of papers dealing with *Semantics*, even more than on Morphology. Less attention is comparatively paid to Syntax. A reason could be that Semantics on one side is the hot and relatively new - at least with large coverage - topic, crucial for HLT applications, Morphology on the other is the well consolidated level where many practical tools/systems appear for many languages, while Syntax is neither hot or new, nor yet - despite years of both theoretical and applied work - robust enough to be a product, at least in a widespread way as morphology.

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

Innovation vs. Consolidation

There are quite a number of relatively innovative trends - even though not completely new approaches -, also with respect to the previous LREC:

- *acquisition techniques and machine learning*, also for semantic and multilingual information;
- *annotation 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;
- *multilingual aspects*, for resources, tools, applications;
- *Web-based resources and tools*.

Novelty sometimes lies more in moving from toy systems towards robustness and large-scale. This is crucial in LR, involving - contrary to what is felt - research and innovation. LR is not just a sector where mostly compilative, repetitive work is involved, but it requires a strong research effort to get new types of LR - self-adaptive, flexible, robust - critically needed by HLT applications.
LREC is however a conference where it is important to hear not only what is methodologically new, but also what exists, for which languages, in which state of development, and evaluate what is usable in applications. That constitutes its strong industrial relevance.
Here consolidation is at least as relevant as innovation. "Mature" aspects emerged in Athens are:

- *tagging*, described for about 20 languages;
- *treebanks*, recently a must for every language;
- *large scale resources*, i.e. lexicons, variously annotated corpora, grammars;
- *standards*, such as XML, EAGLES, TEI, CES, open architectures, rightly felt as a priority.

An important feature is that, both for Lexicons and Corpora, *large-coverage applies* - contrary to Granada - also to semantic and multilingual LR, no longer considered at an experimental level.
Also *integration of Lexicon and Corpus* is at the basis of many papers, as already in Granada, as are descriptions of large *WLR projects*. In this respect the *crucial role played by the EC, recently 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 is an impressive number of papers on development of systems, tools, components, and related resources. The main applicative areas - where again multilingual issues and semantics are at stake - are:

- *Cross-lingual Information Retrieval*;
- *Information Extraction*;
- *Machine Translation*, with renewed interest;
- *Word Processing*;
- *Word Sense Disambiguation*, important component technology in various applications.

Policy Issues and Infrastructural Initiatives

Main issues of infrastructural nature, recognised as critical for a real advancement in HLT, are:

- *standards*, or de-facto sort of standards emerging from large resources built for many languages, as (Euro)WordNet, PAROLE/SIMPLE;
- *multilinguality*, not only a technical issue, but presenting aspects of organisational, strategic, political nature;
- *open architectures* for LR, to allow reusability of available LR;
- *minority languages*, urgent issue both in Europe and world-wide;
- *large-scale resources*, presenting both technical and strategic challenges;
- *distribution of LR*, for which exemplary are ELRA and LDC.

These are the more important issues for international cooperation, where national or EC support is critical. For some, concrete American/European cooperation already officially started with cooperative EU-US projects, such as ISLE/EAGLES for standards and Network-DC for distribution.

Overall Assessment: the field is in a good state

LREC itself seems very well consolidated at only its second round. It provides an important perspective of the *level of maturity* of the field, in those areas where:

- 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 LR 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 must be 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. This year we notice:

- *acquisition systems*, because it is evident that "static" resources are insufficient;
- *multilingual resources*, because of globalisation and world-wide communication;
- *semantics and conceptual aspects*, because of the criticality of content management;
- *web related aspects*.

At last I just touch two aspects probably not yet reflected enough in this LREC:

- *industrial requirements*, to feed future activity. Many companies' representative were present, a very important feature of LREC with respect to other conferences (e.g. Coling, ACL), but more as observers than with an active role;
- *use of existing resources in applications*. Here a question can be asked: is there a gap between available resources and systems' ability to use them? It is true that in general we don't have yet enough resources to cover application needs, but sometimes it seems that there are resources with more information than what systems have the ability to exploit.

We could consider these remarks for the next LREC. We should find a way to have an even more active industrial involvement, and more interaction between the communities of researchers and industrials.

Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR Pisa, Italy
Email: glottolo@ilc.pi.cnr.it

Terminology and Written Language Evaluation

Bente Maegaard, Center for Sprogteknologi, Denmark

Terminology is one of the fields of language resources which has a long tradition. LREC-2000 featured the integration between terminology work and natural language processing in a series of presentations. The excellent keynote speech by Klaus-Dirk Schmitz and Alan Melby focussed on terminology standards and described how standards support the terminology community. The other 13 presentations fell in four main classes, on standard work, term extraction, ontologies and summarisation.

In the area of written evaluation 30 presentations were given. MT 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. The session on MT evaluation showed various interesting ways of approaching the problem, complementing the workshop preceding the conference, specifically devoted to MT evaluation. It is to be hoped that some of this work will find its way into the general methodology currently being developed in the ISLE project.

Other sessions covered the evaluation of tools, grammar and system evaluation, and evaluation and semantics. Most of the presentations obviously took their point of departure in the evaluation of a specific project, but then took the discussion to more general methodological issues. It is very encouraging to see how seriously evaluation is being taking during development, i.e. evaluation is being integrated into the development cycle and much attention is paid to establishing the right way of performing testing and evaluation. The presentations showed that as research fields evaluation as well as NLP are becoming mature.

The session on Information Retrieval and Question Answering systems was concentrated on the type of evaluation campaign which originated in the United States and now is spreading to other continents. Indeed, most of the presentations were provided by participants in or organisers of the American campaigns. It is interesting to follow the evolution of these campaigns, and their spread in popularity: e.g. Europeans have for some time been participating in the American campaigns.

Turning the attention to the themes in evaluation that came up during the conference, there was a discussion of the methodologies to be used for evaluation of products, i.e. evaluation for end users versus evaluation during development. In the evaluation of products, many themes are relevant which are of a non-linguistic, and non-technical nature, e.g. ergonomics. The methodologies for evaluation of the linguistic and technical characteristics of a system will probably be the same for end-user evaluation and development evaluation, and the important contribution of end-user evaluation is therefore to set the priorities so as to make sure that those functionalities are given priority which are relevant for the users.

Another theme which came up several times during the presentations was the question of fully automatic versus semi-automatic evaluation. This question is closely related to the problem of metrics, i.e. the definition of the measurements which can be reliably made. Automatic evaluation is always to be preferred as it gives the possibility of large testing materials, and of a fully objective evaluation. However, in cases where it is very difficult to find an automatic method for the exact measurement, a semi-automatic method may be a good solution.

When setting up an evaluation of a system or a set of systems, three main points have to be considered:

1) Set the goal (the purpose of the evaluation),

2) Define the functionality you want to obtain,

3) Define the metrics.

As a summary of these sessions on evaluation, we can conclude that

- *NLP is becoming mature*, this is the reason that evaluation of NLP is developing
- *Evaluation as a science is becoming mature*, there is an understanding of the issues in defining a reliable evaluation, and many good contributions
- *Standards for evaluation are emerging*, but more research is still needed and a consolidation may only be reachable in a few years' time, not immediately.

So, we are looking forward to seeing the progress at the next LREC!

Bente Maegaard
Center for Sprogteknologi, Njalsgade 80,
2300 Copenhagen S
Denmark
Email: bente@cst.ku.dk

Industrial aspects of LREC2000
Connecting industry players with academic partners

Khalid Choukri, ELRA/ELDA, France

ELRA has been willing and working towards bridging the gap between industry and academia in the HLT area. In establishing a major conference such as LREC, which addresses specific issues on LRs and Evaluation, ELRA contributes to revitalize the field. This is done through the organization of the conference, its satellite pre- and post-workshops, and an exhibition. A number of key players participated to the exhibition organised in parallel to the conference, in order to demonstrate recent advances in HLT. The exhibitors were:

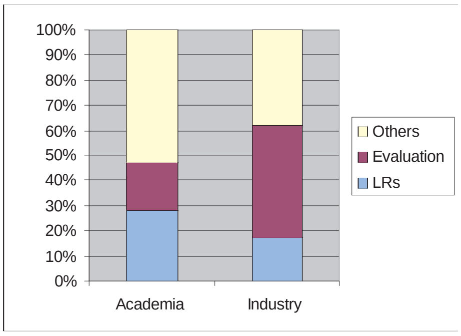
- ELRA
- ILSP
- Lernout & Hauspie Speech Products
- Dialogos Speech Communications and Nuance Communications
- LexiQuest Inc.
- Knowledge S.A
- EXODUS S.A
- WCL-University of Patras
- GENER-X
- Athens Technology Center S.A
- ITACA

This second LREC Conference was attended by more than 600 participants, including 124 from organisations that are members of ELRA, and 98 students. A substantial part of all participants belongs to academic institutions (over 510). 72 participants came from industrial institutions. The most represented countries were: Greece (117 participants), USA (70), France (59), Germany (45), UK (43), Japan (35) and Italy (29).

As already established in Granada, a set of workshops were organised as satellite events to the LREC Conference. These workshops and the number of participants are given below.

If we consider the number of papers presented during the conference, most of them are from academic institutions (247), whereas 29 papers were presented by speakers from industry. These papers addressed two main subjects: Language Resources (LR) and Evaluation. 74 papers presented work carried out in the field of LRs (69 in an academic context, 5 in industry). 61 papers reported on different projects and work conducted in the field of Evaluation (out of these 61 papers, 48 came from academia and 13 from industry). 141 papers dealt with other subjects of HLT (130 from academia and 11 from industry). These are illustrated on the following histogram:

WS	Nb
WS 1: From Spoken Dialogue to Full Natural Interactive Dialogue. Theory, Empirical Analysis and Evaluation	53
WS 2. Very Large Telephone Speech Databases	25
WS 3. Meta-Descriptions and Annotation Schemas for Multimodal/Multimedia Language Resources	67
WS 4. Terminology Resources and Computation	40
WS 5. Workshop on the Evaluation of Machine Translation	37
WS 6. Information Extraction Meets Corpus Linguistics	74
WS 7. Language Resources and Tools in Educational Applications	23
WS11. Using Evaluation within HLT Programs: Results and Trends	46
WS8. Data Architectures and Software Support for Large Corpora DATA: Towards an American National Corpus	61
WS 9. Developing Language Resources for Minority Languages. Reusability and Strategic Priorities.	38



If we consider ELRA distribution activities from the beginning of 2000, ELRA distributed 64 resources for R&D and 56 for industry. Despite this balance regarding sales number, industry represents about 96% of the ELRA revenues compared to 4% for research. Such results show that it is still our challenge to attract more participants and more submissions from industry to the next LREC, regarding their key position on the LR market.

It also appears that industry and research do not request the same type of LRs. Industry is more interested in speech databases, whereas research rather needs multimodal - multimedia corpus, as well as dialog corpus. Being aware of these different needs, ELRA is commissioning the production of new resources according to a preference list established with the results of users' surveys that have been conducted by ELRA (see lists below). ELRA will pursue its efforts to commission new LRs in order to meet the needs of both industrial and academic LR users.

Khalid Choukri
ELRA/ELDA
55-57, rue Brillat Savarin
75015 Paris
France
Email: choukri@elda.fr

Preference lists	ELRA commissioned LRs
• SpeechDat-like database (a language or/and an application area not yet covered within the SpeechDat family - 1000 to 5000 speakers), • Speech database for embedded systems (basically 16kHz sampling, noisy environment, 500 to 1000 speakers), • Pronunciation lexica (for speech recognition and speech synthesis, including extent of proper names), • Dialog corpus, • Enrichment of existing SLRs within the ELRA catalogue, • Multilingual speech synthesis database, • Large monolingual corpora, • Parallel texts, • Bi/multilingual computational lexica, • Multimedia corpus, • Multimodal corpus.	• *Corpus of written Business English* (Ruslan Mitkov; University of Wolverhampton) • *Sets of bilingual LR dictionaries for English and Russian* (Vera Semenova-Fluhr; SCIPER) • *Crater 2 - Expanding Resources for Terminology Extraction* (Tony McEnery; Lancaster University) • *Italian Broadcast News Corpus* (Marcello Federico; ITC-IRST) • *Pronunciation lexicon of British English place-names, surnames and first names* (Marc Fryd; Université de Poitiers) • *Scientific Corpus of Modern French* (Béatrice Daille and Geoffrey Williams; Université de Nantes) • *German-French Parallel Corpus of 30 Million words* (Wolfgang Teubert; Institut für deutsche Sprache, University of Mannheim • *Columbian Spanish SpeechDat-like* (Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya)

Closing Session Remarks

Antonio Zampolli, Istituto di Linguistica Computazionale del CNR, Italy

It is now the moment of closing the Conference. We are all very tired, in some cases even exhausted, and I will be very brief.

Tonight we will have the occasion to thank everyone who worked on this Conference. In the meantime, my colleagues of the Organising Committee have offered a first survey of the outcome of the major thematic sessions of the Conference.

Their remarks and conclusions, and the summaries of the various Panels and possibly of the Workshops too, will certainly appear in one of the next ELRA Newsletter issue.

My Colleagues, whom I want to heartily thank for the intelligent smooth co-operative work in the Program Committee, have already summarised highlights and trends in the respective areas.

It seems to me very significant that so many common points and issues have emerged. Due to time constraints, I can indicate only some examples: the requirement of a continuous effort for developing standards; the recognition of the impact of the WEB on LR and HLT; the numerous consequences, in terms of strategy, policy choices, organisational problems of the infra-structural role of LR in the ICT-based Society; the need of international co-operation to comply with this role and with the requirement of globalisation and multilinguality; the need of co-ordination between national efforts and activities supported by international Funding Agencies; the recognition, in the different areas, of various signs of progress towards maturity of HLT (for ex. the programmatic inclusion of evaluation in the development cycle of NLP systems and in the design and production of spoken and written LR); the need of developing acquisition methods to implement LR flexibly adaptable to specific domains and applications ("corpora and lexica go together"); the trend toward combining rule based technologies with corpus based, data driven approaches both in speech and in NLP; the combination and integration of speech and NLP in building badly needed complete/complex systems.

This last issue provides a clear example of integration between R and D and evaluation, speech and NLP, empirical and rule-based methods, etc.; the promotion of this integration seems to me, as I pointed out in the opening, a characterising feature of LREC.

Speech technology, which has been very successful in recognition tasks, should today cope with applications like dialogue, speech to speech translation, broadcast news on demand, voice interface with WEB, etc., which clearly requires the recognition of linguistic structures and the processing and use of semantic and pragmatic knowledge, which have been until now the realm of NLP.

All that requires (and we already have interesting examples provided by papers presented here) not only the conception, design and construction of new types of LR, optimally suited for building speech systems, but also the creation of standards in new areas, innovative annotation methods, grammar processing, dealing extensively with "meaning", methods to flexibly adapt grammars, lexica, dialogue management rules to domains and sub-languages, acquisition and "discovery" procedures, even for semantics, experiments for transferring models and technologies between languages, etc.

We can already observe the effect of this integration of speech and NLP on the evaluation paradigm, in efforts, reported in this conference, to combine two different traditionally distinct approaches, namely user-oriented and developer-oriented.

I hope that in this way we will also acquire more evidence to answer the vexed question of the reason for the gap between available LR and industrial systems: scarcity of LR, in particular for some languages, inadequacy for the intended applications of the information provided by available LR, the inability of industrial systems to exploit the information provided, etc..

In any case, we will do our best to prompt companies to take a more active role in the next LREC: i.e. not only watching what is already available, but actively co-operating or driving the realisation of new types of LR. For example, multimodal LR, including more than written and spoken language, will probably have a more prominent role in the next LREC

I hope that the results of this Conference will contribute to the advance of the state of the art in our field of HLT and, in general, to the improvement of the Society where we live, in which information, communication, multilinguality play an increasingly central role.

The outcome of the Conference has reinforced my belief that our work is an essential part of the effort of the R&D community to open the possibility of a democratic access to an increasingly large part of the citizens of our world.

This awareness has motivated the Programme Committee to continue its efforts. We ensure you that we will organise a third LREC, possibly in an historical attractive place, and we hope that your presence will contribute to its success. We will announce the venue very soon.

Let us also hope that the many voices who have witnessed the relevance of LR and evaluation will persuade the public Authorities to ensure an adequate support to our field.

You are all invited to the LREC gala dinner which will take place tonight at the Hilton Hotel at 21.00.

Many people still remember - I hope with pleasure - the "El Relicario" happening in the Alhambra Gardens in Granada. We will certainly do our best to conclude our stay in Athens with songs and dances.

After all, Greece is not only the land of Apollo and Minerva, but also of Bacchus and Venus!

Seeing you tonight at the dinner and thanks again to all of you for your participation.

Antonio Zampolli
Istituto di Linguistica Computazionale
del CNR,
Via della Faggiola 32
56100 Pisa, Italy
Email: pisa@ilc.pi.cnr.it

LREC-2000 Concluding Remarks

George Carayannis, Institute for Language and Speech Processing, Greece

The success of the Second LREC Conference is the sign of an explosive activity in R&D and system development. As a result, LRs appear to be more and more necessary for core technology development and as new methods and techniques (numerical and statistical, machine learning …) become more and more successful and widely adopted, larger and larger LRs are required.

It is now obvious that quality of LRs has direct implications to the quality and performance of the systems.

LR Development needs human resources and dedication. It is a time consuming task to obtain high quality data.

A lot of work still needs to be done on the following points:

- Semantic annotation, ontology building and thesaurus production,

- Bilingual and multilingual corpora creation,

- Multilmodal / multimedia resources collection and annotation.

It is of importance to identify the availability of LR for the development of core technologies in the various countries / languages.

Methods have to be developed regarding the following points:

- Time decrease to customise LE systems,

- Time decrease to transpose to another language.

The Internet is of uppermost for LRs LR (LR <=> INTERNET <=> LE)

- The Internet is a big deposit of LRs,

- Internet's functionality can be improved through LE-techniques,

- LR emerging standards will help to structure information in the Internet,

LRs and LE will be more and more required regarding the improvement of the following functionalities:

- information extraction

- information retrieval

- document routing and classification

A European effort is necessary in the terminology field for the standardisation of both the technical and the procedural aspects (collection, quality control…)

Several measures need to be conducted at national and international levels in order to achieve the following:

NATIONAL	INTERNATIONAL
Not to exclude A language from Information Society and Man-Machine Communication	Not to delay market penetration of sophisticated products in some countries and delay progress (Industrial Importance)

That means, at a national level:

- Collection of monolingual resources
- The maintenance of resources,

And at an international level:

- Resources handling tools,
- Standardisation,
- Multilingual resources production,
- Evaluation and validation of resources.

General remarks on the 2nd LREC Conference

Education training in the field has been addressed (fortunately, we benefit from the Elsnet activities),

The Integration of LRs and LE - tools in "CALL" software was only addressed in one specialised workshop. A very positive evolution is the long-term design in EU from now on.

George Carayannis
Institute for Language and Speech Processing (ILSP),
Artemidos & Epidavrou Str.
Paradisos Amarousiou
151 25 Athens, Greece
Email: gcara@ilsp.gr

LREC Panels Summaries

Resources for the Millennium

Catherine Macleod, New York University, USA

Participants: Jeffrey Allen (ELRA/ELDA), Lin Chase (Speechworks), Sadaoki Furui (Tokyo Institute of Technology), Lynette Hirschman (Mitre), Sadao Kurohashi (Kyoto University), Nils Lenke (Philips Speech Processing), Masumi Narita (RICOH), Antoine Ogonowski (LEXIQUEST) and Marilyn Walker (AT&T Research)

The intent of this panel was to assemble researchers from various fields of natural language processing to discuss the resources that they believe will be needed in this millennium. The discussion covered a number of diverse types of resources. We hope it will give some direction to the development of future resources.

Professor Furui, Professor Kurohashi and Ms. Narita discussed various resources being developed in Japan. Among these projects are: the tagging of the Kyoto Text Corpus, a new MT Project, a spontaneous speech corpus with processing technology adapted to it, and the Japanese Learner's Corpus. The researchers involved in these projects are commercial, university and governmental. The GSK has been recently formed to collect and distribute natural language resources (much on the lines of the European ELRA and the LDC in the U.S.).

Spoken dialogue systems need common resources for automatic training of dialogue systems, according to Marilyn Walker. These include dialogues for different domains, cross-system logfiles (logged with a common tool), standards for representing system behaviors and module metrics and standards for cross-system evaluation. These resources would enable training in dialogue management and natural language generation. Lin Chase also discussed the need for spoken dialogue resources.

Mr. Ogonowski wanted to emphasize the need for large sharable, structured standardized resources in many languages. In Y2K we need standardized encoding, linguistic description and semantic/conceptual description. Mr. Allen's report on ELDA's survey on the NLP resources needs of the European community supports this conclusion. Users want a variety of resources in speech and text processing. Speakers from 33 languages responded, underscoring the need for multi-lingual resources. New ELRA supported resources include special text corpora, speech corpora, and lexica.

Lynette Hirschman (Mitre) wondered whether, for some applications, resources could be developed on the fly. For this, we need common tools for data cleanup, cheap storage, cheap annotation modules and methods for sparse annotation. Possibilities other than labor-intensive corpora must be explored to enable access to "live" information.

To summarize, many of the panelists were involved in developing corpora. They represent a cross-section of researchers who are cooperating in the venture of making this labor-intensive and time-consuming task of resource creation possible. It is clear that besides needing common resources with common annotation, we need common tools for accessing this information and common programs to utilize it. The novel idea of developing "on the fly" resources is not alien to the panel's point of view. It also requires common notation and common tools to take advantage of the reams of "raw" data available nowadays. Cooperation among different research entities and different countries is needed to ensure that the resources we develop today will be useful in the years to come.

Catherine MacLeod
New York University
251 Mercer Street, NY 10012 New York
USA
Email: macleod@cs.nyu.edu

Human Language Technology Resources for Central European Languages

Zygmunt Vetulani, Adam Mickiewicz University, Poland

The panel discussion on "Human Language Technology Resources for Central European Languages: European Integration Issues" was intended as the opening of a public debate on the state of the art and on the future developments in the domain of language resources for Central European languages. Particular emphasis was on how to encourage vigilance, active co-operation, and co-ordination of HLT resources, which will be essential prerequisites for integration in the near future.

The invited panelists represented Central European countries[1], EU countries[2] and the European Commission[3].

The session was opened by Zygmunt Vetulani (chairman) who gave a short presentation of the main problems identified within the pre-event discussion among the panelists (published in the LREC proceedings). The central problem was the existing technological gap in the field of LE resources between EU and CEE countries in the context of European integration. The term "technological gap" appeared hardly acceptable for some of the participants (Boitet, Gibbon, Maegaard). It should therefore be explained that this term was not intended to apply to the expertise level of individual researchers but rather to the overall technological level biased by systematic under-investment in the HLT domains: non-existence of some important resources, low awareness level (industry), deficits in the area of well-trained personnel.

Maria Gavrilidou pointed at Greece as a country whose experience may be helpful for the CEE candidate countries because of its size and the fact of being a relatively "young" member of the Community. The main point of her presentation was that the need to meet the demands stemming from the evolution of the LR field and the effort to keep up with the state of the art constituted a target which proved to be demanding for a small country with relatively recent presence in this field. The observation of Gibbon about needs for human resources worth noting here. This aspect was also raised by Támas Váradi who noticed the lack (in Hungary[4]) of formal training in computer linguistics within the higher education system. Concerning traditional resources, Dafydd Gibbon estimated that the main goals in the coming years lie in the acquisition of aligned multimodal and multilingual corpora for lingware and system development. The role of ELRA for the CEE language resources was one of the issues for discussion. Commenting on this problem Bente Maegaard shortly presented the philosophy of ELRA focussing on the awareness mission and on the new EUROMAP initiative. She also noted the fact of limited budgets for commissioning language resources so that ELRA needs to look at the market value of resources. For some countries this may be a real problem: e.g. Váradi, speaking about obstacles facing the HLT R&D work in Hungary (actively developing HLT resources within EU funded projects), singled out the relatively small size of the Hungarian market. Also Eva Hajicova contributed with the observation that in the CEE countries the local software companies still struggle with financial problems of their own and are not able to support research and development to a degree comparable with companies in the West. This is another reason why financial support has to be looked for at the EU level. Hajicova also observed that for some kinds of resources, as e.g. parallel corpora for comparative research, it is hard to get national support and therefore an international support is necessary. Roberto Cencioni was the last of "official" panelists. He confirmed his good understanding of the challenge for Europe connected with future "new" languages on the European map and problems connected with the scale of the integration process (there is clearly no possibility of financing large scale MT projects for 110 language pairs). He pointed that to act in favour of enhancement of the HLT in the EU-accessing countries is among his main priorities.

At the end of the session, a public debate was opened. The question was raised whether the European Commission envisaged completing its own language resources with the languages of the candidate countries. Dimitrios Theologitis[5], replied that there existed a plan (pending official acceptance) to create "Pre-Eurodicautom" terminology and "Pre-Euramis" translation memories based on the work done at the Translation Co-ordination Units translating Community legislation, the "Acquis". On accession day, these resources would be merged with the existing ones.

Christian Boitet[6], arguing with Cencioni, made an implicit association between the European policy and state of technology. He noticed that (perhaps) there would be more groups working on MT in Europe "if the EU had not rejected almost all proposals concerning MT from the end of Eurotra until very recently." He also presented the UNL project (Universal Networking Language) for developing of a computer format to represent the linguistic content of documents in a language independent way and fervently proposed that EU should now support the effort of building UNL-related language resources for all the European languages.

There are good reasons to think that the panel discussion was well appreciated by the audience: the number of participants increased constantly during the session and practically nobody left before the end[7].

Zygmunt Vetulani
Adam Mickiewicz University
ul. Matejki 48/49, 60769 Poznan
Poland
Email: vetulani@math.amu.edu.pl

[1] Eva Hajicova, Charles University, Prague, Czech Republic,
Támas Váradi, Hungarian Academy of Sciences, Budapeszt, Hungary,
Zygmunt Vetulani, Adam Mickiewicz University (UAM), Poznan, Poland.

[2] Maria Gavrilidou, Institute for Language and Speech Processing (ILSP), Athens, Greece,
Dafydd Gibbon, University of Bielefeld, Bielefeld, Germany,
Bente Meagaard, Center for Language Technology (CST), Copenhagen, Danemark.

[3] Roberto Cencioni, European Commission, DG Information Society, Luxembourg.

[4] But this a general situation in most of CEE countries.

[5] Dimitrios Theologitis, EC Translation Service, Luxembourg.

[6] Christian Boitet, University of Grenoble, GETA, Grenoble, France.

[7] Because of space limitation, it is not possible to present the standpoints of the panelists in a more complete way, we will publish the more complete abstracts on the WEB page at http://main.amu.edu.pl/~zlisi/news/lrec2000.htm

Speech Database Processing Tools: the state of the art in automatic labeling of speech

Nick Campbell, ATR-ITL, Japan

The panelists were: Stephen Bird (LDC), Alistair Conkie (AT&T), Edouard Geoffrois (CTA/GIP), Dafydd Gibbon (University of Bielefeld), Bruce Millar (Australian National University), Vincent Pagel (MBROLA), Jan van Santen (OGI/CSLU), and Kåre Sjölander (KTH/CTT).

The session was organised by Nick Campbell (ATR).

The panelists discussed the extent to which presently-existing tools can be used in the creation and annotation of large speech corpora, and proposed the development of an open-source toolkit for the segmentation, annotation, and visualisation of acoustic and prosodic characteristics. Specific topics that arose included the difficulty of standardising data formats, the use of annotation graphs and data models, evaluation standards, software licensing conditions, visualisation software, and speech-specific programming languages such as Tcl/Tk's Snack speech processing extensions.

Although a high degree of co-ordination and integration can be reached at the programming level for software and tools, it may be unrealistic to try to achieve consensus on physical file formats. Rather, the community should seek general-purpose data models which can be stored and visualized in a number of ways. Such shared data models with their associated application programming interfaces, can provide the foundation for wide-ranging integration of tools and databases.

Following the panel discussion, a COCOSDA technical topic domain ‘Corpus Annotation Tools’ was instigated to co-ordinate efforts in this area. Steven Bird is to be the Rapporteur for this new topic domain. Results and progress will be announced under the new COCOSDA Website at www.slt.atr.co.jp/cocosda.

Nick Campbell
ATR-ITL, Japan
Email: nick@itl.atr.co.jp

International Cooperation in the Field of Language Resources and Evaluation

Antonio Zampolli, Istituto di Linguistica Computazionale del CNR, Italy

1. Background

The issue of international co-operation was extensively discussed at the first LREC in Granada (1998), with emphasis on the following issues:

- Language resources (LR) are essential components of HLT activity, supporting research, system development and training, and evaluation in both the mono- and multilingual context.

- A key enabling condition of integration of different technologies and languages requires that LR are shared among different sectors and applications.

- The richness of the multilingual capabilities associated with a language depends on the number of languages for which adequate LR exist.

- The high cost and effort of the production of LR should be shared, in order to make them more affordable. The creation of multilingual LR requires agreement on a co-ordination policy, to ensure the reuse of existing monolingual resources and to facilitate access to native speakers of the various languages.

The situation in the field of evaluation is rather different in Europe and in the United States, where American and European expertise seem to be complementary. The question of co-operation in the field of evaluation therefore arises very naturally, in particular because many experts believe that it is often only through such evaluations as TREC and MUC that research finds a common focus and makes easily quantifiable progress.

Three events of the first LREC have particularly stimulated discussion on these topics:

(1) the Panel on "Co-operation between EU and Other Countries in the Field of Language Resources and Evaluation" [see A. Zampolli, "Panel of the Funding Agencies", in ELRA Newsletter, Vol. 3, No. 3 (August 1998, Special Issue on the 1st LREC)];

(2) the Panel on "International Co-operation" [see A. Servantie, Panel on " International Co-operation", in ELRA Newsletter, Vol. 3, No. 3 (August 1998, Special Issue on the 1st LREC), p. 12];

(3) the Closing Session of the post-Conference Workshop on "Cross-lingual Information Management" [see E. Hovy, A. Zampolli, "Governments: Policy and Funding", Chapter 10, in E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, (Eds), "Multilingual Information Society: Current Levels and Future Abilities", to be found at http://www.cs.cmu.edu/~ref/mlim/index.html].

The following areas of Language Technology emerged in the Granada debates as being in urgent need of international co-operation:

- Standards: de facto, best practices.

- Language Resources and Related Tools.

- Core Technologies.

- Evaluation.

- Selected vertical sector domains.

These aspects were endorsed in the session dedicated to HLT at the International Conference on "New Vista in Transatlantic Scientific and Technical Cooperation," organised on the occasion of the signing of the transatlantic technical and scientific co-operation agreement (Washington DC, June 1998).

2. Objectives of the Panel

The panel aimed, in a sense, at putting together the main issues which were the focus of the first LREC events quoted above: a survey of the current programs, initiatives and underpinning policies of the Funding Agencies in different parts of the world, a discussion of the needs and opportunities for a world-wide co-operation in the field.

3. Overall Structure of the Panel

The panel was structured in four parts: introduction, panelists (presenting the situation in various parts of the world), discussants (commenting on specific issues) and a general discussion involving the audience.

3.1. Introduction

Antonio Zampolli (University of Pisa, ILC-CNR)

GENERAL FRAMEWORK

The issue of international co-operation was extensively discussed at the first LREC in Granada (1998), with emphasis on the following issues:

- Language resources (LR) are essential components of HLT activity, supporting research, system development and training, and evaluation in both the mono- and multilingual context.

- A key enabling condition of integration of different technologies and languages requires that LR are shared among the different sectors and applications.

- The richness of the multilingual capabilities associated with a language depends on the number of languages for which adequate LR exist.

- The high cost and effort of the production of LR should be shared, in order to make them more affordable. The creation of multilingual LR requires agreement on a co-ordination policy, to ensure the reuse of existing monolingual resources and to facilitate access to native speakers of the various languages.

The situation is different in the field of Evaluation in USA and in Europe.

- The complementarity of expertise can be an issue of co-operation.

Many experts believe that it is often only through such evaluations as TREC and MUC that research finds a common focus and make easily quantifiable progress.

INTERNATIONAL AND NATIONAL FUNDING AGENCIES

- The interest of national and international Funding Agencies in the social, economic, industrial and strategic impact of HLT has decisively contributed to the directions of evolution of our field.

- This interest is bound to grow in the current context of the global Multilingual Society.

- HLT (in particular, LR) involves not only R&D issues but also cultural and political aspects.

INTERNATIONAL CO-OPERATION

As the Speech and NLP field matures, as technology is increasingly commercialised, international co-operation is increasingly important. It:

- enhances advance in the state of the art by combining more effectively the strengths and excellence developed in different regions;
- facilitates the integration of LT across languages, surely one of the key aspects which makes this field relevant to the society at large.

In the light of such arguments, the US government and the EC have recently (June '98) signed an agreement for scientific and technological co-operation.

HLT has been (one of) the first sectors to implement this agreement.

MULTILINGUAL LR

In particular, the production of multilingual LR poses:

- research issues and challenges;
- organisational problems:

who has the responsibility of promoting the co-operation of R&D communities speaking different languages and how this should be done?

The situation is different for:

- types of LR: corpora, lexicons etc.;
- large/general multilingual LR;
- applications specific LR;
- customisation;
- different types of information (data VS analytical/interpretative features).

TOPICS FOR DISCUSSION ON INTERNATIONAL CO-OPERATION

- Needs, themes, priorities:
  - for HLT,
  - for other IS sectors,
  - for different types of LR/EV,
  - for different phases of LR development (research standards; specifications; construction; maintenance; updating; technology transfer; etc.);
- reasons;
- different roles, responsibilities, challenges.

3.2. Panelists

Roberto Cencioni (European Commission, DGXIII, E4)

The core of the mission of the Units he is heading in Luxembourg is to promote advanced technologies for HLT and natural interfaces to access, assimilate and use multimedia content.

The programs include both spoken and written language(s) and address human-computer interaction, interpersonal communication, information management and encompass R and D, demonstration and market stimulation activities.

International co-operation is - so to speak - directly built in the very nature of the programs: they represent widely recognised focal points: 200 million Euro have been dedicated since 1992 and 90 projects have been supported since 1997. By the end of the year 30 new projects will be underway, involving about 400 participants from more than 20 countries.

International co-operation is, from this point of view, "easy", because multination, multiparty collaborations are the norm.

This approach is rather "Eurocentric" and can be compared with the world-wide approach of the US Agencies, which have come to realize the potential of multilingual ITC.

International co-operation is essential for LR: it will be more and more important to take into account that affiliated and new accession countries bring their languages with them.

Another crucial issue is the co-ordination between EU and national activities: in particular, it is obvious that the EU can not, alone, support the development of adequate LR for all the European languages. Initiatives and proposals in this direction will be welcome.

LR are an essential component for reaching the targets of the programs, determined by the overall social and technological framework.

E-commerce should provide instant access to global markets; business should speak the language of the customer; mobile communications, wireless multimedia etc. provide new opportunities for e-business.

Internet is increasingly multilingual: 50% of surfers speak languages other than English and bi- and multi-lingual Web sites are slowly becoming the norm. We should move towards an inclusive Information Society, overcoming exclusion factors due to language, culture, computer literacy, disabilities etc.

Today enterprises should be IT and knowledge bound: hence, the relevance of content-based and cross-lingual information.

The LR, needed for as many languages as possible, can be built and made available only through concrete international collaboration activities.

At all project levels, collaboration with third countries is unproblematic, per se, if matching resources are available. In fact, HLT has been the first IST sector to launch joint programs (currently five) with NSF, following plans discussed at the first LREC in Granada: it should be noted that less than one year elapsed between these discussions and the first call including co-operation with NSF.

From another point of view, international co-operation proves, in concrete terms, to be "difficult".

Government and agency level collaboration presupposes well-established programmes on each side, similar policy and research agendas, ambitious and sizeable endeavours, balanced participation and synchronized operations, good will and personal trust at personal level, continuity over time.

It will be very interesting to hear what the situation is in other parts of the world.

Lynette Hirschman (MITRE Corporation, Bedford) presented the US perspective, speaking also for *Gary Strong (DARPA, Washington, former NSF)*, unable to participate, as intended, for managerial duties.

The vision of US technology directions, as defined by DARPA, is to move beyond document access, towards providing "just-in-time", "just-right" information to the user: the goal is to connecting the user with world class expertise via natural, conversational interaction with on-line, distributed resources. These resources may be free text, broadcast news, formatted databases - or other people with appropriate expertise or information. The information must be presented to the user in the appropriate form (short answer, graph, table, summary) and in the appropriate medium. By providing conversational access over mobile devices, we can bridge the digital divide, making Internet connectivity globally available. By focusing on the issue of multilingual and spoken language access, we can begin to bridge the language divide, providing translingual processing for the major world languages and preserving cultural heritage for non-written and minority languages.

DARPA's two major human language programs address these goals. The DARPA Communicator focuses on a plug-and-play architecture for conversational interaction to distributed resources. It is making available an open-source implementation of this framework (http://www.fofoca.mitre.org), and has put into place a DARPA Affiliate structure, to encourage international collaboration. The DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program focuses on translingual information access. Major goals are speech-to-speech translation, a toolkit to develop machine translation capabilities in a day or a week, and translingual question answering systems (see http://www.darpa.mil/ito/research/tides).

These research programs, together with other international programs, such as the joint US-EU Multi-lingual Information Access and Management (MLIAM) program, and the developing Western Hemisphere Alliance for Information Technologies program, are funding the creation of shared infrastructure and resources. In addition to these opportunities, many opportunities for informal sharing or exchange of resources exist through the Linguistic Data Consortium, through open source tools, and through the extensive series of technology evaluations supported by DARPA that are open to international participation.

Jun'ichi Tsujii (University of Tokyo) presented his view of the Japanese situation.

Mutual understanding is an essential prerequisite for international co-operation to be fruitful. Each region has its own historical and cultural background, which influences research interests and the whole direction of research projects. In his talk, Tsujii briefly summarized the Japanese experience from the early '80s till now and explained what kinds of research programs are under way now in Japan and why. In particular, he emphasized that the Japanese research community has focused on basic generic NLP techniques throughout the '90s after the period of exploratory integration of basic techniques of the '80s. As a result, the Japanese community now feels to have reached the stage where another integration of basic technologies will be fruitful as well as possible. This type of research, i.e. exploratory integration needs public support for close international co-operation, while basic research of generic technologies as well as application-oriented development can be pursued in a looser co-operation form.

International co-operation in NLP seems more difficult than in those sciences such as brain science, physics, human genome, space science, etc. This is because our field is more tightly linked with social goals of individual countries as well as commercial interests of private sectors. Therefore, natural fields of co-operation would be in those fields independent of particular applications. International co-operation will be increasingly important in the field of collection/gathering and integration of multi-lingual resources, which support exploratory integration of basic technologies in the early 21st century.

Feng Zhiwei (State Language Commission of China, Beijing; currently at the University of Trier) presented a detailed inventory of LR (Text Corpora, Tools for Corpus Processing, Machine Dictionaries, Grammar Knowledge Base, Terminology Data Bank) available or under construction for Chinese, discussed channels of Chinese government funding for HLT, investments of private companies and the needs and opportunities for international co-operation.

Chinese language is the most important language of Sino-Tibetan language family. Now nine hundred forty million people in the world speak Chinese language as their mother tongue. Not only Chinese people speak Chinese language, some people in Singapore and Malaysia also speak Chinese language. Chinese language is one of the working languages for United Nations.

Chinese language resources and evaluation must deal with the Chinese characters. It is a remarkable feature for Chinese Language Technology (CLT). CLT is an important part of Human Language Technology (HLT).

Standards are an obvious priority issue for international co-operation.

For text corpora, international co-operation is mainly promoted through joint projects with foreign countries. "People's Daily" corpus processing is a joint project between ICL-PKU (China) and FUJITSU Company (Japan).

For other types of language resources, international cooperation is mainly achieved by sharing resources, data and tools.

Machine dictionary GKBCC: sharing with Intel (USA), Matsushita (Japan), XRCE (Xerox Research Center Europe, France), CiTaL (Centrum für Terminologie Internationale und Angewandte Linguistik, Germany), KAIST (Korea Advanced Institute of Science and Technology, Korea), Pecan (a sub-company of CANON).

Corpus processing tool Slex: sharing with Intel (USA), Matsushita (Japan), XRCE (France), CiTaL (Germany), KAIST(Korea), NUS (National University of Singapore).

Terminology Data Bank: sharing with CiTaL (Germany).

3.3. Discussants

According to *Joseph Mariani (LIMSI-CNRS, Paris)*, the LREC 2000 conference on Language Resources and Evaluation in Athens was the opportunity for the international community to meet, report on the present situation and propose cooperative actions.

The present situation in Human Language Technologies evaluation is that the US keeps on organizing large comparative evaluation campaigns embracing speech and natural language, with a large European participation which is not funded by US or EC funds, but it appears that the interest in participating is strong enough to prompt this free participation. DARPA starts new programs (Communicator and TIDES on Translingual Information Detection, Extraction and Summarization) using the evaluation paradigm within a common architecture, and several European laboratories join those programs as affiliates. In Japan, forces on Text processing systems evaluation have been gathered in a single entity, the National Institute for Informatics (NII). Apart from those large programs, several initiatives are taking place in various places around the world, such as the evaluation campaigns in France (AUF, Amaryllis, French DoD...), in Germany (within the Verbmobil or SmartKom programs) or at the international level (Senseval, for example). Such a tool is still lacking in the European Commission programs.

Two questions then raise.

- Is there room for several initiatives around the world?

The answer seems to be yes, as there are different languages to be covered, there may be different ideas based on different cultures and therefore discussing those ideas may help defining the best way to handle the question, and finally because the size of the effort is very large, thus necessitating shared efforts to cover the various tasks in the various languages.

- If so, should it be coordinated?

The answer seems also to be yes. It is obvious that science and technology are international, and that evaluation should therefore be conducted at the international level. Laboratories find it difficult to participate in all initiatives due to lack of time and manpower. Thirdly, it appears in the present situation that it is difficult in the various initiatives to get the necessary language resources in the various languages aimed at, and also it would avoid reinventing the wheel in the design of evaluation methodologies.

We should therefore try to find a way to install a truly international human language technologies evaluation scheme, one of the problem being that it doesn't fit in so well with the EC programs Call for Proposals mechanisms, and that creating an institute comparable to NIST or NII in Europe will be a very difficult task, which may take a long time and a large amount of efforts.

Harald Hoege (Siemens, Munich) started considering that in the last five years a successful infrastructure to produce, disseminate, standardize and validate SLR has been set up within Europe and US. This infrastructure becomes visible through ELRA and LDC. Also activities in Japan start working in this direction. Due to the different funding strategies of the national bodies no common international approach exists.

He proposed to start such a common production and dissemination strategy through the following actions:

- International production of SLR for Speech-to Speech translation for 50 languages at an international level.
- Each funding agency (Europe, US, Asia) supports this action by 20MECU (ca. 1 Million ECU per language).
- of the SLR through a common dissemination policy on a license free basis.

On the basis of the previous interventions, *Volker Steinbiss (Philips, Aachen)* asked various questions on the role that ELRA can play for the development of LR through international co-operation, focusing in particular on overall policy issues.

Núria Bel (gilcUB, Barcelona) stated that, as HLT components are more and more being included in all kind of IT applications, Language Resources should be considered as a basic infrastructure for current and future Information Society. As any other basic infrastructure, these resources need to be created, maintained and updated, and this means a planning based on a long term strategy and a long term funding. Besides, there are already examples (such as software localization) that have proven that availability of all kind of applications in local market languages becomes to be considered a further user requirement. There is such a demand. Hence we should not expect a full deployment of HLT in the world without addressing all kind of local languages, independently of its number of speakers.

This infrastructure is, with no doubt, a very expensive investment, and because of the social and economical interests which are behind of the area of HLT applications, it is commonly agreed that there should be public support for them. Until now, in Europe there has been two strategies: to appeal to the subsidiarity principle, so that each state should care of covering its language, or, as a more strategic international policy, to fund such initiatives in the form of EU R&D projects. Some of these projects, though, have given support to very concrete multinational industries in this area, resulting in a non widely sharable infrastructure, and, more crucially, an infrastructure that is only available for those languages that have an interest for these industries because they have a large market, major languages which are not spoken only in one country, languages which are not only EU languages.

It seems, hence, that on the one hand the EU is investing public funds in languages that have a clear market even though they are spoken in many different countries around the world, a fact that one would expect to be the basis for international co-operation outside the EU. On the other hand, some other European languages are left to national initiatives because of its low interest in terms of short term marketing. These national initiatives exist, but they lack common organization and normally they count on low funding because the arguments used to defend them are mainly based on supporting cultural diversity, which is, as we know, a non very attractive argument in terms of funds. For them, international co-operation will mean political support.

If we look at other areas where economic, social and politic interests play a role, such as health, nuclear, space, aeronautic research and development, we can see that the different administrations have managed, in co-operation with interested industries, to create special agencies or large projects, with fixed contributions from the different participants, and, what it is really important, long term planning and funding. Hence, do not we go for such an international agency for Language Resources? An overseas international body that organises, plans and fixes long term strategies for the development, maintenance and update of this HLT infrastructure for all languages.

Lori Levin (Carnegie Mellon University, Pittsburg) presented NICE (Native language Interpretation and Communication Environment) as an example of collaboration between United States and Latin American countries. The project, dealing with MT between Spanish and indigenous languages, was conceived by U.S. funding agencies (NSF and DARPA) along with the Organization of American States in the context of a larger project on Western Hemisphere collaboration in multilingual contexts.

There was a concern about disenfranchisement of speakers of indigenous languages from goverment and the Internet.

The first Latin American partner is the Universidad de la Frontera in Chile. It was learned from them that the Mapuche people would view a machine translation project in the context of community development, which in their villages is centred around the schools.

As a result, we are working primarily through the Ministry of Education in Chile.

This is in contrast to other countries where machine translation projects are centred around government, industry, or defense.

3.4. General Discussion

A general agreement emerged on the need of international co-ordination and co-operation, which appears the only way to provide the LR required to answer the challenges and the expectations of the contemporary evolving multilingual ICT-based Society.

In Europe, an explicit co-ordination should be established between the initiatives of the EU and the activities of the member States: in fact, the prevision of LR is a common target for the various European national projects, and initiatives of the type of ENABLER should be developed and maintained.

Several interventions highlighted specific needs, calling the attention on opportunities for international co-operation offered by planned or on-going initiatives.

Due to lack of space, we can quote only a few examples here.

Zygmunt Vetulani (University of Poznan) observed that creation of LR for languages of eastern countries is a priority for HLT development in these countries and represents an uncontroversial logical starting point for eastern-western co-operation.

Piek Vossen (Sail Labs GmbH, Munich) and *Christian Fellbaum (Princeton University)* announced a new international association aiming at fostering co-operation among researchers and developers interested in lexical semantic networks.

Tarcisio Della Senta (United Nation University, Tokyo), offered UNL, and in particular the wealth of LR-corpora, lexica, knowledge developed for languages of five continents, as an example and a forum for international co-ordination.

Gerhard Budin (University of Vienne) and *Rute Costa (President of EAFT)* observed that the situation of terminology is ripe now, both from the organizational and the technical point of view, to realise the co-operation with computational lexicography, well recognized as a need but never practically firmly established.

Steven Krauwer (University of Utrecht) briefly summarised the institutional vocation of ELSNET to promote international co-operation, and offered the expertise and the infrastructure of ELSNET, in particular the ELSNET task force for LR, for helping implementing a world-wide co-operation.

Ideas and suggestions emerged during the Panel were immediately taken in consideration, already during the remaining of the Conference, in particular the proposal for establishing an overall world-wide initiative, involving existing infrastructures like ELRA, LDC, COCOSDA.

A first meeting will be organized, in co-operation with a workshop sponsored, at the ACL Conference in Hong Kong (October 2000), by ELSNET, to address questions like: (1) what are the existing infrastructures which should be involved world-wide, and how they can be optimally exploited to foster global co-operation; (2) what infrastructure and interconnections are missing, and which are the main actors (institutions, organizations) to be involved to build and operate a truly overall international infrastructure; (3) what are the mandate and more urgent priorities for such an infrastructure. A second discussion will be organized at the occasion of the next COCOSDA meeting (which takes place two weeks after in Beijing).

Post-Panel Discussion

Those wishing to further contribute to the discussion, for example reporting on experience of international co-operation, highlighting general or specific needs, suggesting priorities, or commenting on policy and organisational problems, are invited to send messages to the discussion list intpan@ilc.pi.cnr.it. If appropriate, we will channel comments and suggestions to the relevant funding agencies.

The same Web site will make available the transparencies used at the COLING Panel on "International Co-operation" (Saarbrüchen, August 2000), and the following discussions.

Antonio Zampolli
University of Pisa
Department of Linguistics
Istituto di Linguistica Computazionale del CNR
Pisa
Email: intpan@ilc.pi.cnr.it

LREC Technical Sessions Summaries

Dialogue Evaluation Methods

Sadaoki Furui

The following four papers were presented in this session.

Carine-Alexia Lavelle et al.: "Dialogue and prompting strategies evaluation in the DEMON system"

Helen Bonneau-Maynard et al.: "Predictive performance of dialog systems"

Niels Ole Bernsen et al.: "A methodology for evaluating spoken language dialogue systems and their components"

Marilyn Walker et al.: "Developing and testing general models of spoken dialogue system performance"

The first paper discussed prompting strategies for spoken dialogue systems. A set of measures to evaluate three different confirmation strategies was presented. Five criteria were then used to evaluate the systems' question complexities and their effect on users' answers were investigated.

The second paper investigated predictive performance measures of dialogue systems by measuring the system's performance using an objective cost function. Using the PARADICE paradigm, a performance function derived from the relative contribution of various factors was obtained for two different systems. It was found that the most important predictors of user satisfaction were understanding accuracy, recognition accuracy and number of user repetitions.

The third paper presented results of the European DISC project concerning technical and usability evaluations of dialogue systems and their components.

The fourth paper presented the PARADISE methodology for developing predictive models of spoken dialogue performance, and showed how to evaluate the predictive power and generalizability of such models. A number of models for predicting system usability (as measured by user satisfaction) was developed for two dialogue systems. The results showed that the models generalized well across the two systems.

Various interesting and lively debates were pursued concerning how to evaluate and predict performances of spoken dialogue systems. There are still many possible avenues for improving the models of user satisfaction and the performance measures of spoken dialogue systems.

Professor Sadaoki Furui
Tokyo Institute of Technology, Department of Computer Science
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: furui@cs.titech.ac.jp
http://www.furui.cs.titech.ac.jp

Data Centers/Major Projects

Lin-Shan Lee

This session is to address the various issues, considerations and experiences with data centers and major projects. 1 paper is from LDC, 2 from ELRA, 1 from COCOSDA, 1 from Motorola Center and 1 for a general platform.

The first paper, Issues in Corpus Generation and Distribution: the Evolution of the Linguistic Data Consortium (LDC) by C. Cieri and M. Liberman, presented the recent LDC efforts regarding the creation and distribution of language resources. The increased demand for larger corpora with more sophisticated annotation for a wide variety of languages were reported. Distribution of resources via different channels and quite several new projects were mentioned. The fourth paper, Recent Developments within the European Language Resources Association (ELRA) by K. Choukri, A. Mance and V. Mapelli, illustrated the various developments in ELRA. On ELRA catalogue there are 111 speech resources, 163 monolingual and multilingual lexica, 24 written corpora and 275 terminological databases. The sale of language resources grew from 33 in 1997, 180 in 1998 to 217 in 1999. The membership of ELRA was also increased significantly. Various issues including identification of resources, legal problems, distribution channels, validations and quality assessment, etc., were all discussed. The sixth paper, Survey of Language Engineering Needs: A Language Resources Perspective by J. Allen, K. Choukri, presented the summary of an on-going survey on language engineering needs conducted by ELRA. Statistical data for many issues were available indicating possible directions for development of language resources. The fifth paper, COCOSDA - A Progress Report by N. Campbell, provided the updated progress of the Coordinating Committee for Speech Database and Assessment (COCOSDA), from its history to recent developments, including the renewal of the Central Coordinating Committee (CCC) and the re-structure of the functionalities by a matrix whose two dimensions are the topic domains and the regions. The second paper, The Establishment of Motorola’s Human Language Data Resource Center: Addressing the Criticality of Language Resources in the Industrial Setting by J. Talley, addressed the various issues and experiences for a 1-man center for language resources in an individual company, while the third paper, A Platform for Dutch in Human Language Technologies by E. D Halleweyn, E. Dewallef and J. Beeken, presented experiences and considerations for the establishment of a common, convenient and efficient platform for Dutch language technologies.

Speech Recognition and related issues

Herman J.M. Steeneken

In this session five papers were devoted to the evaluation of speech recognition systems or its application. One paper concerns speech coding and one paper the evaluation of multi-model multi-user systems.

The paper by Bengler discusses speech-input and speech-output system aspects for use in cars. Rather than the usual performance measures the usability is determined. From a manufacturer point of view usability aspects are of great importance for successful application. It is clear that dialogue design is a major topic. Error driven approaches are discussed.

Distribution of reference recognizers is still a topic as it was in the time of the former EU-funded SAM-project. Two papers describe software based reference systems. The Japanese IPA dictation system is designed for large vocabulary continuous speech. The performance for the given examples of speech tokens covers 5 - 8% word error rates.

Multilingual speech recognition is offered by the COST 249 reference recognizer. This system is designed for language independent training procedures for the phonetic recognizer system. As expected the performance is vocabulary dependent, examples are given for a number of conditions: e.g., from isolated digits to city names (500-1100) and for five languages (Danish, Norwegian, Slovenian, Swedish and Swiss German).

A methodology was presented for the evaluation of multi-modal multi-user group-ware systems. The evaluation of two systems is scenario based and made use of specific evaluation metrics. The examples concern a Map Navigation Experiment and an experiment on Information Management.

This session on speech recognition initiated some fruitful discussions on evaluation paradigms. Two examples on assessment were presented, There seems to be a need for some standardized guidance on experimental design and the related statistics in speech technology .

Information Retrieval and Question Answering Evaluation

Stella Markantonatou

The seven papers presented at the Session entitled "Information Retrieval and Question Answering Evaluation" focused on issues concerning the resources and the automatic means needed to evaluate IE and QA systems. Currently, several monolingual and multilingual IE and QA systems are either available or at the stage of development. Such systems become popular as the number or users who need to regularly retrieve information from a multilingual (text or speech) document collection is steadily increasing, especially with the use of WEB facilities. At this successful Session, papers were delivered lively and there was interesting interaction between the paper presenters and their audience.

The first four papers reported on four different projects which have provided corpora and tools for the automatic evaluation of monolingual and multilingual IE systems. The paper entitled "The Evaluation of Systems for Cross-language Information Retrieval" by Martin Braschler, Donna Harman, Michael Hess, Michael Kluck, Carol Peters and Peter Schauble reported on the approach adopted and the issues that had to be taken in consideration while constructing an infrastructure for testing cross-language text retrieval systems within the framework of the Text REtrieval Conferences (TREC) organised by the US National Institute of Standards and Technology (NIST). On a similar track was the paper entitled "IREX: IR and IE Evaluation Project in Japanese" by Satoshi Sekine and Hitoshi Isahara and focused on Japanese. The paper entitled "Textual information retrieval systems test: the point of view of an organiser and corpuses provider" by Patrick Kremer and Laurent Schmitt reported on the experience of INIST as providers of corpora for testing IR systems: the difficulties encountered in obtaining the material for building such corpora and the need for a wider collaboration among the providers of evaluation systems and their users. The paper entitled "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation" by Charles L. Wayne reported on DARPA-sponsored research on the evaluation of automatic techniques for locating topically related material in streams of data such as newswire and broadcast news. The program has provided well-designed corpora and objective performance evaluations. The next two papers report on the evaluation of Question Answering Systems. The paper entitled "How to Evaluate your Question Answering System Every Day … and Still Get Real Work Done" by Eric J. Breck, John D. Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light and Inderjeet Mani reported on Qaviar, an experimental automated evaluation system for question answering applications which provides an automatically calculated measure that correlates well with human judges' assessment. The paper entitled "The TREC-8 Question Answering Track" by Ellen M. Voorhees and Dawn M. Tice summarized the results of the TREC-8 Question Answering track offering an overview of the approaches taken to the problem and an analysis of the evaluation methodology. Finally, the paper entitled "Cardinal, nominal or ordinal similarity measures in comparative evaluation of information retrieval process" by Christine Michel addressed the issue of evaluating IE systems which return totally ordered and partially ordered answers.

Multilingual resources and applications

Ruslan Mitkov

I chaired LREC session WO13 "Multilingual resources and applications". The papers were well balanced and attracted considerable interest. To start with, Jorge Kinoshita (Escola Politécnica da Universidade de São Paolo, Brazil) presented an approach to grammarless bracketing in an aligned bilingual corpus based on the difference of word sequences in two languages. The second session speaker, Masumi Narita (Ricoh Co., Japan) explained how she constructed a tagged English and Japanese parallel corpus of sample abstracts which was employed in the development of an English abstract writing assistance tool. Next, Marta Villegas (GILCUB, Barcelona) presented on behalf of a team from GILCUB, Instituto di Linguistica Computazionale (Pisa) and Institut d'Estudis Catalans, a procedure for converting the PAROLE-SIMPLE monolingual lexicons into bilingual interrelated lexicons where each word sense of a given language is linked to the pertinent sense of the right words in one or two target lexicons. Finally, Elliott Macklovitch (Laboratoire RALI, Université de Montréal) reported on the Web-based version of TransSearch which over the last three years has given Internet users a free access to a large English-French translation database made up of Canadian parliamentary debates.

Each paper was followed by a question session, the most questions attracting Elliott Macklovitch's presentation.

LREC Workshops Summaries

XLDB - Very Large Telephone Speech Databases

Christoph Draxler, University of Munich, Germany

The XLDB workshop on Very Large Telephone Speech Databases was held as a pre-conference satellite event to LREC 2000 in Athens. The workshop programme consisted of one keynote speech, nine oral paper presentations and a tutorial.

The keynote speech was held by Christian Dugast, head of the Paris office of Nuance Communications. In his presentation, he convincingly argued for the use of speech technology to automate teleservices. He claimed that speech databases are a good starting point for bootstrapping speech technology products, but that true performance increases come solely from the real-world data gathered in running systems. Despite their limited use, speech databases are a very important resource. In the near future, the two most important requirements speech databases must meet are multi-linguality and non-native speakers.

In the first paper presentation, Harald Höge of SIEMENS AG, discussed the technicalities of speech database production and coined the term "Speech Database Technology". He focused on the requirements concerning speaker demographics, dialects vs. languages, and acoustical environments.

The following four papers gave an overview of speech data collections. Bruce Millar of the Australian National University outlined the economical basis, the variety of languages, and the status of telephone speech data collections in Oceania. Asuncion Moreno of the Catalonian Polytechnical University in Barcelona presented SALA, a SpeechDat-type data collection in Latin America. This region is economically very interesting, but the project faced enormous technological and political obstacles. In his presentation of SpeechDat-E, Petr Pollak of the Czech Technical University in Prague summarised the experiences of collecting telephone Speech in Eastern Europe; he discussed in detail the problems of applying the SpeechDat specifications to the Slavic languages, e.g. for names and numbers. Finally, Gael Richard of Matra Communication (now L&H France) presented SpeechDat-Car, a large speech database collection with synchronous high band with recordings in a car and GSM recordings via mobile phone.

The annotation of large speech databases is a time-consuming and expensive task. In his tutorial session, Christoph Draxler of the University of Munich introduced WWWTranscribe, a web-based annotation tool. Due to its client-server architecture, it is completely platform independent and can be adapted to different annotation systems easily.

In SpeechDat, validation by a project partner not involved in speech collection is the method of choice to guarantee a certain standard of quality and thus defines an exchange value for a database. Henk van den Heuvel of the Dutch Speech Expertise Centre SPEX is responsible for the validation in the SpeechDat projects, and he described the validation process itself and the lessons learnt from validation SpeechDat databases.

SpeechDat presents a unique opportunity to evaluate the performance of speech recognisers across many languages. In the COST 249 project, led by Finn-Tore Johansen of Telenor, a reference recogniser based on HTK has been defined, and in different labs recognisers have been trained on a subset of the SpeechDat-II databases.

The last two papers contained extensions to SpeechDat to new languages: Catalan and Austrian German. For the Catalan database, first results of recognition experiments, performed by the Catalonian Polytechnical University in Barcelona, were presented. In Austria, SpeechDat databases were collected both via the fixed and the mobile telephone network by the Telecommunications Research Centre in Vienna.

The workshop attracted a total of 30 participants. From the organiser's point of view the workshop was rather Euro-centric. However, especially participants from outside Europe found the workshop to be very instructive and to provide an excellent overview of the ongoing work in the area of large telephone speech databases.

Christoph Draxler
Department of Phonetics and Speech Communication, Ludwig-Maximilian University Munich, Schellingstr. 3, D 80799 Munich
Tel: +49 +89 28669968
Fax: +49 +89 280 0362
Email:draxler@phonetik.uni-muenchen.de
http://gspot.phonetik.uni-muenchen.de/draxler.html

Meta-Descriptions for Multi-media Language Resources

P. Wittenburg, H. Brugman, D. Broeder

Contributions

First the White Paper of the meta-description initiative within the EAGLES/ISLE project was presented by Peter Wittenburg. It argues that it is time to create a structured sub-space for the language resource community in the Internet to easily locate resources of interest. This will be achieved by describing the resources with meta-descriptions (header information), making them available for structured searches and by using them to create browsable hierarchies. The White Paper describes the problems to be solved such as defining a widely accepted set of meta-elements and describes an organisational structure to achieve that. Then Henry Thompson gave his view about meta-data. His great perspective is that all data is open data in the Internet and that it is machine exploitable. In this scenario XML is one of the key components for representing "simple tree-structured" documents. Finally he argued that we need many projects, since we still don't have a stable "ontology" of the field. A key for the success of meta-descriptions will be the control of its quality. Steven Berman discussed the requirements with respect to meta-data form the point of searching. He made a distinction between the needs for local information and web-based information, since having searching agents crawling through the web to find hits is still a very expensive operation. He also suggests to come to a meta-data standard where meta-data is available as attribute-value pairs.

After these more general talks three talks were given which presented the ideas from a users perspective. Caroline Wilners described the corpus related work at Lund university and argued that even for internal purposes the availability of a browsable & searchable universe of meta-descriptions would help the researchers a lot. The current practice is that no one knows exactly which corpora are available and in what status they are. Nelleke Oostdijk explained why the Dutch National Corpus project relies on structured meta-descriptions to organize the project's data and allow better access to it for arbitrary users. She especially stressed the need of flexibility with respect to meta-descriptions. Pirkko Suihkonen referred to the work at Helsinki university where a web-site was setup to help interested people to get an overview about available corpora. She also presented a highly detailed list of meta-data categories and the meta-elements she needs to describe the resources at Helsinki university and MPI for evolutionary Anthropology.

Daan Broeder presented the meta-description project at the MPI for Psycholinguistics. A unified meta-scheme based on XML syntax is the basis for describing the many resources and for preventing a chaotic situation where only individuals know how to access the resources. He also explained what kind of tools were programmed to create meta-descriptions and browse through the universe of such descriptions. Finally, Khalid Choukri discussed the role of a resource agency like ELRA in distributing language resources and how this role may change over time. Internet will have its great impact, but still existing channels of accessing data via the ELRA catalogue and media distribution will the preferred method by many users.

Discussion

The discussion after the talks and at the end of the session resulted in a number of interesting points:

- Meta-descriptions will only be accepted when a high quality is guaranteed.
- Some people or institutions urgently need methods to prevent a complete chaos where only few individuals know about the state of corpus projects and the ways to access them. In these institutions typically many resources are created continuously.
- Many meta-description related aspects are highly dynamic, i.e. we will need several attempts and projects to fully understand the problems, unify the terminology, and come to a stable state.
- Some argue that it is better not to separate meta-descriptions and annotations. One reason is that the content of the annotations might change such that meta-data is effected. When meta-data and content data is separated it might be difficult to keep the meta-data up-to-date. Another reason may be the enhanced possibilities of search operations which could combine header and body search.
- The meta-descriptions must have a mechanism to allow flexible extensions for sub-communities. The problem with such extensions, however, is how to allow the search engine operate on them and how to inform the user of their existence and meaning.
- All resources should be openly accessible in the Internet.
- How to prevent an endless discussion about meta-elements?
- Are other initiatives such as RDF from W3C of any relevance for the meta-project?

Summary Statement

- We can identify an extreme increase in the number of language resources being produced world-wide. We urgently need ways to capture the knowledge about their content and construction and to make them browsable and searchable by the interested community. Some users from well-known Research institutions have expressed their wish to start the meta-project and soon have a description standard available and tools operating on them. In companies working with many resources it is self-evident to have a database which describes them.
- The community is skeptical whether we will achieve the goal to have all resources freely available in the Internet rather soon. There are too many obstacles which will limit the general accessibility of the resources themselves. However, meta-descriptions could be openly available. In fact, the Talkbank project designed an elaborated access right system which may be taken as an indicator of how sensitive these aspects are.
- Although the authors see the problem which can occur when separating meta- and content information, there are 5 reasons which advise us to go ahead with the meta-description project:

  – There are many resources in native formats such as CHAT. It cannot be seen how all this data will soon be converted to XML-based formats. This means that there is no such hierarchically structured document describing meta-data and content. Separate meta-descriptions can easily be created based on the header information.

  – As already mentioned many resources will not be freely available on the net. Nevertheless, it is very useful for the community to know whether there are some available with certain characteristics and whom to contact to get access.

  – Searching in a web-based meta-universe will be much simpler and much less compute intensive than searching in a universe of the resources themselves.

  – There is an extremely high pressure to start creating a standard for meta-descriptions independent of the question whether they are integrated with the content or not.

  – There is no special problem not to hook up complete resources to the meta-universe, if the meta-description schema is identical and if the tools can cope with this. On the other hand it is a simple operation to extract meta-data from a complete document and integrate it in the meta-universe.

- If separation is done, of course, one has to set up a scheme which allows the provider to automatically adapt the meta-descriptions after the content was changed. Since the meta-descriptions are part of a distributed scheme there is no reason not to maintain them at the places where the resources themselves are stored.
- With respect to combined header and body searches methods can be imagined to allow these even in case of separate meta-descriptions.
- The quality assurance needs an appropriate organizational approach. No one may be allowed to hook up meta-descriptions to the universe without quality check. There has to be a clear system of authorization.
- It was clear that the structure of the meta-descriptions has to cope with flexibility and dynamics. The right technical mechanisms have to be worked out within the project.
- Continuously analyzing the progress of other initiatives such as XML-schemas, RDF, DC, MPEG7 etc. is a must. When it is possible to join or to take profit from these initiatives then the project should do.
- Agencies such as ELRA will be needed to control the quality of the meta-descriptions and to help integration and usage.

As a result of the workshop a Steering Board and a Technical Board for this meta-initiative could be setup and a the SB had its first meeting. An Advisory Board is in the process of being setup.

Comments and questions should be addressed to: ISLE@mpi.nl
http://www.mpi.nl/world/ISLE/

Language Resources and Tools in Educational Applications

Eleni Efthimiou, Institute for Language and Speech Processing, Greece

The workshop on "Language Resources and Tools in educational Applications" was organized by ILSP-Institute for Language and Speech Processing, Department of Educational Technologies and the Department of Computing, School of Electronic Eng., IT and Mathematics, University of Surrey and addressed the technological state of the art, needs and near future perspectives of exploitation of LRs in the context of development of tools and applications for educational purposes.

With the concepts of *Standardization and Reusability* underlying the design and creation of large scale lexica, grammars and corpora, the workshop attempted to touch upon some central questions in respect to whether language resources and tools developed for the Human Language Technology (HLT) sector may be (re)used also in educational applications either in the INTERNET/INTRANET or in the CD-ROM environment.

In order to fulfill this goal, the workshop call had invited papers on topics such as *Internet and/or network applications for educational software, Distance Learning, integration of language resources in multimedia educational environments, evaluation of language resources for educational applications, development of language tools based on language resources, legal aspects and problems in the access and use of available language resources and customization of language resources.*

The result was a program of nine presentations covering a wide spectrum of the above topics including demos (Proceedings available by ELRA), as well as a round table discussion which was planned to close the workshop session.

The workshop managed to bring together specialists from the areas of both language engineering (including theoretical as well as computational linguists and computer engineers) and multimedia technologies, with experience in the creation of educational software. The working group addressed a number of issues related to innovative and reflective approaches to the exploitation, integration and evaluation of LRs in respect to educational applications not only by means of the paper presentations but also with significant contribution from a very active audience during the round table discussion.

The content of the round table discussion was divided into four sections: a) current technical aspects of developing educational software, b) positive vs. negative aspects of (over)using electronic means in the educational process, c) reported reactions by the side of young users in respect to acceptance or rejection of design principles of various tested educational products, d) perspectives for the future; how far can we go with the CD-ROM, Network and Internet environment? At this point special notice was given to standardization attempts provided through mechanisms such as the IMS and the IEEE Learning Technologies Standards Committee.

In general, the workshop achieved its primary goal which was to contribute to the exchange of ideas and experience and add to knowledge and insight in respect to its relevant domains (theoretical and best practice). It is encouraging that a big part of the audience expressed the wish to have a workshop on the addressed issues repeated on some sort of permanent basis.

Eleni Efthimiou
Institute for Language and Speech Processing
Artemidos & Epidavrou Str.
Paradisos Amarousiou
151 25 Athens, Greece
Email: eleni_e@ilsp.gr

Developing Language Resources for Minority Languages: Reusability and Strategic Priorities

Bojan Petek, University of Ljubljana, Slovenia

Developing Language Resources for Minority Languages: Reusability and Strategic Priorities was one of the ten pre-LREC 2000 satellite workshops, organized on 30 May 2000 in Athens, Greece. The workshop programme included four invited talks, 14 poster presentations, and the first meeting of the International Speech Communication Association Special Interest Group Speech And Language Technologies for MInority Languages (ISCA SALTMIL SIG). 39 participants registered for the workshop from various countries: Australia (1), Austria (1), Canada (1), Czech Republic (1), Estonia (1), France (2), Germany (1), Greece (3), Italy (2), Japan (1), Korea (1), Malta (1), Netherlands (1), Slovenia (2), Spain (7), Sweden (1), Switzerland (1), UK (3), and USA (8).

Kepa Sarasola, University of the Basque Country, reported on twelve years of experience in the development of Human Language Technologies (HLT) for the Basque language. Since minority languages usually experience serious challenges with respect to machine readable linguistic resources as well as lack of critical mass of human and research resources, he stressed the difference between development of the HLTs for minority and prevalent languages. His proposal included detailed long-term strategy divided into three phases: foundations, tools and applications. Each of the phases has been further subdivided in accordance to the proposed work on the minority language lexicon, morphology, syntax, semantics and speech. Additionally, he discussed what the HLTs for minority languages should pay special attention to, eg, as wide availability of resources and tools as possible.

Harold Somers, University of Manchester Institute of Science and Technology, UK, concentrated on computational resources and development of new language engineering resources for non-indigenous minority languages (NIMLs). He stressed that in order to aid translators in working with NIMLs there exists a big lack of computational resources. Discussed were ways and means to overcome this problem in the short term.

Steven Bird, Linguistic Data Consortium, USA, presented Linguistic Exploration as a mode of investigation in computational linguistics. Languages under study may range from the undescribed to the very well studied with the aim of generating reusable language resources, new tools, and continuously evolving datasets. This avenue of research is further detailed at URL http://www.ldc.upenn.edu/exploration/ .

Bojan Petek, University of Ljubljana, Slovenia, focused on the funding issues for research on less prevalent languages. The presentation stressed a necessity to search for funding beyond the current EU mainstream programmes, eg, the 5th Framework Programme. Proposals extended beyond a narrow view of financial funding issue itself, and included a call for intensive international collaboration in education and research. Lively international student exchange was proposed to ensure critical mass of qualified young scientists capable of narrowing the gap in technological maturity between the less prevalent and major languages. Additionally, discussed were innovative project proposals to philanthropic and less prevalent language funding organizations in a concerted manner, in order to overcome various problems such as the lack of appropriate resources and insufficient information technology infrastructure.

The poster presentations stressed particular HLT-related research issues considering the Galician, Catalan, Basque, Maltese, Breton, French Creole, Finnish Romani, Czech, Tatar, Nenet, and the Graecanic Dialect in Southern Italy.

After the presentations the first ISCA SALTMIL SIG meeting contained the following agenda:

- Presentation of past activities in 1999-2000.
- Proposals for future activities and discussion on them.
- Election of the new SALTMIL committee.

Past activities 1999-2000: It was pointed out that the ISCA board approved the SALTMIL SIG proposal submitted by the Founding Committee Members (Briony Williams (Univ. of Edinburgh), Climent Nadeu (UPC, Barcelona) and Donncha Ó'Cróinín (ITE, Dublin)) and transferred Euro 1000 to the SALTMIL SIG. Funds were deposited on a bank account in Ireland by Donncha Ó'Cróinín. Other activities included creation of the SALTMIL website (now at http://isl.ntftex.uni-lj.si/SALTMIL/), the SALTMIL electronic discussion list (at http://www.egroups.com) and the organization of the LREC-2000 Workshop.

Proposed future activities: These included background information and discussion of plans to:

- Maintain a database of active researchers.
- Organize conferences and workshops with a focus on minority languages.
- Intensify international collaboration in education and research.
- Promote active international student exchange.
- Issue a call for the best student project or essay.
- Design and coordinate proposals for joint projects in minority languages.
- Initiate collaboration with the ACL, IEEE, ACM.

Election of the new SALTMIL Committee: Unfortunately, two of the SALTMIL founding members (Briony Williams, Climent Nadeu) resigned as SALTMIL Committee Members. Since Donncha Ó'Cróinín could not be present at the meeting it was decided to fill the positions of the SALTMIL Committee at a later date, eg, through voting on the SALTMIL electronic discussion list.

Discussion: In the discussion that followed, Dafydd Gibbon (University of Bielefeld) pointed out that the recent activities under the COCOSDA project are expected to address explicitly the local languages of Africa. He proposed collaborating on the topic. David Graff (LDC, University of Pennsylvania) offered the LDC infrastructure and experience to enable acquisition of databases through the "Voice of America" broadcasts in minority languages. Funding for such projects at the moment, however, cannot be provided by the LDC itself. Additionally, Jeff Allen (ELRA) offered to distribute any existing minority language resources, and welcomed submissions of such material. Lori Levin (Language Technologies Institute, CMU) offered support to visiting student researchers, through the provision of expertise on the design and implementation of speech processing systems. However, the funding of such activities remains a challenging issue to be addressed in the future.

Bojan Petek
University of Ljubljana
Faculty of Natural Sciences and
Engineering
Snezniska 5
1000 Ljubljana
Slovenia
Email: bojan.petek@uni-lj.si

The Evaluation Paradigm

Patrick Paroubek, LIMSI-CNRS, France

Among the workshops that took place before the second Language and Resources Evaluation Conference (Athens May 31th- June 2nd 2000), the one of the morning of Tuesday, May 30th, entitled "Using Evaluation within HLT Programs: Results and Trends" and organized jointly by the CLASS team in charge of evaluation, the European Network of Excellence in Human Language Technologies and the European association ELRA, was aiming at drawing a picture of the current status of the evaluation paradigm in Language Technology programs, such as the ones of the European Commission, the north-American programs of the Darpa and NIST, the programs of the recent NII (National Institute for Informatics) in Japan, which now provides a unifying framework for the domain in this country, or in national or transnational programs like Senseval. There were 28 registered participants and around 60 people present in the room, among which one could notice the inventor of Tree Adjoining Grammars, Prof. Aravind Joshi of University of Pennsylvania. On the other hand, no representative of the European Commission attended the meeting.

In his introductory speech, Joseph Mariani recalled what was at stake for the development of evaluation in Natural Language Processing (identifying new research direction, technological and scientific progress, better visibility for the domain). He also commented on the large diversity of types (experiment reports, evaluation campaign reports, theoretical and prospective studies) and topics offered by the workshop presentations. The theme of the first session, addressed by 9 presentations, was experience reports. Their respective titles were: *Evaluating the Coverage of LTAGs on Annotated Corpora* (Fei Xia and Martha Palmer, IRCS/UPENN (USA)), *Comparing Test Suite-based Evaluation and Corpus-based Evaluation of a Wide Coverage Grammar for English* (Rashmi Prasa and Anoop Sarkar, IRCS/UPENN (USA)), *Evaluating a Multi-Word Term Indexing System: Method, Implementation and Report* (Béatrice Daille, IRIN/U. Nantes (France)), *Evaluation of the Machine Translation of Financial*

Documents (Rémi Zajac, CRL/NMSU (USA)), *Amaryllis: an Evaluation-based program for Text Retrieval in French* (Stéphane Chaudiron and Laurent Schmidt, Ministère de la Recherche and INIST (France)), *Evaluation of Document Retrieval Systems* (Claude de Loupy and Patrice Bellot, LIA/U. Avignon (France)), *Implementing a Question Answering Evaluation* (Ellen Voorhees and Dawn Tice, NIST (USA)), *IR/IE/Summarization Evaluation Projects in Japan* (Kyo Kageura, NACSIS (Japon)) and lastly *Is That a Good Spoken Language Dialogue System?* (Ole Bernsen and Laila Dybkjaer, NIS/U. Southern Denmark (Danemark)). Unfortunately Monika Höge (U. Helsinki/Finlande) could not be there to present her paper entitled: *A Framework for the Quantitative and Qualitative Evaluation of Translator's Aids Systems*. During this session, the discussions on parsing were of a rather technical nature, while Text Retrieval and its related issues got the lion's share with lots of comments from the audience on methodological or technical aspects as well as infrastructural ones. From the debates, it seems that automatic translation evaluation makes its come-back, particularly in the context of Information Retrieval; and that evaluation of Spoken Language Dialog Systems is still facing the same challenges despites ambitious programs like Communicator in the United-States or Smartkom (Verbmobil follow-up) in Germany.

The next session was more theory-oriented with the following 3 presentations : *Categorical Data-Specification for Control Task Formalization and Validation in Quantitative Black Box Evaluation* (Patrick Paroubek, Limsi/CNRS (France)), *Reading Comprehension and Question-Answering New Evaluation Paradigms for Human Language Technology* (Lynette Hirschman/MITRE (USA)), and *To Validate or Not To Validate? - Some Difficulties for a Scientific Evaluation of Natural Language Processing* (Gérard Sabah/Limsi/CNRS (France)). A more philosophical orientation given to the last presentation was well received by the audience, which also showed its interest for the pragmatism and prospects offered by the idea of reusing reading comprehension tests for evaluation. The last session took the form of a panel session with: Donna Harman (NIST (USA)), Stéphane Chaudiron (MR (France)), Adam Kilgariff (ITRI (Grande-Bretagne)), Édouard Geoffrois (DGA (France)), Khalid Choukri (ELRA), Gerhart Budin (U. Vienna (Autriche), SALT project), Rémi Zajac (New Mexico State University (USA), Transaccount project), Lazaros Polymnenakos (IBM, (Grèce), Catch-2004 project). Each panelist presented briefly his views on evaluation before the general discussion with the public took place. The issues raised during the debates were: the opposition between Technology evaluation and Usage Evaluation (both kinds appear to be complementary), the well-foundedness of a European infrastructure for evaluation (in particular in the context of ongoing cooperation with the United-States), copyrights and access to resources, portability (across languages), evaluation models for terminology, applications and packages for evaluation.

In his concluding speech, Joseph Mariani said that he had seen the workshop as renewed proof of the fully scientific nature of evaluation in Language Engineering, which requires solving both technological and theoretical issues; he also commented on the participation of north-American researchers, pioneers of the domain, Donna Harman and Dave Pallett from NIST, saying that thanks to their efforts, Language Technology had evolved from the Middle-Ages to the Renaissance because they had brought the means to objectively measure advances and progress in the field.

Patrick Paroubek
Spoken Language Processing Group /
Human-Machine Communication Dept
LIMSI-CNRS
Bâtiment 508, Université Paris XI
BP 133, 91403 Orsay Cedex
France
Tel: +33 (0)1 69 85 81 91
Fax: +33 (0)1 69 85 80 88
Email: pap@limsi.fr

LREC Opening Session Speeches

George Carayannis

Institute for Language and Speech Processing (ILSP), Greece

Καλημερα σας, Good morning

Thank you for coming so many from so many countries in order to participate to the 2nd International Conference on Language Resources and Evaluation.

My name is George Carayannis and I am the Chairman of the organizing committee. On behalf of this Committee I would like to say to all of you welcome to Athens and welcome to Zappeion, which becomes for some days your place. I am very happy to see here many long-term friends.

As you know I am the Director of ILSP, which is one of the institutions involved in the organisation of the conference. ILSP is a specialized Institute working under the auspices of the Greek Ministry of Development and the General Secretariat for R&T. The mission of ILSP is to develop innovative aspects of languages engineering, methods, tools, platforms which are useful in the modern information society era. Part of its mission is also to develop the necessary infrastructure and language dependent tools related to the Greek language. We give emphasis to language resources and have developed all kinds of resources for Greek. There are five departments: Electronic Lexicography, Machine Translation, Speech Technology, Educational Technology and Language Applications for the Modern Office.

An important application for our tools and algorithm is educational software for languages learning. ILSP has an industrial orientation and is active both in the framework of national and international projects. In the exhibition area you can have more information about ILSP and its activities.

In Greece we have a proverb: To learn what life means, you have to build your house and to marry your children. I think that we can add, a third item, that is you need to organize a conference. We learnt a lot through the preparation of LREC-2000. As you know a conference is the fruit of the work of many persons. I would like to thank all of you who have contributed with dedication in the organization of LREC-2000.

LREC-2000 has similar style with the first

LREC in Granada. Both the microstructure and the scientific content are similar. The Athens organizing committee has benefited from the Granada experience. I would like to thank my colleagues from the Programme Committee who have already had the Granada experience, for their help.

The concept of the LREC conference is being tested for the second time here in Athens. It is very reconforting to see that you are so many to attend the conference. Your participation creates a very rich event not only in the number of scientific presentations but also the quality of the topics and their innovative aspects. During the paper selection procedure we were very happy to see many important papers. All of us involved in LE R&D know how useful LR is in system design. Improving the various LR practices is equivalent to improving the LE system themselves.

The creation, the maintenance and the distribution of resources have their secrets, and ELRA has accumulated this specific know-how.

I would like to focus on some practical issues. You have an extensive programme in your hands. We hope that the programme will be followed without major modifications. Slight modifications of the programme will be announced promptly on the conference messages board on the left handside of the main entrance. There is a major modification. This evening we have a welcome reception here at the PERISTYLION area, which is not announced in the programme. You are all invited. Don't forget the Gala dinner at the Hilton hotel after the Closing Session the 2nd of June. I hope you will be all of you present for the final comments and the fiesta.

There is an Internet room available on the left handside of the main entrance. The Internet set-up brings a new look to this old historical building. It was offered by our sponsors COMPAQ - GE-CAPITAL and GRNET. Thank you very much for this offer.

There is a post office with various post services inside the secretariat room on the right handside of the main entrance. You can send your proceedings to your office, exchange money, etc.

Poster sessions are organized in the Peristylion. Peristylion has a symbolic value in Zappeion. It is consider to be the place where the big achievements have been presented. I hope you will enjoy this place. Coffee breaks are offered in the Peristylion area as well. You know that we have an exhibition. You will find there many important exhibitors.

Finally there is a post conference workshop on the "simple" programme on Saturday 3 of June in concurrence with an excursion in Delphi.

There are two important points.

- Timing is very strict. No presentation can be longer than 20 minutes. I would like to ask to the session chairpersons to be precise.
- Request to all the participants. Try to meet with your session chairperson 5 minutes before the session starts. It is important both for the session chair and the technical staff to know your presentation requirements.

I think you are fully informed now about the various events. The organising committee staff is here: you can recognize the members of the staff by the green contour of their badge. We all believe in the LREC-2000 success and we are continuously available to solve any problem.

I would like to thank all our sponsors who gave to us the possibility to benefit from some comfort and quality of life COMPAQ- GE Capital, GRNET.

I hope that LREC-2000 will initiate many new research initiatives and many new collaborations.

Thank you.

Petros Efthimiou

Greek Minister of Education, Greece

I am especially glad that I have been given the opportunity to inaugurate your conference such as your workshop can assist the educational process, through the development of new tools and systems, which will provide improved human-machine communication solutions and fascinate young pupils, thus leading to a greater enjoyment of educational software.

In politics, we have come to understand that computer technology after the user-friendly ergonomy era, will begin a natural interactivity era which is being prepared by Human Language Technology.

It is impressive that computers are learning to communicate by using pieces of human speech. It is also impressive that computers are able to combine language rules with the frequencies of occurrence of specific linguistic types in order to understand different messages just as man does.

I am responsible for a Ministry, which possesses substantial collections of different types of corpora from a variety of sublanguages and text types. They comprise of pupils' books and notes from Primary School through to University. We are currently in the pleasant process of converting these into electronic format in order to include all this knowledge in the electronic libraries, which we are creating.

Perhaps these texts will be of use to you some day in your research. We are also planning the creation of electronic glossaries and terminological dictionaries for the educational process, something, which may also be of use to you. Language resources you know are constantly being systematised and evaluated. I feel that we in the Ministries of Education of various countries, are living parallel lives with your scientific community and ELRA (European Language Resources Association) as we also collect language resources and evaluate them.

In Greece we are extremely interested in Human Language Technology developments and as you already know we have various research organisations which have been active in this field. Our interest in this technology stems from two reasons.

The first is linked to the policy of our government. We seek the solutions which will allow us to make the organisational leaps necessary for the alignment with other European economics within the framework of the European Union and in light of the Single European Currency objective.

The second is related to our policy on issues concerning the Greek language, aiming to secure its presence and its participation throughout the creation of the Information Society. Thus we welcome the European Union political directive calling for equal development opportunities in the Information Society for all European languages. We are especially pleased that Human Language Technology is included in the 5th Framework Programme both as a core technology and as an applications technology.

It seems that the time when multilinguality could be considered as an impediment to the communication between European peoples or as an obstacle to trade is now ending due to effective Language Education and the development of Human Language Technology development.

In Greece we are beginning to speak a minimum of three languages (our mother tongue included) and it seems that this will become the European standard of the new millennium. Foreign languages are introduced at the appropriate times in Greek schools; the second language is introduced in Primary School and the third in High School, while an effort is made to encourage foreign language learning with the use of multimedia software.

Due to the development of Language Technology, the matching of linguistic codes will be facilitated so that even cases of impossible communication between people in the past will become possible potential solutions are found even when they speak no common language. We are receiving positive messages from the translation technology sectors. With much work, the quality of translation is improving, especially in texts of a technical and managerial nature. It seems that Language Technology will contribute in improving the accuracy of our communication.

"Speech Interfaces" hold a prime position among Language Technologies. I hope that soon we will have what is known as "natural interactivity", which will improve the quality of life for information technology users. I believe that these interfaces will make a significant contribution to the reduction of what is known as "information technology illiteracy". Because many users are discouraged by the potential difficulties of using a computer or a keyboard, they may be more willing to use a computer enhanced with a speech interface which will provide direct communication. I have the belief that speech interfaces will become soon reality due to the lengthy and intensive research efforts in your laboratories, which are slowly converted, to applications. I believe that if this complex technology becomes available for all European languages, it will be a great achievement and will allow to a great extent, improved organization within Europe, both for the creation of texts and the management and extraction of information. Now that the World Wide Web has become the transporter of all the information and knowledge used by man, we must deal with the organisation of this information and the retrieval of the correct information with specialised technologies. I know that Language Technology is at the centre of these technologies in order to improve the efficiency of exploitation of the World Wide Web.

Therefore, due to the advances made, only the positive aspects of multilinguality will remain in our continent: It is part of our cultural heritage and provides a rich diversity, which we hope to preserve. Thus the European Union will posess a variety in its expression and communication, as well as in its literary creations.

As I have already mentioned, we follow a specific political direction for the Greek language. It should be a language of the information technology of tomorrow, it should be present in the World Wide Web, easily translated, recognised, in short, and it should be "spoken" in the Internet environment. To this end, a series of software tools are being developed for the effective presence of Greek in the human-machine communication field as well as for language education.

The political directions of the Greek Government on Language Technology issues in the last decade have been:

a) Participation in translation programmes of the European Union

b) Support for the creation of the critical mass of scientists carrying out research on language and speech issues

c) Encouragement of the participation of Greek laboratories in language and speech programmes of the European Union with the sponsorship of the Ministry of Development (what is known as "matching funds")

d) Creation of National Programmes by the General Secretariat for Research and Technology, funded by the structural funds of the European Union and the Greek Ministry of Finances

An important factor contributing to the success of technology is cooperation, and in the Language Technology field we seek the close collaboration of research centres in Europe and other countries. Especially in research issues related to the Language Technology field, the funding which is required exceeds what we are able to provide. We therefore require the cooperation of established laboratories in these areas in order to build systems for the Greek language based on specific core technologies.

The Greek government is making an attempt to increase the research budget and the participation of the private sector in the research budget. I hope that in the coming years we will be able to improve the percentages of GNP which are spent for research and development and especially for new technologies, information technology and therefore language technologies.

On behalf of the Greek Research and Educational Community, I would like to thank you for coming to Athens and to open your Conference.

I wish you success in your work during this conference.

Stefano Stefanile

Ambassador of Italy, Greece

Ladies and Gentlemen,

After the introductory words in Greek, I have the pleasure of adding a few words in English, the language known from the majority of you.

I do not need to insist on the potential input of Human Language Technology in the framework of the rapidly evolving and all-pervasive Information Society, already very clearly and authoritatively described by His Excellency the Minister of Education of Greece.

We, in Italy, entirely share this view, and in particular we agree on the need of fostering international co-operation, an essential condition for the development of a truly multilingual, cohesive Information Society.

The issue here goes beyond economic and business competitiveness.

Languages and cultures are linked on many levels. If the modes of communications are restricted, we shall arbitrarily inhibit the participation of the full range of human inspiration in the Information Society. This is implicitly a threat to one of our most valuable human assets, our diversity, both linguistic and cultural. The only way to avoid this danger is to take the necessary measures in order to support multilinguality.

Authoritative sources have already warned that languages for which Language Technologies are not adequately developed run the risk of losing their status as media of communication within the electronic sphere.

The Commission of the European Union has already taken important initiatives in the field of Human Language Technology, and its efforts should be complemented by national activities in the member Countries, in particular for what concerns Language Resources.

The availability of language resources (LR) is the single most important condition for the extension of language technology to different languages: language resources, in fact, provide to systems the specific knowledge for dealing with a language and its relation with the other languages.

Language resources are the most expensive component in any language technology system. Today, for most languages, only embryonic nuclei of LR exist, which cannot be effectively used in real systems without a substantial enlargement of their coverage.

To make this a reality, duplication of effort is a luxury we cannot afford. We must ensure and enhance reusability of resources as they are developed. We must exploit existing LR and the technical knowledge specific to them. Wherever possible, we must look to derive maximum advantage from economies of scale.

EAGLES/ISLE and ELRA are notable examples of initiatives which have as their mission, in this framework, respectively the promotion of standardization efforts and the design and execution of an overall distribution policy for Language Resources.

And language resources are an indispensable part of the infrastructure. It follows from this that they should be made available, in time, for as many languages as possible, in the public domain.

It is urgent and necessary that International Organizations assign a clear priority to the development of Language Resources, and that different countries co-ordinate actions between them and with the international authorities.

The Italian Ministry for Universities and Research in Science and Technology has recently approved a proposal for a national programme in the field of Language Resources, presented by a group promoted by the Italian Ministry of Telecommunication, and co-ordinated by Professor Antonio Zampolli, Chair of this Conference, and his Institute, the Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche.

The Group, formed by representatives of various Ministries, research organizations, universities, professional associations, industries, service providers, public administrations, has recognized the need for Language Resources to be available in Italian as the most urgent priority for the Italian research and development community. On this basis, it has established the general lines for provision of an adequate range of language resources for Italian. It will develop annotated corpora, mono- and multi-lingual, for written and spoken language. It will also pursue the development of innovative methods to extract from them new linguistic knowledge. It will develop structured lexical knowledge bases to include phonological, morphological, syntactic and semantic information. There will be grammars developed and also tools to assist their use in applications. It shall also elaborate practical methods to transfer language resources and basic components from the technology providers to products and services developers.

Although these are for the most part technical tasks, they will be undertaken with full regard to the Italian cultural heritage.

But the trends we are talking of will extend well beyond the confines of the European Union. So solutions to the challenges of multilinguality need to be planned globally too.

This leads us to the need for international co-operation. It will be a key factor for the success of this endeavour.

Therefore, we consider this Second International Conference on Language Resources and Evaluation a most timely event, and one of key importance.

We wish to all participants a very successful Conference, not only in discussing the state of the art and future research and development directions, but also in discussing opportunity, promoting concrete actions, planning initiatives for ensuring the international co-operation needed to enable Human Language Technology answering to the expectations of our Society and to the epochal challenge we are all facing.

Before concluding, I want to express our gratitude to the Greek Authorities, to ELRA, to the Institute for Language and Speech Processing, to Professor Zampolli and his colleagues of the Program Committee for their effort to organize this very important event.

We trust that this Conference will be a major occasion for stimulating and fostering international co-operation in this field of strategic relevance for our future.

I have the privilege and pleasure to add to my words the message of welcome sent to the participants to LREC by Professor Tullio De Mauro, the Italian Minister of Public Education and an internationally well-known linguist.

"E' con particolare piacere che mi rivolgo ai partecipanti del congresso che si apre oggi ad Atene. Impegni di carattere istituzionale mi impediscono di essere presente, come avrei voluto.

Gli argomenti che verranno discussi nel corso dei lavori sono per me di grande interesse. E' un interesse non solo doverosamente istituzionale. A questi temi ho dedicato la mi attività di ricerca prima di essere chiamato qualche settimana fa a ricoprire l'incarico di Ministro dell'Istruzione.

Vorrei quindi inviare a tutti i presenti gli auguri più cordiali di buon lavoro, come collega più ancora che come titolare del Ministero Italiano della Pubblica Istruzione. I risultati di questo incontro internazionale, a due anni dalla prima edizione a Granada, segneranno una tappa importante negli studi sulle risorse linguistiche".

The meaning of this message, can be formulated in English as follows:

"It is with a particular pleasure that I address the participants of the Conference inaugurated today in Athens. Some engagements linked to my governmental duties prevent me to attend it, as I wished I did.

The topics which will be discussed during the Conference have a very strong interest for me.

Obviously, this interest is not only linked to my institutional duties. I devoted my professional research activity to these topics before being called, a few weeks ago, to hold the office of Minister of Education.

I wish therefore to send my best wishes of good work to all those who are present, not only as the holder of the Italian Ministry of Public Education, but even more as a colleague. The results of this international meeting, two years after its first edition in Granada, will mark a very important step for the advancement of the research and studies on Language Resources".

Roberto Cencioni, on behalf of Vicente Parajon-Collada

Deputy Director of DG XIII of the European Commission

Mr Minister, Mr Ambassador, dear members of the organising committee, Ladies and gentlemen,

INTRODUCTION

I am especially glad to participate in the opening session of LREC 2000, as two years ago I had the pleasure to participate in the first edition of what has since become a widely recognised and truly international event.

This year several major conferences devoted to written and spoken language research, technology and applications are scheduled to take place in Europe and elsewhere. They will without doubt show the constant growth and dynamic of what has emerged in recent years as a socially and economically relevant domain, which mobilises in Europe only some 10 000 researchers and many more professionals.

As the RIAO conference held a couple of months ago in Paris clearly showed, speech and language technologies play a crucial role in multimedia information access and management. Upcoming international events in the areas of computational linguistics and speech technology will highlight recent advances and ongoing developments, hard research problems still awaiting solution, and further opportunities for multinational collaborations, thus contributing to the consolidation of a truly global research space in your areas of work.

OVERALL SCENE

The incredibly fast development of the Internet, both in Europe and worldwide, the emergence of virtually universal mobile communications, the promise of faster and more versatile multimedia appliances, the dramatic rate at which electronic commerce is developing in all our countries … set the scene against which both researchers and developers must set their own agenda, and assess their progress and results.

As the number of mobiles and other communicating appliances exceeds that of Internet PCs, the Web becomes a *multilingual and cross-cultural space*. Industry analysts suggest that the number of non-English speaking Internauts is expected to reach 70% of the total population within the next few years.

As most of you know already, the European Commission is committed since the mid 1970s to fostering both basic and applied research and operational applications, in a wide range of speech and language areas that you are going to review in the next three days.

Over the last few years, virtually all of these activities have been *regrouped* within my directorate generate, the DG for the Information Society, where they represent an integral part of our research and market orientated programmes.

It can be estimated that the total European spending in language and speech projects amounted to more than 170 million euro between 1994 and 1999. Some 110 EU sponsored projects have been launched since 1996.

Let me now concentrate on three themes that are especially relevant for this conference: 1) the IST programme, 2) the eEurope initiative and one of its constituent programmes - eContent, and finally 3) how I see *language resources* within European programmes.

IST PROGRAMME

The Information Society Technologies programme (IST in short), is the single largest R&D programme under the 5th framework, with a total budget of 3,6 billion euro over 4 years.

IST provides more room than ever for truly global collaborations in the field of information and communication technologies. Most European countries, within and without the Union, are *fully integrated* within the programme. Bilateral agreements exist or are being finalised with many third countries, allowing both companies and research centres to participate in European projects on an equal footing with their EU counterparts.

The broad geographical scope of the IST programme is witnessed by last year's call for proposals. In the Human Language Technologies field alone, project proposals were submitted by almost 800 organisations established in 30 different countries.

While there is still much room for improvement, especially regarding a fuller integration of our colleagues from candidate member states, we are heading towards an open, border-less research space, where laboratories from Europe, America and Asia will co-operate more closely than ever.

A concrete example of such opportunities are the projects jointly funded towards the end of last year by the Commission and the National Science Foundation of the USA, setting a precedent that other research communities now intend to exploit.

Thanks to its broad scope and flexible framework, and despite some drawbacks inherent in any large scale operation, the IST programme and more specifically its Human Language Technologies sector, provide financial support for progressive and ambitious multinational endeavours.

This year alone, three calls for proposals are planned in areas such as natural interactivity and multimodality, cross-lingual information management and multilingual communications, which all have a prominent place in the programme of this conference.

Let me finally mention that the 6th framework programme, from 2002 on, will take shape towards the end of this year, and that the Commission will circulate working documents outlining the intended scientific and technical content in the first quarter of next year.

E-EUROPE

I would now like to turn to another major initiative that the Commission is setting in motion in collaboration with the Member States, and which has become known as e-Europe.

The eEurope's action plan, which will be discussed and hopefully adopted by the EU leaders in the coming weeks, calls for an integrated package of actions at national and transnational level, in support of a quick move towards the new knowledge-based economy. The action plan proposes that Member States and the Commission bind themselves to achieving three main objectives by 2002: 1) a cheaper, faster, more secure Internet;

2) investing in people's skills and access; and 3) stimulating the use of Internet.

This last dimension, Stimulating the use of Internet, is especially relevant for your conference. The eEurope plan foresees actions addressing e-commerce, online access to public services, and the provision of European digital content.

Within this framework, and in order to provide additional impetus for the implementation of eEurope, the Commission has approved on 24 May last and forwarded to the European Parliament and Council, a proposal for a multi-annual market stimulation programme addressing the provision of innovative information and transaction services over the Internet and other global networks (including the upcoming mobile multimedia service, UMTS).

The proposed programme, which has become known as eContent, mainly addresses:

- the wider use and commercial exploitation of public sector information, also in multilingual settings,
- and the cross-lingual 'customisation' of content-rich, rich-content digital services, where customisation is taken to mean the adaptation, from a cross-lingual and cross-cultural standpoint, of information content, interfaces and access points.

In the Commission's view, eContent should have a budget of 150 million euro over 5 years, with 60 million euro being earmarked for the linguistic customisation action line, a substantial increase with respect to previous non-research activities, most recently the MLIS programme.

'Customisation' would encompass both market orientated demonstration projects and more horizontal actions, including the provision of multilingual language resources for less widely spoken languages, and the languages of the new accession countries.

Pending a final decision on the main programme, which may be taken towards the end of the year, the Commission has launched in April last a call for preparatory actions, which are intended to test the market, measure the demand and prepare the ground for the follow-on programme and large-scale multinational projects in the years to come.

The call, which will close on 7 July next, invites proposals from private and public sector providers of digital content, suppliers of multilingual services and solutions, and providers of network access and delivery platforms. It is expected to yield ten or so collaborative projects in the areas of multilingual public-sector information and linguistic customisation of commercial, corporate and e-commerce products and services.

LANGUAGE RESOURCES within EUROPEAN PROGRAMMES

Let me know finish by making a few remarks on the place and function of Language Resources within European programmes, now and in the near future.

We all know that electronic collections of linguistic data are a sine-qua-non for researching, building, testing and operating language and speech enabled systems, but also for language learning, technical writing and business communications, translation and localisation, etc.

As you know far too well, the scale of the problem and the number of languages involved are however such that no single organisation, be it the European Commission, can afford to build and maintain a potentially infinite series of data repositories.

The solution therefore resides in selectivity and co-operation, bearing in mind that the yearly budget for European actions will be the region of 30-40 million euro, and that this represents a very small fraction of the total European expenditure in the field.

As for Selectivity, and in the foreseeable future, EU sponsored research actions can be expected to concentrate on 4 themes:

- Research into and models, methods and tools for building truly multilingual resources underpinning language-transfer applications;

- Language resources in support of novel research strands, most notably in the area of multimodal interactions and communications;

- Language resources as an enabler for applied research and technology up-take, especially for Internet services and appliances.

and

- Encoding standards, interchange protocols, open architectures and APIs.

As regards non-research activities, in particular within the eContent programme, EU sponsored actions will concentrate on:

- Multilingual language resources that can readily serve as a basis for cost effective globalisation and localisation processes, to the benefit of content providers and suppliers of language services;

- Especially for those languages where market forces prove insufficient to create the initial momentum and critical mass.

As for research, common standards and protocols, professional and expert forums, etc. will play an important role in the consensus building process.

In order the make such an endeavour feasible and indeed meaningful, we are going to need:

- efficient and reliable data-driven models and tools for acquiring and extracting linguistic knowledge, automatically or semi-automatically;

- widely accepted data capture, representation and labelling protocols, including shared software primitives and tools;

- broad agreements facilitating the distribution and reuse of valuable data sets; and more importantly

- the commitment of all the parties involved (industry, academia, national projects and sponsoring agencies) to extend, maintain and use the resources thus produced.

Finally, and as already mentioned, we need to promote and actively exploit any opportunities arising from international collaborations, within Europe, around Europe and globally, thus sharing risks and costs and exploiting complementary skills and interests.

I wish you a fruitful discussion in the coming days.

Th. Xanthopulos

Rector of NTUA, Greece

1. It is very challenging to take part in the organisation of LREC-2000 Conference. Language Engineering & Language Resources are innovative aspects of the information technology and therefore are very interesting, as our policy is to participate and to contribute in innovative aspects of technology.

NTUA has been always and will continue to be in a pioneering position in research and education in the country.

2. The National Technical University (NTUA) is the oldest and most prestigious educational institution of Greece in the field of technology, and has contributed unceasingly to the country's scientific, technical and economic development since its foundation in 1836. It is closely linked with Greece's struggle for independence, democracy and social progress.

NTUA took its present form in 1917 by special law that organised it into the Higher Schools of Civil Engineers, Mechanical & Electrical Engineers, Chemical Engineers, Surveying Engineers and Architecture. Up to the 1950s, NTUA was the only University in Greece offering degrees in engineering.

Student numbers at NTUA have increased very rapidly in recent years. In 1937, the total number of students registered in all departments was approximately 500. By the early sixties this figure reached 2,000, and today there are more than 7,000 students. This rather sudden increase in conjunction with new requirements in science and technology created urgent needs regarding personnel, equipment and facilities. As far as the faculty is concerned, NTUA has at present a teaching and research staff of approximately 700 members, all holding doctorates (Professors, Associate Professors, Assistant Professors and Lecturers). It is worth pointing out, for the sake of comparison, that in the 1930's there were approximately 40 Professors at all levels, and about 30 assistants.

NTUA is able to select top rated students from all over Greece through highly competitive national entrance examinations. All degrees last for five years (10 semesters) and provide students with a variety of courses and laboratory practices. Students are required to submit a Diploma Thesis before graduation that is usually based on active research work performed by NTUA faculty. The level of study and the standard of the degrees awarded are considerably high. An appreciable number of NTUA graduates are accepted by foreign Universities for doctoral studies and a large percentage settles abroad, engaged in either lecturing or research work.

The academic level of the faculty is exceptionally high, as all of them have studied to an advanced level both in Greece and abroad, published a considerable number of papers in scientific journals and actively participated in sponsored research programmes. Over the years, NTUA researchers established the excellence of the University in international R&D efforts. Currently, NTUA attracts funding for research from National and European sources that place it on the top of all Greek Academic and Research Institutions.

Research is carried out in about 100 laboratories belonging to the various Departments and Sections of the institution. All departments now offer graduate programmes that lead to Doctorates. There are approximately 1100 doctoral students presently enrolled.

Teaching and research activities are carried out in nine Departments.

The annual budget of NTUA is approximately 4 billion Greek Drachmas ($16,000,000). This covers operational costs, laboratory equipment, research, investment in buildings and other works, staff salaries (for approximately 1,400 employees) and other expenses.

In addition R&D funded programs are administrated by the Research Committee with an annual budget of more than 4.5 billion Greek Drachmas ($18,00 0,000). NTUA employs about 1800 researchers in more than 700 R&D projects supported by National and European Union funds.

3. Especially in Information Technology, NTUA is very active, almost 90 Professors teach related issues. Pioneering projects in the country have been led by NTUA. We keep close cooperation with ILSP in the Language Engineering field.

4. NTUA has participated in the organisation effort of LREC-2000 with the persons of its staff:

- Vice Rector Professor Galanis who is a member of the International Advisory Committee
- Professor G. Carayannis, Chairman of the Organising Committee
- Professor G. Papakonstantinou member of the Scientific Committee

5. We will all of us be interested to know the scientific results of your Conference. I wish you success in your work.

Antonio Zampolli

Istituto di Linguistica Computazionale del CNR, Italy

Ladies and Gentlemen,

First of all let me express my warmest gratitude to the Authorities who have honoured our Opening Session, 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 at this second edition of the International Conference on Language Resources and Evaluation (LREC).

The first edition of the Conference, two years ago in Granada, was truly a success, as the number of submissions to the present one clearly indicates.

I hope that this Conference here in Athens 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 on the 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, LR 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.

That poses, as clearly emerged in the discussions in Granada, a number of organisational and policy problems, for a large part yet unsolved.

On the other hand, the provision of adequate LR 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 LR.

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, to the Local Organising Committee, to the International Advisory Board, to the ELDA staff, to the sponsors which have generously contributed to the financial efforts, to the various Organisations which have accepted our invitation to co-sponsor the Conference.

Of course the personnel of the ILSP deserve a particular mention. Under the guidance of Professor George Carayannis, they have dedicated an incredible effort to preparing the Conference.

As you have certainly noticed, the Zappeion Megaron looks very different from the Palacio de Exposiciones y Congresos in Granada. This building is a very prestigious one and the location is splendid. I am told that it is considered as THE venue per excellence for the major official events in Greece. But it has been necessary to adapt it to the needs of as large and complex a Conference as LREC. Our Greek colleagues have spared nor energies nor financial efforts to offer the best logistic facilities compatible with the nature of the building itself.

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 will 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 of resources, joining knowledge and efforts to search for solutions.

I wish all of you a successful Conference and a pleasant stay in Athens.

I hope you will accept with benevolence any inconvenience or problem our organisation might cause to you.

Khalid Choukri

ELRA/ELDA, France

On behalf of the board of ELRA, I am very happy to welcome you at this 2nd LREC.

I will not elaborate on ELRA's mission and services as this is in the leaflet that is in your bag.

I endorse Professor Antonio Zampolli statements and also his congratulations to the ILSP staff for their efforts to organize this conference. I will say that better by the last day of the conference on Friday.

The number and varieties of issues and subjects (including new and emerging ones) to adress during the next days is reflected by the number of participants and the number of presentations. It shows the strength of our field, Language Resources and Evaluation both in academic and industry worlds. This is also reflected by the actions of the European Commission and other funding agencies. I am very glad to see that ELRA is actively contributing to its development including through this forum and the satellite workshops. ELRA through a collaboration and networking effort, is trying to set up and extend the basis of true partnership with other organisations like LDC (USA), GSK (Japan), oriental Cocosda, etc… for the benefit of the field.

A special word of thanks to each of the authors of the 281 papers which are to be presented. In order to better plan for the delivery of the presentations, the program committee has placed 129 papers in oral sessions and 152 papers in poster sessions, taking into consideration the adequacy of the presentations to the oral versus poster communication mode. Selection criteria for oral and poster presentations have been identical.

It is not usual but I would like to extend a special thank to my collegues of the Program Committee, Antonio Zampolli, Nicoletta Calzolari, Joseph Mariani, Bente Maegaard, Harald Hoege, George Carayannis, for the wonderful job they did.

ELRA & ILSP staffs are prepared and willing to make this conference a very successful event, if you need any help please contact us.

Thank you George (Carayannis) for the wonderful event we have ahead of us, thank you all for joining us and enjoy the conference and Athens.

New Resources

Keys: R. = Research use by an academic organisation - RC. = Research use by a commerical organisation - C. = Commercial use

ELRA-L0042 PAROLE Spanish lexicon

The PAROLE Spanish lexicon follows standard PAROLE architecture which includes morphological and syntactic layers. It includes the most frequent words found in a 1 million word corpus, coded according to the PAROLE specifications. The lexicon contains about 22,000 morphological units, of which 12,209 are common nouns, 3,367 verbs, 4,996 adjectives. Closed classed categories are fully covered. The information associated with each morphological unit concerns part-of-speech and subtype, inflection paradigm (with morphosyntactic information for the endings organised in about 132 models), possible stems in relation with the relevant endings, linking with syntactic layer. In the syntactic layer, information regarding subcategorisation for verbs and insertion context for nouns is encoded following the PAROLE model.

	ELRA Members	Non Members
For research use	3,400 Euro	5,100 Euro
For commercial use	9,000 Euro	13,500 Euro

ELRA-S0085 BABEL Bulgarian Database

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

The BABEL consortium included six partners from Central and Eastern Europe (who had the major responsibility of planning and carrying out the recording and labelling) and six from Western Europe (whose role was mainly to advise and in some cases to act as host to BABEL researchers). The five databases collected within the project concern the Bulgarian, Estonian, Hungarian, Polish, and Romanian languages.

The Bulgarian database consists of the basic "common" set which is:

· Many Talker Set: 30 males, 30 females; each to read twice the five blocks of numbers (each of which contains 10 numbers), 3 connected passages and one "filler" passage.

· Few Talker Set: 5 males, 5 females, selected from the above group: each to read 5 times the blocks of numbers, 15 connected passages and 2 "filler" passages, and 5 repetitions of the lists of monosyllables.

· Very Few Talker Set: 1 male, 1 female, selected from Few Talker set: each to read blocks of monosyllables in carrier sentences and five repetitions of the context words.

And the extension part: semi-spontaneous answers to questions: the answers were recorded by the 10 Few Talker Set speakers.

The other languages will be available soon.

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

The EDR Electronic Dictionary

The EDR Electronic Dictionary is a result of combining the information of conventional paper-based Japanese and English dictionaries, thesauri and corpora. The words treated in the dictionary include basic or commonly used words and technical terms from the field of information processing. The EDR Electronic Dictionary includes 6 monolingual dictionaries and 2 multilingual dictionaries. Each subdictionary shares the same basic design, including the record number, headword information, co-occurrence constituent information, syntactic information, semantic information, co-occurrence situation information, and management information. A basic descriptive format makes use of a portion of SGML (Standard Generalized Mark-up Language); this is not exactly SGML.

ELRA-L0036 The Japanese Word Dictionary

The Japanese Word Dictionary is composed of 260,000 Japanese word records arranged alphabetically according to the Japanese syllabary. Each record of the Japanese Word Dictionary is composed of the record number, headword information, grammatical information, semantic information, pragmatic and supplementary information and management information. The main role of the Japanese Word Dictionary is to describe the correspondence between the Japanese word and the concept represented by the word and to provide the grammatical information for the word when used with the given meaning. Commonly used word are the subject of the Japanese Word Dictionary.

Prices: R. 1,488 EURO RC. 14,884 EURO C. 29,768 EURO

ELRA-L0037 The English Word Dictionary

The English Word Dictionary is composed of 190,000 English word records arranged alphabetically. The record of the English Word Dictionary is composed of the record number, headword information, grammatical information, semantic information, pragmatic and supplementary information and management information. The main role of the English Word Dictionary is to describe the correspondence between the English word and the concept represented by the word and to provide the grammatical information for the word when used with the given meaning. Commonly used words are the subject of the English Word Dictionary.

Prices: R. 1,488 EURO RC. 12,899 EURO C. 25,799 EURO

ELRA-L0038 The Concept Dictionary

The Concept Dictionary provides 400,000 concepts that are made reference to in the Japanese and English Word Dictionaries (ref. ELRA-L0036 and L0037), the Japanese-English and English-Japanese Bilingual Dictionaries (ref. ELRA-M0023 and M0024) as well as in the Japanese and English Co-occurrence Dictionaries (ref. ELRA-L0039 and L0040). The Concept Dictionary is composed of three separate dictionaries: the Headconcept Dictionary, the Concept Classification Dictionary and the Concept Description Dictionary. The Headconcept Dictionary gives a description of each concept in words and the Concept Classification Dictionary contains a classification of concepts that have a super-sub relation. The Concept Description Dictionary provides all other information regarding the relation between concepts.

Prices: R. 1,488 EURO RC. 14,884 EURO C. 29,768 EURO

ELRA-M0023 The Japanese-English Bilingual Dictionary

The Japanese-English Bilingual Dictionary is composed of 230,000 word records arranged alphabetically according to the Japanese syllabary. Records of the Japanese-English Bilingual Dictionary are composed of the record number, headword information, grammatical information, semantic information, English correspondence information and management information. The main role of the Japanese-English Bilingual Dictionary is to provide an English correspondence word for Japanese headwords based on the meaning of the headword.

Prices: R. 1,488 EURO RC. 12,403 EURO C. 24,807 EURO

ELRA-M0024 The English-Japanese Bilingual Dictionary

The English-Japanese Bilingual Dictionary is composed of 160,000 bilingual word records arranged alphabetically according to the headword. The record of the English-Japanese Bilingual Dictionary is composed of the record number, headword information, grammatical information, semantic information, Japanese correspondence information and management information. The main role of the English-Japanese Bilingual Dictionary is to provide the Japanese correspondence word for English headwords based on the meaning of the headword.

Prices: R. 1,488 EURO RC. 12,403 EURO C. 24,807 EURO

ELRA-L0039 The Japanese Co-occurrence Dictionary

The Japanese Co-occurrence Dictionary is composed of 900,000 headphrase notations arranged according to the Japanese syllabary. The phrases are the abstracted portions of actual sentences contained in the EDR Japanese Corpus. The results of the parsing analysis of these sentences indicates that the constituents of the sentence have a dependency structure. That is, the constituents have a governing-dependent relation. It is these constituents that form the headphrases of the Japanese Co-occurrence Dictionary. Records in the Japanese Co-occurrence Dictionary are composed of the record number, headword information, co-occurrence constituent information, syntactic information, semantic information, co-occurrence situation information, and management information. The main role of the Japanese Co-occurrence Dictionary is to show actual examples of how autonomous words are appropriately combined based on the co-occurrence situation information obtained from the Japanese Corpus.

Appendix to the Japanese Co-occurrence Dictionary: The Japanese Corpus:

The Japanese Corpus is composed of records arranged according to EUC (Extended Unix Code). The records of the Japanese Corpus are composed of the record number, sentence information, constituent information, morpheme information, syntactic information, semantic information and management information. The basic role of the Japanese Corpus is first to identify the sentence constituents of sentences, and then to indicate how the constituents combine to form the semantic, syntactic and morphological structure of the sentence using a large number of actual examples as the source data.

Prices: R. 1,488 EURO RC. 13,892 EURO C. 27,783 EURO

ELRA-L0040 The English Co-occurrence Dictionary

The English Co-occurrence Dictionary is composed of 460,000 alphabetically arranged of headphrases (plus 160,000 extra sentences on the English Corpus). The phrases are the abstracted portions of actual sentences contained in the EDR English Corpus. The results of the parsing analysis of these sentences indicates that the constituents of the sentence have a dependency structure. That is, the constituents have a governing-dependent relation. It is these constituents that form the headphrases of the English Co-occurrence Dictionary.

Records of the English Co-occurrence Dictionary are composed of the record number, headword information, co-occurrence constituent information, syntactic information, semantic information, co-occurrence situation information, and management information. The main role of the English Co-occurrence Dictionary is to show actual examples of how autonomous words are appropriately combined based on the co-occurrence situation information obtained from the English Corpus.

Appendix to the English Co-occurrence Dictionary: The English Corpus:

The English Corpus is composed of records arranged alphabetically. The records of the English Corpus are composed of the record number, sentence information, constituent information, morpheme information, syntactic information, semantic information and management information. The basic role of the English Corpus is first to identify the sentence constituents of sentences, and then to indicate how the constituents combine to form the semantic, syntactic and morphological structure of the sentence using a large number of actual examples as the source data.

Prices: R. 1,488 EURO RC. 12,403 EURO C. 24,807 EURO

ELRA-L0041 The Technical Terms Dictionary (Information processing)

The Technical Terms Dictionary contains 80,000 technical terms in English and 120,000 technical terms in Japanese from the field of information processing. The Technical Terms Dictionary is composed of the following subdictionaries: the Japanese Technical Terms Dictionary, the English Technical Terms Dictionary, the Japanese-English Bilingual Dictionary of Technical Terms, the English-Japanese Bilingual Dictionary of Technical Terms, the Headconcept Dictionary of Technical Terms, the Concept Classification of Technical Terms, the Japanese Technical Terms Co-occurrence Data, and the English Technical Terms Co-occurrence Data.

A basic descriptive format has been selected to illustrate the contents of the dictionary. An attempt has been made to select a format that eliminates the possibility of misunderstanding or misinterpretation. A record is composed of a number of fields. The correspondence between a field and its sub-fields is indicated by indentations in which the name or specifications of the field are given. The role of the field or the description of the contents that compose the field is given to the right of the indentation. This description method makes use of a portion of SGML (Standard Generalized Mark-up Language) but it is not the SGML.

Prices R. 1,488 EURO RC. 12,403 EURO C. 24,807 EURO

ELRA-S0084 SALA Spanish Colombian Database

The SALA Spanish Colombian Database comprises 1000 Colombian speakers (475 males, 525 females) recorded over the Colombian fixed telephone network. Corpus design, recruiting of speakers, annotation and formatting was done by the Universitat Politècnica de Catalunya (UPC). Collection was performed at Siemens Colombia.. Six speakers repeated the same prompt sheet in different calls. This database is partitioned into 4 CDs, each of which comprises 300 speakers sessions (except for CD 4, with 100 speakers sessions). The speech databases made within the SALA project were validated by SPEX, the Netherlands, to assess their compliance with the SALA format and content specifications.

The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SALA. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

Corpus contents:

· 6 application words;
· 1 sequence of 10 isolated digits;
· 4 connected digits: 1 sheet number (6 digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits);
· 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression;
· 1 spotting phrase using an application word (embedded);
· 1 isolated digit;
· 3 spelled-out words (letter sequences): 1 spelling of surname; 1 spelling of directory assistance city name; 1 real/artificial name for coverage;
· 1 currency money amount;
· 1 natural number;
· 5 directory assistance names: 1 surname (out of 500); 1 city of birth / growing up (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 "forename surname" (set of 150)
· 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question;
· 9 phonetically rich sentences;
· 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
· 4 phonetically rich words.

The following age distribution has been obtained: 11 speakers are below 16 years old, 486 speakers are between 16 and 30, 305 speakers are between 31 and 45, 163 speakers are between 46 and 60, and 35 speakers are over 60.

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

	ELRA Members	Non Members
For research use	13,000 Euro	16,000 Euro
For commercial use	16,000 Euro	20,000 Euro

Up-date on Language Resources from the ELRA Catalogue

ELRA-S0034 Verbmobil

This resource consists of spontaneous speech recorded in a dialog task (appointment scheduling). The BAS edition of the German part is fully labelled and segmented into phonemic/phonetic SAM-PA by the MAUS system and partly segmented manually.

New corpora available via ELRA (for the complete list, please contact ELRA or visit ELRA or BAS Web sites):

VM CD 30.1 - VM30.1 (BAS edition)
Verbmobil II - German, 58 spontaneous dialogues (33 close mic, 0 room mic, 25 phone line (GSM) recordings), 3024 turns, transliteration (Verbmobil II Format)

VM CD 31.1 - VM31.1 (BAS edition)
Verbmobil II - American English, 32 spontaneous dialogues (32 close mic, 0 room mic, 0 phone line (GSM) recordings), 2512 turns, transliteration (Verbmobil II Format)

VM CD 32.1 - VM32.1 (BAS edition)
Verbmobil II - Multilingual, 17 spontaneous dialogues (17 close mic, 0 room mic, 0 phone line (GSM) recordings), 992 turns, transliteration (Verbmobil II Format)

Price for ELRA members:	127,82 Euro/CD-Rom
Price for non members:	255,65 Euro/CD-Rom

Founding of the Global WordNet Association

We are pleased to announce the founding of the Global WordNet Association. The Global WordNet Association is a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world. The aims of the association are:

- To establish distribution facilities for the dissemination of the Association and Association publications and information materials:
- To promote cooperation and information exchange among related professional and technical societies that build or use wordnets.
- To provide information on wordnets to the general public.
- To promote the standardization of the specification of wordnets for all languages in the world, including:
- the standardization of the Inter-Lingual-Index for inter-linking the wordnets of different languages, as a universal index of meaning
- the development of a common representation for wordnet data
- To promote the development of sense-tagged corpora in all the linked languages.
- To promote sharing and transferring of data, software and specifications across wordnet builders for different languages
- To promote the development of guidelines and methodologies for building wordnets in new languages
- To promote the development of explicit criteria and definitions for verifying the relations in any language
- To promote the development of consistency checking, comparison and evaluation modules
- To promote research into the psychological adequacy of models of the mental lexicon

The Global WordNet Association (GWA) builds on the results of Princeton WordNet and EuroWordNet. Everybody with an interest in Wordnets can become a member of GWA.
Look us up at http://www.hum.uva.nl/~ewn/gwa.htm. Information about membership will be posted soon.
On behalf of the GWA board,
Dr. Piek Vossen, President: Piek.Vossen@sail-labs.be / Dr. Christiane Fellbaum, Vice-President: fellbaum@clarity.princeton.edu