The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

August 1998

Vol.3 n.3

Special Issue 1st LREC

CONTENTS

Editor in chief: Khalid Choukri

Editors: Khalid Choukri & Deborah Fry

Layout: Rébecca Jaffrain

ISSN: 1026-8200

Dear ELRA Members,

This issue of our newsletter is devoted to the First International Conference on Language Resources and Evaluation (First LREC) held in Granada during the last week of May 1998. This was organised by ELRA with the support of the major organisations involved in the language engineering area. Of course, the new resources secured by ELRA for distribution are also featured in this issue, as usual.

As you know, the main LREC conference took place from 28 May to 30 May, with several satellite workshops: 8 pre-workshops (lasting a half-day each on 26 and 27 May), and a post-workshop about potential co-operation issues across the Atlantic (31 May - 1 June).

According to the feedback we received and other echoes, it seems that the First LREC was a major event (and hopefully a significant milestone) in the life of language engineering. Its success can be summarised in just a few figures: over 197 papers, about 510 registered participants from over 38 different countries and all the continents. Among these, the largest group came from Spain (81 participants), followed by France (75), the USA (73), Germany (47), the UK (43) and Italy (41). Other “small” contingents came from Belarus, Croatia, Morocco, Taiwan, and Tunisia.

According to the registration forms, the participants belonged to over 325 different organisations, of which 210 were academic institutions (universities and research centres).

A major outcome of the conference is what is becoming well known as the ''Declaration of Granada'', which highlights the paramount importance of language resources. This declaration is enclosed.

In order to give you an idea of what happened in Granada, this issue is structured in three main parts.

The first part consists of general summaries drawn up by the Program Committee during the closing sessions. The summaries relate to spoken language resources (H. Höge), written language resources (N. Calzolari), evaluation in the spoken area (J. Mariani), evaluation in the written area (B. Maegaard), involvement of industrials in LREC (K. Choukri) and some concluding remarks from the chairman (A. Zampolli).

The second part attempts to give you some details of several sessions, as reported by the chairpersons. This part also includes short summaries of the panel discussions. At LREC we had three general panels, one with representatives from funding agencies in Europe and the USA, a second with representatives from non-European countries (Eastern-European and Arabic countries), and a third one with representatives of major industrial companies. We had also two technical panels, one about maintenance of LRs and another about EAGLES' work on semantics. There is also a summary of the post-LREC workshop. The topic was "Translingual Information Management: Current Levels and Future Abilities", the goal being to discuss past, present and future orientations and perspectives and to elaborate on potential areas for co-operation between the EU and the USA in the framework of the new scientific co-operation agreement signed by the Commission and NSF.

The last part is devoted to the important speeches given by some of the key political guests and supporters during the opening session. They addressed the crucial issues of LRs, evaluation and new information technologies. We are pleased to enclose the welcome speech given by Angel Martin-Municio, President of the Royal Academy of Sciences of Spain (and also Vice-president of ELRA), the speech of Mr. Vicente Parajon-Collada, Deputy Director of DGXIII, in which he elaborated on the prospects for language engineering from the European Commission’s point of view, the speech of Professor Bernard Quemada, Vice-president of the "Conseil Supérieur de la langue française", in which he addressed two key issues: the paramount importance of multilinguality when tackling language resources issues and the importance of co-operation between the various disciplines of language processing, in particular between producers of machine readable language resources and producers of the more classical dictionaries and lexicographies. The speech by Mr. Giuseppe Tognon, the Italian Sottosegretario di Stato al Ministero dell'Università e della Ricerca Scientifica e Tecnologica, is a fundamental statement about the global information society, the importance of language resources and language engineering in order to enable "the provision of universal access to the sources of information, offering opportunities for all citizens to use their own language". In his statement, he clearly points out that this issue goes beyond economic and business competitiveness and has an international dimension.

In his introductory speech, A. Zampolli, President of ELRA and Chairman of the Conference, draws a picture of the language engineering field, from the language resources and evaluation perspective of the last decades. According to him, LREC constituted a world premier conference where over 500 participants would focus on the very specific item of language resources. He said that LREC should be "a forum for exchanging information and exploring possible synergies and co-operation between teams, institutions, and funding organisations".

We would like to take this opportunity to thank all the authors and participants who facilitated the very interesting discussions and debates. We would also like to thank the local organising committee for its invaluable support.

Last but not least, we continued to carry out our regular activities even while we prepared the first LREC (and while starting on plans for the next one). We have updated our catalogue to include the new resources that are briefly described in this issue. These include the new speech databases developed within the framework of SpeechDat(M) and SpeechDat(II) and which cover German, Italian, and Slovenian. Other resources developed according to the SpeechDat specifications, are also available for Chilean Spanish, Russian, and Shanghai Mandarin. The speech database RVG 1 (Regional Variants of German), prepared by our partner BAS, is also available.

Antonio Zampolli, President Khalid Choukri, CEO

LREC Closing Session Summaries

Spoken Language Resources

Harald Höge

Globalization and the evolving technology of voice-driven man-machine interfaces are the driving forces for the growing demand for spoken language resources (SLRs), i.e. for:

- annotated speech databases,
- pronunciation lexica,
- tagged and raw text corpora.

At the LREC, the status of the field of SLRs was presented by the keynote paper (Höge, 1998). The main demand for SLRs stems from speech recognition technology based on statistical (i.e. data-driven) approaches. In future, new demand for SLRs will also come from new data-driven approaches in speech synthesis, where large speech and text databases are needed for acoustic synthesis (Campbell, 1998) and linguistic analysis.

The main goals within the field of SLRs can be described as follows:

- provide the requested SLRs for all relevant languages,
- provide standards for specifying SLRs,
- provide the tools needed to produce SLRs efficiently.

Contributions were presented on all three topics at LREC. However, it was clear that the goals are still far away. The biggest bottleneck is the production and distribution of SLRs.

Due to the success of ELRA and LDC, new attempts are underway to establish an "Oriental ELRA" for the distribution of "Oriental" SLRs (Tanaka, 1998).

In addition, several national and international projects are underway for the production of SLRs. The main contributions at LREC regarding SLRs for commercial use were covered by the European project SpeechDat, which addresses many languages and a number of application areas:

- project SpeechDat (Draxler, 1998), telephone applications, West European languages,
- project SALA (Moreno, 1998), telephone applications, Latin American languages,
- project SpeechDat-Car (Draxler, 1998), car application, West European languages.

Many SLR projects were described at LREC investigating new research topics such as prosody, dialects, translation, speaker verification, language identification and child voices.

In order to cover new languages or extend the existing SLRs to include new resources, the status of many national funded projects was reported. The SLRs produced in those projects are mostly oriented towards research use:

- African languages: SASPEECH project (Roux,1998),
- Dutch: several projects (Bouma, 1998),
- Eastern and Central European languages: BABEL project (Roach, 1998)
- French: several projects (Mariani, 1998)
- German: BAS project (Schiel, 1998),
- Italian: CLIP project (Leoni, 1998) ,
- Japanese: several projects (Ithaca, 1998)
- Russian: several projects (Semenova, 1998)
- Spanish: ALBAYZIN project (Diaz-Verdejo, 1998)

Regarding all these activities connected with producing SLRs, it is evident that no common international standard for specifying SLRs exists. The reason for this unsatisfactory situation is to be found in two facts:

- the field of SLRs is new,
- the new demands for SLRs for research and development of speech technology means that standards have to follow these issues rapidly.

As a result, the standards proposed at LREC were isolated actions focusing on certain aspects. Consequently, the tools for producing SLRs which were presented cannot operate on the basis of common accepted standards but have to be developed for non-standardised SLRs.

In order to improve this situation, discussions took place at the LREC to use the European EAGLES group together with ELRA and LDC as a platform to establish international accepted standards. An example of a widely accepted standard for commercial telephone speech databases is given by the SpeechDat project (Draxler, 1998).

Summarising the results achieved at LREC within the field of SLRs, the main findings are:

- SLRs is a fast growing field,
- there is a major shortage of SLRs for many languages,
- a basic set of SLRs for each language has to be produced, based on common accepted standards.

Bibliographical References

Bouma, G., Schuurman, I.(1998) Intergovernmental language policy for Dutch and the language and speech technology infrastructure, LREC (509-513)

Campbell, N. (1998). Design of Speech Corpora for Use in Concatenative Speech Synthesis Systems, LREC 1998 (1309-1312)

Diaz-Verdejo, J. et. al. (1998): a Task Oriented Spanish Speech Corpus, LREC 1998 (497-502).

Gibbon, D., Moore, R., Winski, R. (1997). Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Berlin, New York

Draxler, C. et. al (1998). SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices, LREC 1998 (361-366)

Draxler, C. (1998). URL SpeechDatCar: http://speechdat.phonetik.uni-muenchen.de/SP-CAR

Höge, H. (1998). Spoken Language Resources for Voice Driven Man Machine Interfaces, LREC 1998 (209-216)

Itahashi, S. (1998). On Speech and Text Databases Activities in Japan, LREC 1998 (355-360)

Leoni, F., et. al (1998). Corpus della Lingua Italiana Parlata, LREC 1998 (503-508)

Mariani, J. (1998). The Aupelf-Uref Evaluation-Based Language Engineering Actions and Related Projects, LREC 1998 (123-128)

Moreno, A. (1998). SpeechDat Across Latin America Project SALA, LREC 1998 (367-370)

Roach, P. et. al. (1998) Babel: A Database of Central and Eastern European Languages, LREC 1998 (371-376)

Roux, J.C. (1998). SASPEECH: Establishing Speech Resources for the Indigenous Languages of South Africa, LREC 1998, (351-354)

Schiel, F. (1998). Speech and Speech related Resources at BAS, LREC 1998 (343-350)

Semenova, V. (1998) Russian Resources in Language Engineering: Evaluation and Description, LREC 1998 (1123-1127)

Tanaka, H. & Fujisaki, H. (1998). LREC 1998 oral presentation, Oriental COCOSDA group

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

Written Language Resources

Nicoletta Calzolari

Parameters for Classification

The following parameters can be applied to classify the LREC papers on Written Language Resources (WLR): i) research vs. development, ii) type of resource/tool/etc., iii) level of linguistic description, iv) language(s). They are subdivided, in the table below, into sub-classifications for which the relative order is given in terms of number of Oral and Poster presentations. This quantitative overview of the distribution of the papers provides a rough idea of the relative weight of different aspects related to WLR.

Parameters for Classification

	Oral	Poster
Research vs. Development		
(Innovative) Research	3°	
Large Projects	1°	2°
System Development	3°	1°
Policy Issues	2°	
Type of Resource/Tool/... described		
Lexicon	1°	3°
Corpus	1°	1°
Methods	3°	5°
Task/Component	6°	4°
System	4°	2°
Infrastructural Aspects	4°	6

	Oral	Poster
Level of Linguistic Description		
Morphology	3°	1°
Syntax	1°	2°
Semantics	2°	2°
Ontology/Conceptual	5°	
Terminology	4°	2°
Language(s)		
One Language	1°	1°
More Languages	3°	2°
Bi- Multi- Lingual	2°	2°

Research vs. Development and type of Resource/Tool/etc. described

Development wins wrt research by a large margin, as expected given the nature of the topic (WLR), where the focus is on actual building of resources and related tools.

Within the few research papers the innovative aspects are: *acquisition techniques, word sense disambiguation, standards in lexical semantics*. There are no completely new trends, but papers describing, experimenting and using recently designed good quality approaches.

As for development there are more papers on resources than on tools, tasks, systems. It is important to note that most of them describe WLR (both Lexicons and Corpora) of (relatively) *large-coverage* (this is not true, however, for semantic annotation of corpora, still at an experimental level). Another relevant issue, in many papers, is *integration of Lexicon and Corpus*. Finally there are descriptions of resources *for Evaluation purposes*, while a new topic is emerging, due to ELRA: *evaluation of WLR*.

It is no longer true that Europe has only "feasibility studies", as we were criticised for until a few years ago. The global impression that we get from LREC is that European groups not only have understood the critical role of LRs within Language Engineering (LE) and Human Language Technologies (HLT), but are very active in *building large-scale WLR*. In this respect, I want to underline the *crucial role played by the European Commission (EC) - only recently complemented by national initiatives* - in the field of WLR. Without the EC support this plethora of initiatives could not have happened.

Levels of Linguistic Description

All the classical levels of linguistic description are represented with varying degrees of frequency. It is worthy of mention that – unlike in the recent past – there are more papers on *Syntax and Semantics* than on Morphology, both for Resources and Tools. Morphology is somehow taken for granted and no longer worthy of report (at least for many languages).

The clear meaning is that the state-of-the-art of R&D in our field is advancing and growing.

Surprising vs. Natural/Obvious Aspects

Some "surprising" aspects while considering the LREC papers: i) the number of submissions has largely overcome what foreseen in the most optimistic vision; ii) the quality is also much better that envisaged; iii) the proportion between Written and Spoken LRs leans towards WLR; iv) many papers - not to mention panels - concern "policy issues". This is a proof of the strategic importance attributed to LRs.

On the contrary, the following aspects are only "natural" or "obvious" in a conference such as LREC: i) many papers describe ongoing (often large collaborative) projects (e.g. PAROLE); ii) few papers report on really innovative research; iii) "technology transfer" is an important topic of quite a few papers, which are not just repetitions of existing technology, but usually present innovative aspects (e.g. those on EuroWordNet wrt WordNet, and on Treebanks for different languages wrt the Penn Treebank).

Policy Issues and Infrastructural Initiatives

The relatively large number of papers discussing policy issues and/or infrastructural initiatives is a very strong sign of the strategic and critical role recognised to LRs for a real advancement in LE and HLT.

The main topics touched on are: i) Standards. EAGLES was mentioned in quite many presentations, showing that its results are becoming the de-facto standards in Europe; ii) Means for accessing and making resources available; iii) Maintenance of LRs; iv) Distribution of LRs. ELRA is the obvious European response; v) Validation of Resources; vi) Multilinguality. One of the crucial topics for the near future, not only in Europe, but world-wide.

Wrt all these topics, one of the major concerns was to promote *cooperation* among different communities, projects, countries, and – in particular – between Europeans and Americans.

Maturity of the field

The average "good" quality of the presentations is an indicator of an *overall level of maturity in our field (LE)*: globally we have reached a common – technically good - basic platform.

Wrt conferences as COLING and ACL, at LREC many papers report on mature enough aspects to become consolidated through actual building of resources, tools, etc. If there is not much wrt "looking towards the future", LREC extraordinarily useful contribution consists in providing the first extensive overview of the field of WLR with a very good picture of "where we are now".

LREC makes clear what already is or can become in the immediate future a "product", ready to be sold and used in different applications: i) Morphological Lexicons; ii) Syntactic Lexicons; iii) WordNets; iv) Corpora Morphologically annotated; v) Corpora Syntactically annotated; vi) Taggers; vii) Robust/shallow Parsers; viii) Extractors from Textual Corpora.

A sort of "meta-product" or "meta-resource" available and used by the community at large are the EAGLES de-facto standards.

What Next?

In the next LREC Conference we would expect more of : i) Integration (of resources, tools, components, etc.); ii) Innovative Research; iii) Semantics and Contents related aspects; iv) true Multilingual resources, tools, etc.; v) Web, Multimedia, Multimodality; vi) new Standards for different aspects of LRs.

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

Spoken Language Systems Evaluation

Joseph Mariani

General issues : US vs EU

As mentioned by **C. Wayne** in the Else-Elsnet Pre-Workshop, the evaluation paradigm has been used by Darpa since 1984 to monitor its program. The need for an infrastructure in this framework was underlined, and this was achieved through the participation of the National Institute for Science and Technology (NIST) and the creation of the Linguistic Data Consortium (LDC). The Darpa program was opened to non-US laboratories in 1992, and already targeted, at that time, at the evaluation of multilingual language processing systems by the year 2001.

At the European level, we may mention a one-shot project (Sqale (1993-1995)), and several on-going projects supported by the European Commission (Else, Disc, Eagles...), or by other Funding Agencies (ARC Aupelf-Francil,...). Several projects, such as LE-ARISE, include an evaluation component, both technology and application oriented.

Technology vs User evaluation

There was a big discussion on the topic of Technology versus User evaluation. **M. Blasband** mentioned a decrease of speech recognition performances when going from laboratory conditions to field conditions (94% to 66%). Also, **J. Polifroni** mentioned a similar decrease in similar conditions (93% to 66%), but mentioned that after adaptation to the application conditions, the system performance went back to 93%.

In this framework, one may wonder if it is possible to design good applications with insufficient technology. But it also appears that having a good technology is sometimes not enough to address an application.

If it appears that Technology Evaluation, based on black-box quantitative evaluation, is feasible and helps conducting research programs, such as in the US Darpa experience, conducting Application Evaluation raises the problems of the size of the effort necessary to adapt a system to a specific application, of the genericity of the task which may not be general enough to attract a sufficient number of participants with enough interest, especially if it considers specific languages.

The usability of a system is an important topic, which was developed within the LE-Eagles Evaluation Working group, as mentioned by **M. King** and **B. Maegaard**.

What is a user?

While it appears that user evaluation is important, it is also important to carefully define what a user is. The user should have a goal in agreement with the task to be completed. **M. Blasband** mentioned the fact that a "user" was misrecognized by a system on two acoustically very similar cities (Mantes and Nantes), and kept on the dialog, as his task was not actually to get a train ticket, but rather to check if the dialog system was acceptable.

In the same way, evaluation experiments on dialog systems were carried out at the Eurospeech conference in September 1997 (Elsnet Olympics), and the results were reported by **G. Bloothooft**. Here, the "users" were speech scientists (and eventually even those who developed the systems, as no control was made on this), who had nothing to do with achieving a real task, such as getting a train ticket, but had the task to appreciate the quality of the dialog system.

More reliable results could probably be inferred from experiments with real users, such as the ones reported by **C. Dugast** (Philips) in the Swiss railway query system, which rose 200,000 Calls per annum, or those related to the use of speech technology in Car Navigation, as reported by **L. Hitzenberger**.

Confidence measures

Confidence measures appear to be a hot and important topic. It is related to the confidence that the system puts on the fact that a word, or a sentence, has been correctly recognized. It may be used in the presentation of the performances of a system, in order to have a finer analysis, as reported by **L. Chase**, in a dialog strategy or in order to have more natural human-machine dialogs, as reported by **G. Bouwman**, but even more interestingly, it may be used to facilitate systems training, especially for building up Language Models (**G. Zavaliagkos, S. Wegman**).

Resources for evaluation

The importance of having IPR-cleared resources for evaluation, both for training and testing, has been underlined. Producing resources for evaluation induces that the data should be of good quality (in agreement with the initial requirements, distributed in due time, properly recorded, with enough speakers, etc), as the participants in the evaluation campaign will not accept to get bad results because of the insufficient quality of the training or test data!

The data used for evaluation may be used afterwards by laboratories which didn't participate in the evaluation campaign to compare the quality of their systems with the state-of-the-art, and by laboratories which participated in the evaluation campaign to measure the progress achieved since they participated in the evaluation.

M. Liberman raised the dream of "Plug and Play" Linguistic Data, that you could plug into your system and get the corresponding application, with quality measures, overnight.

Another contribution was on the way to achieve rapid language model development, when not enough training data is available for constructing the language model **(L. Galescu et al.)**.

Annotation of resources

It is necessary to annotate the spoken language resources, either to build Acoustic Models or Language Models. For this, tools are fortunately available, that may be manual, semi-automatic or automatic. **E. Geoffrois** and **M. Liberman** proposed their Transcriber freeware, allowing easy corpus annotation and encouraging laboratories to produce data with de facto standards. **D. Fohr** proposed a software for annotating the speech signal. **G. Zavaliagkos** reported speech recognition results using untranscribed data for training the system. He mentioned that results similar to those obtained with a transcribed corpus may be obtained by using a much larger untranscribed corpus, if available. This is of course interesting in the case of radio or TV-Broadcast data, which are of unlimited size and easy access...

S. Wegmann and coll. described the use of a speech recognition system, with a confidence measure, for transcribing data.

Several papers dealt with specific speech processing systems evaluation (speech and speaker recognition systems evaluation, text-to-speech synthesis systems evaluation and dialog systems evaluation,). Please check those papers for more information.

Speech + NL evaluation

The LREC conference was therefore an excellent forum, where Speech and Natural Language specialists could meet on a topic of

shared interest, with a problem-oriented approach (designing the best tools for solving a problem), not a theoretically-oriented one.

Cooperation between the two scientific communities has been reported or would be desirable in many different domains : on the design of Language Models (especially in order to outperform the Bigram or Trigram approaches presently used), on the proposal of the DQR measure, initially proposed for Text Understanding in Aupelf actions, on the use of a reference tagged data for grapheme-to-phoneme conversion which was requested by the Aupelf[1] B3 TTS action, while it has been achieved by the CNRS GRACE project, on lexical semantics, which now comes into practice for training speech understanding systems on semantically annotated data. NL experts may work on transcribed data, as it has already been experimented in ATIS, SDR or TDT.

L. Hirshmann proposed to use a reading comprehension test, which could be of interest for conducting both spoken and written language processing experiments.

Multilingual Evaluation

Multilingual evaluation has been experimented in the EU LE-SQALE project (Large Vocabulary Speech recognition on American English, British English, French and German), in the US Call Home (Spanish, Arabic, Mandarin, German, Japanese), and TDT (Spanish, Chinese) tasks. Evaluation experiments have been conducted on Language Identification, and there exists since 1991 an international Working Group on Speech Databases and Speech I/O Systems Assessment (Cocosda).

Conclusions

In conclusion, we will stress the importance of using the evaluation paradigm as a necessary tool to accompany research and development in Language Engineering. It is a good area for promoting cooperation between speech and NL. Both technology and user evaluation should be considered. It indirectly produces Language Resources of quality and it is indubitably an excellent candidate for international cooperation in the field of Language Engineering.

1. Association des universités partiellement ou entièrement de langue française.

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

Written Language Evaluation

Bente Maegaard

Evaluation of NL projects and systems started in the 1960s with the MT evaluations, most notably the ALPAC report, but only in the last 10 years the importance of evaluation and evaluation techniques has really been getting attention from researchers on both sides of the Atlantic.

At the LREC conference we have witnessed that the importance of evaluation and of evaluation methodologies is now obvious, and we have heard a wealth of presentations which I will first briefly describe by a few headlines.

The first session concentrated on the broad issues in NLP evaluation: The US and French evaluation campaigns, their methods and results were described, as well as issues in text retrieval and fact extraction evaluation. This session was concluded with a presentation of the use of evaluation methodologies to validate, in casu to validate lexica.

Evaluation methodology and the importance of standardisation was discussed next. There is an emerging agreement on the important elements of an evaluation methodology. This was one of the very positive results of the conference.

The session on evaluation tools and tools for evaluation covered spelling and grammar checkers, alignment tools, terminology extraction tools, tokenizers, taggers, parsers. Finally, we heard about evaluation of generation, summarisation and other NLP components.

These sessions showed a multitude of approaches to evaluation, of resources for evaluation, of tools for evaluation and of tasks that were evaluated. It was also positive to see that the systems and projects treated worked on many different languages. The presentations and discussions underlined the fact that there are a number of very different types of interests in evaluation: researchers, funding agencies, industry and consumer associations/consumers each have their special purpose with evaluation and therefore need a different approach.

It might be feared that this whole multitude of approaches and purposes would necessarily lead to a very heterogeneous picture. But the fact is that whoever the interested party is, evaluation consists of three main points: 1) Set the goal (the purpose of the evaluation), 2) Define the functionality you want to obtain, 3) Define the metrics.

We have seen different ways of looking at evaluation:

- Evaluation as a science. This involves methodologies, metrics, resources for evaluation, validation of resources, e.g. how can you measure a certain feature of a system in a reliable way, how can you build good and usable resources for evaluation?
- Evaluation as a means to advance research. This is more organisational, but there are also technical considerations, e.g. on the nature of test data, etc.

These are the two basic, different ways of looking at evaluation. Additionally, we have seen:

- Evaluation perspectives. Evaluation of technology, versus evaluation with respect to user needs, and evaluation with respect to industrial needs.
- Measurements. A few papers concerned measurements, but measurements alone do not form an evaluation, cf. points 1) and 2) above.

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

- NLP is becoming mature, this is the reason why 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 progress is still needed.

So, I am looking forward to the next LREC!

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

Involvement of Industrials in LREC

Khalid Choukri

From its inception, LREC was designed as a forum in which industrial players would meet major R&D actors. "The aim of this Conference is to provide an overview of the state-of-the-art; discuss problems and opportunities; exchange information regarding ongoing and planned activities, language resources and their applications; discuss evaluation methodologies and demonstrate evaluation tools; and explore possibilities and promote initiatives for international co-operation ..."

The goal in terms of participation has been achieved. We welcomed over 500 registered participants with at least 120 from industry. Out of the 325 different organisations, 215 were research laboratories or universities, and 110 were industrial companies (roughly 2/3-1/3).

If we consider the number of papers accepted by the programme committee, we see that most of these are from universities and research centres (about 160), 12 papers reported the work carried out jointly by academic and industrial teams. Over 25 papers reported on work done in industry, and around 10 papers reported the activities of other types of organisations, such as associations and agencies.

If we consider the industry contributions split over four major areas, we obtain the following table:

	Speech		Written	
	Resources & related tools	System Evaluation	Resources & related tools	System Evaluation
Papers from industrial organizations	4	5	10	8
Papers published jointly (industrial and Academic teams)	4	1	4	3

It is our challenge to attract more participants from industry, as well as more submissions from industry to the 2nd LREC.

In my capacity as ELRA CEO, I highlighted the importance of ELRA pursuing its role in supporting both R&D labs and commercial entities. ELRA's pricing and distribution policy is the most obvious means to do this. The policy clearly defines different membership fees for not-for-profit organisations and for commercial ones. It also distinguishes two prices for the resources: a low one for data acquired for research purposes and a higher one for commercial use, whenever this is a negotiable issue with the producer. If we take a look at our distribution activities during the first two quarters of 1998, we can see that ELRA distributed over 74 resources for R&D, including 5 at no cost, and about 60 items for commercial use (a major step forward when compared to the same period of 1997 with 13 resources distributed for R&D and 9 for commercial use). About 10% of ELRA's sales revenues are accounted for by R&D entities and 90% by commercial enterprises.

The exploitation of the data assumes fair use in line with the clauses of the agreed "user-license". The user who needs to use the data purely for research, without any intention of technology transfer, agrees that: "Within this Agreement DISTRIBUTOR grants END-USER, engaged in bona fide language engineering research, the non-exclusive right to use the Language Resources, exclusively for the purposes of their language engineering research activities. END-USER is not permitted to reproduce the Language Resources for commercial or distribution purposes and to commercialise (or distribute for free) in any form or by any means the Language Resources or any derivative product or services based on all or a substantial part of it" (Article 1 and 2 of our end-user license).

We assume that for commercial organisations the ultimate goal is to develop new technologies and products, and therefore they are granted "the non-exclusive right to create derivative products or services from the LRs for internal research purposes and/or internal technology development and the non-exclusive right to distribute and market, according to VAR's commercialisation policies, any derivative product or service from the LRs by VAR." (Article 2 and 3 of our VAR agreements).

It is obvious that if none of our customers infringe the license they are granted, then the rights of the data owners will be sufficiently safeguarded, without any need for courts and lawyers. A consequence of that may be that more providers will entrust ELRA with the distribution of their valuable resources. ELRA can then devote its funds to technical matters such as joint-ventures, commissioning the production of new and useful resources.

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

Closing Session Remarks

Antonio Zampolli

Needs

What has clearly emerged in all the LREC events is that LRs are a top priority in both academic research and industrial development.

LRs have a particular role to play in the integration of Speech and NLP: this Conference has been the largest planned effort so far for promoting and effectively fostering the integration of the two communities.

LRs are the key to an effective multilingual information society. The availability of adequate LRs in a particular language is the critical factor in the development of applications and services, informed by LT, in that language. LRs provide language-specific linguistic knowledge, as well as the cross-linguistic knowledge necessary for successful multilingual links among languages. In many cases, it will be possible to transfer methods, technology and, in particular, software tools, from one language to another provided that adequate LRs exist for the second language.

Evaluation and LRs are closely related in many ways and share several research issues.

A current issue is whether both should be supported within the same organisational structure.

All types of LRs are needed: general, domain specific, and for individual applications. At the organisational level, these types differ in various essential aspects: funding modalities, legal status, availability to the users, degree or need for reusability. But they are closely interrelated and must be designed to allow for efficient and cost effective customisation in building the specific domains on the general ones, and to join them within a common and shareable linguistic knowledge base.

Requirements

Common services are required.

The need for standards in LRs to be continuously maintained and updated is universally recognised. The Conference witnessed the wide dissemination and use of the EAGLES guidelines.

Continuity is an essential feature, in particular for updating and maintaining LRs.

LRs constitute a proper research field in themselves: new methods (for customisation, knowledge extraction from corpora, etc.); new types of LRs should be conceived, designed, and trialled to promote, anticipate, and accommodate the evolution of science and technology in HLT.

Key Organisational Issues

The requirement of a basic set of LRs for as many languages as possible clearly emerged at the Conference, not only as a political, social, and cultural need, but also as an industrial one.

In designing global priorities, we should take not only market forces into account, but also the needs of the research community, the preservation of linguistic and cultural diversity, and the principle of offering citizens equal access to the benefits of the IS.

LRs design and production requires specialised professional expertise and dedicated skills. LRs are the most expensive component of any LT system. Today, only embryonic nuclei of LRs exist for the majority of languages, which cannot be effectively used in real systems without a substantial enlargement of their coverage and the addition of new layers of linguistic information. This requires that efforts are cumulated and not duplicated, reusability of LRs ensured and enhanced, and that existing LRs and specific technical knowledge are exploited. The provision of LRs, and, consequently, the development of the products and services required by the IS are feasible only if we are able to reach a substantial economy of scale.

It is vital that the results achieved in the last decade through co-operative efforts and a sometimes painful process, are not dispersed or lost but preserved and put to use.

Otherwise we risk the epochal mistake of losing 5 – 10 years' worth of progress, which could be fatal for the role of LT in the global multilingual IS.

Evolution Toward a Global Organisational Model

We should be very grateful to the Commission who, through the Steering Committee chaired by Mr. Parajon and with the inspiration of Mr. Cencioni, has promoted the foundation of ELRA and has designed its mission and structure as they are today.

We must work together to adapt this model to the evolution of LRs in the framework of the IS.

A model for the participation of industrial partners must be found: their direct involvement will be a key condition for the future of ELRA. ELRA is ready to enter a joint venture with a network of industrial developers, users, and researchers.

In doing that, we should take into account the general technological, scientific, industrial, commercial, social, cultural, and political requirements of the field, the mission of ELRA, and the various actors and forces which are operating in the international context.

We should not forget that innovative applications require the continuous development of core technologies, and that technology is still immature for certain LT sectors and classes of applications: the organisational model should also take the needs of developers and of researchers into account. The model should also make provision for languages which are not privileged by market forces, that is, if we don’t want to give up the principle of offering, through the potential of LT, a friendly “democratic” access to the possible benefits of the IS in as many languages as possible, in a truly global multilingual context.

A recent EUROMAP draft survey shows that support of LT is extremely uneven across Europe at a national level. Several Member States have no policy on the support of their national language within the IS, "a situation which could adversely affect the survival of those languages in the mainstream". This problem is particularly acute for the provision of language-specific LRs.

As far as we know, six member states of the EU are launching national programs in LT: significantly 5 of them are explicitly dedicated to LRs.

Even if national authorities were to take responsibility for the provision of the monolingual LRs for their own languages, in this way countering the market forces which tend to promote only the more widely-used and economically-important languages, the problem and responsibility for a multilingual LR policy remains.

Individual application projects, even if clustered, cannot answer all these requirements alone. Co-ordinated projects and initiatives explicitly dedicated to the various aspects and phases of the life-cycle of LRs are needed.

In the current framework, the EU has recognised the need for 3 stages in LRs development: standards for LRs, creation of LRs, and distribution of LRs.

European networks are already in place for these tasks: for example, EAGLES, PAROLE, SPEECHDAT, TELRI, ELSNET LR GROUP, and ELRA.

We should take inspiration from the model NSF has offered for consideration: to use and support networks of professional organisations and specialised institutions in order to share efforts, costs, and know-how; to reuse existing expertise and skills; to ensure continuity, maintenance, updating, and synergies between research , production, and distribution of LRs.

We need to consolidate and support a distributed co-ordinated European infrastructure in which existing networks are reinforced, and co-operate in the different phases and aspects of LRs, and the European Commission and Member States need to interact according to the subsidiarity principle.

International co-operation is essential. Multilinguality involves languages of all continents. The globalization of the society is already breaking organisational, institutional and political barriers. The recently signed Science and Technology Agreement between the government of the United States and the European Commission seems to us a unique opportunity to promote co-operation in the field of LRs in a truly multilingual context.

A common strategy could help in overcoming possible differences in the bureaucratic,

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

LREC Panels Summaries

Panel of the Funding Agencies

Antonio Zampolli

Introduction

The Funding Agencies (FAs) had a decisive role in the emergence of the paradigm in which Language Resources (LRs) play a central role.

LRs seem to involve policy-related issues more than any other sector of Human Language Technologies (HLT):

- LRs are largely specific to individual languages.
- LRs are a key prerequisite for the application of a given technology in a given language.
- LRs require continuous maintenance and update.
- LRs are inextricably connected with culture.
- International cooperation is essential to ensure the creation and availability of multilingual LRs for as many languages as possible.

This consideration raises the following important questions:

- Whose responsibility is it to develop core LRs for a language?
- Can the provision of LRs be left to market forces alone?
- What is the best way to ensure international cooperation?
- What are the implications of the pre-competitive infrastructural nature of LRs?
- How can the best use of the scarce human/financial means available be ensured?
- How can national priorities be reconciled with international cooperation?
- What are the FAs planning in terms of LR development?
- How can the need for continuity in developing LRs be reconciled with the lifecycle of individual FA programs?
- Can the FAs support infrastructure and activities in this regard?

The strategy the FAs choose to adopt for LRs will have decisive impact not only on the future of LRs, but, even more importantly, on the place of language technology in the information society.

PANELISTS

Roberto Cencioni (European Commission, DG XIII-E-5):
Language Engineering Progress and Prospects

After having summarized the aims and results of the 3rd and 4th Framework Research Program (FRP), and having characterized the current situation in the field of LE and LRs, R. Cencioni discussed future perspectives.

Future perspectives

Challenges which inspired the EC action in LE for 1999 – 2002 include:

Adopt an integrated approach from research to market launch (avoid gaps)

Build on strengths & concentrate on global challenges (top-down R&D)

Increase reaction to industry-driven developments (bottom-up validation)

Facilitate the transfer of key technologies into multiple languages (new forms of partnership)

Support shared networks & facilities (best practice, market intelligence, standards,…)

The program will include:

- Five application areas:

Business information services, Services of public interest, International commerce, Telecommuting, Business (language) training.

- Prospective accompanying actions:

Focused, goal-oriented research, Resources (written & spoken language databases), Best practice & de-facto standards, Network of national focal points (European LE scope).

- Three technology lines:

Fully multilingual capabilities, Ability to work and communicate in one's language, Natural interaction, Natural and keyboard-less interfaces, language input-output, Active Content, Information retrieval, extraction, filtering, clustering and delivery.

Gary Strong (National Science Foundation):
Language Resources and Evaluation

The NSF (DARPA) Human Language Resources Program highlights some crucial needs in this field:

Large data resource centres; Annotated, shared data to fuel data-centred research; Connection between data and evaluation plans, Multiple, overlapping data resource centres.

An NSF workshop on 8/16/97 identified the following needs:

Speech from a broader population, multimedia archives, and corpora of new kinds of computer-mediated communication; New, shared multilingual resources (including monolingual text and speech in languages other than English, and parallel text corpora); Resources to support research and development in generation and synthesis of spoken language.

The following key points should be considered when discussing the ways and means of creating and distributing LRs:

- **Intellectual property rights**

The language research community needs effective strategies and tactics to deal with IPR issues.

- **New modes of outreach**

Language technology researchers are vastly outnumbered by researchers, teachers, and students in all language-related subjects. New channels of communication and resource-sharing should attempt to take advantage of this large reservoir of talent and energy.

- **Improved infrastructure for data sharing**

Most government-sponsored LRs never leave the lab. Modern computer networks make sharing such resources both easier and more valuable. The various approaches to data sharing should be made available to the public. There should be more pressure from sponsors to publish or otherwise share useful LRs.

Charles Wayne (National Security Council)

The basic components of the LE paradigm are general research, exploratory development, advanced development; technology-oriented and task-oriented evaluation metrics, and LRs for multiple languages.

These valuable LRs demand 100% public support for their development and maintenance.

The DARPA-sponsored text and speech projects from 1987 until today have proved that evaluation is essential to the advancement of research and development.

The "evaluation + data + tasks/algorithms" paradigm stimulates and enables progress, represents a manageable framework, and is cost effective.

The infrastructure is essential.

The main issue in international co-operation should be multilingual LRs, evaluation and research.

Catherine Macleod (New York University)

Catherine Macleod pointed out that funding new resources was not enough if the money is not allocated for a project right from the start. This is needed to ensure the continued maintenance and development of the resources.

Nuria Bel (Fundacion Bosch Gimpera)

It should be a priority to link existing speech/lexical/corpus/knowledge databases in what could potentially become a real multilevel, multilingual network of LRs. A network of this kind would guarantee that applications made for one language could also easily be transferred to other languages cost-effectively.

Joseph Mariani (LIMSI-CNRS)

In the US, efforts are structured using the evaluation paradigm, and have focused mostly on improving technology. A permanent infrastructure has been used or at least installed, comprising NIST and LDC, 100% funded to organise or provide data to feed the evaluation.

In the EU, a project-oriented approach has been used, which is more of an application-oriented approach. There is no permanent infrastructure, with the exception indirectly of ELRA and Elsnet, and is only 50% funded.

Are both the US and the EU happy with their respective methodology or do they plan to adapt their approaches?

If so, how does the EC plan to install a permanent infrastructure for language resources and evaluation, as well as overcome the problem of the 50% funding scheme ?

Nicholas Ostler (Linguacubun Ltd)

When a large alliance (such as the European Union) adopts a serious technical programme, or when an agency (such as ARPA) sets itself technical goals that survive several political administrations, continuity and consistency are possible. The UK policymakers who made the SALT Programme and its predecessors possible had long since gone before the real fruits of the programme could be assessed.

Nancy Ide (Vassar College)

Standards for LRs and creation of LRs should be the priority of trans-atlantic co-operation. National corpora and lexica should be developed for American English, compatible with their European counterparts.

Summary

The results of the discussion can be summarised as follows:

- Priorities for co-operation: LR standards, development of harmonised monolingual and multilingual LRs, research in core technologies and co-operative evaluation.
- There is a need for public support in the development of general and domain-specific LRs in as many languages as possible.
- There is also a need for infrastructure continuity and stability to ensure standards, production and maintenance, as well as the distribution of LRs.

Industrial Panel

Khalid Choukri

The main objective of this panel was to discuss major issues of interest to industry players and to bridge the gap between them as commercial users/producers of language resources, "academic" producers, funding agencies, and distribution agencies such as ELRA or LDC.

The panel was chaired by **Khalid Choukri** (ELRA CEO) with the participation of heavy weight players in the LE field: **M. Hunt** (Dragon, UK), **S. Kunzmann** (IBM, Germany), **J. Odijk** (L&H, Belgium), **D. Brooks** (Microsoft, USA), **N. Lenke** (Philips, Germany), **J.P. Chanod** (Xerox, France).

The following section reflects some common views of the panelists and the floor. In addition, the panelists were asked to summarize their contributions in a few sentences.

The panel discussion focused on industrial companies' need for "pre-competitive" language resources with potential for the development of a large class of applications, and for "competitive resources", which can be used to tune real applications. In the case of the first category, if the resources are well designed and if the IPR issues are cleared, they may become shareable data, both for development and as a benchmark for evaluation.

The needs of industrial companies are not identical to those expressed by the academic researchers. For the commercial sector, languages are chosen according to business criteria with a clear orientation toward lucrative languages. In considering less-lucrative languages (the "minority" languages), public funding remains a key factor, except where specific knowledge can be derived and generalised to more lucrative languages.

Several companies expressed their intent to co-invest in language resource productions in order to share investments; the co-funding would come from language engineering funding agencies, in particular the European Commission.

The commercial enterprises, when asked about the role of organisations such as ELRA, insisted on the need for a data collection and distribution infrastructure. None of them were of the opinion that such organisations can be run as "private businesses".

Siegfried Kunzmann (IBM Speech activities)

1997 can be considered to be the year where automatic speech recognition became a commodity for millions of people in several languages, including British and American English, Chinese, Japanese, German, French, Italian and Spanish. The various products set the bar for highly accurate, large vocabulary, continuous, speaker-independent speech recognition systems for the PC market. Besides a lot of algorithmic improvements in basic speech recognition technologies, the progress was driven through the availability of a lot of acoustic and linguistic data enabling the proper estimation of parameters for the statistical methods.

To be able to achieve the same progress in many more languages and to make further progress in natural language understanding systems and/or telephones as input devices, a lot of data needs to be available for the various disciplines. To make rapid progress, concerted efforts on data collection are needed to enable the European and world-wide speech community to make key inventions in basic speech technologies, as well as to enable the deployment of speech processing systems in more languages.

Melvyn Hunt (Dragon Systems, UK)

Dragon Systems is interested in a wide range of speech and text corpora. We obtain them by internal collection, collection by partners and sub-contractors, private purchase, and from LDC and ELRA. Dragon Systems Inc. is a member of LDC, and Dragon Systems UK is a member of ELRA. We consider that such organisations have a very important role to play. We are particularly keen to see reciprocal arrangements between LDC and ELRA.

ELRA is seen to be stronger in its speech corpora than in its text corpora. Dragon would be particularly interested in specialised text corpora (e.g. with legal or medical texts). We feel that not all the ELRA speech corpora represent good value for money: A corpus obtained from an external supplier is considerably less valuable than a similar corpus collected internally, because the external corpus is never tailored exactly to a company's needs; one has less information about it; and there is always work to be done in converting it to local formats and standards. The exception to this is when an external corpus can be used at several sites to compare performance. It then becomes more valuable than an internal corpus.

In working with LDC, ELRA and other external resource suppliers, there are grey areas concerning what constitutes commercial -- as opposed to research -- use. Merely re-stating a definition of what constitutes commercial use does not solve the problem, because there are many gradations between using a resource directly in a product and using it highly indirectly, perhaps discovering an abstract property which is later used in a commercial product by the same organisation. Similarly, there are anomalies in charging commercial as opposed to research organisations for access to resources. Should a five-person start-up company really pay more for access to resources than a large educational or government research organisation, which might well pass on the fruits of its research to associated companies?

In collecting speech material, merely collecting examples of spoken words is not enough. It is important to remember that speech is communication. Recognition performance can depend strongly on the attitude of the speakers. The ideal speech recordings for speech recognition research would be taken from people operationally using a similar speech recognition system.

David Brooks (Microsoft)

Microsoft is committed to making computers easier to use by providing users with a natural, linguistic user interface. To achieve this, we will utilise state-of-the-art linguistic technologies to provide speech recognition, speech synthesis, natural language understanding and similar capabilities. In addition to the linguistic research efforts under way at Microsoft, we look forward to working toward this goal with the linguistic research communities in Europe, Asia and America.

Microsoft's products are widely used around the world, and we are committed to delivering this linguistic interface to our global community of users. In addition to the technical barriers, there are many thorny issues that need to be resolved before this goal can be achieved. Among them are clarification of intellectual property rights vis-à-vis language, valuation of linguistic resources, survival of minority languages, resolution of regional differences among speakers, and many others. We look forward to working with the academic community, legal scholars and political leaders to resolve these issues and smooth the way for the transition to a linguistic user interface.

Nils Lenke (Philips)

Philips Speech Processing is active in the speech recognition markets of IT-based applications like dictation, telephony-based dialogue systems and voice-controlled consumer electronics. Many different languages are supported, therefore we have a large interest in various language resources. This can also be seen from our membership in ELRA and LDC. I think that organisations like these have an important task in distributing language resources, especially those stemming from EU or otherwise publicly funded projects.

However, companies like Philips Speech Processing will also always have to collect speech resources for their own use (or the use of their customers), only these resources will never appear in e.g. ELRA. What sometimes seems to be missing is an infrastructure of organisations and small companies who can assist us in performing data collection campaigns for the various languages in various countries. This is especially true for complex dialogue systems.

In this case, a methodology for collecting spontaneous utterances in an efficient way is still missing. This could certainly be an area where organisations like ELRA could play a role.

Jean-Pierre Chanod (Xerox)

From an industrial perspective, two important aspects of language technology deserve our attention: language resources and integration of language technology into broader user environments.

Creating and maintaining language resources is a complex and expensive process. Much progress has been accomplished in the last 10 years, especially for broad-coverage monolingual dictionaries and annotated corpora, as well as in the area of standardisation.

Still, we are facing continuing challenges, multilinguality among others. With the Internet growing, citizens will need and require services, ranging from e-commerce to health or education, in their own languages. Multilingual support is an economic, political and cultural necessity. We must create monolingual and bilingual resources for a wide variety of European and non-European languages (Xerox has so far developed resources and tools for more than 15 different languages). As basic language technology aims at more ambitious goals, we must also create new types of very large resources, such as thesauri.

This is a complex and expensive process. To co-ordinate the construction and maintenance of such resources, to ensure their quality and to optimise costs, Xerox put in place a new infrastructure, the Language Resources Group. Needless to say, collaborative research and joint projects with external partners play an important role in that strategy, as we cannot rely on our own skills to cover so many different languages.

Another important industrial concern is the integration of language technology into multiple-user environments.

Our technical approach to language engineering relies on a core technology (mostly finite-state) and a unified language-independent architecture (XeLDA) into which modular linguistic components (e.g. morphological analysers, part-of-speech taggers, shallow parsers) are integrated. Basic linguistic components are then integrated into multiple language applications such as information retrieval, terminology extraction, translation aids or translation memory.

Our industrial approach relies on the smooth interaction of R&D with business entities. The relationship between research and the market is cyclical. Each entity may elicit requirements or constraints, or more broadly inspire the other. As specific interests are expressed, new opportunities are created more quickly for the market, as well as for research. This is innovative tension.

But beyond this technical and organisational integration, we need a vision to sustain the future of language technology. Language technology will be integrated in multiple environments, the Web, e-mail, paper, digital libraries, knowledge brokers, multimedia applications. Language technology will be embedded wherever there is a need to access, share or disseminate distributed knowledge, no matter the medium, no matter the language.

Panel on International Cooperation

Alain Servantie, European Commission, DG XIII

The special session on international co-operation mainly dealt with the advantages and difficulties of co-operation between researchers of Central and Eastern Europe or Southern Mediterranean Countries with the Community researchers in the field of linguistics. The session was chaired by **Alain Servantie** (European Commission, DG XIII) who mentioned that about 15 R&D co-operative projects involving researchers of those countries had been financed representing a total amount of 4.18 million ECUs.

Prof. Eva Hajicova (Charles University, Prag) said that Eastern researchers had a better imagination and were better organised than their Western counterparts, which allowed them to take a lead in some of the joint projects. Summer Schools on language engineering have been particularly useful, but their financing is becoming more and more complicated. The 5th Framework Programme should be oriented more towards supporting the preparation of concrete products.

Prof. Dan Tufis (Romanian Academy) recognised that advances in the field of speech technology in his country would not have been possible without the western co-operation. Awareness actions and particularly summer schools are welcome to sensitise the public at large and responsible persons in particular of the interest of such research. However delays in procedures and payments, cumbersome calculation of overhead costs for universities reduce the efficiency of the co-operative research.

Prof. Klara Vicsi (Technical University of Budapest) said that taking part in the Babel project taught several people to work in team. However delays in payments were such that the project was nearly completed when the money arrived.

Prof. Mohamed Chad (Université de Fez) emphasised that co-operation with Europe was necessary for Southern Mediterranean countries in order to be brought to a satisfactory technological level, and pleaded for the creation of joint EU/Med research teams. There were unfortunately too few actions between the EU and Mediterranean Countries: Med Campus had been suspended and should be resumed. Europe should not forget its South and give the impression that it leaves them on their own; resentment would create intolerance.

Prof. Zygmunt Vetulani (Mickiewicz University, Poznan) emphasised the importance of awareness actions, particularly useful to increase the so far low support of local industries. Co-operation between Western researchers and researchers of countries speaking slavonic languages was particularly useful as those languages are highly inflected and constitute a good reference for comparative studies, testing general formats and algorithms, etc. Intensive human mobility is necessary to establish a good social climate for co-operation – Internet is great but this is not a panacea. International co-operation within research projects helps local structures (university administration, enterprises - to get familiar with EC rules and regulations. Problems laid with heavy European bureaucracy, different treatments of currencies, high co-operation costs, low level of awareness of local industries.

Prof. Salem Ghazali (IRSIT, Tunis) also said that Europe should co-operate with Southern Mediterranean countries to help Arabic to pass the computer test and build on the future: a specific adaptation is necessary; and this will condition the survival of the Arab national heritage. So far research teams in Southern Mediterranean countries work like underground organisations acting on personal initiatives; no coordinated policy exists in the Arab world. Arab research institutions are not developed enough to be able to compete in calls like the ones launched by the European Union; they should be helped through seed money or a special status to help preparing projects.

A communication of **Mr Daniel Martin Mayorga** (Telefónica, Argentina) was also read, emphasising the interest of researchers of that country to cooperate with Europe in the linguistic field, using such instruments as the René Thalmann Foundation and promoting the creation of hispano-american joint archives.

List of projects with involvement of non European Union teams

MULTEXT-EAST, ELSNET GOES EAST, GLOSSER, GRAMLEX, PRACTEAST, BILEDITA, BABEL, LANGELEC, AGILE, CONCEDE, ARAMED, AREF

For more information please visit the CEC Website:
http://www2.echo.lu

EAGLES in Granada

Antonio Sanfilippo (Sharp, UK and Linglink, Luxembourg) and Nicoletta Calzolari (ILC-CNR, Pisa)

The LE EAGLES (Expert Advisory Group on Language Engineering Standards) Project organised a Panel on *Lexical Semantic Standards for Information Systems*. The panel's objective was to discuss issues concerning the provision of guidelines in standardising the encoding of semantic information in lexical resources with specific reference to multi-/cross-lingual document management applications. The basis for this discussion was the ongoing work by the EAGLES *Lexical Semantic Interest Group* (see http://www.ilc.pi.cnr.it/EAGLES96/rep2).

The group includes researchers from a large variety of language technology institutions across Europe, both industrial and academic: SHARP (UK), ILC (Pisa), IRIT (Toulouse), University of Amsterdam, Sheffield University, Manchester Metropolitan University, GILCUB (Barcelona), ISSCO (Geneva), DFKI (Saarbruecken), Institut d'Estudis Catalans (Barcelona), Facultad de Lenguas Aplicadas (Madrid), IRST (Trento), XEROX (Grenoble), University of Gothenburg, LINGLINK (Luxembourg).

Natural language meaning has always been thought of as one of the hardest problems for standardisation. However, the increasing use of conceptual classification in the development of language technologies is rapidly changing this perception. At the same time, the growing need for dealing with semantics and contents in LE applications is pushing towards more powerful and robust semantic components. Within the last decade, the availability of robust tools for language analysis has provided an opportunity for using semantic information to improve the performance of applications such as *Machine Translation*, *Information Retrieval*, *Information Extraction* and *Summarisation*. As this trend consolidates, the need of a protocol which helps normalise and structure the semantic information needed for the creation of reusable lexical resources within the applications of focus becomes more pressing. Times are thus mature to start tackling the question of how to formulate guidelines for lexical semantic standards.

That was the core message that the panelists **Antonio Sanfilippo, Nicoletta Calzolari, Rob Gaizauskas, Patrick**

Saint-Dizier and **Piek Vossen** set out to convey to an audience of well over a hundred participants.

The panel program began with a brief introduction about the group's work and its interactions with previous results, its objectives, and background including current synergies with other R&D consortia. Standards are not of interest if they are not actually used. It was stressed how existing EAGLES results in the Lexicon and Corpus areas are currently adopted by an impressive number of European - and recently also National - projects, thus becoming "the de-facto standard" for LR in Europe. This is a very good measure of the impact – and of the need – of such a project in the LE sector. The major project implementing the EAGLES standards is LE PAROLE, with Corpora for 14 languages and Lexicons for 12 languages adopting the same model.

Three presentations followed, on the following topics:

• *Semantic Requirements for LE applications* with reference to: Machine translation, Information Systems and related enabling technologies (e.g. word clustering, word-sense disambiguation, multi-word and proper noun recognition).

• *Lexical Semantic Resources*, including: wordnets, thesauri, ontologies, bilingual and monolingual dictionaries, large-scale and experimental NLP lexicons.

• *Linguistic aspects of lexical semantics* concerning: lexical aspect, semantic relations, semantic roles, lexicalisation, verb semantic classes, nominals, adjectives, prepositions and adverbs.

These presentations had a primary focus on the survey phase carried out by the group from May 1997 through February 1998. The survey phase aimed at identifying basic notions, which constitute the building blocks of lexical semantic encoding.

The next stage of the work, which is concerned with the deliberation of standard guidelines in lexical semantic encoding and whose results will be available after August 1998, was also broadly outlined with reference to three bands of priority:

• use in real-world and experimental applications;

• information available in lexical resources, not yet used in applications,

• notions which could be encoded in LR to improve performance of applications.

All discussants had words of praise for the goals and coverage of the work carried out by the EAGLES *Lexical Semantic Interest Group* and provided constructive criticism which stimulated an interesting and sustained discussion with many questions and comments from the floor.

Lin Chase (LIMSI-CNRS, Paris) related the concerns expressed by panelists to an increasing need to use lexical semantic information to improve the performance of Spoken Language Systems. She pointed out that the use of lexical semantic information may be instrumental in increasing the precision of language model for speech recognition. An active involvement of representatives of the Spoken Language community in this Group was considered desirable both by the discussant and the panelists. An integration of Written and Spoken Language in the field of semantics is quite natural, given the common pressing interests in this area.

Sergei Nirenburg (CRL, Las Cruces) commented on the panel presentation on semantic requirements for LE applications. He criticised current practices in the development of concept-based applications for offering either small-scale domains of application or only providing limited uses of lexical semantic information with a stronger emphasis on corpus- or syntax-based techniques.

Ed Hovy (ISI & USC, Marina del Rey) commented on the panel presentation about lexical semantic resources with appraisal for the large coverage of the Survey, a very good platform which gives an idea of the current range of possibilities. In the field of Lexical Semantics we know many little pieces, and it is timely to try to put them together.

Ralph Grishman (NYU, New York) raised the question of whether the standardisation of linguistic aspects of lexical semantics will succeed in providing sufficient criteria and guidelines for assessing the lexical semantic resources of different types, both internal evaluations, and evaluations relative to an application.

Comments and questions by both discussants and participants to the panel have already proved to be useful in shaping the ongoing work of the group, and we look forward to a further validation event at COLING/ACL in Montreal where further developments of this work will be discussed.

Panel on the Need for Maintenance of Language Resources

Catherine Macleod, NYU, COMLEX

Participants

Chair: Catherine Macleod (NYU, COMLEX), Lou Burnard (BNC), Khalid Choukri (ELRA), George Doddington (SRI/NSA), David Graff (LDC), Nancy Ide (Vassar, TEI, CES), John McNaught (UMIST, EAGLES), Antoine Ogonowski (ERLI, Parole-Simple), Richard Piepenbrock (Max Plank, Celex), Hozumi Tanaka (Tokyo Institute of Technology, GSK)

This panel resulted from a paper by Catherine Macleod, "A Plea for Consideration of Maintenance of Language Resources" (Macleod, 98). This paper was written with the input of many designers of Language Resources (LRs), therefore it seemed appropriate to schedule a panel on this topic to let the resource creators speak for themselves.

The panelists gave 5 minute talks, in four areas: lexicons, corpora, standards, and funding and distribution organizations. It ended with 20 minutes of discussion, with questions and statements from the floor.

Catherine Macleod gave an introduction and also described dictionary maintenance needed for COMLEX Syntax. This consists in:

1. ongoing maintenance for errors, perhaps a yearly update, relatively inexpensive.
2. additions/enrichments, e.g. classification as to British usage, signi ficant funding required.

The next speaker **Antoine Ogonowski**, spoke of the "Le Parole" and Simple projects. The maintenance concerns for these lexica are:

1. after the development period, the project structure disappears
2. maintenance of direct contacts with users
3. feedback from Simple to Parole is desirable. Through ELRA?

He proposed that these resources be supported by national funding, selling directly to external customers and through ELRA.

Richard Piepenbrock spoke about the Celex dictionary and its need for maintenance.

Next, **Lou Burnard** spoke about the British National Corpus (BNC). The commercial partners of the BNC are not interested in

maintenance or development and the academic partners only get funding for research.

This is a common problem for all the resources. Given some support, the BNC could improve its tagging and documentation, link the speech and transcription, develop access software and expand the user base.

Standards were discussed by **Nancy Ide** and **John McNaught**. Nancy points out that standards also need maintenance. The TEI Guidelines have not been updated to fix small bugs nor have any extensions been added. For CES, no funding is available to continue development, though the development cycle of the CES was intended to include several phases of use, feedback and modification.

John McNaught spoke for EAGLES about the need for standards and the need to maintain and develop them. He noted that standards evolve and are superseded and thus need to be maintained. He also mentioned quality checking by standards and suggested that funding for maintenance of resources might be minimally conditional on meeting available standards. Standards should become more important as more and different resources are developed. Because they are critical for utility, maintenance and long-term viability of resources, work on standards must be extended.

The next speakers were from funding and distribution organizations.

David Graff from the U.S. Linguistic Data Consortium (LDC) noted that different types of corpora can be divided into low maintenance (collections of speech data), medium maintenance (newswire text collections, existing text archives) and high maintenance (speech transcripts, lexicons, manually annotated text). Maintenance for the latter group is labor intensive and often requires specialized workers.

Khalid Choukri spoke about the missions and contributions of ELRA. He insisted that maintenance be understood in a broader sense as it involves technical, commercial, and legal issues among others. He asked if a new copyright should be required when "corrective maintenance" is performed by a user. Should maintenance as customization of a generic resource, done by the owner at the request of a user, lead to new licenses? The owner usually has the knowledge needed to maintain the resource but he needs the feedback of the users and funding. Who supplies the funding: customers, producers, funding agency?

Hozumi Tanaka spoke of the creation in Japan of a new consortium, Gengo Shigen Konsooshiamu (GSK Language Resource Consortium) which hopes to gather and maintain language resources which he sees as part of a global consortium network including ELRA (Europe), the LDC (USA) and the GSK (Asia).

The last speaker, **George Doddington**, suggested that resource development and maintenance be done as part of sponsored research projects. He noted that research and resources interact, that language science and technology are immature and therefore that resource definition and development must proceed in concert with research.

Language Resources are inherently always "works in progress" and that, given the current state of immaturity and the (hopefully) rapid progress in the area, it is inevitable that resources will evolve dramatically. Therefore, he argues against "maintenance" and for resource development being part of research. Doddington asserted that computer mastery of language will depend on an evolutionary accretion of knowledge. This whole process in turn will depend on broad adoption of and contribution to linguistic resources and the generally accepted conventions and standards on which they are based. To attack the language understanding problem without building on prior contributions is to ensure failure. Thus the ultimate success of language research absolutely and critically depends on maintenance of language resources.

The ensuing discussion touched on the problem of maintenance funding, it was urged that users become involved. Another important concern is finding a way to evaluate LRs to determine whether to continue funding or to abandon the resource.

Bibliography

Macleod Catherine
"A Plea for Consideration of Maintenance of Language Resources" Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Spain, May 1998 pp.35-43.

LREC Technical Sessions Summaries

Machine translation Evaluation

Steven Krauwer

Three papers, three different views, but all based on the same observation: Machine Translation is hard, evaluating MT systems is even harder, so let us try not to do everything at once.

In their paper "A Task-Oriented Metric for Machine Translation", **John White** and **Kathryn Taylor** have chosen for a very pragmatic approach to MT evaluation. Rather than asking whether the output of an MT system is good (a very subjective question to answer) they ask themselves what the output might be good FOR. If the result isn't good enough to serve as a proper, publishable translation, it might still be good enough to serve as a basis for summarizing, for capturing key information, for ranking documents by importance, for identifying documents of interest, or for discarding irrelevant documents.

These five possible usages of translation output, in the order given, define a quality ranking of MT output (in decreasing order). This ranking alone, based on a range of typical text-handling tasks, is already an interesting contribution to the MT evaluation discussion in that it very clearly reflects current views on MT, where MT is not just seen as an isolated activity, aiming at simulating some process normally carried out by skilled humans, but rather as a step in a chain of processes, where the usefulness of the output of one step is determined by the requirements of the next one.

But the authors go one step further, in that they have determined empirically what sort of translation problems will cause translation output to be suitable for which text handling tasks, and which patterns in the source text will make these problems likely to occur. On the basis of these patterns a test suite was constructed.

The test suite is run through the MT system to be evaluated, and the result is scored by

human experts. The result will be a score on the MT Proficiency Scale, which will indicate what sort of text-handling tasks the MT system is suited for. It is clear that this procedure is crucially dependent on the validity of the test suite for this purpose, but once this has been established, it provides a quick, inexpensive and portable diagnostic set to predict the suitability of an MT system's output for real use in specific text-handling tasks.

Ed Hovy, in his paper "Creating Useful Evaluation Metrics for machine Translation", opts for a different approach, although it seems to be based on the same philosophy that MT isn't just one monolithic activity, but rather a collection of possible activities located in different places in a multidimensional space. The various dimensions are organized into a taxonomy of ever-increasing specificity, with appropriate evaluation measures associated with each level of each branch. This picture allows for the definition of various types of users of MT or usage context, each of which will have their own requirements in terms of the desired specificity, and to the relative importance of each of the dimensions for different types of users.

It was interesting to observe that this approach comes very close to the views on evaluation developed in the context of EAGLES and the related projects. It is unfortunate that the paper is not included in the Proceedings, and one can hope that it will be published elsewhere in the near future.

In their paper, "Evaluating Text-type Suitability for Machine Translation: a case study on an English-Danish MT System", **Claus Povlsen**, **Nancy Underwood**, **Bradley Music** and **Anne Neville** present a tool to predict how well an existing MT system would perform on a new text type it was not originally designed for.

Here again, the approach is very pragmatic in that the aim is not to give the ultimate evaluation of the quality of an MT system, but rather to answer the much more modest question "Would this system be good enough for this task".

The basis of their approach lies in the notion one could describe as "post-editing complexity": the seriousness of errors according to the extent to which MT post-editors found them disturbing (and presumably time-consuming to correct).

As a first step, users and post-editors were asked to identify typical classes of errors. As a second step the post-editors were asked to score the error-types according to their disturbingness.

The error types were not restricted to syntactic or linguistic errors, but could also include very superficial aspects such as layout.

Error types were described in terms of the way they manifested themselves in the source text.

The most interesting aspect of this paper is the semi-automatic tool that was developed in order to scan the source text, detect sources of possible errors as included in the list, and assign an overall "post-editing complexity score" to the source text.

This score would help to predict for new text types the post-editing complexity, and hence the suitedness of the system to translate texts of the given type.

The results of this enterprise looked very interesting, but could not yet be properly interpreted as the validity of the post-editing complexity score as a true predictor for the actual complexity of the post-editing process had not yet been established.

Steven Krauwer
ELSNET, Utrecht, the Netherlands
Email: steven.krauwer@let.ruu.nl

Language Resources: Policy Issues

Gary W. Strong, U.S. National Science Foundation

This session of the First International Conference on Language Resources and Evaluation featured five speakers from a variety of institutions dealing with language resources. **Dimitrios Theologitis** spoke on "Linguistic Resources at the European Commission Translation Service". A unique aspect of this service is that they translate legal documents that must have exact translations, even of the jargon that may occur in the documents. The result is a large quantity of parallel linguistic data. Some current issues of concern are how to deal with draft versus final documents, pricing of the service, and copyrights held on original work.

Poul Anderson spoke on "Language Engineering and Multi-lingual Issues: Cooperation with Central and Eastern Europe". As language engineering efforts proceed, there is increasing demand for resources that bridge the Central and Eastern European countries' languages with those of the European Union.

Khalid Choukri's talk was entitled "ELRA: From Infrastructure to Market Demands" and concerned problems in the distribution of language resources to value added retailers in the European Language Resources Association. There is a distribution of activities involved, from licensing of various kinds to the commissioning of new resources that support both research and commercial users. Currently, ELRA makes available 70 speech databases, 133 written databases, 361 terminology databases, and 2 tools.

Mark Liberman discussed "The Creation, Distribution and Use of Linguistic Data: the Case of the Linguistic Data Consortium". This consortium is hosted by a university and has a rich collection of resources that it makes available through membership fees and piece sales. The LDC holds the principle that no resources will be denied to researchers who need them. A primary effort of the consortium has been that of negotiating intellectual property rights so that members may have access to data. There are several new efforts underway, notably the collection of Voice of America broadcasts and Radio Marti, data whose collection by the LDC was recently permitted by the US Congress. One novel feature of the LDC is the ability to search data and access samples over the World Wide Web.

Finally, **Tarcisio Della Senta** discussed "UNL: A New Electronic Language for the Internet". This is an effort within the UN University, now based in Japan at the Institute for Advanced Studies. The United Nations spends a great deal of money on translation services since there are six official languages in which its activities are conducted. As a result of translation activity, there is an enormous thesaurus available at the UN in 6 languages. The Universal Networking Language (UNL) is an intermediary electronic language that serves as an interlingua between translated languages. The domains of its use include science, health, and engineering. The effort is two-years old and expects to conduct its first test near the end of 1999.

Gary W. Strong
U.S. National Science Foundation
Email: gstrong@nsf.gov

Lexical projects

Eva Hajikova

The session concentrated on descriptions of several projects, both on the international and national levels, concerned with issues of creating large morpho-lexical and syntactic resources. The broadest project reported was that of Multext-East, describing morpho-lexical specifications of six CEE languages, including different language types and families (Romance, Finno-Ugric and Slavic). The PAROLE Italian syntactic lexicon, as described by collaborators of ILC in Pisa, was a very important step forward, leading to syntactically tagged corpora. The PAROLE project was also one of the sources of the large-scale lexicon for Danish, as developed in the Center for Sprogteknologi in Copenhagen. The large-scale onomasticon, as developed by the Computing Research Laboratory of New Mexico State University, is a broadly conceived multi-lingual project, which is intended to help a NLP system to process proper names. The Habanera knowledge based managements system, developed at the same Laboratory, is supposed to be used as a central repository of multilingual lexical data based on most different resources. An applicationally oriented project on text editing in Japanese, called Writer's Helper (Yokohama), was aimed at a user friendly tool encouraging the Japanese user to expand his vocabulary and improve his ability to express himself in English.

The papers presented and the discussions demonstrated that to make resources really reusable one has to base the annotations or the information included in the lexicon on a reliable and well-founded linguistic analysis of the given languages, even if not all the information gained by such a research is immediately applied.

Corpus projects

Truus Kruyt

The presentations in this section involved four reports on corpus projects and two on corpus encoding and data architecture.

Diana Santos (diana.santos@ilf.uio.no) reported on the Oslo Corpus of Bosnian texts, which is accessible via a web service (http://www.tekstlab.uio.no/Bosnian/Corpus.html). The corpus consists of a variety of text types. Use is restricted to research purposes. Focus was on the architecture and functionalities of the service system. A clear-cut distinction is made between the proper corpus contents, the technical corpus encoding scheme (for loading the corpus data in a query system) and the web user interface. The query system used is the IMS Corpus Query Processor (http://www.ims.uni-stuttgart.de/CorpusToolbox/). The functionalities of the web interface are described and evaluated according to the parameters ease of use, availability of documentation and help functions, query power, speed, and display of results. The corpus is not (yet) linguistically annotated, which implies that queries based on POS or other linguistic levels are not possible.

Two presentations concerned corpus development in Japan. The Text Subgroup of the Real World Computing (RWC) Database Workshop has been building monolingual Japanese text databases since 1994, for the sake of research and evaluation of various technologies. Five text databases were characterized with respect to contents, morphological analysis, partial syntactic analysis and text classification. The texts mainly concern modern-Japanese reports and newspapers. The morphological tagset, including 16 POS, is designed to serve as the basis for many purposes (convertable to other tagsets). The partial syntactic analysis involves the transformation of real-life (complex) sentences into 'simple sentences', being bundles of dependency relations between nouns and predicates (verbs and adjectives). For text classification, the UDC (Universal Decimal Classification) has been used allowing for a multi-dimensional encoding of texts. Much of the morphological annotation is manually post-edited. For the future, the aim is the extension of linguistic annotation.

The other Japanese project, presented by **Hitoshi Isahara** (isahara@crl.go.jp), is JEIDA's (Japan Electronics Industry Development Association) English-Japanese bilingual corpus, a sentence-aligned corpus for NLP research. After a pilot project in 1996/1997, a new project aiming at a very large and improved corpus started in April 1998. The Japanese texts are white papers from Japanese Ministries. The English translations are merely sentence-to-sentence or paragraph-to-paragraph translations. The corpora are converted to TEI-format. Sentence-aligned data is developed by automatic processing using an alignment tagger and by manual post-editing. The aim is to add more precise tags to the bilingual corpora, including word and clause alignment tags. The corpora are developed to be available without charge to the public for research and evaluation of NLP technology.

The BAF corpus, presented by **Michel Simard** (simardm@IRO.UMontreal.CA), is a sentence-aligned corpus of English and French translations (ca. 800,000 words covering four genres). It is available from RALI's Web site http://www-rali.iro.umontreal.ca. In contrast to other aligned corpora, it has been aligned manually, so as to be a reference corpus functioning as an evaluation tool for automatic bilingual text alignment methods. Alignment is conceived as the parallel segmentation of the two texts, into an equal number of segments. "Sentences" include not only syntactically autonomous sequences of words, but also titles, enumerators, items of an enumeration and separate cells of a table. Various strategies have been used for situations where the order of sentences is not the same in the two texts, or where in one of the texts a sentence is omitted or inserted. Text segmentation and alignment were performed in parallel, in a computer-assisted environment. The corpus is available in the original COAL-format, a CES-format and a HTML-format. Apart from improvements of the present methods, word-level alignment is also aimed at.

Rather than on a corpus project, **Nancy Ide** (ide@cs.vassar.edu) reported on the Corpus Encoding Standard (CES). The CES is being developed to provide encoding conventions for corpora intended for use in language engineering ('corpus' defined as 'any collection of linguistic data'). The design principles include processability, validatability, consistency, and recoverability of the source text. The CES is an application of SGML. With respect to the TEI, parts appropriate for corpus encoding were selected and some extensions were made. The CES covers three levels of encoding. The minimum level required for a corpus to be standardized is the encoding of the gross document structure down to the level of the paragraph. The other two levels concern the paragraph and sub-paragraph level. Linguistic annotation is preferably retained in separate SGML documents linked to the SGML-encoded original text. There are different CES DTD's for corpus text, linguistic annotation and parallel text alignment. The CES is being updated for conformance to XML. Full documentation is available at http://www.cs.vassar.edu/CES/.

Dan Cristea et al. (dcristea@infoiasi.ro) presented an encoding scheme for discourse structure and reference, based on the TEI and CES and realized in an SGML/XML format. The annotation architecture enables multiple views on a document. A "hub" document (HD), encoded according to a slightly extended CES level 1, is referenced via inter-document links by a family of documents, each containing an additional view of the HD (directed acyclic graph with the HD as its root). An annotation tool GLOSS supports this view-based scheme. The encoding conventions for reference annotation

and discourse structure are based on the principle of separation of segmental markup (identifying the units of interest for a given study) and relational markup (identifying structural constraints between the units). The annotation scheme has been implemented to encode a small corpus of texts in English, French and Romanian. This corpus was used to study a model of discourse cohesion based on Veins Theory.

Some conclusions

1. Text corpora become more easily accessible, provided that copyright does not impose restrictions on what is technically possible.

2. Morphosyntactic annotated corpora are available to a fair extent. Focus will be on syntactic and semantic annotation.

3. Encoding standards like SGML, TEI, CES and XML are becoming more widely applied. They need to be extended cq. established for, among other things, the higher linguistic levels and discourse.

4. As for parallel corpora, sentence alignment is to be refined to clause and word alignment.

Truus Kruyt
Institute for Dutch Lexicology INL
Leiden, The Netherlands
Email: kruyt@rulxha.leidenuniv.nl

Tools for Natural Language Processing

Ulrich Heid

This report summarizes the main lines of the discussion held in the section on tools which was organized in the framework of the first international conference on language resources and evaluation, in Granada, in May 1998.

The section comprised two papers on part-of-speech-tagging, one on parsing of unrestricted text (This paper could unfortunately not be presented. See however the Proceedings for more details.), two on Natural Language Processing (NLP) application development environments and one on the customisation of linguistic resources by semi-automatic means.

This section in particular, but also in a more general way the whole congress showed that that actual development of corpus-related tools, such as part of speech taggers, lemmatizers, chunkers or robust parsers is in itself becoming a bit less of a research issue, than it used to be, for example a few years ago, whereas the evaluation of such tools, their embedding in application development environments or their use for languages so far less "well-resourced" than the major western European ones seem to gain in interest. Evidently, such topics are closer to the heart of a linguistic resources congress, than the actual development of techniques and methods; but in a more general way it seems that there is a basic familiarity in the community with technologies underlying, for example, part-of-speech tagging.

The two papers about tagging both dealt with ways of improving tagging quality, on inflecing languages (Romanian in the case of **Tufis/Mason**, Spanish in the case of **Pla/Prieto**). Quality improvements are sought by inclusion of some type of low-level syntactic information, be it by use of a linguistically informed guesser and "tiered tagging" (first use a small tag set with 89 tags then a larger one with over 600 tags), or, in the case of **Pla** and **Prieto**, by use of grammatical inference based on a regular grammar with statistical information, which allows to include more linguistic information into a statistical tagging process.

Along with the methods, practical implications, from the point of view of tagger training and assessment, were discussed: **Tufis** suggests to measure text complexity with respect to tagging difficulty, by means of identifying the number of potential ambiguities, in order to be able to correlate tagging precision measures with the measure of text complexity. **Pla** discussed the question of the size of the training set necessary to obtain good results (more than 80,000 word forms are necessary): what we gain by introducing more linguistic information is indeed obtained through more work on the preparation of training material.

The papers on NLP development environments focused on aspects of language engineering and the constructive use which can be made of software engineering principles in the development of NLP applications. Aspects of resuability, customization and easy combination of different components are the focus of work by **Prodanof** et **Al**. This is exemplified by a whole range of NLP applications, going from phrasal analysis (chunking) with a view to the extraction of linguistic information from large corpora, over parsing of a restricted fragment with the aim of a translation into predicate argument structures, to a module for the analysis of queries in the framework of conceptual information retrieval. In all cases, variants of the same very large word form lexicon (represented in a data base) are combined with different types of grammars, in a development interface. The work on GEPPETTO, presented by **Pianesi Etal**, focused more on the methodology of the development of NLP applications than on individual examples of the development process. The authors implemented a methodology based on the software engineering life cycle for the development of LE applications, including a very detailed definition of the tasks of the individual participants in such a development process, a clear definition of requirements and specifications, as well as tools and methods to support application development in practice. A case study from the development of an information extraction system is used to exemplify the approach, which has been applied successfully to numerous development tasks already.

The last presentation in this section was devoted to questions of the engineering of the reuse of lexical resources, focussing on possibilities of automatically constructing a large lexical resource for French from the tables provided in the framework of Lexicon Grammarfor the syntactic description of French verbs, by the group of **Maurice Gross** (LADL). The idea is to provide a translation tool which takes the tables produced by *Lexicon Grammar* as an input and converts the content into a lexical representation from where application lexicons for different NLP-oriented theories can be derived, such as HPSG or Tree Adjoining Grammar. The tables are rather complex specifications containing among other defaults, blocking and overwriting, as well as negative constraints. The first task is thus, very much like ten years ago in the reuse of printed dictionaries, to provide a thorough and consistent interpretation of the descriptive devices used by *Lexicon Grammar* tables. This allows to define translation rules towards the internal lexicon format, from where then reuse by means of an additional mapping towards the application specific formats is easy, including different ways of packaging syntactic information, as required by the two, partly rather different application systems. So far one third of the tables of *Lexicon Grammar* has been converted; methods and measures for a detailed evaluation are discussed in the last part of the paper.

According to the orientation of the congress, the section on tools was maybe not the most central one. However, it clearly shows the move towards refinements of the existing technologies, for part-of-speech tagging, for example, as well as towards the embedding of existing technology and modules into NLP application development. The reuse of existing theory-based resources is a very interesting topic, all the more because it nicely fits the overall attempt of the community to provide better resources in a more efficient way.

Ulrich Heid
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Computerlinguistik, Azenbergstr. 12
D 70174 Stuttgart - Germany
Email: uli@ims.uni-stuttgart.de

Spoken Language Systems Evaluation

Nick Campbell

There were five papers in this session, which was chaired by **Nick Campbell** from ATR-ITL in Japan.

The first paper, by **L. Pols et al**, described the use of large corpora for evaluating text-to-speech systems. This presentation described a TTS Web-site that has been set up as part of the COCOSDA/ESCA 3rd Speech Synthesis Workshop to be held in Jenolan (Australia) in November this year. The talk compared different text types and stressed the need for both diversity and random-selection in order to provide fair coverage for potential test materials. Discussion from the floor stressed the need for including non-text (i.e. speech and discourse) materials for evaluating speech synthesisers, and for modular evaluation of synthesis components in general.

The second paper was jointly written by 15 researchers representing 9 labs in 4 different French speaking counties. It reported joint work on the Evaluation of Grapheme-to-Phoneme Conversion for French Text-to-speech Synthesis under the aegis of the Francil network. In all 8 systems were compared. The paper presented a state-of-the-art review of problems still remaining and established benchmarks for future development. Following on from the previous paper, this presentation offered fruitful suggestions as to how a component evaluation can be performed. It also provided useful pointers to resources and data which are now placed at the disposal of the synthesis community. An important point arising from the discussion concerned the mutual benefits to the groups involved from such communal evaluation.

The third paper, from the ICP in Grenoble, was presented by **Yann Morlec** and discussed a methodology for evaluating the quality of prosody in synthetic speech. The paper presented clear results of well-designed experiments, but might have been better as a poster, allowing the author more time to explain the methodology and background. The written version of the paper is very clear, and reading is recommended. Discussion mainly concerned ambiguity of equally acceptable but meaningfully different possible alternative prosodic patterns.

The fourth paper discussed speech quality evaluation in Slovenian TTS, presented by **Jerneja Gros**. The talk focussed on differences from an earlier version of the Slovenian synthesis system and showed using the results of MOS tests that both intelligibility and naturalness were improved. In addition to the opinion scores, tests also included a dictation component using frame sentences and fillers suitable for an airline announcement system. A demonstration comparing the old and the new systems unfortunately failed because only samples from the old system were found on the tape.

The final paper in this session was presented by **Lise van Haaren** and concerned Evaluating the quality of Spoken Dialogue Systems, comparing a technology-focussed and user-focussed approach. Two different sets of evaluations of the same Train Scheduling Information Service showed the different expectations of users and developers. However, in spite of the immediate apparent differences (as outlined in the talk) it was encouraging how much the different groups actually agreed. The paper presents an interesting study not just of the methodology of system assessment, but also of the psychology of interpreting the results.

In summary, this was a lively session with plenty of discussion. We were fortunate in the selection of papers as all continued a single theme and discussion topics could be carried forward from each paper to the next. The theme of the first paper set the tone for the session, and a variety of useful suggestions and comments ensued.

Evaluation is certainly a major topic for speech synthesis and we are encouraged that COCOSDA and ESCA will be featuring it strongly in the forthcoming 3rd International Speech Synthesis Workshop later this year.

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

Post-LREC Workshop Summary

Multilingual Information Management : Current Levels and Future Abilities

Antonio Zampolli, on behalf of the Editorial/Organizing Committee
(E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Martin-Municio, A. Zampolli)

Over the past 50 years, a variety of language-related capabilities has been developed in machine translation, information retrieval, speech recognition, text summarization, and so on. These applications rest upon a set of core techniques: it is a puzzling fact that although all of this work deals with language in some form or other, the major applications have each developed a separate research field.

The most effective way to change this situation, and to ensure that the various techniques and applications fit together, is to start talking across the artificial research boundaries. Extending the current technologies will require integrating the various capabilities into multi-functional and multi-lingual natural language systems.

However, at this time there is no clear vision of how these technologies could or should be assembled into a coherent framework.

The purpose of the workshop was to address these questions, in an attempt to identify the most effective future directions of computational linguistics research and, in particular, how to address the problems of handling multilingual and multi-modal information. Experts in various subfields from Europe, Asia, and North America were invited to present their views regarding the following fundamental questions:

1. What is the current level of capability in each of the major areas of the field dealing with language and related media of human communication?

2. How can (some of) these functions be integrated in the near future, and what kind of systems will result?

3. What are the major considerations for extending these functions to handle multi-lingual and multi-modal information, particularly in integrated systems of the type envisioned in (2)?

The experts were invited to represent the following areas:

- multilingual resources

- information retrieval, especially cross-lingual and cross-modal
- machine translation
- automated (cross-lingual) summarization and information extraction
- multimedia communication, in conjunction with text
- speech processing, especially multilingual
- evaluation and assessment techniques for each of these areas
- methods and techniques (both statistics-based and linguistics-based)
- government: funding and development policy

In a series of ten sessions, one session per topic, the experts explained their perspectives and participated in a panel discussion that attempted to synthesize the material and hypothesize about where we can expect to be in a few years' time.

Each thematic session has been the responsibility of an area editor, whose task was to compile all the presentations, notes, and comments, into a chapter of a report which, after public discussion and critique at subsequent conferences (in particular, the COLING-ACL in Montreal), will be presented to representatives and funders of the US and European Governments and other relevant associations and agencies.

The US Government and the EU have recently signed a transatlantic Science and Technology Agreement. The need for international cooperation was a recurring theme throughout LREC.

The goal of the last session of the workshops (on "Governments") was to consider and compare the organization principles and the goals inspiring LT R & D programs of major Funding Agencies, on the basis of the outcome of the preceding sessions, to discuss and collectively identify issues for which transatlantic cooperation is primarily needed, and eventually to indicate concrete proposals for joint initiatives, providing, in this way, suggestions to the Funding Agencies which have the task of defining the cooperation policy.

The session, introduced by Antonio Zampolli, started with two panelists (Giovanni Varile, EC-DGXII and Gary W. Strong, NSF), followed by 5 discussants (Charles Wayne, National Security Agency; Lynn Carlson, US Department of Defense; Khalid Choukri, ELRA; Joseph Mariani, LIMSI-CNRS; Nicoletta Calzolari, ILC) and then a general discussion in which the following issues were unanimously recommended for transatlantic cooperation:

1.Standards (de facto, best practice)

Standards for language resources are seen as essential for LT and for the development of mono and, in particular, multilingual applications. Unified standardization efforts are required, one of which exists in Europe: EAGLES (whose results and recommendations are already adopted in other countries). Several participants have proposed that the U.S. join EAGLES as soon as possible, which is the initial critical step.

2. Language Resources and Related Tools

International cooperation for Language Resources is the key that can open the door to a true multilingual society. Language Resources, mono and multilingual, multifunctional (i.e. to be shared across different types of LT applications) are unavoidable issues for cooperation.

Priorities shall be given to:

- Computational lexica (mono and multilingual, general and domain specific, but possibly based on compatible models) both written and spoken.
- Mono and multilingual corpora, both general and task/domain specific, spoken and written, and especially national corpora developed in close coordination among the countries involved.
- Related research and methods and tools for acquisition, annotation, maintenance, development, customization, etc.

The inclusion of semantic knowledge (semantic annotation of corpora, semantic information in lexical resources) is an urgent need: in this area coordination both between ongoing development activities and on research aspects is crucial.

The value of Language Resources suggests that Language Resources are considered as a research and development area in itself and the production of Language Resources is financed 100%. An international distributed networked infrastructure should be in place.

3. Evaluation

Evaluation is required by the HLT researchers and developers, in order to measure the status of technology and the progress made. Evaluation applies to methods, technologies, components, systems, applications and both the developers and the users should be considered. A good basis for cooperation exists, with a complementarity between the US (competitive evaluation) and the EU (standards, general methodology, and the user and usability perspective - considered in particular in the EAGLES Evaluation Working Group) experiences.

Several participants also mentioned the need for cooperation in the development of core technologies both in speech and written areas (for example, integration of statistical and rule-based approaches, word sense disambuguation, dynamic acquisition of linguistic knowledge from corpora, transfer of technologies between application domains and languages, reference architectures for integrable systems) and in vertical domains (in particular education, tourism, access to cultural resources, language learning, digital libraries, e-commerce) where priority should be given to the integration of application systems and to multilingual applications.

Proposals for cooperative projects can be summarized as follows:

Standards: American participants to join EAGLES Working Groups immediately

Language Resources: Cooperation in building lexica (e.g. Framenet–PAROLE/SIMPLE) and corpora (e.g. BNC-ANC-PAROLE); tool development; research (identification of priorities, innovative LR, e.g. semantic, multimodal); networks of Language Resources centers and organization (e.g. ELRA, LDC, PAROLE, SPEECHDAT, etc.).

Evaluation: Integrating cooperative evaluation and EAGLES approach expertise; topic spotting from broadcast news; multilingual TREC with European participation).

Core Technologies: Automating learning methods from corpora; robust analysers; customization of Language Resources.

A Conference was organized in Washington, D.C. June 8 – 9 at the National Academy of Sciences, to celebrate the signature of a new Science and Technology Agreement, which should be ratified in October by the European Institutions. A section of this Conference was dedicated to "Translingual Information Management", with the participation of invited American and European experts. In reporting to the plenary session on the discussion (which also took into consideration the outcome of the Post LREC Workshop basically reaffirming its suggestions), Gary W. Strong (NSF), on behalf of the participants, summarized the results in 4 points:

Motivations: Removing language barriers in the global information society (people-people, people-data); building and developing on complementary efforts.

Goals: More rapid international progress; standards for interoperability and Language Resources sharing and integration; networking for Language Resources centers; fuse user-centered and technology-based evaluations; develop reference architectures.

Plans for early cooperation: Develop common Language Resources; create Language Resources development tools; cooperative work in planning reference architectures and developing standards; joint evaluations.

Applications Domains: Education (in a cross-cultural framework; access to cultural resources); environmental data sharing; international digital libraries; e-commerce.

Language Resources and Evaluation
''Declaration of Granada'': 10 Articles

1. At this moment, language resources are one indispensable key to unlock the potential of the global Information Society.

The most important single fact in the world economy at the end of the second millennium A.D. is the massive growth, and growing interpenetration, of information and communication technologies. Within that growth, there are changes in profile: value-added services are growing relative to straight transmission and storage of information (even in telecommunications companies); and the highest growth is concentrated outside English-speaking domains. All these trends place a new focus on automatic processing of content expressed in human languages, spoken and written. And the experience of the last decade has shown that effective processing of language content at any level is impossible without extensive use of authentic language resources, for look-up, experiment, and training of systems. In future, new degrees of run-time accessibility may extend their role into language understanding itself. And the quality of these resources is crucial to the quality of applications built on and with them. They constitute an essential infrastructure.

"Language Resources" are understood to include full-scale dictionaries, software to analyse and generate language structures, and extensive selections of language in use, as well as systems to access, manage and update all of these.

2. All sectors of society, and all languages, have an interest in seeing these resources developed, for a variety of purposes, economic, social, industrial and cultural.

Since human languages are the primary vehicle for all business, social policy, education and culture, at local, national and global levels, the new modes of language use are of concern and potential value to everyone. There are vast commercial opportunities, and not only for large corporations. Democratic governments will wish to ensure that all their citizens' interests are protected, at all ages of life. Special interests and regional communities also have a concern for representation of their languages, varieties and terminologies in the data and tools which are essential to language processing. The public interest arguments require that the core of language resources should remain available in the public domain under warranty of public authorities, although there is a place for private ownership in those that will support specific product areas.

3. Like human languages themselves, such resources are necessarily large-scale, and require a wide range of participants.

Language resources are by their nature large objects. No common authority is capable of making full central provision for them. There are too many imponderables, such as the balance between active use and passive exposure, between individual variety and codes characteristic of communities and sectors, and the very different traditions and present status of languages. Hence these resources cannot fulfil their potential if large corporations or sectoral interests end up dictating their provision, or dominating their funding. They may in fact be non-sectoral, independent of application, and universal in scope. As such, they are properly a field for co-operation among governments, companies and others.

4. Although they are essential to realize the growth of private enterprise, they will not, indeed cannot, emerge simply from the sum of individual projects.

By now, considerable experience has been accumulated in scientific centres around the world, but particularly in North America, Europe and Japan. Using this is a necessary condition for rapid progress in spreading this expertise into applications within new sectors, and into new languages. Increasingly, language resources, and the skill to design and provide them, will be a precondition for more specific applications. One immediate example lies in speech processing, where some commercial applications (e.g. to medical pathology) are specialized, but rely on a general-purpose dictionaries and recognition algorithms. Bridges need to be built between work in individual languages and more general, language-independent, applications (e.g. in topic identification, library search and thesaurus building). It is not reasonable to expect such common resources to emerge simply from projects focused on immediate application. Instead, this work must be complemented by large-scale co-operative projects, at national level and above. National organizations, although free in their strategic choices, are encouraged to agree upon common lines. Private industry in telecommunications has already understood this, collaborating on speech data collection before competing to provide new services.

5. For each language, there is a need for strategy to co-ordinate existing resources and create new ones.

Language resources are often specific to individual languages. There is great variety in the situations of these languages. Such differences will include:

- legacy and the current stock of language technology available;
- their relation to government bodies, regional, national and international;
- their speaker populations: size, average wealth, use of ICT (Information and Communications Technology), familiarity with other languages (and hence the level of interest from large corporations);
- role in international communication;
- technical issues of character-sets and coding, and the state of art of local R&D.

There is then a need for planning of priorities at the level of individual languages. Paradoxically this planning is most necessary for the least spoken and least developed languages, where expertise will be least readily available.

6. When resources have been created, there is a continuing requirement for support and maintenance.

Language use continues to evolve, so that maintaining the currency of resources that represent it is a persistent task. More urgently, the use made of the resources in language technology will expand, no doubt in unforeseen ways. Standards will evolve to meet these new demands, and the resources will need to be revised to maintain a high level of accessibility.

7. These efforts for each language will benefit by taking into account, and profiting from, progress made in providing resources to underpin others.

Language technology for any language has a common scientific basis, although most languages pose some distinctive challenges, and for some, major technical innovations have been necessary (especially for Japanese, and the languages of the Far East). Hence developers of resources for any language can profit from our past experience, both technically and for project management. All languages also have the practical challenge of giving access to other languages (e.g. by translation). This puts a premium on widespread adoption of common technical standards (especially in dictionaries). Above all, the vast costs of large-scale resource selection and preparation demand that exchange and reuse of data must be a priority. (There may be especial benefits among closely related languages). And besides saving money, this will contribute

to empowering even smaller communities: all should have access to the language of each.

8. Understanding of the role, usefulness and optimum means of preparation for language resources is a research theme in itself.

Although language resources are amply proving their usefulness, and the encompassing scope of recent LE projects have shown that we have workable standards already for the challenging task of building workable resources, there remain major challenges in conceiving, demonstrating and standardizing resources for aspects of language use: for example, the role of language in multimedia. We are still searching for the best methods to customize resources, acquire linguistic knowledge for language resources dynamically in a continuous process, to incorporate new levels of annotation. Above all, effective design and preparation of multilingual data is an unsolved problem. Providing the best language resources will require intellect as well as unremitting toil.

9. This co-operative understanding will benefit greatly from the use of common standards for evaluation of resources.

In this time of explosive growth and dynamic large-scale change, the different techniques of evaluation become especially important. These enable R&D workers to compare different results, policy makers to assess the state-of-the-art, developers to identify real capabilities on which to base applications, and users to compare performance of products, especially for specific applications. Common evaluation requires common standards, at some level, in its input material: it is hence a stimulus, in itself, to inter-operability between systems. Moreover, the generality which results when resources use common standards aids reusability, and so increases economy in the production of resources.

10. Cooperation can take many forms.

The provision of language resources is a task which needs to respect the diversity of languages, and the diversity of purposes for which each of them is used. However, it is possible for any developers to benefit from efforts going on concurrently, or in the past, elsewhere.

Some examples:

- International, but regional, institutions such as the Linguistic Data Consortium and European Language Resources Association can compare the practical effects of their different structures, and where useful forge links for co-operation.
- National governments will have a role in mediating the results of national R&D to the business community, for example through the network of Chambers of Commerce.
- Leading nations can make their national programmes open to participation from outside, thus building de facto international standards through common endeavour. (The US and Japanese governments have a particularly good record in this respect, especially through the creative use of competitions against given evaluation standards.)
- Lesser used languages, as well as technologically emerging large linguistic communities, can take over technical infrastructure from those in a leading position, but also learn from each other how smaller languages may be sustained within larger political and social units.
- The institutions of the European Union can propagate technical standards and knowledge of best practice among groups throughout Europe; they can also create and finance partnerships for constructive work, as for evaluation and future collaborations that might carry on from current Resource building under the Fourth Framework programme.
- Explicit joint actions can be defined between the projects of national and international governments, putting the weight of national policy behind speculative research actions.
- Most generally, networks of projects and of developers can be set up at the grass-roots level (with scope ranging from local to trans-continental): links so forged may survive the lifetime of individual projects. There is already a wealth of links to build on, both within and beyond the European Union.

All of these are different mechanisms. Together, they and others will provide mutual support for the development of these expensive common goods, Language Resources. And they will serve to keep global aspects before developers' eyes as they devise new, and uniquely fitting, solutions for their own application, and their own language.

LREC Opening Session Speeches

Angel Martin-Municio's Speech

President of the Royal Academy of Sciences of Spain

Dear ELRA president and representatives of national and European institutions.

At this moment, I have no other mission than expressing gratitude together with the introductions. Certainly, in both cases, reiteration helps situate everyone and every institution in the place it deserves according to responsibilities and efforts devoted to the organisation of this first conference.

In addition, this conference is among the first international events that is taking place after the recent European economic treaty. There is no doubt that the content and development of this conference will constitute a key factor for the linguistic and industrial policies at national and European levels.

When the president and the board of ELRA have selected the city of Granada to hold this conference, an enthousiastic group of professors from the University of Granada, supervised by Natividad Gallardo, Rosa Castro and Antonio Rubio, set up all the necessary actions in motion, and benefited from an immediate financial support of the Fundaci¢n del Banco Central Hispano. They have also benefited from the support of la Junta de Andalucia through its General Director, Mrs Elena Angulo, to whom we owe the organisation of the visit of the emblematic monument of Granada: The Alhambra.

In addition to all these thanks, I would like to point out the importance of this conference for the Spanish language and obviously for the Universitary and Academic Institutions which are involved in its dissemination, for the political and administrative organisations in charge of its protection and for those responsible to position it in the modern field of science and technologies.

Demography of Spanish language, the number of its linguistic communities, the vitality of its litterature, and the high level of its normalisation process, including the level of the scientific research in various fields do not correspond to the attention given to the use of Spanish in the communication and information technologies, and to the promotion of initiatives and international cooperations. The crucial issues of the conference aim to lead to solutions or initiatives regarding a large number of problems under investigation and development of the language, including issues of interest to basic research activities.

In these times, when the interdisciplinary activities are a common motivation to the knowledge progress, when academic institutions are facing instability, when the multimedia cultural industry is emerging strongly in all the didactic resources, the Language Engineering area offers nowadays a large range of scientific, economic and social solutions.

As the representative of the Local Committee, I hope that our efforts and ambitions will lead to the most successful conference in favor of the largest European cooperation for the development of Language Resources.

Vicente Parajon-Collada's Speech

Deputy Director of DG XIII of the European Commission

Ladies and Gentlemen,

I am pleased and indeed honoured to be with you today as this Conference takes place in a most symbolic town, which witnesses so vividly the getting together of two civilisations. I am confident that this conference will help to lay down the *foundations of new, lasting collaborations* between the many countries and groups that you represent.

Over the last few years what has become widely known in Europe as Language Technology and Industry has achieved a broader recognition than ever before. The fact that Europe's Information Society, itself part of a truly global digital Village, can only be built upon *the mutual recognition of cultural and linguistic values and identities*, is now widely recognised.

Likewise, the *impact of language technologies and applications on business and everyday life* is largely undisputed. Every week, leading magazines and market analysts report on new applications of spoken and written language technologies. Major trade shows, both in Europe and elsewhere, feature new products exploiting language technologies. While we all know that many research problems still await solution, and that today's commercial solutions are far from being perfect, we are witnessing the acceleration of a process that will turn language systems and solutions into Key *enables* of an open, pluralist and truly *human-centred information age*.

European programmes

In recent years, the European Union has made a major effort and provided a *comprehensive framework* for research and technology development in the language field. Several of our programmes, including Telematics, Esprit, MLIS and Leonardo feature projects and other collaborative actions that directly or indirectly contribute to Europe's technical leadership in this area.

In the Language Engineering programme alone, *some 100 projects* have been launched since 1992, and more than *50 are underway* at this very moment, with the participation of some 500 research centres, companies and administrations. As tonight's panel will show, a small but important part of the Union's *International Cooperation* programme is devoted to linguistic research and engineering. Overall, one can estimate that since 1995 some 100 MECU have been invested in language R&D in European programmes.

However impressive, these figures must be seen in relation to Europe's share of worldwide multilingual services, which according to an OVUM report is expected to reach 6 billion US $ by the year 2000. Bear also in mind that the European language *research base* consists of some 10,000 specialists, and that the total R&D expenditure can thus be estimated at around 1 billion ECU per annum.

What was still regarded a few years ago as an immature research field, has now got its credentials and attracted the interest of *global market players*.

In parallel, major US and Japanese corporations have established language research facilities in Europe.

Thanks to the spectacular growth of the *Internet*, the importance of language in general and of language technologies in particular, has surfaced on the political agenda. Countries as diverse as Italy and Norway, The Netherlands and Finland have announced or are preparing *national programmes*. It is worth noting in this respect that by the year 2000, only 40 % of the Internet users will be based in the USA, as opposed to 55% today.

As you all know, electronic repositories of language knowledge, or *Language Resources* as we call them nowadays, play a crucial role in building, training, testing, and operating systems that can analyse, process or generate *human language in all its forms*. Here again, the European Union has invested some 20 MECU over the last few years, and launched large-scale projects such as PAROLE or SPEECHDAT, which have led to fruitful collaborations between public institutions and primary companies. I am delighted to see that these partnerships now provide a platform around which *new initiatives are being undertaken* at national and industrial level.

Multi-party collaborations

For language resources to serve their purposes, they must obey some common specifications, and be disseminated as widely as possible. The pioneer work done by groups like EAGLES and ELRA is worth a special mention in this respect, in that they have addressed what appeared just a few years ago as a major gap in the European research arena, and prepared the ground for larger-scale operations based upon private-public partnerships.

DG XIII of the European Commission has supported the establishment and early operation of ELRA, the European Language Resources Association, a forum open to all the parties interested in a wider availability of language databases and tools. ELRA's basic tasks - the collection and re-distribution of language resources of general interest, could now be extended to encompass both the creation and the validation of high-quality, multi-purpose resources. I am confident that this conference, initiated by the ELRA members, will provide a unique opportunity for reviewing current and future collaborations.

Indeed, for this process to continue and scale up in the coming years, *national agencies, industrial providers and commercial users* of language-enabled systems and services must play more fully their role, and contribute more actively to the creation and distribution of multi-purpose, multi-language resources. *Nobody can expect a single party*, let alone the European Commission, *to tackle a challenge of this scale*.

The new framework programme

If we now turn our attention to the Union's *upcoming research programme* that will take us into the new millennium, the Commission has released two weeks ago its proposal for the specific programmes to council and the European Parliament. All the research and technological development activities relating to computing, telecommunications and media are going to be *grouped together within a single specific programme*, which has become known as the Information Society Technologies (IST in short) programme.

One of the IST constituent elements (or Key Actions as we call them) is intended to address those research lines which are geared towards the creation, manipulation and delivery of *digital content*, in all its forms. *Human language technologies*, applications and resources are expected to be placed under this roof, along with other important research strands such as multimedia publishing and education & training.

Project clusters centred around language-enabled content processing, will bring together R&D work, demonstration projects and *infrastructural actions*, including those aimed at providing *shared language utilities and resources*.

I am convinced that the new programme will help forge new alliances and stimulate the development of *new skills*. It will provide a flexible and effective framework for *global endeavours*, more specifically for actions bringing together organisations from all parts of Europe, and from other regions of the globe.

Conclusion

Information Society Technologies are the d*riving force* for radical

transformations in business and society worldwide. Economies will experience an "Internet multiplier effect" where successful network-based applications, products and services are developed. To fully exploit this potential, it is essential that *Europe's Information Society* provides accessibility, usability and language appropriateness; this cannot be done without the widespread application of language technologies. While the language research base in Europe is unrivalled, the *challenge for Europe* is to turn this research advantage into useful and profitable *applications for* economy and society.

To conclude with a slightly provocative remark, you all know that the impact and effectiveness of *publicly funded programmes* are routinely questioned. Our political masters, and the tax-payer, rightly expect the financial resources poured into research programmes to bear fruit and contribute visibly to Europe's prosperity and competitiveness. I am confident that this Conference will help re-assure them as to the value and *cost-effectiveness of ongoing and future R&D efforts* in such a fascinating and demanding field.

I wish you intense but enriching discussions over the next few days, and thank you for your attention.

Bernard Quemada's Speech

Vice-Président du Conseil supérieur de la langue française

Mr President, Ladies and Gentlemen, dear Colleagues,

In response to the honour made to me by the organisers of this Conference, who asked me to speak during this session, I would like to add two messages to the two former speeches. Do not worry, they will not be too long :

- the first message will come from one of the actors working alongside the Prime Minister on the French linguistic policy,

- the second message will come from a linguist-lexicologist who has been using and producing for many years a large quantity of computerised text resources.

For more than 10 years now, I have had the role of advisor for the French government on the subject of *Language Industries* which has now become the sector of *Linguistic Engineering*. And for even longer than that, I have, along with several specialists here present, in particular the President of ELRA, being waiting for a major mobilisation around the field of Linguistic Resources, which is so essential in our mind. That is why it is with great pleasure that I can state the success of the present Conference, organised for a scientific community which is both extending and diversifying : the participation of 500 people (which has largely by-passed the most optimistic expectations), but also the quality of the articles and the multitude of topics which are being addressed speak for themselves. I could see in this a comforting response to the recommendations made by the High Council for the French to its chairman, the Prime Minister, when it was created in 1989: *"Progress in automatic language processing demands a large amount of computerised linguistic data. These have to have a wide coverage and be representative of the varieties of the ways in which they can be used. At the same time, they have to be of top quality. Their return will be all the more effective as they will fully reflect the potential evolutions of the language, so that it is possible to update them on a permanent basis"*. Such demands could not be assumed in a satisfying manner by isolated initiatives from researchers or private companies; this underlined the need to organise joint activities in the field. However, large scale efforts in this direction have been implemented quite slowly.

This need, which had drawn the attention of the French and French-speaking bodies, is now one of the priorities of the Délégation Générale à la Langue Française, of which Mrs Anne MAGNANT would have talked this morning, if she had been able to come. The strong interest of France for this action has resulted in the outstanding support that we gave to the creation of ELRA. The success of this Conference seems to justify a posteriori the validity of this support and the fact that we would like other European partners to add theirs to ours.

I also believe that some fears that arose in the past, on the creation of ELRA, have now disappeared, even though some vigilance is still justified, in a field as vast and complex as this one. However, I feel that the decision-makers within the DG XIII and the French authorities, who made it possible for ELRA to be created (Délégation Générale à la Langue Française, Ministry for Higher Education and Research, Ministry of Industry) can be satisfied with the result of their intervention in favour of an initiative which was not risk free. In its action plan concerning the position of France in the Information Society, which was published in January, the French Prime Minister indicated the real importance of linguistic resources in the development and the evaluation of new software systems. He asked the Délégation Générale à la Langue Française to lead and co-ordinate the necessary actions for the French language in close collaboration with ELRA and the European programmes in the field.

The first goals given to ELRA have been achieved and the services provided by the Association satisfy both data suppliers and users. It can be legitimately hoped that the missions assigned to ELRA will receive the international acknowledgement they deserve from all disciplines in the field, over the following days, as well as those who have managed and co-ordinated these missions, who will find many reasons to be satisfied. I am thinking in particular about ELRA's tireless President, Pr. Antonio ZAMPOLLI who, along with Khalid CHOUKRI and all those in the University of Granada as in the Istituto di Linguistica Computazionale in Pisa, have contributed to the organisation before and throughout the conference, with the very efficient support of the President of the Academy of Sciences of Madrid, Pr. MARTIN-MUNICIO. All of these people deserve our sincere acknowledgement and I have the pleasure of expressing it to them here.

The importance given by the High Council for the French Language to "the computerisation of French" stemmed from the belief that failing to participate fully to the mutations affecting the information and communication technologies would result in very negative effects on the destiny of our language. In fact, we do know the disqualification that affected those languages that did not reach the written stage or the printing stage. This has been fatal for most of them.

But without minimising the economical, technical, social and educational stakes associated to the computerisation of our society, I shall insist more particularly on the cultural consequences. Already, the use of our language is receding, or even tends to disappear from major knowledge sectors, in particular those concerning the most recent fields in science and technology. Nobody can deny that if, tomorrow, science and innovation were no longer written up in French, this would result in a great intellectual loss for all French-speaking people as it would mean that, at the short term, we could no longer think about these topics in French. Here I am talking about French, but this is also true for most European languages. But the intellectual wealth of humanity relies on the diversity in the ways of thinking shaped by each language and on the various visions of the world that they convey. We can not willingly accept the decline of this common patrimony. Our duty is to continue to enrich it.

Therefore, we must join our forces against uniformisation, in order to preserve all languages, including, I insist, those that dominate today world-wide exchanges.

These are the reasons for which France has committed itself with determination in the PROMOTION OF EUROPEAN PLURILINGUALISM

and we expect the contribution of new technologies and of linguistic engineering to resist to some lethal forces currently in action within the Information and Communication Society.

It is therefore necessary to have access to LINGUISTIC RESOURCES IN EVERY LANGUAGE in order to support ALL CULTURES. The production of these resources, their standardisation, their evaluation and their distribution constitute major challenges that our community must face without delay. Taking into account the progress that has to be made in this direction, these challenges can not be taken on without joining efforts through co-operation, in order to facilitate standardisation, exchanges and reusability of what is produced, or of what will have to be produced, by each of the actors involved. Thus, my best wishes go to the ELRA project and to all the follow-ups which may arise from it.

My second message is going to be a more personal one.

Indeed, I can not forget the difficult times, which are not too far back, when the domain of linguistic applications of computer sciences was split between, on the one hand, *computer scientist engineers* (who often referred to themselves as *applied mathematicians*) dealing exclusively with the design of algorithms and, on the other hand, *linguistic data producers*, lexicographists, terminologists or speech and text analysts, who were rather weary of "machines". And, except in marginal cases, neither group communicated with each other. Today, all one needs to do is to glance through the volumes of the proceedings of this conference, to be convinced of the intensity of the exchanges between the various disciplines involved and of the real links that have been developed. Limiting myself to my own research field of lexicography and terminology, I can confirm that no qualified dictionarist would dare to ignore the work carried out to develop *machine dictionaries*; the dictionarist knows that he can find in them many elements that his own analyses could have missed out on. But as is only fair after all, the *machine dictionaries* owe a large part of their data to *traditional dictionaries*.

What satisfaction also for those who pleaded and acted in favour of co-operation and standardisation of the work carried out since the prehistory of data processing as well as for the exchange and the reuse of data thus produced, with such difficulty and at such heavy costs. I am very happy that these recommendations made over and over again have now become reality and I hope that they will progress even faster in spite of the legal and administrative obstacles which still exist nowadays.

I will conclude by expressing my warmest wishes for the success of this conference which I am sure will be remembered for a long time, all the more since our exchanges will benefit from the stimulating context of the magnificent town of Granada, which is as prestigious as it is symbolic. What is more, I am sure that our work will be enriching thanks to the representation of the various fields of our international community. And it will be even more enriching if researchers and theoreticians do not under-estimate the constraints which weigh heavily on the developers of industrial applications and if, from their side, private companies take on with professionalism the legitimate demands regarding the quality of the language, because they will surely be able to meet these demands tomorrow - or in the near future…

Thank you for your attention.

Speech of His Excellence Giuseppe Tognon

Italian Sottosegretario di Stato al Ministero dell'Università e della Ricerca Scientifica e Tecnologica

We live in times of rapid change, where electronic technology, applied both in the handling of information, and in its rapid transmission, is pervading the way we work, and more and more the way we live.

Throughout the world, we are rapidly evolving towards a global model of information and communication society. In this new model, economic, political and cultural life will rely on the availability of information at any time, from any place. The information so available is stored as text, as video, in sound recordings, in computer programs and in structured databases, and in ever more languages. The information and communication technologies that store and access this will make a substantial, and perhaps the greatest, contribution to future economic growth. Product and process innovations here are likely to give rise to radical changes to aspects of our social life, but in particular to business and the global economy. And meanwhile, the quantity of the information available from public and private sources and the means of physical access to it keeps on growing and growing, at an exponential rate.

Nowhere does this fact have more importance than in Language Engineering, in the technologies that can analyze a speech signal, that can work out the reference of text, and that can increasingly assist, and even automate translation. Natural languages, the languages we all think in and use to express ourselves every day, are the vehicles of choice for information: this is why there is a parallel growth in the need for tools to automate, or increase the ease and efficiency of using language, through which information is received, understood and applied. It is the task of Language Engineering to provide these tools, and there is a wide and growing range of language technologies to support it.

Classically, we have conceived the goal of enlightened policy as to enable the provision of universal access to these sources of information. But how should this access be understood? We believe it should be extended beyond a guarantee of physical access to the information channels; it should in fact include opportunities for all citizens use their own language for this access. This will make access to information easier, and processing it more effective. The issue here goes beyond economic and business competitiveness.

It has implications for the development of the social cohesion, and all the more when this has an international dimension.

In fact, the increasingly effective globalization of Information Technology, to the extent where an 'Information Society' is being created, has brought multilinguality to the forefront as a crucial issue: we urgently need strategies in order to rise to the challenge of multiple language barriers.

And these strategies will go beyond the technical realm to take in organizational and political aspects of the problem.

In fact, there are two complementary aspects of the challenge of multilinguality for language engineering.

One is to give citizens in their own language all the features, functions, tools and services, which are so far practically possible.

The other to assist citizens in operating across languages: translation and interpretation are just particular cases of the capabilities required.

One of the main objectives of the R&D activities in the 5th Framework Programme should be the provision of the basic language processing capabilities in all the official EU languages. This would be a useful first step towards truly universal support.

This imperative is underlined by both the dynamic trends we see in the EU: by the progressive enlargement of the Union to include more countries and languages, and the reinforcement of the links between the European countries.

But the trends we are talking of 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 cooperation. It will be a key factor for the success of this endeavor.

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.

Therefore, we consider this First International Conference on Language Resources and Evaluation a most timely event, and one of key importance: not only does it offer an open international forum to discuss the state-of-the-art and future directions, but also, by bringing together political and industrial decision makers, researchers, technology developers, service providers and users, it has a unique potential to spread awareness of the epochal challenge we are all facing. In so doing, it may also initiate a new effort, on an international scale, to rise to this major challenge facing our society.

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 coordinated by Prof. 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 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 multilingual, 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.

In practice, only languages for which adequate LR products and systems have been developed will be available over the network, certainly globally, but in practice on local networks too. In the worst case, citizens who are not able to communicate in the languages implemented in the global network could be denied full participation in their own institutions and media.

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

This will be more than a purely technical drawback. 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 measures necessary in order to support multilinguality.

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. No, we must build on past successes. 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.

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

All these considerations bring us to the question of whose responsibility it is to make LR available for a given language.

A recent survey promoted by the Commission has shown that the support of language technologies is at present extremely uneven across Europe at the national level. Several member States have no policy on the support of their national language within the Information Society, "a situation which threatens the survival of those languages in the mainstream". This problem is particularly acute for the provision of LR, which are bound to be specific to individual languages.

Even if national authorities would take responsibility for the provision of the monolingual LR for their own languages, in this way countering the market forces which privilege the more widely used and economically important languages, the problem of the responsibility for multilingual LR policy remains.

The Commission, naturally, has to consider the consequences of future extension of the Union to new countries.

But quite aside from this, the growing global scope of the Information Society is already posing the problem of interaction with language communities outside Europe. In short, the problem is mushrooming. Any solution needs to be adequate to a network economy and social relationships that stretch across the continents and oceans, regional blocs levels of education and development.

Sheer scale will require, on the one hand, an increasingly selective approach in deciding the best order of priorities for technological development. To this end technology evaluation could be very useful. On the other, this same logic imposed by the scale of the necessary work will call for open and well organized international cooperation in the field of LR.

We are only in the first phase of the process of responding to these challenges.

The presently embryonic infrastructure will need to be reinforced. It needs to be able to coordinate and perform complementary tasks without unnecessary duplications; to provide and update common repertories of linguistic data and knowledge which are available for the maximum possible number of languages ; to produce at reasonable cost and in due time customized resources to answer specific requests of developers; to offer the services that the Language Engineering community urgently needs.

The participation of international and national Funding Agencies in this Conference is a sign that they are aware of the key role and relevance of Language Resources.

The strategy they adopt will, no doubt, have decisive consequences for the place and contribution of language technology in the Information Society.

It is urgent and necessary that international organizations assign a clear priority to the development of Language Resources, and that different countries coordinate actions between them and with the international authorities.

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

Speech of Antonio Zampolli

President of ELRA

Ladies and Gentlemen,

It is a great honour and pleasure for me to welcome you to the "First International Conference on Language Resources and Evaluation" on behalf of the LREC Programme Committee.

The current landscape of this field, the field of Language Resources and Language Systems Evaluation, is very rich and complex. Evaluation and LRs are closely connected in many ways. Both play central roles of the infrastructure for natural language and speech processing: it was to underline this role that I proposed, in 1991, the term "Language Resources", today widely used.

Even a cursory analysis of the present situation shows how rapidly the whole field is evolving, both at the technical and organizational level. Unfortunately, it has often happened that research and organizational activities have developed without a proper level of synergy between them. The state of technical advancement can vary widely in different sectors and even in different countries. This of course leads to the risk that efforts may not receive the reinforcement they deserve, and or that their results may be delayed in coming to maturity. This is the basic reason why we need more efficient organization and easier exchange of technical expertise and information.

One recent development in this complex field has been the establishment of ELRA, the European Language Resources Association. It offers a good point from which to survey the variety and complexity of the various initiatives, present and future. It also gets a plain view of the needs of the R & D communities still remaining unsatisfied.

There is a profusion of teams working in different sectors, on different aspects of LRs, focusing on issues of particular relevance to their respective professional interests. Since they belong to different communities, they have their own specific organizations and conferences. And so they seldom have the possibility of a common forum and meeting-place, where they can exchange information and explore possible synergies and cooperation.

LREC aims to provide such a venue, promoting the awareness that all those working for LRs will benefit from considering themselves as members of a well-identified field. As stated in the Conference Announcement, the aim of this Conference is "to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding ongoing and planned activities, language resources and their applications, discuss evaluation methodologies and demonstrate evaluation tools, explore possibilities and promote initiatives for international cooperation in the areas mentioned above".

The variety of Associations and Consortia who have joined ELRA in promoting the Conference is in itself a demonstration of the variety of the activities related to LRs and of the perceived need for a common venue.

We are very grateful for the participation of national and international Funding Agencies at LREC: the strategy they will adopt will play a key role for the future of LRs and evaluation, and, as a consequence, of the human centered global Information Society. In fact, at this moment, language resources are the crucial key to unlock the potential of the global Information Society.

The most important single fact in the world economy at the end of the second millennium A.D. is the massive growth, and growing interpenetration, of information and communication technologies.

These trends place a new focus on automatic processing of content expressed in human languages, spoken and written. And the experience of the last decade has shown that effective processing of

language content at any level is impossible without extensive use of authentic language resources, for look-up, experiment, and training of systems.

The globalization of the society makes multilinguality an inescapable social and economic need.

Only languages for which adequate LR products and systems have been developed will be available over the IS network. On the worst hypothesis, citizens who are not able to communicate in the languages implemented in the global network would be denied full participation in the IS. Authoritative sources have already warned that languages for which LT will not be adequately developed run the risk of losing their status as media of communication in the IS. Because languages and cultures are inextricably linked, that will seriously threaten one of our most valuable human assets, linguistic and cultural diversity. To avoid this danger it is necessary to support multilinguality. Multilinguality has two obvious aspects: a citizen should be able to access the services of the IS in his or her own language; but should also be able to communicate and use information and services across language barriers.

The availability of adequate LRs in a language is the key condition for the development in it of applications and services that are informed by LT. LRs have the function of providing the linguistic knowledge specific to a language, and the linguistic knowledge needed to ensure the multilingual links among languages.

International cooperation in HLT, and in particular in LR, is the key that can open the door to a true multilingual society. One of the major goals of this Conference is to promote this cooperation, not only within researchers, but also at the institutional level.

This has been the goal of my working life in the last decade, and I feel compulsory to mention here the cooperation of D. Walker, who has dedicated his whole life to promote international and interdisciplinary cooperation in our field.

It is important to note that NSF and EC have signed a cooperation agreement a few weeks ago, and HLT is in the agenda.

We hope that the participation to this Conference of outstanding representatives of these two organisations is a sign that the role of LRs will be a priority in the future cooperation.

The number of participants, more than 500 from 35 countries, seems to confirm that this Conference was timely and answering to a perceived need. This number largely exceeds the 150-200 we had in mind organising this Conference, and if this will not, as we hope, have consequences on the adequacy of the organisation, it will be due to the efficiency and energy of the local organizers and the help of the supporting Organizations.

I feel it is my duty, in particular, to express our gratitude to the Authorities whom, honoring this Opening Session with their presence, are witnessing the large interest in the scientific, cultural, social and economic relevance of our field.

I wish to all of the participants a useful and enjoyable Conference in this marvellous town of Granada.

New Resources

Keys: R: for research use - C: for commercial use

ELRA-S0052 FIXED0IT - Italian Fixed Network Speech Corpus DB1 Phonetically rich sentences & application oriented utterances

The Italian Fixed Network Speech Corpus version 1.0 was recorded within the scope of the SpeechDat(M) project (LRE-63314), funded by the European Commission. Recording was done by using a primary rate ISDN interface, yielding 8 kHz, 8 bits per sample, A-law coded signal. The data files are formatted according to the SAM European project. The speech data are compressed with the GNU gzip program. All software needed to use the corpus is provided on the CDs.

The corpus contains the speech of about 1000 speakers (about 500 male and 500 female) and was designed to support the creation of voice-driven teleservices. The callers spoke at least 39 items, comprising:

Isolated and connected digits, natural numbers, money amounts, spelled words, time and date phrases, yes/no questions, city names, common application words, application words in phrases, phonetically rich sentences.

Most items are read, some are spontaneously spoken.

The recordings come with extensive and standardised documentation. All speech is carefully transcribed at the orthographic level; in addition, a number of clearly audible non-speech events are included in the transcription. Moreover, age and regional background of the speakers are provided. A pronunciation dictionary is added, containing all words that occur in the corpus, with a corresponding SAMPA broad-class phonemic transcription.

Validation and premastering of the CD-ROMs were performed by the Speech Processing Expertise Centre (SPEX), Leidschendam, The Netherlands.

Price for ELRA members: R: 11000 ECU C: 14000 ECU Price for non members: R: 20000 ECU C: 20000 ECU

ELRA-S0053 FIXED0IT - Italian Fixed Network Speech Corpus DB2 Phonetically rich sentences sub-set

See ELRA-S0052 for description. DB2 is a sub-set of DB1; it contains only the phonetically rich sentences items.

Price for ELRA members: R: 8800 ECU C: 14000 ECU Price for non members: R: 14000 ECU C: 20000 ECU

ELRA-S0054 Siemens Chile Spanish FDB-250

This speech database gathers Spanish data as spoken in Chile. All participants are native speakers. The corpus consists of read speech, including digits and application words for teleservices, recorded through an ISDN card. The whole database consists of 6.45 hours of speech, with 24 utterances per speaker. There is a total of 250 speakers (68 male, 80 female, 102 untagged). Except for the 102 untagged speakers, the age class is divided as follows: 15 speakers are less than 16 year old, 72 speakers are between age 16 to 30, 44 speakers are between age 31 to 45, and 14 speakers are between age 46 to 60 (and 102 untagged).

The callers spoke 74 different items in total: isolated digits, yes/no, common application words.

The data is provided with orthographic transliteration for all 6,000 utterances including 4 categories of non-speech acoustic events. A phonetic lexicon with canonical transcription in SAMPA is also included.

The speech files are stored as sequences of 8 bits 8 kHz A-law samples. Data are stored in a SAM file format.

Date of availability: end of September 1998

Price for ELRA members: 5000 ECU Price for non members: 7500 ECU

ELRA-S0057 Siemens Shanghai Mandarin FDB-1000

This acoustic database gathers Mandarin data, as spoken in Shanghai as a first or second Chinese dialect/language. The corpus consists of read speech, including digits and application words for teleservices, recorded through an ISDN card. A total of 70 utterances was prompted by each speaker. About 1000 speakers were recorded (500 male, 500 female).

The callers spoke the following items: isolated digits, yes/no, city names, common application words and phrases.
The data is provided with Chinese characters and English translation, canonical Pinyin transcription including tone markers, and several categories of non-speech events.

The speech files are stored as sequences of 8 bits 8 kHz A-law samples. Signal and annotation files are stored separately.

Date of availability: end of September 1998

ELRA-S0055 Siemens Russian FDB-1000

This speech database gathers Russian data. The corpus consists of read and spontaneous speech, recorded through an ISDN card, and was validated and accepted according to the SpeechDat(II) database exchange format. The whole database consists of 72 hours of speech, with approx. 49 prompted utterances per speaker. A total of 1000 speakers was recorded (500 male, 500 female). These are native speakers from 5 regions, mainly from Moscow and St. Petersburg (803 speakers). The speakers age class is divided as follows: 16 speakers are less than 16 year old, 340 speakers are between age 16 to 30, 345 speakers are between age 31 to 45, 255 speakers are between age 46 to 60, and 44 speakers are above age 60.

The callers spoke the following items:

Isolated and connected digits, natural numbers, money amounts, spelled words, time and date phrases, yes/no, city names, common application words, application words in phrases, phonetically rich sentences.

The data is provided with orthographic transliteration for all 48,812 utterances including 4 categories of non-speech acoustic events. A phonetic lexicon with canonical pronunciation is also provided.

The speech files are stored as sequences of 8 bits 8 kHz A-law samples. The data is stored in a SAM file format (4 CD-ROMs).

Date of availability: end of August 1998

Price for ELRA members: 14000 ECU Price for non members: 20000 ECU

ELRA-S0056 Slovenian SpeechDat(II) FDB-1000

The Slovenian SpeechDat(II) FDB-1000 consists of read and spontaneous speech, recorded through an ISDN card, and was validated and accepted according to the SpeechDat(II) database exchange format. The corpus includes about 1000 speakers (about 500 male and 500 female) who called over the Slovenian fixed network. All are native speakers of Slovenian from all dialect regions of Slovenia.

The callers spoke the following items: isolated and connected digits, natural numbers, money amounts, spelled words, time and date phrases, yes/no, city names, common application words, application words in phrases, phonetically rich sentences.

The speech files are stored as sequences of 8 bits 8 kHz A-law samples. The data is stored in a SAM file format (CD-ROMs). A phonetic lexicon with canonical transcriptions in SAMPA is also provided.

Date of availability: end of July 1998

Price for ELRA members: 14000 ECU Price for non members: 20000 ECU

ELRA-S0058 RVG1 (Regional Variants of German 1, Part 1)

The corpus consists of single digits, connected digits, phone numbers, phonetically balanced sentences, computer command phrases and spontaneous speech. Each speaker has read a subcorpus of 85 items:

· 11 single digits (0-9, with the two pronunciations of 2 (‘zwei’, ‘zwo’)),

· 19 connected digits (10-19, 20-100 in steps of ten),

· 12 computer command phrases,

· 30 phonetically balanced sentences,

· 5 6-digit phone numbers,

· 5 7-digit phone numbers,

· 2 phone numbers with area code,

· 1 minute spontaneous speech (monologue).

The speaker was placed in front of a standard IBM-compatible PC. The backround noise was limited to the usual noise in office environment, eg. door slam, backround crosstalk, phone ringing, paper rustle, PC noise, etc. The head of the speaker is in a range between 2-4 feet to the screen, 1-2 feet from the desktop microphones. The speaker is not forced into a special position. The speaker is wearing a Sennheiser HD 410 and is free to use the keyboard or the mouse in front of him. The three desktop microphones are: Sennheiser MD 441 U, Telex (Soundblaster) and Talk Back (AT&T). Speakers were selected to achieve the demoscopic density of the German spoken areas in Europe (including Austria and Switzerland).

The recorded sound samples are stored in NIST SPHERE format. The resolution is 16 Bits. The sampling frequency is 22.050 Hz except for speakers 001 to 036 which were recorded with 11.025 Hz. Each microphone channel is stored into a separate file. A transliteration of spontaneous speech according to Verbmobil Format is also provided.

RVG1, Part 1 contains 197 speakers recorded through 2 microphones.

(RVG1, Part 2, with 303 speakers recorded through 2 microphones will be available from the beginning of 1999).