The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

July-September 2001

Vol.6 n.3

Contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri &
Valérie Mapelli

Layout:
Magali Duclaux-Jeanmaire

Contributors:
Suzanne Lervard
Joseph Mariani
Patrick Paroubek
Antonio Ribeiro

ISSN: 1026-8200

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

Dear Members,

As 2001 progresses, the Language Resources and Evaluation Conference 2002 (LREC 2002) is getting more and more concrete. ELRA is much involved in the preparation of this third edition, which will take place in Las Palmas, Canary Islands - Spain, from 27th May 2002 to 2nd June 2002. The second Call for Papers has been distributed very recently and has been published on the web site dedicated to the LREC conferences (www.lrec-conf.org). You can also find in this issue a presentation of the LREC 2002 Announcement & Call for Papers.

Now considering the projects ELRA/ELDA is involved in, the recordings for the *Speecon project*, funded by the European Commission's Information Society Technologies (IST) programme, actually started at the end of July, and we are currently proceeding to the recruitment of some temporary staff to help and work on the project. In the framework of Speecon, ELRA/ELDA, who will provide the linguistic data to be used for the development of voice-controlled products, is not only responsible for the French recordings but is also in charge of supervising the recordings for the Italian and Swedish languages. For the three languages (French, Italian & Swedish), 600 speakers, distributed in 6 dialectal regions (in each country), will be recruited. The recordings will then take place in five different environments (entertainment, office, public place, car, children room).

Another project started recently for ELRA/ELDA: *HOPE 2001/EUROMAP Language Technologies* was officially launched at the beginning of July. A meeting was organised in Copenhagen, which brought together the "old" partners and the "new" ones. ELRA/ELDA, as a new partner, will have to develop and set up its National Action Plan to promote Language Engineering among the various and potential players, at a national level, who could take part into the development of language technologies and integrate such new technologies into their own systems and organisations.

Besides, on 31st May, ELDA, LDC and SPEX met together in the framework of the *Network-DC* project (Network of Regional and International Data Centres) to discuss the various aspects of the project, and particularly to draft a convention of partnership and review the final design documentation of the Broadcast News Speech Corpus (BNSC). The Network-DC project aims at setting up a network of data centres, thus facilitating the access to electronic language resources, which are currently managed by many different regional data centres.

Last but not least in this short review and update of the projects, the workshop of *CLEF* (Cross-Language Evaluation Forum) has taken place on 3rd & 4th September in Germany to report and analyse the results obtained following the first CLEF 2001 evaluation campaign. More information is available at the following address: www.iei.pi.cnr.it/DELOS/CLEF/.

Moreover, we are very glad to announce that the new ELRA & ELDA web sites should be made publicly available very soon. They have been both completely redesigned, and include new features such as a search tool on our catalogue of Language Resources (LRs). These two new sites are currently being internally tested.

As for the content of this issue, the first section comprises four articles: the first one written by Antonio Ribeiro, entitled "Making Use of Translated Texts to Identify Translations", as implied by its title, navigate through the field of textual language resources, especially the written corpora. Two other papers, devoted to speech processing and evaluation, have been written by Patrick Paroubek: "An Expert Bird's Eye View on Evaluation in Speech and Language Engineering" and "Workshop on Evaluation for Language and Dialog Systems at EACL'01", and the last one, by Joseph Mariani, is a report of the last meeting for the TIDES (Translingual Information Detection Extraction Summarization) US programme in July.

A second section is dedicated to LREC 2002, including the announcement and call for papers, and to the summit organised by the AEFT (European association for terminology). Of course, a third part is dedicated as usual to the resources added to our catalogue.

We hope that you have enjoyed these summer months, and wish you a good reading.

Please do not hesitate to contact us if there are any comments or suggestions that you would like to make, and if you wish to contribute to the next newsletter.

Antonio Zampolli, President　　　　Khalid Choukri, CEO

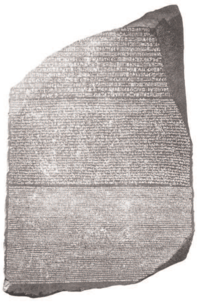
Making Use of Translated Texts to Identify Translations

Antonio Ribeiro

Texts which are mutual translations, usually shortened to *parallel texts*, have proven to be excellent sources of information in order to identify translations of words, terms or expressions. Human translators themselves acknowledge that these texts help them with the translation of unusual or unknown words or expressions, especially, in technical domains. They provide examples on the use of those words in the appropriate contexts. We have used ELRA's Multilingual Corpora of parallel texts to make correspondences between parts of the parallel texts - *Text Alignment* - and extract translations of words, terms and expressions - *Extraction of Translation Equivalents*.

Introduction

It is impossible to talk about Parallel Texts without making a reference to the Rosetta Stone. The Rosetta Stone is one of the most famous parallel texts. This parallel text helped Jean-François Champollion to decipher the Egyptian hieroglyphs in 1822 from a parallel text written in three different scripts: Hieroglyphic at the top, Demotic in the middle and Greek at the bottom. This is a clear example of the value and usefulness of these texts. In 1997, ELRA made a set of parallel texts from the Official Journal of the European Communities publicly available: "Multilingual Corpora for Co-operation - MLCC", version 1.0, disk 2 of 2.



It consists of texts translated in nine official languages of the European Union: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. The data were provided by the European Commission and comprise two sub-corpora from the Official Journal of the European Communities:

- The C Series of the Journal: written Questions (1993) asked by members of the European Parliament to the European Commission and their corresponding answers. It amounts to approximately 1.1 million words per language;
- Annexes: written records of Debates in the European Parliament covering a period between 1992 and 1994. These are written transcripts of oral discussions, which make them richer texts. There are between 5 to 8 million words per language.

Although Finnish and Swedish are also official languages of the European Union, ELRA's Multilingual Corpora do not provide translated texts for these languages. The data in the CD refer to texts between 1992 and 1994 and it was not until 1995 that Finland and Sweden joined the European Union.

The Canadian Parliament also produces parallel texts, in English and French. However, as far as we know, there is no other publicly available set of parallel texts translated in so many languages as in this resource available from ELRA. It is one of the best resources of parallel texts available.

Description

ELRA's Multilingual Corpora have been used for the PhD work of António Ribeiro at the Department of Informatics of Universidade Nova de Lisboa, in Lisbon, Portugal, supervised by Gabriel Lopes from the same department, and João Mexia from the Department of Mathematics. The goal of the work has been the development of a statistically supported and language independent methodology which allows the compilation of a bilingual "dictionary" for any two languages starting from parallel texts in those languages and without using any other linguistic knowledge. These bilingual "dictionaries" can then be used for bilingual lexicography, machine (-aided) translation, cross-language information retrieval, multilingual question-answering systems, to name but a few applications. As far as we know, this was the first attempt ELRA's Multilingual Corpora was used for such a task.

Parallel texts on their own are not of much use unless it is possible to know what is the translation of a text passage, found in the texts in the other languages. In order to do so, they must be aligned first, i.e. the various text passages of two parallel texts must be put into correspondence. The figure below shows an example of alignment of a parallel text in Portuguese and English:

A Written Question of the Official Journal of the European Communities.

Em	On
9	9
de Julho de	July
1992	1992
, o Tribunal de Justiça das Comunidades Europeias proferiu um acordao relativo ao processo	the Court of Justice of the European Comunities delivered a judgment in Case
C-2/90	C-2/90
(	(
Comissao contra o Reino de Belgica	Commission v. Belgium
)	)
sobre a proibicao de armazenar, depositar ou descarregar, mandar armazenar, depositar ou descarregar na regiao da Valonia residuos provenientes de um outro Estadomembro ou de uma regiao distinta da regiao em causa	concerning the ban on the storage, tipping or dumping in Wallonia of waste from other member States or other regions of Belgium
.	.

The alignment can be done taking advantage of sequences of characters which are the same in both texts, like 1992, which look similar, like *Bélgica* and *Belgium*, or which have already been identified as translations, like *acordão* and *judgement*.

In order to reduce the number of possible pairs of parallel texts from 72 sets (9 languages×8) to a more manageable size of 10 sets, Portuguese was taken as the source language of all language pairs. This PhD work has developed a methodology not only to align texts but also to build automatically a bilingual dictionary of words or expressions. The key issue for the automatic extraction of translations is to find a correlation between the occurrences of words or expressions in aligned parallel texts. In general, if two expressions appear more often together in aligned text passages than isolated, then they should quite probably be translations. We made a statistical analysis of the aligned texts in order to identify expressions and extract their corresponding translations.

This methodology is currently being used in the European project Tradaut-PT. This project aims at developing the Portuguese-English and Portuguese-French language pairs for the Machine Translation System Systran.

The MLCC resource from ELRA has been quite useful for this type of research work. Unfortunately, there are not many parallel texts available for research use, especially multilingual corpora. Up until recently the Canadian Hansards (the proceedings of the Canadian Parliament) have been one of the main sources of parallel texts for this type of research. The United Nations also publishes texts in several languages but they are not publicly available. The availability of this resource from ELRA has made possible a cross-language research which benefits to the international co-operation.

References

- António Ribeiro, Gabriel Lopes and João Mexia (2000) "Linear Regression Based Alignment of Parallel Texts Using Homograph Words". In Werner Horn (ed.) (2000) *ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence*, volume 54, © IOS Press 2000, Amsterdam, The Netherlands, pp. 446-450. Berlin, Germany, 2000 August 20-25.
- António Ribeiro, Gabriel Lopes and João Mexia (2000) "Using Confidence Bands for Parallel Texts Alignment". In *Proceedings of the 38th Conference of the Association for Computational Linguistics (ACL 2000)*, © Association for Computational Linguistics 2000, pp. 432-439. Hong Kong, China, 2000 October 3-6.
- António Ribeiro, Gabriel Lopes and João Mexia (2000) "A Self-Learning Method of Parallel Texts Alignment". In John White (ed.) (2000), *Envisioning Machine Translation in the Information Future - Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000 - Lecture Notes in Artificial Intelligence*, volume 1934, © Springer-Verlag 2000, Berlin, Germany, pp. 30-39. Cuernavaca, Mexico, 2000 October 10-14.

Antonio Ribeiro
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática
Quinta da Torre
Monte da Caparica
P-2829-516 Caparica - Portugal
Phone: +351-21-294 8300, ext. 10743
Fax: +351-21-294 8541
Email: ambar@di.fct.unl.pt
Web page: centria.di.fct.unl.pt/~ambar

An Expert Bird's Eye View on Evaluation in Speech and Language Engineering

Patrick Paroubek

Despite a quite busy calendar at this period of the year, on 2nd and 3rd July, took place in Paris the "International Course on Speech and Language Engineering Evaluation", jointly organized by CLASS (http://www.class-tech.org/) and ELSNET. During the course, a group of international experts of the field, who, for some of them, have been involved at the highest level for more than a decade in the largest evaluation programs for speech and language technology worldwide, presented to a limited number of participants from various backgrounds (both research and industry), the current state of deployment of evaluation in the field, along with the most pressing issues. How does evaluation relates with pre- and post research and development activities? What are the interests and benefits of evaluation for language engineering? What are the existing methodologies and how are they deployed? What is the relationship with basic research, development and market prospection activities? How is evaluation deployed in the different domains (speech versus text)? What form should it take? Should it be more technology-oriented or user-oriented? Are the existing methodologies and metrics satisfactory and sufficient? How can we take into account or abstract from the subjective factor introduced by human operators in the process? What about the needs for language resources? What are the required infrastructures? Those were the questions addressed by the presenters in front of the audience, which was initially expected to bring together a majority of high-level executives, decision makers, project officers and project managers in addition to engineers and scientists; but which finally gathered people with preoccupations focused more on practical issues (e.g. how do I evaluate the speech recognition interface of an on-board navigation system for cars) than decision-making oriented ones. The course started with an introductory presentation by Joseph Mariani of the French Ministry of Research and Limsi-CNRS, who drew a picture of the past and present programs that use the Evaluation Paradigm in Speech and Language Technology across the world, with an emphasis on what has been done in the USA, and the current situation in Europe, in particular with the articulation between the European programs and national initiatives. He was followed by Herman Steeneken of TNO Human Factors, who addressed the problem of methodology and international standardization for assessment activities of speech technology. He talked about the complexity stemming from the interplay between speech and language processing in human-computer interface development, and described the range of possible conditions of deployment for speech technology evaluation, from laboratory condition to field tests, remarking that progress have been impaired by a serious lack of agreed protocols for specifying language-based computer interfaces and for assessing their overall effectiveness. Dave Pallett, from NIST, then reviewed the development and implementation of benchmark tests for automatic speech recognition technology as they have been conducted for 15 years in the United States and the tremendous impact such tests had on the field. Herman Steeneken's and Dave Pallett's presentations were punctuated by recounts of interesting anecdotes that brought a note of humor and were invaluable for conveying to the audience the reality behind the organization of large scale evaluation campaigns. The first half-day of the course was concluded with a presentation of Valerie Mapelli, from ELRA, on the issue of linguistic resources, along with the current activities of LDC and ELRA. She took advantage of the event to make the first public disclosure of the third venue of the LREC conference (Language Resources and Evaluation Conference) which will take place in Las Palmas, in the Canary Islands (Spain) in May 2002. The morning of the second day started with more text-oriented evaluation issues, with a presentation by Philip Resnik, from UMIACS, who talked enthusiastically on the evaluations of meaning for Word Sense Disambiguation and Machine Translation. "What are the issues involved in creating a technology of meaning?" and "Does the market need it?" were the two main questions he used as a lead for his interesting presentation which benefited from his insider's view of the ongoing Senseval evaluation campaign for which the closing workshop was to be held at EACL'01 in Toulouse at the end of the week. I then had the opportunity to show how the evaluation paradigm can provide a meeting ground for all the actors of the domain (both from research and industry), through the example of the GRACE evaluation campaign for Part Of Speech tagging of French. and how the paradigm of evaluation can function as a language resource producer for high-quality and low-cost validated language resources. The issue of "Meaning" was back with the next talk, as Beth Sundheim, from SPAWAR Systems Center, told us how "Message Understanding" came to mean "Information Extraction" throughout the course of the series of seven United States government-sponsored evaluations of text analysis technologies that was carried out between 1987 and 1998 and known as the Message Understanding Conference (MUC) evaluations. Afterward, the course switched back to a more speech-oriented track with the last three presentations of the afternoon which were also the conclusion of the course. Kathleen Stibler, from Lookheed Martin Co, presented us the "Three-tiered Evaluation Approach for Interactive Spoken Language Dialog Systems" which measures user satisfaction, system support of mission success and component performance, and how it was applied in numerous fielded user studies conducted with the U.S. military. Then John Garofolo, from NIST, brought us a taste of the future, with his presentation of how NIST plan to integrate Human Language Technologies via Common Evaluations within the TREC Spoken Document Retrieval Track and the Automatic Meeting Transcription Project, which will help in creating technologies for the automatic production of meeting minutes using a combination of video and audio sensors and language technologies. Finally Niels Ole Bernsen, from NIS Labs at University of Odense, presented his views and experience on user-oriented evaluation of Spoken Language Dialog Systems, using in particular the results of the DISC and DISC-2 European projects. During the course the occasions were numerous for the audience to interact with the presenters, in particular the issue of how to evaluate Spoken Language Dialog Systems raised a lot of interest from the audience since some of the attendees had came to the course to find concrete answers to questions they were facing in the pursuit of their professional activities. The general feeling was very positive and the audience expressed its gratitude to the organizers for being provided with a complete picture and with a glimpse of what has been and what will be done in the field of evaluation in speech and language engineering, by the people who are directly involved in the action.

Note: The slides presented during the bullet course will be available at: www.limsi.fr/TLP/CLASS/class_events.html

Patrick Paroubek
Spoken Language Processing Group / Human-Machine Communication
Limsi - CNRS
Batiment 508 Universite Paris XI
BP 133 - 91403 ORSAY Cedex - France
Fax: (33) (0)1 69 85 80 88
Phone: (33) (0)1 69 85 81 91
Email: pap@limsi.fr

Workshop on Evaluation for Language and Dialog Systems at EACl'01

Patrick Paroubek

Toulouse is reputedly sunny and very hot in the summer, but this time it was a cool weather and rain which greeted at EACL'01 the participants to the two parallel workshops that had evaluation on their agenda: the Senseval workshop closing the current evaluation campaign on Word Sense Disambiguation and the workshop on Evaluation for Language and Dialog Systems organized by David Novick (U. of Texas), Joseph Mariani (French Ministry of Research and Limsi-CNRS), Candy Kamm (AT&T), Nils Dahlbäck (Linköping University), Frankie James (NASA), Karen Ward (U. of Texas) and myself. This two day workshop (July 5th and 6th) with 38 registered participants, was split into three informal sessions: dialog systems evaluation, evaluation for language engineering in general and another session more focused on probabilistic issues and classification. After the sessions, a final debate on the current needs of the field in relation with evaluation took place, gathering all the participants. The first session started with a flashy presentation made by Anton Nijholt (U. Twente); he offered us exciting views on multi-modal and multi-party contexts, illustrated with glimpses of avatars acting in their virtual environment. It raised the question brought forth by Dave Pallett

(NIST), and that is bound to come to the forefront of the scene in the coming years: how to perform evaluation in a multi-modal communication? Since Tim Paek (Microsoft) could not be with us, I took the responsibility of presenting his paper, entitled "Empirical Methods for Evaluating Dialog Systems", where he advocates the use of reference data built with a carefully crafted Wizard of Oz methodology, like "Gold Standard" (hypothesized maximal performance target for a given task), in conjunction with basic statistical metrics. The session ended with Laila Dybkjaer presentation focusing on the usability issue in evaluation of dialog systems, which was followed by a short panel discussion with Frankie James, Anton Nijholt, Niels Ole Bernsen (NIS Labs), John Garofolo (NIST) and myself as panelists. Among the topics addressed during the discussion were the dichotomy existing between user-oriented evaluation practices and black-box quantitative metrics, the fact that even without considering unrestricted dialogs, a dialog whose scope reaches beyond simple booking tasks requires a good emulation of human understanding, and, mentioned by David Novick, the surprising fact that no paper during this workshop raised the question of standard architecture for dialog systems.

The second session grouped more diverse papers about evaluation, starting with a rather technical presentation of Stephen Watkinson (U. of York) about the automatic translation of the Penn Treebank annotations into Categorial Grammar formalism, which raised the question of annotation standards for evaluation data and of resource re-usability. It was followed by the presentation of Martine Hurault-Plantet (Limsi-CNRS) with a two-level evaluation scheme applied to a system that participates in the Question & Answer track of TREC (Text REtrieval Conference). Then Widad Mustafa El Hadi (U. Lille 3) gave to the audience an insight on the problems the organizers of a terminology extraction evaluation campaign have to cope with, with her recounting of the ARC A3 evaluation campaign of the AUF (International association of French-speaking Universities). The afternoon ended in high point with the duet presentation of Valerie Barr (Hofstra University) and Judith Klavans (Colombia University) who captivated the audience when they explained that linguists and software engineers do not mean the same thing when they talk about evaluation.

In her invited talk for opening the last session, Donna Harman (NIST) advocated the benefits of focused evaluation, with the examples of TREC and DUC (Document Understanding Conference, see http://www-nlpir.nist.gov/projects/duc/main.html). Then Yuval Krymolowski (Bar-Ilan U.) gave an interesting talk, where he showed that one can use the distribution of performance to study statistical NLP systems and corpora. Michèle Jardino (Limsi-CNRS) concluded the session with her presentation on the comparison of two clustering methods. A short panel discussion ensued with Donna Harman, Yuval Krymolowski, M. Jardino, Widad Mustafa El Hadi, and Martin Rajman (EPFL) with the problem of classifiers evaluation in language engineering as an opening question.

The workshop ended in the afternoon of the second day with a general work session about the deployment of evaluation in Language Engineering. Joseph Mariani opened the discussion with an overview of the current situation across the world (mostly United-States, Japan and Europe). The general consensus was that standards and data were crucial assets for the development of evaluation in language engineering, in particular the audience recognized the important role played by resource repositories like LDC and ELRA, as well as the need to have available evaluation packages. It was also said that an international framework should be set up in order to cooperate on promoting Language Technologies evaluation as a good practice, on ensuring standard metrics, methods and protocols, on conducting studies on areas where the evaluation methods are still open (dialog, spoken language translation...). The set up of a permanent entity in Europe, comparable to NIST, for the organization of evaluation activities was recognized as an essential need of the field.

Patrick Paroubek
Spoken Language Processing Group / Human-Machine Communication
Limsi - CNRS
Batiment 508 Universite Paris XI
BP 133 - 91403 ORSAY Cedex - France
Fax: (33) (0)1 69 85 80 88
Phone: (33) (0)1 69 85 81 91
Email: pap@limsi.fr

Translingual Information Detection Extraction Summarization programme

Joseph Mariani

The DARPA TIDES (Translingual Information Detection, Extraction and Summarization) Principal Investigators (PI) meeting took place at the Inn at Penn hotel in Philadelphia on July 23-25, with a crowd of about 80 participants, most of them from DARPA sponsored projects, or from US governmental agencies. There were few foreign participants : only Karen Sparck Jones (Cambridge University), Sadaoki Furui (Tokyo Institute of Technology) and myself, all from the TIDES Advisory Committee (TACO). The organization of the meeting was under the responsibility of Martha Palmer (UPenn).

The general goal of the TIDES program is to develop tools which would allow an english-speaking (or more exactly -hearing) analyst to understand information which is encoded on various media (text, OCR, speech, image…) and in various languages that he may not understand.

The meeting was introduced by Charles Wayne, who took over the responsibility as the TIDES program manager, following Gary Strong's return to NSF. He considered the two first years as an "exploratory research" phase, and mentioned that he wished that the workshop would help in strengthening the links within the TIDES community (changing therefore the name to TEAM TIDES) and in installing the whole program on solid tracks. In order to do so, he nominated several individuals as responsible for each part of the program: James Allan for Detection (D), Ralph Weidschedel for Extraction (E), Donna Harman for Summarization (S), Kevin Knight for Translation (T) plus Mark Liberman for language resources and Allen Sears for Integration (including experimentation within Integrating Feasibility Experiments (IFE)). He also announced that the program already obtained a one-year extension until mid-2005.

The program was very dense, as usual for DARPA workshop. On the first day, each project was given a five (5) minute slot to present in one slide (with four parts) its activity. We had 28 of such presentations. It got much more comfortable in the afternoon with 8 minutes for each task coordinator (but I only got 5 minutes to present international activities on evaluation in Language Technologies…), followed by a session on the "Enabling Infrastructure", including resources and evaluation for each task. 8 demos were organized after dinner until a final end of this first day at exactly 9:27 pm. The second day was much much more relaxed with 10 minutes for each selected scientific presentation in each of the 4 areas, followed by a general "Other research" session. The last day was for breakout sessions in parallel on each of the four tasks, followed by two breakout sessions : one on resources and one on integration, followed by a final wrap-up by the program manager. Oooopps ! I forgot to mention that breakfasts served from 7:00 to 8:30 am were the place to conveniently start the day with some first scientific discussions…

There are still some basic questions which remain open: researchers were still arguing on what is the exact content of detection (which may be considered as related to TREC or TDT-like activities), extraction (which may considered as a continuation of the MUC (Message Understanding Conference) task, or DUC (Document Understanding Conference)) and summarization (a brand new tasks, with obviously many open issues on the way to evaluate and to produce the data necessary for that).

The big move in my opinion is the increase of the place of translation in the program, as it appeared in a movie which even says that the "T" of TIDES stands for "Translation", not "Translingual" … The content of this task is however very specific as English is considered as the only target language, and as a few languages are considered as source language: a first circle, including languages of large importance where resources are or will be available for development in large quantities (typically 10 Mwords of parallel texts and 100 Mwords of similar texts, for two languages, Chinese and Arabic), a second circle where data is available in lower quantity (Japanese, Spanish and Korean were mentioned here, with 1 Mwords of parallel texts and 10 Mwords of similar texts) and a third cycle with "low density languages", those which are not well studied, automatized and resourced (100 Kwords of parallel texts and 1 Mwords of comparable texts), with a "surprise" language which could be proposed to experiment "developing a machine translation system in a week" for example.

The translation field actually obtained the largest success of the workshop with a proposal made by IBM of a metric for measuring translation quality, a proposal which was considered by some of the attendees as having the same potential impact on that field than the word accuracy measure had for speech recognition evaluation and development. The score uses reference human translations, and is based on n-gram matching, including n-gram weights and length penalty. There was a spontaneous proposal to install a server at LDC, in order to distribute this Language Translation evaluation scoring software, in cooperation with MITRE.

It was amazing to see that although Speech is not very present in TIDES, which mostly focus on the translingual aspects of natural language processing, including only transcribed speech, many researchers who used to work on speech processing were present at the workshop and are shifting their research activity towards translingual language processing (people like Makhoul, Roukos, Schwartz, Jelinek or Waibel) and are now discussing with NL people. This will probably modify the "Natural Language Processing" landscape in the near future.

The initial TIDES mission was to process information regardless of language and medium. It seems that the number of languages has been much restricted and that the medium only extends from written language to transcribed speech due to the size of the budget and to the necessity to focus the activity in agreement with the available manpower. However, speech won't be forgotten in the long run, as it is obviously a major medium for human communication.

It was mentioned that there is a need to better articulate the efforts conducted on T, D E and S. Translation is transversal and its position should therefore be taken into account in the overall organization of the program, including in the deliverables and milestones. Also translation could be considered as a task per se, or as a component for other tasks (D, E or S). This also applies for evaluation, which should take into account the various relationships, in terms of linked modules, and in terms of the coherence of the development and test data made available for various tasks. Summarization may also appear as having a transversal position as a follow-up of the detection or extraction tasks. Articulating the four tasks is a very appealing challenge for the program manager, both in terms of component development and evaluation and in terms of integration within demonstrators.

It is however a pity that other groups than those sponsored by DARPA could not attend the event, while technology development also takes place in other parts of the world, and as other languages could benefit from the TIDES infrastructure and generic tools, and while the effort to address all languages is obviously out of reach of any single country or agency. The TIDES program would actually much welcome affiliates from overseas. The whole area of Language Technology would benefit from a distributed share of the necessary effort. Therefore international cooperation in the field of Language Technology Evaluation should be encouraged and I proposed to install a committee (SCILITE: Supporting Committee for the Internationalization of Language and (possibly) Image Technology Evaluation) for addressing this issue, at least as an information exchange forum in the first step.

Joseph Mariani
LIMSI-CNRS
BP 133
91403 Orsay Cédex, France
Tel: +33 1 69 85 80 85
Fax: +33 1 69 85 80 88
E-mail: mariani@limsi.fr

LREC 2002 Third International Conference on Language Resources and Evaluation

DATES

Main Conference:	**29-30-31 May 2002**
Pre~ & Post-Conference Workshops:	**27-28 May 2002 & 1-2 June 2002**

LOCATION **Las Palmas, Canary Islands - Spain**

With support of TELEFONICA Foundation (of Spain) and support sought from the
Commission of the EU and other institutions.

The Third International Conference on Language Resources and Evaluation is organised by ELRA in cooperation with other Associations and Consortia, including ACL, AFNLPA, ALLC, CLASS, COCOSDA, ORIENTAL COCOSDA, EAFT, EAGLES/ISLE, ELSNET, ENABLER, EURALEX, FRANCIL, ISCA, LDC, ONTOWEB, PAROLE, TEI, etc., and with major national and international organisations, including the Commission of the EU - Information Society DG, DARPA, NSF, and the Japanese Project for International Co-ordination of East-Asian Spoken Language Resources and Evaluation. Co-operation with other organisations is currently being sought.

CONFERENCE AIMS

In the framework of the Information Society, the pervasive character of Human Language Technologies (HLT) and their relevance to practically all the fields of Information Society Technologies (IST) has been widely recognised.

Two issues are considered particularly relevant: the availability of language resources and the methods for the evaluation of resources, technologies, products and applications. Substantial mutual benefits can be expected from addressing these issues through international cooperation.

Languge resources are i. e. written, spoken and multimodal corpora and lexica, grammars, terminology databases, multimedia databases, basic software tools for the acquisition, preparation, collection, management, customisation and use of these and other resources.

The evaluation, fully recognised in the field, involves assessment of the state-of-the-art for a given technology, measuring the progress achieved within a programme, comparing different approaches to a given problem and choosing the best solution, knowing its advantages and disadvantages, assessment of the availability of technologies for a given application, product benchmarking, and assessment of system usability and user satisfaction.

In the recent past, language engineering and research and development in language technologies have led to important advances in various aspects of written, spoken and multimodal language processing. Although the evaluation paradigm has been studied and used in large national and international programmes, including the US DARPA HLT programme, the EU HLT programme under FP5-IST, the Francophone AUF programme and others, particularly in the localisation industry (LISA and LRC), it is still subject to substantial unsolved basic research problems. The European 6th Framework program (FP6), planned for a start in 2003, includes multilingual and multisensorial communication as one of the major R&D issue, and the evaluation of technologies appears as a specific item in the Integrated Project instrument presentation.

The aim of this Conference is to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding language resources, their applications, ongoing and planned activities, industrial use and requirements, discuss evaluation methodologies and demonstrate evaluation tools, explore possibilities and promote initiatives for international cooperation in the areas mentioned above.

LREC 2002
Third International Conference on Language Resources and Evaluation

CONFERENCE TOPICS

ISSUES IN THE DESIGN, CONSTRUCTION AND USE OF LANGUAGE RESOURCES (LR)

- Guidelines, standards, specifications, models and best practices for LR,
- Methods, tools, procedures for the acquisition, creation, management, access, distribution, use of LR,
- Organisational issues in the construction, distribution and use of LR,
- Legal aspects and problems in the construction, access and use of LR,
- Availability and use of generic vs. task/domain specific LR,
- Methods for the extraction and acquisition of knowledge (e.g. terms, lexical information, language modelling) from LR,
- Monolingual and multilingual LR,
- Multimodal and multimedia LR,
- Integration of various modalities in LR (speech, vision, language),
- Documentation and archiving of languages, including minority and endangered languages,
- Ontological aspects of creation and use of LR,
- LR for psycholinguistic and sociolinguistic research in human-machine communication,
- Exploitation of LR in different types of applications (information extraction, information retrieval, vocal and multisensorial interfaces, translation, summarisation, www services, etc.),
- Industrial LR requirements and community's response,
- Industrial production of LR,
- Industrial use of LR,
- Analysis of user needs for LR,
- Internet-accessible metadata descriptions of LR,
- Mechanisms of LR distribution and marketing,
- Economics of LR.

ISSUES IN HUMAN LANGUAGE TECHNOLOGIES EVALUATION

- Evaluation, validation, quality assurance of LR
- Benchmarking of systems and products; resources for benchmarking and evaluation
- Evaluation in written language processing (text retrieval, terminology extraction, message understanding, text alignment, machine translation, morphosyntactic tagging, parsing, semantic tagging, word sense disambiguation, text understanding, summarization, localization, etc.)
- Evaluation in spoken language processing (speech recognition and understanding, voice dictation, oral dialog, speech synthesis, speech coding, speaker and language recognition, spoken translation, etc.)
- Evaluation of document processing (document recognition, on-line and off-line machine and hand-written character recognition, etc.)
- Evaluation of (multimedia) document retrieval and search systems (including detection, indexing, filtering, alert, question answering, etc)
- Evaluation of multimodal systems
- Qualitative and perceptive evaluation
- Evaluation of products and applications, benchmarking
- Blackbox, glassbox and diagnostic evaluation of systems
- Situated evaluation of applications
- Evaluation methodologies, protocols and measures
- From evaluation to standardisation of LR

LREC 2002
Third International Conference on Language Resources and Evaluation

GENERAL ISSUES

- National and international activities and projects
- LR and the needs/opportunities of the emerging multimedia cultural industry
- Priorities, perspectives, strategies in the field of LR national and international policies
- Needs, possibilities, forms, initiatives of/for international cooperation
- Open architectures for LR

PROGRAMME

The Scientific Program will include invited talks, papers accepted for oral presentations, papers accepted for poster presentations, referenced demonstrations and panels. A special workshop will be organised on National Projects in LR and evaluation.

FORMAT FOR ABSTRACT SUBMISSION

Submitted abstracts of papers for oral and posters presentations should consist of about 800 words.

Demonstrations of LR and related tools will be reviewed as well. Please send an outline of about 400 words. If a demo is connected to a paper, please attach the outline to the paper abstract.

A limited number of panels and workshops is foreseen. Proposals are welcome and will be reviewed. For panels please send a brief description, including an outline of the intended structure (topic, organiser, panel moderator , tentative list of panelists). For workshops, see below.

All submissions should include a separate title page, providing the following information: type of proposal (paper for oral presentation, paper for poster presentation, demo, paper plus demo, panel); the title to be printed in the programme of the Conference; names and affiliations of the authors or proposers; the full address of the first author (or a contact person), including phone, fax, email, URL; the required facilities for presentation (overhead projector, data display; other hardware, platforms, communications); and 5 keywords. All submissions will be reviewed by the Scientific Committee.

Electronic submission

Electronic submission of abstracts should be in ASCII file format. This file should be sent to:
lrec@ilc.pi.cnr.it
Attn: Antonio Zampolli
LREC chairman

Submission in hard copy

Please send five hard copies to:
Antonio Zampolli - LREC Chairman
Istituto di Linguistica Computazionale del CNR
Area della Ricerca di Pisa
Via G. Moruzzi 1
56124 Pisa - ITALY

Exhibits

An exhibit area will also be made available at LREC2002. This is open to companies and projects wishing to promote, present and demonstrate their language resources and evaluation products and prototypes to a wide range of experts and representatives from all over the world who will be participating at the conference. Please note that the exhibits of LR are different from system demonstrations. The exhibits will run in parallel with the Conference for 3 days and the exhibit hall will be located near the general conference rooms. For more information, please contact the ELDA office at: choukri@elda.fr

LREC 2002 Third International Conference on Language Resources and Evaluation

IMPORTANT DATES

Submission of proposals for oral and poster papers, referenced demos, panels and workshops:
20 November 2001

Notification of acceptance of workshop and panel proposals:
10 December 2001

Notification of acceptance of oral papers, posters, referenced demos:
2 February 2002

Final versions for the Proceedings:
2 April 2002

Conference:
29-30-31 May 2002

Pre Conference Workshops:
27-28 May 2002

Post Conference Workshops:
1-2 June 2002

Proposals for the pre- and post-conference workshops should be sent to Antonio Zampolli (see address above), be no longer than three pages and contain:

(1) a brief technical description of the specific technical issues that the workshop will address;
(2) the reasons why the workshop is of interest this time;
(3) the names, postal addresses, phone and fax numbers and email addresses of the Workshop Organising Committee, which should consist of at least three people knowledgeable in the field coming from different institutions;
(4) the name of the member of the Workshop Organising Committee designated as the contact person;
(5) a time schedule of the workshop and a preliminary agenda;
(6) a summary of the intended workshop Call for Participation;
(7) a list of audio-visual or technical requirements and any special room requirements.

The workshop proposers will be responsible for the organisational aspects (e.g. Workshop Call preparation and distribution, review of papers, notification of acceptance, etc.). Further details will be sent to the proposers.

CONSORTIA AND PROJECT MEETINGS

Consortia or projects wishing to take this opportunity for organising meetings, should refer to the website for details on assistance in arranging meetings' facilities (www.lrec-conf.org) or contact the ELDA office, choukri@elda.fr.

The complete and detailed announcement, including the composition of the committees, some more information regarding the organisation and the registration, and the conference addresses is available on the web site dedicated to the LREC conferences, at the following address:
www.lrec-conf.org

Announcement:
Pre-summit of the European Association for Terminology (EAFT)

The European Association for Terminology is planning to organise a meeting with world wide terminology networks to prepare the terminology summit which will take place in June 2002.

This summit has been decided in March 2000, when the EAFT and TDCNet met together in Paris. Some key players expressed the need to expand the exchanges and share as many experiences as possible between the different players. Eight national and regional associations are already members of the EAFT. They will be attending the planned terminology summit in June 2002: Assiterm (Italia), Termip (Portugal), Eleto (Greece), Danterm (Denmark), TermMRom (Moldavia), TermRomBucarest (Roumania), NL-Term (The Netherlands), SFT (France).

Other associations and networks such as ELRA, RIFAL, NORDTERM, TERMNET, GTW, IITF, INFOTERM or EAFTerm (Asia) are invited.

The pre-summit meeting will take place from 22nd November to 24th November 2001 in Brussels, in the ISTI premises.

Six workshops are organised, with the following themes:

Workshop 1: Terminology in society/ political grounds
Workshop 2: Terminology training
Workshop 3: Infrastructure and co-operation
Workshop 4: Information and documentation
Workshop 5: Commercial aspects of terminology
Workshop 6: Terminology production

The workshops will be conducted by the representatives of some terminology networks such as Nordterm (Northern European countries), Riterm, Realiter and Rifal (latin countries) and TermNet (on a world wide scale).

The main themes which should be developed in June 2002 are:

- Terminology and language teaching
- Terminology processing and production
- International cooperation
- Improvement of the multicultural and multilingual aspects by:
  - creating some terminology guidelines at a European level
  - streamlining the role of terminology in language teaching
  - elaborating a convention on terminology infrastructure
  - inviting actors from private sectors (not only terminologists but also political representatives)
  - improving co-operation

Over 150 participants will attend the conference, which will be diffused live on the Internet.

The languages of the conference are English, French and Spanish.

The details/information about the terminology summit (2002) and pre-conference (2001) are available on the web, at the following address: **www.eaft-aet.net**

ERRATUM

The previous issue of the ELRA newsletter, published in June 2001, included an article dealing with the European Terminology Information Server (ETIS).

Only 14 partners to this project (out of 16 partners) were listed: INFOTERM, ASS.I.TERM, CINDOC, CTB, CTN, DANTERM, DEUTERM, ELOT, IM, NTU, TNC, TSK, UL/DTIL, UZEI.

The other two partners, missing in the original article are: TERMCAT (Centre de Terminologia) & RTT (Rådet for teknisk terminologi).

New Resources

ELRA-S0111 Eleftherotypia Journal Speech Database

The Eleftherotypia Speech Database (13 CD-ROMs) consists of read material collected in order to be used for the development of continuous speech recognition systems for the Greek language. All recorded sentences were selected from extracts of the Elefterotypia-journal text corpus and provide a vocabulary of about 40,000 words. The total number of utterances is over 32,000 (aproximately 72 hours of speech material from 120 different speakers, male and female).

Detailed orthographic transcription files are also included in the distribution. There are markings for the utterance's orthography and several speech and non-speech events (e.g. mispronunciations, truncation, noise etc).

The recording procedure took place in three different environments : a sound proof room, a quiet environment and an office environment. Two different microphones were used : a desk microphone and a head-mounted close-talking microphone.

The format of the waveform files is NIST. Waveforms are encoded using PCM coding format, 16000 sampling rate, 2 bytes per sample.

	ELRA Members	Non Members
Price for research use	2,500 Euro	4,000 Euro
Price for commercial use	10,000 Euro	15,000 Euro

ELRA-S0112 Persian Speech Database - Farsdat

The Persian Speech Database Farsdat comprises the recordings of 300 Persian speakers, who differ from each other with regards to age, sex, education level, and dialect (10 dialect regions of Iran were represented: Tehrani, Torki, Esfahani, Jonubi, Shomali, Khorassani, Baluchi, Kordi, Lori, and Yazdi). Each speaker uttered 20 sentences in two sessions, and 100 of these speakers uttered 110 isolated words. 6000 utterances were segmented and labelled phonetically and phonemically manually, including 386 phonetically balanced sentences, using IPA characters. The acoustic signal has been stored with a Wave file standard, so that it can be used by any other application software. The used sampling frequency reaches 22.5 KHz, and the signal-to-noise ratio 34 dB. The ambiguities in segmentation have been solved by reference to the corresponding spectrograms extracted from DSP sona-Graph KAY 5500.

	ELRA Members	Non Members
Price for research use	800 Euro	1,200 Euro
Price for commercial use	2,500 Euro	5,000 Euro

ELRA-S0113 Spoken Dutch Corpus

The Spoken Dutch Corpus will upon completion contain approximately ten million words, two thirds of which originate from the Netherlands and one third from Flanders. The Spoken Dutch Corpus comprises a large number of samples of (recorded) spoken text. In all about 1,000 hours of speech. The entire corpus will be transcribed orthographically, while the transcripts will be linked to the speech files. The orthographic transcript is used as the starting-point for the lemmatization and part-of-speech tagging of the corpus, which is manually verified. For a selection of one million words it is envisaged that a (verified) broad phonetic transcription will be produced, while for this part of the corpus also the alignment of the transcripts and the speech files will be verified at the word level. In addition, a selection of one million words will be annotated syntactically. Finally, a more modest part of the corpus, approximately 250,000 words, will be enriched with a prosodic annotation. Parts of the corpus are made available in the course of the project through intermediate releases that appear at regular six to eight month intervals. The first release came out in March 2000. The complete corpus will be available by June 2003.

Release 1 (March 2000):

62 hours speech samples orthographically transcribed (615,000 words), 90,000 words enriched with Part-of-Speech tags; annotation CD with first version of PRAAT (annotation tool) and first version of documentation (in Dutch) among which relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment) (information on the speakers in anonymous form);

Release 2 (October 2000):

Over 150 hours of speech samples, orthographically transcribed (over 1,500,000 words), approximately 750,000 words enriched with Part-of-Speech tags; annotation CD with annotation protocols and relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment) is available (information on the speaker in anonymous form);

Release 3 (April 2001):

More orthographically data enriched with Part-of-Speech tags; the first broad phonetic transcriptions, word alignments, syntactic annotations, lexicon link-up is available; annotation CD with documentation among which relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment); this release encompasses the first version of Corex, the exploitation tool. The next intermediate release (Release 4) is planned for October 2001.

	ELRA Members	Non Members
Price for research use	1,000 Euro	1,750 Euro
Price for commercial use	11,000 Euro	15,000 Euro
Annotations-only version for non-commercial use	50 Euro	

ELRA-W0027 An-Nahar Newspaper Text Corpus

The An-Nahar Newspaper Text Corpus comprises articles in Arabic (Lebanon) from 1995 to 2000 (6 years) stored as HTML files on CDRom media. Each year contains 45 000 articles and 24 million words. Each article includes information such as title, newspaper's name, date, country, type, page, etc. The size in byte is 150 MB per year except for 1999 (533 MB).

	ELRA Members		Non Members	
Price for research use	1 yr	336 Euro	1 yr	504 Euro
	2 yrs	672 Euro	2 yrs	1008 Euro
	3 yrs	1008 Euro	3 yrs	1512 Euro
	4 yrs	1344 Euro	4 yrs	2016 Euro
	5 yrs	1680 Euro	5 yrs	2520 Euro
	6 yrs	2016 Euro	6 yrs	3024 Euro
Price for research use by a commercial organisation	1 yr	672 Euro	1 yr	1008 Euro
	2 yrs	1176 Euro	2 yrs	1764 Euro
	3 yrs	1680 Euro	3 yrs	2520 Euro
	4 yrs	2100 Euro	4 yrs	3150 Euro
	5 yrs	2520 Euro	5 yrs	3780 Euro
	6 yrs	3192 Euro	6 yrs	4788 Euro

ELRA-W0028 Wolverhampton Business English Corpus

The WBE was created by the Computational Linguistics Group at University of Wolverhampton through a funding from ELRA in the framework of the European Commision project LRsP&P (Language Resources Production & Packaging - LE4-8335). A survey of electronic language resources in the business domain carried out at Wolverhampton revealed that there are very few business corpora in existence, and almost none of them are widely accessible. There is significant demand for a business corpus, from both the NLP and pedagogic (language, business communication, and linguistics teachers and students) communities. The Wolverhampton Corpus of Written Business English is:

- A synchronic corpus, including only texts available on the web during a 6-month period in 1999-2000 AD.
- A monolingual English corpus: it comprises only texts written in English; but no restriction was applied as regards the variety of English used. On the contrary, the WBE deliberately tried to capture a wide range of varieties of English, by including documents from websites in Britain, USA, Pakistan, Netherlands, Belgium, Switzerland, Hong Kong, etc.
- A written corpus: it contains only written materials. However, a few of the documents are transcripts of speeches.
- A business corpus: the texts were selected manually, and care was taken to ensure that all the texts were from the business domain.

The corpus consists of 10,186,259 words from 23 different Web sites.

The data can contribute to a wide range of NLP tasks, including information retrieval, information extraction, summarisation, etc.

The WBE was built using materials solely from the Web. However, this does not mean that the corpus gives access only to a restricted range of categories of texts. On the contrary, the amount of information available online allowed us to select from a wide variety of categories. These range from product descriptions, company press releases, and annual financial reports, to business journalism, academic research papers, political speeches and government reports. The texts have been grouped according to the source site. The corpus is distributed in three formats.

The first one is the original encoding of the text. The majority of the texts are in HTML and plain text format. There are a few in PDF format or Microsoft Word DOC format.

The second format is plain text. The files were converted automatically if they were not in plain text format, and manually checked.

The corpus is also provided as SGML encoded files, using the Corpus Encoding Standard (http://www.cs.vassar.edu/CES/). The header of each file provides information about the title of the file, length in words, etc. The paragraph and sentence boundaries, and part of speech tags for each word are marked using SGML tags.

All the available files were converted to 8-bit ASCII format using ISO 8859-1. Characters with ASCII codes from 127—255 (also known as Extended ASCII) were manually checked in order to ensure the correct representation of the characters.

The corpus was checked for spelling errors, but special care was taken to ensure that any variant spellings specific to the business domain were not wrongly corrected.

A validation work was carried out by an external validator. It consisted of checking text files, tools, tagging and documentation.

	ELRA Members	Non Members
Price for research use	750 Euro	3,000 Euro
Price for commercial use	3,000 Euro	10,000 Euro

ELRA-S0034 Verbmobil: new resources added

Verbmobil is a long-term project of the German Federal Ministry of Education, Science, Research and Technology (BMBF, Projektträger DLR). Its aim is to give Germany an international top position in language technology and its economical application in the next millenium by cooperation and concentration of as many as possible specialists from industry and science. The long-sighted aim is the development of a mobile translation system for the translation of spontaneous speech in face-to-face situations.
The following resources are spontaneous speech databases recorded in a dialogue task (appointment scheduling).
See next page for details on the new Verbmobil resources added to the catalogue.

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

VM CD 16.1 - VM16.1 (1 CDROM, new edition)
Verbmobil II - Japanese, 200 dialogues, 200 appointment schedulings - 3311 turns.

VM CD 17.1 - VM17.1 (1 CDROM, new edition)
Verbmobil II - Japanese, 200 dialogues, 200 appointment schedulings - 2741 turns.

VM CD 18.1 - VM18.1 (1 CDROM, new edition)
Japanese, 200 dialogues, 200 appointment schedulings - 2345 turns.

VM CD 19.1 - VM19.1 (1 CDROM, new edition)
Japanese, 200 dialogues, 200 appointment schedulings - 2911 turns.

Verbmobil - VM CD 48.1 - VM48.1 (BAS edition)
Verbmobil II - German, 28 spontaneous dialogues (28 close mic, 28 room mic, 27 phone line (GSM) recordings), 4516 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 49.1 - VM49.1 (BAS edition)
Verbmobil II - German, 24 spontaneous dialogues (24 close mic, 12 room mic, 12 phone line (GSM) recordings), 2597 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 50.1 - VM50.1 (BAS edition)
Verbmobil II - American-English, 8 spontaneous dialogues (8 close mic, 0 room mic, 0 phone line (GSM) recordings), 679 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 44.1 - VM44.1 (BAS edition)
Verbmobil II - Japanese, 19 spontaneous dialogues (19 close mic, 0 room mic, 0 phone line (GSM) recordings), 920 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 45.1 - VM45.1 (BAS edition)
Verbmobil II - Japanese, 21 spontaneous dialogues (21 close mic, 0 room mic, 0 phone line (GSM) recordings), 1293 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 46.1 - VM46.1 (BAS edition)
Verbmobil II - Multilingual Japanese/German, 11 spontaneous dialogues (11 close mic, 0 room mic, 0 phone line (GSM) recordings), 607 turns, transliteration (Verbmobil II Format).

Verbmobil - VM CD 47.1 - VM47.1 (BAS edition)
Verbmobil II - Multilingual with human interpreter (3 channels) English/German, 18 spontaneous dialogues (18 close mic, 0 room mic, 0 phone line (GSM) recordings), 902 turns, transliteration (Verbmobil II Format).

Verbmobil - VM Bonus CD - VMBONUS (BAS edition)
Additional data and documentation that is not included in the regular VM volumes.

Verbmobil - VM Lexicon database - VMLEX (BAS edition)
Verbmobil lexicon database of the University of Bielefeld..

Verbmobil - VM CD 15.1 - VM15.1 (new edition)
Verbmobil II - Multilingual mit Simultanbersetzer (3 Kanle) English/German, 18 spontaneous dialogues (18 close mic, 0 room mic, 0 phone line (GSM) recordings), 902 turns, transliteration (Verbmobil II Format).

ELRA-S0034 Verbmobil: new resources added

VM CD 53.1 - VM53.1 (BAS edition)
German, 16 spontaneous dialogues (16 close mic, 8 room mic, 8 phone line (GSM) recordings) - 1771 turns, transliteration (VM II Format).

VM CD 60.1 - VM60.1 (BAS-Edition)
Japanese - 10 spontaneous dialogues (10 close mic, 0 room mic, 0 phone line (GSM) recordings) - 501 turns, transliteration (VM II Format).

VM CD 61.1 - VM61.1 (BAS-Edition)
Japanese - 19 spontaneous dialogues (19 close mic, 0 room mic, 0 phone line (GSM) recordings) - 946 turns, transliteration (VM II Format).

VM CD 62.1 - VM62.1 (BAS-Edition)
Japanese - 21 spontaneous dialogues (21 close mic, 0 room mic, 0 phone line (GSM) recordings) - 981 turns, transliteration (VM II Format).

VM CD 51.1 - VM51.1 (BAS-Edition)
Multilingual German/English with human interpreter (3 channels) - 15 spontaneous dialogues (15 close mic, 0 room mic, 0 phone line (GSM) recordings) - 873 turns, transliteration (VM II Format).

VM CD 52.1 - VM52.1 (BAS-Edition)
Multilingual German/English with human interpreter (3 channels) - 13 spontaneous dialogues (13 close mic, 0 room mic, 0 phone line (GSM) recordings) - 728 turns, transliteration (VM II Format).

VM CD 55.1 - VM55.1 (BAS-Edition)
Multilingual German/English with human interpreter (3 channels) - 11 spontaneous dialogues (11 close mic, 0 room mic, 0 phone line (GSM) recordings) - 518 turns, transliteration (VM II Format).

VM CD 56.1 - VM56.1 (BAS-Edition)
Multilingual German/English with human interpreter (3 channels) - 12 spontaneous dialogues (12 close mic, 0 room mic, 0 phone line (GSM) recordings) - 620 turns, transliteration (VM II Format).

VM CD 57.1 - VM57.1 (BAS-Edition)
Multilingual German/Japanese with 2 human interpreters (4 channels) - 11 spontaneous dialogues (11 close mic, 0 room mic, 0 phone line (GSM) recordings) - 702 turns, transliteration (VM II Format).

VM CD 58.1 - VM58.1 (BAS-Edition)
Multilingual German/Japanese with 2 human interpreters (4 channels) - 7 spontaneous dialogues (7 close mic, 0 room mic, 0 phone line (GSM) recordings) - 421 turns, transliteration (VM II Format).

VM CD 59.1 - VM59.1 (BAS-Edition)
Multilingual German/Japanese with 2 human interpreters (4 channels) - 7 spontaneous dialogues (7 close mic, 0 room mic, 0 phone line (GSM) recordings) - 354 turns, transliteration (VM II Format).

VM CD 63.0 - VM63.0 (original edition)
German - 14 WOZ dialogues designed to evoke emotions (mainnly anger) - transliteration, emotion labeling.

VM CD 64.0 - VM64.0 (original edition)
German - 13 WOZ dialogues designed to evoke emotions (mainnly anger) - transliteration, emotion labeling.

VM CD 65.0 - VM65.0 (original edition)
German - 13 WOZ dialogues designed to evoke emotions (mainnly anger) - transliteration, emotion labeling.