The ELRA Newsletter

October - December 2000

Vol.5 n.4

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

Editor:
Khalid Choukri &
Valérie Mapelli

Layout:
Emilie Marquois

Contributors:
Key-Sun Choi
Christian Galinski
Ana García-Serrano
Henk van den Heuvel
Paloma Martínez
Enrique Vidal

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,

This is the last issue of the ELRA Newsletter to be printed on the 20th century! The first issue was printed in March 1996, five years ago. This is an important milestone in our professional lives and it is useful to try and draw a picture of the progress of the Human Language Technologies (HLT) area, from the multilingualism perspective which is the core contribution of ELRA. Through the supply of Language Resources, we expected to boost the deployment of new technologies as well as the transfer of proven technologies to new languages.

In a recent survey, we collected information about the evolution of HLT with respect to the languages that were/are handled. The survey shows that for speech applications (Text-to-Speech, Dictation, Telephony), only 2 companies reported the availability in 1995 of deployed technologies in about 9 different languages while in 1999, 8 companies reported the availability of such systems for about 31 different languages. All anticipate to offer more than 200 products in different languages by 2005. Out of this, 30 products will focus on English, 90 on the main Western European languages, 13 on Eastern European languages, 34 on the other European languages, 37 on the main Asian languages and only 13 will focus on the other languages. Similar figures were obtained for MT (a more detailed report is available on our site - "Members only" section). We are very proud to bring in our own contribution.

During the last quarter we continued our work on the LRsP&P Project (a EU project granted to ELDA). This project led to the production of several key resources which are now ready for distribution. Some of them went through an external validation to check the quality of the data with respect to the specifications.

We also devoted some time to the preparation of new proposals submitted within the European Commission IST program. In particular a proposal entitled "Coral-rom" has been accepted for funding. It aims at building a large database of aligned corpora for 4 spoken romance languages. We will report on this proposal and the others in coming issues of this newsletter.

A Board meeting took place in Paris on October 23. A major theme debated during the meeting was about the validation of Language Resources being distributed via ELDA. It is generally agreed that we need to add a "quality flag" to our catalogue to ensure that our customers get reliable information about the data they purchase. It has also been agreed to set up and run a "bug reporting" procedure, using our web facilities, to get feedback from data users. This will be detailed in a next issue of the newsletter, but as an introduction, Henk van den Heuvel, from SPEX (our Spoken Language Resource Validation Unit) elaborates on the major problems related to this important topic, in a paper enclosed herein.

The GEMA project, in which ELDA is involved, is progressing as planned. It aims at providing a central and organised access point for the linguistic sector, by building and developing a linguistic portal. A number of technical aspects have been addressed such as the conversion of various formats of terminological resources into a standard one, the implementation of e-commerce techniques for accessing language resources (in particular terminology databases) and other related services. The GEMA project should lead to a referential portal and is expected to go public by the first quarter of 2001.

During this quarter, we continued our efforts to secure new resources for distribution. As usual, these resources are described in the last section of this newsletter and concern the Hungarian and Estonian speech databases produced within the Babel project, the Albayzin corpus of Spanish produced in a large Spanish national effort, the Portuguese part of the SpeechDat-II databases, the Polish part of the SpeechDat(E) project, a very interesting speech database consisting of recordings of twin's speech, tuned to speaker identification/verification problems, a new French corpus with scientific texts (with SGML markup).

A first set of very interesting resources produced by ELRA in the scope of LRsP&P project are now available and are described in this volume; these are: a British English onomasticon dictionary (a pronunciation lexicon of over 160,000 entries of british place names and proper names), a multilingual Russian-English English-Russian dictionary (XML-based). We are particularly proud to announce our first broadcast news corpus, 30 hours of Italian data.

Last but not least, there are new releases of Verbmobil resources (of spontaneous speech recorded in a dialog task in German, Japanese and American English).

We have also concluded agreements with some speech data providers to supply us with data to be used for evaluation purposes within the Aurora project (see the announcement of the EuroSpeech special event enclosed in this issue).

In addition to the paper on "The Art of Validation" and the announcement of the EuroSpeech special event, this issue contains an annoucement of the 8th MT Summit, an article on the EuTrans project achievements (Example-based speech-to-speech translation), a paper on the work being carried out on Document retrieval systems at the University Carlos III and the Technical University of Madrid. We also continue our brief summaries of LREC event, through a report on the LREC 2000 pre-conference workshop on "Terminology resources and computation".

In a few days, we will be starting a new century and millenium. On behalf of the ELRA Board and the ELDA staff, we wish you a happy new year, a happy century and a wonderful HLT odyssey. A century that will probably see most of dreams become reality (maybe not in 2001 !).

Antonio Zampolli, President

Khalid Choukri, CEO

Conference Announcements

MACHINE TRANSLATION SUMMIT VIII

September 18-22, 2001, Santiago de Compostela, Spain

The 8th Machine Translation Summit, organized by the European Association for Machine Translation (EAMT), will be held in Santiago de Compostela, Spain, from 18 to 22 September 2001. MT Summit VIII, which is the first conference of the century in the premier series of conferences on machine translation, will provide a forum for discussing the prospect of MT and related areas in the coming century. MT Summit VIII will feature an expanded programme including research papers, reports on users' experiences, discussions of policy issues, invited talks, panels, exhibitions, tutorials, and workshops. EAMT invites all who are interested in any aspect of machine translation and tools for translation support - researchers, developers, providers, users, and watchers - to participate in the conference.

Conference Schedule

18-19 September 2001	Tutorials, workshops, excursions
20-22 September 2001	Papers, panels and exhibitions

Important Dates

15 Dec. 2000	Workshop and tutorial proposals
15 Jan. 2001	Notification
31 Jan. 2001	Speaker and panel suggestions
15 April 2001	Paper submission deadline
15 April 2001	Exhibition registration
30 May 2001	Notifications
1 July 2001	Final camera-ready copy deadline

Further Information

For more details, please visit the Web-site: http://www.eamt.org. You may also send a request for information to summitVIII@eamt.org.

EuroSpeech Special Event

NOISE ROBUST RECOGNITION

Robust Algorithms and a Comparison of their Performance on the "Aurora 2" Database

In conjunction with the EuroSpeech 2001 Conference - http://eurospeech2001.org

Noise robustness is an important area of scientific investigation with commercial relevance. Many novel and interesting algorithms have and continue to be developed to address this problem. Each technique is often evaluated in a different way and on a different database making cross comparison of their relative effectiveness difficult to assess. The objective of this special event is for researchers to present leading edge algorithms for noise robustness and their results measured on the same database. It is hoped that not only will the research community benefit from comparing techniques and reviewing scientific progress but also the process of evaluating on a common database will stimulate new ideas.

What makes this special session different from the main conference is that each paper will be required to submit results on the evaluation database. The Aurora 2 database has been chosen for this .

While the database was designed for the evaluation of front-end algorithms, and there is a reference HMM back-end configuration of HTK to enable this, the Aurora 2 database is also suitable for other noise robustness techniques including the back-end. Note that there is also a reference Mel-Cepstrum Front-end.

Conference Schedule

September 3 - 7, 2001	Eurospeech 2001 - Scandinavia

EuroSpeech web site : http://eurospeech2001.org/information/eurospeech_special_event.htm

Important dates

30 March 2001	Paper submission deadline (results on Aurora 2 must be included)
Until June 15, 2001	Early registration
8 June 2001	Notification of acceptance
Until August 1, 2001	Advance registration
After August 1, 2001	Late and on-site registration
3-7 Sept 2001 (day TBD)	Eurospeech Special Session

The Aurora databases have been made available publicly through ELRA. Aurora has also prepared real-world noise databases using subsets of the Speechdat-Car project collections: the Finnish and the Spanish subsets are available, Danish and German languages will be available on 1st Feb 2001 from ELRA.

Further Information

Please send an email to David Pearce (bdp003@email.mot.com) in advance if you intend to submit a paper so we can keep you informed of any updated information.

The Art of Validation

Henk van den Heuvel, SPEX, The Netherlands

Introduction

An increasing number of Spoken Language Resources (SLRs) in ELRA's catalogue contains a remark such as: "The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications." Some may read such a sentence in "dustbin-mode", so without paying attention to it, but others may be interested in the background and contents of such a validation procedure. This article serves to satisfy the curiosity of the latter group of readers, at least to some extent.

Validation of SLRs may refer to a variety of actions:

1. checking a SLR against a fixed set of requirements;

2. putting a quality stamp on a SLR as a result of the aforementioned check. If the database passes the check, then we say that it has been "validated";

3. the evaluation of a SLR in a field test, thus testing the usability of the LR in an actual application.

4. …

SLR validation, as carried out by SPEX (acronym for Speech Processing Expertise Centre) , typically refers to the first type of action: the quality evaluation of a database against a checklist of relevant criteria. These criteria are typically the specifications of the databases, together with some tolerance margins in case deviations are found.

The validation of language resources in general, and SLRs in particular, is a rather new type of activity in the area of language and speech technology. As more and more SLRs are entering the market, the need for validation of these resources increases, and therefore the best ways to accomplish validation need to be established. Validation of SLRs is of particular interest to the European Language Resources Association and its distribution agency ELDA (http://www.elda.fr/). ELRA offers a wide range of SLRs in its catalogue. Before distribution can proceed, the products must be subjected to quality control and validation. ELRA has established manuals for validation and has been actively persuading producers of Language Resources to adopt these as a means of adding value to the marketability of their products. ELRA, therefore, has started instituting a system that, in the long term, will yield a specification and quality control document to be issued with every product that ELRA sells or licenses. In order to evaluate the quality of the SLRs in the ELRA catalogue, a procedure to describe and validate these SLRs has to be developed. ELRA entrusted this task, after an open call, to SPEX. SPEX constitutes the first SLR validation unit of ELRA's Validation Network.

In this contribution I will give an overview of various aspects of SLR validation and present some future directions in this field, especially with respect to SPEX's validation mission for ELRA.

What is there?

The first SLRs that were formally validated were the databases of the collaborative EC funded SpeechDat(M) project. An important internal motivation for this SLR validation was the idea that all partners should exchange equivalent databases within a project. For this reason, validation also was used in the sense of the second interpretation given above: validation as a binary quality stamp: pass or reject. Only databases which passed the validation were released by the consortium. SpeechDat has created an impressive off-spring. Table 1 presents an overview of the projects in, what is nowadays called, the SpeechDat "family".

The SpeechDat formula was, in addition, also used for a number of other data collections, as shown in Table 2. Also here, a formal SLR validation was carried out by SPEX.

Also the SLRs collected in the Speecon project (Siemund et al., 2000) will be collected more or less according to the SpeechDat standards. All SLRs mentioned above will be offered to ELRA for distribution.

How do we do it?

As I see it, SLR validation operates along two dimensions with two points on the axis of each dimension. The first dimension concerns the integration of validation into the specification phase. Along this axis validation can be performed in two fundamentally different ways: (a) Quality assessment issues are already addressed in the specification phase of the SLR. That is, throughout the definition of the specifications, the feasibility of their evaluation and the criteria to be employed for such an evaluation are taken into account. (b) A SLR is created, and the validation criteria and procedure are defined afterwards. In this way, validation may boil down to reverse-engineering and the risk is faced that the validation of some parts of the specification may become infeasible. As for the second dimension, validation can be done (a) in-house by the SLR producer (internal validation) or (b) by another organisation (external validation). The two dimensions thus identified are shown in Table 3.

Compartment (1) in this table points to an essential element for proper database production: Each database producer should safeguard the database quality during the collection and processing of the data in order to ascertain that the specifications are met. In this way, each producer is his own validator. An internal final check (2) should be an obvious, be it ideally superfluous, part of this procedure. Alternatively, or in addition, an external organisation can be contracted to carry out the validation of a SLR. In that case the best approach is that the external validator is closely involved in the definition of the specifications (in order to assess the feasibility of corresponding validation checks), and performs quality checks for all phases of the production process (3), followed by a final check after database completion (4). (3) and (4) are more objective quality evaluations, and should be considered important for that reason.

Table 1. Overview of SpeechDat projects. CDB = Car databases; FDB = Fixed (telephone) Network databases; MDB = Mobile network (telephone) databases; SDB = Speaker Verification databases.

Project	**SLR**	**Period**	**Ref.**
SpeechDat(M)	8 FDB	1994-1996	Höge & Tropf (1996)
SpeechDat(II)	20 FDB 5 MDB 3 SDB	1995-1998	Höge, et al. (1999)
SpeechDat-Car	9 CDB	1998-2001	Van den Heuvel, et al (1999)
SpeechDat-East	5 FDB	1998-2000	Pollak, et al. (2000)
SALA	4-5 FDB	1998-2000	Moreno, et al. (2000)

Table 2. Overview of projects collecting data according to SpeechDat protocols.

Language	**SLR**	**Producing Company**	**Ref.**
Russian	1 FDB	Auditech (for Siemens), Petersburg, Russia	Pollak, et al. (2000)
Austrian German	1 FDB 1 MDB	FTW, Vienna, Austria	Baum et al. (2000)

Table 3: Four types of validation strategies

Validator	**Validation scheduling**	
	During production	After production
Internal	(1)	(2)
External	(3)	(4)

The optimal strategy is to have all (1), (2), (3), (4) done. In fact, this strategy was adopted by the SpeechDat projects, where all producers performed internal quality checks, whilst SPEX served as an independent external validation centre, being closely involved in the specifications and performing intermediate and final quality assessments.

Validation Procedure

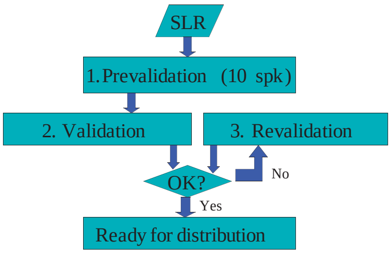

Figure 1. SLR validation procedure in SpeechDat-related projects

As shown in Figure 1, validation in "SpeechDat style" proceeds in three steps:

1. Prevalidation of a small database of about 10 speakers shortly after the design specifications have been established and the recording platforms installed. The objective of this stage is to detect serious (design) errors before the actual recordings start. This stage also allows partners to build their database compilation software in an early stage of the project. This corresponds to strategy (3) in Table 3.

2. Validation of the complete database. The database is checked against the SpeechDat specifications and a validation report is edited. This stage corresponds to strategy (4) in Table 3.

3. Revalidation of a database. In case the validation report shows that corrections of a database are necessary or desirable, then (part of) the database can again be offered for validation, and a new report is written. In horrendous cases this phase may show some iterations.

In SpeechDat projects the eventual decision about the approval of a database is not made by SPEX, but by the consortium concerned. In fact, the consortium performs validation in the second interpretation mentioned in the introduction: putting a quality stamp on a product.

Back to Table 3. For obtaining the highest SLR quality the numbers in the compartments in the table reflect the order of importance of validation strategies: The internal quality control during production is the most important quality safeguard. In contrast, to have only an external validation after the database is produced is the least preferable option.

ELRA resources are distributed "as-is with all defects" as stated in the licenses. The databases are created (and sold), but a thorough validation has yet to be carried out for the majority of the SLRs in the catalogue. Of course, one may have some faith that internal quality checks in the spirit of (1) and (2) took place for individual databases, but an objective external validation is a valuable, if not necessary, additional means of quality assessment.

Validation and improvement

A principal issue concerns the difference between validation and improvement of a SLR. At first sight, both seem closely intertwined. Who could better rectify the errors in a database than the person (or institute) that was smart enough to detect the errors? Nonetheless, a principal stance should be taken here. In SPEX's view, validation and improvement should be clearly distinguished. There are differences with respect to:

1. Nature of the actions: Validation is a quality assessment procedure and therefore a diagnostic operation.

2. Chronology: Validation yields the diagnosis; the improvement is the cure. Therefore, SLR validation should obviously precede SLR improvement.

3. Responsible institutes: In principle, the validator and the corrector should be different institutes, in order to avoid the undesirable situation that the validating institute assesses its own work. The correction of a SLR is accordingly in principle a responsibility of the SLR owner.

What is checked?

SLR validation criteria come in the following categories:

1. Documentation. It is checked if all relevant aspects of a SLR (see 2-8 below) are properly described in terms of the three C's: clarity, completeness and correctness.

2. Database format. It is checked if all relevant files (documentation, speech files, label files, lexicon) are present in the appropriate directory structure and with the correct format.

3. Design. The appropriateness of the recorded items for the purpose of the envisaged application(s) and the completeness of the recordings should be checked.

4. Speech files. The acoustical quality of the speech files is measured in terms of (e.g.) (average) duration, clipping rate, SNR, mean sample value. Also auditory inspection of signal quality belongs to this category.

5. Label files. The label files should obey the correct format. Ideally, they can be automatically parsed without yielding erroneous information.

6. Phonemic lexicon. The lexicon should contain appropriate phonemic (or allophonic) transcriptions of all words in the orthographic transcriptions of a SLR.

7. Speaker & environment distributions. The recorded speakers should present a fair sample of the population of interest in terms of (typically) sex, age and dialectal background. Also the recording environments should be representative for the targeted applications.

8. Orthographic transcriptions. A (native) speaker of the language should check a sufficiently large sample of the orthographic transcriptions by comparing these to the speech in the signal files and the transcription protocol.

An example of an extensive list of validation criteria in terms of specifications and tolerance intervals is given in Van den Heuvel (1996).

Rank order of validation check points

The acoustic quality of the speech files is of utmost importance. Although the desired quality may to a great deal depend on the wishes of the customer or on the targeted applications, it is obvious that recordings containing rubbish disqualify for being included in a speech database. Further, the clarity, completeness and the correctness of the documentation is a first order requirement for any SLR that deserves this name. Also, only a proper transcription of the speech qualifies the database as more than a mere collection of speech signals. In summary, at SPEX we consider documentation, transcription, and good speech signals as the core ingredients of a SLR, which should have the highest validation weight.

On the second level in the validation rank order follow: completeness criteria for the design of the SLR and for the recordings actually contained in the database, and completeness criteria for distributions of speakers and environments, etc.

The third level of priority concerns SLR aspects that can be easily corrected afterwards, such as the phoneme lexicon, the formatting of the annotation files and the directory tree structure and file nomenclature of the database. Of course, errors on this level may be very frustrating when one uses the database, but the important thing for database validation is that they can be relatively easily fixed. In fact, also the documentation files could be considered as part of this third priority level, since they can be easily modified as well. The reason why we in contrast consider documentation as a priority 1 matter is that a good documentation is a prerequisite for a sensible database validation.

Quality labels can be attached to each aspect of the database. Our quality labels have three possible values: 1. not acceptable; 2. not OK, but acceptable; 3. OK.

Table 4 gives a summary of the priority weights and quality values that can be attached to the SLR characteristics. SPEX regards this scheme as the key framework to validate SLRs in the ELRA catalogue.

Who is responsible for what?

The validation and improvement of a SLR involves two players: (1) The validation institute which assesses the quality of a database and reports its deficiencies; (2) the database owner taking care of the improvements that become necessary after such a report. In the specific case of SPEX performing the validation for ELRA, ELRA is a third player. As a matter of fact, SPEX as a validation institute acts as the intermediary between ELRA and the database owner. The Board of ELRA is represented by the

Speech College members of the Board. The ELRA Board strives for a validation of the SLR in its catalogue; the database owner may be asked to supply an improved database if deficiencies of the database show up, and SPEX carries out the validations and takes care of the communication between ELRA and the database owner. Further, the ELRA Board decides or affirms the priority list with which SLR have to be validated (i.e. priority in time); it determines the corrections that have to follow after a validation.

The procedure can be captured by the action list given in Table 5. In vertical direction this table reflects a rough time axis. For SPEX, the role of intermediary between A and C holds for the full validation process.

Bug reports

Errors in a database do not only emerge during the validation procedure. Errors are also typically detected by clients once they use the database. An efficient means of bug reporting and appropriate procedures for updating a SLR and distributing a new release should, therefore, be an integral part of permanent quality maintenance.

Below is presented the procedure for ELRA that we see as the most promising for the time being, and which SPEX intends to start with. This procedure can easily be combined with the validation/correction procedure presented just before.

1. A link to a *bug report sheet* is created at ELRA's WWW home page

2. The bug report sheet is a frame based sheet, with slots for the information like: Database name; Code in ELRA's catalogue; Coordinates (name, affiliation, e-mail address) of the reporter; Errors to report.

3. Lists of all reported bugs for each SLR in the catalogue are made available through ELRA's home page and can be accessed by ELRA members.

4. Depending on the seriousness and the number of the bugs reported, SPEX recommends a SLR for validation and/or correction. The decision is made by the ELRA Board, and the steps indicated in Table 5 are followed.

Who comes first?

The order in the priority list of SLRs to be validated is driven by several factors. First of all the number of copies sold through ELRA gives a good indication of the market value of a database and hence of the need to have this database in an optimal condition. On the other hand, if this database has already been validated before (as it is the case with the databases in the SpeechDat projects), then a (new) validation should have lower priority (but this is something that practice should prove).

Furthermore, the bug reports are also indicative of the condition of a database. If many and serious bugs are reported for a SLR, then rapid action should be taken. In that case, we recommend to give a database a thorough validation first in order to have the major shortcomings detected at once. This is in agreement with the general strategy pointed out above to precede SLR improvement by a validation. To insert a validation between bug reports and SLR improvements serves two purposes:

1. Verification of the reported bugs

2. Guarantee that the most serious other bugs are found in one action

Therefore, in summary, the following determinants for prioritising SLR validation are considered:

- The numbers of copies sold / expected to be sold through ELRA
- The number and seriousness of errors reported via bug reports
- Availability of reports of previous validations

Table 4: Quality assessment methodology for existing SLRs in ELRA's catalogue. See the text for clarifications for rank orders and quality labels.

Database part	**Rank order**	**Quality value**		
Documentation	1	1	2	3
Transcription	1			
Speech signal	1			
SLR completeness	2			
Speaker distributions	2			
Recording conditions	2			
Annotation files	3			
Lexicon	3			
Formats & file names	3			

Table 5:General procedure and responsibilities for the validation and improvement of SLRs in the ELRA catalogue.

A. ELRA	**B. SPEX**	**C. SLR owner**
	Makes priority list (see section 8 below)	
Decision of SLR validation		
	Intermediary between A and C Performs validation and makes report	
		Reaction to validation report/results
Decision on necessary corrections		
		Corrects and updates the SLR

Future plans

SPEX has established a first priority list of SLRs in ELRA's SLR catalogue that need validation. The idea is to validate various SLRs this year, following the quality chart presented in Table 4. Plans are being developed in order to make a validation protocol for Broadcast News databases, as part of the new MLIS project NETWORK-DC.

References

Baum, M., et al. (2000): *SpeechDat AT: Telephone speech databases for Austrian German*. Proceedings of the LREC'2000 Satellite workshop on XLDB - Very large Telephone Speech Databases, Athens, Greece, pp. 51-56.

Höge, H., Tropf, H.S. (1996): *Final Report. SpeechDat(M) Technical Report D0.6 & 0.7*. http://www.icp.grenet.fr/SpeechDat/home.html

Höge, H., et al. (1999): *Speechdat multilingual speech databases for teleservices: across the Finnish line*. Proceedings EUROSPEECH'99, Budapest, Hungary, 5-9 Sep. 1999, Vol. 6, pp. 2699-2702

Moreno, et al. (2000): *SALA: SpeechDat across Latin America. Results of the first phase*. Proceedings LREC2000, Athens, Greece, pp. 877-882.

Pollak, P., Czernocky, J., Boudy, J. et al. (2000): *SpeechDat(E)- Eastern European telephone speech databases*. Proceedings of the LREC'2000 Satellite workshop on XLDB - Very large Telephone Speech Databases, Athens, Greece, pp. 20-25.

Siemund, R., et al. (2000) *SPEECON - Speech Data for Consumer Devices*. Proceedings LREC2000, Athens, Greece, pp. 883-886.

Van den Heuvel, H. (1996): *Validation criteria*. SpeechDat Technical Report SD1.3.3. http://www.speechdat.org/SpeechDat.html

Van den Heuvel, H., et al. (1999): *The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results*. Proceedings EUROSPEECH'99, Budapest, Hungary, pp. 2279-2282.

Dr Henk van den Heuvel
SPEX / A2RT
Dept. of Language and Speech
University of Nijmegen
P.O.Box 9103
NL-6500HD Nijmegen
E-mail: H.v.d.Heuvel@spex.nl
http:/lands.let.kun.nl

The EuTrans Speech-to-Speech Translation Project

Enrique Vidal, Universidad Politécnica de Valencia, Spain

The EuTrans project (**E**XAMPLE-BASED LANG**U**AGE **TRANS**LATION **S**YSTEMS) has come to its successful completion on August, 2000. It has entailed a tight three-year collaboration between four partners: the *Instituto Tecnológico de Informática* (ITI, Valencia, Spain), the *Rheinisch-Westfälische Technische Hochschule Aachen Lehrstuhl für Informatik VI* (RWTH, Aachen, Germany), the *Fondazione Ugo Bordoni* (FUB, Rome, Italy) and *ZERES GmbH* (Bochum, Germany). It started within the *Open Domain* of the *Long-Term Reseach (LTR)* ESPRIT programme as a continuation of a short first-phase LTR ESPRIT project which was also called EuTrans. This first phase of the project will be referred to as EuTrans-I.

In EuTrans-I the viability and adequacy of using Example-Based, Finite-State technology for limited-domain (text and speech) Language Translation was assessed. While good results were obtained in a relatively simple task (called "Traveler Task"), it also pointed out the necessity of extending the baseline techniques in order to deal with increasingly complex, natural and spontaneous tasks.

In this direction, the second-phase of EuTrans aimed at exploring alternative Example-Based Machine Translation approaches that (a) are useful by themselves and/or (b) can be adequately combined with the Finite State approaches. The main goal was to demonstrate useful performance in medium-complexity, limited-domain real-world applications; i.e., applications involving *spontaneous (spoken)* language with a vocabulary of a few thousands of words (or much larger in the case of text-input). Aiming at these general goals, the following objectives were proposed :

1. *To collect two adequate text-input and speech-input MT corpora.*

2. *To further develop finite-state learning techniques introduced in EuTrans-I.*

3. *To investigate complementary example-based translation techniques and statistical approaches in particular.*

4. *To implement appropriate text-input and speech-input translation prototypes.*

The most significant work carried out throughput the project towards these objectives is summarized below.

Speech and text bilingual data acquisition

The following MT tasks have been defined and the corresponding corpora have been collected (see Table 1) :

The ZERES corpus corresponds to a natural German-English *text-input* MT application which entails the translation of different text types belonging to the domain of tourism: bilingual Web pages of hotels, bilingual tourism brochures and business correspondence. Data collection was based on semiautomatic processing and alignment of scanned documents, web pages and other sources of information.

The FUB *speech-input* corpus corresponds to a person-to-person communication task consisting in the translation into English of queries, requests and complains made through the telephone to the front desk of a hotel in Italian. The collection of this corpus has been based on the Wizard of Oz paradigm. This way, the acquired text and speech data are reasonably realistic for the task considered [DiCarlo99].

In addition, a small subset of the "Traveler Task" corpus produced in EuTrans-I was selected and considered as a *standard benchmark* data set. Since this corpus is simpler and better controlled, even small variations in TWER (Translation Word Error Rate) do reflect true differences in performance. For this reason, it has turned out to be quite useful in experiments requiring careful comparison of different techniques. Spanish telephone-speech utterances corresponding to a part of this corpus have also been collected to allow for *speech-input* experimentation.

Finite-State and Translation Memory technologies

Finite-State (FS) models are particularly interesting for MT because of their great adequacy for speech-input operation. In fact these are the only models known so far that allow for simple, efficient and tight integration of the speech-recognition and translation processes [Vidal97]. On the other hand, Translation Memory (TM) techniques are among the most promising approaches for practical text-input MT.

The work on FS modeling departed from the baseline models used in EuTrans-I; namely Subsequential Transducers which were learned by the "Onward Subsequential Transducer Inference Algorithm" (OSTIA) [OGV97]. A crucial idea was to rely on bilingual alignments provided by statistical techniques to assist the learning of the FS transducers. This has dramatically reduced the amount of training data originally required by OSTIA, directly leading to the so-called "OMEGA" algorithm [Vilar00].

On the other hand, a new training approach has been introduced which is not based on the OSTIA state-merging paradigm. The new technique uses the alignments to obtain a homomorphic image of each training pair in the form of a standard string of "meta-words" which combine input and output lexical tokens. Using these training strings, conventional N-Gram language models are learned. The final step consists in computing an "inverse morphism", which converts the N-Gram into a FS transducer. This new technique, which is called "Morphic Generator Transducer Inference" (MGTI) [Casacuberta00], has yielded the best results among all FS techniques.

In addition to the standard statistical bilingual alignments, other more specific tools have been developed to assist the training of FS transducers; namely, *Error-Correcting*, *Word/Phrase Reordering*, *Automatic Categorization* and *Bilingual Segmentation* [ABC+97, AV98, VJA+98]. A summary of the best results achieved by FS techniques is shown in Table 2.

Work on Translation Memory, finally, has been devoted to improve and test the existing ZERES TM search engine. In particular, the use of grammatical representations, as provided by an HMM Part Of Speech tagger, has been explored.

Statistical Translation technology

The most interesting models and techniques developed are summarized here :

* *Quasi-Monotone alignment model*: This model and the associated search assume that input and output sequences of words admit an (approximate) monotonous, left to right alignment. The search has been extended to handle word re-ordering, if only a limited number of source sentence positions are actually reordered [NNO+00].

* *Alignment templates*: This alignment model allows matching of contiguous word groups rather than single words. The current formulation of this technique, which is explicitly based on statistical arguments [NNO+00], has consistently provided the *best results in all the corpora* and experimental settings tested throughout the project.

* *Iterative DP-based search*: This search algorithm is based on a dynamic programming-like algorithm which attempts to solve the basic MT Bayes equation using an iterative process. This process produces a series of solution refinements in which better solutions are built from the solutions achieved in previous iterations [GCN98].

A summary of the best results achieved by the different techniques is shown in table 3.

Integrating Speech Recognition and Translation

Work in this area has been done in the following main directions :

* *Acoustic modeling*: Using Italian (Spanish) speech input sentences of the FUB (Traveler) corpus, adequate acoustic models were trained. These models have been used in the speech recognition/translation experiments and in the speech-input EUTRANS prototypes. To check the quality of the trained models, speech recognition(-only) experiments were also performed using conventional trigram language models trained on input-language text sentences of the corpus.

* *Direct Coupling of Recognition-Translation*: In most systems, there is only a loose, serial interface between recognition and translation. In contrast, in this project a full, tight integration of the speech Recognition and Translation processes has been pursued. Work in this topic has led to a sound formulation of the problem which puts forward the sources of the difficulty and explains how a tight coupling between recognition and translation can be obtained [Ney99]. Following these ideas under the statistical MT framework, a concrete technique called *Speech-Input Iterative DP-search* has been developed and some initial tests have been carried out [GSC00].

Table 1 : Summary of corpora features

Corpus	ZERES (text only)		FUB		Traveler	
Input Output languages	German	English	Italian	English	Spanish	English
Sentences pairs	27 204		3 338		13 000	
Running words	501 655	565 023	61 423	72 689	132 154	134 882
Vocabulary	58 323	33 882	2 459	1 701	686	513
Bigram Test-Set Perplexity	-	121	31	25	8.6	5.2

Table 2 :Text-Input translation results (in %) achieved by different Finite-State MT methods. "TWER", "PER" and "SSER" stand for Translation Word, Position-independent translation word and Subjective Sentence Error Rates, respectively. In all cases, training was assisted by (IBM2 or IBM4) statistical alignments and, in some cases, Error-Correcting (EC), Automatic Bilingual lexical Categorization (ABC) and/or Automatic Bilingual Segmentation (ABS).

Task	Method	Assisted by	TWER	PER	SSER
Traveler	OMEGA/2gram	IBM2, EC, ABC	3.9	3.7	-
	MGTI/4gram	IBM4	8.0	7.6	-
FUB	OMEGA/2gram	IBM2, EC, ABC	36.5	30.0	-
	MGTI/4gram	IBM4, ABS	25.3	19.9	27.5

Table 3: Text-input Translation results (in %) achieved by different statistical MT methods. Systran results were obtained using the WWW interface accessible via http://babelfish.altavista.com/cgi-bin/translate ?. All the experiments with Alignement Templates used automatic Bilingual Lexical Categorization.

Task	ZERES			ZERES			Traveler	
Error rate	TWER	PER	SSER	TWER	PER	SSER	TWER	PER
Iter DP Search	-	-	-	61.0	37.1	-	13.9	12.8
Quasi-Monotone	68.9	58.3	61.8	29.6	22.4	29.4	10.8	10.0
Align Templates	64.2	52.70	57.4	25.1	19.0	24.2	4.4	2.9
Systran	74.0	65.9	59.9	-	-	-	-	-

* *Full integration of Recognition-Translation using Finite-State models*. Due to their finite-state nature, FS transducers are particularly appropriate for a complete integration of recognition-translation [Vidal97]. So far, this approach is the only one that has led to working integrated systems. The EUTRANS MT speech-to-speech translation prototypes are based on this approach.

For the FUB task, best recognition-only WER range from 22% to 35%, depending on the Language/Translation models used. Using the best integrated FS translation models, a TWER lower than 45% is obtained. It should be taken into account that, in unconstrained-speech tasks like this one, TWER can really be (pessimistically) misleading. In many cases, the system gives reasonably good translations with very high TWER. This is consistent with formal subjective tests and with the subjective experience with the on-line speech-input prototypes. Better results are obtained for the simpler Traveler task: 15.5% TWER for telephone speech and 6.8% TWER for microphone input.

Assessment

A reliable and non-expensive method for measuring the progress in the quality of the MT systems has been developed. In MT research a test-set is generally used many times over a relatively long period of time to keep track of system improvements and/or to compare the relative performance of different systems. In most cases, the results differ only in a small number of words. Based on this idea, a method was devised which takes advantage of previously obtained TWER and SSER scores. To this end, test sentences are stored in a Data Base (DB) [NOLN00], along with corresponding translations and scores evaluated so far. Then a user-friendly tool has been implemented which, by searching this DB, offers the following opportunities:

* automatically returns previous scores

* facilitates evaluation of new translations

* extrapolate scores

* offers new types of quality criteria

The SSER values reported in Tables 2 and 3 were obtained by using this tool.

Speech and text automatic translation prototypes

Four prototypes have been implemented:

* EUTEXT, a demonstrator of text-input Italian-English translation for the FUB task. It is based on a TclTk Graphical User Interface (GUI) and is used to demonstrate the practical performance of the Alignment Templates statistical technique and all the Finite-State MT techniques developed in the project.

* WTRANS, a WEB-based demonstrator of text-input translation for both the German-English ZERES task and the Italian-English FUB task. It is written in Java and is also general-purpose. Currently it has interfaces for two statistical techniques: Alignment Templates and Quasi-Monotone Alignments.

* EUTRANS, a speech-to-speech Italian-English translation demonstrator, which supports all kinds of finite-state translation models and is fully operational for the FUB task through standard telephone lines (plus an optional GUI).

* EUTRANS-I, a speech-to-speech Spanish-English translation demonstrator, similar to EUTRANS, which works in the domain of the EuTrans-I Traveler Task and is fully operational through standard telephone lines (plus an optional GUI).

All these prototypes run on low-cost hardware platforms such as Intel PC under Unix or Linux operating systems. Both speech-to-speech prototypes relay on the ATROS finite-state-oriented speech Recognition/Translation engine, developed in part and improved in this project. On the other hand, the output English speech is obtained by using a free-software Text-To-Speech synthesizer ("Festival") which offers understandable speech and reasonably good quality. The prototype for the FUB task achieves quite acceptable response time (about two or three times real time), while that for the EuTrans-I task often runs in less than real time, even on low-performance Pentium machines.

Dissemination and exploitation

Machine Translation in limited domain applications is of great interest for industry, especially in the European Community due to its multi-lingual nature. There is a huge market for text-input and speech-input Machine Translation, even in limited domain tasks : translation of product manuals, phone assistance services in a multi-lingual scenario, tourist information and services, reservations of hotels, trains, flights, etc., weather forecast, business letters, business conversations, etc. The main exploitation potential of EUTRANS rests in several lines of development which ZERES GmbH is currently following up.

Home Page and Web Presentation

The EuTrans home page is located at *http://www.zeres.de/Eutrans*. Its public area contains the project identification and description, pointers to the demonstration systems, informations about the partners of the project, etc.

References

[AV98] J.C. Amengual, E. Vidal: "Efficient Error-Correcting Viterbi Parsing". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.20, No.10, pp.1109-1116. October, 1998.

[ABC+97] J. C. Amengual, J.M. Benedí, F. Casacuberta, A. Castaño, A. Castellanos, V.M. Jiménez, A. Marzal, F. Prat, E. Vidal, J.M. Vilar (ITI): "Using Categories in the EUTRANS System". In *Procs. of the Spoken Language Translation Workshop*, pp. 44-53. Association for Computational Linguistics and European Network in Language and Speech. Madrid, Spain 1997.

[Casacuberta00] F. Casacuberta. "Inference of finite-state transducers by using regular grammars and morphisms". In "Grammatical Inference: Algorithms and Applications", A.Oliveira Ed. Springer-Verlag, LNCS, Vol.1891, pp.1-14, 2000. (Proc. of the 5th International Colloquium on Grammatical Inference. Lisbon, Portugal, September, 2000).

[Di Carlo99] A. Di Carlo (FUB) : *Telephone corpus collection in the EUTRANS project*. COST249 Meeting, Tampere, Finland, May 1999.

[GCN98] I. García-Varea, F. Casacuberta, H. Ney (ITI) : An Iterative, *DP-Based Search for Statistical Machine Translation*. Proceedings ICSLP98, pp.1135-113, Sydney Australia, Nov. 1998.

[GSC00] Ismael García-Varea, Alberto Sanchis and Francisco Casacuberta "A New Approach to Speech-input Statistical Translation". ICPR2000 Proc., Vol.2, pp.907-910. (15th International Conference on Pattern Recognition, Barcelona, Spain, September, 2000).

[NOLN00] S. Niessen, F. J. Och, G. Leusch and H. Ney. "An evaluation tool for machine translation: Fast evaluation for MT research". In *Proceedings of 2nd International Conference on Language Resources and Evaluation* (LREC), pp.39-45. Athens, Greece, May 2000.

[Ney99] H. Ney. (RWTH) "Speech translation: Coupling of recognition and translation.". In Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP99, pp.517-520, Phoenix, AR, March 1999.

[NNO+, 00] H. Ney, S. Niessen, F. J. Och, H. Sawaf, C. Tillman, S. Vogel. "Algorithms for Statistical Translation of Spoken Language". IEEE Trans. on Speech and Audio Processing. Vol.8, No.1, Jan. 2000.

[OGV97] J.Oncina, P.Garca, *E.Vidal*: "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol.PAMI-15, No.5, pp.448-458. Mayo, 1993.

[Vidal97] E. Vidal (ITI): "Finite-State Speech-to-Speech Translation". In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, Munich, April, 1997.

[VJA+98] J.M. Vilar, V. M. Jiménez, J. C. Amengual, A. Castellanos, D. Llorens, E . Vidal (ITI): "Text and Speech Translation by means of Subsequential Transducers". Cambridge University Press, Extended Finite State Models of Language, ACL Studies in Natural Language Processing series, 1998.

[Vilar00] J.M. Vilar "Improve the learning of Subsequential Transducers by using Alignments and Dictionaries". In "Grammatical Inference: Algorithms and Applications", A.Oliveira Ed. Springer-Verlag, LNCS, Vol.1891, pp.298-311, 2000. (Proc. of the 5th International Colloquium on Grammatical Inference. Lisbon, Portugal, September, 2000).

Enrique Vidal
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Spain
E-mail: e.vidal@iti.upv.es

MESIA: A Prototype of a Document Retrieval System that incorporates Linguistic Resources

Paloma Martínez, University Carlos III of Madrid & Ana García-Serrano, Technical University of Madrid, Spain

1. Introduction

The growing use of Internet has motivated additional demands of new information management techniques and effective search methodologies. The goal of this article is to show the work in progress in the MESIA[1] project, focused on the development of a metasearch engine with semantic capabilities for Web information retrieval (IR). The approach is based on the extraction of pragmatic knowledge from the documents retrieved by a conventional search engine. It is tested and validated on documents delivered by the Altavista search engine from the "Comunidad de Madrid" (Madrid Region) web site. By the usage of natural language processing (NLP) tools, the results of existing commercial search engines could be enhanced not only in the treatment of the user queries but also in handling the content of retrieved Web pages.

MESIA system expands the normal search (query and presentation of results) with new semantic capabilities and other aspects that consider the structure of WWW pages, the linguistic treatment of several text units automatically selected and the experience of usage. Currently, due to the fact that there are neither complete and correct natural language understanding systems nor general purpose linguistic resources, it is only possible to work in specific knowledge domains. That is why it is necessary to have domain-dependent knowledge that facilitates information search. Please, note that we do not propose a solution based on incorporating semantic knowledge in Web documents (such as extending HTML tags) but to extract semantic knowledge from the documents located by a traditional search engine.

Other active research features are including user profiles in IR as well as the ranking of results. The aim is to find other useful search methods apart from the full-text search and to develop a system closer to the user than to the IR system.

2. Description of the project

The aim of this project is to develop a metasearch engine that works both in natural language querying and in the post processing of the results obtained in a traditional search engine. Currently, existing search engines (AltaVista, Yahoo and others) are based on statistical analysis, to discriminate and select Web pages as an answer to a query. Purely statistical methods used in IR do not achieve optimal results. Usually, the result of keyword-based searches delivers many more documents than requested, i.e. irrelevant information, mostly caused by the exponential growth of information in Internet. On the other hand, other documents do not appear in the answer because they do not explicitly contain the query terms but other semantically related words. Thus, new strategies that profit from document data content as well as new mechanisms that may help user to define search criteria until the query is completely specified, taking into account the acquired experience, are required.

Three steps are proposed to face the information retrieval task:

(a) *Modifying the original query*. The system transforms the user's query, which is close to natural language, into a formal query by extracting the significant terms and expanding them by including morphological variants and synonyms. The result of this process is stored in a structure that contains information about the original query and those performed by the search engine.

(b) *A classification of documents that compose the search results*. The metasearch sys-

[1] This work is supported by the MESIA project (CAM-07T/0017/1998).

tem obtains from the document the required information for its identification from a query. This information supports the application of different criteria gathered both from a domain analysis and in an experimental way due to the lack of a unifying structure that stands for a classification of documents with an absolute certainty rate. Currently, there are two types of these criteria, structural (documents have four types of formats according to their content structure) and semantic (subject, disseminated purpose, etc.). The result of this process is a frame of features generated for each analysed document.

(c) *The accumulation of experience*. The system includes a Knowledge Manager for the documents it handles and also for the information collected from them. It is also foreseen the use of user profiles that will allow the system to decide whether the query is sent to the Knowledge Manager or a new search is launched. Knowledge Manager will incorporate knowledge about the most frequent queries made by each type of users as well as the outcome of document analysis corresponding to the previously performed queries. The currently available User Model is very simple (an ontology with a description based on the foreseen usage of the system for each type of user) but it supports, in some cases, to incorporate conditions into the formal query in order to delimit the metasearch answers.

Automatic handling of both query and significant texts included in documents makes possible to generate a conceptual structure that contains the relevant features obtained from analysis. This process requires the organisation of linguistic knowledge (general and specific terminology) as well as domain and process control knowledge. In order to design a knowledge-based system to be used in selective search through Spanish language, three types of knowledge have been identified from a manual analysis and the foreseen utilisation of the system:

1. Knowledge about documents structure and their classification according to different criteria.

2. Knowledge about the users that perform the queries: preferences (taking into account the historic database) and other positive or negative constraints.

3. Linguistic knowledge about domain sublanguage, specific vocabulary and expectative-based analysis considering significant expressions.

A software system that incorporates and articulates previous types of knowledge has been designed. Figure 1 displays the modules that compose MESIA system.

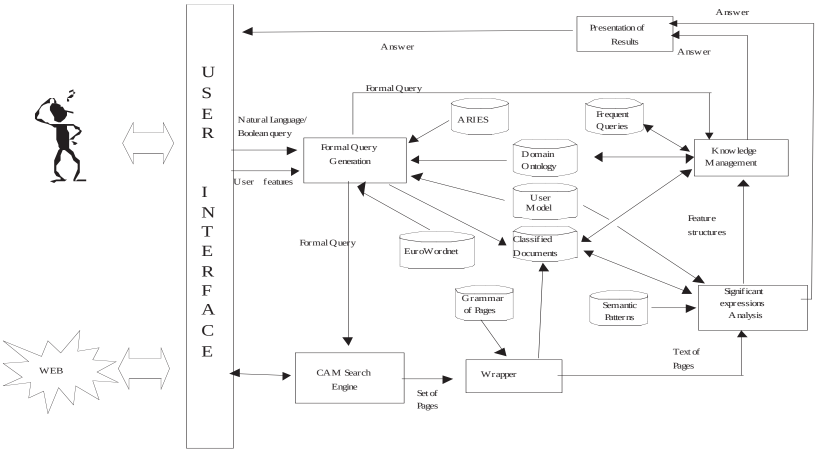

Figure 1: MESIA architecture

With the purpose of achieving robustness, user can input natural language queries (simple sentences) as well as boolean queries.

An overview description of the more relevant NLP techniques and resources included in MESIA system includes:

- ARIES (http://www.mat.upm.es/~aries/), Goñi et al. (1997), is a Spanish lexical platform developed by the Universidad Politécnica de Madrid and Universidad Autónoma de Madrid. ARIES is composed of a Spanish lexicon with around 38,000 lemma entries, including 21,000 nouns, 7,300 verbs, 10,000 adjectives and around 500 entries for prepositions, conjunctions, articles, adverbs and pronouns; some access utilities and a morphological analyser/generator are also included. Particularly, a DCG morphological generator for deriving words variants is being incorporated in MESIA system. This generator allows, for instance, obtaining number and gender forms from a nominal lemma.

- EuroWordNet (http://www.let.uva.nl/~ewn/), Vossen (1997), Gonzalo et al. (1998), is a lexical database that is structured as a top concept ontology that reflects different explicit opposite relationships. It can be seen as a representation of several vocabulary semantic fields. Moreover, it contains a hierarchy of domain tags that relate concepts in different subjects. The EuroWordNet database enables the user to work in different domains (a hierarchy of domains labels which relate concepts on the basis of scripts or topics) to separate the generic from the domain-specific vocabularies which is important to control the ambiguity problem in NLP.

- The *Wrapper* module is in charge of analysing the HTML pages retrieved by the CAM search engine in order to extract the textual information that these pages contain. Afterwards, the *Significant Expressions Analyser* treats these textual units. The Wrapper is based on a parser that uses a grammar describing the different relevant sections and subsections of HTML pages.

- A set of *Semantic Patterns* is used to guide the partial linguistic analysis of significant expressions trying to profit the most promising information according to domain terminology and keywords. The patterns also profit from a shallow parser that carries out a partial segmentation of specific textual units, Martínez y García-Serrano (1998). This analysis produces a structure of semantic features that superficially describes the text of a page. The structures of semantic features obtained are sent to the *Knowledge Manager* to be stored for future queries as well as to the Presentation of *Results* module in order to be organised and displayed to the user.

- The user query is stored along with the MESIA generated formal query in the *Classified Document* database that also contains structural information of the retrieved pages (title, paragraph, links, etc.) and their XML format, significance order according to several criteria, etc.

- Finally, *Knowledge Manager* handles the *Domain Ontology*, where the structures of semantic features are inserted once the document analysis has been performed. This ontology is a consensual and formal specification of a vocabulary used to describe the specific domain and contains the URLs linked by a set of domain concepts along with their semantic features.

Up to now, a first version of the MESIA system core has been implemented using JAVA language programming and MS Access as DBMS; the logic programming environment is CIAO-Prolog, Bueno et al. (1999).

References

Bueno et al. (1999), F. Bueno, D. Cabeza, M. Carro, M. Hermenegildo, P. López, and G. Puebla (1999) ``The Ciao Prolog System: A Next Generation Logic Programming Environment, REFERENCE MANUAL" The Ciao System Documentation Series Technical Report CLIP 3/97.1, The CLIP Group School of Computer Science Technical University of Madrid.

Gonzalo et al. (1998), J. Gonzalo, M.F. Verdejo, I. Chugur, Fernando López, Anselmo Peñas. "Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet". *Revista SEPLN*, nº 23, 1998.

Goñi et al. (1997), Goñi, J. M., González, J. C. y Moreno, A. ARIES: A lexical platform for engineering Spanish processing tools. *Natural Language Engineering*, 3 (4), pp. 317-345.

Martínez y García-Serrano (1998), Martínez, P. and García-Serrano, A. A Knowledge-based Methodology applied to Linguistic Engineering. In R. Nigel Horspool Ed., *Systems Implementation 2000: Languages, Methods and Tools*. London: Chapman & Hall, pp. 166-179, 1998.

Vossen (1997), Vossen, P. EuroWordNet: a multiingual database for information retrieval. DELOS workshop on Cross-language Information Retrieval, Zurich, 1997.

Paloma Martínez
Department of Computer Science
University Carlos III of Madrid, Spain
E-mail: pmf@inf.uc3m.es

Ana García-Serrano
Department of Artificial Intelligence
Technical University of Madrid, Spain
E-mail: agarcia@dia.fi.upm.es

Report on the Workshop on Terminology Resources and Computation

Pre-conference workshop to LREC2000, Athens, 29 May 2000

Key-Sun Choi and Christian Galinski

After a general introduction into the focus of the Workshop individual presentations are analysed (with authors indicated in square parentheses). The report ends with a conclusion.

1. Overview

Terminology Science (TS) deals with concepts (which are the main 'object' of TS represented by terms or other linguistic and non-linguistic symbols), conceptual relations (which are difficult to represent in language), the layout of terminologies, definitions or other kinds of descriptions of concepts, the computer representation of these concept representations as well as concept relations, etc. TS today thus comprises a concept theory (which can be considered as part of general epistemology), a representation theory, terminography (i.e. the methodology dealing with data elements, data models, layout, etc.), a theory of terminology management, and their practical applications in language and terminology planning, terminology work, specialized lexicography, etc. Following the scope of Sub-committee 3 "Computer applications in terminology" of the Technical Committee ISO/TC 37 "Terminology (principles and coordination)" this workshop focused on the computational (or engineering) viewpoint of terminology science.

One of the key issues in terminology computation is automatic terminology recognition and extraction from (text and speech) corpora, and detecting conceptual relations between concept representations in texts. Texts in this connection can mean any 'traditional' or electronic document or databases (largely containing alphanumeric-textual data) - preferably tagged or marked-up in a systematic way. Terminology resources include primarily terminology collections in conventional or electronic form, specific-domain corpora and their annotation.

The first problem of automatic terminology recognition is to identify word boundaries from strings of symbols of alphabetic letters, syllabaries or other script symbols [see Potipiti]. The next problem is automatic term selection from words or word combinations. Every word in principle can be a term - but in practice not every word is a term. Many papers are related to this term selection problem that is later described. The third problem is how to organize the identified terminological 'units' into some relation (hierarchy or non-hierarchical relations), whether it is logical relations or partitive relations or other types of relations. There are concept systems composed of a mixture of types of relations. Last but not least, the data on concepts, their representations and relations between each other should be kept in an easily accessible form - compatible with an internationally harmonized terminology interchange format. Some papers deal with a theory of machine learning that starts from the world of terminology.

This Workshop, therefore, can be subdivided into two main aspects: terminology computation and terminology resources. Automatic terminology recognition and related papers will be summarized in section 2. Papers about terminology resources are summarized in section 3.

2. Terminology computation

There is no paper concentrating on concept relations as such. However, almost all papers focus on how to recognize and extract (mono- and multi-word) terms or other concept representations from text corpora. Hereunder, a general introduction to automatic terminology recognition is given.

2.1. Automatic Terminology Recognition

A general introduction to automatic terminology recognition is given in section 2.1.1, and then this workshop's presentations are summarized.

2.1.1. General introduction to automatic terminology recognition

Automatic terminology recognition (ATR) is classified according to the type of corpora: e.g. monolingual or bilingual. The respective methodologies come from linguistic and statistical processing. Shallow syntactic processing is mainly employed to extract (complex) nominal units (nominal terms or phrases) under the assumption that the majority of terms are (complex) nominals [Bourigault92]. From the statistical point of view, relative frequency of terms is calculated on the basis of term frequency per domain. A terminological 'unit' may be a single word, a compound word or a combination of words [Damerau90,93]. Many mixed approaches have been reported on the use of both statistical and linguistic information [Justeson *et al.* 95, Lauriston96, Frantzi *et al.* 99, Maynard *et al.* 99]. The common approach in this discipline today is to apply various statistical approaches after shallow syntactic processing. Typical approaches are frequency-based [Justeson *et al.*95, Lauriston96], co-occurrence [Frantzi *et al.*99], and semantic information on context [Maynard *et al.*99]. The idea of the co-occurrence based approach is that words co-occurring with a certain terminological unit may provide environments for other similar or conceptually related terminological units. While this approach utilizes only surface patterns of words, Maynard *et al.*(99) investigates the semantic information on context: terms and their contextual words share similar word senses.

The current state-of-the-art in bilingual approaches is not in terminology recognition in itself, but in identifying the corresponding translation equivalents, assuming that terms have been already recognized in monolingual corpora. We expect that identified translation pairs help find the exact boundary of a given terminological unit in each language involved. Major approaches in this discipline are as follows: One is about the full automatic frequency-based alignment of translated pairs after monolingual term recognition [Daille, *et al.*94]. The other is about the machine-aided human detection of terminological units in monolingual text and automatic alignment of translated pairs [Dagan *et al.* 94].

2.1.2. This Workshop's focus on terminology computation

Major efforts in this workshop are geared towards automatic terminology recognition, presented by eight papers. Among them, three papers were about terminology recognition from monolingual domain-specific corpora. Further five papers discussed the term (or word) identification by comparison (or alignment) between bilingual corpora, e.g., Chinese-English, Japanese-English (two papers), German-English, and Swedish-English.

Li, et al. shows a web document based alignment using similar behavior of HTML tags and word forms. Tsujii et al. shows that a morpheme in Japanese can be one unit for alignment with an English term. Bilingual approaches are concerned with term recognition, but they seem not to show what is terminology. Carl proposes a new method called "invertible translation" for the English-German- language pair. Statistical methods are investigated by Tiedermann for English-Swedish, and Nakagawa for English-Japanese.

In terminology recognition from monolingual corpus, Oh et al. tries to find terminological units that consist of single word as well as multi-word units. They used (1) term frequency, (2) partial strings (or expressions) found in existing terminology dictionaries and foreign words, and (3) parenthetical expressions that expose abbreviations or translation pairs, e.g., GIS (Geographical Information System). Estopa, et al. identifies single-word terminological units from their contexts, Greek/Latin compounds, and semantic information. In Hisamitsu, *et al.*, a distribution based method is presented in the form of terminology distribution characteristics, assuming that the words of the general language (general purpose language - GPL) are distributed differently from domain-specific terms (in special-purpose languages - SPLs).

A fundamental study on word recognition was presented for unsegmented strings of Thai language corpora.

2.2. Application to dialog system

Bagga et al. shows an application of terminology detection by substring match in transcribed dialog after speech recognition in the medical domain.

3. Terminology resources

There are two presentations on terminology resources. Johnson et al. introduces the current European infrastructure for terminology resources. This overview of European projects mentions two aspects: shared management and dissemination of terminology resources as a basis for integrating terminology into the translation process. Shioda's work is a result of a comparative study of Japanese and Korean terms based on the investigation of occurrences in web pages. Terms of both languages are analyzed comparing their structure and constituents.

4. Conclusion

A terminological unit is a term (or other linguistic or non-linguistic concept representation) confined to a specific domain. A term may consist of a single word or combination of words. How can we differentiate homonymous terms having a special meaning in each domain where they occur? While terms are the most important terminological units in any given domain, keywords refer to important terms in a document. Such differentiation between different types of terminological units and their roles in cognition, communication, technical and scientific writing as well as specialized translation provides an insight to the variety of factors which have to be taken into account in order to 'find' a terminological unit in a given document and in existing language resources.

This study on automatic terminology recognition begins with automatic complex nominal recognition [Bourigault 92]. The assumption behind this idea is that a series of nouns tend to be a term. However, this general theory is not intended to identify what is a domain-specific terminological unit, but what is a term in general. Although it is a preprocessing phase for terminology recognition, we have to mention this stage under "shallow syntactic processing".

Contributions

Bourigault, D. (1992) *Surface grammatical analysis for the extraction of terminological noun phrases*. In Proceedings of the 14th International Conference on Computational Linguistics, COLING'92 pp. 977-981.

Dagan, I. and K. Church. (1994) *Termight: Identifying and terminology* In Proceedings of the 4th Conference on Applied Natural Language Processing, Stuttgart/Germany, 1994. Association for Computational Linguistics.

Daille, B., Gaussier, E., and Lange, J.M (1994) *Towards Automatic Extraction of Monolingual and Bilingual Terminology*,. In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94 pp. 515-521.

Damerau, F.J. (1990) *Evaluating computer-generated domain-oriented vocabularies*. Information Processing and Management, 26(4):791-801.

Damerau, F.J. (1993) *Generating and evaluating domain-oriented multi-word terms from texts*. Information Processing and Management, 29(4):433-447.

Frantzi, K.T. and S.Ananiadou (1999) *The C-value/NC-value domain independent method for multi-word term extraction*. Journal of Natural Language Processing, 6(3) pp. 145-180.

Justeson, J.S. and S.M. Katz (1995) *Technical terminology : some linguistic properties and an algorithm for identification in text*. Natural Language Engineering, 1(1) pp. 9-27.

Lauriston, A. (1996) *Automatic Term Recognition : performance of Linguistic and Statistical Techniques*. Ph.D. thesis, University of Manchester Institute of Science and Technology.

Maynard, D. and Ananiadou, S. (1998) *Acquiring Context Information for Term Disambiguation* In First Workshop on Computational Terminology Computerm'98, pp 86-90.

Key-Sun Choi
KORTERM
373-1, Kusung-dong, Yusong-gu, Taejon
305-701 Korea
E-mail: kschoi@cs.kaist.ac.kr

Christian Galinski
INFOTERM
Simmeringer Hauptstraße 24
A-1110 Vienna - Austria
E-mail: Christian.Galinski@chello.at

New Resources

ELRA-S0086 BABEL Estonian Database

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

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

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

- Many Talker Set: 30 males, 30 females; each to read 50 numbers, 1-2 connected passages, 1 block of "filler" sentences, and 1 block of syllables.
- Few Talker Set: 4 males, 4 females; each to read 50 numbers, 10 connected passages, 1 block of "filler" sentences, and 2-3 blocks of syllables.
- Very Few Talker Set: 1 male, 1 female; each to read 2 blocks of 50 numbers, 40 connected passages, 4 blocks of "filler" sentences, and 9 blocks of syllables.

And the extension part: a short description of Estonian sound system.

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

ELRA-S0087 BABEL Hungarian Database

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

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

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

- Many Talker Set: 30 males, 30 females; each to read 50 numbers, 1-2 connected passages, 1 block of "filler" sentences, and 1 block of syllables.
- Few Talker Set: 4 males, 4 females; each to read 50 numbers, 10 connected passages, 1 block of "filler" sentences, and 2-3 blocks of syllables.
- Very Few Talker Set: 1 male, 1 female; each to read 2 blocks of 50 numbers, 40 connected passages, 4 blocks of "filler" sentences, and 9 blocks of syllables.

And the extension part: a short description of Hungarian sound system.

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

The BABEL Bulgarian (ELRA-S0085), Estonian (ELRA-S0086) and Hungarian (ELRA-S0087) databases are available at ELRA. The BABEL Polish and Romanian databases will be available soon.

ELRA-S0089 Albayzin corpus

This corpus consists of 3 sub-corpora of 16 kHz 16 bits signals, recorded by 304 Castillian speakers.

The 3 sub-corpora are:

- Phonetic corpus: 6,800 utterances of phonetically balanced sentences, including 1,000 with phonetic segmentation.
- Geographic corpus: 6,800 utterances of sentences extracted from a Spanish geographic database.
- "Lombard" corpus: 2,000 utterances from various corpora.

	ELRA Members	Non Members
Price for Spanish research organisations	100 Euro	120 Euro
Price for Other research organisations	1,000 Euro	2,000 Euro
Price for Commercial organisations	10,000 Euro	12,000 Euro

ELRA-S0092 Portuguese SpeechDat(II) FDB-4000

The Portuguese SpeechDat(II) FDB-4000 comprises 4,027 Portuguese speakers (1,861 males, 2,166 females) recorded over the Portuguese fixed telephone network. The SpeechDat database has been collected and annotated by Portugal Telecom. This database is partitioned into 11 CDs. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:
- 1 isolated single digit,
- 1 sequence of 10 isolated digits,
- 4 numbers : 1 sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits),
- 1 currency money amount,
- 1 natural number,
- 3 dates : 1 spontaneous (date or year of birth), 1 prompted date, 1 relative or general date expression,
- 2 time phrases : 1 time of day (spontaneous), 1 time phrase (word style),
- 3 spelled words : 1 spontaneous (own forename), 1 city name, 1 real word for coverage,
- 5 directory assistance utterances : 1 spontaneous, own forename, 1 city of birth / growing up (spontaneous), 1 frequent city name, 1 frequent company name, 1 common forename and surname,
- 2 yes/no questions : 1 predominantly "yes" question, 1 predominantly "no" question,
- 3 application words,
- 1 keyword phrase using an embedded application word,
- 4 phonetically rich words,
- 9 phonetically rich sentences.

The following age distribution has been obtained: 241 speakers are below 16 years old, 1,404 speakers are between 16 and 30, 1,532 speakers are between 31 and 45, 711 speakers are between 46 and 60, and 139 speakers are over 60.

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

	ELRA Members	Non Members
Price for research use	28,000 Euro	48,000 Euro
Price for commercial use	40,000 Euro	56,000 Euro

ELRA-S0090 Polish SpeechDat(E) Database

The Polish SpeechDat(E) Database comprises 1000 Polish speakers (488 males, 512 females) recorded over the Polish fixed telephone network. The database was collected at the Wroclaw University of Technology (Poland). This database is partitioned into 5 CDs, each of which comprises 200 speakers sessions. The speech databases made within the SpeechDat(E) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat(E) format and content specifications.

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

Corpus contents:
- 6 application words;
- 1 sequence of 10 isolated digits;
- 4 connected digits: 1 sheet number (5 digits), 1 telephone number (8-11 digits), 1 credit card number (15-16 digits), 1 PIN code (6 digits);
- 3 dates: 1 spontaneous date (birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 spotting phrase using an application word (embedded);
- 1 isolated digit;
- 3 spelled-out words (letter sequences): 1 spelling of surname, 1 spelling of directory assistance city name, 1 real/artificial name for coverage;
- 2 currency money amounts: 1 Polish money amount, 1 International money amount (USD, EURO);
- 1 natural number;
- 6 directory assistance names: 1 surname (out of 500), 1 city of birth / growing up (spontaneous), 1 most frequent city (out of 500), 1 most frequent company/agency (out of 500), 1 "forename surname" (set of 150), 1 "surname" (set of 150);
- 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question;
- 12 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

The following age distribution has been obtained: 9 speakers are below 16 years old, 428 speakers are between 16 and 30, 291 speakers are between 31 and 45, 254 speakers are between 46 and 60, and 18 speakers are over 60.

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

The Czech, Russian and Slovak SpeechDat(E) databases will be available soon.

Price for research use	12,500 Euro
Price for commercial use	16,000 Euro

ELRA-S0091 Pronunciation lexicon of British place names, surnames and first names

The *Pronunciation lexicon of British place names, surnames and first names* was produced by the University of Poitiers (France) within the European Commission funded project LRsP&P (Language Resources Production & Packaging - LE4-8335). This lexicon is an SGML-encoded database of British proper names. All entries belong to one or several of the following categories: place-names (a quasi-exhaustive list of toponyms from England, Scotland and Wales), and surnames or first names (a selection of names based on an extensive survey of bibliographic sources in the field of British onomastics combined with lists compiled by the author of this lexicon).

The database is composed of 160,000 entries, breaking down as follows:

	Number of entries	**Number of transcriptions**
Place-names		
England	31,657	45,380
Wales	5,086	10,000
Scotland	15,406	20,444
Total 1	52,149	75,824
Surnames	92,253	115,918
First names	15,598	36,732
Total 2	107,851	152,650
Total 1+2	160,000	228,474

	ELRA Members
Price for research use	5,000 Euro
Price for commercial use	25,000 Euro

	Non Members
Price for research use	15,000 Euro
Price for commercial use	40,000 Euro

All phonemic transcriptions in the database are based on the SAMPA phonetic alphabet.

ELRA-M0025 Bilingual English-Russian Russian-English Dictionaries

The Bilingual English-Russian Russian-English Dictionaries were produced by the SCIPER company within the European Commission funded project LRsP&P (Language Resources Production & Packaging - LE4-8335).

In this work were used linguistic resources produced originally in Russia. It is well-known that during the Soviet period, a number of linguistic resources of very high quality have been developed in Russia. Among those are dictionaries and especially bilingual dictionaries which generally have much more entries than those found in Western countries.

Bilingual language resources produced within the above-mentioned LRsP&P project contain, in total, more than 350,000 pairs of words (in tabular form). In XML format which corresponds to the DTD, the dictionaries have the following volumes:

1) Russian-English dictionary - more than 130 000 entries

2) English-Russian dictionary - more than 95 000 entries

In this format, a dictionary entry corresponds to more than one pair of words because it may contain several semantically equal translations in target language.

Each dictionary entry contains the following information: source word (lemma); part of speech of source word; target word(s) (lemma(s)), grouped by same meaning; part of speech of target word(s); domain(s);

Both dictionaries contain as "source words" only significant parts of speech: nouns, adjectives, verbs, adverbs and nominals. Stop-words (prepositions, articles, pronouns, etc.) have not been included into the dictionaries because of the intended use in multilingual search and cross-lingual interrogation.

Both dictionaries are presented as XML files, with the same DTD. These files are coded in UNICODE - UTF-8.

The dictionaries are consistent, i.e. each of them presents the inverted version of the second one. This feature proves to be very useful for aligners, multilingual search engines, etc.

The list of domains may be found in DTD. It contains more than 100 domain names.

	ELRA Members	Non Members
Price for research use	2,000 Euro	4,000 Euro
Price for commercial use	12,000 Euro	16,000 Euro

ELRA-S0093 IBNC - An Italian Broadcast News Corpus

The Italian Broadcast News Corpus (IBNC) was produced by the ITC-IRST (Italy) within the European Commission funded project LRsP&P (Language Resources Production & Packaging - LE4-8335).

RAI, the major Italian broadcast company, supplied studio quality recordings of radio news programs sampled from its internal digital archive. The collection consists of 150 programs, for a total time of about 30 hours, issued in 36 different days, between 1992 and 1999.

Recordings were supplied by RAI on Digital Audio Tapes (DAT), with 44kHz sampling rate and 16 bit resolution. Each DAT was manually processed to transfer each single program issue into a single file. During this operation, the signal was down-sampled to 16kHz with a resolution of 16 bits, and encoded into the NIST Sphere PCM format.

Speech recordings present variations of topic, speaker, acoustic channel, speaking mode, etc. The corpus has been segmented, labelled and transcribed manually using the tool developed by DGA (Délégation Générale pour l'Armement, France) and LDC (Linguistic Data Consortium, USA), called "Transcriber", with conventions similar to those adopted by LDC for the DARPA HUB-4 corpora.

The transcription text consists of mixed-case ASCII characters of the ISO-8859-1 extended set.

A validation work was carried out by an external validator. It consisted of checking audio files, documentation and transcriptions.

	ELRA Members	Non Members
Price for research use by an academic organisation	5,000 Euro	8,000 Euro
Price for research use by a commercial organisation	15,000 Euro	25,000 Euro

ELRA-W0025 “Scientific” corpus of modern French

This “Scientific” corpus of modern French was produced by the University of Nantes (France) within the European Commission funded project LRsP&P (Language Resources Production & Packaging - LE4-8335).

The corpus contains all articles published in La Recherche magazine in 1998, including issues 305 (January) to 315 (December), which amounts to 447,244 tokens and 30,238 types. It is aimed to be used within text analysis and related applications.

The texts, provided in XML (Extended Markup Language) format, have been marked-up into the SGML standard (Standard Generalized Markup Language). XML contained a structure where only the constituant parts of the text were coded (title, body, etc.), whereas SGML marking up , richer, goes up to the word level, including the grammatical category and the canonical form for each word. The annotation work is conformant with the TEI (Text Encoding Initiative) international project's guidelines.

Raw data (XML):

	ELRA Members	Non Members
Price for research use	240 Euro	310 Euro
Price for commercial use	1,200 Euro	1,500 Euro

Complete version (XML + SGML):

	ELRA Members	Non Members
Price for research use	400 Euro	500 Euro
Price for commercial use	3,000 Euro	5,000 Euro

ELRA-S0088 Twin database - TWINDB1

The Twin database named TWINDB1 includes recordings of 45 French speakers, consisting of 9 pairs of identical twins (8 males and 10 females) with similar voices, and 27 other speakers (13 males and 14 females) including 4 none-twin siblings. Each twin or sibling spoke for a total of 24 to 30 minutes in three sessions conducted with at least one week interval between sessions.

In each session subjects were asked to read three different texts of one page. These texts consist of one paragraph of about 10 lines extracted from the French journal SVM Mac July 1994, and some short phrases, digits, credit card numbers, etc. extracted from the Polyphone Swiss-French database corpus (ELRA-S0030). The speakers called from their office or from their home. Subjects were recorded over the telephone using an OROS AU32 PC-board at 16 bits linear form, 8KHz sampling frequency.

	ELRA Members	Non Members
Price for research use	200 Euro	400 Euro
Price for commercial use	400 Euro	800 Euro

Up-date on Language Resources from the ELRA Catalogue

ELRA-S0034 Verbmobil

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

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

VM CD 33.1 - VM33.1 (BAS edition)

Verbmobil II - Japanese, 25 spontaneous dialogues (25 close mic, 0 room mic, 0 phone line (GSM) recordings), 1050 turns, transliteration (Verbmobil II Format)

VM CD 34.1 - VM34.1 (BAS edition)

Verbmobil II - Japanese, 28 spontaneous dialogues (28 close mic, 0 room mic, 0 phone line (GSM) recordings), 1437 turns, transliteration (Verbmobil II Format)

VM CD 35.1 - VM35.1 (BAS edition)

Verbmobil II - Japanese, 27 spontaneous dialogues (27 close mic, 0 room mic, 0 phone line (GSM) recordings), 1645 turns, transliteration (Verbmobil II Format)

VM CD 38.1 - VM38.1 (BAS edition)

Verbmobil II - German, 33 spontaneous dialogues (33 close mic, 0 room mic, 28 phone line (GSM) recordings), 3483 turns, transliteration (Verbmobil II Format)

VM CD 39.1 - VM39.1 (BAS edition)

Verbmobil II - German, 28 spontaneous dialogues (28 close mic, 0 room mic, 20 phone line (GSM) recordings), 2475 turns, transliteration (Verbmobil II Format)

VM CD 29.1 - VM29.1 (BAS edition)

Verbmobil II - German, 25 spontaneous dialogues (25 close mic, 0 room mic, 20 phone line (GSM) recordings), 1870 turns, transliteration (Verbmobil II Format)

VM CD 42.1 - VM42.1 (BAS edition)

Verbmobil II - American English, 20 spontaneous dialogues (20 close mic, 0 room mic, 0 phone line (GSM) recordings), 1874 turns, transliteration (Verbmobil II Format)

VM CD 43.1 - VM43.1 (BAS edition)

Verbmobil II - American English, 11 spontaneous dialogues (11 close mic, 0 room mic, 0 phone line (GSM) recordings), 633 turns, transliteration (Verbmobil II Format)

Price for ELRA members	127.82 Euro	Price for non members	255.65 Euro