The ELRA Newsletter

December 1997

Vol.2 n.4

Table of contents

Letter from the President and the CEO *page 2*

ELRA Board Profiles *page 3*

ELRA Annual General Meeting, La Villette, Paris 28 November 1997 *page 4*

Lexicon Validation
Nancy Underwood *page 4*

Techniques for Evaluation of Language Corpora: A Report from the Front, *Lou Burnard and Tony McEnery* *page 5*

Probabilistic analysis of pronunciation with MAUS
Florian Schiel *page 6*

Integrating Arabic into a Western MT System
Catherine Pease *page 9*

LISA Workgroup on Tools Benchmarking
Leon Rubinstein *page 11*

"ELSNET in Wonderland" *page 11*

New resources *page 12*

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:
Deborah Fry

Layout:
Rébecca Jaffrain

Contributors:
Lou Burnard
Tony McEnery
Catherine Pease
Leon Rubinstein
Florian Schiel
Nancy Underwood

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain

87, avenue d'Italie
75013 Paris
Phone: (33) 1 45 86 53 00
Fax: (33) 1 45 86 44 88
E-mail:
elra@calvanet.calvacom.fr
WWW:
http://www.icp.grenet.fr/
ELRA/home.html

Dear ELRA Members,

At the end of ELRA's second full year of operation, it is time to take stock of the progress we have made in 1997 and to give a preview of activities for the coming year. At the Annual General Meeting held at the end of November we were able to document a wide range of activities. 66 Speech, 126 Written, and 361 Terminology resources are now available from ELRA, and the number of agreements with resource providers is increasing. The 1997 management and financial reports and the budget for 1998 were all approved by the General Assembly, as was the revised membership fee structure, giving us a solid basis on which to work in 1998. More detailed information on the AGM is given in this Newsletter, and the full minutes have been sent to all members.

The Annual General Meeting also saw the election of a new Board for the next two years. We would like to take this opportunity to thank the retiring Board members, Robin Bonthrone, Lou Boves, Guiseppe Castagneri, and Christian Galinski for their valuable contribution to the work of the Association in its decisive early stages. At the same time, we would like to welcome the new Board members: Daniel Tapias from Telefónica I&D, Henk van den Heuvel from SPEX, and Volker Steinbiss from Philips GmbH Forschungslaboratorien. For those members who did not attend the AGM, profiles of the new Board members are given in this issue.

Turning to ELRA's day-to-day business, we are happy to report that a number of new resources have been acquired, including PolyVar and the SpeechDat speaker verification database from IDIAP, the Bilingual Collocational dictionary acquired from Horst Bogatz, the Karl-May-Korpus made available by Karlheinz Evert, and new corpora from the Verbmobil spoken dialogue collections. Detailed descriptions of these resources are available further on in this Newsletter and on our Web site. A number of other contracts are in the pipeline, and we would like to repeat our offer to members and others to distribute their resources. For more details, please get in touch with the ELDA office.

The current issue of the Newsletter also carries a number of interesting articles, including one by Catherine Pease on issues in Arabic machine translation, and by Florian Schiel on the probabilistic analysis of pronunciation with MAUS. Further there is a report from Leon Rubinstein on the new LISA Workgroup on Tools Benchmarking and summaries on the work for ELRA on validation manuals from Nancy Underwood and Tony McEnery & Lou Burnard.

One of ELRA's key activities in 1998 will be the Conference on Language Resources and Evaluation, to be held in May in Granada, Spain. Interest in the Conference has been considerable; all in all, proposals for 250 papers and 12 workshops were submitted. The preliminary Conference programme will be decided on 3rd of February and published on our Web site and in the newsletter.

In conclusion, all that remains for us to do on behalf of the ELRA Board and the ELDA staff is to wish all our members and partners a merry Christmas and a happy New Year. We look forward to working with you in 1998!

Antonio Zampolli, President

Khalid Choukri, CEO

ELRA Board Profiles

Henk van den Heuvel

Henk van den Heuvel was born in Zeist, The Netherlands, in 1963. He studied German language and literature at the University of Utrecht, with a specialisation in phonetics.

In 1988 he joined the Department of Language & Speech at the University of Nijmegen, where he worked in the field of computer-aided instruction for students in phonetics. He stayed in the same department while writing his PhD thesis entitled "Speaker variability in acoustic properties of Dutch phoneme realisations", which he defended in February 1996.

Henk van den Heuvel also worked on the European ONOMASTICA project (LRE-61004) during the years 1994-1995 (developing pronunciation lexicons for Dutch TTS systems). During 1995, he was involved in the MLAP MAIS project on speech recognition in automatic inquiry systems, while during 1995-1996 he worked for SPEX in the SpeechDat(M) project on database validation. He is Workpackage manager for the SpeechDat(II) project, again with respect to database validation. He is also working on the improvement of automatic speech recognition systems.

He sees his contribution to ELRA being mainly in the field of the validation/quality control of speech databases.

Volker Steinbiss

Born in Rheydt, Germany, in 1957, Volker Steinbiss studied mathematics in Goettingen, Germany, and Nice, France, specialising in complex analysis and receiving Diplom-Mathematiker and Dr. rer. nat. degrees in 1983 and 1985 respectively.

Starting in 1986, he worked on automatic speech recognition at the Philips Research Laboratories in Hamburg and Aachen, where his primary interest has been in search techniques. He was in charge of SPICOS, a joint project between Siemens, Philips, and the Institute of Perception Research in Eindhoven that led to the first German 1,000-word continuous speech understanding system, and of the Philips internal large-vocabulary speech recognition project, which provides the technology for automatic transcription of naturally spoken dictation (65,000 words continuous speech). The head of the Aachen-based Philips speech recognition research group since 1994, he is responsible for the definition and execution of Philips' scientific program in speech recognition and understanding, and for transfer to its commercial outlets.

His public engagement is reflected in his membership of a number of organisations (DMV, ESCA, ELRA, IEEE, LDC, and GI) as well as in his participation in various committees, steering committees and Boards (ITG, Verbmobil, ELRA, and DAGA).

Daniel Tapias

Daniel Tapias obtained a degree in Telecommunications Engineering with a specialisation in "Communications & Transmission" (Comunicación-Transmisión) from the Universidad Politécnica de Madrid in 1987. He later joined Page Iberica S.A., where he worked as a software engineer before becoming a research and development engineer in the Speech Processing Group in 1988. In this position he researched robust isolated speech recognition in automotive environments. He also participated in the ESPRIT-II A.R.S. project (Adverse-environment Recognition of Speech).

Since 1991, he has been working in the Speech Technology Division of Telefónica Investigación y Desarrollo, where he is currently the technical manager of the Speech Recognition Group.

Daniel Tapias' areas of research are speaker-independent automatic speech recognition through telephone and GSM channels, speaker adaptation, noise and channel compensation, and conversational systems. He is also involved in the design of evaluation methodologies and in speech database design, collection and labelling. A visiting scientist at both Bell Labs (1991) and Carnegie Mellon University (1995), he focused there on speech-to-speech translation and large-vocabulary continuous speech recognition respectively.

The author or co-author of more than 20 papers, Daniel Tapias has participated in several University conferences and seminars. Together with the Speech Technology Division, he was awarded the first AHCIET '92 prize for innovation in the field of telecommunications, as well as the Actualidad Electronica '93 award for the SAITMAP speech technology-based service.

The ELRA Board 1998-1999

President:
ANTONIO ZAMPOLLI

Vice-presidents:
NORBERT KALFON
JOSEPH MARIANI
ANGEL MARTIN-MUNICIO

Treasurer:
HARALD HÖGE

Secretary:
BENTE MAEGAARD

Members:
GEORGES CARAYANNIS
HENK VAN DEN HEUVEL
THOMAS SCHNEIDER
VOLKER STEINBISS
DANIEL TAPIAS

ELRA Annual General Meeting, La Villette, Paris 28 November 1997

The third ELRA AGM started with lunch organised by ELDA staff member Rébecca Jaffrain. Following this, there was a welcome from the ELRA President and meeting chairman, Antonio Zampolli, followed by the examination of the proxies (a total of seven) and the approval of the agenda and the 1996 AGM minutes. The CEO, Khalid Choukri, then presented the management report for the period between October 1996 and September 1997, which had also been mailed to all members. In his presentation, he gave an update on the ELDA staff as well as on ELRA membership statistics for 1995-1997 (which show an equal balance between the three colleges). Touching on the 1997 special membership offer, he urged members who wanted to take advantage of the offer to send back their consent forms. 66 Speech, 126 Written, and 361 Terminology resources are now available from ELRA, with most resources distributed coming from the Speech college. The number of agreements signed with resource providers has increased.

Khalid Choukri also said that the validation manuals would soon be available, before going on to present the methodology of and initial results from the marketing survey. He encouraged those members who had not yet filled in their questionnaires to do so. Marketing would be the most important task performed by ELRA in the coming months, and would be co-ordinated by ELDA's new Marketing Assistant, Malin Nilsson. The CEO then went on to give a status report on the LREC conference, emphasising that one person from each ELRA member organisation was entitled to attend the conference free of charge. Turning to the Association's detailed plans for 1997/1998, Khalid Choukri said that these would concentrate on increasing the number of members, improving sales, organising the LREC, and distributing the validation manuals.

Following this, the financial report and audited accounts were presented. In a brief opening statement, Antonio Zampolli commented on the Association's cash flow problems, which had been caused by late payment of moneys due from the European Commission, and stated that these would be solved. Khalid Choukri then presented the income and expenditure statements for the period under review, while the Treasurer, Thomas Schneider, stated that the accounts had been audited by an external company and had been found to be in order. After the approval of the financial report, the 1997-1998 budget was then presented by the CEO and approved. A proposal for new membership fees based on the type and size of organisation concerned (see below) was presented by the Secretary and approved by the General Assembly. Renewal notices will be sent to all members in January.

After this, the nominations of candidates for the Board which had been received were approved, and the elections to the Board were held. Since only 11 nominations had been received for a twelve-person Board, the missing member would be nominated as foreseen in the statutes. The CEO and President then thanked the former Board members, Robin Bonthrone, Louis Boves, Giuseppe Castagneri and Christian Galinski for their work and presented them with gifts. Closing the meeting, Antonio Zampolli thanked everyone for attending.

New membership fees for 1998

Non-profit making organisations:	750 ECU
European SMEs < 50 employees:	1000 ECU
European profit making organisations ≥ 50 employees:	1500 ECU
Non-European profit-making organisations:	5000 ECU

Lexicon Validation

Nancy Underwood

As part of its contract with the European Commission, ELRA has to produce validation manuals for resources in each of the colleges (Speech, Text, Terminology). The following article gives an overview of the draft manuals on lexicon validation, produced by Center for Sprogteknologi.

The work on lexicon validation performed at Center for Sprogteknologi has resulted in two reports: a draft manual for the validation of lexica and a draft proposal for a standard for the creation of lexica (the latter being based on EAGLES 1996a, 1996b). The draft validation manual includes a step-by-step guide to lexicon validation, followed by a more discursive section discussing each of the various steps and the characteristics which are to be checked. The validation process relies heavily on good documentation, and so a section is devoted to what the latter should contain. Finally, a validation report template is included for validators to complete.

Lexicon validation has two main aspects: content validation, which is concerned with the linguistic soundness of the coding in a lexicon, and formal validation. Formal validation itself is further divided into technical validation and conformity with specifications. Technical validation will be first carried out by, or under the supervision of, ELDA before the lexicon is passed on to expert validation sites to complete the validation. Expert validation sites will be chosen for their expertise in the lexicography of the language(s) treated by the lexicon in question.

Technical Validation

Perhaps the most important task involved in technical validation is parsing the lexicon to check its syntactic consistency, in order to ensure that it is in fact usable in an NLP system. In addition, a number of technical characteristics of lexica have been identified, such as the medium on which it is delivered, the character set used, the number and type of entries and the format, all of which must be checked by the validator. For some of these characteristics certain requirements have been defined (e.g. the preferred format is SGML), whilst others (e.g. the number of entries) are quite open.

Conformity with Specifications

The next stage in the process is to validate the lexicon's conformity with its specifications, that is, to check that the lexicon contains all the legal features specified for it, and only these. Such specifications may either be the producer's own or an external standard. In fact, at the beginning of the project we had provisionally defined this stage as checking conformity with "standards". However, there is a potentially very large number of different types of lexica, based on a variety of linguistic formalisms, which could currently be distributed by ELDA. As a result, it became clear that there could not be one single standard against which all lexica should be validated, and different standards for the many different types of lexica do not exist. Whilst the use of standards such as those being developed under EAGLES is certainly to be encouraged, at this stage it would not be reasonable to reject an otherwise acceptable lexicon because it does not conform to such an external standard.

Content Validation

In the final stage, content validation, the aim is to assess how the specific linguistic features are applied in the lexicon and how far the correct values are assigned in the entries. Because different languages typically pose their own specific problems in constructing a lexicon, the manual does not give a definitive list of all the features to be checked in validating the content of a lexicon. Rather, it provides an overall methodology and guidelines for the validator in selecting samples to be checked and in designing a validation which is pertinent both to the language in question and the intended coverage and purpose of the lexicon. In developing a general sampling technique, the manual takes a somewhat pragmatic approach, taking into account both the need to ensure representativity and the time and costs involved. It also includes a number of indicative examples, from different languages, of the type of phenomena which a validator may need to check.

Feedback

During the development of the manual we received invaluable feedback from members of the ELRA panel for Validation of Written Resources. The current manual has the status of a draft and once the validation procedure has been tested on specific lexica, we look forward to receiving feedback on all aspects of the validation procedure and manual. In particular, with respect to the language-specific aspects of content validation, it is hoped that the criteria developed by expert validators for a particular language can be provided as feedback to the manual and possibly incorporated as appendices. Such appendices could then serve as an aid to validators in designing new validations for other lexica, especially those treating closely related languages.

The two reports mentioned in this article are available from ELDA, free of charge:

Underwood N & C Navarretta, "A Draft Manual for the Validation of Lexica. Final Report". June, 1997.

Underwood N & C Navarretta, "Towards a Standard for the Creation of Lexica". June, 1997.

References

EAGLES (1996a) "Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages". EAGLES Document EAG-LSG/IR-T4.6/CSG-T3.2

EAGLES (1996b) "Subcategorisation Standards, Report of the EAGLES Lexicon/Syntax Group. Sharp Laboratories of Europe, Oxford Science Park, Oxford, UK.

For more information, please contact:
Nancy Underwood
CST (Center for Sprogteknologi)
Njalsgade, 80
DK 2300 Copenhagen S - Denmark
Tel: +45 35 32 90 90 - Fax: +45 35 32 90 89

Techniques for Evaluation of Language Corpora: A Report from the Front, *Lou Burnard and Tony McEnery*

This brief report describes the work we are currently undertaking for ELRA to develop guidelines for the validation of corpus encoding. Until recently, such guidelines would have been meaningless, since almost every new corpus developed used a new encoding scheme. Today, however, with corpus encoding slowly converging on the use of SGML and the availability of detailed recommendations such as those of TEI and EAGLES, the task is not merely possible, but also necessary.

The necessity for such guidelines is best understood if we take a look at what has happened in other areas where products with a wide market have been developed. For example, the need for validation with respect to software, or other consumer products with defined outputs and defined goals, is relatively uncontentious, and is indeed the subject of important ongoing work in standardisation (e.g. the EAGLES extensions to ISO 9126). However, the development and application of encoding standards for language corpora seem to be at an earlier stage of development. Although the applicability of a corpus resource is likely to be far greater than the uses originally envisaged for it, and indeed may often be unpredictable, corpus developers are only slowly beginning to see how this unpredictability makes the need for agreed and well-defined practices in encoding more urgent. For the producer of a corpus, validation may simply be a form of quality control, akin to traditional proof-reading; while for the user of a corpus, validation should provide a rapid and explicit account of what a corpus contains, and hence its likely usefulness in a given task.

Our view is that validation procedures for language corpora should thus concern themselves chiefly with the relationship between what is actually present in a corpus, and what claims are made about it. The primary goal of such procedures should be to establish that a corpus is accurately and completely described by its associated documentation, and secondarily to assess whether the features present conform with reasonable user expectations, i.e. whether they are "fit for use".

With this in mind, we are working with a tripartite description of validation:

- internal validation: for example, whether the encoding scheme used is selfconsistent, and conforms to a formal description;

- external validation: for example, whether the corpus conforms to some external standard such as the TEI/EAGLES recommendations. Note that conformance to such a standard may exist, even when no explicit claim that this is the case is made;
- user-oriented validation, or "fitness for use": for example, whether the features encoded form a reasonable subset of expressed user needs.

To identify those needs, we began our work by attempting to define an appropriate analytic framework for the validation of language corpora. Our first approach was to derive this empirically by examination of a large sample of existing corpora and their documentation, and by a user survey. Indeed, it is quite likely that you have seen and answered one of the Web questionnaires that we have distributed over the past few months. (If so, then may we thank you once again!) Our examination of the data allowed us to compare the features proposed by several related standards with actual user requirements as solicited by questionnaire, and actual user practice as demonstrated in a wide sample of corpora.

At the heart of our work is a cross-tabulation of three sets of features: those recommended by European standards (EAGLES in particular), those specified by users and, finally, the actual features found in the sample corpora. In doing this we are arriving at a view of where 'reality gaps' are emerging; for example, where current corpus encoding standards do not encode features felt to be essential by the user community, or where corpus builders are not encoding corpora in line with current standards and best practice.

Obviously, in order to carry out such a study we have had to select a range of corpora from the many that are currently available. Our sampling procedure aimed to maximise variability in such features as language, delicacy/method of mark-up, commercial interest, size, topic, etc. Attention was paid to a range of features, including technical characteristics (delivery media, physical encoding, etc.) and documentary characteristics (usability and accuracy of documentation), as well as inherent linguistic properties made explicit in the corpora.

Following this overall survey, we will proceed to define a staged series of validation procedures:

1. those concerned with detecting the presence of a given feature;
2. those concerned with identifying the syntactic correctness and consistency of the feature's representation;
3. those concerned with semantic correctness, i.e. whether the feature is correctly stated to be present in a given context.

Work on each of these is currently in progress. The degree to which these three levels of procedure may be automated is being assessed, and informal descriptions of the various tools available to perform such automatic validation at each level are being provided.

Our results so far seem to indicate that (with a few notable exceptions) current standards are somewhat in advance of current practice, and are also falling somewhat short of user expectations. This suggests to us that development of better and more exacting validation tools should be given a high priority.

Reports from this project will be made available via ELRA in the near future. In the mean time, we would be very interested in your comments or feedback: please contact either of us at the addresses given. Draft versions of the project reports are available at the following URL: http://users.ox.ac.uk/~lou/wip/ELRA/

For more information, please contact:

Lou Burnard, Oxford University Computing Service, United Kingdom
lou.burnard@oucs.ox.ac.uk

Tony McEnery, Lancaster University, United Kingdom
mcenery@comp.lancs.ac.uk

Probabilistic analysis of pronunciation with MAUS

Florian Schiel

This paper was first presented orally at the CWPU workshop in Berlin, Germany, 22-23 October 1997. It describes a method to automatically detect pronunciation variants in large speech corpora within the framework of the MAUS project. MAUS stands for 'Munich Automatic Segmentation System' a general purpose tool for automatically labelling and segmenting read or spontaneous German speech into phonetic/phonologic segments. MAUS output can, for example, be used to build probabilistic models of pronunciation of fluent German as reflected by the analysed corpus. These models can be the basis for phonetic investigations or can be incorporated into classic speech recognition algorithms.

Introduction to MAUS

The MAUS system was developed at the Bavarian Archive for Speech Signals (BAS) to facilitate the otherwise very time-consuming manual labelling and segmentation of speech corpora into phonetic units. Initially funded by the German government within the Verbmobil I project, MAUS has now been extended by BAS with the aim of automatically improving all BAS speech corpora by means of complete broad phonetic transcriptions and segmentations. The basic motivation for MAUS is the hypothesis that automatic speech recognition (ASR) of conversational speech, as well as high quality 'concept-to-speech' systems, will require huge amounts of carefully labelled and segmented speech data for successful progress.

Traditionally, a small part of a speech corpus is transcribed and segmented by hand to yield bootstrap data for ASR or basic units for concatenative speech synthesis (e.g. PSOLA). Examples of such corpora are the PhonDat I and II corpora (read speech) and the Verbmobil corpus (spontaneous speech). However, since this labelling and segmentation is done manually, it takes about 800 times as long as the utterance itself, e.g. to label and segment a 10-second utterance, a skilled phonetician spends about 2 hours and 13 minutes at the computer. It is clear that such an enormous effort makes it impossible to annotate large corpora such as the Verbmobil corpus, which contains over 33 hours of speech. On the other hand, such large databases are urgently needed for empirical investigations at the phonological and lexical level.

Input to the MAUS system takes the form of the digitised speech wave and any kind of orthographic representation that reflects the chain of words in the utterance. Optionally there may be markers for non-speech events as well, but this is not essential for MAUS. MAUS output consists of a sequence of phonetic/phonemic symbols from the extended German SAM Phonetic Alphabet, together with the time position within the corresponding speech signal.

Example:

Input:

Speech Wave + 'bis morgen wiederhoeren'

Output:

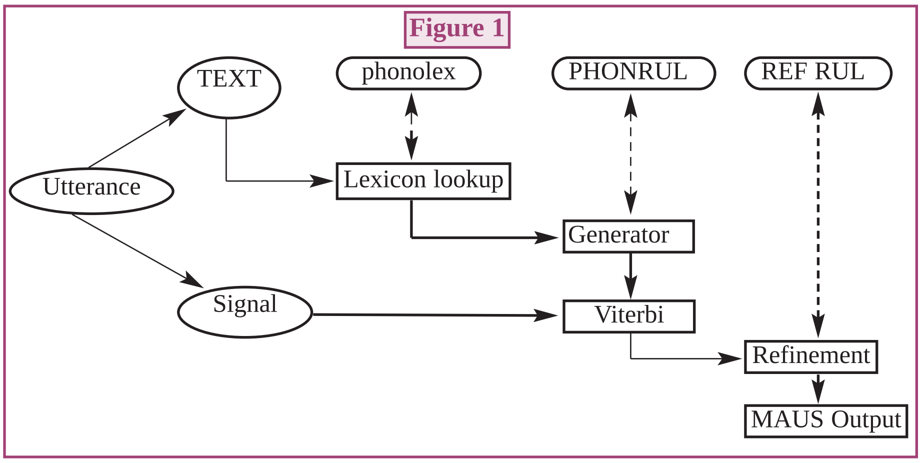
MAU: 0 479 -1 <p:>
MAU: 480 480 0 b
MAU: 961 478 0 I
MAU: 1440 1758 0 s
MAU: 2720 959 1 m
MAU: 3680 799 1 O
MAU: 4480 2399 1 6
MAU: 6880 2079 1 N
MAU: 8960 799 2 v
MAU: 9760 959 2 i:
MAU: 10720 479 2 d
MAU: 11200 2239 2 6
MAU: 13440 799 2 h
MAU: 14240 639 2 2:
MAU: 14880 1439 2 6
MAU: 16320 1599 2 n
MAU: 17920 1759 -1 <p:>

The output is written as a tier in the new BAS Partitur format. 'MAU:' is a label to identify the MAUS tier; the first integer gives the start of the segment in samples counted from the beginning of the utterance; the second integer shows the length of the segment in samples, while the third number is the word order and the final string is the labelling of the segment in extended German SAM-PA.

MAUS is a three-stage system (see Figure 1):

In the first step, the orthographic string of the utterance is looked up in a canonical pronunciation dictionary (e.g. PHONOLEX) and processed into a Markov chain (represented as a directed acyclic graph) containing all possible alternative pronunciations using either a set of data driven microrules or using the phonetic expert system PHONRUL.

Figure 1



A microrule set describes possible alterations of the canonical pronunciation within the context of ∓ 1 segments, together with the probability of such a variant. The microrules are automatically derived from manually segmented parts of the corpus. Hence, these rules are corpus-dependent and contain no a priori knowledge about German pronunciation. Depending on the pruning factor (observations are very rarely discarded) and the size of the manually segmented data, the microrule set consists of 500 to 2,000 rules. In this paper we use a set of approximately 1,200 rules derived from 72 manually segmented Verbmobil dialogues from the Kiel Corpus of Spontaneous Speech.

The expert system PHONRUL consists of a rule set of over 6,000 rules with unlimited context. The rules were compiled by an experienced phonetician on the basis of literature and generalised observations in manually transcribed data. There is no statistical information within this rule set; all rules are treated with equal probability. PHONRUL is therefore a generic model and should be considered independent of the analysed speech corpus.

The second stage of MAUS is a standard HMM Viterbi alignment in which the search space is constrained by the directed acyclic graph from the first stage (see Figure 2 for an example). Currently we use HTK 2.0 as the aligner with the following pre-processing: 12 MFCCs + log Energy, Delta, Delta-delta every 10 msecs. Models are left-to-right, 3 to 5 states and 5 mixtures per state. No tying of parameters was applied to keep the model as sharp as possible. The models were trained to manually segmented speech only (no embedded re-estimation).

Probabilistic pronunciation model

Aside from the many other uses of MAUS output, in this paper we will show how to derive a simple but effective probabilistic pronunciation model for ASR from the data. There are two obvious ways to use the MAUS results for this purpose:

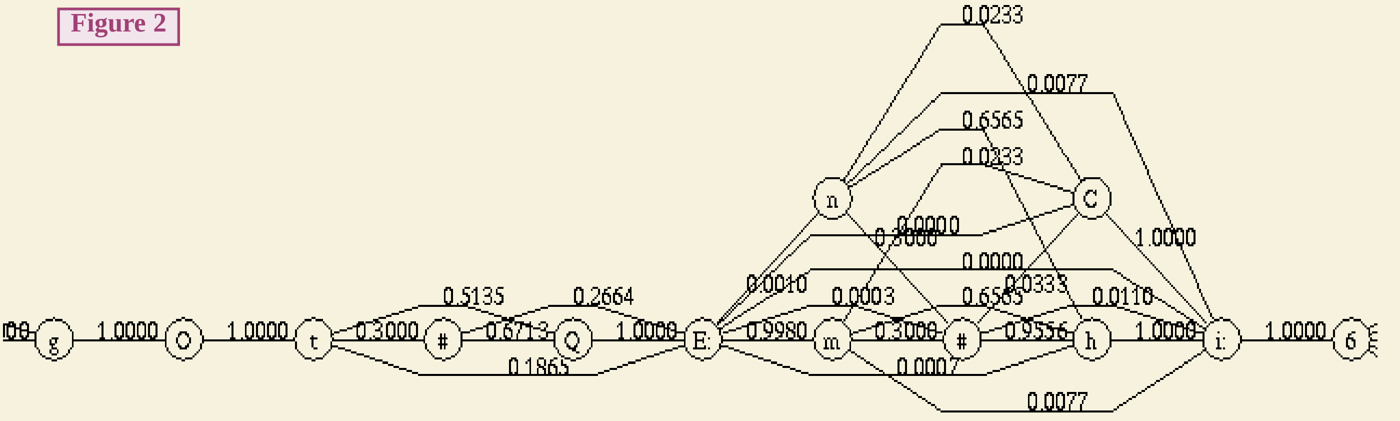
A) use direct statistics on the observed variants;

B) use generalised statistics in the form of microrules.

Direct Statistics

Since in the MAUS output each segment is assigned to a word reference level (Partitur Format) it is quite easy to derive all observed pronunciation variants from a corpus and collect them in a PHONOLEX style dictionary. The analysis of the training set of

Figure 2



the 1996 Verbmobil evaluation (volumes 1-5, 7, 12) led to a collection of approximately 230,000 observations.

Obviously many of the observations are not frequent enough for statistical parameterisation, which is why the baseline dictionary is pruned in the following way:

- Observations with a total count of less than N per lexical item are discarded.
- From the remaining observations for each lexical word, L, the a posteriori probabilities, P(V|L), that the variant V was observed are calculated. All variants that have less than M% of the total probability mass are discarded.
- The remaining variants are re-normalised to a total probability mass of 1.0.

Generalised statistics

The use of direct statistics has the disadvantage that most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation because of lack of data. An easy way to generalise to less frequent (or unseen) words is to use not the statistics relating to the variant itself, but the underlying rules applied during the MAUS segmentation process. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it is the number of times these rules are applied that counts.

Since there is no formal distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows:

A) Derive a set of statistical microrules from a subset of manually segmented data (see “Introduction to MAUS”) or use the rule set PHONRUL.

B) Apply this rule set to segment the training corpus and count all appliances of each rule forming the statistics of the recognition rule set.

Note that the recognition rule set might be a subset of the PHONRUL/microrule set, although this is very unlikely in the latter case.

This approach has the great advantage that the statistics are more compact (and therefore robust), are independent of the dictionary used for recognition (which will certainly contain words that were never seen in the training set) and generalise knowledge about pronunciation to unseen cases. However, the last point may be a source of uncertainty, since it cannot be foreseen whether the generalisation is valid in all cases in which the context matches. We cannot be sure that the context we are using is sufficient to justify the usage of a certain rule in all places where this context occurs.

Automatic Speech Recognition (ASR)

There have been several attempts to incorporate knowledge about pronunciation into standard methods for ASR. Most of them (with a few exceptions) did not yield any improvements. The argument was that the advantage of better modelling on the lexical level is offset by the fact that the search space and/or the dictionary ambivalence in-creases. However, most of the literature did not take reliable statistics into account (because they were simply not available) and used acoustic models that were trained using canonical pronunciations. Our hypothesis is that an increase in recognition performance can only be achieved if the following conditions are satisfied:

1. A reliable statistical model for pronunciation (which very likely will turn out to be adapted to the task) and

2. Acoustic models that match the modelling at the lexical level.

We are currently conducting several experiments on this basis with a standard HTK recogniser for the 1996 Verbmobil evaluation task. In this paper we will only report on preliminary results using the direct statistics approach.

A standard HTK 2.0 recogniser with the following properties was designed for the experiment:

The speech signal is mean subtracted, emphasised and pre-processed into 12 MFCCs + log Energy, Delta, Delta-delta every 10 msecs. Training and test sets are defined in the 1996 Verbmobil evaluation task ('Kuer', test corpus: 6,555 words). The canonical dictionary contains 840 different entries. The language model is a simple bigram calcu-lated exclusively from the training set. The acoustic models are monophone left-to-right HMMs with 3-5 states of 7 mixtures each without tying. We use 46 models from the extended German SAM-PA, including one model for silence and one model for non-speech events.

We trained and tested the recogniser with the same amount of data in two different fashions:

A) Baseline System

Standard bootstrapping to manually labelled data and iterative embedded re-estimation (segmental k-means) until the performance on the independent test set converged (note: performance in terms of word accuracy, defined by (number of words - inser-tions - replacements - deletions) / number of words). The re-estimation process utilised a canonical pronunciation dictionary with one pronunciation per lexical entry. The system was tested with the same canonical dictionary.

B) MAUS System

This system was bootstrapped to one third of the training corpus (approximately 10 hours of speech) using the MAUS segmentation and then iteratively re-estimated using the transcripts of the MAUS analysis instead of the canonical dictionary (note that the segmental information of the MAUS analysis was NOT used here). The system was tested with the probabilistic pronunciation model described in the section on direct statistics using the pruning parameters N=20 and M=0%.

Figure 3 shows the performance of both sys-

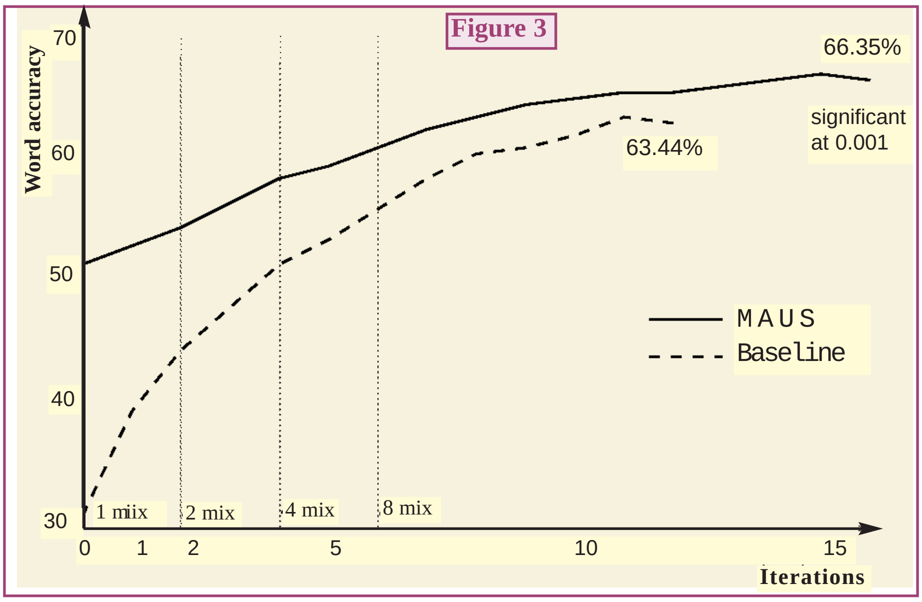


Figure 3

tems during the training process. Note that the MAUS system starts out with a much higher performance level because it was bootstrapped to 10 hours of MAUS data (compared to 1 hour 40 minutes of manually labelled data for the baseline system). After training, the MAUS system converges on a significantly higher performance level of 66.35%, compared to 63.44% for the baseline system.

Conclusions

The MAUS system can be used effectively to label and segment read and spontaneous speech corpora into broad phonetic alphabets completely automatically. This enables us for the first time to derive statistical models on different processing levels (acoustic, phonetic, lexical) on the basis of very large databases. We have shown that the usage of this data can significantly improve ASR for spontaneous speech.

The MAUS principle is not language-dependent (although the required resources are!). We therefore strongly encourage colleagues in other European countries to adopt the MAUS principle for their specific languages and to produce similar resources to those currently being produced at BAS for the German language. The first joint project (MIGHTY MAUS) for American English and Japanese is scheduled for 1998 together with the International Computer Science Institute (ICSI), Berkeley, California, and Sofia University, Tokyo.

For more information, please contact:
Florian Schiel
Bavarian Archive for Speech Signals
University of Munich, Germany
schiel@phonetik.uni-muenchen.de/Bas
http://www.phonetik.uni-muenchen.de/Bas

Integrating Arabic into a Western MT System

Catherine Pease

Experiments in the automatic translation of Arabic were carried out at the Institute for Applied Information Science, Saarbrücken, as part of the Aramed project, which was financed by the European Commission's INCO programme. The aim of this project was to develop a system which translates medical classifications (based on the SNOMED medical codes) from English into Arabic, and German and English medical texts into Arabic. The system consists of two main components: a transfer-, constraint- and unification-based machine translation system (CAT2), and an Arabic morphological generator (written at the Electronic Research Institute in Cairo).

Introducing Arabic into the CAT2 system and translating the (at least linguistically) Western-dominated field of medicine into Arabic both raised a number of interesting issues, such as to what extent the `universal' linguistic phenomena implemented in CAT2 were really universal, and whether Latin-based words or Arabic should be used for the translation of medical terminology (Latin medical terms are often simply transliterated for use in Arabic). However, the issue addressed in this article is that of lexicographical organisation, and the difficulties faced when trying to integrate Arabic into a framework written for Western European languages.

The CAT2 MT system was first developed in 1987 as a sideline to Eurotra. The system has two basic parts, the formalism and its implementation (the software) and the lexica, grammars and translation modules (the lingware). The basic architecture of the CAT2 system is a classic stratificational, transfer-based one, and uses tree structures at all levels. The approach followed in CAT2 the abstracting away from surface features and the reliance on semantic and pragmatic aspects in transfer (attained by the inclusion of general cognitive categories relating to time, space and cause in the Interface Structure representation) is located somewhere between a normal word-based transfer and an interlingua, and presupposes a fully competent language component which can relate the semantic and pragmatic content to its surface representations. The backbone of this implementation can be found in the organisation of lexical concepts: dictionary entries in CAT2 are lexeme-based, which means that a lexeme forms the key of a lexical entry, regardless of morphosyntactic information. This information is then encoded by describing different realisations of this lexeme: i.e. lemmas.

Figure 1 shows the CAT2 entry for the lexeme **sell** (VERB,VN_AGENT,VN_ING and VN_IRREG are macros containing grammatical and semantic information). The first lemma is the verb itself, the next two are nominal forms of the verb, all of which fill the zero argument slot (i.e. the process itself), and the last is the agentive derivation, which therefore fills the first argument slot of the subcategorisation frame.

Figure 1: The CAT entry for the lexeme sell

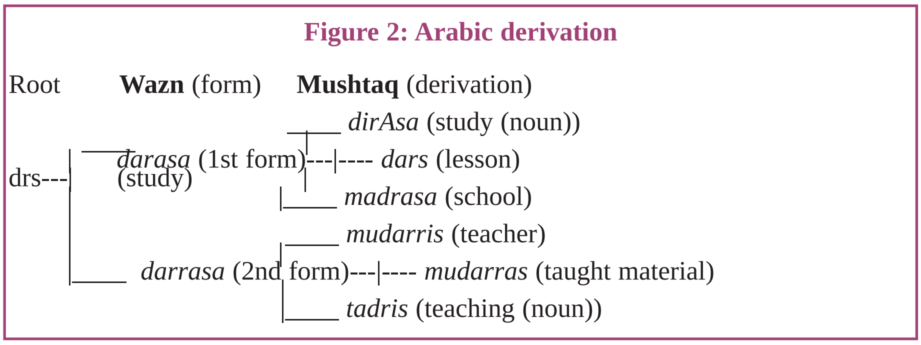
```
sell={lexeme=sell,head=({lemma=sell,VERB};{lemma=selling,VN_ING};
{lemma=sale,VN_IRREG};{lemma=seller,VN_AGENT}),
frame={arg1={role=agent},arg2={role=theme}, arg3={role=goal}}}.[ ].
```

Coding the various lemmas was relatively easy as differences between derivations of a lexeme in German, English and French (from here on the `CAT2 languages') are limited either to a variation in syntactic category of the lexeme (i.e. they denote the same concept as the lexeme from which they are derived, but may differ in category, and possibly also in aspect and modality), or else they denote an argument in the subcategorisation frame (one of the thematic roles) of the lexeme if it is predicative. The only condition for entering different lemmas (derivations of the same lexeme) in one entry is that they share the same subcategorisation frame.

By far the most interesting aspect within this experiment was the comparison of the lexical structures found in Arabic and in the Western languages for which the system was written. Arabic is morphologically very rich, and contains countless possibilities for expressing conceptual phenomena morphologically. Paradoxically, however, this very richness often defied attempts to exploit the resulting derivational richness in Arabic, which ideally should be advantageous in this form of lexicon-writing as it enables one stem and its derivations to be more quickly and consistently coded.

As shown in Figure 2 (next page), Arabic derivation functions structurally in two ways: the awzaan (plural of wazn), formed from the root and the mushtaqaat (plural of mushtaq), which are derivations from the awzaan. (Wazn is often translated as `form', which I shall use here interchangeably with wazn, but as mushtaq is translated as `derivation', I shall maintain the Arabic word mushtaq to avoid confusion with `derivation' in other languages, or as a general concept). However, Arabic derivation also takes place along another axis, which divides it into two categories: grammatical and semantic deri-

Figure 2: Arabic derivation



vation. This semantic derivation makes it difficult to establish what the `lexeme' is (the minimal unit of meaning of a word). Let us tentatively assume, for the sake of argument, that these two types coincide: the mushtaqaat represent grammatical derivation, and the awzaan semantic derivation. The mushtaqaat were easy to encode as this type of derivation is common in many languages, including the CAT2 languages, and can be encoded as shown in the CAT2 entry above.

The grammatical derivations which involve different mushtaqaat relating to one wazn present no real problem in CAT2 implementation, for the important thing is the inheritance of the subcategorisation frame (which allows "I like to read" to be translated as ahub al-qarAa (I like the reading), even though the derivational forms and syntactic categories of read/qarAa are different). In addition, as mentioned above, this phenomenon has already been implemented for other languages in the CAT2 system. However, we did discover grammatical derivations in Arabic for example locative derivations which were rarely found in the other languages, and adjustments had to be made in order to incorporate these.

The other derivational `direction' semantic derivation proved much more problematic, however. Up to now we have been assuming that the lexemes from which grammatical derivations are made (we have only considered mushtaqaat) are the various awzaan. So the mushtaq Akl (food) is derived from Akala (eat) (first form); the mushtaq astamaal (usage) is derived from AstAmala (use) (eighth form). If, however, we want to claim that the different awzaan are derivationally linked, we may need to take our derivation a step further. As the lexeme is the smallest unit of meaning in a word, we obviously cannot say that one wazn is derived from another whilst also claiming that they are both lexemes. This raises the question of whether it is the root which is the lexeme in which case the first form is strictly speaking also a derivation (a zero derivation) or whether the root is merely a string of letters used by different awzaan for creating meaning. The obvious answer is to look on the root itself as being our lexeme. Take for example the first and the fifth forms of the root khrj: kharaja, meaning to exit, and takharaga, meaning to graduate. These are obviously connected in meaning, for the fifth form literally means to exit from university (successfully). Because it adds another semantic feature to the first form, though, we cannot say in CAT2 that it is derived from either this first form or the root, as derivation in CAT2 only involves morphosyntactic changes. It would be necessary to establish a framework for semantic derivation first.

However, there are problems with this view. How do we treat two forms of a root which are not related in meaning (such as hadatha (happen), and haddatha (talk)). Or even two mushtaqaat of one form (such as shaariA (street) and mashrUA (project))? We could say that a root can represent more than one lexeme. We certainly allow a stem to be two different lexemes in other languages e.g. bank in English. However, if we end up having to claim that a root can represent many lexemes, it would seem more reasonable to say that the root itself is merely a string of letters, without meaning, from which different lexemes can be built. Another point is that some of the awzaan are related grammatically, rather than semantically. The second form, for example, is often a transitivisation of an intransitive first form of the verb (e.g. mAt (die) and mUt (kill)). The sixth form is often used to express reciprocality (e.g. fahama (understand), tafaahama (understand one another)). The seventh form also involves a grammatical derivation from the first form, as it indicates the passive i.e. it gives the verb a passive sense without it being grammatically passive (e.g. kasara (break), inkasara (be broken)).

If we maintain that the root is the lexeme (possibly more than one), in an attempt to link kharaga to takharaga, these links need to be formally described before being implemented in an automatic application. Are these links clear enough to be formally represented? Is it useful to define such relationships for translation?

It proved possible to encode in one entry (i.e. to treat as derivations of one lexeme) those derivations which involve linking one form to another (e.g. second form) by suppressing certain arguments in the frame for certain derivations. The first (agentive) argument is suppressed for the first form of mAt (die), for example, which is needed in the derivation mUt (kill - second form). Likewise, the passive form, the seventh form, can also be easily incorporated by setting the second argument (the `theme' in this case) to be the grammatical subject of the surface structure.

Without a framework for semantic derivation, however, it is not possible to relate takharaga to kharaga. What exactly is the change in meaning, and how can we represent it? One possibility would be to use semantic predicates i.e. to analyse lexemes componentially. This is achieved by reducing meaning to the smallest possible semantic units commonly known as `sense components' and then describing words in terms of these sense components. This has been implemented by Bonnie Dorr in the UNITRAN MT system for English, Spanish and German. Kharaga and takharaga would then be described in the following way:

kharaga: + Movement + Out (Building) + Sentient Subject

takharaga: + Movement + Out (Building, University) + Sentient (Human) Subject + Success

It would seem that this analysis is the best answer. This method of semantic representation is, however, not used in CAT2 at present (and would also be difficult to incorporate within the present implementation). What would be the advantage of these complex semantic descriptions? If carried out consequently, they would mean a much deeper analysis of the intricate links within Arabic morphology, which might result in better translations this would have to be tested. For the present, modest aims of the inclusion of an Arabic component in CAT2, it is easier just to write an additional transfer rule:

kharaja <=> exit
takharaja <=> graduate

It is, however, worth thinking along these lines for future NLP projects as this type of semantic representation certainly appears to suit the derivational patterns found in the Arabic language, and, what is more, may be the only way in which these patterns can be exploited in formal applications.

Catherine Pease
Institut fuer Angewandte Informationsforschung
an der Universität des Saarlandes
Martin-Luther-Strasse 14
D-66111 Saarbrücken - Germany
Tel: +49 681 3895126-Fax: +49 681 3895140
E-mail: cath@iai.uni-sb.de

LISA Workgroup on Tools Benchmarking

Leon Rubinstein

The most recent LISA (Localisation Industry Standards Association) Forum, held in Geneva on 4-5 December 1997, hosted a Workgroup on Tools Benchmarking. About 30 tools users, both localisation service providers and high-tech product developers, came to learn more and share their experiences in the field.

The Workgroup was organised by Mr. Leon Rubinstein from the global outsourcing company McQueen, who launched the ToBe (Tools Benchmarking) Special Interest Group (SIG) initiative at a previous LISA Forum in Washington DC this summer. At the end of the Workgroup session, a core team of participants set up a launch target of one month to meet and define the framework of the SIG.

The ToBe SIG has the following core goals:

- compilation of a list of tools for benchmarking (e.g. terminology management, terminology extraction, text alignment, TM, workflow, localisation project management, controlled language authoring, MT, etc.)
- development of user profiles for benchmarking
- development of operations profiles for benchmarking
- collection of existing valid information on tools comparison
- definition and contracting of independent evaluation of tools against pre-defined profiles and real-life scenarios
- initiation of an annual benchmarking review process, based on the evolving tools market.

It was already clear that members would be interested in defining specific real-life scenarios to be tested with various tools, in order to compare such aspects as functionality, performance, usability, and operational and technical complexity, etc.

The initial activity will be focused on translation memory/translator's workbench products and could grow to encompass other tools further down the road. Members also agreed that this activity has to be ongoing in order to provide a continuous view of this evolving market, utilise diverse real-life scenarios, make testing applicable to different operations set-ups, and better understand the different tools' behaviour as a function of a set of external parameters (to be defined by the members).

The SIG will consist of representatives of the user community, but will also co-operate closely with tools developers, both as experts and "watchdogs", who will provide self-interested quality control of the evaluation methods. The tests themselves will be performed by "independent" organisations (e.g. the academic community, consultants, etc.) and/or by member organisations themselves, depending on the specific project.

The SIG does not seek to re-invent anything and will therefore actively try to collect all valid information on tools comparison that already exists on the market.

The results will be published and distributed to SIG members and will probably also be re-sold as a LISA product (this point will be confirmed in the SIG's statutes).

For further information, please contact:
Leon Rubinstein
McQueen France - ZAC du Pont Blanc
26-28, rue Henri Becquerel
93275 Sevran Cedex
France
Tel. +33-1-49 36 53 23
Fax +33-1-49 36 53 33
E-mail: leon.rubinstein@mcqueen.com

“ELSNET in Wonderland”

How can we turn ELSNET into a showcase of Language and Speech technology?
March 25-27, 1998, Utrecht, the Netherlands

ELSNET members are invited to register for “ELSNET in Wonderland”, a two-day conference (lunch-to-lunch format) for the entire ELSNET community. The conference will consist of a mix of practical and theoretical discussions, plenary sessions, and small working group sessions.

Given the current state of Language and Speech technology: which available facilities could ELSNET offer in principle, for example via its web pages? and what are the main research problems to be addressed in order to facilitate and promote the implementation of L&S technology in the emerging Multilingual Information Society?

Conference results will include:

- a number of concrete project proposals (pilot studies), leading to the implementation of new L&S technologies on ELSNET’s web pages;
- the identification of significant research strands for the future (e.g. in the Commission’s Fifth Framework Programme);
- identification of commercial or research systems resulting from EC funded projects, suitable for inclusion in a permanent electronic exhibition with the look and feel of a real exhibition (in collaboration with Linglink).

Registration forms will be distributed via elsnet-list and via our WWW pages (http://www.elsnet.org/wonderland/form.html), or will be sent to you upon request.

ELSNET
Trans 10, 3512 JK Utrecht, The Netherlands
phone +31 30 253 6039, fax +31 30 253 6000
elsnet@let.ruu.nl - http://www.elsnet.org
Up-to-date information can be found at http://www.elsnet.org/wonderland.

New resources

ELRA-S0046 PolyVar

PolyVar is a speaker verification database comprising native and non-native speakers of French, mainly from Switzerland but also from other European countries. It consists of read and spontaneous speech recorded by 143 speakers (85 male and 58 female) amounting to 160 hours of speech. Each speaker recorded from 1 to 229 sessions, giving a total of 3,600 recorded sessions. The data are provided with orthographic annotation.

The number of calls per speaker is as follows: 13 speakers called 100 times; 9 speakers called from 51 to 100 times; 16 speakers called from 21 to 50 times; 3 speakers called from 11 to 20 times; 31 speakers called from 2 to 10 times; 71 speakers called only once

Each speaker uttered up to 53 different items per session, including: 3 sequences of digits (1 ID number, 1 credit card number and 1 sequence of 6 digits); 24 application words (17 words about tourism – Martigny); 10 read sentences; 4 numbers (2 natural numbers, 2 amounts), 2 items with dates (1 read/1 spontaneous), 2 items with hours (1 read/1 spontaneous), 2 spelled words; 3 spontaneous answers (questions about their gender, native language and the weather); 1 comment; 1 telephone enquiry

File format:	8-bit a-law		
Standard in use:	NIST	Price for ELRA members:	Price for non members:
Sampling rate:	8 kHz	for research use: 1,000 ECU	for research use: 2,000 ECU
Medium:	8 CD-ROMs	for commercial use: 2,000 ECU	for commercial use: 4,000 ECU

ELRA-S0047 SpeechDat Speaker Verification database

This subset of PolyVar consists of 20 speakers which recorded 50 sessions. The format in use is a-law with SAM headers.

Medium: 3 CD-ROMs

Price for ELRA members:	Price for non members:
for research use: 750 ECU	for research use: 1500 ECU
for commercial use: 1500 ECU	for commercial use: 3000 ECU

ELRA-W0016 Karl-May-Korpus (KM corpus)

The "Karl-May-Korpus" is a monolingual German corpus, available in an SGML-tagged ASCII text format. It contains the works of the German author Karl May (1842-1912) and consists of around 1.6 million words (divided into 9 subcorpora of about 180,000 words each). The corpus was created between 1993 and 1997.
Each word form is tagged with a word class (1 out of 43 classes) and appropriate lemma.

File format:	Text	Price for ELRA members:	Price for non-members:
Standard in use:	SGML	for research use: 400 ECU	for research use: 800 ECU
Character set:	8-bit ASCII	for commercial use: 2,500 ECU	for commercial use: 3,500 ECU

ELRA-S0034 Verbmobil

This resource consists of spontaneous speech recorded in a dialogue task (appointment scheduling). The German corpus has a total of 13,910 utterances (turns). The BAS edition of the German part has been 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 the ELRA or BAS Web sites):
VM CD 13.0 - VM13.0 (original edition)
American/'Denglish'* - 90 speakers - 1,714 turns - 200 spontaneous dialogues.
VM CD 13.1 - VM13.1 (new edition)
American/'Denglish'* - 90 speakers - 1,714 turns - 200 spontaneous dialogues - transliteration.
VM CD 14.0 - VM14.0 (original edition)
97 speakers - 1,891 turns - 156 spontaneous dialogues - transliteration.
VM CD 14.1 - VM14.1 (new edition)
97 speakers - 1,891 turns - 156 spontaneous dialogues - transliteration - PhonDat 2 headers - Partitur Files**.
* 'Denglish': English spoken by Germans.
** Partitur files: files describing the different parts which constitute the corpus word order, phrase order, etc.