

Table of contents

Vol.2 n.3

Letter from the President and the CEO _____ page 2

First International Conference on Language Resources and Evaluation - Second call
Granada, Spain, 28-30 May 1998 _____ page 3

ELRA Distribution activities as of 30/09/97 _____ page 4

Defining a methodology for designing evaluations: Standards and Sharing
Maghi King _____ page 5

SALT Workshop on Evaluation in Speech and Language Technology
Robert Gaizauskas _____ page 8

ELRA Validation report _____ page 9

The Babel project - Speech databases from Central & Eastern Europe
Peter Roach _____ page 9

ELRA Market Segmentation Survey - update _____ page 10

Announcements _____ page 11

New resources _____ page 13

Membership application
ELRA 1997 Members' Special Offer _____ page 16

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:
Robert Gaizauskas
Maghi King
Peter Roach

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,

Much has happened at ELRA in the past few months. To start with, preparations for the First International Conference on Language Resources and Evaluation, which ELRA initiated, are now at an advanced stage. The Conference will be held in Granada, Spain on 28-30 May, 1998, and will be sponsored by DG XIII and the Fundación Banco Central-Hispano. In addition, it has gained the support of a large number of government agencies world-wide, as well as leading language industry associations in Europe and beyond. Further information, including submission and registration details, are provided in this issue.

In keeping with the theme of the Conference, we are also featuring a report by Robert Gaizauskas on the Workshop on Evaluation in Speech and Language Technology organised in the UK by SALT (the Speech and Language Technology Club), and an article by Maghi King on the approach to validation and evaluation in the EAGLES project. Other highlights include reports on the Babel project, on the validation work commissioned by ELRA, and on ELRA's distribution activities.

The new version of the ELRA Catalogue of Resources, which can be found on the ELRA Website, now provides over 500 offerings. Four new licenses signed during the September-October period a text corpus from "Le Monde", the Dutch lexical database CELEX, the POLYCOST speech database and the Onomastica Copernicus speech database are described in this issue. Along with these you will find the following new resources from our provider BAS (Bavarian Archive for Speech signals): the SPINA speech corpus (a set of words and utterances for robot commands), the set of pronunciation rules for German, PHONRUL 9.0, and new corpora from the Verbmobil spoken dialogue collections. Samples of these and other resources can be found on our Web Site.

Work on the validation manual packages has also continued, with the first reports on written resources (lexica and corpora) now being available. The work done in this area for speech was discussed at an international forum during Eurospeech (Cocosda meeting), while terminology, as agreed, will build on the validation manual to be produced by the Interval project when this becomes available.

To close, we would like to urge members to attend the 1997 Annual General Assembly on 28 November in La Villette in Paris. A letter with full details of the agenda and venue will be sent to you shortly. In addition to electing a new Board, the meeting will discuss past, present and future activities and new subscription structures. We would remind those members who have not yet renewed their subscriptions that payment is a prerequisite for voting, and would also like to mention that new members can still take advantage of the special introductory offer of two free resources. We look forward to welcoming you in Paris.

With best wishes

Antonio Zampolli, President

Khalid Choukri, CEO

PS: Starting with the next issue, we will be running a series of member profiles in the Newsletter. If you would like to be featured, please send an outline of your organization's activities and any marketing or similar materials to ELRA/ELDA. We shall then get in touch with you.

First International Conference on Language Resources and Evaluation - *Second call*

Granada, Spain, 28-30 May 1998

The First International Conference on Language Resources and Evaluation was initiated by ELRA and is being organised in co-operation with many other associations and consortia. These include COCOSDA, EAFT, EAGLES, EDR, ELSNET, ESCA, FRANCIL, LDC, PAROLE, and TELRI, etc., as well as major national and international organisations such as the European Commission (DG XIII), ARPA, NSF, and Chinese, Japanese and Korean programs. Co-sponsorship and support from other institutions is currently being sought.

The Conference will be hosted by the University of Granada's Departamento de Traducción e Interpretación and Departamento de Electrónica y Tecnología de Computadores, with the support of the European Commission's DG XIII and Fundación Banco Central-Hispano.

Conference aims

The pervasive character of language technologies in the information society and their relevance to practically all fields of information and communication technologies (ICT) is now widely recognised.

Two issues are currently considered particularly relevant to international co-operation: the availability of language resources and the methods for evaluating resources, technologies and products.

The term "language resources" (LRs) refers to sets of language data and descriptions in machine-readable form which are used specifically for building, improving, or evaluating natural language and speech algorithms or systems, and in general as core resources for the software localisation and language services industries, for language studies, electronic publishing, and international transactions, as well as by subject area specialists and end users.

Examples of linguistic resources are written and spoken corpora, computational lexicons, grammars, terminology databases, and basic software tools for the acquisition, preparation, collection, management, customisation and use of these and other resources.

The relevance of evaluation in language engineering is becoming increasingly clear. It involves assessing the state of the art for a given technology; measuring the progress achieved within a program; comparing different approaches to a given problem and choosing the best solution; knowing the advantages and drawbacks; assessing the availability of technologies for a given application and, finally, product benchmarking. Evaluation accompanies research and development in human language technologies and has driven

important advances recently in various aspects of both written and spoken language processing. Although the evaluation paradigm has been studied and used in large national and international programs, including the US ARPA HLT program, EU Language Engineering projects, the Francophone Aupelf-Uref program and others (particularly LISA and LRC within the localisation industry), it is still subject to substantial unresolved basic research problems.

The aim of this Conference is to provide an overview of the state of the art, discuss problems and opportunities, exchange information on ongoing and planned activities, present language resources and their applications, discuss evaluation methodologies and demonstrate evaluation tools, and explore possibilities and promote initiatives for international co-operation.

Conference Topics

The following list gives some examples of topics which could be addressed:

- Issues in the design, construction and use of LRs (theoretical and best practice).
- Guidelines, standards, specifications and models for LRs.
- Organisational issues in the construction, distribution and use of LRs. Legal aspects and problems in the construction, access and use of LRs.
- Methods, tools and procedures for the acquisition, creation, management, access, distribution and use of LRs.
- Availability and use of generic versus task-/domain-specific LRs. Monolingual versus multilingual LRs.
- Methods for the extraction and acquisition of knowledge (e.g. terms, lexical information, language modelling) from LRs.
- National and international activities and projects. Needs, possibilities, initiatives for and forms of international co-operation. Priorities, perspectives and strategies in national and international policies on LRs.
- Integration of various modalities in LRs (speech, vision, language).
- Exploitation of LRs in different types of applications (language technology, information retrieval, vocal interfaces, electronic commerce, etc.).
- Industrial production of LRs. Industrial LR requirements and the community's response.

- Analysis of user needs for LRs. The needs/opportunities of the emerging multimedia cultural industry.
- Evaluation, validation, quality assurance of LRs.
- Evaluation and benchmarking of systems, applications and products, and resources for benchmarking and evaluation.
- Qualitative and perceptive evaluation. Evaluation methodologies, protocols and measures.
- Evaluation in written language processing (text retrieval, terminology extraction, message understanding, text alignment, machine translation, morphosyntactic tagging, parsing, text understanding, summarisation, localisation, etc.). Evaluation in spoken language processing (e.g. speech recognition and understanding, voice dictation, oral dialogue, speech synthesis, speech coding, speaker and language recognition).
- Evaluation of document processing (document recognition, on-line and off-line machine and hand-written character recognition, etc.). Evaluation of (multimedia) document retrieval and search systems.

Important Dates

Summaries for proposed papers (c. 800 words) should be submitted by **1 December 1997**.

E-mail submission in ASCII format is encouraged. Otherwise, five hard copies should be submitted.

E-mail submissions should be sent to:
lrec@ilc.pi.cnr.it
Attn: Antonio Zampolli - LREC

Postal submissions should be sent to:
Antonio Zampolli - LREC
Istituto di Linguistica Computazionale
del CNR
via della Faggiola, 32
56100, Pisa, ITALY

Notification of acceptance will be given by **15 February 1998**, and the final papers must be submitted by **20 April 1998**. Accepted papers will be included in the Conference Proceedings.

Program and Workshops

The program will include both papers and poster sessions, plus invited speakers and a number of panels on major themes of the Conference. In particular, there are plans for a panel on aspects of and perspectives in international co-operation, featuring representatives of the major European, North American and Asian sponsor agencies.

Half-day pre- and post-conference workshops

can be organised at the request of presenters to permit discussion and debate on important topics. The format of each workshop will be determined by the workshop organiser, who will set any necessary deadlines for participants. Various platforms will be available for language resources, tools presentations and systems demos.

For general information on the conference please contact: reli98@goliat.ugr.es
LREC Secretariat
Facultad de Traducción e Interpretación
Dpto. de Traducción e Interpretación
C/ Puentezuelas, 55 - 18002 Granada, SPAIN
tel. +34 58 24 41 00 - fax. +34 58 24 41 04

The demonstration organizer is Prof. Antonio Rubio:

rubio@hal.ugr.es
Antonio J. Rubio Ayuso - LREC
Grupo de Investigación
Procesamiento de Señales y
Comunicaciones
Dpto. Electrónica y Tecnología de
Computadores
Universidad de Granada
18071 Granada
SPAIN
tel. +34 58 24 31 93
fax. +34 58 24 32 30

Information on travel, accommodation and general information on Granada can be obtained from:

Carmen CANO
El Corte Inglés
C/ Arabial, 97, 1
18003 Granada, SPAIN
tel. +34 58 28 26 12
fax. +34 58 20 30 90

**Registration is free of charge
for ELRA members**

ELRA Distribution activities as of 30/09/97

The distribution of ELRA resources are highlighted in the following tables. We distinguish the resources sold and resources distributed for free (mainly ACCOR and TED, distributed to members of the original consortia). The figures from the last reporting period, 30 June 1997, are indicated in brackets.

Distribution to members and non-members

Distribution	Including free distribution	Real sales	Free data	Price in ECU	ELRA margin
Members	33 (29)	30 (26)	3 (3)	99584	30870
Non-members	24 (22)	15 (13)	9 (9)	30868	13723
Total	57 (51)	45 (39)	12 (12)	130452	44593

Distribution with respect to each type of resource

The below table, shows the distribution figures from the colleges; speech, written, terminology and tools.

Distribution	Including free distribution	Real sales	Free data	Price in ECU	ELRA margin
Speech	47 (42)	37 (32)	10 (10)	117342	39032
Written	7 (6)	7 (6)	0 (0)	12466	5239
Terminology	1 (1)	1 (1)	0 (0)	644	322
Tools	2 (2)	0 (0)	2 (2)	0	0
Total	57 (51)	45 (39)	12 (12)	130452	44593

Distribution for commercial versus research use, according to the agreement signed by the user (End-user or VAR)

Usage	Including free distribution	Real sales	Free data	Price in ECU	ELRA margin
Research	39 (37)	27 (25)	12 (12)	10509	3681
Commercial	18 (14)	18 (14)	0 (0)	119943	40912
Total	57 (51)	45 (39)	12 (12)	130452	44593

Distribution in Europe and outside

The below table indicates the distribution of resources to European organisations (including European subsidiaries of American or Japanese companies) and to Non-European organisations.

Geographic area	Including free distribution	Real sales	Free data	Price in ECU	ELRA margin
Europe	49 (45)	37 (33)	12 (12)	102652	32153
Outside Europe	8 (6)	8 (6)	0 (0)	27800	12440
Total	57 (51)	45 (39)	12 (12)	130452	44593

Defining a methodology for designing evaluations: Standards and Sharing

Maghi King

The EAGLES initiative was launched by the European Commission in 1993 to work towards the establishment of standards in language engineering. Five working groups were involved in the first period between 1993 and 1995, covering the areas of text corpora, computational lexicons, grammar formalisms, spoken language and evaluation. This article will concentrate primarily on the work of the Evaluation Group. Full information on EAGLES work, including that of the other groups, can be found by visiting: <http://www.ilc.pi.cnr.it/EAGLES96/home.html>

During the first round of EAGLES work, the Evaluation Group was primarily concerned with defining a methodology for designing evaluations. An associated LRE project, TEMAA, aimed at putting flesh on the theory-oriented bones of EAGLES work by building a small prototype evaluator's workbench in which the EAGLES framework was applied to the evaluation of authoring aids. Within EAGLES itself, preliminary work was done on applying the framework to translation aids. The final report of the first round of EAGLES work can be found at: <http://issco-www.unige.ch/ewg95/ewg95.html>.

A second round of EAGLES work has just started, and will run until the end of 1998. For the Evaluation Group, this is primarily concerned with dissemination, information and consensus building. Two workshops are planned, the first for November 1997, and the group is in the process of setting up a Web site which will serve both as a focal point for information exchange and to discuss evaluation issues. Towards the end of this second period, it is hoped to draw together the results of the workshops and the discussion in the form of a manual of best practice. The group also offers an information service to interested parties. The new Web site in its embryonic form can be found at: <http://www.cst.ku.dk/projects/eagles2.html>

A methodology for evaluation design

One main idea accepted right from the start of EAGLES work was that it was impossible to set up a single evaluation scheme applicable to all situations in which evaluation was required. It was, however, both feasible and desirable to foster a common way of thinking about evaluation, to develop a set of guidelines that could be followed by the designer of any individual specific evaluation. In order

to focus thinking about high-level issues, and to prevent it drifting off into hyperspace, it was decided to adopt the adequacy evaluation of language industry products as a test case.

Adequacy evaluation here means evaluating a system or a product to see if it does what it is supposed to do and if it matches a particular set of user needs. Adequacy evaluation stands in contradistinction to progress evaluation – assessing whether a system has progressed towards some defined goal and to diagnostic evaluation, which primarily aims at finding out why a system fails to give the results expected of it. These distinctions, whilst far from being clear-cut or even completely mutually exclusive, help clarify the goal of an evaluation.

Adequacy evaluation is strictly tied to the expression of user needs, and therefore immediately leads to an interest in user profiling and the description of user profiles. We shall return to this point shortly.

A second important decision was to take an existing standard, ISO 9126, as a basis for the work. ISO 9126 is concerned with software product evaluation; language engineering products are a special type of software product, and as such, the standard should be applicable to them. The standard concerns and defines a set of quality characteristics – functionality, reliability, usability, efficiency, maintainability and portability. In addition, it gives suggested guidelines for using the quality characteristics in the form of an evaluation process model, which was also used to guide EAGLES work. The definition of the quality characteristic on which EAGLES I work mostly concentrated is given as an example in Box 1.

A revised version of the ISO 9126 standard is currently being prepared, and will be an important part of the EAGLES workshop in November.

User needs: the consumer report paradigm

There is clearly a tension between wanting to define a general methodology for evaluation design, and two elements of the starting points above. First, the focus is on adequacy evaluation, which implies a knowledge of user needs. Secondly, ISO 9126 makes the critical assumption that specific user needs can be defined and set out as a quality requirements definition. However, those defining a general methodology have no a priori knowledge of any particular set of users. The apparent paradox was resolved by assuming that it was possible to identify specific classes of users and to define their needs, in much the same way that consumer organisations do when they report on classes of products. If, for example, the consumer report concerns cars, characteristics such as the size of the boot, the petrol consumption, the number of doors, the presence of air bags and so on are picked out and reported on. These characteristics are not chosen at random: behind them is the assumption that there are users who need to transport whole families, and therefore require a certain amount of space, users who travel long distances in the course of their work and therefore need comfort and performance, users whose primary concern is economy, or environmental friendliness, and so on. In other words, there are classes of typical users, who can be profiled and whose needs can be identified.

In the same kind of way, EAGLES work assumes that typical users of language engineering products and their needs can be identified.

Box 1: The ISO 9126 Quality Characteristic for Functionality

"Functionality:

4.6 A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs.

Notes:

1. This set of attributes characterises what the software does to fulfil need, whereas the other sets mainly characterise when and how it does.
2. For the stated and implied needs in this characteristic, the note to the definition of quality applies (see 3.6)."

The note referred to is of some importance, and so is reproduced below:

"3.6

Note: In a contractual environment, needs are specified, whereas in other environments, implied needs should be identified and defined"

(ISO 8402: 1986, note 1).

An addition to the ISO 9126 standard

The EAGLES work made one modification to the ISO 9126 standard's quality characteristics by introducing a seventh quality characteristic, customisability. This is due to the special nature of language engineering products. It is very rare for a customer to be able to buy a language engineering product off the shelf which exactly suits his or her particular requirements. This is true even of very modest products, such as spelling checkers, since the customer will almost certainly have to add words to the dictionary supplied. As the product gets more complex, the need for modification to fit specific requirements becomes even greater, and the difficulty of making the modifications can sometimes increase in consequence, potentially to the point where modification becomes so difficult that it nullifies the potential utility of the product. (Older machine translation systems offered some very good examples of this).

It would have been possible to include customisability as a sub-characteristic of maintainability, but a deliberate choice was made not to do so, partly due to the importance of customisability in language engineering applications and partly to a perception that doing so meant twisting the definition of maintainability somewhat. It should be said immediately that with the definition of maintainability given in the new draft version of ISO 9126, it would have been much harder to justify the creation of a new quality characteristic, since a note to the definition of maintainability explicitly states: "2. If the software is to be modified by the end user, changeability [author's gloss: changeability is given as a sub-quality of maintainability] may be a prerequisite for operability." (ISO 9126 rev.)

Extension of ISO 9126

ISO 9126 contains the following sentence, which proved to be seminal in EAGLES thinking:

"Features are identified properties of a software product which can be related to the quality characteristics" (ISO 9126, p.1)

Perhaps not surprisingly, given that many of those involved in the Evaluation Group came from a computational linguistics background and were familiar with grammar formalisms based on feature structures, this suggested to the group that all quality (sub)-characteristics, software products and users could be formally described through the use of feature structures, where a feature, as is familiar from computational linguistics, is an attribute value pair, with each value being either atomic or itself a feature structure. Formalising descriptions in this way allows us both to

give a more precise content to the notion of quality (sub)-characteristics, and to create the basis for a prototypical evaluator's workbench (a piece of software which, on the basis of formal descriptions of products, quality characteristics and users, can semi-automatically carry out an evaluation of one or more products and produce a report taking user needs into account). The TEMAA project produced such a prototype.

In the EAGLES model, attributes are typed by the type of value they may have. Attributes of the type "fact" have values which are factual. In the case of a spelling checker, for example, factual attributes describe what language the spelling checker deals with, or whether personal dictionaries can be defined.

At first sight factual attributes might seem rather banal. But when used to build up a check list of desirable or undesirable features, they can be quite powerful descriptive tools. Box 2 gives an excerpt from a check list for translation memory systems produced as part of the EAGLES work.

Box 2: Excerpt from the EAGLES Translation Memory Checklist

"D.1.1.3 Adding an SL-segment and its translation to a TM while translating in TM mode.

1. Is it possible to have the SL segment and its translation added to a TM database automatically? If so:

Is it possible to select another TM database to add the sentence to (i.e. to indicate another database as the active one)?

Is it possible to de-activate the automatic updating function in individual cases?

2. How does the program react if an SL-segment and its translation are added to the TM database and one of these segments has already been stored in that database?

- The new segment is added to the TM database
- The new segment is not added to the TM database
- The old segment is deleted from the TM database
- A warning appears indicating that the user has to make a choice".

The reader will easily be able to imagine that when such check lists are worked out carefully and in meticulous detail, they can give a very fine-grained picture of what a system can or cannot do.

Attributes of type "judgement" have a value which is determined on the basis of human judgement. A common example might be "Is the user interface pleasant to work with?". Although judgement type attributes have a bad reputation in the history of evaluation because they rely on human judgement, which is notoriously subjective, there are some situations where they are unavoidable, such as in many sub-characteristics related to usability. There are even situations where judgement attributes, when part of a careful-

ly designed set of metrics, can be argued to give more reliable evaluation results than metrics based on other types of attribute. Johnston (1997), for example, argues that this is the case in the evaluation of text-to-speech and automatic speech recognition systems.

The values of attributes of type "test" are obtained by applying a test to the product, and are typically expressed in quantitative terms. This is important: judgement attributes may also be based on applying a test, but the results of the test will be values on e.g. an opinion scale, rather than a quantitative value which is in itself intended to be directly informative. An example from the evaluation of spelling checkers is:

"What percentage of the 10,000 most common words of the language are included in the spelling checker's dictionary?"

It is perhaps worth mentioning in passing that obtaining values for attributes of type "fact" or "judgement" relies critically on human input, whilst obtaining values for attributes of type "test" can sometimes be completely automated.

Metrics, measures, methods and validity

The 1991 version of ISO 9126 says only this about metrics and their evaluation:

"Due to the high level nature of figure 1, a number of detailed procedures such as analysis and validation of metrics are not shown" (p. 6).

The decision to consider practical test cases meant that the first round of EAGLES work had to face the issue of choice and validation of metrics. A metric was taken to involve a measure and a method to be followed to obtain a value for a particular attribute with respect to that measure. When more than one product or system is evaluated, testing different products will normally result in different values for a single attribute across them. Thus one spelling checker may have 95% of the

most common words of the language in its dictionary, whilst another may have 98%. It is the difference between the values obtained which helps to determine whether a product is best suited to a (prototypical) user's needs.

Not all attributes are of equal importance to all users. Thus the TEMAA prototype Evaluator's Work Bench allowed the values obtained for specific attributes to be combined in ways which reflected the relative importance of an attribute to a class of users. We shall not go into further detail on this topic here.

Measures and the methods used to do the measuring must be both valid and reliable, informally glossed as meaning that the metric must measure what it is supposed to measure (validity) and that it must do so consistently (reliability). Defining valid and reliable metrics is one of the most delicate tasks in evaluation design, and can often require considerable ingenuity.

EAGLES' thinking about validity and reliability was substantially based on work in the social sciences, which frequently distinguishes between internal and external validity.

A metric with internal validity adequately measures an appropriate attribute of the object to be evaluated. An example can be found in reading tests, where the test is based on the construction of texts with a suitable vocabulary for each particular level. The method employed is to ask the person being tested to read the text aloud.

As is the case with this test, internal validity relies only on the judgement of experts. Validity here can only be challenged by challenging the choice of vocabulary, a challenge that could only be made successfully by an expert in the subject.

If a metric has external validity, it is based on a correlation between the results of applying the metric and some external criterion. An example can be found in the measures used by insurance companies to decide what premium an applicant for life insurance should pay. The criteria will include factors such as age, height, weight, the existence of hereditary diseases in the applicant's family, whether the applicant has undergone major surgery, previous serious illnesses and so on; all factors which are held to correlate strongly with life expectancy.

The validity of a metric based on external validity is challenged by challenging the correlation. It is an old saw in statistics that correlation is not causation, and most of us can come up with examples of false correlations. (I cannot be the only person who was told as

a child that if I went to sleep, morning would come more quickly. Until I was old enough to have a quite sophisticated notion of time and its measurement, I believed firmly that the length of the night changed according to my sleeping habits).

A metric may be valid, but have low reliability. The measurements attached to clothing offer a good example of this. In theory, such measurements are valid: size 42 corresponds to so many centimetres round the chest, the waist and the hips. In practice, relying on what the label says is normally ill-advised. I recently came across an example where the label actually showed a diagram of what the size corresponded to in terms of body measurements. Unfortunately, when trying on two garments exactly similar except for a labelled difference in size, the size 46 proved to be a considerably tighter fit than the size 44.

Even if a measure is valid, the method for obtaining it may be invalid or unreliable. For example, if the person administering the reading test is biased, drunk, or in some other way incapable of perceiving accurately the performance of the person reading aloud, he is unlikely to produce consistent test results over a number of tests.

It is also possible for a metric to have high reliability but low validity. Helberg (95) cites IQ tests as an example, since people tend to achieve consistent scores over time, but the scores do not correlate well with, say, job performance in certain types of job.

Space constraints prevent us from discussing all the possible combinations of valid and invalid, reliable and unreliable measures and methods, as well as a number of other problems such as bias and ensuring independence of observations, but enough has been said to justify the earlier claim that ensuring the validity and reliability of metrics is not only of paramount importance, but can also be extremely difficult.

Before leaving this section, though, it should be noticed that the new ISO 9126 draft does contain substantial discussion on metrics and on validity. It would not, however, be appropriate to discuss the draft in any detail here, especially since it is still under discussion within ISO.

Current work

EAGLES work is intended to help designers produce an evaluation that is both

valid and informative, i.e. to avoid some of the pitfalls that previous work on evaluation has fallen into. It is also intended to encourage sharing of evaluation experience and of evaluation resources. Eventually, as the EAGLES methodology gains acceptance, it should also make it possible to share evaluation results, since the theoretical underpinnings of any specific evaluation will be well understood and accepted by others than the individual evaluator.

The second round of EAGLES work, which has only just started, aims at creating a community within which evaluation matters can be discussed and debated and from which a consensus about well-founded ways to go about evaluation should emerge. All who are interested in such matters are invited to a workshop to be held in Brussels on 26-27 November 1997, the theme of which is "Evaluation: Standards and Sharing". Work aiming at the formulation of standards will be presented, as will work providing support for evaluation across projects and system development. More detailed information about the workshop can be found by visiting the EAGLES II web site, which will grow and evolve throughout the life of the project: <http://www.cst.ku.dk/projects/eagles2.html>

Maghi King
ISSCO and ETI
University of Geneva
E-mail: Margaret.King@issco.unige.ch

References

- EAGLES Evaluation Group Final Report, 1995. Paper copy can be obtained from CST, Njalsgade 80, DK-2300 Copenhagen. Electronic copy at: <http://issco-www.unige.ch/ewg95/ewg95.html>*
- Helberg, C., 1995. "Pitfalls of Data Analysis (or how to avoid lies and damned lies)". Paper presented at the Third International Applied Statistics Conference, Dallas, TX. Available on the World Wide Web at: <http://maddog.fammed.wisc.edu/pitfalls/>*
- ISO 1991: ISO/IEC 9126 Information technology - Software Product Evaluation: Quality Characteristics and Guidelines for their use. First Edition 1991-12-15*
- Johnston, R.D., 1997. "Towards a Reference System for Speech Technology", SALT Club Workshop, Sheffield, June 1997.*
- TEMAA project Final Report, 1996. Paper copy can be obtained from CST, Njalsgade 80, DK-2300 Copenhagen.*

SALT Workshop on Evaluation in Speech and Language Technology

Robert Gaizauskas

Conference Report;

17-18 June 1997, Sheffield, UK

The UK Speech and Language Technology Club (SALT) held one of its periodic workshops at Halifax Hall, University of Sheffield, on 17-18 June 1997. The workshop was sponsored by the UK Department of Trade and Industry, the UK Engineering and Physical Sciences Research Council, and by ILASH - the University of Sheffield's Institute for Language, Speech and Hearing.

The theme of the workshop was evaluation in speech and language (S&L) technology. Judging by the keen response in terms of both submissions and attendance, it struck a chord throughout the S&L community, both within the UK and beyond. Record non-UK attendance for a SALT event gave the workshop a truly international flavour and ensured that most strands of current work on evaluation in S&L were represented. Researchers from France, Germany, Denmark, the Netherlands and the US were present, as well as from most centres in the UK involved in S&L work.

In keeping with SALT's mandate to bring speech and language workers together, sessions at the workshop were not, for the most part, divided either into specifically speech- or specifically language-related topics, and all sessions were plenary. The first day started with a review of the best known and highest profile S&L evaluation exercises - the DARPA programmes in the US. Steve Young of Cambridge University gave an overview of the DARPA Continuous Speech Recognition (CSR) programme, and Lynette Hirschman of Mitre Corp. reviewed the DARPA Message Understanding Conference (MUC) written language evaluations and the Air Traffic Information Systems (ATIS) spoken language understanding evaluations. Following this was a session on other multi-site comparative evaluation exercises, three of them French and one Anglo-Irish: Patrick Paroubek of LIMSI reviewed the French GRACE programme for part-of-speech taggers; Lauren Schmitt (INIST-CNRS) reviewed the French AMARYLIS programme for evaluating French language information retrieval systems; Christophe Jouis (Université de Lille 3) described a qualitative programme for evaluating terminology

and semantic relation extraction sponsored by the French AUPELF initiative; and Eric Atwell (Leeds University) described a low-overhead parser evaluation exercise which he organised with Richard Sutcliffe of Limerick University for parsing software manuals.

Two of the sessions at the workshop dealt with system evaluation and two with component evaluation. This division was meant to highlight the difference between evaluating a system which has functionality which a user requires (e.g. an information retrieval system for newspaper texts or a spoken language interface to a railway timetable) and evaluating component technology within such a system which is of no direct interest to an end user, but whose performance will affect the overall behaviour of the system (e.g. a statistical language model or a part-of-speech tagger). The first system evaluation session was on spoken language dialogue systems and included presentations by Gavin Churcher, Leeds University, on a qualitative approach to ranking features of spoken dialogue management systems, and Niels Ole Bernsen on the newly initiated DISC project, which aims to identify best practice in current evaluation of dialogue systems and propose a detailed reference model.

The first component evaluation system included papers by Adam Kilgarriff (Brighton University) reviewing proposals for the evaluation of word sense disambiguation algorithms; Jeremy Crowe (Harlequin Ltd.) on evaluating techniques for recognising and distinguishing between multiple events in discourse; and Gerit Sonntag (University of Bonn) on a novel method for evaluating the prosodic component of a speech synthesis system. The first day ended with a poster session which included presentations by another dozen researchers.

The second day started with a talk by Maghi King on the European Commission-sponsored work on evaluation through the EAGLES project. The EAGLES approach to evaluation stands in contrast to the DARPA-sponsored work. Instead of concentrating on comparative evaluation between sites, each of

which has developed a system to attempt a standardised task, EAGLES has, on the one hand, promoted a user-centred approach, developing checklists of features to assist users in assessing systems (for example, translation memories). In addition, it has standardised resources, such as the Test Suites for Natural Language Processing (TSNLP), against which system developers can benchmark their systems.

The second session on system evaluation followed, and included presentations by Karen Krueger-Thielmann (University of Tübingen) on an evaluation scheme for a multilingual information retrieval system for multimedia documents (part of the Twenty-One and Pop-Eye CEC projects) and by Frances Johnson (Manchester Metropolitan University) on a user-centred approach to evaluating automatic abstracting systems.

The second session on component evaluation started with a paper by Stephen Cox (University of East Anglia), who proposed a novel method for rating speech recognisers based on the idea of systematically degrading a human recogniser's performance until it matches that of an automatic system and using the amount of impairment as a measure of the automatic recogniser's performance. This paper was followed by presentations by Lynette Hirschman on the status and plans for co-reference annotation and evaluation within the DARPA MUC framework; by Douglas Beeferman (Carnegie Mellon University) on a new probability-motivated error metric for segmentation tasks (phoneme, word, sentence, paragraph, document division) that complements precision and recall metrics; and finally by Peter Rodgers (Sheffield University) on the effort to develop generic component evaluation tools within the TIPSTER text processing architecture, as implemented in Sheffield's Generic Architecture for Text Engineering (GATE).

The final session, entitled "Perspectives on Evaluation", included papers which adopted a broader perspective. Nicholas Ostler (Linguacubun Ltd.) spoke on the evaluation of S&L products, projects, and programmes in an effort to shed light on the slow take-off of language technology in the marketplace. John Tait (University of Sunderland) made

an impassioned plea for user-centred evaluation of language technologies not to be forgotten in the face of the formalistic, metricated approaches popularised by the DARPA exercises. Philip Arden of British Telecom, standing in for Denis Johnston, presented a paper describing how traditional methods for rating the speech quality of transmitted speech have found new applicability in rating text-to-speech and automatic speech recognition. Finally Uwe Jost (University of Hamburg) discussed the experiences and lessons learned in first phase evaluation efforts of the Verbmobil speech-to-speech project.

The workshop ended with a panel discussion between Lynette Hirschman, Maghi King, Francoise Neel of LIMSI, David

Pallett of NIST and Karen Sparck Jones of Cambridge University presenting personal views of, and then discussing, "the way forward" in S&L evaluation.

While no consensus concerning new evaluation activities emerged from the workshop, the event clearly played a valuable role in spreading awareness of the breadth of activity and interest in evaluation, and served to forge links between researchers in related areas.

Finally for some credits: members of the local programme committee were Prof. Y. Wilks, Dr. P. Green, Dr. S. Renals and Dr. R. Gaizauskas; local organisation was carried out by Ms. G. Callaghan and Mrs. G. Wells; Mr. M. Crawford and Dr. P.

Rodgers helped immensely in producing the proceedings; and the warden of Halifax Hall, Mrs. N. Taylor, and her staff provided accommodation, meals and conference facilities in a thoroughly friendly and more than satisfactory manner.

NB: Copies of the workshop proceedings are

Dr. R. Gaizauskas
Workshop chair/organiser and
Academic language representative
SALT Steering Committee
robertg@dcs.shef.ac.uk

still available and may be obtained for £10. Please e-mail the author for details.

ELRA Validation report

Work on the validation portions of the ELRA contract is proceeding. The lexicon manuals, which were subcontracted to CST in Copenhagen, are now available at ELDA. The validation criteria and methodology, based on work performed by SPEX in the framework of the SpeechDat project, were discussed during the Cocosda workshop in Rhodes (26-27 September) and the manuals will be available as soon as all the received feedback has been taken into account, while the written corpus work, which was subcontracted to OTA in Oxford, will be ready in November. The terminology manual will be based on the results of the INTERVAL project when these become available.

Work on the lexicon manuals involved defining the task and compiling a list of relevant references (project reports, proposals, etc.), followed by the definition of an initial set of formal validation criteria and related tasks. A first part was concerned with technical validation and conformity with standards (both the producer's own and ELRA's).

After this came content validation and the drafting of the chapters of the ELRA validation manual, the ELRA standard and the validation methodology (including validation schemata).

Preparatory activities performed for the written corpus manual include the compilation of a bibliography of relevant reference materials, the agreement on task definitions, the identification of a preliminary set of corpus resources, and the identification and acquisition of the relevant software tools. Following this, work on the analytical framework has looked at technical and descriptive characteristics and linguistic properties, while appropriate validation procedures have also been defined. As with the lexicon manual, the production of the validation manual is the last step in the process.

For speech, an initial set of formal validation criteria will be defined along with general methodological guidelines for formal validation. Technical validation

encompasses issues such as the size of the Spoken Language Corpus, the number of speakers recorded, the type of speech (extemporaneous, read; monologue, dialogue, group discussion; etc.), the signals that have been recorded (audio, physiological time signals, video signals, etc.), the way in which these signals have been recorded (analogue or digital; if analogue, the bandwidth, signal-to-noise ratio, etc.; if digital, sampling frequency, number of bits per sample, etc.); the annotation coming with the signals, the medium on which the data is delivered, character sets used, and the accompanying documentation.

The formal validation of conformity with standards for Spoken Language Corpora concerns validating the quality of the signals, the precision of the annotations and the use of only legal features, as defined in the relevant standard.

Once finished, the manuals will be widely disseminated and will also be made available on the Web.

The BABEL project - *Speech databases from Central & Eastern Europe* *Peter Roach*

The BABEL project is a Joint Research Project funded by the European Commission as part of the COPERNICUS programme (project no. 1304), and was started in 1995. Its objective is to create a speech database of a number of languages of Central and Eastern Europe, following as closely as possible the design of

the EUROM1 database produced by the ESPRIT SAM project and associated research. The standard workstation adopted is the PC-based SESAM system using the OROS AU-21 board for data acquisition and signal processing calculations; transcriptions are made using versions of the SAMPA machine-rea-

dable phonetic alphabet adopted for the BABEL languages with the assistance of Professor John Wells of University College London. The languages involved are Bulgarian, Estonian, Hungarian, Polish and Romanian, and it is intended that the recorded material will be distributed by ELRA in the form of two CD ROM disks per language.

Not surprisingly, the project has encountered many technical, political and economic problems, but progress is good in most cases. The Hungarian recordings are already complete, and those of Estonian and Romanian nearly so. Each language component of the database will only be regarded as complete when the agreed phonemic labelling has also been completed.

Following the EUROM1 design, each set of language recordings consists of a many-talker set (30 female and 30 male), a few-talker set (5 female and 5 male) and a very-few-talker set (one female and one male), reading material comprising word lists, number sets, phonemically-based sentences and connected-speech passages. The connected-speech passages of the many-talker set of each language (which on average represent 1.5 hours of material per language) are being phonemically annotated by expert transcribers.

The project is co-ordinated by Reading University, with Peter Roach as Project Co-ordinator and Elizabeth Hallum as Project Assistant. The BABEL group is made up of six Eastern and six Western partners. The Eastern group comprises Bulgaria (work being done by a group of phoneticians at the University of Sofia after the tragic death of Anastasia Misheva, our Bulgarian project leader), Estonia (led by Einar Meister with the assistance of Arvo Eek), Hungary (lead by Klara Vicsi), Poland (work divided between the Polish Academy of Sciences in Warsaw, where Ryszard Gubrynowicz is

responsible for quality control, and Lublin, where Wiktor Gonet is responsible for producing part of the Polish data and transcriptions) and Romania (where Marian Boldea carries out the work in Timisoara). The Western European partners receive very little financial support, and function mainly as advisors and as hosts to visiting BABEL researchers from Central and Eastern Europe. In France the partners are LIMSI (Paris) with input from Lori Lamel and also Joseph Mariani, and CNRS where Alain Marchal (previously at Aix-en-Provence and now in Caen) contributes. In Germany we have Bill Barry at Saarbrücken and Krzysztof Marasek at IMS Stuttgart, and in the UK in addition to Reading (where Elizabeth Hallum is Project Administrator) we have University College London with contributions from Adrian Fourcin and John Wells. We hope that in the near future our Polish group will be joined by Professor Basztura of Wrocław, who will carry out essential recording work.

The project began with a kick-off meeting of all partners in Reading in 1995, and later we had a mid-term technical review hosted by Wiktor Gonet in Lublin, Poland. A number of visits have taken place in both directions. Recently, Bill Barry hosted a small workshop on phonemic labelling in Saarbrücken, and in summer 1997 three of the Bulgarian group spent time working on various

aspects of their project in the Speech Research Laboratory at Reading. A number of conference presentations have been given by various project members.

Until recently, the Project Director in Luxembourg was José Soler, who has now moved on to another department in Brussels. We are sorry to lose our contact with him, but look forward to meeting his successor.

The BABEL project has a Web site, and we encourage fellow researchers to look us up. The URL is <http://midwich.reading.ac.uk/research/speechlab/BABEL/>.

We intend to complete the project in 1998, though we have decided to extend our time scale to near the end of the year. Our work plan has always included an end-of-project workshop at which we shall present our methods and data to interested researchers. It is possible that the ELRA conference in Granada in May 1998 could present us with an excellent opportunity to produce a satellite workshop.

Peter Roach
Professor of Phonetics
University of Reading
Department of Linguistic Science
White Knights
PO Box 217
Reading RG6 6AH
United Kingdom
tel. +44-118 931 8138
email: p.j.roach@reading.ac.uk

ELRA Market Segmentation Survey - update

Following the first report on the ongoing study of language resource needs and market segmentation, all ELRA members and some of our partners will now have received the material relating to the study. This includes a questionnaire on different topics such as acquisition and use of language resources, company activities and thoughts on future market development. Also included is a list of applications from different resource fields referred to in the article in the June issue of the ELRA Newsletter.

The purpose of the study is for companies and organisations which use language resources when developing systems, or embed language resources in systems or tools, to inform ELRA about their current and future needs in relation to such resources. In addition, they have the chance to put on record what they expect or wish

from ELRA, as well as their views on the language engineering market in general. As a result, ELRA will be able to improve its collection and distribution of language resources. The results from the study will also form the basis for the development commissioning of new activities, one of which will be production and packaging of language resources.

The companies and organisations participating in the survey are either ELRA members or other major actors on the Language Engineering market. While most are located in Europe, some also come from other parts of the world. We have attempted to create a balance between the different language resource fields, and hence to reach users of speech, text and terminology resources alike.

We would like to take this opportunity to call on everyone using language resources or interested in the development of the language engineering sector, to participate in the study. The finalised results will be made available to ELRA members and all participants, and we hope that the results will make both useful and interesting reading.

If you still have not received the questionnaire, you can download it from the ELRA Website, <http://www.icp.grenet.fr/ELRA/home.html>, or have it sent to you by the ELDA office.

If you have any questions on the study or comments on other ELRA activities, please contact us at:

ELRA/ELDA office
Phone: +33-1-45 86 53 00
Fax: +33-1-45 86 44 88
e-mail: elra-elda@calva.net

Announcements

European Telematics Conference: Advancing the Information Society Barcelona, 4-7 February 1998

Sponsored by the European Commission, DG XIII Information Market and Exploitation of Research, Directorate XIII C "Telematics applications" and Directorate XIII E "Information industry and market and language processing", plus the Spanish Ministerio de Industria y Energía (CDTI), Ministerio de Fomento, Fundacio Catalana per a la Recerca and Ayuntamiento de Barcelona.

Aims of the conference

The Telematics Applications Programme has been the driving force for the development of societal applications for information and communication technologies in Europe over the last 10 years. The programme has brought together users, industry and researchers in shaping leading-edge technologies into applications for the European information society.

With its focus on meeting the needs of users, the programme has helped promote the competitiveness of European industry, improve the delivery of services of public interest and stimulate job creation.

Now the results and achievements of the Telematics Applications Programme and visions of future perspectives will be featured during this significant four-day event in Barcelona.

This event will be of key relevance to some 2-3,000 people: Programme participants, industrialists (in informatics, telecommunications, transport, health, multimedia, etc.), users and decision makers. It will be a platform for:

- disseminating and demonstrating the achievements, results and impacts of the Telematics Applications Programme (TAP),
- exploring visions of the future for societal applications of telematics,
- explaining the role of the Fifth Framework Programme of R&D (1998-2002) in realising these future scenarios,
- bringing project participants together to share expertise on the state of the art in telematics technologies and applications.

To meet the needs of a large and diverse audience, the event will run over three days of formal sessions and include a large exhibition of demonstrations and telematics developments.

There will be a large exhibition offering visitors an opportunity to view and follow up on the projects highlighted, as well as demonstrations and displays of the Programme work and projects. The exhibition will be laid out according to thematic sector villages highlighting the key projects, surrounded by smaller exhibitions of projects and other stands for digital sites, 5th Framework Programme information and the information/help desk.

For further information please e-mail: conference@cscdc.be

If you wish to ensure that you receive an invitation, please send your full address and contact details to: invite-request@cscdc.be

Conference Steering Committee

Michel Richonnier, European Commission, Director DG XIII C "Telematics Applications"

Frans de Bruine, European Commission, Director DG XIII E "Information Industry and Market and Language Processing"

Giangaleazzo Cairoli, European Commission, Head of Unit DG XIII C1 "Programme Management"

"Managing Global Business Challenges": LISA Forum, Geneva, 4-5 December 1997

Hosted by The McQueen Group in collaboration with Xerox Limited Technical Center, the next LISA Forum will focus on the business and technology issues associated with the trend towards consolidation in the translation and localisation industry. The keynote presentation, "Acquisition is the Easy Part" by Florita Mendez (President, Mendez Language and Technology) will take a candid look at the business and organisational management challenges facing those involved in consolidation. Panel discussions and interactive workgroups will put special emphasis on the Internet, quality, benchmarking and other aspects of business management.

Preceding the Forum on 2-3 December will be a LISA workshop on "Creating Localizable On-screen Information", run by Richard Ishida, Globalization Consultant at Xerox Limited Technical Center.

Registration is limited to 20 participants on a first-come, first-serve basis.

A full agenda and registration details can be downloaded from the LISA Web site, LISA Administration
<http://www.lisa.unige.ch/progen.html>

COLING-ACL'98: First Announcement and call for papers

On behalf of the International Committee on Computational Linguistics (ICCL) and the Association for Computational Linguistics (ACL), we are pleased to announce a major joint conference COLING-ACL'98 which will be held on the campus of l'Université de Montréal, Canada on August 10-14, 1998. The RALI laboratory of the Computer Science and Operations Research Department of l'Université de Montréal will be hosting the first North American COLING since the joint COLING-ACL'84 was held at Stanford University in 1984.

We welcome submission of papers describing substantial, original and unpublished research contributions on all aspects of computational linguistics. Program subcommittees will be organised around the following main areas:

- Linguistic issues & associated formalisms
- Linguistic resources & computational methods
- Applications
- Projects

Submissions may be of two different types: 1) regular papers; and 2) project notes. Regular papers should report the results of original completed research. Project notes, on the other hand, should describe ongoing research or demonstrate a system. Regular papers will be presented in three parallel sessions that do not overlap with the presentation of project notes.

All submissions and questions regarding submissions should be sent to:

COLING-ACL'98 submissions
Professor Christian Boitet
GETA, CLIPS, IMAG BP 53
38041 Grenoble cedex 9
France
e-mail: ColingACL98.program@imag.fr

Deadlines

Submission announcement (ID page): e-mail before January 20, 1998

Submission (6 copies + ID page):
to arrive in Grenoble no later than January 30, 1998
Notification to authors: April 17, 1998

Final camera-ready copies (2):
to arrive in Montreal no later than May 30, 1998

The organising committee is being chaired by:

Dr. Pierre Isabelle
RALI, DIRO
Université de Montréal
PO Box 6128, Succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
e-mail: coling-ac198@iro.umontreal.ca
Tel: (514) 343-6161; Fax: (514) 343-2496

EURALEX'98

The Eighth International Congress of the European Association for Lexicography (EURALEX) will be held at the University of Liège Belgium, from 4-8 August, 1998. The EURALEX Congresses bring together scholars, professional lexicographers, publishers and others interested in dictionaries of all types. The programme will include workshops (among others on linguistic resources for NLP and on dictionary use), plenary lectures, parallel sessions of individual papers, software demonstrations and a poster session. The congress will be preceded by two tutorials, one on "creating a bilingual dictionary" and one on "preparing a terminological database".

Papers are invited on all aspects of lexicography but the main topics are: computational lexicology/lexicography, lexical combinatorics, the dictionary-making process, bilingual lexicography, lexicographical and lexicological projects, terminology and dictionaries.

First and second circulars, call for papers and registration forms are available from:

<http://engdep1.philo.ulg.ac.be/euralex.html>
E-mail: amichiels@ulg.ac.be
Fax: +32-4-3665721
Address: EURALEX'98 Congress Organisers University of Liège,
Department of English Language and Linguistics,
Building A2,
Place Cockerill 3,
B-4000 Liège, Belgium

New resources

ELRA-S0034 Verbmobil

This resource consists of spontaneous speech recorded in a dialog task (appointment scheduling). The German corpus has a total of 13,910 utterances (turns). 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 4.0 - VM40 (1 CD-ROM, original edition)

72 Dialogues, 181 Appointments, 1,588 Turns.

VM CD 4.1 - VM41 (1 CD-ROM, new edition)

72 Dialogues 181 Appointments 1,588 Turns

This new edition contains the transliterations of all dialogues, signal files with PhonDat 2 Header structure, software and speaker documentation and partitur files*. All files were evaluated according to BAS guidelines.

VM CD 5.0 - VM50 (1 CD-ROM, original edition)

101 Dialogues, 256 Appointments, 2,154 Turns.

VM CD 5.1 - VM51 (1 CD-ROM, new edition)

101 Dialogues, 256 Appointments, 2,154 Turns.

This new edition contains the transliterations of all dialogues, signal files with PhonDat 2 Header structure, software and speaker documentation and partitur files*. All files were evaluated according to BAS guidelines.

VM CD 6.0 - VM60 (1 CD-ROM, original edition)

American/English and 'Denglish'**. 146 Dialogues, 191 Appointments, 1,828 Turns.

VM CD 6.1 - VM61 (1 CD-ROM, new edition)

American/English and 'Denglish'**. 146 Dialogues, 191 Appointments 1,828 Turns. This new edition contains the transliterations of all dialogues, signal files with PhonDat 1 Header structure, software and speaker documentation. All files were evaluated according to BAS guidelines.

VM CD 7.0 - VM70 (1 CD-ROM, original edition)

68 Dialogues, 238 Appointments, 1,739 Turns.

VM CD 7.1 - VM71 (1 CD-ROM, new edition)

68 Dialogues, 238 Appointments, 1,739 Turns. This new edition contains the transliterations of all dialogues, signal files with PhonDat 2 Header structure, software and speaker documentation and partitur files*. All files were evaluated according to BAS guidelines.

VM CD 8.0 - VM80 (1 CD-ROM, original edition)

American/English 167 Dialogues, 167 Appointments, 1,181 Turns.

VM CD 8.1 - VM81 (1 CD-ROM, new edition)

American/English 167 Dialogues, 167 Appointments, 1,181 Turns. This new edition contains the transliterations of all dialogues, signal files with PhonDat 1 Header structure, software and speaker documentation. All files were evaluated according to BAS guidelines.

VM CD 12.0 - VM120 (1 CD-ROM, original edition)

207 Dialogues, 207 Appointments, 2,154 Turns.

VM CD 12.1 - VM121 (1 CD-ROM, new edition)

207 Dialogues, 207 Appointments, 2,154 Turns. This new edition contains the transliterations of all dialogues, signal files with PhonDat 2 Header structure, software and speaker documentation and partitur files*. All files were evaluated according to BAS guidelines.

Price for ELRA members: 76 ECU per CD-ROM

Price for non members: 152 ECU per CD-ROM

* partitur files: files describing the different parts which constitute the corpus - word order, phrase order, etc.

** 'Denglish': English spoken by Germans.

ELRA-L0029 CELEX Dutch lexical database

The Dutch CELEX data is derived from R.H. Baayen, R. Piepenbrock & L. Gulikers, The CELEX Lexical Database (CD-ROM), Release 2, Dutch Version 3.1, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

Apart from orthographic features, the CELEX database comprises representations of the phonological, morphological, syntactic and frequency properties of lemmata. For the Dutch data, frequencies have been disambiguated on the basis of the 42.4m Dutch Instituut voor Nederlandse Lexicologie text corpora.

To make for greater compatibility with other operating systems, the databases have not been tailored to fit any particular database management program. Instead, the information is presented in a series of plain ASCII files, which can be queried with tools such as AWK and ICON. Unique identity numbers allow the linking of information from different files.

This database can be divided into 5 different subsets:

- orthography: with or without diacritics, with or without word division positions, alternative spellings, number of letters/syllables;
- phonology: phonetic transcriptions with syllable boundaries or primary and secondary stress markers, consonant-vowel patterns, number of phonemes/syllables, alternative pronunciations, frequency per phonetic syllable within words;
- morphology: division into stems and affixes, flat or hierarchical representations, stems and their inflections;
- syntax: word class, subcategorisations per word class;
- frequency of the entries: disambiguated for homographic lemmata.

Price for ELRA members:

* for research use: Contact ELRA

* for commercial use: complete set: 56,182 ECU; orthography subset: 6,000 ECU; phonology subset: 12,273 ECU; morphology subset (inflectional): 6,000 ECU; morphology subset (derivational): 13,636 ECU; syntax subset: 6,000 ECU; frequency subset: 12,273 ECU.

Price for non members:

* for research use: Contact ELRA

* for commercial use: complete set: 93,636 ECU; orthography subset: 10,000 ECU; phonology subset: 20,454 ECU; morphology subset (inflectional): 10,000 ECU; morphology subset (derivational): 22,727 ECU; syntax subset: 10,000 ECU; frequency subset: 20,454 ECU.

ELRA-S0042 Polycost

The POLYCOST speech database was recorded during January-March 1996 as a common initiative entitled "Speaker Recognition in Telephony" within the COST 250 action. The main purpose of the database is to compare and validate speaker recognition algorithms. The data was collected via international telephone lines, with more than five sessions per speaker, and with English spoken by foreigners.

The database contains around 10 sessions recorded by 134 subjects from 14 countries. Each session contains 14 items. All items, except the last two, are expressed in English. The speakers come from the European countries taking part in the action. Approximately 10 speakers per country were provided by each partner.

Each session comprises 15 prompts, including one prompt for DTMF detection, 10 prompts with connected digits uttered in English, 2 prompts with sentences uttered in English and 2 prompts in the speaker's mother tongue. One of the prompts in the speaker's mother tongue consists of free speech.

English:

- 4 prompts distributed throughout the session in which the speaker pronounces his or her 7-digit client code;
- 5 prompts distributed throughout the session in which the speaker pronounces a sequence of 10 digits (the same from session to session and from speaker to speaker);
- 2 prompts in which the speaker pronounces the sentences: "Joe took father's green shoe bench out" and "He eats several light tacos", as fixed password phrases which are common to all speakers;
- 1 prompt in which the speaker is supposed to give his or her international phone number.

Mother tongue

- 1 prompt in which the speaker gives his or her first name, family name, gender (female/male), town and country;
- 1 prompt with free speech.

The database was collected through the European telephone network and was recorded through an ISDN card on XTL SUN platform with an 8 kHz sampling rate. Most of the calls were automatically classified by DTMF detection. Manual classification has been used in the case of no DTMF or wrong DTMF PIN code (circa 10% of the database).

The English prompts are segmented and labelled at the word level (orthographic transcription and word stretches). The prompts in mother tongue are simply labelled (an orthographic transcription will be given). The conventions used for the annotation are those defined within the SpeechDat project.

Character set: ISO-8859-1

Medium: 2 CD-ROMs. The first CD contains speech data from speakers M001-M069, and the second CD contains data from speakers F001-F060 plus M070-M074.

Total size CD1: 636 MB

Total size CD2: 610 MB

File format: A-law, 8 kHz sampling rate, 8 bits/sample, with no file header.

Price for ELRA members:

- * for research use: 500 ECU
- * for commercial use: 1,000 ECU

Price for non members:

- * for research use: 600 ECU
- * for commercial use: 1,200 ECU

Price for COST 250 partners: 100 ECU

ELRA-S0044 SPINA Corpus ("Robots Commands")

The corpus contains German read speech of 22 different speakers (6 male, 16 female). The corpus consists of 10 robot command sentences and 62 robot command words. Each speaker reads the whole corpus 5 times, except one speaker who reads the sentence corpus 16 times and the word corpus 51 times. The speakers were recorded at two different sites in Germany (University of Goettingen, University of Bochum).

The corpus contains a total of 10,810 recorded utterances. All speakers are within the age of 25-30. Two speakers are non-native speakers. A file gives information about the speakers (speaker ID, recording site, sex). The task for the speaker was to read carefully but fluently. If an error occurred, the recording was interrupted by the supervisor and the sentence was repeated. The signal files are raw files without any header, 16 bit per sample, linear, most significant byte first, 16 kHz sample frequency. The orthography of the corpus is given in two distinct files which contain the prompted words and the prompted sentences as an ordered list.

The recording conditions are as follows:

Microphone: AKG acoustics, C414B-TL, condensator microphone omnidirectional, built-in attenuator and high pass filter switched off, distance to mouth 50 cm.

Environment: Studio Quality, echo cancelled room, about 121 qqm

Preamplifier: John Hardy, M-1

Sampling rate: 48 kHz to DAT recorder, filtered to 16 kHz

Resolution: 16 Bit, most significant byte first

The speech data were digitally filtered to 8 kHz cut-off frequency and downsampled to 16 kHz.

The corpus consists of 1 volume, total size 266,361 KB uncompressed data.

The signal of each utterance is stored in a separate file. Symbolic information like segmentations or labelling (e.g. Phonological Segmentation of words or Word Segmentation of sentences) are stored in files with the same prefix but with different extensions.

Prices for ELRA members: 76 ECU

Prices for non members: 152 ECU

ELRA-S0043 Onomastica-Copernicus database

The ONOMASTICA project was a European-wide research initiative within the scope of the Linguistic Research and Engineering Programme, the aim of which was the construction of a multi-language pronunciation lexicon of proper names. That project covered eleven European languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish.

Although the ONOMASTICA project ended in June 1995, the work continued with the introduction of new partners, addressing names in Eastern and Central European languages: Czech, Estonian, Latvian, Polish, Romanian, Slovakian, Slovenian and Ukrainian, in a new project funded by the European Commission's Copernicus Programme.

Though the result of the Onomastica project related to Western languages is not available (except for the German), the result of this new project is available. It consists of a collection of 1,783,390 transcriptions of 1,705,653 names, broken down as follows:

- **Czech:** 257,700 entries consisting of 244,025 names prepared by Dr. Pavel Kolar of the Language Institute, Silesian University, Opava, Czech Republic.
- **Estonian:** 209,515 entries consisting of 208,380 names prepared by Dr. Peeter Päll of the Institute for the Estonian Language, Estonian Academy of Sciences, Tallinn, Estonia.
- **Latvian:** 258,214 entries consisting of 245,331 names prepared by Dr. Andrejs Spektors of the Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia.
- **Polish:** 285,412 entries consisting of 244,632 names prepared by Prof. Wiktor Jassem of the Institute of Fundamental Technological Research, Polish Academy of Sciences, Posnan, Poland.
- **Slovak:** 228,257 entries consisting of 228,257 names prepared by Dr. Peter Durco of the Department of Foreign Languages, Police Academy of the Slovak Republic, Bratislava, Slovak Republic.
- **Slovenian:** 285,862 entries consisting of 283,449 names prepared by Dr. Zdravko Kacic of the Faculty of Technical Sciences, University of Maribor, Maribor, Slovenia.
- **Ukrainian:** 258,430 entries consisting of 251,579 names prepared by Dr. Yevgeniy Ludovik of the Institute of Cybernetics, Ukraine Academy of Sciences, Kiev, Ukraine.

The databases are presented in Microsoft Access format and in ASCII text format, together with database browser software prepared by Keith Edwards of the Centre for Communication Interface Research, The University of Edinburgh.

More details are available on the ELRA Web site.

Price for ELRA members:

- * for research use: 400 ECU
- * for commercial use: 3,000 ECU

Price for non members:

- * for research use: 800 ECU
- * for commercial use: 6,000 ECU

ELRA-S0045 German Pronunciation Rules Set - PHONRUL 9.0

PHONRUL is a collection of computer-readable underspecifying pronunciation rules of standard German. This set describes the most common known effects in German pronunciation if deviating from the so-called canonic or citation form of words. The knowledge of this rule set was derived from empirical analysis of speech corpora as well as from a multitude of publications about German phonetics. The set does not contain any dialect-specific rules, however the line between Standard German and dialects is indistinct. Presently, this rule set is used at the University of Munich to aid automatic segmentation and labelling of unknown speech utterances.

The rule set, in its present form, consists of approximately 1,500 complex rules which expand to 5,546 simple replacement rules. The rule set was designed for extended German SAM-PA, but can be translated into other alphabets (e.g. Worldbet, IPA) without much effort.

Price for ELRA members:

- * for research use: 76 ECU
- * for commercial use: 482 ECU

Price for non members:

- * for research use: 152 ECU
- * for commercial use: 964 ECU

ELRA-W0015 "Le Monde" Text corpus

Electronic archiving of "Le Monde" articles started on 1 January 1987. Some 200 articles are added every day, and as of October 1997 the database contains more than 500,000 articles, making it the biggest of its kind for all French daily newspapers.

The corpus is available in an SGML-tagged ASCII text format. Each month consists of some 10 MB of data (circa 120 MB per year). Data ranging from 1987 until present date are available through ELRA (each buyer may purchase up to 5 years of data).

Price for ELRA members:

- * 1 year 291 ECU
- * 2 years 581 ECU
- * 3 years 872 ECU
- * 4 years 1,163 ECU
- * 5 years 1,454 ECU

Price for non members:

- * 1 year 378 ECU
- * 2 years 756 ECU
- * 3 years 1,134 ECU
- * 4 years 1,512 ECU
- * 5 years 1,890 ECU

MEMBERSHIP APPLICATION

Organisation
 Department
 Name of Designated Representative.....
 Address

 Town Postcode.....
 Country
 Telephone.....
 Fax
 E-mail
 College : () Spoken () Written () Terminology

I agree to the information above appearing in the ELRA Directory

Signature _____ Date _____

Notes :

- 1) The annual membership fee for all European organisations and for non-profit organisations outside Europe is ECU 1,000. For profit-making organisations outside Europe the annual fee is ECU 5,000 per year. An invoice for the amount will be sent upon receipt of the completed application form.
- 2) Applicants may apply for membership in one or more of the Colleges, but will be required to pay multiple membership fees if entering more than one College. European members are entitled to voting privileges in each College in which they are registered.
- 3) Payment may be made by bank transfer or cheque in ECU made out to ELRA. Bank: BNP (Luxembourg) S.A, Bd. Royal, L-2953 Luxembourg; Account number 63-114418-57-6102-997. Bank charges shall be paid by the subscriber.

ELRA 1997 MEMBERS' SPECIAL OFFER

ELRA is offering two language resources free of charge to all new members. Please choose from the resources below.

<input type="checkbox"/>	ELRA-S0031	TED (Translanguage English Database) - Recordings made of 188 oral presentations in English given at Eurospeech'93 in Berlin (high percentage of non-native English speakers).
<input type="checkbox"/>	ELRA-W0003	CRATER Multilingual aligned corpus - 3 x 1,000,000 token corpora for English, French and Spanish, morphosyntactic annotations, lemmatisation and term extraction routines for English, French and Spanish.
<input type="checkbox"/>	ELRA-W0006 ELRA-W0007	MLCC - Multilingual corpus - Contains articles from 6 European newspapers: <i>Het Financieele Dagblad</i> (Dutch, 8.5 million words), <i>The Financial Times</i> (English, 30 million words), <i>Le Monde</i> (French, 10 million words), <i>Handelsblatt</i> (German, 33 million words), <i>Il sole 24 Ore</i> (Italian, 1.88 million words), <i>Expansion</i> (Spanish, 10 million words). MLCC - The Official Journal of the European Communities - Parallel corpus of translated documents in the nine official European languages (1992-1994), divided into 2 subcorpora: written questions (10.2 million words) and parliamentary debates (5 to 8 million words per language).
<input type="checkbox"/>	Samples from ELRA-T0001 to ELRA-T0088	MULTIDOMAIN MULTILINGUAL TERMINOLOGY DATABASE - Over 20,000 terms from several domains including Finance, Telecommunications, Energy, Environment, etc. A guide to terminology consolidation will also be supplied (<i>Please contact ELRA for details</i>).