The ELRA Newsletter

January - June 2007

Vol.12 n.1&2

Contents

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

ISSN: 1026-8200

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

Dear Colleagues,

The ELRA Annual General Assembly was held on May 22, in Paris. This meeting was chaired by Bente Maegaard, ELRA President, and 13 members attended the General Assembly. The ELRA activities were reviewed for 2006 and the plans for 2007 were presented with a focus on the strategic committees, PCom, VCom and ECom, whose activities have been thoroughly discussed. The ELRA regular activities will continue over 2007 and especially the preparation of LREC 2008. During this meeting, the Board members reasserted that ELRA visibility, through the **Language Resources and Evaluation Journal** but also with the usual information dissemination actions, should be a strong concern for 2007.

Two events, in which ELRA is strongly involved, have also been discussed:

- The "**Automatic Procedures in MT Evaluation**" workshop, organized within the framework of MT Summit XI, Copenhagen (Denmark) on September 11, 2007 by the ELRA Evaluation Committee.
- **LREC 2008**, which will take place on May 26 - June 1, 2008 in Marrakech (Morocco).

During these last few months, ELRA and ELDA have continued to work on a number of projects supported by funding agencies, in particular, both European-funded projects TC-Star and CHIL.

As for this newsletter, it contains a comprehensive and detailed overview of all the services offered by ELRA to the community and to its members.

New resources have been secured for distribution. As usual, these are announced in the last section of this newsletter and consist of:

- *Evaluation Packages from the French national research programme Technolangue for various language technologies:*
  - Transcription of Broadcast News: ESTER Corpus and Evaluation Package (catalogue ref. ELRA-S0241 and E0021)
  - Alignment of Multilingual Corpora: ARCADE II (catalogue ref. ELRA-E0018)
  - Acquisition of Terminological Resources: CESART (catalogue ref. ELRA-E0019)
  - Machine Translation: CESTA (catalogue ref. ELRA-E0020)
  - Question-Answering Task for Information Retrieval: EQueR (catalogue ref. ELRA-E0022)
  - Speech synthesis evaluation: EvaSy (catalogue ref. ELRA-E0023)
  - Speech Dialogue for tourist information-oriented servers: MEDIA (catalogue ref. ELRA-E0024)
- *Monolingual and Multilingual Lexicons from the general domain:*
  - POLEX Polish Lexicon (catalogue ref. ELRA-L0074)
  - Bulgarian Linguistic Database (catalogue ref. ELRA-L0075)
  - English-German Bilingual Dictionary (catalogue ref. ELRA-M0038)
- *Written Corpora from news agencies:*
  - Update of the text corpus of "Le Monde" with the addition of 2005 and 2006 data (catalogue ref. ELRA-W0015)
  - Catalan Corpus of News Articles (catalogue ref. ELRA-W0047)
- *A Broadcast News Speech Corpus resulting from LDC (Linguistic Data Consortium, USA) and ELRA cooperation in the European-funded NetDC (Network of Data Centres) project:*
  - NetDC Arabic BNSC (Broadcast News Speech Corpus) (catalogue ref. ELRA-S0157)
- *Phonetic lexicons from the European-funded LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) project:*
  - LC-STAR Turkish phonetic lexicon (catalogue ref. ELRA-S0229)
  - LC-STAR Russian phonetic lexicon (catalogue ref. ELRA-S0230)
  - LC-STAR English-Russian Bilingual Aligned Phrasal lexicon (catalogue ref. ELRA-S0231)
  - LC-STAR Hebrew (Israel) phonetic lexicon (catalogue ref. ELRA-S0235)
  - LC-STAR English-Hebrew (Israel) Bilingual Aligned Phrasal lexicon (catalogue ref. ELRA-S0236)
  - LC-STAR US English phonetic lexicon (catalogue ref. ELRA-S0237)
- *Speech Microphone resources from the European-funded Speecon project (for the development of voice controlled consumer applications):*
  - Swiss-German Speecon database (catalogue ref. ELRA-S0232)
  - US English Speecon database (catalogue ref. ELRA-S0233)
  - The French-Canadian Speecon database (catalogue ref. ELRA-S0240)
- *Speech Microphone Resources from the NATO research group:*
  - MIST Multi-lingual Interoperability in Speech Technology database (catalogue ref. ELRA-S0238)
  - N4 (NATO Native and Non Native) database (catalogue ref. ELRA-S0239)
- *Speech Telephone Database from the European-funded SALA (SpeechDat Across Latin America) project:*
  - SALA Spanish Chilean Database (catalogue ref. ELRA-S0234)

Once again if you would like to join ELRA and benefit from its services (that are also summarized at www.elra.info), please contact us.

Bente Maegaard, President

Khalid Choukri, CEO

INTRODUCTION

ELRA, through its operational body ELDA, carries out a wide variety of activities related to Language Resources. These activities range from identification, to distribution, promotion, validation, production, as well as evaluation of technologies, where each of them consists of a long series of sub-activities.

The first three, namely **identification**, **distribution** and **promotion**, and their related sub-activities can be illustrated by a two-direction relationship with two different types of entities, namely the providers and the customers, as shown in Figure 1. This implies managing all discussions, negotiations and legal matters with the providers as well as with the customers, or merely potential customers (if no purchase takes place). As far as providers are concerned, ELRA goes through tasks of (a) identification (searching for resources unless their owner has offered them directly to ELRA), (b) discussion of legal matters, pricing and revenues (i.e. negotiations to establish a contract and define all its conditions), and (c) royalty reporting (concerning all regular notifications together with payments to be done to the owner of the resource once a sale has been formalised). For the customers, ELRA carries out tasks of (d) promotion, in order to advertise the resources and any relevant activity (with the maintenance of catalogues, edition of the ELRA Newsletter, the organisation of the LREC Conference, and the maintenance of the HLT Portal, among others), as well as of (e) distribution, which comprises all sales management matters (e.g., information and data sample exchanges, financial discussions, etc.).

Further to its interaction with both providers and customers, ELRA also invests considerable efforts on the **validation** and **production** of LRs, with direct implication from its Validation Committee (VCom) and Validation Centres (CST and SPEX), as well as its Production Committee (PCom). All these activities are described in detail in the following sections.

Although not the focus of the current Newsletter, ELRA also plays an important role in the area of t**echnology evaluation**, which is referred to here through a description of the evaluation packages distributed and the HLT evaluation portal.

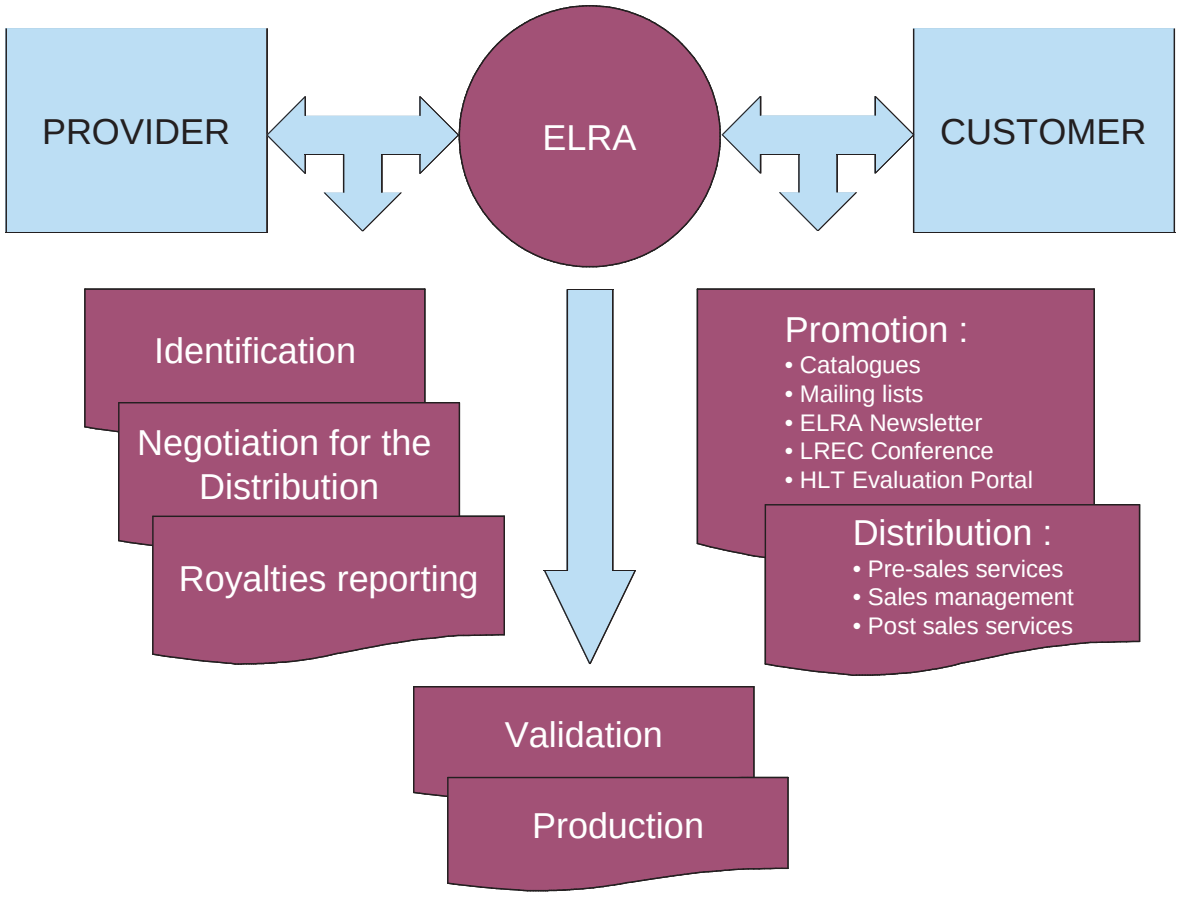


Figure 1: ELRA activities

LRs Identification and the Universal Catalogue

The Identification of Languages Resources

The efforts on the identification of LRs have been increased considerably at ELRA in the past few years. This has been directly linked to ELRA's emphasis on identifying already existing resources, wherever they may be developed and stored, easing their availability to their potential users within the Language Technology R&D community and, consequently, helping to reduce the efforts to produce already-available resources.

As a consequence, and further to our Catalogues of ready-for-distribution LRs, intensive work is taking place on the identification of all existing LRs in order to compile them in what we refer to as the ***Universal Catalogue***, which provides all available information for its users.

A number of methods and procedures are already used by ELRA for this identification task, which require a constant monitoring or search for relevant information, not only in terms of language resources themselves, but also in terms of projects, conferences or simply consortia gatherings that may be taking place. In general, these methods and procedures can be summarised as follows:

- "LR watch" through ads, announcements in linguistic related mailing lists: Linguist list, Corpora, LN, ELSNET, etc.
- Hiring the service of internal and external experts with the mission of identifying LRs.
- Cooperation with institutions, participation in national, European and international projects related to LR creation.
- Development of expertise on national, European and international research programs and their relevant projects, including:
  - European Community Initiatives and Programs: INTAS, CORDIS, IST, COST, EUREKA.
  - EU/International Networks and Associations: AFNLP, ALTA, ATILF, Dutch Language Union, ELSNET, LT World, NEMLAR, OntoWeb, SEPLN.
  - LRs distribution agencies: LDC, CCC, GSK, Indian LDC, OGI, TELRI.
  - Information Portals: COLLATE.
  - Project Consortia: National & International.
  - Metadata Infrastructure: IMDI, INTERA, OLAC.
  - Standards: ISLE, ISO/TC 37/SC 4.
  - International Coordination Committees: ALRC, ICCWLRE, COCOSDA.
- Participation in HLT events and conferences.
- Search for information in technical documentation and publications, such as conference proceedings (LREC, Eurospeech, ICSLP, Coling, ACL).
- Information about the existence of LRs by ELDA contacts.
- ELRA's and ELDA's websites.
- Mouth-to-ear information.

Further to this general identification task, ELRA is also carrying out identification upon demand for its members. This latter activity implies searching for specific LRs that our members need and that are not available in our catalogues. Thus, in addition to feeding us with information for our Universal Catalogue, this activity allows us to enrich our own ELRA catalogues and to continue in the forefront of the distribution of LRs for the HLT community. A number of interesting LRs are currently being searched and discussed under this framework. A further positive feature of this modality is that new resources are sometimes created to meet our members' needs, which helps the field of Language Resources to remain productive and growing.

The Universal Catalogue

The Universal Catalogue comprises information regarding Language Resources identified by the ELRA team. It aims to be a repository for all identified Language Resources so as to provide the ELRA members with information on what resources there exist and what their characteristics are. New information is constantly being added.

The Universal Catalogue functions as a pointer to interesting resources, a search help that allows users to have an initial insight on a resource and to enquire further about it. Our team is always pleased to provide its users with as much information as required. Furthermore, priority for the acquisition of LRs for the ELRA

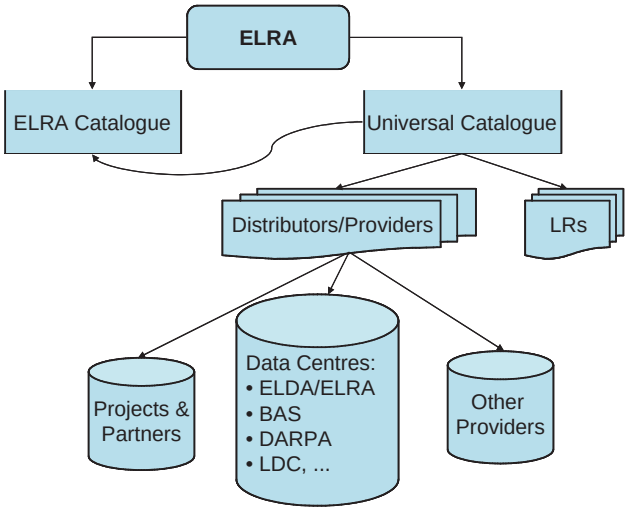


Figure 1: Relation between the ELRA Catalogue and the Universal Catalogue

Figure 2: Sample of screenshot from ELRA's Universal Catalogue

Catalogue will be given to those LRs required by its users, so we welcome any queries or proofs of interest.

Thus, further to its role of repository for all identified LRs, the Universal Catalogue plays a very important role towards the ELRA Catalogue. Both catalogues have a relation of self-information where the Universal Catalogue is an important source of information for the ELRA Catalogue to acquire new LRs. Figure 1 shows how the Universal Catalogue comprises information regarding existing LRs and their Distributors and potential Providers. Such Distributors or potential Providers can be of the following nature: Data Centers (BAS, LDC...), Projects and their Consortia Partners (LC-STAR, OrienTel,...) or other types (e.g., research groups producing LRs outside the context of a project, etc.).

The design of the Universal Catalogue is based on that of the ELRA Catalogues, aiming to provide also a dynamic MySQL-based interface that offers the same features of the other catalogues and can easily interact with the existing catalogues, as it is planned for future developments. Figure 2 shows a screenshot of the Universal Catalogue.

As it can be observed on its left-hand side menu, further to the standard categories established and generally used by ELRA ("Speech and Related Resources", "Written Resources", "Terminological Resources", and "Multimodal/Multimedia Resources"), the Universal Catalogue also offers a 5th category called "Tools". This category comprises a list of language processing tools (e.g. lemmatizers, tokenizers, parsers, etc.) that are also available for the community and may be of use to researchers or developers needing such tools and not aware of their existence.

The Universal Catalogue also allows for feedback to be sent directly to its administrator by means of the easy-to-fill-in form illustrated in Figure 3.

As a result of an initial but thorough search for LR sources, and further to the over 900 LRs already distributed through the ELRA Catalogue, the following figures can already be presented in terms of identified resources to be further processed and sub-classified for the Universal Catalogue: a total number of 841 potential LRs have been identified and are being classified.

Last but not least, we look forward to hearing from all members of the Human Language Technology community who would like to share any information about existing resources they may have or they may know about (and which they may consider interesting for the community). The more exhaustive the Universal Catalogue is the more useful it will be for all its users. Moreover, we would also like to encourage all our readers to contribute to both Universal Catalogue and ELRA Catalogue and to contact us to help them distribute their resources. In particular, we will be happy to help you distribute free and low-cost resources specifically oriented to R&D, through both our Catalogue of LRs and through our R&D Catalogue.

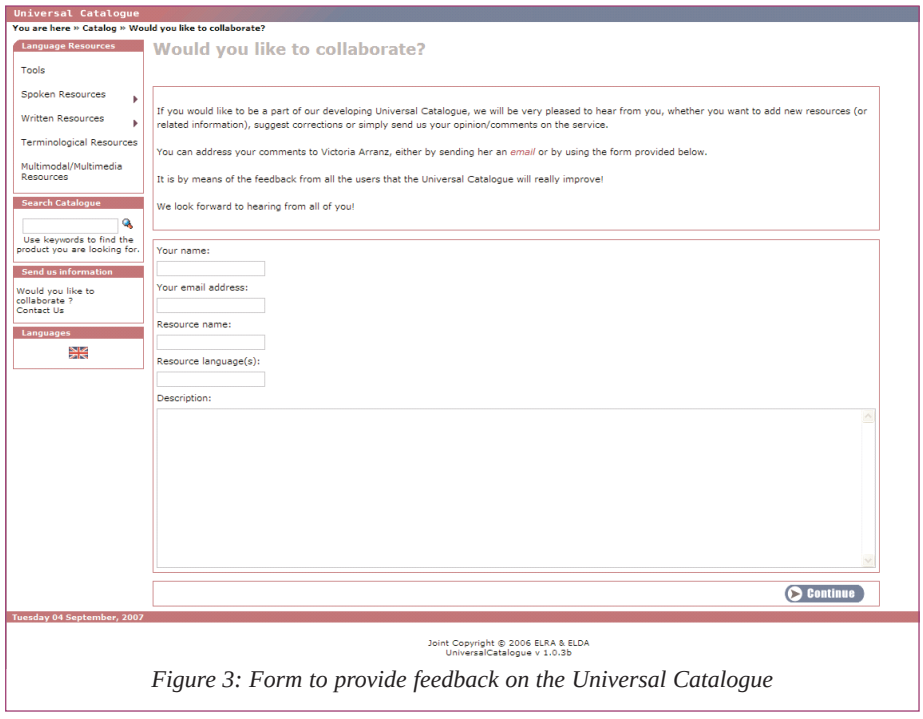


Figure 3: Form to provide feedback on the Universal Catalogue

LRs Distribution: Licensing and Pricing Policy

Licensing (Generic contracts)

ELRA is a useful conduit for the distribution of Language Resources (LRs) that are produced at several production sites and channels, enabling HLT players to have access to those LRs. In order to effectively provide such resources to research and development groups in academic, commercial and industrial environments, it is necessary to address a number of issues related to distribution activities, in particular licensing issues and pricing policy.

Simplify the relationship between producers/providers and users of LRs

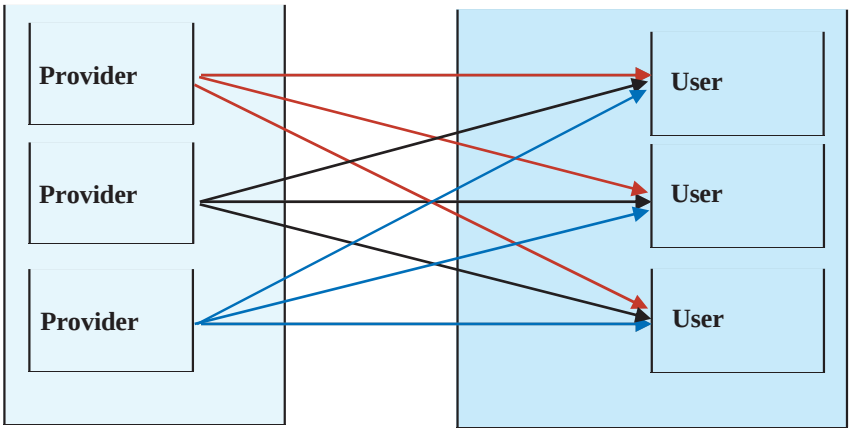


Figure 1: Relationship between producers/providers and users without ELRA

The basic principles of Language Resource licensing have been worked out with the support of lawyers. At the beginning, marketing Language Resources was a new activity, and creating an equitable and balanced framework was not easy. It was agreed that one of the priority tasks of ELRA was to simplify the relationship between producers/providers and users of LRs (illustrated in Figure 1).

In order to encourage producers and/or providers of LRs to make such data available to others, ELRA has drafted generic contracts defining the responsibilities and obligations of both parties.

To minimize variations in agreements and to keep things simple, these contracts are based on the model shown in Figure 2.

This model takes into account the interest of both parties (owners/producers and users) in keeping ELRA's role as a neutral, non-profit making organization, dedicated to promoting the Human Language Technology field.

Let us illustrate the role of ELRA with the MLCC Multilingual and Parallel Corpora, referenced as ELRA-W0023, the first set of which consists of 6 written corpora of similar nature, provided by six different newspapers through Europe (Le Monde from France, Financial Times from UK, Handelsblatt from Germany, Expansion from Spain, Il Sole 24 Ore from Italy, and Het Financieele Dagblad from The Netherlands). ELRA has signed one contract with each provider. If this resource is purchased via ELRA, the customer needs to sign one agreement. If the customer rather chooses to go to each individual provider, he/she needs to sign 6 licenses in 6 different judicial systems and will probably have to pay at least 6 different lawyers plus his own!

Distribution Agreements are drawn up between the resource provider and ELRA, while User licenses are drawn up between ELDA, on behalf of ELRA, and the resource user (either a Value Added Reseller-VAR, or an End-User, or an Evaluation license). Since 1996, they have evolved in the light of feedback from our members, customers and resource providers.

ELRA considers the production and distribution of these licenses as one of its contributions to the development of LR brokerage, so the licenses are available on the Web (as copyrighted documents) and we encourage all actors to use them. One can get electronic copies from the ELRA Web site.

Distribution agreement between ELRA and LR providers

Contracts between ELRA and LR providers need to be concluded so that provider may grant distribution licenses to ELRA. In other words, the purpose of the contract is for the provider to supply the LRs and to receive payment, royalties or other compensation in return for any further sales made by ELRA on the provider's behalf. The contract lays down that the LRs must be delivered with any necessary documentation in a specified language. It also stipulates that ELRA is allowed to publicize the existence and availability of the LRs in its catalogues, and to reproduce, and duplicate them for distribution purposes in accordance with its marketing, distribution, and commercialization policy. Alternatively, these tasks can be undertaken by the provider (in this case the terms and conditions must be specified).

Through this distribution agreement, ELRA is entitled to distribute the LRs and to grant its users -i.e. members and customers- the right to use them, in full or in part, for the purposes defined in the agreement between ELRA and the provider, at the user's institution or site.

User licenses between ELRA and LR users

The contract between ELRA and the users grants the latter a non-exclusive, non transferable right to use, rework and build on the LRs for the purposes agreed upon bet-

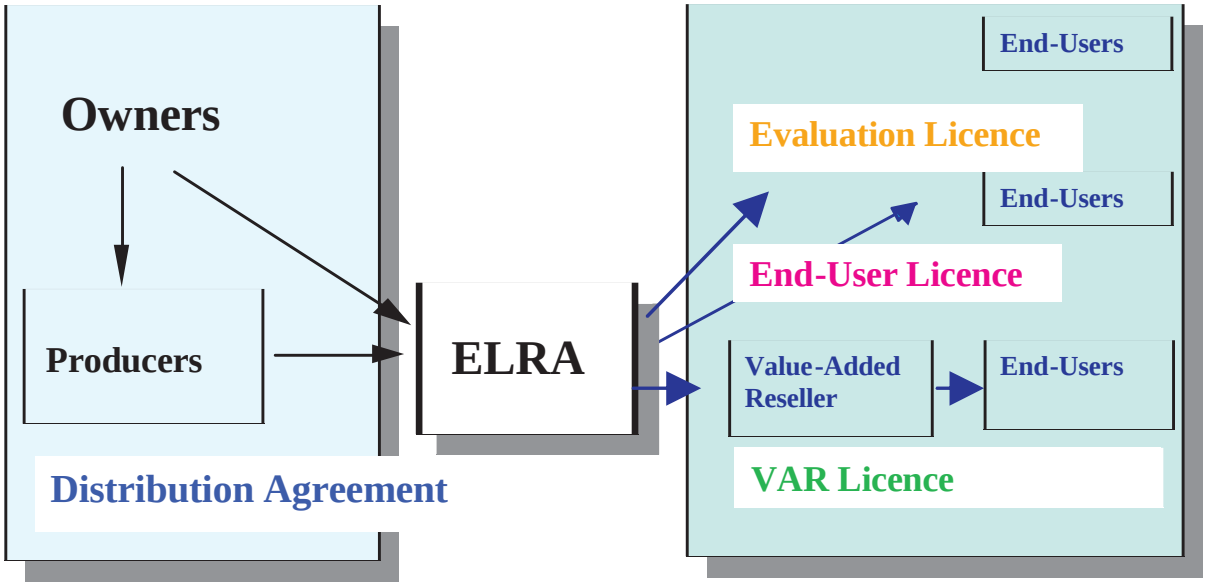


Figure 2: ELRA licensing model

ween the provider and ELRA within the user's institution. To this extent the user is allowed to create derivative works or software for his/her own internal research and development activities from the LRs or any component of them.

The agreements make no provision for the user to acquire any ownership, rights, title, or interest in the LRs. User acknowledges the right of the provider in the LRs and related materials including support documentation, and he will not infringe them in any way.

ELRA states that in all cases the user shall not copy or redistribute the LRs, although backup copies may be made. Any use of LRs by an affiliate, subsidiary, or other entity outside the user's place of business must be negotiated. In particular, the use of LRs or parts of them in any documentation, application or service which is charged for by the user, may be subject to a separate agreement. The contract may also state that the user has to obtain prior permission to sell or redistribute any product or derivative work based on the LRs.

A clause may stipulate that ELRA is authorized to reproduce and duplicate the LRs and any relevant documentation, in whole or in part, plus any summaries, compilations, or translations of the documentation for the purpose of distribution in accordance to the agreement. This has to be made explicit in the contract.

In general, LRs are provided on an "as is with all defects" basis. ELRA and the provider make no representation or warranties of any kind, either expressed or implied. In particular, all warranties for merchantability and/or fitness for a particular purpose are expressly excluded.

A detailed technical specification of the resources is given in a separate appendix to ensure clear identification and avoid subsequent differences of opinion.

Fees payable, which are determined in accordance with a schedule, are also attached as appendices. It is possible to update them yearly.

Agreements are entered into in good faith. If a dispute arises between the parties to the contract, they are bound to give notice of it in writing and try to reach an agreement. If no agreement is possible, the parties can either agree to arbitration or refer the matter to litigation.

One point that has to be clearly defined is what usage is allowed: some providers allow their resources to be used for research and technology/product development, while others only allow distribution for research purposes. As an answer to these different needs, three types of User Licenses were drafted:

- **End-User Agreement**: Within this Agreement, the user is engaged in *bona fide* language engineering research activities. The user is not permitted to distribute and market any derivative product or service based on all or a substantial part of the Language Resources.

- **Evaluation Packages End-User Agreement**: Within this Agreement ELRA grants the user the non-exclusive right to use the Evaluation Packages, exclusively for the purposes of evaluating their Human Language Technologies. The user is not permitted to reproduce the Evaluation Packages for commercial or distribution purposes and to commercialise (or distribute for free) in any form or by any means the Evaluation Packages or any derivative product or services based on all or a substantial part of it.

- **Value-Added Reseller Agreement (VAR)**: ELRA grants the user the non-exclusive right to distribute and market any derivative product or service based on all or a substantial part of the Language Resources (according to VAR's commercialization policies).

Pricing Policy

The pricing policy is also a crucial issue that needed careful attention. This had to take into account the fact that ELRA was establishing a new market in which LRs should be traded like any other commodity, bearing in mind the requirements and restrictions imposed by the provider (or the producer) when it comes to the issue of financial compensation. ELRA's approach is to simplify the price-setting, to clarify possible uses of LRs, and to reduce the restrictions imposed by the producer.

The prerequisite of acting as a broker is that each purchase renders a payment, covering the compensation claimed by the owner of the resource. In general, ELRA is not the owner of the resources, and can therefore only set a fair price in co-operation with the owner. This co-operation in setting the price is often based on conventional pricing methods like production costs. The pricing must also take into account ELRA's distribution policy, which is always to try and offer a discounted price to its members.

In some cases, providers accept to have their resources distributed for free. This is sometimes the case when production of LRs is already financed by the European Commission or by national governments. When browsing the catalogue, you will notice that the ELRA members benefit from price reductions which may go up to 70% on the public price. Exceptionally, ELRA is able to offer price reductions even if these are not financially supported by the providers. This is one of the services offered to our members, proving that ELRA is unique in its way of offering services and distributing LRs. The restrictions on the distribution, sometimes imposed by providers, are generally of two kinds: it is either a restriction on the user profile or a restriction on the usage. Providers may limit the distribution to members only, or they may restrict the use of their resource to research at large or even to academic research. When the restrictions are connected with the type of use, the reason is often that providers do not want their resource to be used in technical (commercial) development.

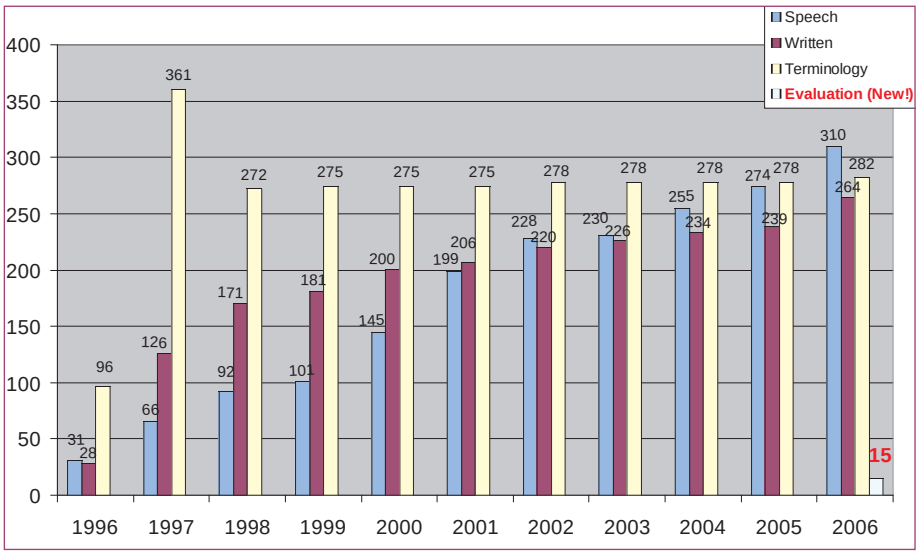
To sum up the different types of pricing offered by ELRA, let us use a real example extracted from the ELRA catalogue. Figure 1 presents the different prices offered for the NEMLAR Written Corpus. Two distinct price blocks are shown, one dedicated to ELRA members and another one to non members. Each block distinguishes the different types of use (research and commercial use) according to the type of user organisation (academic or commercial organisation). It happens that some resources may offer complementary prices, most of the time for bundle purchases. This can be seen on the right-hand side of the same example, where discounts ("Special Prices") available for the NEMLAR Written Corpus are highlighted. In this specific case, the discounts stand for a combined purchase of resources coming from the same project (NEMLAR).

Another type of pricing was added recently to the current offer. ELRA included "Evaluation Packages" in its catalogue., which comprise not only Language Resources but also protocols, methodology, tools, etc. that may be used to evaluate Language Technologies. For these Evaluation Packages, a different user license has been created by ELRA in order to limit the use of the packages for the purposes of evaluating Human Language Technologies. Prices are then considered differently depending on whether they are offered for evaluation, research or commercial use. Figure 2 presents the CLEF Test Suite for which only Evaluation licenses are proposed.

Figure 1: Examples of special prices

Figure 2: Examples of pricing policy for Evaluation Packages

THE ELRA CATALOGUE OF LANGUAGES RESOURCES



Graph 1: Distribution of Language Resources available in the ELRA Catalogue

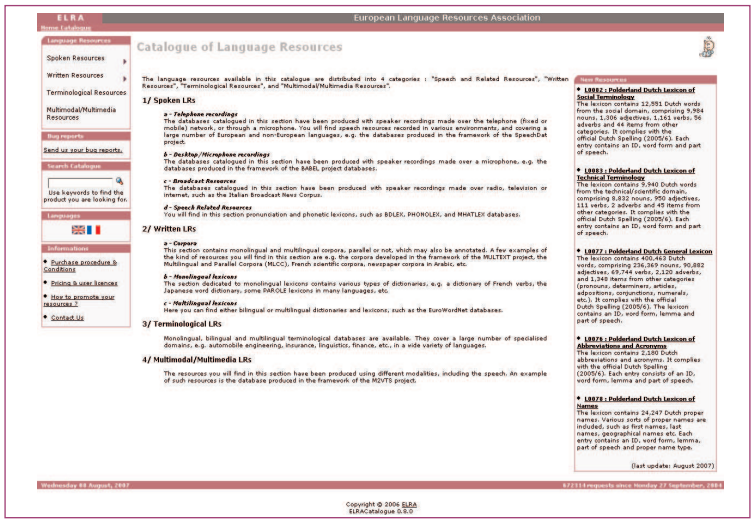
After 12 years of activity, ELRA has managed to make available, worldwide, a large set of marketable resources. Since October 1996, over 200 agreements with providers of language resources have been secured by ELRA, saving both users' and providers' time on numerous contractual agreement negotiations. In order to add value to the resources it distributes, ELRA also initiated the production of validation manuals for each resource type, namely spoken resources, written corpora, and lexica.

The increase in the number of resources available at ELRA over the years is illustrated in Graph 1.

To distribute this large range of resources and make them more visible worldwide, an electronic support was needed. First available as a set of static web pages, the ELRA Catalogue is now dynamic, based on a MySQL database, so as to improve browsing and retrieving of information within the catalogue of LRs. The redesign of the ELRA catalogue of LRs has benefited from the work done within the INTERA project (funded by the European Commission[1]). One of the aims of this project was to adopt a common set of metadata for the description of LRs. This would allow to harmonize the information available from the different data centers, thus focusing on the compatibility with the international standards. As a consequence, ELRA has reviewed its own system, working on its LRs description forms, and developing a new metadata set for the description of LRs. Figure 1 illustrates the latest MySQL-based ELRA catalogue of Language Resources.

Language Resources may be used, for instance, in the following two different ways: system development and system evaluation:

- *System Development*: spoken or written language processing systems are generally based on the use of corpora. For example, the performance of the systems available for text retrieval and filtering or machine-assisted translation tools mainly depends



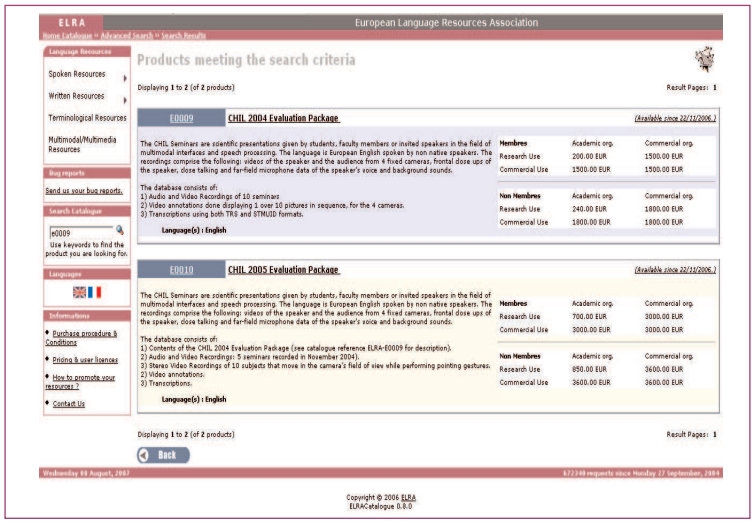


Figure 1: Screenshots from the ELRA Catalogue of LRs

(1) http://www.elda.org/rubrique22.html

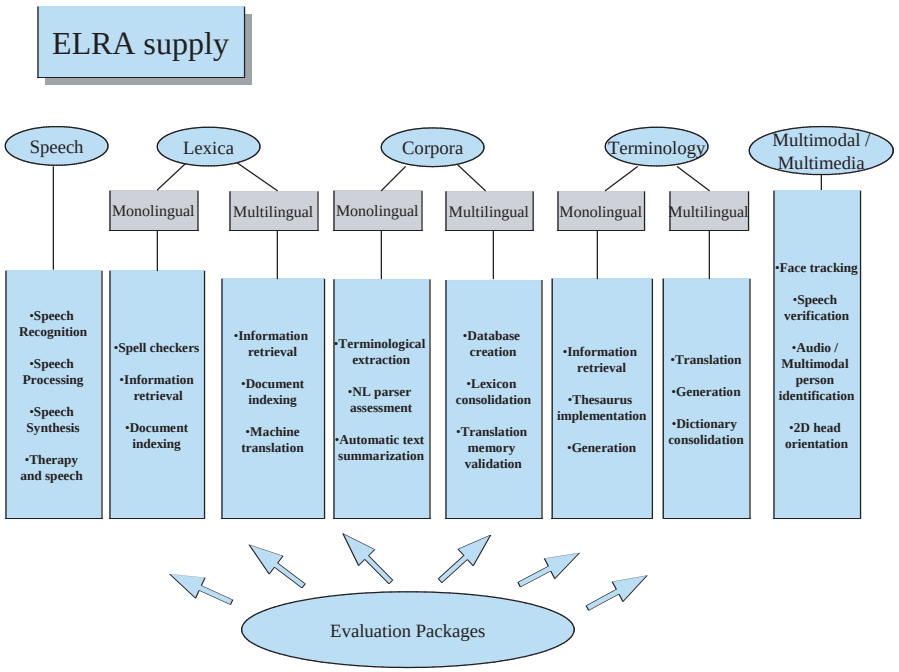


Figure 2: Distribution of ELRA supply of Language Resources and related applications

on the amount of linguistic data that are to be used to train the system. Corpora also allow you to build indirect language resources, i.e. specialised lexicons based on a group of technical texts.

• *System Evaluation*: Language Resources, e.g. large-size corpora, are also used to evaluate and compare systems which have already been developed. This is so for technologies, such as information filtering, orthographic and grammatical check, text retrieval, etc. Efficient and useful evaluations, which are based on appropriate and large corpora, are particularly important to measure the changes and progress made, and to disseminate and increase the value of the search results.

Figure 2 illustrates a wide range of applications of the Language Resources supplied by ELRA, as organised according to the different catalogue types (speech, written, terminology and multimodal/multimedia).

For an easier browsing in the catalogue, the Language Resources are distributed into four categories: "Speech and Related Resources", "Written Resources", "Terminological Resources", and "Multimodal/Multimedia Resources". Each category, along with sub-categories, when relevant, are described hereafter:

- Speech and Related LRs:

• *Telephone recordings*:

The databases catalogued in this section have been produced with speaker recordings made over the telephone (fixed or mobile) network. Speech resources recorded in various environments and covering a large number of European and non-European languages can be found here, e.g. the databases produced in the framework of the SpeechDat project.

• *Desktop/Microphone recordings*:

The databases catalogued in this section have been produced with speaker recordings made over a microphone, e.g. the databases produced in the framework of the BABEL project.

• *Broadcast Resources*:

The databases catalogued in this section have been produced with speaker recordings made over radio, television or internet, such as the Italian Broadcast News Corpus.

• *Speech Related Resources*:

This section comprises pronunciation and phonetic lexicons, such as BDLEX, PHONOLEX, and MHATLEX databases.

• *AURORA*: A specific section was created for the offer of the AURORA Project Database series, which is intended for the evaluation of algorithms for front-end feature extraction algorithms in background noise but may also be used more widely by speech researchers to evaluate and compare the performance of noise robust speech recognition algorithms.

- Written LRs:

• *Corpora*:

This section contains monolingual and multilingual corpora, whether parallel or not, which may also be annotated. Some examples of the kind of resources that can be found in this section are the corpora developed within the framework of the MULTEXT project, the Multilingual and Parallel Corpora (MLCC), French scientific corpora, newspaper corpora in Arabic, etc.

• *Monolingual lexicons*:

The section dedicated to monolingual lexicons contains various types of dictionaries, e.g. a dictionary of French verbs, the Japanese word dictionary, some PAROLE lexicons in many languages, etc.

• *Multilingual lexicons*:

Both bilingual and multilingual dictionaries and lexicons can be found here, such as the EuroWordNet databases.

- Terminological LRs: Monolingual, bilingual and multilingual terminological databases are available. They cover a large number of specialised domains, e.g. automobile engineering, insurance, linguistics, finance, in a wide variety of languages.

- Multimodal/Multimedia LRs: These resources have been produced using different modalities, including speech, hand gestures, facial expressions, etc. An example of such resources is the database produced within the framework of the M2VTS project.

Visit the ELRA Catalogue of Language Resources:

http://catalogue.elra.info

Focus on new resources: Evaluation Packages

If at the very beginning the main activity of ELRA & ELDA in the framework of the evaluation task was to supply Language Resources appropriate for test and evaluation, both are now getting involved in the evaluation process itself, the evaluation of products, systems, and applications developed for HLT.

Evaluation has become a major activity in the field of HLT. This activity is highly critical, as its main objectives are to check the quality of the developed applications and systems, and ensure that these are ready for the market. Evaluating a specific technology means measuring the progress achieved, comparing different approaches to a given problem, and choosing the best solution, assuming that its advantages and disadvantages have been analysed. Evaluation also involves the assessment of the availability of technologies for a given application, product benchmarking, and assessment of system usability and user satisfaction.

ELDA, acting on behalf of ELRA, actively participates in evaluation projects, at French, European and world levels. As a natural consequence of this participation, ELRA worked on making the results of those projects more visible and exploitable by the whole community.

Indeed, several sets of resources and tools, developed within these evaluation projects, are now available in our catalogue. Furthermore, as we are getting more and more involved in the evaluation activity, ELRA will continue to include further resources and tools related to evaluation in its catalogue, and will be soon adding a new section of the catalogue dedicated to this new offer.

Below are provided some key Evaluation Packages which are already available in the catalogue:

Amaryllis Corpus - Evaluation Package (Catalogue reference: ELRA-W0029)

Launched at the end of 1995, the AMARYLLIS project aimed at evaluating information retrieval software for French text documents (corpora and questions and answers) in order to provide a common methodology for the evaluation of other similar tools. AMARYLLIS was organised by the *Institut de l'Information Scientifique et Technique (INIST)* with the support of the *Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF)* and the French *Ministère de l'Education Nationale, de la Recherche et de la Technologie (MERT)*. Its goal was to create document corpora, questions and answers, in the framework of the *Action de Recherche Concertée* (ARC A1, renamed as Amaryllis- Access to text information in French), in order to get similar works to those obtained by the United States TREC project. All corpora are structured as SGML files with isolatin character-encoding.

The CLEF Test Suite for the CLEF 2000-2003 Campaigns (Catalogue reference: ELRA-E0008)

The main task of the Cross-Language Evaluation Forum (CLEF) consists of providing an infrastructure for the evaluation of information retrieval systems that operate on European languages, in multilingual, monolingual, and cross-language contexts. The CLEF Test Suite contains the data used for the main tracks of the CLEF campaigns carried out from 2000 to 2003: Multilingual text retrieval, Bilingual text retrieval, Monolingual text retrieval, and Domain-specific text retrieval.

The AURORA Project Database series (Catalogue references: AURORA/CD0002, /CD0003, /CD0004)

The Aurora project aimed at establishing a worldwide standard for feature extraction software in a DSR (Distributed Speech Recognition) system: evaluation of algorithms for front-end feature extraction in background noise on the one hand; on the other hand, the evaluation and comparison of the performance of noise-robust speech recognition algorithms.

The TC-STAR Evaluation Packages (Catalogue references: ELRA-E0002 to E0007, ELRA-E0011 to E0016)

TC-STAR is a European-integrated project focusing on Speech-to-Speech Translation (SST). To encourage significant breakthrough in all SST technologies, annual open competitive evaluations are organized. Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text-To-Speech (TTS) are evaluated independently and within an end-to-end system. Each evaluation package includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the evaluation campaigns. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself. The TC-STAR Evaluation Packages currently available are dedicated to the evaluation of Automatic Speech Recognition (ASR) and Spoken Language Translation (SLT).

The Technolangue Evaluation Packages (Catalogue references: ELRA-E0018, E0019, E0020, E0021, E0022, E0023, E0024)

The Evaluation Packages resulting from the following projects: ARCADE II, CESART, CESTA, ESTER, EQueR, EvaSy and MEDIA, were produced within the French national research programme Technolangue funded by the French Ministry of Research and New Technologies (MRNT). The technologies that could be evaluated within these projects are as follows: Speech synthesis, Transcription of Broadcast News, Dialogue Systems (for tourist information-oriented servers), Alignment of Multilingual Corpora, Machine Translation, Acquisition of Terminological Resources, Question-Answering for Information Retrieval. Each package includes the material that was used and/or produced for each evaluation campaign: resources, protocols and metrics, scoring tools, results of the campaign, etc. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results against the state of the art.

The CHIL Evaluation Packages (Catalogue references: ELRA-E0009 and E0010)

The CHIL Evaluation Packages were produced within the European-integrated CHIL Project (Computers in the Human Interaction Loop), financed under the European Commission Sixth Framework Programme. The objective of the project is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. Instead of computers operating in an isolated manner, and Humans [thrust] in the loop [of computers] we will put Computers in the Human Interaction Loop (CHIL).

R&D Catalogue - Language Resources with favourable conditions for R&D

Considering the needs expressed by several academic institutions of the Human Language Technology field, ELRA offers access to a version of its Catalogue of Language Resources dedicated to academic research. Indeed, at various occasions, while discussing with the players of the R&D academic community, we concluded that it was important to allow an easy and fast access to a list of resources more specifically produced for R&D purposes in Human Language Technology.

Thus, we now provide a list of Language Resources, available at very affordable prices, and dedicated to a research use. Like the full version of the catalogue, the Language Resources available here are distributed into four categories: "Speech and Related Resources", "Written Resources", "Terminological Resources", and "Multimodal/Multimedia Resources". So as to facilitate the access to this list, we preserved the interface and browsing tools of the ELRA Catalogue. Of course, at any time, one may choose to return to the full version of the catalogue.

Visit the R&D Catalogue: **http://catalogue.elra.info/retd**

VALIDATION SERVICES

Language Resources (LRs) must offer a certain degree of quality to be useful and suitable to the largest community. Validating LRs consists of checking their quality against several standardised or best practice criteria. According to its statutes where validation is mentioned as one of its goals, ELRA takes care of the validation of the LRs available in its catalogue.

ELRA Validation Committee (VCom) and Validation Centres (VCs)

In order to ensure a qualitative distribution, Language Resources (LRs) must be subject to quality control and validation. The term "validation" in ELRA is used in reference to the activity of checking the compliance to standards, and the quality control of the LR product. The contribution of ELRA can be seen as a validation of existing and newly developed resources and documentation of the results in the catalogue.

To perform this task, *a validation committee*, *VCom*, initiated by Harald Höge, was set up by the ELRA Board, on 23rd October 2000. At that time, the founding members of VCom were members of the ELRA Board as well. Since then, VCom has been open to non Board members.

The aim of VCom is to maximize the "ease of use" and "suitability" of the LRs which may be needed for HLT-systems, such as speech recognition, character recognition or information retrieval systems. For promoting "ease of use", VCom pushes forward the quality of LRs, i.e. by providing Language Resources with optimal documentation and minimal errors. For promoting "suitability", VCom supports standards and best practices for LRs leading to the best performance of state-of-the-art HLT-systems.

Due to the generic work of VCom, most issues are usually handled by all VCom members. At the beginning of VCom, ELRA also established two *Validation Centres (VCs)*, one for Spoken Language Resources (SLRs) and one for Written Language Resources (WLRs). Those VCs are controlled and co-ordinated by VCom and the ELRA Board. The procedures of validation are supported by validation committees linked to the Board, on the one hand, and by the VCs themselves, on the other hand. ELDA (http://www.elra.info/services/operationalbody.html), as the operational body of ELRA, implements services and supports strategic tasks as defined by VCom.

The work of ELRA related to validation applies to each of the two areas of activity:

- Speech Language Resource Validation: SLR validation covers the quality evaluation of a database against a checklist of relevant criteria. These criteria are typically the specifications of the databases, together with some tolerance margins for deviations.
- Written Language Resources Validation: Aiming to fulfil its objectives regarding the production of validation manuals, the ELRA VC for WLR is working in close co-operation with highly recognised research centres in order to produce such manuals. The work being carried out capitalises on previous projects including, but not limited to, EAGLES, Parole, Simple, Multext, and previous work done by ELRA.

Currently, the validation centres appointed by ELRA are:
- For Speech LRs: SPEX, SPeech EXpertise centre, The Netherlands (http://www.spex.nl/validationcentre/).
- For Written LRs: CST, Center for Sprogteknologi, Denmark (http://cst.dk/validation/index.html). CST cooperates with a network of expert centres in Europe.

VCom's tasks are multiple and can be itemised as follows:
- Define and supervise tasks performed by the operational units.
- Establish and optimize the organisation of VCom.
- Propose and control the budget of VCom.

- Prepare contracts with operational units.
- Report to the ELRA Board.

As a result of VCom's tasks and decisions, the tasks of the Operational units (validation centres and ELDA) are:
- Describe the quality of existing LRs.
- Improve the quality of existing LRs.
- Communicate with users and producers of LRs.
- Promote standards and best practices.
- Maintain the ELRA web pages concerning validation, according to the progress achieved within VCom.
- Maintain their LR validation portal.

Quick Quality Checks (QQCs) and Full Validation reports

To validate any type of LR, it is necessary to follow a strict procedure conformant with the LR production procedures in practice in the HLT world. Procedures were defined thanks to the work of ELRA's Validation Centres and put down in validation manuals, publicly available through its web site. Those documents describe validation principles that should be taken into account by producers of new LRs, specifically those who aim to distribute LRs through ELRA. Subsequently, ELRA's work has been to convince producers to adopt those procedures as a means of adding to the marketability of their products.

As a first step, a full validation protocol was implemented. This was mainly followed in an already agreed upon production protocol, which involved both the Validation Center and the database producer in an interactive way all along the production project. At the end of this procedure, a full validation report was produced. Validation checks typically include the following elements of a SLR:

- Documentation: correctness and clarity.
- Formats: directory structure and formats, and names of files.
- Design: completeness of recordings.
- Speech files: quality in terms of clipping, SNR, etc.
- Lexicon: completeness and correctness of formats and transcriptions.
- Speakers: realistic distributions over gender, age, accents.
- Recording environments.
- Orthographic transcriptions: format and correctness.

Many other LRs in the catalogue were not subjected to such an extensive (external) validation scenario. Since extensive validations are time-consuming and costly, VCom instructed SPEX to develop a method for a quick validation of a database. As a result, SPEX introduced the Quick Quality Check (QQC) for Speech LRs, whose methodology was also adapted to Written LRs as a second step.

As a starting point for the Quick Quality Check, two principles are taken:

1. The QQC mainly checks the database contents against its documentation. The main purpose of a QQC is to check if the documentation of the LR gives a correct account of the contents of the LR, in other words, if the LR meets the internal standards set up in the documentation.

2. Generally, the QQC of a LR should not take more than half a day's work.

The topics checked in a QQC are basically the same as those in the list of validation elements presented above. The crucial difference with a full validation is that a QQC only comprises a number of formal checks to see if the database contains what the documentation promises. There are no checks on the contents, that is, the correctness of, say, for SLRs, orthographic and phonemic transcriptions. The report is concluded by a brief advice to the producer from the Validation Centres.

To this date, over 50 LRs have undergone a QQC (Speech LRs, Monolingual and Multilingual Lexica) and over 140 LRs of the ELRA catalogue are provided with a full validation report.

Bug Reporting

The checks carried out by the Validation Centers are one method of finding errors in a LR. The users play an important role in detecting and reporting the bugs they find when actually using the LRs. To streamline the process, ELRA activated a bug report service. At present, this service is open only to users of Speech LRs. Those users who identified errors (bugs) in a database can report these via the ELRA web site through the Bug report section by clicking on:

Bug Report Service

There, a simple form is proposed in which users can add their comments and report bugs in the most detailed manner. This form looks as shown in Figure 1.
Based on the bug report received, once a substantial number of errors has been collected for a particular LR, ELRA will create a correction tool (patch) for users to update the LR concerned by removing the reported errors.

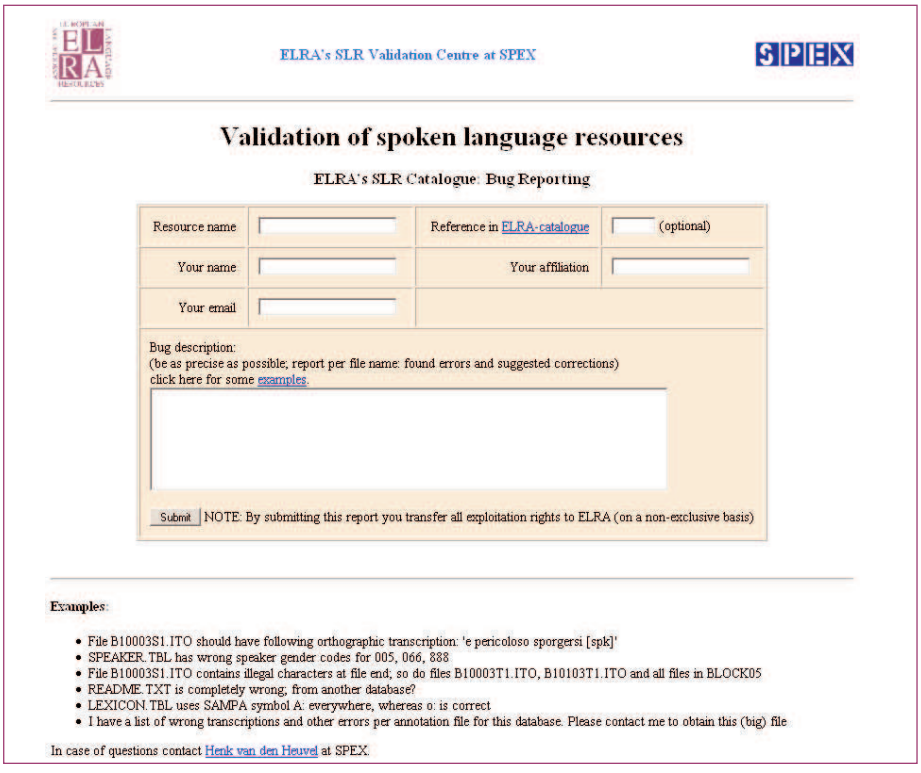


Figure 1: Online bug reporting form

PRODUCTION OF LRS

The production of new Language Resources is a major activity of ELRA and its operational body ELDA. The ELRA Board agreed to support this activity with internal funding. This production activity is achieved in several projects dedicated to the production of LRs but also in the framework of the evaluation projects in which ELDA is involved.

ELRA Production Committee (PCom)

The aim of the ELRA Production Committee, *PCom*, is to promote and support the production of Language Resources (LRs) needed for commercial and research use. Those activities are based on a business plan which takes into account a Return On Investment (ROI) analysis. Furthermore, PCom supervises the implementation of production centres and partnerships.

In order to achieve this goal, PCom identifies the needs of LRs both for commercial/industrial and research activities and in relation with the BLARK (Basic Language Resources Kit) / ELARK (Extended Language Resources Kit) concept. The BLARK concept was defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) and first launched with a Dutch initiative called Dutch Human Language Technologies Platform that was initiated in April 1999. Then, in the framework of the ENABLER thematic network (European National Activities for Basic Language Resources -Action Line : IST-2000-3.5.1), ELDA elaborated a report defining a (minimal) set of LRs to be made available for as many languages as possible and mapping the actual gaps that should be filled in, so as to meet the needs of the HLT field. A good illustration of the BLARK concept is the work done to gather information about Arabic resources within the NEMLAR project. The BLARK concept may be extended to many other languages, which is made feasible thanks to the online matrices implemented on the BLARK web site: http://www.elda.org/blark.

Based on this, PCom aims to provide a framework or even different possible scenarios for the LR production, promoting in particular new initiatives for producing LRs with a strategic impact.

In addition to this, several joint tasks are being carried out between PCom and VCom. The first of these tasks aims to collect, optimize, develop and promote standards and best practices for LR specifications. Another task is to manage the production of LR patches, based on the Formal Error Lists (FELs) provided by VCom.

First mission: Unified Lexicon Project

PCom started a test case called Unified Lexicon Project, using the Italian LC-Star and PAROLE lexicons, as a proof-of-concept of the feasibility of mapping the two tagsets and of unifying the entries coming from the two lexica. The ILC CNR, in Pisa, Italy, was the member organisation who implemented this task. Below is given an overview from the ILC CNR report.

This task falls within ELRA's PCom interests in promoting the development of techniques for merging two different lexica to generate a fused unified lexicon. The objective here is the merging of the LC-Star and Parole/Simple/Clips (henceforth Parole) Italian lexica and the mapping of their respective morpho-syntactic categorization systems. The overall aim, by exploring what is in common between the two, is to create a Unified Lexicon and propose a set of morpho-syntactic specifications common to spoken and written Italian lexica.

The mapping was performed on a common sample of about 1,000 lexical entries extracted from the two lexicons. The two categorization systems were mapped not only at a label level, but also via a careful checking of how these labels are applied: for each label, it has been investigated which word forms it applies to. In this way, prospective differences in application criteria may also emerge.

The benefit of this work lies on the possibility of importing complementary information present in the two source lexica (e.g. inflectional codes can be obtained "for grant" from PAROLE) into the new unified resource. Another advantage for LC-Star can be to hook further layers of the PAROLE lexicon and access syntactic and semantic lexical information.

The Unified Lexicon Project can be considered as an example of promoting new ways of creating and distributing LRs. It is also a way of promoting common standards shared by the Spoken and Written communities, both based on EAGLES.

Production services by ELRA's operational body, ELDA

Strengthened by its experience and knowledge in Language Resources, ELDA offers production services available to everyone interested in Human Language Technologies.

You may contact us to produce new LRs to help you build, improve or evaluate natural language and speech algorithms or systems. Those Language Resources may be also used as core resources for the software localisation and language services industries, language studies, electronic publishing, international transactions, subject-area specialists and end users.

ELDA benefits from a large network of partners all over the world and can provide Language Resources in various languages. ELDA has already compiled Language Resources in more than 25 languages. Besides, we offer the highest quality of resources through a strict validation procedure.

Production Services

Services

ELDA is involved in every stage of the production of Language Resources :

• ***Speech/Video Data Collection:***

- Management of the data collection team and recruitment of speakers all over the world
- Collections of recordings in different scenarios: in cars, public places, offices, etc.
- Transcriptions and annotations of audio and video data
- Validation

• ***Written Data Collection (corpora and lexica*)**:

- Preparation of raw data
- Annotations
- Validation

• ***Data creation for specific technologies and/or evaluation campaigns (see also the HLT Evaluation Portal maintained by ELRA: www.hlt-evaluation.org)*:**

- Machine translation
- Speech language translation
- Automatic summarization
- Question Answering
- Automatic Speech Recognition (ASR)
- Spoken Language Translation (SLT)
- Text To Speech synthesis (TTS)

Customers & Partners

Our services target all types of organisations or consortia willing to commission the production of Language Resources. They can be categorized as follows:

• ***Academic or private organizations needing to develop linguistic tools or consumer products*:**

The resources produced in the framework of a contract with individual organisations may become available in the ELRA catalogue. We offer special prices to customers who agree to let us distribute widely the resources produced within their own projects.

• ***Consortia participating in cooperation projects*:**

We participate in many consortia whose aim is to create Language Resources. These consortia may be a part of EU-funded projects, national projects or International cooperation projects. Speech and Written LRs are also produced within some of the evaluation campaigns of the EVALDA platform of the Technolangue program; ELDA has launched the creation of new LRs, mainly for the French language. This is the case for the Media and Ester campaigns, with the production of transcribed and not-transcribed speech databases, or the Cesta campaign, with the production of text corpora in Arabic and French. The same process is implemented for EU projects such as CHIL, TC-STAR or CLEF.

• ***Commissioning the production of LRs*:**

ELDA also commissions the production, packaging and customisation of Language Resources so as to take part in the development of the HLT market and thus, offer the Language Resources needed by the HLT community.

Language Resources and Evaluation Journal

The Language Resources and Evaluation Journal is the first publication devoted to the acquisition, creation, annotation, and use of language resources, together with methods for evaluation of resources, technologies, and applications. It is published by Springer. Nicoletta Calzolari, from ILC-CNR in Pisa (Italy) and Nancy Ide, from Vassar College in Poughkeepsie, NY (USA) are the editors in chief.

Hereafter is a quick overview of the latest issues (those not published yet appear in italics):

• Volume 40, no. 1, a regular issue dedicated to Data Resources, Evaluation and Dialogue Interaction, guest-edited by Laila Dybkjaer and Wolfgang Minker.

• Volume 40, no. 2, a regular issue edited by Nicoletta Calzolari and Nancy Ide.

• Volume 40, no. 3 & 4, a special issue on Asian Languages, guest-edited by Chu-Ren Huang (Academia Sinica, Taiwan) and Takenobu Tokunaga (Tokyo Institute of Technology, Japan).

• *Volume 41, no. 1, a regular issue edited by Nicoletta Calzolari and Nancy Ide.*

• *Volume 41, no. 2, a special issue on Asian Languages.*

• *Volume 41, no. 3, a special issue on Multimodal Corpora, guest-edited by Jean-Claude Martin et al.*

• *Volume 41, no. 4, a special issue on Multilingual Resources guest-edited by Gilles Sérasset and Andreas Witt.*

The ELRA members are granted complimentary access to the journal through the society on the condition that they subscribe to the publisher's table-of-contents alert service. During LREC 2006, special conditions were offered to participants and in anticipation of LREC 2008, the Editors will try to arrange with Springer some special conditions for subscription to the LRE Journal to be offered to conference participants.

After LREC 2006 in Genoa, the flow of submissions to the Journal sensibly raised and it looked like the interest for the Journal spread worldwide thanks to the visibility and promotion given by the conference.

Editorial Board

• **Editor-in-Chief:**

Nicoletta Calzolari, Istituto di Linguistica Computazionale, CNR, Pisa, Italy

Nancy Ide, Dept. of Computer Science, Vassar College, Poughkeepsie, NY, USA

• **Assistant Editor:**

Sara Goggi, Istituto di Linguistica Computazionale, CNR, Pisa, Italy

• **Book Review Editor:**

Alessandro Lenci, Istituto di Linguistica Computazionale, CNR, Pisa, Italy

• **Advisory Board Members:**

Bente Maegaard, Center for Sprogteknologi, University of Copenhagen, Copenhagen, Denmark

Khalid Choukri, ELRA, Paris, France

Joseph Mariani, LIMSI-CNRS, Orsay, France

Jan Odijk, Nuance Belgium, Merelbeke, Belgium and Utrecht University, Utrecht, The Netherlands

LANGUAGE RESOURCES AND EVALUATION CONFERENCE

Overview

The Language Resources and Evaluation Conference is an international **scientific** event, which aims at providing an overview of the **state of the art**, exploring new **R&D** directions and emerging trends, and exchanging information regarding Language Resources and their applications, **evaluation methodologies** and tools, on-going and planned activities, industrial uses and needs, requirements coming from the **e-society**, both with respect to policy issues and to technological and organisational ones.

LREC provides a unique forum for researchers, industrials and funding agencies from across a wide spectrum of areas to discuss problems and opportunities, find **new synergies** and promote **initiatives** for international cooperation, in support to investigations in language sciences, progress in **language technologies** and development of corresponding products, services and applications, and standards.

The conference generally covers a full week and LREC's programme is organised around parallel oral and poster sessions during the main conference, and 4 days, before and after the conference, are dedicated to workshops and tutorials.

Previous editions

In 10 years, LREC, which is organized every other year, has become the major event on Language Resources and Evaluation for Human Language Technologies.

Here are the places where the last editions have taken place:

• **Granada (LREC 1998)**: the first Language Resources and Evaluation Conference (LREC) attracted over 510 attendees from 325 organisations in 38 countries.

• **Athens (LREC 2000)**: the second edition became a major event in the overall area of HLT with 600 participants.

• **Las Palmas (LREC 2002**): for the third edition, the number of registered participants increased to reach 730 participants.

• **Lisbon (LREC 2004)**: for the fourth edition, the number of participants increased significantly and reached 950.

• **Genoa (LREC 2006)**: around 800 participants from 44 countries attended the fifth edition.

The partners and sponsors

A number of associations from all over the world have been supporting LREC from the beginning. In 2006, LREC was organised in cooperation with a wide range of international associations and consortia, including AAMT, ACL, AFNLP, ALLC, ALTA, COCOSDA and Oriental COCOSDA, EACL, EAMT, ELSNET, ENABLER, EURALEX, Forum TAL, GWA, IAMT, ISCA, KnowledgeWeb, LDC, NEMLAR Network, SENSEVAL, SIGLEX, TEI, Technolangue French Program, WRITE and with major national and international organisations including the European Commission - Information Society and Media Directorate General, Unit "Interfaces".
In addition, each edition is also sponsored by large companies, such as Telefonica, IBM or Microsoft. The conference organisation also receives the support of local sponsors and local authorities, just like in Genoa, where the Liguria Region had provided the conference with computers.

The participants

Over the years, the number of participants to LREC has increased from 500 to 900. The participants of LREC come in their large majority from the scientific community of language processing. In 2006, nearly 85% of the participants came from universities, whether researchers or students. Industrials, in other words, those who develop and/or market the solutions, represent 10% of the participants.
Even if the majority of the participants come from Europe, and more specifically from Germany, Italy and France, on the whole, 45 countries are represented, out of which the United States and Japan bring a considerable quota of participants.
Last but not least, LREC is also very oriented towards national, European and international research projects. A certain number of sessions and workshops are devoted to them and decision-makers, in particular from funding agencies, take an active part in these sessions.

Antonio Zampolli Prize

The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language

Technologies through all issues related to Language Resources and Evaluation. In awarding the prize, ELRA is seeking to reward and encourage innovation and inventiveness in the development and use of language resources and evaluation of HLT.

The prize has been created by ELRA to honour the memory of its first President, Antonio Zampolli, who died in 2003. This 10,000€ prize is awarded every two years, within LREC. So far, it has been awarded twice to:

1. Fred Jelinek from John Hopkins University (USA), during LREC 2004 in Lisbon.

2. Christiane Fellbaum and George A. Miller, from Princeton University (USA), during LREC 2006 in Genoa.

LREC 2008

LREC 2008, the sixth edition of the Language Resources and Evaluation Conference, biennially organized by ELRA, will take place in Marrakech (Morocco) from May 26th to June 1st 2008.

Marrakech, the southernmost Imperial City, once the capital of the Saadian dynasty, is a unique place, perhaps the most fascinating in Morocco. It has also become a very attractive business destination hosting many international events, such as conferences and summits. In addition to impressive conference facilities, Marrakech proposes a huge accommodation offer and is very well served with international flights from and to Europe, America and Asia by both low-cost companies and regular services.

The main conference will take place on May 28, 29 and 30. Various themes such as design, construction and use of Language Resources, their exploitation and evaluation, but also the integration of different types of resources will be dealt with. Every day, a great number of parallel sessions, whether oral or poster, will be proposed to the participants. In addition, pre- and post-conference workshops and tutorials will be organised before (May 26-27) and after (May 31-June 1) the main conference.

Finally, the Antonio Zampolli Prize that rewards the Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation within Human Language Technologies will be awarded to the winner during the closing ceremony.

LREC 2008 Second Call for Papers

The sixth international conference on Language Resources and Evaluation (LREC) will be organised in 2008 by ELRA in cooperation with a wide range of international associations and organisations.

CONFERENCE AIMS

In 10 years (the first LREC was held in Granada in 1998), LREC has become the major event on Language Resources (LRs) and Evaluation for Human Language Technologies (HLT). The aim of LREC is to provide an overview of the state-of-the-art, explore new R&D directions and emerging trends, exchange information regarding LRs and their applications, evaluation methodologies and tools, ongoing and planned activities, industrial uses and needs, requirements coming from the e-society, both with respect to policy issues and to technological and organisational ones.

LREC provides a unique forum for researchers, industrials and funding agencies from across a wide spectrum of areas to discuss problems and opportunities, find new synergies and promote initiatives for international cooperation, in support to investigations in language sciences, progress in language technologies and development of corresponding products, services and applications, and standards.

CONFERENCE TOPICS

Issues in the design, construction and use of Language Resources (LRs): text, speech, multimodality

- Guidelines, standards, specifications, models and best practices for LRs
- Methodologies and tools for LRs construction and annotation
- Methodologies and tools for the extraction and acquisition of knowledge
- Ontologies and knowledge representation
- Terminology
- Integration between (multilingual) LRs, ontologies and Semantic Web technologies
- Metadata descriptions of LRs and metadata for semantic/content markup

Exploitation of LRs in different types of systems and applications

- For: information extraction, information retrieval, speech dictation, mobile communication, machine translation, summarisation, web services, semantic search, text mining, inferencing, reasoning, etc.
- In different types of interfaces: (speech-based) dialogue systems, natural language and multimodal/multisensorial interactions, voice activated services, etc.
- Communication with neighbouring fields of applications, e.g. e-government, e-culture, e-health, e-participation, mobile applications, etc.
- Industrial LRs requirements, user needs

Issues in HLT evaluation

- HLT Evaluation methodologies, protocols and measures
- Validation, quality assurance, evaluation of LRs
- Benchmarking of systems and products
- Usability evaluation of HLT-based user interfaces, interactions and dialog systems
- Usability and user satisfaction evaluation

General issues regarding LRs & Evaluation

- National and international activities and projects
- Priorities, perspectives, strategies in national and international policies for LRs
- Open architectures
- Organisational, economical and legal issues

Special Highlights

LREC targets the **integration of different types of LRs** - spoken, written, and other modalities - and of the respective communities. To this end, LREC encourages submissions covering issues which are common to different types of LRs and language technologies.

LRs are currently developed and deployed in a much wider range of applications and domains. LREC 2008 recognises the need to encompass all those data that interact with language resources in an attempt to model more complex human processes and develop more complex systems, and encourages submissions on topics such as:

- Multimodal and multimedia systems, for Human-Machine interfaces, Human-Human interactions, and content processing
- Resources for modelling language-related cognitive processes, including emotions
- Interaction/Association of language and perception data, also for robotic systems

PROGRAMME

The Scientific Programme will include invited talks, oral presentations, poster and demo presentations, and panels.

There is no difference in quality between oral and poster presentations. Only the appropriateness of the type of communication (more or less interactive) to the content of the paper will be considered.

ABSTRACT SUBMISSION

On-line submission form for abstracts is now available: please go to the "Abstract submission" section on the LREC 2008 web site:

(http://www.lrec-conf.org/lrec2008/Home.html) and follow the procedure instructions.

Submitted abstracts of papers for oral and poster or demo presentations should consist of about 1500-2000 words.

WORKSHOPS, TUTORIALS AND PANELS

Submission of workshop, tutorial and panel proposals should be made by e-mail to the following e-mail address:

lrec@lrec-conf.org

and will be reviewed by the Programme Committee.

Proposals for workshops and tutorials should be no longer than three pages, and include:

- **For workshops:**
  - The title
  - A brief technical description of the specific technical issues that the workshop will address
  - The reasons why the workshop is of interest
  - The names and affiliations, postal addresses, phone and fax numbers, email and web site addresses of the organising committee, which should consist of at least three people knowledgeable in the field, coming from different institutions
  - The name and the e-mail address of the member of the workshop organising committee designated as the contact person
  - The desirable duration of the workshop (half day or full day)
  - A summary of the intended call for participation
  - An estimate of the approximate audience size
  - A list of audio-visual or technical requirements and any special room requirements

  The workshop proposers will be responsible for the organisational aspects (e.g. workshop call preparation and distribution, review of papers, notification of acceptance, assembling of the workshop proceedings using the ELRA specifications, etc.).
- **For tutorials:**
  - The title
  - A brief technical description of the tutorial content
  - The reasons why the tutorial is of interest
  - The names and affiliations, postal addresses, phone and fax numbers, email and web site addresses of the tutorial speakers, with brief descriptions of their technical background
  - The name and e-mail address of one tutorial speaker designated as the contact person
  - The duration of the tutorial (half day is the expected usual length)
  - An estimate of the approximate audience size
  - A list of audio-visual or technical requirements and any special room requirements

  The tutorial proposers will be responsible for the organisational aspects (e.g. assembling of the tutorial material, etc.).

Proposals for panels should contain the following information:

- The title
- A brief technical description of the specific technical issues that the panel will address
- The reasons why the panel is of interest
- Name of the panel organiser/s; affiliation and postal address; phone and fax numbers; e-mail address; web site address
- The name and the e-mail address of the designated contact person

IMPORTANT DATES

- Submission of proposals for panels, workshops and tutorials: **31 October 2007**
- Submission of proposals for oral and poster/demo papers: **31 October 2007**
- Notification of acceptance of panels, workshops and tutorials proposals: **22 November 2007**
- Notification of acceptance of oral papers, posters: **4 February 2008**
- Final versions for the proceedings: **25 March 2008**
- Conference: **28-30 May 2008**
- Pre-conference workshops and tutorials: **26 and 27 May 2008**
- Post-conference workshops and tutorials: **31 May and 1 June 2008**

CONSORTIA AND PROJECT MEETINGS

Consortia or projects wishing to take this opportunity for organising meetings should contact well in advance the organisers at the following e-mail address:

lrec@lrec-conf.org

PROCEEDINGS

The **Proceedings on CD will include both oral and poster papers,** in the same format.

In addition, a **Book of Abstracts** will be printed.

CONFERENCE PROGRAMME COMMITTEE

Nicoletta Calzolari, Conference Chair, Istituto di Linguistica Computazionale del CNR, Pisa, Italy

Khalid Choukri, ELRA, Paris, France

Bente Maegaard, CST, University of Copenhagen, Denmark

Joseph Mariani, LIMSI-CNRS, Orsay, France

Jan Odijk, Nuance Communications International, Belgium and UIL-OTS, Utrecht, the Netherlands

Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athens, Greece

Daniel Tapias, Telefónica Móviles España, Madrid, Spain.

HLT-Evaluation.org: a Portal for Human Language Technology Evaluation

The general mission of HLT evaluation is to assist in improving the quality of language engineering products. It is essential for validating research hypotheses, for assessing progress and for choosing between research alternatives. The language technology and neighbouring technology communities range from academic to industrial partners who share an interest in evaluation.

Building a portal for these communities, which serves as a platform for communication between partners, and offers a permanent infrastructure for the development of evaluation activities in Europe is the main goal of the HLT Evaluation Portal.

The portal provides all kinds of information related to the evaluation for the language technology community, and also for the general public, and helps users who need to have quick and easily understandable information on evaluation protocols, including evaluation methodologies, metrics, evaluation tasks, resources and worldwide evaluation activities, such as research projects and campaigns. Moreover, permanent and generic protocols and packages for the major language technologies, such as machine translation, information extraction and retrieval, speech processing (including speech recognition, speech synthesis and speech translation) will be also provided.

The online HLT evaluation web portal, available at http://www.hlt-evaluation.org, is structured around three main sections:

- **Overview**
- **State of the art**: Information on HLT evaluation
- **Evaluation services**: Evaluation services offered by ELRA
- **Evaluation resources**: List of evaluation data and evaluation packages (e.g. methodologies, scoring software, test data)

State of the art: HLT evaluation information

Here, the portal provides a free information service to the R&D community, potential users of language technologies and other interested parties. The information and knowledge resources are organised as follows:

- The **Introduction** gives a general overview of HLT evaluation including the state of the art.
- Summary of methods, measurements and related events of evaluation by different HLT technologies including written language, speech and multimodal interfaces are presented in **HLT evaluations**.
- The **Activities by technology** section provides the list of evaluation events regarding campaigns, conferences and workshops including those explicitly focused on HLT evaluation and those offering specific sessions on evaluation.
- The **Activities by geographical region** section provides the list of national programs or projects focused on HLT evaluation or partially contributing to the development of the HLT evaluation.
- The **Players** section provides a non-exhaustive directory of HLT evaluation centres, organisations and institutions that are involved in evaluation activities and links to organisations which provide relevant resources such as corpora, lexicon and annotations for evaluation research and development.

Text processing	Speech processing	Multi-modal interface
Information Retrieval	Speech Synthesis	Multimodal Person Tracking
Question Answering	Speech Translation	Audiovisual Speech Recognition
Machine Translation	Automatic Speech Recognition	Multimodal Person Identification
Automatic Summarization	Broadcast News Transcription	
Parsing	Acoustic Person Tracking	
Multilingual Text Alignment	Acoustic Speaker Identification	
Terminology Extraction	Speech Activity Detection	

Table 1: Technologies evaluated by ELRA

HLT evaluation services

Evaluation services for each of the language technology listed in Table 1 are offered through the portal. Any institution or company wishing to evaluate its HLT-based system is welcome to contact ELRA (evaluation@elda.org).

HLT evaluation resources

The Resources section offers pointers to available commercial and research toolkits which allow one to perform comparative evaluation. A number of evaluation resources are now available from the ELRA catalogue. For example, Evaluation Packages have been produced within the French national research programme Technolangue and include resources, protocols and metrics, scoring tools, together with results of the campaign. These packages allow external players to evaluate their own system.

NEW RESOURCES

Evaluation Packages from the French national research programme Technolangue

The Evaluation Packages resulting from the following projects: ARCADE II, CESART, CESTA, ESTER, EQueR, EvaSy and MEDIA, are now available. These Evaluation Packages were produced within the French national research programme **Technolangue** funded by the French Ministry of Research and New Technologies (MRNT). Each package includes the material that was used and/or produced for each evaluation campaign: resources, protocols and metrics, scoring tools, results of the campaign, etc. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results against the state of the art.

The technologies that could be evaluated within this programme are as follows:

- Speech Technologies:

  - Speech synthesis:
  The **EvaSy** project enabled to carry out a campaign for the evaluation of speech synthesis systems using French text data. Three actions were considered during the campaign: evaluation of grapheme-to-phoneme conversion, evaluation of prosody, global evaluation of the quality of speech synthesis systems. The **EvaSy** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0023**.
  - Transcription of Broadcast News:
  The **ESTER** project enabled to carry out a campaign for the evaluation of speech recognition systems that produce Broadcast News enriched transcription for French. Three actions were considered during the campaign: orthographic transcription, segmentation and information extraction (named entity tracking). The **ESTER** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0021**. For research or commercial use of this database, please refer to **ELRA-S0241 ESTER Corpus**.
  - Speech Dialogue (for tourist information-oriented servers):
  The **MEDIA** project enables to carry out a campaign for the evaluation of man-machine dialogue systems for French. The campaign is distributed over two actions: an evaluation taking into account the dialogue context and an evaluation not taking into account the dialogue context. The **MEDIA** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0024**.

- Written Technologies:

  - Alignment of Multilingual Corpora:
  The **ARCADE II** project enables to carry out a campaign for the evaluation in the field of multilingual alignment. Two actions were considered during the campaign: sentence alignment and translation of named entities. The **ARCADE II** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0018**.
  - Machine Translation:
  The **CESTA** project enabled to carry out a campaign for the evaluation of machine translation technologies. Two actions were connsidered during the campaign: evaluation on a general vocabulary and evaluation on a specialised domain (evaluation after terminology enrichment). The **CESTA** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0020**.
  - Acquisition of Terminological Resources:
  The **CESART** project enabled to carry out a campaign for the evaluation of tools for the acquisition of terminological resources. Two actions were considered during the campaign: term extraction and extraction of relations between terms within text data. The **CESART** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0019**.
  - Question-Answering Task for Information Retrieval:
  The **EQueR** project enabled to carry out a campaign for the evaluation of Question-Answering systems in French. Two actions were considered during the campaign: one generic task and one specialised task (medical domain). The **EQueR** Evaluation Package is available in the ELRA catalogue under reference **ELRA-E0022**.
  - Syntactic Parsers:
  The **EASy** project enabled to carry out a campaign for the evaluation of syntactic parsers. Two actions were considered during the campaign: annotation within constituents and annotation between dependency relations.
  The Evaluation Package will be published soon…

ELRA takes the opportunity of this announcement to thank all scientific coordinators and participants who made this work possible.

About Technolangue

Technolangue is a French inter-ministerial-funded action (Ministry of Research and New Technologies, Ministry of Industry and Ministry of Culture and Communication). This programme provides access to a number of tools and data necessary to develop technologies and to enable the use of standard methodologies in the field. For more information on the Technolangue programme, please visit: http://www.technolangue.net.

For more information on the Technolangue Evaluation Packages, visit the ELRA Catalogue: **http://catalogue.elra.info**

Monolingual and Multilingual Lexicons from the general domain

ELRA-L0074 POLEX Polish Lexicon

The POLEX Polish Lexicon is a morphological dictionary of Polish language. It comprises about 100,000 entries. The POLEX dictionary includes the core Polish vocabulary of general interest. It is based on a precise machine-interpretable formalism (coding system), the same for all categories (classes of speech). The dictionary entries are of the following form:
BASIC_FORM+LIST_OF_STEMS+PARADIGMATIC_CODE+DISTRIBUTION_OF_STEMS
It contains more than 42,000 nouns, 12,000 verbs, 15,000 adjectives, 25,000 participles, and about 200 pronouns. A simple lemmatiser (in form of PROLOG prototype) is also included.

	ELRA members	Non-members
For research use	80 Euro	100 Euro
For commercial use	1,500 Euro	3,000 Euro

Special price for students willing to acquire the Language Resource on a personal basis for their research: 40 Euro

ELRA-L0075 Bulgarian Linguistic Database

The Bulgarian Linguistic Database contains 81,647 entries in Bulgarian with a linguistic environment tool (for WINDOWS XP). The data may be used for morphological analysis and synthesis, syntactic agreement checking, phonetic stress determining.

	ELRA members	Non-members
For research use	2,000 Euro	4,000 Euro
For commercial use	10,000 Euro	16,000 Euro

ELRA-M0038 English-German Bilingual Dictionary (SCI-AN-ALL)

The English-German Bilingual Dictionary (SCI-AN-ALL) contains 59,758 pairs of English-German terms, with their part of speech. The data are presented in a table format, where information related to each entry is separated by ";". See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036, ELRA-M0037.

	ELRA members	Non-members
For research use	1,000 Euro	1,800 Euro
For commercial use	8,000 Euro	11,000 Euro

Written Corpora from news agencies

ELRA-W0015 Text corpus of "Le Monde"

Years 2005 and 2006 from the Text Corpus of "Le Monde" are now available. Years 1987 to 2002 from the Text corpus from "Le Monde" newspaper are available in an ASCII text format. Years 2003 to 2006 are available in .XML format. Each month consists of some 10 MB of data (circa 120 MB per year).

	ELRA members	Non-members
For research use	240.91 Euro per year	313.18 Euro per year

ELRA-W0047 Catalan Corpus of News Articles

The Catalan Corpus of News Articles comprises articles in Catalan from 1 January 1999 to 31 March 2007. These articles are grouped per trimester without chronological order inside.

	ELRA members	Non-members
For research use	2,975 Euro	3,930 Euro
For commercial use	14,855 Euro	19,315 Euro

A Broadcast News Speech Corpus resulting from LDC (Linguistic Data Consortium, USA) and ELRA cooperation in the European-funded NetDC project

ELRA-S0157 NetDC Arabic BNSC (Broadcast News Speech Corpus)

The NetDC Arabic BNSC (Broadcast News Speech Corpus) is a corpus developed by ELDA in the framework of the European-funded project Network of Data Centres (NetDC). The project was done in collaboration with the LDC (Linguistic Data Consortium), which has produced a similar corpus from the news broadcasted by Voice of America Arabic in the United States. The database contains ca. 22.5 hours of broadcast news speech recorded from Radio Orient (France) during a 3-month period.

	ELRA members	Non-members
For research use	100 Euro	200 Euro
For commercial use	1,350 Euro	2,700 Euro

Phonetic lexicons from the European-funded LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) project

ELRA-S0229 LC-STAR Turkish phonetic lexicon

The LC-STAR Turkish lexicon comprises 104,513 words, including a set of 59,213 common words, a set of 45,300 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 7,498 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	21,250 Euro	27,625 Euro
For commercial use	28,000 Euro	36,400 Euro

ELRA-S0230 LC-STAR Russian phonetic lexicon

The LC-STAR Russian lexicon comprises about 128,000 words, including a set of 77,154 common words, a set of 51,074 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 12,012 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	23,750 Euro	30,875 Euro
For commercial use	30,000 Euro	39,000 Euro

ELRA-S0231 LC-STAR English-Russian Bilingual Aligned Phrasal lexicon

The LC-STAR English-Russian Bilingual Aligned Phrasal lexicon comprises 10,519 phrases from the tourist domain. It is based on a list of short sentences obtained by translation of a US-English 10,000 phrasal corpus. The lexicon is provided in XML format.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

ELRA-S0235 LC-STAR Hebrew (Israel) phonetic lexicon

The LC-STAR Hebrew (Israel) phonetic lexicon comprises 109,580 words, including a set of 62,431 common words, a set of 47,149 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 8,677 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	30,500 Euro	46,000 Euro
For commercial use	50,000 Euro	62,500 Euro

ELRA-S0236 LC-STAR English-Hebrew (Israel) Bilingual Aligned Phrasal lexicon

The LC-STAR English-Hebrew (Israel) Bilingual Aligned Phrasal lexicon comprises 10,520 phrases from the tourist domain. It is based on a list of short sentences obtained by translation from a 10,449 US-English phrase corpus. The lexicon is provided in XML format.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

ELRA-S0237 LC-STAR US English phonetic lexicon

The LC-STAR US English phonetic lexicon comprises 102,310 words, including a set of 51,119 common words, a set of 51,111 proper names (including person names, family names, cities, streets, companies and brand names) and a list of 6,807 special application words. The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

	ELRA members	Non-members
For research use	21,550 Euro	32,500 Euro
For commercial use	35,000 Euro	43,750 Euro

Speech Microphone resources from the European-funded Speecon project (for the development of voice controlled consumer applications)

ELRA-S0232 Swiss-German Speecon database

The Swiss-German Speecon database comprises the recordings of 550 adult Swiss-German speakers and 50 child Swiss-German speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0233 US English Speecon database

The US-English Speecon database comprises the recordings of 550 adult US-English speakers and 50 child US-English speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0240 French-Canadian Speecon database

The French-Canadian Speecon database comprises the recordings of 550 adult French-Canadian speakers and 50 child French-Canadian speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

Speech Microphone Resources from the NATO research group

ELRA-S0238 MIST Multi-lingual Interoperability in Speech Technology database

The MIST Multi-lingual Interoperability in Speech Technology database comprises the recordings of 74 native Dutch speakers (52 males, 22 females) who uttered 10 sentences in Dutch, English, French and German. These sentences comprise 5 sentences per language which are identical for all speakers and 5 sentences per language which are unique for each speaker. Dutch sentences are orthographically annotated.

	ELRA members	Non-members
For research use	400 Euro	500 Euro

ELRA-S0239 N4 (NATO Native and Non Native) database

The (NATO Native and Non Native) database comprises speech data recorded in the naval transmission training centers of four countries (Germany, The Netherlands, United Kingdom, and Canada) during naval communication training sessions in 2000-2002. The material consists of native and non-native speakers using NATO Naval English procedure between ships, and reading from a text, "The North Wind and the Sun," in both English and the speaker's native language. The audio material was recorded on DAT and downsampled to 16kHz-16bit, and all the audio files have been manually transcribed and annotated with speakers identities using the Transcriber tool.

	ELRA members	Non-members
For research use	400 Euro	500 Euro

Speech Telephone Database from the European-funded SALA (SpeechDat Across Latin America) project

ELRA-S0234 SALA Spanish Chilean Database

The SALA Spanish Chilean Database comprises 1,024 Chilean speakers (477 males, 547 females) recorded over the Chilean fixed telephone network.

	ELRA members	Non-members
For research use	13,000 Euro	16,000 Euro
For commercial use	16,000 Euro	20,000 Euro

ELRA Membership Fidelity Program

ELRA has implemented a fidelity program to reward its loyal members. The principle of the fidelity program is to earn miles by joining and remaining member of our association.

Miles, what for?

The awarded miles can be used by members, once earned, for:

- The payment of membership fees
- The payment of registration fees to LREC and other events organized by the association
- The purchase of Language Resources from the ELRA Catalogue with additional discount

How many miles?

This depends on the type of the institution, and therefore on the membership fee. The number of miles per year that can be earned is currently as follows:

- Not-for-profit organization:	200 miles
- European small/medium-sized companies (< 50 employees):	250 miles
- European profit making organizations (>= 50 employees):	375 miles
- Non-European profit making organizations:	1250 miles

Rules

The use and earning of miles is subject to the following rules:

- If the membership fee is paid before July 1st, the member gets the annual number of miles immediately.
- If the membership fee is paid after July 1st, the member is entitled to keep the miles acquired so far but the member will not earn miles for the current year (all other member benefits still apply, e.g. reduced price on resources, on LREC registration fees, etc.). The rule does not apply to new members who can join anytime and earn miles for that 1st year.
- If the membership fee has not been paid by December 31st, all miles acquired so far are lost. Miles can be used as soon as they are earned.

Examples

After 2 years, a not-for-profit institution will have earned 400 miles (200 each year) and will be able to register 3 students for LREC.

After 4 years, any institution member of ELRA will have earned enough miles to get a free membership for one year.