Latest Developments in ELRA's Services

Valérie Mapelli, Victoria Arranz, Hélène Mazo, Khalid Choukri
ELDA (Evaluations and Language resources Distribution Agency)
55-57 rue Brillat Savarin, 75013 Paris, France
E-mail: mapelli@elda.org, arranz@elda.org, mazo@elda.org, choukri@elda.org

Abstract

This paper describes the latest developments in ELRA's services within the field of Language Resources (LR). These developments focus on 4 main groups of activities: the identification and distribution of Language Resources; the production of LRs; the evaluation of Human Language Technology (HLT), and the dissemination of information in the field. ELRA's initial work on the distribution of language resources has evolved throughout the years, currently covering a much wider range of activities that have been considered crucial for the current needs of the R&D community and the "good health" of the LR world. Regarding distribution, considerable work has been done on a broader identification, which does not only consider resources to be immediately negotiated for distribution but which aims to inform on all available resources. This has been the seed for the Universal Catalogue. Furthermore, a Catalogue of LRs with favourable conditions for R&D has also been created. Moreover, the different activities in what regards identification on demand, production within different frameworks, evaluation of language technologies and participation in evaluation campaigns, as well as our very specific focus on information dissemination are described in detail in this paper.

1. Introduction

For more than 12 years, since its creation in February 1995, ELRA (the European Language Resources Association) has focused its activities on a central point of interest: Language Resources (LRs). One of the main rationale that lays behind that orientation is to bring into focus the need for a mutual exchange and use of the LRs that are required for research and development works in the Human Language Technology world. Thanks to the funding of the European Commission during its first three years of activity and with the support of very active experts of the field, ELRA succeeded in providing the HLT community with a now internationally known platform of services for the identification, collection, validation and distribution of LRs. In order to answer to the ever increasing needs of this growing community, ELRA worked out at its early stage of development (October 1995) on the creation of an operational body to help with its everyday life activities: ELDA (the Evaluations and Language resources Distribution Agency).

ELRA, through ELDA, carries out a wide variety of activities related to LRs. These activities have been further developed over the past few years and these developments are the aim of the current paper.

The ELRA activities, also illustrated in Figure 1, can be distributed over four main groups of services, where each of them consists of a long series of sub-activities:

- **Identification and Distribution of LRs**, and related sub-activities: these can be represented by a two-direction relationship with two different types of entities, namely the providers and the users.
- **Production of LRs**: further to its interaction with both providers and users, ELRA also invests considerable efforts on the production of LRs.
- ELRA plays an important role in the area of **Technology Evaluation**, combining its skills acquired with LRs and adapting them for the improvement of Language Engineering products.
- Both for its members and general users, ELRA carries out tasks of **Dissemination**, in order to inform on the resources available 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).

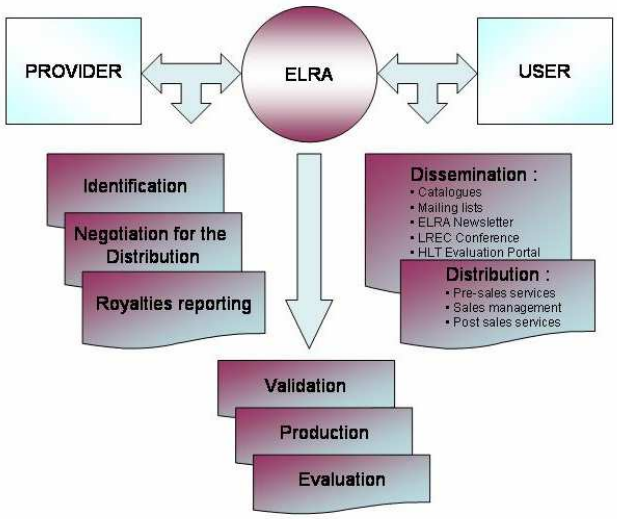


Figure 1: ELRA activities

2. From LR Identification to Distribution

Identification and distribution work implies searching for resources and, then, 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). Two tools are currently used for this purpose: the Universal Catalogue and the ELRA Catalogue of Language Resources.

2.1 The Universal Catalogue

The efforts on the identification of LRs have been increased considerably at ELRA in the past few years. This is directly linked to ELRA's emphasis on identifying already existing resources, easing their availability to their potential users within the HLT 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 and compilation of the found LRs in what we refer to as the Universal Catalogue. This catalogue aims to be a repository for all identified LRs and plays a very important role towards the ELRA Catalogue since it represents an important source of information for the latter to acquire new LRs. Figure 2 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, Projects and their Consortia Partners or other types (e.g., research groups producing LRs outside the context of a project, etc.). This figure also illustrates, in a simplified manner, how LRs initially identified for the Universal Catalogue may end up in the ELRA Catalogue. The distribution step comprises both the complex process of discussion and negotiation with the data providers (and licensing establishment) as well as the distribution procedure through the ELRA Catalogue, which also deals with licensing issues (but these in regard to the future users of the LRs).

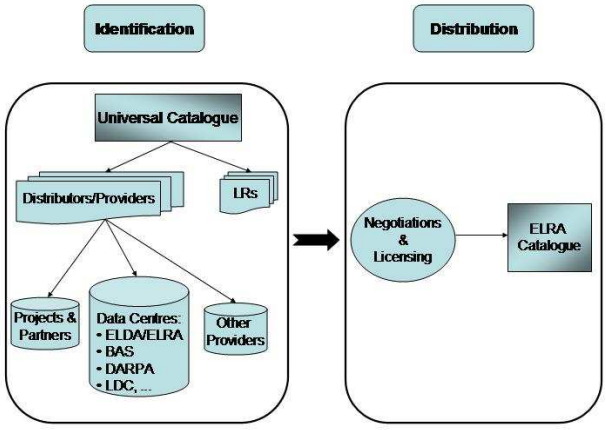


Figure 2: Relation between the Universal Catalogue and the ELRA Catalogue

At present the Universal Catalogue comprises 1,139 identified resources. This figure is constantly being updated as LR additions and revisions (with potential deletions) take place on a regular basis.

The distribution of these figures according to the type of resources is as follows in Table 1:

LR Type	# of LRs
Tools	58
Speech	389
Written	648
Terminology	22
Multimodal/Multimedia	22

Table 1: LR figures in the Universal Catalogue

The Universal Catalogue is available through the ELRA web site: http://www.elra.info.

2.2 From the Universal Catalogue to the ELRA Catalogue of LRs

Further to the general identification task described in the previous section, ELRA is also carrying out identification upon demand. This implies searching for specific LRs that are required in a number of situations. For instance, identification may be done for ELRA members who need some specific resource not available in our catalogues, thus, contributing to the enrichment of the Universal Catalogue, and once negotiated and distributed, also to that of the ELRA Catalogue. Identification may also be done to search for LRs needed within the framework of some R&D project or even resulting from national or international projects ELDA may have been involved in. This has been the case of the evaluation campaigns and projects of the French Technolangue programme, packages which are currently being made available through the ELRA Catalogue.

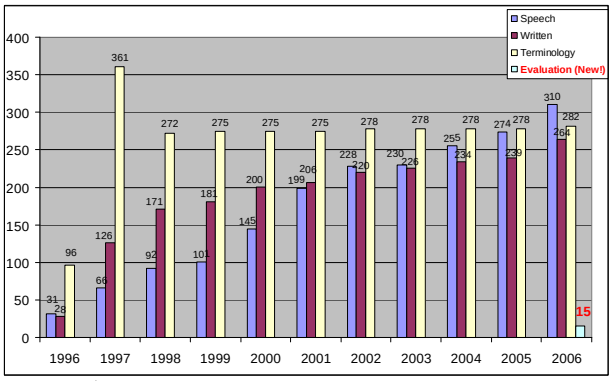
Furthermore, ELRA may focus on community needs, such as those for specific technologies or languages and, thus, search for resources to satisfy these needs. All these scenarios aim to enrich the Universal Catalogue while leading towards the enrichment of the ELRA Catalogue, making the LRs available for the community.

It should also be mentioned that, in a number of occasions, no resource is found for a specific requirement. In that case, the production of that seemingly non-existing LR is foreseen and in some cases carried out to meet our members' needs.

The link between the Universal Catalogue and the ELRA Catalogue takes place by means of a number of legal steps, where licensing is the key point. Some simple standard licenses have been specifically created for the distribution of LRs, and it is through these licenses that distribution rights are obtained to make the LRs public for the HLT community through our ELRA catalogue.

2.3 The ELRA Catalogue of Language Resources: Focus on Evaluation Packages and LRs with Favourable Conditions for R&D

After 12 years of activity, ELRA has managed to make available, worldwide, a large set of LRs through its catalogue: http://catalogue.elra.info. The increase in the number of resources within this Catalogue over the years is illustrated in Graph 1.



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

Two interesting novelties can be pointed out in this regard: (a) ELRA's implication in evaluation, as well as (b) ELRA's work towards a Catalogue of LRs for R&D.

While at the very beginning the main activity of ELRA & ELDA in the framework of evaluation was to supply LRs 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. ELDA, acting on behalf of ELRA, actively participates in evaluation projects, at French, European and international levels. As a natural consequence of this participation, ELRA has worked on making the results of those projects more visible and exploitable by the whole community. Thus, several sets of resources and tools, developed within these evaluation projects, are now available in our catalogue.

Further to the insertion of Evaluation Packages in the ELRA catalogue, we also had to consider a new type of use and pricing policy. This was implemented through the production of a new type of agreement, namely the Evaluation Packages End-User Agreement, as well as another type of pricing that was added to ELRA's current offer. As Evaluation Packages do not only comprise LRs but also the protocols, methodologies, tools, etc. that were used to evaluate a certain Language Technology, and that are required to reproduce such evaluation under the same conditions, different prices were then considered depending on whether the packages were offered for evaluation, research or commercial use.

Moreover, after considering the needs expressed by several academic institutions of the Human Language Technology field, ELRA has decided to give access to a version of its Catalogue of Language Resources dedicated to academic research (http://catalogue.elra.info/retd). This focuses on the importance to allow an easy and fast access to a list of resources specifically produced for R&D purposes and available at very affordable prices.

[1] Note that the decrease in the number of terminology resources between 1997 and 1998 is due to the deletion of obsolete resources from the catalogue, which also shows ELRA's interest in providing users with up-to-date LRs.

3. Production of LRs

In the framework of European and international projects, ELRA has produced – or commissioned the production of – written and spoken LRs, either for R&D purposes, or for distribution among the HLT community. This has been the case for the resources created within projects such as NEMLAR, Neologos, OrienTel, etc. Should a company or an institution need to produce new LRs, ELRA can help building, improving and/or evaluating natural language and speech algorithms or systems. Those LRs may be also used as core resources for the software localization and language services industries, language studies, electronic publishing, international transactions, subject-area specialists and end users.

3.1 ELRA Production and Validation Committees (PCom and VCom)

Still with the same goal of playing an advisory role for LRs, ELRA implemented two technical committees, composed of experts in the field to think about, promote and support the production and validation of LRs. The Validation Committee (VCom) was set up in 2000 in order to push forward the quality of LRs. This was mainly implemented through the creation of validation manuals for Speech and Written LRs, thanks to which a good number of LRs of the ELRA catalogue could be checked out and validated. The work was implemented through two validation centres, one for Speech LRs (SPEX, the Netherlands) and one for Written LRs (CST, Denmark). Three years later, ELRA considered that some tasks of the VCom could not be dealt with as such within the same committee and needed another level of expertise, i.e. at the production level. Consequently, the Production Committee (PCom) was set up to collect, optimize, develop and promote standards and best practices for LR production specifications. One of the main experiments of PCom was to study and work out the development of techniques for merging two independently created lexicons and thus generate a unified lexicon. The work was implemented by the ILC CNR (Pisa, Italy).

3.2 Production Services by ELRA's Operational Body, ELDA

With the support of a significant network of partners worldwide, ELDA can provide LRs in various languages: ELDA has already compiled LRs in more than 25 languages. Besides, the highest quality of resources is guaranteed by a strict validation procedure.

ELDA is involved in every stage of the production of LRs, and for different types of resources, starting from Speech/Video Data Collection, Written Data Collection (corpora and lexica) up to data creation for specific technologies and/or evaluation campaigns (see also the HLT Evaluation Portal maintained by ELRA and detailed in Section 4.2).

Among a number of projects, we can quote some of the most recent production achievements (more information on some of them can be found at the current LREC conference):

- Speech resources for automatic speech recognition over (fixed or mobile) telephone applications: LILA Hindi and Korean databases, Orientel Moroccan and Tunisian, etc.
- Speech resources for speech recognition over broadcast news applications: NEMLAR Broadcast News Speech Corpus for Arabic, ESTER Corpus for French, etc.
- Speech resources for speech-to-speech translation: TC-STAR for English, Spanish and Mandarin.
- Multimodal resources: CHIL video annotations with audio transcriptions.
- Written corpora: NEMLAR Written Corpus for Arabic.
- Lexica: phonetic lexica within LILA, Orientel or other SpeechDat-family projects, LC-STAR lexica for speech translation applications, etc.

Although they cannot be quoted for confidential reasons, it is also worth raising that, beyond its participation in national or international projects, ELDA also offers production services upon demand.

4. Evaluation

As mentioned earlier, through its participation in major European projects (CHIL, TC-STAR, CLEF) and in French national programmes (Technolangue), ELDA is nowadays involved in the evaluation process itself, the evaluation of products, systems, and applications developed for HLT.

4.1 Collaborative and Customized Services for the Evaluation of Language Technologies

With a strong background in evaluation, from evaluation campaigns management to distribution of evaluation packages, ELDA can now offer on-demand evaluation services and customized LRs to laboratories and/or companies wishing to evaluate their HLT-based systems. Below is given a non exhaustive number of technologies that could be evaluated by ELRA in three different areas of Human Language Technology:

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

Considering this multiplicity of fields, the task requires a number of criteria that must be taken into consideration to carry out a good evaluation campaign:

- Definition of the evaluation task(s),
- Finding or producing LRs to be used for the evaluation: defining the type of resource, checking availability or preparing new resources from scratch,
- Finding or producing evaluation tools: defining the type of tool, checking availability or preparing new tools from scratch.

4.2 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: www.hlt-evaluation.org.

The portal provides all kinds of information related to 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.

5. Dissemination

ELRA has also increased its activities for the dissemination of information on LRs and on the progress of Evaluation, as described below.

5.1 Language Resources and Evaluation Journal

The *Language Resources and Evaluation Journal* is the first publication devoted to the acquisition, creation, annotation, and use of LRs, 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.

The ELRA members are granted complimentary access to the journal through the association under 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 in the Journal spread worldwide thanks to the visibility and promotion given by the conference.

The LRE Journal is available online at: http://wwww.springerlink.com.

5.2 Language Resources and Evaluation Conference

In 10 years (the first LREC was held in Granada in 1998), LREC has become the major event on Language Resources and Evaluation for Human Language Technologies (HLT): http://www.lrec-conf.org/. 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.

6. Conclusions

As it can be seen in this paper, ELRA has expanded its services along the last couple of years. Activities such as identification, evaluation, production and dissemination have seen their efforts considerably increased in a search to support researchers and developers.

The creation of the Universal Catalogue has triggered the existence of a tool that should prevent language resource developers from generating already existing resources, as well as help LR users in their search for material to work on.

Our efforts on evaluation are supporting the advance of evaluation in the area of HLT, in terms of developing evaluation methodologies, making evaluation packages available (with all their data, protocols, etc.), providing information on the field (for instance, through the HLT evaluation portal), participating and setting up evaluation campaigns (through ELDA's participation in projects and R&D programmes) and brainstorming events (such as the different workshops organised for that purpose, including the one to take place at LREC 2008).

Our involvement in LR production has also had a direct impact on the availability of LRs, both regarding the needs of private users, such as companies working on HLT, or public LRs aiming to meet the needs of the community at large.

In what regards dissemination, our activity has been multifold, but always focusing on the dissemination of the activities taking place within the area. This is so for the Newsletter, publishing contributions from different HLT players, who describe their work, their achievements or simply the initiatives taking place in their context. This is also the case for our LREC Conference, which has become an event not to be missed in order to meet the community and have a friendly opportunity to discuss work around its main conference or its many satellite workshops. Last but not least, this is also the case for the other events whose organisation ELRA is involved in, such as LangTech, the latest Evaluation Workshop at the MT Summit 2007, and even its involvement in the publication of the Language Resources and Evaluation Journal.

Although not the focus of the current document, other ELRA regular activities have continued, such as the validation of LRs for the ELRA Catalogue.

Last but not least, the recently created "fidelity program" should also be mentioned. This program has been implemented to reward its loyal members, allowing them to earn miles by joining and remaining a member of the association. 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, or even the purchase of LRs from the ELRA Catalogue with additional discount.