Recent Activities within the European Language Resources Association: issues on sharing Language Resources and Evaluation

Khalid CHOUKRI
ELRA/ELDA
55-57 Rue Brillat-Savarin, 75013 Paris, France
Tel. +33 1 43 13 33 33 - Fax. +33 1 43 13 33 30
Email: choukri@elda.fr
Web: www.elda.fr or www.elra.info

Abstract

This paper aims at describing the recent activities within the European Language Resources Association (ELRA) on issues covering its main missions: making available Language Resources and providing Human Language Technologies Evaluation packages.

1. Introduction

The paper focuses on the issues involved for making Language Resource available to different sectors of the language engineering community as well as contributing to the evaluation of Human Language Technologies (HLT). ELRA has been, since its foundation in 1995, a conduit for the distribution of speech, written and terminology databases, enabling key players to have access to Language Resources (LRs) for technology development and technology evaluation. ELRA's initial mission was to establish itself as a self-supported, centralized Not-for-profit organization for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role, ELRA had to address issues of various natures such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, clearing Intellectual Property Rights (IPR) and/or Industrial rights, etc.), information dissemination (to act as a clearing house). ELRA set up an operational body (ELDA: Evaluations and Language Resources Distribution Agency) to take care of the daily aspects related to these missions. Since its establishment, ELRA has managed to make available, worldwide, a large set of marketable resources. ELRA handled the legal issues through generic license agreements and IPR manuals that were made widely available. A set of Language Resources validation manuals has been produced with the support of appropriate external validation units to promote quality and best practices. They are widely distributed (in particular for speech and written resources), via its web site.

After a first phase devoted to Language Resources, ELRA has been actively promoting the evaluation of Human Language Technologies initially through its involvement in evaluation campaigns by providing adequate Language Resources and more recently by extending this work through the establishment, via its agency ELDA, of an evaluation infrastructure in Europe. Drawing on its experience in national and Europe-wide evaluation projects and also its experience in the production, validation, packaging and distribution of Language Resources, ELDA's evaluation department is working to establish a European clearing house for evaluation, in the same way that ELDA has become the European clearing house for Language Resources. ELDA's vision for a European evaluation infrastructure is inspired by both European and international evaluation initiatives but considers the need to have such work distributed over a number of supporting units consisting of experienced labs all over Europe. A similar distributed infrastructure has been already assessed for the validation of Language Resources trusted to external units under specific contracts.

An important highlight of the association work on promoting the Human Language Technologies area is the set up of this International Conference on Language Resources and Evaluation (LREC). In addition to this, ELRA has strongly contributed to the set up and the organization of LangTech which is the European Forum for Speech and Language Technology. LangTech is organized to promote the business behind the Human Language Technologies.

2. ELRA Foundation

2.1. ELRA structure

As an association, ELRA is governed by a Board, which defines the association's strategy which is then implemented by the Chief Executive Officer (CEO) and his staff. In order to efficiently carry out such mission, the CEO has set up an operational unit called Evaluations & Language Resources - Distribution Agency (ELDA) as the organizational infrastructure which employs all the staff. Whenever appropriate, a number of technical activities are outsourced to experienced institutions that act under its supervision and with its endorsement.

ELRA is a membership based institution open to organizations and not individuals. Since its foundation, ELRA has attracted an important and steady number of members going from 63 in 1995 to over 100 in 2002.

More than 230 different organizations have joined ELRA at least for 1 year. The services offered by ELRA to its members are summarized both on the ELRA web site and brochures. These services go beyond the important discount given on the price of Language Resource.

3. ELRA's Mission and Activities

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by language technologies' developers and/or researchers are made available when they already exist or to produce them in a cost-effective frame. This mission is tuned from time to time to anticipate future requirements. Such a mission can be itemized as:

• Language Resources related issues (issues concerning the identification of useful resources, handling the legal issues related to the availability of Language Resources, the distribution activities and Pricing policy, the validation and Quality assessment, commissioning the production of needed Language, Resources & Market watch, etc).

• Evaluations of Human Language Technologies (issues concerning commissioning the production of needed Language Resources for evaluations, carrying out evaluation campaign, providing evaluation packages for all Human Language Technologies, etc.)

• Promotion of the field (issues concerning information dissemination and awareness, market watch and analysis, Identification of useful resources, metadata and description of LRs, etc.)

In order to play its role, ELRA created a structured and publicly available catalogue of Language Resources. A set of description forms was prepared, aiming to help the providers describe what they propose to ELRA for distribution in a more uniform and consistent way and the users have a quick access to the main features (see the URL corresponding to the catalogue at: www.elda.fr/rubrique6.html) ELRA is revising its work on these meta-data aspects and extensive work is also being carried out worldwide on these "metadata" issues. ELRA is coordinating a major European funded project called Intera (www.elda.fr/INTERA) with the involvement of the Max Planck Institute team behind IMDI and which built on the work achieved within the Eagles/Isle project. A number of key providers of LRs to ELRA are validating the metadata sets before these become publicly available. In addition to this, ELRA is an active player in the OLAC initiative (Open Language Archives Community, www.language-archives.org).

3.1. ELRA Language Resources Catalogue

The current catalogue is compiled with respect to the three colleges of ELRA: speech, written, and terminology. Some tools can also be catalogued if they are available for free. Very recently? we have decided to add a fourth category to our catalogue to take into account new emerging trend of multimodal/multimedia resources. In the under-completion revision of the catalogue we have decided to have a Language Resources Catalogue and a new catalogue of HLT Evaluation packages. The progress of our identification task is illustrated through the number of resources being offered. Since 96 this has increased from 31 Spoken Language Resources , 28 Written LRs, and 96 terminology databases to respectively 228, 220, and 278. This should not hide the fact that many key resources are still not available for a large number of languages (including Western European ones!). Even for a number of basic resources (read speech, phonetic lexica, text corpora, etc.), we could see that either these resources are not available for distribution or (worse) does not exit at all.

ELRA has compiled a matrix representation with a list of resources as the column entry and the languages as the row entry and such matrix shows a large number of empty cells indicating that, for a given language, the resources have not been identified. The ultimate target is to have a large matrix that highlight the "universal catalogue" that should be compiled for all human languages with a pre-requisite to stimulate their production in order to meet the needs and requirements of both academic institutions and industrial users. Based on this, ELRA started to promote the concept of a Basic Language Resource Kit (BLARK) for all languages and later on extended this concept towards an Extended LAnguage Resource Kit (ELARK), which may be useful for some languages that can afford to have more than the basics. A good illustration of the BLARK concept (Basic LAnguage Resource Kit) is the work being done presently to gather information about Arabic resources within the NEMLAR project conducted with the support of the European commission and a large number of players within the Arabic country (see a paper on Nemlar within these proceedings).

3.2. Handling the legal issues related to LRs

The basic principles of LRs 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. 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. 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 and we encourage all actors to use them.

3.3. Distribution activities and Pricing policy

The first two years of activity were devoted to the establishment of the infrastructure and to the identification of valuable resources. This explains the low take off of our sales in 1995-1996.

The pricing policy is also a crucial issue that needed careful attention. This had to take into account the fact that we were establishing a new market in which LRs should be traded like any other commodity, as well as the requirements and restrictions imposed by the provider (or the producer) when it comes to the issue of financial compensation. Likewise, market knowledge and contacts with potential providers allow ELRA to always have reliable and useful information on the demands and needs of the market. The ELRA 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, expected revenues, etc. The pricing must also take into account the ELRA distribution policy, which is to always offer a discounted price to its members.

In some cases, the providers accept to have their resources distributed for free. This is often the case when production of LRs is already financed by the European Commission or by national governments. Exceptionally, ELRA is able to offer price reductions even without this being financially supported by the providers. The restrictions on the distribution, sometimes imposed by the providers, are more often of two kinds: it is either a restriction on the user profile or a restriction on the usage. The providers may limit the distribution to members only or to Europeans only, or they may restrict the use of their resource to research or even to academic research. The statistics on distribution shows that for commercial use this evolved from 88 in 1998 to 134 in 2002 (respectively from 122 to 347 for research use).

3.4. Validation and Quality assessment

The users of LRs, in particular ELRA "customers," need to know about the product they are purchasing: they need to know its technical specifications and need to be assured of quality control. Validation is normally used by ELRA in reference to the activity of checking the suitability for the market, the adherence to standards, and the quality control of the LRs. Of course, "the market" reaction to the product is the ultimate validation. A detailed paper on these issues by our validation unit is included in these proceedings and for an overview of current activities of the Committee: please visit www.elda.fr/article14.html

3.5. Production of useful Language Resources

ELDA is also very active in the production of Language Resources that would end up in its catalogue. A number of partners have outsourced such activity to ELRA which could use its extensive network of production units to carry out such work. Example of such resources for spoken data resources for consumer product embedded applications in the framework of the Speecon project (www.speecon.org), resources for modern standard and colloquial Arabic within the Orientel project (www.orientel.org), or broadcast news for a number of languages (French, Arabic, non-native English, etc.). Other projects for the production of WLR, i.e. lexica or specialized dictionaries are being carried out and will be reported on.

3.6. Human Language Technologies Evaluations

3.6.1. Need for an HLT Evaluation infrastructure.

This section elaborates on the evaluation activity which was initiated at ELRA progressively since 1998, to support the evaluation of Human Language Technologies. ELRA is exploiting the infrastructure set up for Language Resources business to establish an efficient and cost-effective activity on evaluation: both carrying evaluation campaigns and distributing evaluation packages. The main objective is to play the role of a clearing house with the support of a network of institutions willing to ensure that Europe has its own infrastructure for this crucial domain and capitalizing on the model of Language Resources distribution that proved to be efficient. Human Language Technologies Evaluation forms a fundamental part of the development of language engineering. It is essential for validating research hypotheses, for assessing progress and for choosing between research alternatives. Evaluation also identifies promising technology or research directions enabling industry to assess its market value. Evaluation campaigns also provide useful input when deciding whether a technology is mature enough to be considered as a candidate for starting commercial application development. For ELRA, a further side effect of evaluation campaigns is the production of high quality evaluation resources, in the form of training and test data along with evaluation software packages, distributed or produced during evaluation campaigns.

3.6.2. An ELRA Evaluation Infrastructure

ELDA has a proven track record in the efficient and cost-effective distribution of LRs on both a European and worldwide level. It has set up an organizational model of networks dedicated to LRs. Along with its experience in national and European evaluation projects, ELDA's evaluation department capitalizes on this involvement to create an organizational model for efficient and cost-effective evaluation management. This entails the creation of a European, even international, network of evaluation centers providing evaluation resources, software packages, technology, forums of scientific expertise and R&D centers for the independent, ethical evaluation of HLT.

The European infrastructure would be organized along two major principles, proactive and reactive evaluation schemes. ELDA's evaluation department is currently taking part in reactive evaluation in that it has been granted national and European projects, such as the French national programme EVALDA (evaluation of 8 technologies ranging from corpus alignment tools to machine translation: a separate paper is published within

these proceedings), CLEF (Evaluation of Information retrieval systems, see a paper on CLEF within these proceedings), etc. ELRA will be involved in the specification and production of evaluation resources, packages and protocols for the new Integrated Projects of the European R&D FP6 (CHIL, TC-STAR). An exit strategy is defined for each project where the evaluation resources, packages, software and knowledge (final project reports) produced in each evaluation campaign will be made available to external players through ELDA's catalogue for a modest price.

In parallel, the evaluation infrastructure would be proactive. ELDA endeavors to make available evaluation resources and packages for all HLTs in as many languages as possible. At the very least, this European evaluation infrastructure would have to make available evaluation resources and packages for the official EU languages.

3.7. The involvement in FP6 Integrated projects

ELRA, through ELDA, is taking part in two major FP6 Integrated projects, namely CHIL and TC-Star. TC-Star is the follow-up of a one year preparatory action that focused on assessing the best organizational models with respect to the speech to speech issues. The present project aims at advancing research and technology development in all aspects of speech-to-speech translation. ELRA will be acting as the coordinator of aspects related to LRs and Evaluation. A major target for us is the set up of a distributed plug-in evaluation platform that would allow all the participants to assess the progress through the organization of evaluation campaigns addressing performance measures of single technologies as well as end-to-end systems. CHIL is an ambitious and a visionary project that aims to "introduce Computers into a loop of Humans interacting with Humans, rather than condemning a human to operate in a loop of computers, forcing him/her to attend to and interact with an artefact on its artificial terms". The project will have a number of showcases and services, among which lectures and meetings. Among the technology components that will be investigated we may quote: person localization and tracking, person identification, face recognition, speaker identification (and fusion), gesture recognition, "attention" tracking, conversational LVCSR, acoustic scene analysis, emotion identification (facial expression, emotional features, other biometric data), topic identification, interpersonal relationships, conversational style, genre ID, etc. ELRA is in charge of coordinating the work on LRs for several technology components itemized above and the corresponding evaluation through the collection of data and the set up of required infrastructure.

4. Conclusions

This paper briefly reported on several initiatives being carried out within the European Language Resources Association. In addition to LRs and HLT actions, ELRA promotes the field and contributes to information dissemination activities through this conference, LREC, and other events it takes part to such as Langtech. In addition to this, it is important to mention the ENABLER project that has been working on LRs Roadmaps for the next decade to define needs and requirements of HLT R&D and Technology teams. ELRA is also actively promoting the set up of bridges between written and spoken language communities.