ELRA's Services 15 Years on...Sharing and Anticipating the Community

Victoria Arranz, Khalid Choukri
ELDA / ELRA
55-57 rue Brillat Savarin, 75013 Paris, France
E-mail: arranz@elda.org, choukri@elda.org

Abstract

15 years have gone by and ELRA continues embracing the needs of the HLT community to design its services and to implement them through its operational body, ELDA. The needs of the community have become much more ambitious...Larger **language resources** (LR), better quality ones (how do we reach a compromise between price – maybe free – and quality?), more annotations, at different levels and for different modalities...**easy access** to these LRs, **solved IPR issues**, appropriate and adaptable **licensing schemas**...large activity in **HLT evaluation**, both in terms of setting up the evaluation and in need for necessary data, protocols, specifications as well as for more expertise in conducting the whole process...**producing the LRs** researchers and developers need, LRs for a wide variety of activities and technologies...for development, for training, for evaluation...**Disseminating** all knowledge in the field, whether generated at ELRA or elsewhere...keeping the community up to date with what goes on regularly (LREC conferences, LangTech, Newsletters, HLT Evaluation Portal, etc.). Needless to say, part of ELRA's evolution implies facing and anticipating the realities of the new Internet and data exchange era and remaining a LR backbone...looking into **new models** of **LR data centres and platforms**, LR access and exchange via **web services**, new models for **infrastructures** and **repositories** with even higher **collaboration** to make it happen. ELRA/ELDA participate in a number of international projects focused on this new production and sharing schema that will be detailed in the current paper.

1. Introduction

15 years have gone by and ELRA (the European Language Resources Association) continues embracing the needs of the HLT community to design its services and to implement them through its operational body, ELDA (the Evaluations and Language resources Distribution Agency). The needs have evolved, becoming much more ambitious, both following technological trends as well as looking into a reality which certainly goes beyond the simpler initial scope of sharing some existing resource, if this could be considered simple at all.

For those not so familiar with our activities, ELRA has focused its activities around Language Resources (LRs) since its very creation, in 1995. One of the main rationale that has always laid behind it has been bringing 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. During its early days, this was made possible thanks to the funding from the European Commission and to the strong support from a number of experts in the field. At a later stage, ELRA has succeeded in continuing its work and broadening its early scope to cover many other needs and potential worries from the HLT community. Sharing, learning, evolving and anticipating, as well as acting in a sustainable manner, have always been a part of ELRA's motto.

Such evolution and anticipation have demanded endeavouring into new activities, new paths, but all of them directly linked to the LR and HLT world, which have always been the association's backbone. This has allowed both ELRA and ELDA to provide the HLT community with an internationally well-known platform of services for the identification, validation and distribution of LRs. It soon expanded its work into the production of those LRs which were needed by its members and partners, but which were not available, and it has a firm experience in the evaluation of HLT technologies and the running of evaluation campaigns for a wide range of technologies. Last but not least, one further pillar that should not be left behind is that of dissemination. As most of you reading this article will know, ELRA is part of the organisation of the LREC Conference (together with ILC-CNR and with the support of major players in the area), as well as of a number of other international events, and it carries out a number of other relevant activities (maintenance of catalogues, information portals, Newsletter, etc.), which allow it to be a main reference in the area of Language Resources and Technologies.

The needs regarding LRs have become much more ambitious these past couple of years, in terms of both technological development and the realities of the new Internet and data exchange era. ELRA looks ahead with its always-passionate attitude and participates actively in the new ways to support its members, partners, users and observers.

The paper will be structured as followed: it will start with an update on the association's regular activities, showing its new achievements and then, moving onto the new paradigms for the coming years.

2. Latest Developments

Figure 1 illustrates the 4 main pillars under whose umbrella take place the numerous activities and sub-activities at ELRA. As briefly mentioned in the Introduction, these are the following:

- Identification and Distribution of LRs.
- Production of LRs.
- Technology Evaluation.
- Dissemination.

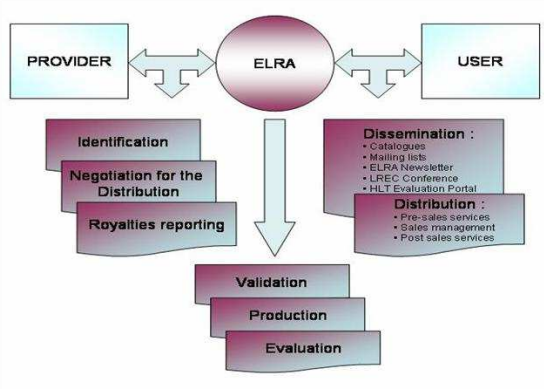

Figure 1: ELRA activities

2.1 Identification and Distribution

ELRA has devoted a considerable effort to the identification and distribution of Language Resources.
The basic principles of language resources licensing has been worked out with the support of lawyers. One of ELRA's priority tasks 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. Figure 2 shows the model these contracts are based on.

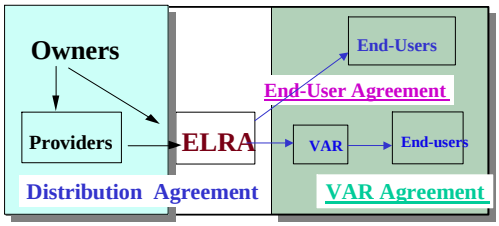

Figure 2: Contract Model Followed by ELRA

Such contracts establish, among other things, what usage is allowed for the LRs, whether for both research and technology/product development or only for research purposes. In any case, these contracts protect the providers and their LRs by stating that the user shall not copy or redistribute the LRs. The production and distribution of these licenses is one of ELRA's contributions to the development of LR brokerage. These licenses are available on ELDA's web site [1] and we encourage all interested actors to use them, even if they were designed before the Creative Commons [2] licenses and future mergings or joint redesigning are a possibility under study. 500 licenses have been signed so far.

Over the past 15 years, more than 1,000 resources have been catalogued and made available, thanks to over 250 distribution agreements. ELRA has distributed over 3,500 resources for HLT development (cf. Figure 3 for further details), out of which 48% were used for research purposes by academia, 37% for research and technology development by industry, and 16% within technology evaluation. Furthermore, an additional 1,500 copies have been distributed within evaluation campaigns. These resources have been made available within an easy, trustable and fair legal and exchange framework, which has undoubtedly supported and boosted the development of HLT and further applications.

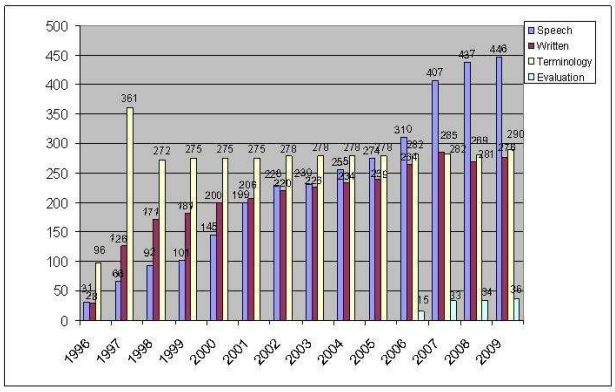

Figure 3: The ELRA Catalogue

Further to our Catalogue [3] of ready-for-distribution LRs intensive work is taking place in the identification of all existing LRs so as to compile them in the Universal Catalogue [4] and provide all possible information to the users of the Language Technology R&D community. In addition to the more than 1,000 LRs available through the ELRA Catalogue, ELRA's identification work has compiled useful information on over 1,700 resources that constitute the Universal Catalogue, an antechamber to the ELRA Catalogue, as shown in Figure 4.

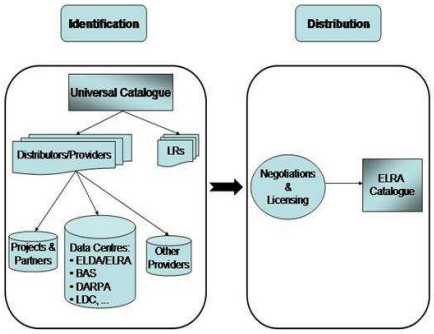

Figure 4: The Universal Catalogue and the ELRA Catalogue of Language Resources

[1] http://www.elda.org/article1.html
[2] http://creativecommons.org/
[3] http://catalog.elra.info/
[4] http://universal.elra.info/

An interesting new feature of the Universal Catalogue is the development of a simplified collaboration form. Following some feedback received from our users, it became clear that establishing a collaborative chain required reducing the contributor's effort to a minimum. Even if it was initially believed that compiling a detailed description on each resource was crucial, it was soon made clear that volunteer users willing to render some information public only participated if it implied a low-cost and low-effort task, which is a reality when considering people's workloads these days.
One further point to be emphasized is the window-shopping nature of the Universal Catalogue, which allows users to realize that some LR exists even if it is not currently available. Numerous people contact ELRA hoping to track down and gain access to these resources, which have been identified by the ELRA team very often in not very thorough sources. This encourages us to continue our archeological work and clear out the resource situation for the user, hopefully wiping out any existing legal barrier.
Last but not least, a few lines should be said about the new LREC Map identification tool. For the first time, a LR identification tool has been set up by ELRA during LREC Conference submission time. This tool has allowed conference contributors to provide information on the resources in their papers, which has provided the community with information on almost 2,000 resources. For further details refer to (Calzolari *et al.*, 2010).

2.2 Production of Language Resources

In line with its regular activities, ELRA has continued being very active with the production or commissioning of LRs. This takes place both within the framework of European and international projects or in support of companies or institutions. With regard to the former, this has been the case for the resources created within projects such as NEMLAR, Neologos, OrienTel, Speecon, C-ORAL-ROM, CHIL, TC-STAR, ESTER, MEDIA, MEDAR, PASSAGE, etc. With regard to the latter, ELRA has also conducted production work on its internal funds or commissioned by partners, sometimes engaging in work of confidential nature.
So far ELRA has compiled LRs in more than 25 languages, following strict validation and quality control procedures to ensure high quality and being involved in every single stage of production, from the establishment and definition of specifications and guidelines to the ultimate quality control detail.
Such LRs target different technologies and the current state of technological development demands more ambitious resources, in terms of size, type of linguistic information as well as quality of the end-result. All these have been main objectives for ELRA, who has produced through ELDA a large number of LRs. Some of its recent achievements comprise: (i) speech data for a variety of languages (e.g., Hindi, Korean, Colloquial Arabic(s), Canadian French, US Spanish, etc.), (ii) Broadcast News Speech Corpus for Arabic, French, Spanish, etc., (iii) corpora for languages such as Catalan, Kazakh, Romanian, Turkish, etc., (iv) aligned textual corpora for Machine Translation in languages such as Arabic, Chinese, English, French, German, Spanish, etc., (v) video annotations with audio transcriptions, (vi) collections of SMS data, (vii) recordings of Wizard-of-Oz based data for dialogue systems, etc.
Most of these resources are intended to be part of the ELRA Catalogue, even if, when commissioned, all the rights including ownership remain with the commissioning party, and, as a consequence, they may remain exclusive for its sponsor, at least for a certain period of time.
Our team is always eager to get your comments and requirements and is ready to assist on any aspect of the production process.

2.3 Evaluation of HLT

Throughout the years, ELRA has also ensured an infrastructure for technology evaluation, which in the last few years has also counted on web-based service platforms.
A number of HLT evaluation campaigns have been initiated, others have been strongly supported, so as to make sure the community is well informed with what goes on in the area, what the state of the art is and they can even engage into evaluation activities thanks to the "evaluation packages" that have been compiled and made available. These evaluation packages contain all required databases, tools, methodologies and protocols to conduct similar experiments and test their work in a comparable scenario.
More than 20 technologies have been evaluated and over 40 evaluation packages are available through the ELRA Catalogue. Some covered technologies are the following:

- 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.

To ensure that the community has access to all available information on evaluation, ELRA has set up a reference portal called *The HLT Evaluation Portal*[5], where all collected information in the area has been compiled. This portal may be used as a quick and easy reference for evaluation protocols and know-how, metrics, tasks, resources, projects, campaigns, whether past or ongoing, etc.

2.4 Dissemination

Last but certainly not least, ELRA continues its work on providing information on the field to all those interested, both as passive observers or as active collaborators. It also offers a discussion or "Speaker's corner" for all those willing to exchange their views, share their findings or their visions for the future on the many topics regarding Language Resources. In this regard, we can mention events such as the current one, *LREC*, on its 7th edition, an

[5] http://www.hlt-evaluation.org

every-time-more-popular event, which has become *the* gathering place by excellence. The *LangTech* events should also be referred to (the European exhibition on language technologies, which took place in Rome in 2008 already looking forward to its most likely 2011 edition). The European Language Resources and Technologies Forum, organized within the framework of the EC funded project FlaReNet (the very last one organised in Barcelona in 2010), and the MEDAR Conferences on Arabic (last one in Cairo in 2009) have also been major events handling a wide variety of topics. Looking into minority languages, the workshops on Less-Resources languages (with their last edition in Poznan, jointly organised with the LTC'09 Conference) should not be missed. All these events allow the community to gather, exchange, discuss, which is needless to say the key for technological learning and advance.

In addition to all these discussion forums, ELRA is also behind the *Language Resources and Evaluation Journal*[6], 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, who we acknowledge as they grant ELRA members with a free subscription.

ELRA's monthly Members' News and its quarterly Newsletter are also very dynamic means to interact with and inform the community with current activities, research, etc.

3. New Visions and a Large International Cooperation

As mentioned early in the paper, part of ELRA's success and evolution has implied facing and anticipating the realities of the new era. Now that many of the original goals have been achieved, others expanded, and some of the activities keep going, ELRA looks into the future facing new challenges, new needs, new trends. With a sound background and a large experience in the field, we reshape our visions and missions, aiming to offer our collaborators and partners with new services derived from the new scenarios of data exchange, Internet and LR sharing, always bearing in mind the rational behind its establishment.

Still keeping the identification and distribution of resources as backbone, we will approach these services from a new modus operandi which will combine features such as access and exchange via web services and web 2.0, inspiring from the "open source" principles as well as business-to-business approaches. New models for LR data centres and platforms will be supported by ELRA, along with LR access and exchange via web services, where each user can satisfy their needs but within a well designed legal framework. This new vision is already being implemented in active collaboration with a number of expert partners. A number of international initiatives have taken off and work is currently taking place to design this new vision, this new trend in LRs. Some newly funded projects such as PANACEA[7] and T4ME[8] will play a key role for the implementation of this large international cooperation. PANACEA (**P**latform for **A**utomatic, **N**ormalized **A**nnotation and **C**ost- **E**ffective **A**cquisition) will allow us to define new legal frameworks for the open production and sharing of LRs, tools and services. It also considers cost-effectiveness in the production of LRs and its automation through web-service "factories". These are challenging issues to tackle. T4ME is a network of excellence to design "an open, integrated, secured and interoperable exchange facility for language data and tools for the Human Language Technologies domain". We anticipate this will help increase the sharing of LRs and the development of the integrated repository concept.

A final non-financed international initiative is also ongoing to secure the largest Global catalogue of LRs, with partners such as LDC[9], NICT[10], OLAC[11], and others, in the perspective of harmonising their catalogues with the ELRA Universal Catalogue, the LREC Map of language Resources and other Maps from other conferences adopting the LREC Map policy.

4. Conclusion

The current paper has aimed to give a quick overview of the activities carried out by ELRA in the last couple of years of its 15-year existence. Both ELRA and ELDA have grown up with the community and have learnt to anticipate its needs in all aspects surrounding LRs and evaluation. From our early archiving and distribution activities, we have successfully set up a LR identification, collection, validation and distribution platform, with clear and well-established legal frameworks to protect LRs. Moreover, we have managed to enhance our work on evaluation, with new techniques, covering more technologies and languages, providing more evaluation packages and setting up an HLT Evaluation portal of information. Furthermore, work has increased considerably in terms of LR production and its coverage, with many more types of LRs being produced and many more languages being covered. Dissemination has also moved on considerably, with a very active organisation of discussion forums and encouraging a large international cooperation.

After years of establishment and consolidation, we look forward to the new challenges emerging from the new trends within our community. We were born to help all our collaborators, partners, etc. and we hope to continue providing them with what they need to go on with their work in an optimal way. Our latest projects, such as PANACEA and T4ME look into all these new aspects and they will surely help move towards the new interoperable exchange we all need.

[6] http://www.springerlink.com/

[7] http://www.panacea-lr.eu/

[8] http://t4me.dfki.de/

[9] http://www.ldc.upenn.edu/

[10] http://www.nict.go.jp/

[11] http://www.language-archives.org/

5. References

Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J. and Piperidis, S. (2010). The LREC 2010 Resource Map. *LREC 2010 Conference*, Valletta, Malta, 2010.