ELRA in the Heart of a Cooperative HLT World

Valérie Mapelli, Victoria Arranz, Matthieu Carré, Hélène Mazo, Djamel Mostefa, Khalid Choukri
ELDA/ELRA
55-57 rue Brillat Savarin, 75013 Paris, France
E-mail: {mapelli; arranz; carre; mazo; mostefa; choukri} @elda.org

Abstract

This paper aims at giving an overview of ELRA's recent activities. The first part elaborates on ELRA's means of boosting the sharing Language Resources (LRs) within the HLT community through its catalogues, LRE-Map initiative, as well as its work towards the integration of its LRs within the META-SHARE open infrastructure. The second part shows how ELRA helps in the development and evaluation of HLT, in particular through its numerous participations to collaborative projects for the production of resources and platforms to facilitate their production and exploitation. A third part focuses on ELRA's work for clearing IPR issues in a HLT-oriented context, one of its latest initiative being its involvement in a Fair Research Act proposal to promote the easy access to LRs to the widest community. Finally, the last part elaborates on recent actions for disseminating information and promoting cooperation in the field, e.g. an the Language Library being launched at LREC2012 and the creation of an International Standard LR Number, a LR unique identifier to enable the accurate identification of LRs. Among the other messages ELRA will be conveying the attendees are the announcement of a set of freely available resources, the establishment of a LR and Evaluation forum, etc.

Keywords: sharing LRs, HLT support and cooperation, Intellectual Property Rights (IPR)

1. Introduction

Over the last few years, ELRA has been involved in a number of international initiatives which are in line with its "original" mission and vision while accounting for the new trends and the community new expectation. These activities focus on the axes that have been ELRA's main concern since its creation in 1995: developing the means for sharing Language Resources (LRs) within the widest community, helping in the development and evaluation of Human Language Technologies, disseminating information and promoting cooperation in the field, and, as a cross-cutting concern, understanding and clearing Intellectual Property Rights in a HLT-oriented context. ELRA, and its operational body ELDA, have been active players of such initiatives from the very beginning. This has been so through their well-established middleman role in the field as well as contributing with their expertise in the setting up of LR repositories, evaluating technologies, producing LRs, establishing standards and best practices, and defining metadata schemas, among others. This paper aims at giving an overview of ELRA/ELDA's recent activities over those dimensions.

2. Sharing Language Resources

2.1 The ELRA catalogues

For 20 years now, the HLT community has seen an ever-growing supply and demand in terms of LRs. Among organisations that have been playing a crucial role in the development of the idea of gathering and sharing LRs in that field, ELRA (European Language Resources Association)[1] has been one of the precursors, in particular thanks to the setting up of its large catalogues of LRs, making them not only visible but also available to the community. The ELRA Catalogue[2] now offers nearly 1100 LRs available for various types of HLT applications in more than 60 different languages. A continuous and considerable effort has been put to publish quality LRs adapted to new requirements and latest trends. For instance, ELRA lastly enlarged its catalogue to offer LRs for Sign-Language, Text-to-Speech, or Audio-visual applications.

In addition to our Catalogue of ready-for-distribution LRs, the Universal Catalogue[3] offers a compilation of world-wide LRs which have been identified by the ELRA/ELDA team. This catalogue represents an antechamber to the ELRA Catalogue, as well as a shop-window of existing LRs for the community. Through the Universal Catalogue, users may discover existing but not-yet-available LRs that ELRA may help them gain access to.

As a complement to this service and with the continuing mission to support the identification of LRs, ELRA launched the *LRE-Map*[4] at LREC 2010. It is a mechanism intended to monitor the use and creation of LRs. This is implemented by collecting information on both existing and newly-created resources during the papers/abstracts submission process [Calzolari et al. 2010]. It thus provides a portrait of the resources behind the community, of their uses and usability. Nearly 2,000 LR forms were filled in in 2010. The feature has been so successful that it has been implemented by other conferences such as COLING 2010 (International Conference on Computational Linguistics) and EMNLP 2010 (Conference on Empirical Methods in Natural Language Processing) and is very likely to be adopted by other major conferences in the future.

Identifying the gaps in LRs has also been one of ELRA's ongoing tasks. ELRA continued the work that was initiated through the implementation and feeding of the BLARK (Basic LAnguage Resource Kit), which aimed at

[1] http://www.elra.info
[2] http://catalog.elra.info/
[3] http://universal.elra.info/
[4] http://www.resourcebook.eu

giving access to matrices[5] that highlighted the needed resources and helped at identifying the gaps with regards to LRs required for specific applications and for as many languages as possible. Such information is compiled and shared with the community, e.g. in November 2011, ELRA co-organised the 2nd Less-Resourced Languages Workshop (a joint LTC-ELRA-FLaReNet-META_NET event) on Addressing the Gaps in Language Resources and Technologies[6].
Along these actions, ELRA is involved in the creation of a distributed repository network, which work is being highlighted in the following section.

2.2 Towards the Integration of LR Repositories

As a step forward on the basis of its work expressed through its catalogues, ELRA is sharing its knowledge and experience by being involved in the very latest initiatives in Europe that have converged on large cooperation networks.
Since 2009, ELRA, through ELDA, is part of the META-NET[7] (Multilingual Europe Technology Alliance) Network of Excellence and is involved in the development of an Open Resource Infrastructure, the META-SHARE action. This infrastructure aims at providing an "open, distributed, secure, and interoperable infrastructure for the Language Technology domain"[8]. This consists in setting up a network of repositories/data centers accessible through a common interface.
Two main actions have taken place towards this sharing:

- Technical implementation of a META-SHARE network of repositories which have imported all LRs from the ELRA Catalogue and other players, and has thus enriched the infrastructure with over 1,000 LRs.
- Work towards the "specification of metadata-based descriptions for language resources and technologies" [Gavrilidou et al. 2011]. ELRA's work, combined with its regular internal effort to improve the description of the ELRA catalogue, has aimed to emphasize the need to review currently existing LR metadata schemas and design a standardised and interoperable one for the current needs of the community. In particular, a large effort has been invested to describe new types of LRs that include modalities such as video or image (for e.g. sign language LRs, multi-sensor and multi-modal data, etc.).

These actions aim to support META-SHARE objective and spirit towards the sharing of resources within the community, as well as the harmonization of licenses and transaction modes.

3. Supporting HLT Enhancement

3.1 Sharing Technologies

With its large experience gained through over 17 years in distributing LRs to the HLT field, ELRA has enlarged its activity to the sharing of LR-oriented technologies. One good example is ELDA's involvement in the PANACEA FP7 project [9] (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies). Work has started towards building a factory of LRs that progressively automates the stages required for the acquisition, production, updating and maintenance of LRs.
ELDA is leading the dissemination and exploitation work as well as the validation of the platform. Furthermore, we are actively involved in the setting up of the platform and the evaluation of the crawling and MT technology developed within the project.
In June 2011, ELDA also organised the first PANACEA Users' workshop[10] which aimed at gathering users of LRs for the development of their own business and technologies. The event was held in conjunction with the META-FORUM organised by META-NET, which again shows the interest in correlating actions around LRs and LR-oriented technologies.

3.2 Evaluation of Technologies

ELRA's dedication to activities in evaluation of technologies started about 10 years ago. Since then, ELRA has been paving the way to a more standardised way of evaluating HLT by offering an evaluation infrastructure and evaluation packages to the HLT community. Thanks to this expertise, ELRA managed to be a core player within innovative evaluation projects and in various technology areas. For the past couple of years, ELRA has been involved in evaluation campaigns in the following fields:

- Cross-language information retrieval and filtering
- Machine translation
- Multimedia and multilingual information systems
- Speech technologies including speech recognition for spontaneous speech
- Topic, opinion and sentiment detection
- Spoken language understanding
- Named entities recognition
- Parsing

These evaluation campaigns could be supported thanks to online automatic evaluation interfaces provided by ELRA. Within recent large-scale projects, we can quote ELRA/ELDA's participation in the PROMISE project[11] (Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation), a EU Network of Excellence that started in September 2010. PROMISE is carrying on the work that was initiated through the Cross-Language Evaluation Forum (CLEF) since 2000[12]. ELDA is responsible for Data acquisition, packaging, and IPR. Through this project, ELRA continues to play a role in experimental evaluations in the field of complex multimedia and multilingual information

[5] http://www.blark.org/
[6] http://www.ltc.amu.edu.pl/content.en.html#lrl-workshop
[7] http://www.meta-net.eu
[8] http://www.meta-net.eu/meta-share
[9] http://www.panacea-lr.eu/
[10] http://workshops.elda.org/panacea2011
[11] http://www.promise-noe.eu
[12] http://www.clef-initiative.eu

systems.
In support to the development of HLT evaluation activities, ELRA created a web portal dedicated to that topic. The HLT Evaluation Portal[13] gathers information on a large number of technologies. Facing the need to keep alive such valuable information for the HLT community, the portal is now supported by the META-NET project. Interested parties are welcome to contact us to enrich the portal with information on evaluation tasks that are not covered.

3.3 Production of LRs

In line with its regular activities, ELRA is very active with the production or commissioning the production of LRs. This takes place both within the framework of European and international projects or in support of companies or institutions.
So far ELRA has compiled LRs in more than 30 languages, being involved in every stage of production, from the establishment and definition of specifications and guidelines (e.g. multimodal annotations of videos for person identification [14]) to the ultimate quality control detail (e.g. quick quality check or validation reports).
Thus, ELRA is a privileged partner for innovative projects that require 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 for a wide variety of languages: English, "Indian" English, German, US Spanish, Catalan, Brazilian Portuguese, French, Canadian French, Moroccan French, Colloquial Arabic(s), Kazakh, Romanian, Czech, Turkish, Hindi, Korean, Chinese... Some of ELDA's recent achievements comprise: (i) broadcast news speech corpora, (ii) newspapers text corpora, (iii) aligned corpora for Machine Translation and Speech Translation, (iv) rich text annotation (named-entities, opinions, feelings, etc.), (v) single and multi-modal annotations (e.g. audio, image, audiovisual), (vi) specific data collections (e.g. SMS, written documents, Wizard-of-Oz based recordings for dialogue systems, etc.).
Out of these achievements, we can mention ELRA's participation within cooperative projects. For instance, we have been involved in the creation of manually-translated parallel corpora in different domains, ranging from medical to transcription data, for language directions such as Arabic-to-French, Chinese-to-English, English-to-German, French-to-German, German-to-English and German-to-French.
Those production activities have also contributed to enrich the ELRA Catalogue with new LRs.

4. Intellectual Property Rights in Question

4.1 Enlightening the HLT field to IPR Issues

One of ELRA's background task and creed since its creation has been the taking care of legal issues related to the exploitation and distribution of LRs. Throughout the years, ELRA has put forward the importance of clearing legal issues that have to be dealt with at each step of the Language Resource lifecycle [Arranz et al. 2008]: specifications, production, validation, distribution, and maintenance.
To enlighten HLT players' knowledge on the topic, ELRA organised a half-day workshop on "Legal Issues for Sharing Language Resources: Constraints and Best Practices"[15] [Mapelli 2010], as a satellite event to LREC 2010. In order to consider an international context from both academic and commercial horizons, the organising committee was constituted of representatives from the following institutions: Linguistic Data Consortium, USA, Institut für Deutsche Sprache, Germany, and ELRA/ELDA, France. It aimed at showing new lines of work in that field, as well as new possible cooperation topics and a good opportunity for participants to share their views on the subject.
Among many interesting issues raised during the talks and following panel discussion, the contribution of lawyers was of great value in order to understand current international legal systems as to be compared to the LR field requirements and present the variety of existing licenses (Creative Commons, GNU, etc.).
Earlier in 2010, ELRA and FLaReNet had already initiated discussions on this issue by organising a special session on "Sharing or Not Sharing: Availability and Legal Issues" at the 2nd European Language Resources and Technologies Forum, on 11th February 2010 in Barcelona, Spain[16]. This session focussed in particular on legal stumbling blocks and possible solutions towards a sustainable sharing of LRs.

4.2 Fair Research Act

Since its creation, ELRA has been promoting the need to give an easy access to LRs to the widest community, in particular for research activities. ELRA has been discussing with a great number of major research institutions in order to identify the legal issues that could block the advances of the field. From these discussions, it became clear that copyrighted resources and intellectual rights as defined in the Berne convention or the European database directive (the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases) had to be reconsidered.
Recently, ELRA and a number of partners have been advocating toward the establishment of a specific exception dedicated to "the fair use of a copyrighted work for research purposes". Such an exception should consider that no rights are infringed as long as the LRs are used exclusively for research purposes. A number of actions have been undertaken to seek harmonization of such a fair use between the countries.

4.3 Clarifying IPR

ELRA has been not only organising events for awareness but is also well involved in international actions to support the clarification of legal questions in the field.
In particular, it is worth mentioning the valuable work carried out within META-SHARE. A detailed document was produced to consider all aspects of legal issues [Choukri et al. 2011]: analysis of current existing licenses,

[13] http://www.hlt-evaluation.org
[14] http://www.defi-repere.fr
[15] http://workshops.elda.org/lislr2010/
[16] http://www.flarenet.eu/?q=S3

consideration, clarification and comparison of IPR, adaptation of licenses to the new HLT requirements. In particular, ELRA focused on the analysis of similarities between its licenses and the ones promoted by Creative Commons, with the intention to harmonize such licenses.
Moreover, ELRA has been carrying out its activities in the identification and negotiation of rights for the use of LRs within cooperative projects. Focusing on Medical Information analysis and retrieval, the KHRESMOI project [17] (Knowledge Helper for Medical and Other Information users) addresses both challenges of trustworthiness and complexity levels in online health information. This FP7 project started in September 2010. In the first year of the project, ELDA has worked at clarifying the IPR framework of the resources that are being exploited in the project both by identifying respective rights and defining lines of action to clear them all throughout the 48 month project activity. This has been done in particular through the drafting of adapted licenses and the negotiation with due owners. It is worth saying that some areas are even more challenging in terms of IPR negotiation, the medical area being one of them mainly for privacy and confidentiality reasons, another one being the TV-radio broadcast where multi-layer rights have to be considered in depth (several players/areas have to be considered, some of them being interlinked: production, distribution, broadcast). As far as the latter is concerned, the French ANR-funded REPERE project (Person recognition in TV shows)[18], which started in March 2011, has been another good IPR-clearing challenge for ELDA. For this project, ELDA has faced a multiplicity of IPR ownership interlinked between producers, distributors and broadcasters.
Recently, ELRA has extended its legal support beyond its membership base to assist producers and users of LRs. The Legal Support Helpdesk[19] is meant to provide support to those who have to deal with IPR issues while using, producing, sharing or distributing LRs.

5. Promoting Cooperation around Language Resources

5.1 Promotion and Information Dissemination

A recent initiative from ELRA towards mass collaboration around LRs is the *Language Library*, a new feature of LREC 2012. The rationale behind this initiative is that accumulation of massive amounts of multi-dimensional data about language is the key to foster advancement in our knowledge about language and its mechanisms. The objective is to gather and share part of the linguistic knowledge the field is able to produce, starting a movement aimed at collecting all possible annotations/encodings at all possible levels.
Since its foundation, ELRA has been active not only at promoting the idea of sharing LRs, but also at gathering common expertise through its participation in multiple networks focusing on LRs. In 2009, ELRA joined the European-funded FLaReNet (Fostering Language Resources Network) [20] network, which aims at "developing a common vision of the area and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide". Some ELRA Board Officers also participate in this project through their respective institutions. Within FLaReNet, ELRA has worked on the definition of standards and best practices for LRs that have been discussed at the three "European Language Resources and Technologies" forums organised by FLaReNet in 2009, 2010 and 2011. Moreover, as a recent output from FLaReNet work, ELRA contributed to the production of the FLaReNet Databook [Baroni et al. 2011] which aims at making visible a set of facts and figures on the LRs and HLT fields.
The organisation of events remains a key activity of the ELRA Dissemination pole. LREC[21] (Language Resources and Evaluation Conference) continues to be our main dissemination event, now gathering more than 1,200 experts. The participation in various European projects as Dissemination leaders has also led us to organise and take part in events over the last months, including the Panacea Users' Workshop in Budapest in June 2011, collocated with the 2nd META-FORUM, the CLEF Conference in Amsterdam in September 2011, as part of PROMISE project, etc.
Published by Springer, the *Language Resources and Evaluation Journal*[22], the official journal of ELRA is devoted to the acquisition, creation, annotation and use of LRs, together with methods for evaluation of resources, technologies, and applications.
The usual dissemination activities continue to be carried out: monthly bulletins sent to members, web sites, newsletter, etc.

5.2 Creation of a LR Unique Identifier

ELRA and a large number of Language Technologies organizations have been debating on the harmonization of the identification of LRs. A consensus seems to emerge regarding the set-up of a small executive committee, steered by a commission representing all key players in the field, data centers (ELRA, LDC, Allagin,/GSK, C-LDC,…), and the stack holders (ACL, IAMT, ISCA,…), to assign each LR an International Standard Language Resource Number (ISLRN) [Park et al. 2012], independently of whether the LR is accessible on Internet, Intranet, available or not, etc… whether it has a DOI, a local PId, etc. Such ISLRN should guarantee that all LRs usable within our field get a unique identifier that can be used to distinguish it from others.
The aim behind this proposal is to ensure the sustainability of LRs by providing them with a unique identification scheme using a standardised nomenclature. This will guarantee that LRs are recognised as proper references in the different activities within Human Language Technologies and within documents and scientific papers. For instance, this will allow resources

[17] http://www.khresmoi.eu
[18] http://www.defi-repere.fr/
[19] http://www.elra.info/Legal-Support-Helpdesk.html
[20] http://www.flarenet.eu
[21] http://www.lrec-conf.org/
[22] http://www.springerlink.com/

that are sometimes named differently to be correctly identified as unique resources, and it will help catalogues (requiring a unique identification format) to manage data correctly, regardless of the LR type and physical location. On a management point of view, such an initiative requires the setting up of a ISLRN attribution body to manage their attribution, storage and consistency. It is thus planned that the ISLRN attribution will be made by a small group of organisations involved in all LRs distribution and sharing issues. This ISLRN attribution body will set up a ISLRN server which will enable the ISLRN attribution and validation of LRs, based on a minimal metadata set that describes them. The ISLRN will be assigned free of charge.

6. Conclusions

As highlighted in this paper, ELRA continues to focus on its regular activities by keeping a close look-up on the HLT evolution and by taking part to new large-scale international cooperative projects thanks to its long-lasting expertise.

Latest trends towards open infrastructures led ELRA to renovate its policy and activities along the following main five axes:

- Easy and free access to LRs:

Anticipating users' expectations, ELRA has decided to offer a large number of resources for free for the research community. Such an offer will consist of several sets of speech, text, multimodal databases that will be released for free regularly, with a particular attention on an easy way to license them.

- Fostering the use of Public Sector Information:

Being among the first to recognize the importance of public sector information, the Association joins forces with all stakeholders in the effort to reinforce and extend the free use of public sector information for research, technology and application development.

- Supporting META-NET / META-SHARE actions:

As a founding member of META-NET and an important player of META-SHARE, ELRA will be involved in the management and standardisation of meta-data schema being built in META-SHARE. It will also support the legal helpdesk and will focus on the commonalities between ELRA licenses and the ones promoted by Creative Commons, with the intention to harmonize such licenses.

- Promoting collaborative and crowd-sourcing-based methods for building LRs:

This task focuses on the new LR production paradigm based on crowd-sourcing. During the last year, the Association has been investigating the set-up of a dedicated platform for crowd-sourcing-based LR building.

- Establishing the Language Resources and Evaluation Forum (LRE-F):

The Forum is open (and not limited) to scientists, students or professors, involved in research activities in universities, small and medium companies or international groups; decision-makers or project managers in large public institutions, etc. It has been established at LREC2012 where participants have been invited to join when registering to the conference. Among the services that members of the LRE-F will be offered, we can quote the following: free downloading of many resources from the ELRA Catalogue and the META-SHARE repository, access to the legal helpdesk, access to the LRE Map, the LR Library, access to LRE Wiki, etc. Members of the community will be also encouraged to join so to upload resources on the ELRA and/or ELRA-META-SHARE repository to share with other colleagues.

7. References

Arranz Victoria, Gandcher Franck, Mapelli Valérie, Choukri Khalid (2008). *A Guide for the Production of Reusable Language Resources*, In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC08), May 2008, Marrakech, Morocco.

Baroni P., Soria C., Calzolari N. (eds.), *The FLaReNet Databook*, with contributions from Arranz V., Bel N., Budin G., Caselli T., Choukri K., Del Gratta R., Desypri E., Francopoulo G., Frontini F., Goggi S., Hamon O., Hinrichs E., Labropoulou P., Lemnitzer L., Krauwer S., Mapelli V., Mariani J., Monachini M., Odijk J., Park J., Piperidis S., Przepiorkowski A., Quochi V., Revilla E., Romary R., Rubinoo F., Russo I., Schmidt H., Uszkoreit H., Wittenburg P., FLaReNet Public Deliverable, 2011.

Calzolari Nicoletta, Soria Claudia, Del Gratta Riccardo, Goggi Sara, Quochi Valeria, Russo Irene, Choukri Khalid, Mariani Joseph, Piperidis Stelios, *The LREC Map of Language Resources and Technologies*, In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), May 2010, Valletta, Malta.

Choukri Khalid, Stelios Piperidis, Prodromos Tsiavos, John Hendrik Weitzmann, *META-SHARE: Licences, Legal, IPR and Licensing Issues*, META-NET Public Deliverable, January 15, 2011.

Gavrilidou Maria, Labropoulou Penny, Piperidis Stelios, Francopoulo Gil, Monachini Monica, Frontini Francesca, Arranz Victoria, Mapelli Valérie, *A Metadata Schema for the Description of Language Resources (LRs)*, In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, November 8-13, 2011.

Mapelli Valérie, *Legal Issues for Sharing Language Resources: Constraints and Best Practices*, In ELRA Newsletter (LREC 2010 Special Issue), Vol.15. N.1-2, January-June 2010.

Park Jungyeul, Hamon Olivier, Arranz Victoria, Choukri Khalid, (2012), *Practical and Technical Aspects for Using the International Standard Language Resource Number*, In Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC 2012), May 2012, Istanbul, Turkey.