Language technology evaluation in Europe
Key achievements and the need for an infrastructure

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.org
***URL:* http://www.elda.org *or* http://www.elra.info**

ABSTRACT

This abstract aims at describing briefly the evolution of language technology evaluation in Europe. This will be directly linked to the activities of the European Language Resources Association (ELRA) and its operational body, the Evaluation and Language resources Distribution Agency (ELDA). The rational behind the foundation of the European Language Resources Association (ELRA) and its Evaluation and Language Distribution Agency (ELDA) in 1995 will be elaborated upon and the HLT Evaluation activities carried out since then highlighted. We would like to focus on the issues to address for making language resources available to different sectors of the language engineering community and, in particular, on those needed to carry out evaluation activities. The presentation will introduce a number of Evaluation projects and Services established through a large number of European and nationally funded projects. In addition, ELRA carries out promotion tasks in the field of Human Language Technology (HLT), in order to advertise resources and any relevant activity (with the maintenance of catalogues, the edition of its Newsletter, the organization of the LREC Conference, which is now a major event for the HLT community, and the maintenance of the HLT Evaluation Portal, among others).

ELRA's Mission

ELRA's initial mission was to set up a centralized Not-for-profit organization for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of a different nature such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights, …), information dissemination (to act as a clearing house). This mission is tuned from time to time to anticipate future requirements. As of today, this can be reflected by the following tasks:

- The identification of useful Language Resources.
- The handling of legal issues related to the availability of Language Resources.
- The Language Resource distribution activities and Pricing policy.
- The validation and Quality assessment of Language Resources.
- The commission of the production of needed Language Resources & Market watches.
- The information dissemination, Promotion and Awareness.
- **Last but not least, the supply of the Evaluation services to the HLT community.**

HLT Evaluation activities

For any HLT research effort to be successful, it is essential that it be assessed through rigorous evaluations of the developed technologies. This allows performance benchmarking and a better understanding of possible limitations and challenging conditions. Since 2000, ELDA had an objective to design and validate evaluation packages for several Human Language Technologies. An ELDA's evaluation package (or Evaluation Kit) comprises the following items:

1) An evaluation protocol specification, including specification of the task to be performed by the systems being evaluated, metrics, data representation formalisms, and the relevant documentation.
2) Development data representative of the task and in sufficient amount to enable a full validation of the evaluation protocol.
3) The test data that will be used to score systems' performance.
4) All the software tools required to run an evaluation campaign implementing the protocol defined in 1), i.e. format standardization and validation tools, measuring tools, result presentation tools, data server, storing and communication tools, etc.

These evaluation packages had to be made available for organizing large evaluation campaigns, involving all key players from laboratories developing technologies targeted for evaluation. The evaluation packages also had to be made commercially available upon request for government agencies or industries wishing to organize other evaluation campaigns. Finally, these packages are distributed for industrial or public research entities wishing to evaluate a technology (possibly the one they develop) and compare it to the state-of-the-art.

In many cases, ELDA also helps to provide data for system training and, since 2006, it provides an on-line evaluation platform for several technologies (web-service and/or UIMA-based platform) to avoid the installation of scoring tools at each site.

In order to achieve such goal, ELDA has established an evaluation Department that takes care of assessing and benchmarking Human Language Technologies both within R&D projects and for customers. In order to do so, ELDA has joined efforts with some of the largest consortia involved in HLT development and it has managed to ensure that the consortia capitalize on the evaluation work through the packaging of all needed pieces to carry out similar initiatives afterwards. A new paradigm refers to this task as the "project exit strategy". Such strategy ensures that the availability of the "evaluation package" as described above (the full documentation, definition and description of the evaluation methodologies, protocols and metrics, alongside the data sets and software scoring tools) is an essential part of each project. An evaluation package can be conditioned so that it can be distributed through ELRA’s Catalogue. This allows any organization to reproduce one of the technology evaluations that were conducted during the project Evaluation Campaign. The exploitation of this outcome is one of the achievements of the project.

ELRA’s Focus on Evaluation

The ELSE project (Evaluation in Language and Speech Engineering, 1999) conducted a study on the possible implementation of HLT evaluations in Europe and compared them to other initiatives conducted in the USA and Japan. Among the findings of this project we can quote the need for comparative evaluations conducted by a truly European infrastructure that would ensure long-term availability of expertise and resources so as to avoid the loss that occurs when projects are funded through R&D programs for a short period of time. The ELSE project also highlighted how important this was for the developers that benefit indirectly from evaluation through the acquisition of the complete evaluation toolkits and by-product data that become available afterwards but also through the knowledge sharing that takes place systematically in the post-campaign workshops during which experts compare approaches and techniques used by each system.

Within these evaluation activities, ELDA has participated in a large number of projects and initiatives which have helped reinforce its expertise in the area and have supported the development of HLT Evaluation in Europe. Among these projects we will mention just a few here to highlight the huge European investment and the crucial need to ensure a serious Return on investment through the exploitation of such packages but also through the support of the ELDA infrastructure to become self-sustainable.

A hot topic these days is Machine Translation technology, including Speech to Speech translation systems. Through its involvement in the FP6 project **TC-STAR**, ELDA has contributed to the evaluation of speech recognition systems, machine translation, and speech synthesis systems. In addition, ELDA conducted end-to-end evaluations and compared the achievements of TC-STAR systems with the work of human interpreters for English and Spanish. A review of such work will be described during the talk. One of ELDA’s tasks was the collection and annotation of huge sets of spoken multilingual corpora and the corresponding written corpora that was used to train the systems. ELDA has also been in charge of elaborating the global evaluation plan for the 3 evaluation campaigns of the project. Today, all packages are being made available to assess system performance and a number of copies have already been distributed.

Another major project worth mentioning here is **CHIL** (“Computers in the Human Interaction Loop“, an FP6 IP). The implication of ELDA made it easy to capitalize on the work conducted within the project and to ensure that all data sets and evaluation toolkits are made widely and immediately available under very fair conditions. CHIL has addressed the largest number of technology components ever done before with the goal to develop computer assistants that attend to human activities, interactions, and intentions. CHIL's thirteen technological components were evaluated, which focused on Vision technologies (Face Detection, Visual Person Tracking, Visual Speaker Identification, Head Pose Estimation, Hand Tracking), on Sound and Speech technologies (Close-Talking Automatic Speech Recognition, Far-Field Automatic Speech Recognition, Acoustic Person Tracking, Acoustic Speaker Identification, Speech Activity Detection, Acoustic Scene Analysis) and on Contents Processing technologies (Automatic Summarization and Question Answering on Spoken Transcriptions, conducted in partnership with CLEF). The corresponding evaluation packages are being made available through ELRA's Catalogue.

A further international achievement resulting from this project, and of great importance, is the establishment of the open international evaluation workshop **CLEAR** - “CLassification of Events, Activities, and Relationships”, in partnership with NIST and other players. So far, two CLEAR evaluation campaigns were conducted, partly with CHIL packages.

Another major area being tackled by most of the HLT key players is the Multilingual/Cross-Lingual Information Access and Retrieval. Through some partial European funding, **CLEF** (Cross-Language Evaluation Forum) was launched in 2000 with the aim to develop an infrastructure for the evaluation, testing and tuning of information retrieval systems operating on European languages in both monolingual and cross-language contexts and, beyond this, to experiment the setting up of a European HLT evaluation infrastructure. The project managed to create test suites of reusable data which are part of the ELRA evaluation catalogue and which are extensively employed by system developers for benchmarking purposes. The exploitation of the methodologies implemented by CLEF for the testing and tuning of information retrieval systems is now part of ELDA's assets and it allows conducting the evaluation of commercial products and applications with a strong and reliable technical and scientific background. More than 11 copies of the CLEF packages have been distributed so far.

Another important initiative, funded by a national agency, is the French programme "**Technolangue/Evalda**": the Evalda projects that ELDA has coordinated consisted of 8 evaluation campaigns with a focus on the spoken and written language technologies for the French language: ARCADE II (evaluation of bilingual corpora alignment systems), CESART (evaluation of terminology extraction systems), CESTA (evaluation of machine translation systems), EASY (evaluation of parsers), ESTER (evaluation of broadcast news automatic transcribing systems), EQUER (evaluation of question answering systems), EVASY (evaluation of speech synthesis systems), and MEDIA (evaluation of in and out-of context dialog systems). As planned and achieved within the other evaluation projects, all Evalda evaluation resources have been packaged and made available and more than 15 copies have been distributed so far.

Further to its participation in such projects, ELDA has also run a number of initiatives, such as discussion and brainstorming events. One such example is ELRA’s 10th Anniversary Workshop on Evaluation, celebrated in Malta in December 2005. The discussions initiated at this occasion have been taken further with the Evaluation Workshop celebrated during the MT Summit 2007 (Automatic Procedures in MT Evaluation), and a coming ELRA Evaluation Workshop (Looking into the Future of Evaluation: when automatic metrics meet task-based and performance-based approaches) to be celebrated jointly with the LREC 2008 Conference, this coming May-June 2008.

Some topics that will be addressed during the talk

During the talk, I will elaborate on the role of evaluation on the research progress, on the need for a truly European infrastructure for HLT evaluation and the potential role of ELRA within such structure, on the main reasons to promote an international dimension of the evaluation, insisting on the multilingual issues. I will introduce and describe some evaluation concepts (comparative evaluation versus competition, Technology evaluation versus Usage/usability evaluation). I will also describe the different types of evaluation and how to ensure that evaluation does not kill very innovative not-yet-mature approaches.

Conclusion

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by Language Engineering players are made available when they already exist or to produce them in a cost-effective frame. It is of paramount importance that regional organizations emerge and co-operate between themselves with respect to the issues described herein. The main common task would be to achieve, all together, a better streamlining of efforts in the development of new Language Resources that are of interest to “local” and “global” players. This role should be extended to Evaluation in particular in geographical areas that do not have a dedicated organization.

At the same time, the paradigm of evaluation should be reconsidered by the funding agencies and funded as a major part of their investments, as it allows both to measure if the money they have invested in technology development has led to significant progress and to identify areas where the technology needs further improvement.

Evaluation also allows application developers/integrators and end-users to understand where the technology is and how it can help them and provide them with new solutions to the problems they face.

ELRA also initiated the HLT Evaluation portal that is designed to be an online information resource about HLT evaluation and related topics of interest to the HLT community at large.