ELRA Contribution to Bridge the Gap Between Industry and Academia

Khalid Choukri

ELRA/ELDA, Paris, France
choukri@elda.fr

Summary

The European Language Resources Association (ELRA) was created in February 1995 to handle all issues related to Language Resources. ELRA's missions and activities include the collection, distribution, validation of speech, text, terminology resources and tools. Very recently ELRA has launched a new activity regarding the evaluation (of technologies, systems, prototypes, services, etc.).

After five years of activity, ELRA has managed to make available, world-wide, a large set of marketable resources, including a large number of speech databases. Since its creation, ELRA has been able to draw a number of conclusions regarding the mutual expectancies between industry and academia. On the one hand, industry expect from academia that the results of their basic research in the field of Human Language Technologies can be exploited to their "greatest" benefit, through an efficient and cost-effective technology transfer process, for the development of systems or products; on the other hand, the research sector expect to get a feedback from the real users (from the "field") and also to get feedback from the industry regarding their true concerns to better tune the research orientations.

Today's HLT technology is based on Language Resources. Language Resources are universally acknowledged to be critical for the development of robust, broad-coverage, and cost-effective applications on all sectors of HLT, in particular those for written and spoken language. Yet the cost of developing such resources is prohibitive, even for large organizations, regardless of the projected market size. Moreover, due to the lack of sufficient coordination, existing Language Resources cannot be easily adapted for multiple users, thereby hindering the rapid deployment of new applications.

The production of such resources is achieved both by industries and academia, while both may be the users of language resources. This is a kind of continual and persistent exchange between the two sectors, and, if ELRA does not "belong" to the academic sector or to the industry, it acts as a mediator between universities or research laboratories and industrial companies, which either provide the resources they have produced, or ask for some specific resources to be used in the framework of their researches or developments.

ELRA offers several other services, such as the identification of LRs, the collection of LRs, the distribution of LRs, the validation of LRs, and the evaluation (of systems, products, etc.), which are adapted to the needs of the industry and of the academia, and where fruitful collaborations could take place.

Below are some statistics which show the proportion of industrial versus academic figures related to ELRA activities: the proportion of sales according to their use, either research or commercial, and the distribution of the providers.

ELRA/ELDA counts 97 providers of Language Resources on the whole. Among these 97 providers, 4 are individuals, 46 belong to the academic sector, and 45 to the industry.

Table 1: Statistics on sales

<table>
<thead>
<tr>
<th></th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>122</td>
<td>104</td>
<td>156</td>
</tr>
<tr>
<td>Commercial</td>
<td>88</td>
<td>74</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>210</td>
<td>178</td>
<td>266</td>
</tr>
</tbody>
</table>

As for the distribution of Language Resources, ELRA has established two prices: one for research use, the other for commercial use. These distribution prices are established with the providers. Since its foundation, ELRA has advocated for the sale of resources at a lower price for research purposes, to encourage and promote the research sector. A higher price for a commercial use sounds acceptable to ELRA whenever justified and allows the provider to gain a financial return on the initial investment.

Some speech databases distributed by ELRA in the framework of the AURORA project are of much interest when one considers what industry and universities expect from each other. The Aurora project was set up to establish a worldwide standard for the feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system.

The Aurora project has released a sub-sets of the SpeechDat-Car databases (which have been collected as part of the European Union funded SpeechDat-Car project). The whole package for the SpeechDat-Car databases will be available at 60-120K€ per language for speech recognizers training. However, the subsets offered for evaluation are priced 1000 €/language and about 45 copies have been sold during the year 2000.

Another point which deserves to be mentioned in this paper concerns the international collaborations and the relationships with the public and private sectors. ELRA has established over the years to promote mutual benefits of R&D centers and industry. ELRA/ELDA collaborates with its American counterpart, the Linguistic Data Consortium (LDC), in the framework of the Network-DC project. The Network-DC (Network of Regional and International Data Centers) aims at establishing a global network through a collaboration agreement between Language Resources distribution centers, starting with a partnership between ELRA and LDC, but other agencies with similar activities, for instance Gengo Shigen Kyonyunkiko (GSK) in Japan, may join this network whenever they are established.

Network-DC is setting up new principles and practices for coordination and cooperation of Language Resources data centers, in areas from licensing terms to data distribution, and will test these innovations in the context of a coordinated data collection effort, resulting in new multilingual broadcast news resources.

To conclude, the role of ELRA as a mediator between the Language Resources providers and the Language Resources users can help academia and industry (which may be, both, provider and/or user) communicate and work together for the development of new HLT technologies and for the promotion of Language Technologies.

Table 1: Statistics on sales