

Language Resources

Go Digital

An update on the European ELRA initiative

by Jeff Allen

This article is a follow-up to First-Person Singular: Is there a Real Market for Europe's Digital Language Resources? (*Language International*, Volume 10, Issue 5, 1998.)

Language Resources, ELRA, and ELDA

For those of you who did not have a chance to read the earlier article, the European Language Resources Association (ELRA) was created in 1995 as a non-profit organization that would collect, market, and distribute language resources (LRs), as well as disseminate general information in the field of language engineering (LEs). LRs are essentially sets of language data and descriptions in machine-readable form, used specifically for building, improving, or evaluating natural-language and speech algorithms or systems. They are core resources for the software-localization and language-services industries, language studies, electronic publishing, international transactions, subject-area specialists, and end-users.

The tasks of the association are to identify and collect existing resources and to educate potential users in Europe and beyond through different media. These include a quarterly newsletter focusing on LE, a regularly updated catalogue of language resources, a Web site, announcements on various electronic lists, etc. ELRA has also simplified the relationship between LR

providers and users (customers) by drafting generic agreements specifying responsibilities and duties of each party when licensing an LR. In order to add value to the resources it distributes, ELRA initiated the production of validation manuals for each type of resource, namely spoken resources, written corpora, lexica, and terminology resources. As of today, the validation manuals for written corpora, lexica, and telephone-based speech data are available upon request, and the manual for terminology is still an ongoing project.

ELRA's steadily increasing member base (nearly 80) comprises three different groups, or colleges, divided by main fields of interest: Speech, Written, and Terminology—each of which is represented on ELRA's 12-member board. To efficiently carry out the distribution activities, all day-to-day operations are run by ELDA, ELRA's operational body. Incorporated as a company and directed by the ELRA CEO, ELDA handles all commercial and business-oriented tasks, and develops and executes ELRA's strategies and plans (i.e., identifying, classifying, collecting, validating, and distributing LRs). In particular, ELDA also handles the professional, practical, and legal concerns of ELRA members and suppliers.

Distribution of LRs

During the first two years, ELRA was predominately focused on the establishment of the infrastructure and identification of

Distribution According to Membership Status:

	1996	1997	1998
Members	11	37	173
Non Members	6	12	31
TOTAL	17	49	204

Distribution According to the Use of LRs:

	1996	1997	1998
Research	12	21	116
Commercial	5	28	88
TOTAL	17	49	204

Distribution According to Geographical Origin of Client:

	1996	1997	1998
Europe	13	40	118
Outside Europe	4	9	86
TOTAL	17	49	204

Sales According to the Type of LRs:

	1996	1997	1998
Speech & Related Resources	16	32	127
Written Resources	1	16	75
Terminological Resources	0	1	2
TOTAL	17	49	204

valuable resources. This explains the low take-off of sales in 1995-1996. The tables provided in this article show the progressive situation of LR distribution via ELDA. They indicate distribution statistics over the last three years, comparing ELRA members to non-members, European to non-European clients, sales within the different colleges (speech, written, terminology, and tools), and the resources distributed for commercial vs. research use.

The ELRA/ELDA fiscal year used to be September through the following October instead of a calendar year. This implies that the following figures are different from the previous ones which were adjusted to the calendar year to make it more easy to understand. Major efforts were devoted to the distribution of LRs which led to a 577 percent increase in our 1997-98 sales over 1996-97. Sales amounted to over 700,000 ECU with 179 items sold in 1997-98, compared with 120,000 ECU for 31 items sold in 1996-97. These figures clearly state that the need for LRs is on the rise and that ELRA/ELDA has been filling the void. Although the majority of past income has been from spoken language resources (86.7 percent of our revenue was generated by speech products, 13 percent by written resources, and 0.3 percent by terminology resources in 1998), ELRA/ELDA has been emphasizing the promotion of written and terminological LRs in 1998 and 1999.

In addition to sales of LRs, the ELRA catalogue is constantly growing with new resources being added on a regular basis. Announcements of such additions are indicated in the monthly ELRA members' news bulletin, the quarterly *ELRA Newsletter*, and several email lists. As of August 1999, our catalog contains a total of 96 speech resources, 48 monolingual lexica, 108 multilingual lexica, 20 written corpora, and 275 terminological databases.

Production of New LRs

In response to the growing demand for LRs, ELRA issued a series of calls for tenders and proposals in 1998 and 1999 to help sponsor the production, and/or the packaging or customization of existing LRs, to fill the needs of the LE community. The intended purpose of these calls has been to ensure that necessary resources are developed within an acceptable framework (in terms of time and legal conditions) by the LE players. These calls target projects with short time scales (duration of up to one year) and modest funding. ELRA funding aims to be an effective tool for producers being both

tactical in their aims for the targeted market, as well as strategic with regard to content and annotation techniques in order to fulfill these needs. ELRA is not supplanting the European Commission nor the national agencies involved in the strategic, long-term creation of resources and infrastructures. Rather fitting within the framework of the D el egation G en erale XIII actions for revitalizing the LE field, ELRA supports the packaging and customization of small sets of key resources that might not be supported through the

ELRA is not supplanting the European Commission nor the national agencies involved in the strategic, long-term creation of resources and infrastructures.

Rather fitting within the framework of the D el egation G en erale XIII actions for revitalizing the LE field,

ELRA supports the packaging and customization of small sets of key resources that might not be supported through the larger European programs.

larger European programs (e.g., Language Engineering, Human Language Technologies).

For the production of new LRs in the framework of the February 1999 ELRA Production and Packaging call, ELRA indicated several key spoken and written resources from 1997-1998 market monitoring efforts in the call for proposals. ELRA subsequently received nearly 30 proposals for the call of proposals alone. The selection process and negotiation procedures undertaken for all three calls have led to nine projects that are now receiving partial or full funding for the development of new LRs. These include monolingual and multilingual corpora, multilingual lexica, and speech databases. ELRA hopes to launch additional calls for tenders and proposals in the future with a similar ob-

jective of promoting the production and (re)packaging of LRs.

Information Dissemination, Promotion, and Awareness

ELRA launched the Language Resources and Evaluation Conference (LREC) in Granada, Spain in 1998. With over 500 attendees, this very specialized conference was a major success. Nearly 200 papers were accepted and appeared in the two-volume set of LREC98 proceedings that is still available through the ELRA/ELDA office in Paris. The second conference of the series (LREC2000) will take place in Athens, Greece from 31 May through 2 June 2000. Further details are available at: www.icp.grenet.fr/ELRA/lrec2000.html

One of the other means to make ELRA more visible consists of a quarterly newsletter, issued in French and English. Articles regularly appear in the *ELRA Newsletter* on many topics including: speech recognition, speech synthesis, machine translation, translation memory, controlled language, terminology management, minority language engineering, evaluation and validation. These articles focus on the importance of LRs in relation to such systems and applications.

For more information visit: www.icp.grenet.fr/ELRA/home.html

Jeff Allen is the technical director of ELRA/ELDA. Email him at jeff@elda.fr

