First-Person Singular

Is there a real market for Europe's digital language resources?

Initial EC funding for the European Language Resources Association (ELRA) was due to end in September. Language International *asked Khalid Choukri to assess progress on making a business out of other people's digital property.*

By the end of the 1980s, the European Commission and various national agencies had encouraged and funded nearly a decade of in-depth language-engineering research and technology-development projects. But the extensive resources developed for these projects (which often lasted three to five years) were usually lost or so customized to a specific application that they were hard to recycle: dictionaries, samples of recorded speech, word lists, small corpora of texts, even terminology moldered on abandoned hard disks or forgotten tapes. Yet once a new project or research program was started, similar materials were built up from scratch all over again—an expensive, time-consuming, and ultimately unnecessary process.

The research community and its sponsors therefore decided to act on this issue of wasted resources. Largely under the influence of Brian Oakley, who led the UK's National Speech and Language Technology Program and was named chairman of the European Language Resources Steering Committee, ELRA was created in March 1995 in Paris with the mission of providing a central clearing house for digital language resources. Khalid Choukri, an experienced speech-technology engineer from the telecom industry, was appointed CEO.

The first issue was to decide on the most appropriate legal structure for an organization devoted to managing other people's property. "We decided on an association of membership," says Choukri, "and started with 16 members each paying an annual subscription which gave them the right to buy at a reduced price various databases of language resources made available by research organizations from academia and industry."

The Resource Broker

The basic idea was for ELRA to sign a contract with a resource provider (for example, someone with a database of Spanish spoken language, or a bilingual glossary of financial terms), giving ELRA the right to supply that database to another party, via another contract with the user. ELRA therefore became a broker for language resources, and in doing so met a genuine need. Sixty percent of its members recently said they still need ELRA as a middle man.

Choukri notes: "As R&D teams increasingly tried to share existing resources, a French lab might have found itself trying to negotiate bilaterally with a German or UK company. This meant trying to understand legalese in an effort to deal with the serious copyright and Intellectual Property Right (IPR) issues governing the exchange of such material. In the end, they preferred to hand over this task to ELRA, especially since different countries have their own copyright laws. The whole thing was too time-consuming for researchers to handle, and none of them, not even the industrial labs, saw themselves making money out of their resources. This offered ELRA a niche for adding value to other people's property."

In fact, when ELRA began, some members refused to contract with an association, so the organization had to set up a company, now known as ELDA, to handle practical matters on behalf of ELRA. Today, contracts are signed between ELDA and those who purchase the data. Members join and subscribe on an annual basis, receiving a 10- to 70-percent discount off the list price of the data.

ELRA's contract with providers usually specifies the potential usage of data in question. This is because certain IPR owners do not want their resources to be distributed for the purposes of technology development, but only for pure research. Others offer their resources at a low price for research use but raise the price for use in technical development. Yet others want their

"WHAT WE NEED IS A MICHELIN GUIDE FOR RESOURCES, WITH A RATING SCALE PROVIDING OBJECTIVE JUDGMENTS OF THE FORM AND CONTENT OF ALL LANGUAGE RESOURCES."

data to be distributed only to European institutions, since they developed it under European Commission funding and cannot give it away outside the EU. Europeans appear to be extremely close-fisted about their language materials, many of them created with public funding in the first place.

UK National Corpus

Choukri: "We are currently confronted with the problem of the UK National Corpus, which is one of the largest collections of English text and transcriptions of speech in the world. Since the UK Department of Trade and Industry funded the Corpus to the tune of several million pounds, they are trying to hang on to it as much as they can. At first, they didn't want it to be released outside the UK, but finally they've extended access to Europe. We are now pressuring them to distribute it outside Europe so that researchers in countries such as Australia can use it."

Although the potential for duplication and redistribution of a database purchased under an ELRA contract exists, no one can physically stop the user copying it. Khalid Choukri is happy to acknowledge that so far no one has infringed the law of copyright or IPR on the association's materials.

Three Colleges

The first two areas for resources that ELRA handled were speech and text. A third so-called "college" was added later, covering terminology. Speech resources are used in such processes as speech processing (recognition and synthesis), speaker verification, and emerging spoken-dialog systems. Text basically covers corpora of written documents for information retrieval applications and computer-aided translation.

Largely as a result of trying to respond to the needs of his customer-members, Choukri is now interested in two other domains which take ELRA slightly further away from its core language-resources task: multimedia information and image data. Image resources are used to explore such applications as face recognition and movement, but also intelligent optical-character recognition where lexica can play a role in disambiguating textual material.

Speaker Verification

"People working on speaker verification through voice recognition realize that they can boost the quality of their systems by building in face recognition, for example. This opens a new market for image data, and since we have to earn a living, we see it as part of our duty to offer access to such resources."

What sort of organizations join ELRA? Choukri is impressed by the growth in membership since the 16 founding members first signed up. "In November last year, over 80 percent of those who participated in the Annual General Meeting were from industry, with representatives from Siemens, Philips, Lernout & Hauspie, Sharp, Ericsson, Xerox, and France Télécom. This suggests that we attract key players with a real need for a resource broker between users and resource developers," Choukri notes.

Quality, Please

One recurring issue in the creation of a resources market is the interrelationship between format standards and quality. In principle, ELRA can better position itself as a market maker in resources if it can play some relevant role in ensuring quality and technical standards for the products it brokers.

Due to the variety of data types in play, the question of "product" standards is somewhat complex. Speech researchers, for example, already benefit from clear standards for their data, mainly as a result of an Esprit program (called SAM) which guided the data format for speech resources. In any case, the very telephone networks that carried the data were themselves standardized, making the process of agreeing on relevant standards relatively straightforward.

OTHERS WANT THEIR DATA TO BE DISTRIBUTED ONLY TO EUROPEAN INSTITUTIONS. EUROPEANS APPEAR TO BE EXTREMELY CLOSE-FISTED ABOUT THEIR LANGUAGE MATERIALS, MANY OF THEM CREATED WITH PUBLIC FUNDING IN THE FIRST PLACE.

In the areas of written text and terminology, the situation is different, but evolving quickly. Common formats such as SGML and the international Text Encoding Initiative (TEI) are emerging as widespread formatting standards for the description and sharing of documents. European players have also made an important contribution towards shared encoding standards through the Eagles program, which Choukri considers "one of the most successful initiatives of its kind."

For terminology, however, the move towards standards in formatting and exchange is a more complex process. "There are now plenty of new initiatives to simplify existing terminology standards," says Choukri, "and a number of organizations are involved, for example in the LISA-promoted Oscar project for translation-memory standardization and the Martif and Martif Lite efforts."

ELRA can best play a role in the standards process at another level—that of establishing best practice in the exchange and sharing of already-standardized resources. In other words, leveraging market standards.

Says Choukri, "Raising the legal issue of who owns what is a way of making people aware of the copyright issues that are often ignored by researchers and users. You simply cannot copy data and think it yours, and we must establish practice in the language-resources area in this respect."

Michelin Guide

A further ELRA ambition aimed at boosting the marketability of resources is to design a useful method of validating them for content quality. "Even if we don't all share the same technical standards, we should allow people to validate the data that is available with respect to its design and specifications. What we need is a Michelin Guide for resources, with a rating scale providing objective judgments of the form and content of all language resources."

Since validation procedures for what makes a good resource differ widely between speech, text, and terminology, ELRA is subcontracting the work to partners. "We don't want providers to validate their data, nor do we want to do it ourselves."

"IF LANGUAGE RESOURCES OFFERED A GENUINE BUSINESS OPPORTUNITY, SOMEONE WOULD CERTAINLY HAVE ALREADY SET UP THE MARKET EQUIVALENT TO ELRA WITHOUT ANY NEED FOR EUROPEAN-COMMISSION FUNDING."

"In speech, we are working with a Dutch institution for the validation manuals. Basically, this means that ELRA will have a data set (e.g., a database of 1,000 speakers of German saying 10 sentences each) together with the description of the data. Our job, through our validators, is to see whether or not there is a mismatch between the objects."

For lexica, the validation process is somewhat different. "You say that your Danish lexicon has 20,000 entries, with the nouns and verbs tagged for morphological and syntactic features. Our job is to see if the tags are correct, so we will need experts in the language in question. This could be an expensive job, so we shall use statistical quality-sampling processes."

Market Making

When ELRA started its job of brokering language resources for the emerging language industry in Europe, the plan was first to create the market, building on the existing informal exchanges and modeling them into a structured activity. Three years later, admits Choukri, "We are still in this same situation. We are doing sales, but the markets are very different in our three colleges."

In the speech area, people are more aware of standards, which means that a German database can easily be sold in the US. The technical aspects of the resources are also relatively easy to grasp, and telecommunication companies with their large potential markets for voice-operated calling systems are good customers.

On the other hand, it is much harder to convince the people who build and use lexica that there are market opportunities for their wares. "Someone wishing to purchase the rights to a lexicon wants to know whether it can be used out-of-the-box, or whether they will have to spend costly man-hours adapting to their specific needs." The best solution in such cases is to encourage working partnerships between users and providers, since the data has to be customized.

"We had a case where an international company needed a French dictionary with synonyms in a very specific format. Some of the information in the available database was useless to them, while the words lacked semantic data that they wanted. So they had to delete part of the information and add other information. This cost them a further Ecu 25,000 (US$30,000). Not a great deal of money, but rather time-consuming."

Emotional About Ownership

Traditional dictionary publishers, who naturally have enormous lexical resources built up over many years of work, are very emotional about their ownership rights, and they are extremely reluctant to put their raw materials on the market. ELRA has no mainstream dictionary producers among its members.

Given the problem of piecemeal, partly-coded datasets, ELRA is trying to think in terms of lexical and corpus resources that are generic enough (yet also rich enough) for handy customization. One key initiative here is to distribute the lexica coming out of the European Commission 'Parole' project, which will cover 15 European languages with a minimal set of generic information on morphology and syntax for some 25,000 entries. If people want more specific grammatical data to be attached (semantics, etc.), then they can negotiate directly with the lexicon producers. "In ELRA, we have to strike a balance between providing the specialized language information different users need from a base of single resources. Our best bet in responding to the market, therefore, is to have rich generic resources that can be customized extremely quickly."

Public Service First

As its initial funding period comes to an end, ELRA is increasingly convinced that its role should be that of a public service rather than a private business. As Choukri says, "If language resources offered a genuine business opportunity, someone would certainly have already set up the market equivalent to ELRA without any need for European-Commission funding."

He feels that if the European grant ends for good, there will still be enough members and customers to keep the association's work going. But he will nevertheless have to apply the famous subsidiary principle and convince national agencies (he has his eye on Italy and the Netherlands) to take over as ELRA backers.

But the one guaranteed present source of national support, not surprisingly, is the French government. France has a vested interest in keeping the invisible hand of the market out of the language arena and is committed to providing further support. ELRA already receives grants from the DGLF (the French Language Authority) and the Ministries of Culture, Education/Research, and Industry.

"France is becoming one of the most active players in Europe in this area," claims Choukri. "Prime Minister Jospin has promised that France will work more closely with ELRA, and the French view of the language industry is that multilinguality should not be a purely market-driven affair. With French backing, we therefore hope to play a much more interesting role in the information society as a whole."

Contact

www.icp.grenet.fr/elra/home.html

ELRA Membership Costs

Non-Profit Organizations (academic R&D):	Ecu 750 (ca. US$750)
Small and Mid-Size Enterprises (under 50 staff):	Ecu 1,000
SMEs with Over 50 staff:	Ecu 1,500
Non-Europeans:	Ecu 5,000