Multi-Lingual Spoken Language Resources

Khalid CHOUKRI, Shuichi ITAHASHI†, *Lori LAMEL*††, *Mark LIBERMAN*‡

ELRA, c/o CL International, 46 Grand Rue, L-1660, LUXEMBOURG
†University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305, Japan
††LIMSI-CNRS, 91403 Orsay, France
‡LDC, Univ. of Pennsylvania, Philadelphia, PA 19104-6305

1 Introduction

Linguistic Resources (LR) are universally acknowledged to be critical for the development of robust, broad-coverage, and cost-effective applications on all sectors of telematics, 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 LRs cannot be easily adapted for multiple users, thereby hindering the rapid deployment of new applications. To address this problem, several activities have been launched over the last 5 years, in particular Coordinating Committee for Speech Databases and Assessement (COCOSDA), the European Language Resources Association (ELRA) launched in early 1995, and the Linguistic Data Consortium (LDC) founded in 1992. In the remainder of this document we summarize the goals of these initiatives and give a snapshot of the types of resources publicly available.

2 Coordinating Committee for Speech Databases and Assessement

COCOSDA is an international group with representatives from Europe, North America, East Asia and Oceania. The role of COCOSDA is to provide a forum for discussion and exchange of information about standards, database design, assessment methods, and database availability. For example, coordination of plans for Polyphone (multi-speaker telephone speech databases) in several countries has been achieved through the medium of COCOSDA. COCOSDA has three Working Groups, on Synthesis, Recognition, and Corpora, as well as a central coordinating committee.[1] The email contact address for the Cocosda convenor is myl@unagi.cis.upenn.edu. More information can be found at http://www.itl.atr.co.jp/cocosda. COCOSDA meetings have been held as satellite workshops to the *Eurospeech* and *ICSLP* conferences since *Eurospeech'91* in Genoa. The most recent meeting was held in Sept. 22, 1995 just after the Eurospeech conference in Madrid. The next meeting is scheduled for Oct. 7, 1996 in association with *ICSLP'96* in Philadelphia, USA.

3 Oriental COCOSDA

At the COCOSDA meeting in Yokohama, Japan, on Sept.23, 1994, it was proposed that the oriental countries to set up an organization through which people concerned can exchange ideas and discuss regional matters on spoken language processing. This organization would take into account the following points:

1. Regional problems should be settled by regional effort
2. There has been growing interest in Oriental languages from Western countries
3. There are many peculiarities in Oriental languages which are different from European languages (there are many varieties, they belong to different language groups; they use different orthographic systems such as Chinese characters, Korean syllabic alphabets and Japanese Kana alphabets; there are various systems of romanization; correspondence between orthography and phonemic or phonetic description is not necessarily clear; Oriental countries have regional continuity).
4. There already exist several organizations in each Oriental country but they do not have any coordinated mutual communication.

[1] Mark Liberman (UPenn, USA) has recently succeeded Adrian Fourcin (UCL, UK) as the convenor of COCOSDA. The European contribution to COCOSDA activities has been partially financed by the LRE project 62-057 EUROCOCOSDA.

Participants agreed to the proposal that Oriental countries should have an organization which coordinate problems related to speech corpora, speech recognition, speech synthesis and speech I/O systems assessment methods which includes representatives from existing organisations such as Chinese COCOSDA, Korean Coordinating Committee for Spoken Language Processing, and the major entities involved in language resources in Japan.

4 European Language Resources Association (ELRA)

The European Linguistic Resources Association was established as a non- profit organization in Luxembourg in February, 1995. The overall goal of ELRA is to provide a centralized organization for the validation, management, and distribution of speech, text, and terminology resources and tools, and to promote their use within the European telematics R&TD community. A non-profit organization, ELRA aims to serve as a central focal point for information related to language resources in Europe, It will help users and developers of European language resources, as well as government agencies and other interested parties, exploit language resources for a wide variety of uses. It will also oversee the distribution of language resources via CDROM and other means and promote standards for such resources. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world.

Language resources include such materials as recorded speech databases, lexicons, grammars, text corpora, and terminological data. ELRA should serve as the repository for a large percentage of linguistic tools and resources developed in different frameworks: public and private actors; past, ongoing and future projects; nationally and internationally funded activities, in particular EU-funded R&D.

During the first year of the project, ELRA will:

- Set up and put into operation an organizational infrastructure for identifying, classifying, collecting, validating, and exploiting European LR;
- Set up a central distribution unit to manage and oversee the activities;
- Set up a group of expert panels, which will assist the CEO in the crucial aspects of operations such as collection, validation, distribution of data, and relationships with other bodies;
- Set up a network of validation units for each of the three major area: speech, written, and terminological LR;
- Address the fundamental organizational, technical, economic problems which constitute the crucial barriers to the development of the market of LR.

ELRA has set the yearly membership fee at 1000 Ecu. This fee will enable members to purchase corpora at a reduced member price. As of September 1995 there are 66 members of ELRA. To find out more information about ELRA contact the executive director Khalid Choukri (email: choukri.acsys.croisix@gmail.gar.no). ELRA information is also available on the RELATOR web pages (http://www.XX.relator.research.ec.org where XX is the two-letter country code of the EU countries (Germany=de, United Kingdom=uk, France=fr)).

5 Linguistic Data Consortium (LDC)

The Linguistic Data Consortium (LDC), founded in 1992, is an open consortium of universities, companies and government research laboratories that creates, collects and distributes speech and text databases, lexicons, and other resources, in support of research and development in human language technologies. The University of Pennsylvania serves as the LDC's host institution, and acts on behalf of the consortium as the holder of necessary licenses and other intellectual property arrangements.

Consortium membership is on a yearly basis, costing $2,000 in the case of not-for-profit members, and $20,000 in the case of for-profit members. Each year's membership gives the joining institution a license to all the databases released in that membership year, and membership is open at all times to any bona fide applicant. Not-for-profit members get a research-only license, while for-profit members get a license that is as broad as the LDC is able to make it, including royalty-free commercial use if possible. In most cases, individual databases are also offered to non-members for a fee ranging from $25 to $10,000, depending on the database in question.

The LDC comprised 92 members in the 1995 membership year, up from 65 members in 1992. More than a third of the LDC's members are from outside the U.S., mainly from Europe. About a quarter of the members are companies, with the remainder being universities or government laboratories. In addition, the LDC has provided data to more than 250 non-member customers.

Most LDC data distribution is now by means of CD-ROM, but a WWW-based search and retrieval service for members is now being provided. The current LDC catalogue of databases includes 64 databases, comprising 280 CD-ROMs. Available databases include text in English, French, Spanish, German, Mandarin, Japanese and a number of other languages; speech (both telephone and wideband) in English, Spanish, Mandarin, and Japanese; and lexicons in English, Spanish, Mandarin, Japanese, German and Dutch. Multilingual parallel text corpora, information retrieval test corpora, and several different kinds of speech corpora are featured. The LDC is sponsoring on-going development of text, speech and lexical databases in many languages, and also continues to act as a publishing and distribution agent for databases developed elsewhere.

More information about the LDC (including the current catalogue of available databases) can be found by consulting the LDC Web page at URL http://www.cis.upenn.edu/ ldc or sending email to ldc@unagi.cis.upenn.edu.

6 Conclusion

Several international initiatives for the construction, validation and distribution of spoken language resources have been launched in recent years. These activities highlight the importance of such resources for the advancement of spoken language technology, both for industrial and research needs. However, there are many outstanding challenges related to spoken language resources. These include design methodology (how to design reusable resources), methodologies for standardization, validation and evaluation of resources, normalization of associated documentation, dissemination media and resolution of legal issues such as IPR and licensing agreements. It is in the interest of the entire community that ongoing initiatives collaborate the coordination of resource production and reutilisation, leading to common actions in the near future.