International Cooperation in the Field of Language Resources and Evaluation

Antonio Zampolli, Istituto di Linguistica Computazionale del CNR, Italy

1. Background

The issue of international co-operation was extensively discussed at the first LREC in Granada (1998), with emphasis on the following issues:

- Language resources (LR) are essential components of HLT activity, supporting research, system development and training, and evaluation in both the mono- and multilingual context.
- A key enabling condition of integration of different technologies and languages requires that LR are shared among different sectors and applications.
- The richness of the multilingual capabilities associated with a language depends on the number of languages for which adequate LR exist.
- The high cost and effort of the production of LR should be shared, in order to make them more affordable. The creation of multilingual LR requires agreement on a co-ordination policy, to ensure the reuse of existing monolingual resources and to facilitate access to native speakers of the various languages.

The situation in the field of evaluation is rather different in Europe and in the United States, where American and European expertise seem to be complementary. The question of co-operation in the field of evaluation therefore arises very naturally, in particular because many experts believe that it is often only through such evaluations as TREC and MUC that research finds a common focus and makes easily quantifiable progress.

Three events of the first LREC have particularly stimulated discussion on these topics:

(1) the Panel on "Co-operation between EU and Other Countries in the Field of Language Resources and Evaluation" [see A. Zampolli, "Panel of the Funding Agencies", in ELRA Newsletter, Vol. 3, No. 3 (August 1998, Special Issue on the 1st LREC)];

(2) the Panel on "International Co-operation" [see A. Servantie, Panel on " International Co-operation", in ELRA Newsletter, Vol. 3, No. 3 (August 1998, Special Issue on the 1st LREC), p. 12];

(3) the Closing Session of the post-Conference Workshop on "Cross-lingual Information Management" [see E. Hovy, A. Zampolli, "Governments: Policy and Funding", Chapter 10, in E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, (Eds), "Multilingual Information Society: Current Levels and Future Abilities", to be found at http://www.cs.cmu.edu/~ref/mlim/index.html].

The following areas of Language Technology emerged in the Granada debates as being in urgent need of international co-operation:

- Standards: de facto, best practices.
- Language Resources and Related Tools.
- Core Technologies.
- Evaluation.
- Selected vertical sector domains.

These aspects were endorsed in the session dedicated to HLT at the International Conference on "New Vista in Transatlantic Scientific and Technical Cooperation," organised on the occasion of the signing of the transatlantic technical and scientific co-operation agreement (Washington DC, June 1998).

2. Objectives of the Panel

The panel aimed, in a sense, at putting together the main issues which were the focus of the first LREC events quoted above: a survey of the current programs, initiatives and underpinning policies of the Funding Agencies in different parts of the world, a discussion of the needs and opportunities for a world-wide co-operation in the field.

3. Overall Structure of the Panel

The panel was structured in four parts: introduction, panelists (presenting the situation in various parts of the world), discussants (commenting on specific issues) and a general discussion involving the audience.

3.1. Introduction

Antonio Zampolli (University of Pisa, ILC-CNR)

GENERAL FRAMEWORK

The issue of international co-operation was extensively discussed at the first LREC in Granada (1998), with emphasis on the following issues:

- Language resources (LR) are essential components of HLT activity, supporting research, system development and training, and evaluation in both the mono- and multilingual context.
- A key enabling condition of integration of different technologies and languages requires that LR are shared among the different sectors and applications.
- The richness of the multilingual capabilities associated with a language depends on the number of languages for which adequate LR exist.
- The high cost and effort of the production of LR should be shared, in order to make them more affordable. The creation of multilingual LR requires agreement on a co-ordination policy, to ensure the reuse of existing monolingual resources and to facilitate access to native speakers of the various languages.

The situation is different in the field of Evaluation in USA and in Europe.

- The complementarity of expertise can be an issue of co-operation.

Many experts believe that it is often only through such evaluations as TREC and MUC that research finds a common focus and make easily quantifiable progress.

INTERNATIONAL AND NATIONAL FUNDING AGENCIES

- The interest of national and international Funding Agencies in the social, economic, industrial and strategic impact of HLT has decisively contributed to the directions of evolution of our field.
- This interest is bound to grow in the current context of the global Multilingual Society.
- HLT (in particular, LR) involves not only R&D issues but also cultural and political aspects.

INTERNATIONAL CO-OPERATION

As the Speech and NLP field matures, as technology is increasingly commercialised, international co-operation is increasingly important. It:

- enhances advance in the state of the art by combining more effectively the strengths and excellence developed in different regions;
- facilitates the integration of LT across languages, surely one of the key aspects which makes this field relevant to the society at large.

In the light of such arguments, the US government and the EC have recently (June '98) signed an agreement for scientific and technological co-operation.

HLT has been (one of) the first sectors to implement this agreement.

MULTILINGUAL LR

In particular, the production of multilingual LR poses:

- research issues and challenges;
- organisational problems:

who has the responsibility of promoting the co-operation of R&D communities speaking different languages and how this should be done?

The situation is different for:

- types of LR: corpora, lexicons etc.;
- large/general multilingual LR;
- applications specific LR;
- customisation;
- different types of information (data VS analytical/interpretative features).

TOPICS FOR DISCUSSION ON INTERNATIONAL CO-OPERATION

- Needs, themes, priorities:
  - for HLT,
  - for other IS sectors,
  - for different types of LR/EV,
  - for different phases of LR development (research standards; specifications; construction; maintenance; updating; technology transfer; etc.);
- reasons;
- different roles, responsibilities, challenges.

3.2. Panelists

Roberto Cencioni (European Commission, DGXIII, E4)

The core of the mission of the Units he is heading in Luxembourg is to promote advanced technologies for HLT and natural interfaces to access, assimilate and use multimedia content.

The programs include both spoken and written language(s) and address human-computer interaction, interpersonal communication, information management and encompass R and D, demonstration and market stimulation activities.

International co-operation is - so to speak - directly built in the very nature of the programs: they represent widely recognised focal points: 200 million Euro have been dedicated since 1992 and 90 projects have been supported since 1997. By the end of the year 30 new projects will be underway, involving about 400 participants from more than 20 countries.

International co-operation is, from this point of view, "easy", because multination, multiparty collaborations are the norm.

This approach is rather "Eurocentric" and can be compared with the world-wide approach of the US Agencies, which have come to realize the potential of multilingual ITC.

International co-operation is essential for LR: it will be more and more important to take into account that affiliated and new accession countries bring their languages with them.

Another crucial issue is the co-ordination between EU and national activities: in particular, it is obvious that the EU can not, alone, support the development of adequate LR for all the European languages. Initiatives and proposals in this direction will be welcome.

LR are an essential component for reaching the targets of the programs, determined by the overall social and technological framework.

E-commerce should provide instant access to global markets; business should speak the language of the customer; mobile communications, wireless multimedia etc. provide new opportunities for e-business.

Internet is increasingly multilingual: 50% of surfers speak languages other than English and bi- and multi-lingual Web sites are slowly becoming the norm. We should move towards an inclusive Information Society, overcoming exclusion factors due to language, culture, computer literacy, disabilities etc.

Today enterprises should be IT and knowledge bound: hence, the relevance of content-based and cross-lingual information.

The LR, needed for as many languages as possible, can be built and made available only through concrete international collaboration activities.

At all project levels, collaboration with third countries is unproblematic, per se, if matching resources are available. In fact, HLT has been the first IST sector to launch joint programs (currently five) with NSF, following plans discussed at the first LREC in Granada: it should be noted that less than one year elapsed between these discussions and the first call including co-operation with NSF.

From another point of view, international co-operation proves, in concrete terms, to be "difficult".

Government and agency level collaboration presupposes well-established programmes on each side, similar policy and research agendas, ambitious and sizeable endeavours, balanced participation and synchronized operations, good will and personal trust at personal level, continuity over time.

It will be very interesting to hear what the situation is in other parts of the world.

Lynette Hirschman (MITRE Corporation, Bedford) presented the US perspective, speaking also for *Gary Strong (DARPA, Washington, former NSF)*, unable to participate, as intended, for managerial duties.

The vision of US technology directions, as defined by DARPA, is to move beyond document access, towards providing "just-in-time", "just-right" information to the user: the goal is to connecting the user with world class expertise via natural, conversational interaction with on-line, distributed resources. These resources may be free text, broadcast news, formatted databases - or other people with appropriate expertise or information. The information must be presented to the user in the appropriate form (short answer, graph, table, summary) and in the appropriate medium. By providing conversational access over mobile devices, we can bridge the digital divide, making Internet connectivity globally available. By focusing on the issue of multilingual and spoken language access, we can begin to bridge the language divide, providing translingual processing for the major world languages and preserving cultural heritage for non-written and minority languages.

DARPA's two major human language programs address these goals. The DARPA Communicator focuses on a plug-and-play architecture for conversational interaction to distributed resources. It is making available an open-source implementation of this framework (http://www.fofoca.mitre.org), and has put into place a DARPA Affiliate structure, to encourage international collaboration. The DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program focuses on translingual information access. Major goals are speech-to-speech translation, a toolkit to develop machine translation capabilities in a day or a week, and translingual question answering systems (see http://www.darpa.mil/ito/research/tides).

These research programs, together with other international programs, such as the joint US-EU Multi-lingual Information Access and Management (MLIAM) program, and the developing Western Hemisphere Alliance for Information Technologies program, are funding the creation of shared infrastructure and resources. In addition to these opportunities, many opportunities for informal sharing or exchange of resources exist through the Linguistic Data Consortium, through open source tools, and through the extensive series of technology evaluations supported by DARPA that are open to international participation.

Jun'ichi Tsujii (University of Tokyo) presented his view of the Japanese situation.

Mutual understanding is an essential prerequisite for international co-operation to be fruitful. Each region has its own historical and cultural background, which influences research interests and the whole direction of research projects. In his talk, Tsujii briefly summarized the Japanese experience from the early '80s till now and explained what

kinds of research programs are under way now in Japan and why. In particular, he emphasized that the Japanese research community has focused on basic generic NLP techniques throughout the '90s after the period of exploratory integration of basic techniques of the '80s. As a result, the Japanese community now feels to have reached the stage where another integration of basic technologies will be fruitful as well as possible. This type of research, i.e. exploratory integration needs public support for close international co-operation, while basic research of generic technologies as well as application-oriented development can be pursued in a looser co-operation form.

International co-operation in NLP seems more difficult than in those sciences such as brain science, physics, human genome, space science, etc. This is because our field is more tightly linked with social goals of individual countries as well as commercial interests of private sectors. Therefore, natural fields of co-operation would be in those fields independent of particular applications. International co-operation will be increasingly important in the field of collection/gathering and integration of multi-lingual resources, which support exploratory integration of basic technologies in the early 21st century.

Feng Zhiwei (State Language Commission of China, Beijing; currently at the University of Trier) presented a detailed inventory of LR (Text Corpora, Tools for Corpus Processing, Machine Dictionaries, Grammar Knowledge Base, Terminology Data Bank) available or under construction for Chinese, discussed channels of Chinese government funding for HLT, investments of private companies and the needs and opportunities for international co-operation.

Chinese language is the most important language of Sino-Tibetan language family. Now nine hundred forty million people in the world speak Chinese language as their mother tongue. Not only Chinese people speak Chinese language, some people in Singapore and Malaysia also speak Chinese language. Chinese language is one of the working languages for United Nations.

Chinese language resources and evaluation must deal with the Chinese characters. It is a remarkable feature for Chinese Language Technology (CLT). CLT is an important part of Human Language Technology (HLT).

Standards are an obvious priority issue for international co-operation.

For text corpora, international co-operation is mainly promoted through joint projects with foreign countries. "People's Daily" corpus processing is a joint project between ICL-PKU (China) and FUJITSU Company (Japan).

For other types of language resources, international cooperation is mainly achieved by sharing resources, data and tools.

Machine dictionary GKBCC: sharing with Intel (USA), Matsushita (Japan), XRCE (Xerox Research Center Europe, France), CiTaL (Centrum für Terminologie Internationale und Angewandte Linguistik, Germany), KAIST (Korea Advanced Institute of Science and Technology, Korea), Pecan (a sub-company of CANON).

Corpus processing tool Slex: sharing with Intel (USA), Matsushita (Japan), XRCE (France), CiTaL (Germany), KAIST(Korea), NUS (National University of Singapore).

Terminology Data Bank: sharing with CiTaL (Germany).

3.3. Discussants

According to *Joseph Mariani (LIMSI-CNRS, Paris)*, the LREC 2000 conference on Language Resources and Evaluation in Athens was the opportunity for the international community to meet, report on the present situation and propose cooperative actions.

The present situation in Human Language Technologies evaluation is that the US keeps on organizing large comparative evaluation campaigns embracing speech and natural language, with a large European participation which is not funded by US or EC funds, but it appears that the interest in participating is strong enough to prompt this free participation. DARPA starts new programs (Communicator and TIDES on Translingual Information Detection, Extraction and Summarization) using the evaluation paradigm within a common architecture, and several European laboratories join those programs as affiliates. In Japan, forces on Text processing systems evaluation have been gathered in a single entity, the National Institute for Informatics (NII). Apart from those large programs, several initiatives are taking place in various places around the world, such as the evaluation campaigns in France (AUF, Amaryllis, French DoD...), in Germany (within the Verbmobil or SmartKom programs) or at the international level (Senseval, for example). Such a tool is still lacking in the European Commission programs.

Two questions then raise.

- Is there room for several initiatives around the world?

The answer seems to be yes, as there are different languages to be covered, there may be different ideas based on different cultures and therefore discussing those ideas may help defining the best way to handle the question, and finally because the size of the effort is very large, thus necessitating shared efforts to cover the various tasks in the various languages.

- If so, should it be coordinated?

The answer seems also to be yes. It is obvious that science and technology are international, and that evaluation should therefore be conducted at the international level. Laboratories find it difficult to participate in all initiatives due to lack of time and manpower. Thirdly, it appears in the present situation that it is difficult in the various initiatives to get the necessary language resources in the various languages aimed at, and also it would avoid reinventing the wheel in the design of evaluation methodologies.

We should therefore try to find a way to install a truly international human language technologies evaluation scheme, one of the problem being that it doesn't fit in so well with the EC programs Call for Proposals mechanisms, and that creating an institute comparable to NIST or NII in Europe will be a very difficult task, which may take a long time and a large amount of efforts.

Harald Hoege (Siemens, Munich) started considering that in the last five years a successful infrastructure to produce, disseminate, standardize and validate SLR has been set up within Europe and US. This infrastructure becomes visible through ELRA and LDC. Also activities in Japan start working in this direction. Due to the different funding strategies of the national bodies no common international approach exists.

He proposed to start such a common production and dissemination strategy through the following actions:

- International production of SLR for Speech-to Speech translation for 50 languages at an international level.
- Each funding agency (Europe, US, Asia) supports this action by 20MECU (ca. 1 Million ECU per language).
- of the SLR through a common dissemination policy on a license free basis.

On the basis of the previous interventions, *Volker Steinbiss (Philips, Aachen)* asked various questions on the role that ELRA can play for the development of LR through international co-operation, focusing in particular on overall policy issues.

Núria Bel (gilcUB, Barcelona) stated that, as HLT components are more and more being included in all kind of IT applications, Language Resources should be considered as a basic infrastructure for current and future Information Society. As any other basic infrastructure, these resources need to be created, maintained and updated, and this means a planning based on a long term strategy and a long term funding. Besides, there are already examples (such as software localization) that have proven that availability of all kind of applications in local market languages becomes to be considered a further user requirement. There is such a demand. Hence we should not expect a full deployment of HLT in the world without addressing all kind of local languages, independently of its number of speakers.

This infrastructure is, with no doubt, a very expensive investment, and because of the social and economical interests which are behind of the area of HLT applications, it is commonly agreed that there should be public support for them. Until now, in Europe there has been two strategies: to appeal to the subsidiarity principle, so that each state should care of covering its language, or, as a more strategic international policy, to fund such initiatives in the form of EU R&D projects. Some of these projects, though, have given support to very concrete multinational industries in this area, resulting in a non widely sharable infrastructure, and, more crucially, an infrastructure that is only available for those languages that have an interest for these industries because they have a large market, major languages which are not spoken only in one country, languages which are not only EU languages.

It seems, hence, that on the one hand the EU is investing public funds in languages that have a clear market even though they are spoken in many different countries around the world, a fact that one would expect to be the basis for international co-operation outside the EU. On the other hand, some other European languages are left to national initiatives because of its low interest in terms of short term marketing. These national initiatives exist, but they lack common organization and normally they count on low funding because the arguments used to defend them are mainly based on supporting cultural diversity, which is, as we know, a non very attractive argument in terms of funds. For them, international co-operation will mean political support.

If we look at other areas where economic, social and politic interests play a role, such as health, nuclear, space, aeronautic research and development, we can see that the different administrations have managed, in co-operation with interested industries, to create special agencies or large projects, with fixed contributions from the different participants, and, what it is really important, long term planning and funding. Hence, do not we go for such an international agency for Language Resources? An overseas international body that organises, plans and fixes long term strategies for the development, maintenance and update of this HLT infrastructure for all languages.

Lori Levin (Carnegie Mellon University, Pittsburg) presented NICE (Native language Interpretation and Communication Environment) as an example of collaboration between United States and Latin American countries. The project, dealing with MT between Spanish and indigenous languages, was conceived by U.S. funding agencies (NSF and DARPA) along with the Organization of American States in the context of a larger project on Western Hemisphere collaboration in multilingual contexts.

There was a concern about disenfranchisement of speakers of indigenous languages from goverment and the Internet.

The first Latin American partner is the Universidad de la Frontera in Chile. It was learned from them that the Mapuche people would view a machine translation project in the context of community development, which in their villages is centred around the schools.

As a result, we are working primarily through the Ministry of Education in Chile.

This is in contrast to other countries where machine translation projects are centred around government, industry, or defense.

3.4. General Discussion

A general agreement emerged on the need of international co-ordination and co-operation, which appears the only way to provide the LR required to answer the challenges and the expectations of the contemporary evolving multilingual ICT-based Society.

In Europe, an explicit co-ordination should be established between the initiatives of the EU and the activities of the member States: in fact, the prevision of LR is a common target for the various European national projects, and initiatives of the type of ENABLER should be developed and maintained.

Several interventions highlighted specific needs, calling the attention on opportunities for international co-operation offered by planned or on-going initiatives.

Due to lack of space, we can quote only a few examples here.

Zygmunt Vetulani (University of Poznan) observed that creation of LR for languages of eastern countries is a priority for HLT development in these countries and represents an uncontroversial logical starting point for eastern-western co-operation.

Piek Vossen (Sail Labs GmbH, Munich) and *Christian Fellbaum (Princeton University)* announced a new international association aiming at fostering co-operation among researchers and developers interested in lexical semantic networks.

Tarcisio Della Senta (United Nation University, Tokyo), offered UNL, and in particular the wealth of LR-corpora, lexica, knowledge developed for languages of five continents, as an example and a forum for international co-ordination.

Gerhard Budin (University of Vienne) and *Rute Costa (President of EAFT)* observed that the situation of terminology is ripe now, both from the organizational and the technical point of view, to realise the co-operation with computational lexicography, well recognized as a need but never practically firmly established.

Steven Krauwer (University of Utrecht) briefly summarised the institutional vocation of ELSNET to promote international co-operation, and offered the expertise and the infrastructure of ELSNET, in particular the ELSNET task force for LR, for helping implementing a world-wide co-operation.

Ideas and suggestions emerged during the Panel were immediately taken in consideration, already during the remaining of the Conference, in particular the proposal for establishing an overall world-wide initiative, involving existing infrastructures like ELRA, LDC, COCOSDA.

A first meeting will be organized, in co-operation with a workshop sponsored, at the ACL Conference in Hong Kong (October 2000), by ELSNET, to address questions like: (1) what are the existing infrastructures which should be involved world-wide, and how they can be optimally exploited to foster global co-operation; (2) what infrastructure and interconnections are missing, and which are the main actors (institutions, organizations) to be involved to build and operate a truly overall international infrastructure; (3) what are the mandate and more urgent priorities for such an infrastructure. A second discussion will be organized at the occasion of the next COCOSDA meeting (which takes place two weeks after in Beijing).

Post-Panel Discussion

Those wishing to further contribute to the discussion, for example reporting on experience of international co-operation, highlighting general or specific needs, suggesting priorities, or commenting on policy and organisational problems, are invited to send messages to the discussion list intpan@ilc.pi.cnr.it. If appropriate, we will channel comments and suggestions to the relevant funding agencies.

The same Web site will make available the transparencies used at the COLING Panel on "International Co-operation" (Saarbrüchen, August 2000), and the following discussions.

Antonio Zampolli
University of Pisa
Department of Linguistics
Istituto di Linguistica Computazionale del CNR
Pisa
Email: intpan@ilc.pi.cnr.it