LANGUAGE TECHNOLOGY AND MULTILINGUALITY – THE EUROPEAN DIMENSION

Poul Andersen
European Commission
Address: EUFO 1197 – rue Alcide de Gasperi – L-2920 Luxembourg
Tel. +352-4301-34324, fax +352-4301-34655
e-mail: poul.andersen@lux.dg13.cec.be

ABSTRACT

The European Commission supports a multilingual Europe, where each citizen can use his native language. With the ongoing integration process in Europe, and the advance of the Information Society, we must create the necessary tools to facilitate communication across language barriers, and make these tools easily available to professionals in the translation sector and information-related industries, as well as to ordinary people.

The Commission pursues these goals, both through promotional action in the MLIS programme, and through support for Research and Technological Development in the Framework Programmes.

The Commission is itself an important user of Language Technology, with more than 1300 translators working in the 11 official languages of the European Union – and the expected extension with 11 new member states (10 CEECs + Cyprus) will increase the number of official languages up to 21.

DISCLAIMER: This article is written as an overview conference paper, to provide general information about Commission-supported activities in an informal way, and does not constitute an official policy statement.

1 THE CHALLENGE FROM EUROPEAN INTEGRATION

The European Union today has 15 Member States, and 11 official languages (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish).

Official documents of legislative character, information material for the public etc. are translated into all 11 languages, and at meetings in the European Commission and other EU institutions, interpretation can be requested from and into any of the 11 languages.

This policy has of course huge costs, both economically, and logistically: if a document is slightly changed after it has been translated, all the 10 other versions also have to be changed; and composing a team of e.g. 3 interpreters for each of 11 languages, who together with their colleagues can secure any of 110 combinations (from each of the 11 languages into the 10 other languages), is not an easy puzzle.

Still, the basis for European integration is a cooperation between politically equal partners, where nobody wants to see a strong dominance of a single nation. Nation and language are closely related concepts – notwithstanding counter-examples such as Belgium, with three official languages (French, Flemish/Dutch, and German) that are shared with other EU member states – and Multilinguality is an official policy, anchored in the Treaties of the European Communities, which is closely linked to the multinational character of European integration, and which is not being seriously contested.

The extension of EU with 10 Central & Eastern European Countries (CEEC), each with its own national language (Bulgarian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Polish, Romanian, Slovak, Slovenian), is not expected to change this policy – if EU has been able to cope with the increase from the original 4 languages (Dutch, French, German and Italian) to 11 languages, it must also be able to cope with 21 languages, if the political will is there, and the appropriate technological means are exploited.

Up to now, we have only talked about *official* languages – let us not forget that there are many more living languages in the European Union – an *"un-official"* EU language such as Catalan is spoken by more people than some of the smaller official languages. The *European Bureau for Lesser Used Languages* [1] lists appr. 40 languages in its publications, all spoken within the present 15 member states, and not including the languages of recent immigrant or refugee populations.

2 THE CHALLENGE FROM THE INFORMATION SOCIETY

The Information Society means an important increase in the flow and exchange of information, and an even bigger increase in the *speed* with which information circulates – and with which information is up-dated – and becomes out-dated. To be of high value, information must be readily available on short notice and with electronic means.

This is easier to manage in one language – multilinguality will easily become a hindrance, if each piece of information has to be translated into a multitude of languages before it is released. In this way, English gets a competitive advantage, as the dominant language in the

Western post-industrial countries where the technological and societal changes linked to the Information Society started, whereas smaller languages come under threat, because it is commercially less interesting to develop and market the technological means needed for these languages to keep up with the advance of the Information Society.

It can therefore not be left to market forces alone to secure equal access to information in their native language for all citizens in the European Union.

3 THE EUROPEAN UNION – GENERAL INFORMATION

Within the European Commission, Directorate General XIII is i.a. responsible for the Information Society (IS) and for Language Processing. DGXIII's activities are presented on the server *i*m europe* [2] (i*m = Information Market).

The best entry point for general information about the EU's R&D activities (not restricted to DG XIII) is the CORDIS database [3], supplemented with subscription (for free) to the newsletter CORDIS*focus* [4].

4 MLIS – MULTILINGUAL INFORMATION SOCIETY

The MLIS Programme [5], which runs from 1997-99, promotes the linguistic diversity of the EU in the Information society. In order to achieve this general aim, the programme will -

- raise awareness of and stimulate provision of multilingual services in the Community, utilising language technologies, resources and standards;
- create favorable conditions for the development of the language industries;
- reduce the cost of information transfer among languages, in particular for Small and Medium-sized Enterprises (SMEs);

This is done by co-financing pilot projects (with public and private organizations), and through concertation and commission studies to improve understanding of the issues for the actors concerned.

The programme is structured along three Action Lines:

1. Supporting the creation of a framework of services for European language resources
2. Encouraging the use of modern language technologies, resources and standards
3. Promoting the use of advanced language tools in the Community and Member States' public sector

- complemented with Accompanying Measures (publications, awareness events etc.).

Compared with other EU programmes, MLIS has a modest budget of 15 mio ECU for the 3 years' duration. MLIS is not a R&D programme, but complements R&D activities, such as projects funded under the Framework Programmes, by promoting the results of these projects.

A large part of the supported activities are shared-cost projects, where the private companies or research institutes involved typically participate with 2/3 of the budget. The common interest of the Commission and the project participants is to arrive at commercially interesting results, such as making already existing dictionaries and terminology databanks accessible for consultation on-line to professionals (translators a.o.) who sign a user agreement with the data providers.

4.1 CEEC in MLIS and other IS programmes

Partners from non-EU member states can participate in MLIS projects, but they cannot receive financial support from the MLIS programme. An official document [6] from 1997 *'invites the Commission to (...) establish opportunities for EU funding for IS activities and for co-financing CEEC participation in EU IS related programmes, like INFO2000 and the MLIS programme, in order to develop the full scope of the IS'.*

At present, we are investigating the possibilities for creating such a co-financing opportunity for CEEC participation in these programmes.

5 R&D – THE FRAMEWORK PROGRAMMES

The European Union's different RTD (Research and Technological Development) programmes and related activities are carried out under the Framework Programmes (FWP).

5.1 Fourth RTD Framework Programme

The 4th FWP runs from 1994-98, with a total budget of appr. 13,000 mio ECU, distributed over

- 15 specific thematic Research programmes aimed at promoting cooperation between companies, research centres and universities (First Activity)
- Promotion of RTD cooperation with third countries and international organizations (Second Activity) = <u>International Cooperation</u>
- Dissemination and optimization of Community-funded RTD results (Third Activity)
- Stimulation of the training and mobility of researchers in the Community (Fourth Activity).

One of the 15 thematic Research programmes is the *Telematics Applications Programme*, which includes *Language Engineering* [7] (LE) as one out of 12 sectors, with a budget of appr. 80 mio ECU for the 5 years' duration of the 4th FWP.

Most of the LE budget is spent on appr. 40 cooperative RTD projects, + a smaller number of projects for development of resources, and projects relating to standardisation and evaluation/testing.

An initiative of general interest is *European Language Resources Association – ELRA* [8], – an infrastructure for identifying, collecting, classifying, validating, distributing, and exploiting language and speech resources, such as basic data (corpora, recordings, terminology), linguistic models (grammars, lexica, HMM) and software tools. ELRA also works on development of evaluation guidelines, and serves as a central clearing house and broker between producers and users of resources.

5.1.1 CEEC in the 4th Framework Programme

Although encouraged by the Commission, direct participation of CEEC in the *LE programme* under 4th FWP has been very low – it is expected that this will change completely under 5th FWP (see below).

The larger part of the *International Cooperation* budget (INCO) under the 'Second Activity' has been used for CEEC-related activities – INCO-COPERNICUS, and for cooperation with Developing Countries, INCO-CD.

The INCO-COPERNICUS programme has made it possible to fund a quite important range of activities in CEEC in the area of Language Engineering and Multilingual Issues. These activities always have a trans-European character, bringing together scientists and industrialists from EU and CEEC, but the main stress is on the CEEC participation :

1. Awareness Seminars on Language & Technology

These seminars are organised in order to raise awareness of the multilingual aspects of the Information Society, and of the need to develop technological tools for language services.

A preparatory seminar for all CEEC was held in Luxembourg in 1994, followed in 1994-1996 by national or regional seminars in Czech Republic, Latvia, Poland, Romania and St. Petersburg.

At present, these seminars are no longer organised by the Commission according to the original model, but the concept of Awareness seminars has been adopted i.a. by Romania, which on its own initiative has conducted similar events, and by Slovenia, as it is shown by the present conference on Language Technologies for the Slovene Language.

2. Concerted Actions

Concerted Actions aim to co-ordinate projects already funded by the Commission – or funded by national public authorities or private bodies. Research & Development work is not supported, but Concerted Actions can be used to fund coordination and concertation measures, such as travel and other meeting costs, and the cost of setting up infrastructure services for the results from ongoing or completed projects.

Two Concerted Actions were funded 1995-97:

2.1. ELSNET goes East – an extension of ELSNET[9] to CEEC. This specific action for CEEC is now finished, but ELSNET remains committed to cooperation with CEEC. There are ELSNET nodes (academic or industrial sites) in 7 CEECs (with 3 nodes each in Russia and Hungary), and there are regular reports by CEEC scientists or industrialists in the newsletter *elsnews*, which recently had a special issue on cooperation between EU and CEEC (*elsnews* 6.1, Feb 1997). Conference Calls for participation etc. are also published on the ELSNET electronic mailing list – both the newsletter and the mailing list can be subscribed to for free [10].

2.2. TELRI [11] – this Concerted Action is now being extended with another three years, 1998-2001. This is the only EU-funded activity within this area that brings together participants from all eligible CEECs and several NIS (New Independent States = ex-USSR), with 27 partners from 22 different countries in EU and CEEC/NIS.

One of the results of TELRI is TRACTOR [12] – the *TELRI Research Archive of Computational Tools and Resources*, which is a kind of complement to ELRA (see above in 5.1) for CEEC languages, but mainly restricted to written language resources (little or no speech resources).

3. Joint Research Projects

The following completed or ongoing projects are funded from the INCO-COPERNICUS programme 1995-1998 – for details, please contact the author of this article :

3.1. Language Resources – Terminology :

PRACTEAST – Preparatory Actions for Terminological Assistance to CEEC

LANGELEC – International Standardised Terminology for Electrical Engineering and Telecommunications

3.2. Language Resources – Corpora :

MULTEXT-EAST – Multilingual Text Tools and Corpora for Central and Eastern European Languages.

MULTEXT-EAST ended in 1997. In 1998 started a new project with the same CEEC partners, but with a shift from corpora building towards dictionary coding :

CONCEDE – Consortium for Central European Dictionary Building

3.3. Language Resources – Dictionaries :

CEGLEX – Central European GeneLEX model

GRAMLEX (morphological dictionaries)

BILEDITA – Bilingual electronic dictionaries and intelligent text alignment

3.4. Language Resources – Speech :

BABEL : A Multi-Language Database

ONOMASTICA – Copernicus. Multi-language pronunciation dictionary of names in CEEC.

SQEL – Spoken Queries in European Languages

3.5. Tools and Applications – CALL / Translation tools:

BALTIC – Basic and advanced language transnational interactive course

GLOSSER (CALL SW / morphological analysis and use of text corpora)

STEEL – Developing Specialised Translation/Foreign Language Understanding Tools for CEEC Languages

AGILE – Automatic generation of Instructions in Languages of Eastern Europe

5.2 Fifth RTD Framework Programme

The final approval of 5th FWP [13] is expected towards the end of 1998, when 4th FWP expires. The new FWP will be structured along a small number of *thematic programmes*:

- Improving the quality of life and the management of living resources:
- Creating a user-friendly Information Society = Information Society Technologies (IST)[14]
- Promoting competitive and sustainable growth:
- Preserving the ecosystem

Each thematic programme groups together several *key actions*. The key actions under the IST programme are:

- Systems and services for the citizen
- New methods of work and electronic commerce
- Multimedia content and tools
- Essential technologies and infrastructures

- and finally, in the third layer, under *Multimedia content and tools*, we find

- Human Language Technologies (HLT),

- with the following description:

"Work will focus on advanced human language technologies enabling cost-effective interchanges across language and culture, natural interfaces to digital services and more intuitive assimilation and use of multimedia content. Work will address written and spoken language technologies and their use in key sectors such as corporate and commercial publishing, education and training, cultural heritage, global business and electronic commerce, public services and utilities, and special-needs groups. Work will also include the development of electronic language resources in standard and re-usable formats.

RTD priorities: adding ***multilinguality*** *to systems at all stages of the information cycle, including content generation and maintenance in multiple languages, localisation of software and content, automated translation and interpretation, and computer-assisted language training; enhancing the* ***natural interactivity*** *and usability of systems where multimodal dialogues, understanding of messages and communicative acts, unconstrained language input-output and keyboard-less operation can greatly improve applications; enabling* ***active assimilation and use of digital content****, where work will apply language-processing models, tools and techniques for deep information analysis and metadata generation, knowledge extraction, classification and summarisation of the meaning embodied in the content, including intelligent language-based assistants; the work will be complemented by* ***take-up*** *actions including validations and assessments, together with first-user actions and other best-practice initiatives."*

The HLT programme is the natural continuation of the LE programme under 4th FWP, see contact information in endnote 5.

The IST thematic programme, with its various components, incl. HLT, will be presented at the IST 98 conference in November/December 1998 in Vienna [15].

5.2.1 CEEC in the 5th Framework Programme

All ten CEEC candidates for EU membership + Cyprus have officially applied for association with the 5th FWP. The Commission supports this request. Negotiations are about to start, and it is hoped that they can be concluded in time for the launch of the new FWP at the start of 1999. This means that partners (companies, research institutes and other legal entities) from CEECs will be able to participate at the same terms as partners from EU, - and associated CEECs will have to contribute to the programme with a certain proportion of their GDP.

The big difference from the 4th FWP is that the participation of CEEC partners will not be financed from a special budget, and will not be managed via special contracts, which in practice until now largely has restricted the cooperation between EU and CEEC partners to specifically CEEC-oriented projects, as explained under item 5.1.1. above.

The 5th FWP will still contain a programme for International Cooperation, which i.a. will support participation in Joint Research Projects of partners from non-associated third countries, such as NIS (= New Independent States, the usual Commission abbreviation for ex-USSR, excl. the three Baltic states).

6 THE COMMISSION AS A USER

As a multilingual institution, with more than 1300 translators working in 11 language units (one for each language), the European Commission is an important user and client on the market for translation tools and other applications to facilitate multilingual text processing and information management in general.

Within the Commission's *translation service*, a common interface is provided through EURAMIS (*European Advanced Multilingual Information System*) to facilities such as:

- *Translator's Workbench*®, by TRADOS, for translation memory;
- *Eurodicautom* [16], the Commission's multilingual terminology database;
- SYSTRAN machine translation

Machine Translation is also available on-line to all other Commission staff. The Systran MT system has been constantly developed and extended with new language pairs, since it was first installed in the Commission in 1976 for translations between English and French. The Commission's SYSTRAN system now covers translation FROM French, English, German, Spanish and Greek as source languages, INTO French, English, German, Italian, Dutch, Spanish, Portuguese and Greek as target languages. These 5 source and 8 target languages add up to 5x8 = 40 language combinations – only 17 of these have been implemented, and we are still far from covering all 110 combinations between the 11 official languages (cf. item 1 at the beginning of this article).

It is the Commission's hope to eventually offer Machine Translation for all – present and future – official languages, although not necessary in all possible combinations – for smaller languages such as Danish or Finnish, or the CEEC languages, the immediate goal is to offer translation into and from at least one better known language, such as English, French or German.

The extension to other EU languages is supported under Action Line 3 of the MLIS programme (see item 4 above) in those cases where it has been possible to define a project of common interest to the Commission and one or more member states.

With special reference to the expected extension of EU with the ten candidate CEECs, the author of this article prepared an overview "Translation Tools for the CEEC Candidates for EU membership"[17].

How the Commission's MT system will be extended to cater for CEEC languages has not yet been decided, but there are good reasons to prefer a solution which fits into the already existing MT architecture in the Commission.

7 REFERENCES

[1] European Bureau for Lesser Used Languages
Brussels Information Centre
Sint-Jooststraat 49
B-1210 Brussel
Tel. +32-2-2182590
Fax +32-2-2181974
E-mail: pub00341@innet.be

[2] http://www2.echo.lu/

[3] http://www.cordis.lu/

[4] to subscribe to CORDIS *focus*, contact:
RTD-Help Desk
European Commission DG XIII/D2
rue Alcide de Gasperi
EUFO 02/2286
L-2920 Luxembourg
Fax : +352 4301 32084
E-mail: RTD-Helpdesk@lux.dg13.cec.be

[5] European Commission DG XIII/E/6
MLIS Office – EUFO 1154
Rue Alcide de Gasperi
Plateau du Kirchberg
L- 2920 Luxembourg
Tel: +352-43 01-34 117
Fax: +352-43 01-34 655
Email: mlis@lux.dg13.cec.be
http://www2.echo.lu/mlis/home.html

[6] Third EU/CEEC Information Society Forum,
Brussels, 9-10 October 1997 – Proceedings
(European Parliament – European Commission)

[7] European Commission, DG XIII/E/5
LE Office - EUFO 0177
Rue Alcide de Gasperi
Plateau du Kirchberg
L-2920 Luxembourg
Fax: +352 4301 34999
Tel: +352 4301 32886
Email: lepostmaster@lux.dg13.cec.be
http://www2.echo.lu/langeng/en/lehome.html

[8] ELRA/ELDA
55-57, rue Brillat Savarin
75013 Paris
FRANCE
Tel: (+33 1) 43 13 33 33
Fax: (+33 1) 43 13 33 30
Email: elra@calvanet.calvacom.fr
http://www.icp.grenet.fr/ELRA/home.html

[9] ELSNET, the European Network in Language and Speech, was established in 1991 with funding from the European Commission's ESPRIT programme.

[10] For information on ELSNET contact:
Utrecht Institute of Linguistics OTS,
Utrecht University
Trans 10
NL-3512 JK Utrecht
Netherlands
Tel: +31-30 253 6039
Fax: +31-30 253 6000
E-mail: elsnet@let.ruu.nl
URL: http://www.elsnet.org

[11] TELRI is short for *'Trans-European Language Resources Infrastructure'*,
see http://www.ids-mannheim.de/telri
– more information in presentation by Wolfgang Teubert

[12] http://solaris3.ids-mannheim.de/~tractor/, expected to become accessible from TELRI's home page, see preceding note.

[13] http://www.cordis.lu/fifth/home.html

[14] http://www.cordis.lu/fifth/src/305b-e-1.htm

[15] IST 98 "Living and Working in the Information Society" – Information Society Technologies Conference & Exhibition, 30 November – 2 December 1998, Vienna, see http://www.cordis.lu/ist98/.

[16] Eurodicautom is publicly accessible at
http://www2.echo.lu/edic/.

[17] to appear in the Commission's journal
Terminologie et Traduction – please, ask the author for a copy, if interested.