MT Research and Development (R&D) in Europe

Roger Havenith
European Commission
Directorate-General XIII
Telecommunications, Information Market and Exploitation of Research
LE Programme
EUFO 0-179
L-2920 Luxembourg

Introduction

MT R&D in Europe is being carried out by a significant number of actors, ranging from private research labs through government research centres to universities and similar institutions. In addition, probably all industrial MT suppliers in Europe are engaged in near-market R&D. Funding for MT R&D comes from a variety of sources.

In my paper, I will concentrate on the programmes, initiatives and projects that are sponsored by the public sector. The focus will be on EC-funded funded research and technology development, but I will also attempt to provide you with a general overview of the characteristics of national initiatives and programmes that provide a framework for MT research, and - by way of exemplification - address a few of them in more detail.

It is worth noting that a range of MT R&D activities in Europe are currently being carried out with private funds. These fall outside the scope of national and European programmes. Some of these activities are rather small-scale basic research projects, others are more significant development projects, typically led by industry. Some of this work has been reported on in scientific papers, conference proceedings and dedicated journals[1]. Other activities were have not been publicised, for a variety of reasons (protection of commercial interests, intellectual property rights etc.)[2]

To provide a comprehensive overview of all this valuable work would have gone beyond the scope of this paper.

National programmes

Both the general framework conditions and the funding opportunities for MT R&D at national level differ from one EU country to another. Some countries have put in place national programmes that provide a dedicated framework for LE-related R&D[3], other EU Member States are discussing plans for such programmes, either at national or at regional level[4]. Finally, a number of EU countries[5] have not yet a dedicated framework or programme for language R&D.

Where dedicated programmes exist, they are typically embedded in a cascading structure of national R&D policy, Programmes on Information and Communication Technologies and Action Plans for the Information Society. These programmes have often been devised taking into account the structures of European RTD.

---

* The opinions expressed in this paper are exclusively those of the author and do not necessarily coincide with those of the European Commission

In the area of LE, national programmes and plans generally show a great awareness of the necessity of creating synergy and achieving co-ordination between national and EU programmes. However, they do not always provide the necessary instruments for achieving these goals.

Financial support at national for R&D projects is typically provided through:

- credits, guarantees for credits, loans or by financing (in full or in part) the costs of R&D projects,
- sponsorship of prizes in science and research, awarded for completed projects,
- scholarships for science and research, allocated to individuals.

The instruments for channelling financial support are quite diverse, ranging from national programmes through sponsoring agencies[6] to national funds[7] and science foundations.

The role that MT R&D plays in the various national environments reflects the great diversity of national policies and schemes.

In the following, I will present some of the MT R&D developments that take place within two EU Member States, Finland and Spain. I singled out these two countries, one in the North and one in the South of Europe, because they exemplify two interesting, different environments and approaches within the broad spectrum of European R&D in MT.

Those who are interested in a more detailed description of all the national programmes, schemes and projects that provide a framework for MT R&D in the European Union, are referred to a series of surveys of EU countries, which are currently being compiled within the EC funded project "Euromap", and are likely to be disseminated via the WWW site of the LE Programme[8] towards the end of this year.

Example 1: Finland

Finland, one of the EU's Nordic countries, is a relatively new Member State, with a private sector playing an active role in R&D: some 60% of research and development in Finland is being carried out by companies using their own funds. The public sector and universities account each for half of the remaining 40%.

There is no specific funding scheme for language engineering in Finland. Financial support to research in this field is provided in the context of long-term national research programmes, e.g. the programme on digital printing technology and the Finnish Multimedia Programme 1995-97.

Some other long-term research programmes in LE are being carried out at the University of Tampere, for example by the Digital Media Institute.

The Technology Development Centre (Teknologian kehittamiskeskus - TEKES), sub-ordinate to the Ministry of Trade and Industry, is the main source of finance for applied and industrial research and development.

The Finnish National Fund for Research and Development, SITRA, has supported LE for many years. Kielikone Ltd, a company specialised in language technology products and software modules, has its origins in a language engineering research project funded by SITRA between 1982 and 1990. R&D work carried out by Kielikone in the area of MT has resulted a.o. in KMTech, a generic MT development system, and the Finnish-English MT system TranSmart.

Lingsoft, a Finnish linguistic software company, has developed a range of language technology products, including proofing tools and components that are relevant as building blocks for MT systems, such as morphological analysers and grammar parsers.

Finally, R&D work in MT has been carried out at the University of Helsinki, in the context of a machine translation project (1987-1992) sponsored by IBM.

As witnessed by the above examples, the Finnish MT R&D community seems to be interested in all the steps in the chain from the development of modules and systems "for research purposes only", to commercialisation of fully-fledged products and the provision of expert services.

Example 2: Spain

Spain, one of the EU's Mediterranean countries, has been a member of the EU for about a dozen years. Since the accession of Spain to the EC on January 1st 1986, Spanish universities and companies have participated in a significant number of European MT-related projects.

At the Spanish national level, language engineering R&D is currently funded under three programmes: (i) the National Programme of Telematic Applications - PNAST, (ii) the National Programme on Information and Communications Technologies - TIC, and (iii) the Sectorial Programme for the General Promotion of Knowledge. In addition, there are regional programmes funded by the *Comunidades Autónomas.*

LE R&D projects can be classified according to four main research strands: speech technology, natural language processing, language resources and information technology in combination with LE technologies. At least two nationally funded R&D projects are concerned specifically with MT and spoken language translation: one is a project on Spoken Language Translation and Comprehension by Means of Learning Techniques (TRACOM, 1995-1997, ref: TIC-95-0984-C02-01), the other is a project on transfer issues in English-Spanish and Spanish-English in an MT system (1993-1996, ref. DGCYT PB92-0668).

Joint R&D initiatives of several EU Member States and regions

In addition to national and regional programmes and EU RTD frameworks, a number of EU countries and regions are setting up joint R&D initiatives, that may provide a home for future MT R&D. The two following examples[9] may illustrate this point:

1. Nordic frameworks and programmes

The Nordic countries have a long tradition of collaboration in many different areas, including research. The issue of Nordic languages in the information society was discussed at the Nordic Council in Oslo on 3-4 March 1997. Although nothing specific has been decided yet, it is expected that the Nordic Council will follow up on this initiative, and possibly initiate a programme addressing the issue of Nordic languages and related technology aspects.

2. Nederlandse taalunie

Dutch is spoken in two EU countries, the Netherlands and Belgium (Flanders). The Dutch speaking communities of both countries have joined forces within the "Nederlandse taalunie" (Dutch language association), and committed themselves to observing the same rules for Dutch spelling and grammar.

Dutch is a less widely spoken language in Europe. In consequence there is a growing awareness that preservation of Dutch as a full language requires coordination of research efforts in the Netherlands and Flanders. Research programmes currently under discussion, for example within the *Flemish Science Policy Council (Vlaamse Raad voor Wetenschapsbeleid (VRWB)),* show that there is a clear will to achieve complementarity with European research efforts and co-ordination with research conducted within the Netherlands.

EC-funded research

The European Commission as a user of MT

Machine translation has a long tradition within the European Commission, which is both a consumer of MT and a sponsor of MT R&D.

The EU has 11 official and working languages, resulting in 110 possible language pairs. All legally binding documents are supposed to be available in all the official and working languages. For in-house purposes, translation is typically limited to a subset of these languages, according to the specific needs. The EC translation service is probably the biggest worldwide, and counts some 1100 translators, 100 linguistic support staff, approximately 100 management staff and close to 500 secretaries and assistants. It processes a significant part of the Commission's total multilingual production, which is close to three million pages a year. Most of the documents are submitted to a combination of in-house, free-lance and machine processing, and have English, French or German either as a source or as a target language. Not all translation needs can currently be met within the deadlines with the human resources available. In particular, there is a latent demand for fast translation of documents at very short notice.

All this shows that the European Commission's inhouse translation needs are indeed very significant. The EC's policy with regard to inhouse translation, the EC's Systran MT system and the developments around EURAMIS have already been reported on at previous conferences, and the most recent developments are the topic of a separate presentation at this Summit. I will therefore confine myself to a few very general remarks, which are based on the results of an in-house MT feasibility study carried out last year by the EC Translation Service.

Among the conclusions of the final report of the study, it is interesting to note the following:

- MT meets a real institutional need.
- MT is considered useful by 67% of the users of the Translation Service, and 95% of the users from other Commission services. MT should therefore be included in the range of tools offered to the EC translators.
- The raw quality delivered by the EC's inhouse Systran system is deemed acceptable by more than 60% of the users within the Translation Service, and 74% of the users from other Commission services.
- There is a latent demand for very fast translation of documents, and MT could play a role in meeting this demand.
- To stimulate further the usage of MT, it is necessary to continue to improve the quality of the system.

The EC as a sponsor of R&D in MT

Early years

At a European level, early work in linguistics was carried out within the context of Esprit and EUROTRA programmes. The first EC-funded R&D project in the area of MT was **EUROTRA**, a very large and ambitious multi-year project (1982-1991) designed to push forward the capabilities of MT. One of the initial aims of EUROTRA was to build a next generation, unification-based operational MT system, that would have been able to replace Systran at the Commission. EUROTRA did not succeed in building such a system. However, it was instrumental in creating a linguistic infrastructure in Europe, and in establishing a truly European community of LE researchers, who know each other and have a tradition of cooperating and communicating across geographic and language borders. Also, the project built up a significant body of linguistic knowledge and descriptions, thus providing the groundwork for much of the work that was undertaken subsequently in the field of M(A)T in Europe. The linguistic specifications produced by EUROTRA ("Eurotra reference manual") were later on rewritten, completed and extended by the MLAP project "**Refman**", and a series of language-specific projects (e.g. **FRELIS**, **MODUL**). This work resulted in guidelines and a very comprehensive, systematic account of linguistic phenomena, which is likely to be useful to not only to academic users, but also to developers of commercial systems.

Building on the Eurotra R&D work, the staff of the Danish Eurotra centre CST developed their grammar modules into an operational MT system, named "**Patrans**". This system has been in operation at the Danish company Lingtech since May 1994 and is being used for the translation of patents from English to Danish. According to Lingtech, Patrans has allowed it to increase productivity by up to 50%.

One of the early R&D activities worth noting is the **CAT**-system developed by the German Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung (IAI) as part of an official Eurotra sideline. This system has been re-used in a number of EC projects and is still being developed further by IAI.

Towards a Language Engineering Strategy

Following up on Eurotra, which was basically a monolithic "one-project-programme", the Linguistic Research and Engineering programme (LRE), launched in 1991, initiates a change in EC's policy in the area of language R&D. LRE establishes a dedicated framework covering a broad range of R&D action lines covering both written and spoken natural language. The aim of LRE is to develop a basic linguistic technology that can be incorporated into a large number of language-aware applications. To that end, work is funded along three lines: general research, common resources (including tools, methods, data and standards) and the development of pilot applications. LRE projects building pilot applications are supposed to pay attention not only to research tasks, but also to the engineering aspects, for example the coding of data, programming and testing by end-users. MT R&D is identified as an important area for LRE projects, but it is no longer the sole purpose of the programme.

Contrary to Eurotra, which was implemented through contracts of association between the EC and Members States, LRE makes use of calls for proposals. This implementation scheme results in a strong increase in the number of project participants. The new constituency encompasses actors from academic institutions, industry and the public sector, with a significant presence of industrial users (both SMEs and multinationals).

In December 1992, the LRE programme provided funding for the **ALEP** project, which was to create a an open, modular and versatile computational environment, for the development of large-scale linguistic resources and applications. Despite a slow start, the project, steered by a new development team

encompassing software integrators, grammar developers and computational linguists, managed to deliver a grammar development platform. This is now relatively widely used by linguists and MT researchers and developers in the academic field. ALEP comes with a number of basic tools for text handling, lingware development, debugging and linguistic processing, and a starter kit of fully documented demonstration lingware, including MT analysis modules for 9 languages that have been developed in the context of the project LS-GRAM (see below).

In preparation of the Fourth Framework Programme for Community R&D, a call for exploratory actions within the language area was launched in 1993, within the context of a bridging action called "the Multilingual Action Plan" (MLAP). Amongst the 15 projects selected for funding under MLAP, two projects are concerned with MT: CAT2-EDS and TRADE.

The aim of **CAT2-EDS** (1994-95) was to bridge the gap between basic research into NLP and MT and the needs of industrial users. Project results include morphology components, MT oriented dictionaries and language modules for three languages: German, French and English.

TRADE (1994-95) was targeted at public administrations, such as the Italian social security and pension administration. In accordance with the user requirements, an MT system demonstrator was developed for the fast translation of well-defined text types, for example social security reports and statistics for communication with other countries.

The LRE project **LS-GRAM** (1994-1996) constitutes probably the largest EC-funded R&D effort in grammar development following Eurotra. LS-GRAM addresses the need for publicly available, high-quality, large-scale grammatical resources that are modular, extensible and fully documented. The project has developed in a set ALEP-based lingware modules for nine EU languages, detailed grammar specifications, and a comprehensive documentation of the results and the approach adopted. The linguistic coverage of all grammar components is based on the investigation of real-life corpora.

These resources have been structured, standardised and packaged to ensure that they can be easily accessed and downloaded via the WWW. There are plans to transform ALEP into a Java-based ALEP toolkit, that would allow access to all ALEP and LS-GRAM tools, components and resources (lingware) by clients using a "standard" browser.

In addition to LS-GRAM and ALEP, I would like to mention two other LRE projects, that were concerned with topics relevant for MT: SECC and TSNLP.

Using the Metal MT development platform and lingware components, **SECC** has developed a controlled language checker and corrector. The application is conceived as a special language pair, English into Simplified English, and supports technical writers producing documentation in the field of telecommunication. SECC checks texts against a grammar and lexicon of controlled English, and provides either suggestions for correction or a fully automatic translation into Bell Alcatel Controlled English.

TSNLP has produced parallel test suites for three languages (German, French and English), with substantial amounts of annotated test items, stored in a variety of data base formats. Building on the results of TSNLP, and of other EC projects in the area of assessment (TEMAA, EAGLES and FraCaS), the recently established LE project **DIET** (1997-1998) will extend the data and provide appropriate user support in terms of database technology, test suite construction tools and graphical interfaces. In addition, DIET will address the issues involved in the customisation of test material to specific domains and applications.

Language Engineering Comes of Age

Launched in 1994, the Language Engineering Programme (LE), while building on the results of LRE, differs from previous EC programmes in the language field in a number of respects:

- LE funds larger-scale applications pilots that provide integrated and cost-efficient solutions in response to user requirements.
- Increased emphasis is placed on user requirements and on field trials, reflecting the user-oriented philosophy of LE R&D.
- LE promotes a more global, multi-disciplinary approach.
- LE addresses a wider range of tasks, felling into four action lines: pilot applications, research, the development of language resources, and support actions, such as standardisation and evaluation.
- The programme actively promotes involvement of industrial partners, both as developers and as users of language-aware application pilots.

User-driven, application oriented projects, designed to stimulate and respond to market needs, provide the backbone of the programme.

Work is focused on projects that integrate language technologies into information and communications systems and services. A key objective is to improve their ease of use and functionality. Pilot applications cover a broad spectrum of tasks, ranging from document creation and management through communication and information services to translation and foreign language acquisition.

Of the 25 tasks mentioned in the LE work programme, which are all concerned with multilingual issues, two tasks address specifically M(A)T related issues. The aim of these tasks is (i) to provide more effective computer-aided support for professional and occasional translators (LE1.10), and (ii) to tune MT for use in companies and administrations (LE1.11). The focus of LE1.10 is on translation aids, translators' toolboxes and telematic links for collaborative authoring and translation by decentralised teams. LE 1.11 puts the stress on process engineering, workflow, system integration and field validation, aiming at “a seamless and cost-effective integration of in-house and networked machine translation facilities into the users’ working environment and the organisations’ information and communications infrastructure”.

Among the LE projects selected for funding, a relatively small number address MT and related topics. One can distinguish between projects where MT is a key issue, and those where MT is part of a more global application. Otelo belongs to the first category of MT projects, whereas LinguaNet and Aventinus belong to the latter.

Otelo (1996-1997) addresses the entire translation cycle by providing an integrated network-based access to both machine and human translation resources. The goal is to integrate MT, translation memories and other translation support tools into the workflow processes and information and communication infrastructure of the users. Otelo will also specify standards for common lexical resource and text handling formats. The purpose is to facilitate the sharing and exchange of existing resources, that have been developed at high costs. A number of existing MT suppliers are involved in the project and will adapt their systems to the common formats.

LinguaNet (Oct. 1995 - April 1998) is working towards an advanced system for multilingual operational communication between regional police forces and emergency services in Europe. The project helps police officers to overcome language and operational barriers by providing templates for the controlled composition, manipulation and exchange of multilingual texts. A number of additional

functions are currently under investigation, including machine translation of controlled language texts, graphics transmission, the use of speech to elicit information (e.g. to check the electronic mailbox) and multilingual speech generation.

Aventinus (Dec. 1995- Dec. 1998) is engaged in the construction of a bundle of advanced language engineering technologies and tools with the aim of strengthening and improving the information systems which are already used in the effort to control the illicit trade in drugs. The Aventinus users have identified two workflows which the project will support: the handling of the inflow of multilingual information (the Indexing Scenario) and the retrieval of relevant information which is available (the Retrieval Scenario).

In the first of these two cases, the users (analysts, for example) are confronted with various types of documents in a foreign language they do not understand and have to decide whether the text they have received is relevant or irrelevant. The aim, here, is to provide the users with tools which will at least enable them to form an initial interpretation of the text and route it intelligently . These tools encompass term substitution, translation memory and raw machine translation.

In the second case, the users are confronted with the problem of filtering out, from the mass of documents and data stored in different languages on a number of databases in different countries, the facts that they actually need. In this case, Aventinus will provide the users with integrated tools enabling them to formulate their queries in natural language, preferably in their own language, and to receive the answers they need in the languages of their choice.

Following the evaluation of proposals submitted in response to the most recent LE call for proposals, a number of shortlisted proposals are currently being discussed by the consortia and the Commission services, including a project proposal in the area of translation tools, called Transrouter.

The **Transrouter** project proposal aims to develop a tool which, on the basis of an analysis of the characteristics of a source text, can aid translation project managers decide whether a document should be dealt with by human translation, by human translation using a translation memory, or by machine translation.

Despite the relatively high number of MT-oriented projects funded by the EC over the years, it is worth noting that MT projects constitute only a small subset of the 38 LE projects in progress and the 23 shortlisted project proposals resulting from the last LE call. None of the MT projects submitted in response to the third LE call survived the technical evaluation by independent experts. Of the MT projects requesting funding under the fourth LE, only one - Transrouter - made it into the shortlist.

There are probably a variety of reasons that explain the relatively poor performance of MT proposals in the EC evaluation process. In the following, I will address a few of the most striking aspects. First of all, MT does not seem to be a particularly popular area amongst LE proposal writers, judging by the relatively low number of the MT proposals received in response to the four LE calls. One of the reasons for this may be that it is hard to reconcile the relatively long development cycles of MT with the relatively short life cycles of EC R&D projects (typically 2-3 years). Secondly, a relatively high number of MT proposals pay insufficient attention to key programme requirements, such as user involvement, validation plans, cost effectiveness, etc. Thirdly, it is not rare that MT proposals are flawed in one of the two following ways:

- either they propose over-ambitious, unrealistic developments that are not supported by any evidence, and lack a clear description of the technology base and the approach to be adopted

- or they propose near-market "deployment" work, that may be relevant from a commercial point of view, but does not provide enough R&D innovation potential to justify funding with taxpayers' money.

Although MT project proposals did not perform very well in recent LE calls, translation applications - including MT - are and will continue to be a prominent area for EC language technology programmes.

Outlook

The growing importance and impact of language technology applications is widely recognised within European programmes. The Commission proposal for the Fifth Framework Programme (1998-2002) foresees a dedicated R&D action within the Information Society Programme. Several expert panels and user seminars held in preparation of a specific research programme on language technologies confirmed that there is indeed a need for MT applications. This means that it will be necessary to maintain some level of spending for basic research into MT.

It is widely believed that innovative, useful solutions will result from the integration of current - imperfect - MT technology with other technologies (eg speech recognition, indexing, human-assisted translation, etc.). Applications incorporating MT should aim at global solutions in response to user requirements, and pay due attention to business processes and workflows.

These findings are also supported by a first analysis of national surveys prepared in the context of the previously mentioned project Euromap. In almost all national R&D surveys, translation applications are listed as one of the relevant application areas for language *research*[10].

LE *market opportunities* are identified in a number of "meta-sectors", such as online information services, education and culture, electronic commerce, public services, and manufacturing. It is interesting to note that MT is expected to play a major role in quite a few of these meta-sectors, in particular in the areas of "electronic commerce" and "online information systems".

The overall picture emerging from the surveys is that there is no single LE "killer application". Instead, useful applications are expected to rely on a combination of technologies, resources and tools. This is also true for MT, illustrating the point made above, namely that useful solutions will probably increasingly rely on the integration of MT with other technologies.

Bringing the various strands of future language-related R&D together, the EC working document "Human Language Technologies: Living and working together in the Information Society" explains the rationale and strategic orientations of the next programme. The focus of future research efforts will be on a human-centered Information Society, in which facilities to access and assimilate information efficiently and effectively are available to all citizens in the appropriate language, format and mode. To that end, emerging language technologies will be integrated and embedded into 'language-enabled' services and products to support global business and to facilitate communication across languages.

R&D and validation activities will address three core themes, in which human language and communication play a central role:

- adding **multilinguality** to information and communication systems, at all stages of the information cycle. This includes content generation and maintenance in multiple languages, content and software localisation, automated translation and interpretation, and computer assisted language training;

- enhancing **natural interactivity** and accessibility of digital services through multimodal dialogues, understanding of messages, unconstrained language input-output and keyboard-less operation;

• enabling an **active use** and assimilation of digital content, through personalised language assistants supporting deep information analysis, knowledge extraction, summarisation, meaning classification and meta-data generation.

The first core theme, multilinguality, covers a range of R&D activities that will provide a home for MT. In addition, one can assume that a number of applications addressing the other two core themes will feature MT technology, components and resources in combination with other technologies.

The next EC programme for language technologies provides thus good prospects for innovative, useful translation applications, and continued MT R&D.

Conclusion

I believe that MT has an important role to play in helping to meet some of the key challenges of the emerging Information Society: it can help bring the information closer to the citizens, in their language, build bridges across languages and cultures, and demonstrate the economic impact of language enabled applications in a broad spectrum of areas, ranging from electronic commerce to online information services.

The opportunities are there. It is up to us to seize them.

---

[1] An example for a privately funded translation application is Telia's demonstrator for a spoken language translation service for Swedish-English-French in the air traffic information domain, produced in collaboration with SRI (Cambridge-UK and Menlo Park-US). The project has been presented at major speech technology conferences, such as ICSLP 96 and ICASSP 97.

[2] One example for a European MT system that was not widely publicised because of contractual restrictions is "LINITEXT". Built between 1984 and 1993 by a small team directed by Edward Johnson at Wolfson College Cambridge UK, Linitext is an English <-> French controlled language MT system for business communication with purpose built grammars (based on an analysis of an English and French corpus of genuine business correspondence), an English/French lexicon of 70.000 items, a disambiguation glossary of 4,000 items and a purpose built thesaurus of 100,000 items. Contact: Edward Johnson, phone +44 1223 276815; e-mail: Edward.Johnson@prolingua.co.uk.

[3] For example Germany and the UK.

[4] In 1993, the Minister of Science and Technology of the Belgian region "Flanders" launched a research initiative concerning speech and language technology for the Dutch language. One of objectives of this initiative is to draw up a long-term research programme concerning the development of technologies for automatic translation of spoken and written Dutch from and into other languages. This programme, which will be developed jointly with the Netherlands, seems to place emphasis on computer-aided translation *(computer-ondersteund vertalen (COV))* rather than *fully-automatic translation.*

[5] For example Austria and Luxembourg.

[6] For example CICYT and SEUID in Spain.

[7] For example the Austrian ERP-Fonds-Technologieprogramm.

[8] (http: //www.echo2.1u/langeng/en/lehome.hmt)

[9] This list is not meant to be exhaustive.

[10] A number of other areas for LE research were also frequently mentioned, such as spoken language recognition and understanding, text retrieval, information extraction, classification, summarisation, generation, authoring, NL input and/or output applications.