[MT Summit IV, July 20-22, 1993, Kobe, Japan]

E.C. Language Projects

L. Rolling

1. Inventory of European Language Projects

A global overview of European activities in the field of computer-based language processing shows a surprising variety of programmes and project types.

Of course, every European country, including E.C. member states, has its own language policy and research programme(s). This is true also for industrial companies, especially multinationals having communication problems between headquarters and their national subsidiaries.

There are trans-national research and development projects with or without non-European partners. This includes the VERBMOBIL project between Germany, Japan and the United States, and a number of EUREKA projects such as GENELEX, GRAAL and EUROLANG.

In the European Community, the Multilingual Action Plan (MLAP) was launched as early as 1977, after the development of multilingual thesauri and term banks in the Sixties and the acquisition of the Systran user rights in 1975.

In 1982 the Eurotra research project for the development of an advanced MT prototype was launched as an offspring of the MLAP, followed in 1991 by the Language Research and Engineering (LRE) project.

In parallel, the ESPRIT R&D programme for information technology, which started in 1984, included a number of highly interesting research projects in the fields of natural language processing and speech technology, such as SAM, SUNDIAL, ACQUILEX and MULTILEX.

All Community activities (with the exception of the MLAP, aimed at improving in-house language processing) are expected to converge into the ECLAT (European Community Language and Technology) programme from 1994 on.

ECLAT will cover basic research as well as application-oriented developments and pilot projects. It will involve a technology watch and a number of promotional and training activities. But its major component, during the 1994-98 period, will no doubt be the common development of standards for all nine Community languages, followed by the creation of language resources and of software tools for their acquisition, generation, adaptation and their implementation for many purposes, including research and industrial applications.

2. Multilingual Action Plan (MLAP)

The MLAP started in 1977 with a limited number of projects. The translators' multilingual terminological glossaries were fed into a huge term bank, EURODICAUTOM, which today has over 500.000 entries and 150.000 acronyms in nine languages.

Multilingual thesauri were developed for the storage and retrieval of information in multi-national data bases. The subject fields covered included nuclear energy, metallurgy, agriculture, nutrition, veterinary science, education and European matters. Software tools were developed, and an inventory of all existing thesauri was published in 1985 and re-published in 1992.

After a number of market studies and comparative evaluations, the Commission acquired a user license for a large number of SYSTRAN language couples for the Community institutions and the government agencies of the E.C. member countries. SYSTRAN has now attained a high quality level for language pairs involving English, French, Italian and Spanish, while language pairs involving German, Dutch, Portuguese and Greek require further improvement.

In the early Eighties it became evident that Europe needed to overcome the language barriers between European and Eastern languages, including Russian, Arabic, Chinese and Japanese, and that M.T. was the best solution to this problem. As a first step, the Commission chose Fujitsu's ATLAS system to translate Japanese grey literature into English for European end-users.

3. EUROTRA

Eurotra is the name of a research project aimed at developing linguistic models and parsers for all Community languages and the prototype of an advanced M.T. system.

Preparatory work started 1978 under the MLAP and the project itself went from 1983 through 1992. The resulting linguistic models and parsers are available for further research and industrial applications. All member states now have teams with linguistic expertise, and a number of follow-up projects were based on Eurotra products and expertise. This includes

- ET-7, a study on the re-usability of lexical and terminological resources in computerized applications,
- ALEP, the implementation of specific formalisms for an advanced linguistic engineering platform, and
- LS-GRAM, an LRE (language engineering) project for the development of core grammars.

4. Language Research and Engineering (LRE)

Starting in 1986, the Commission showed interest in the emerging "Language Industry" and accordingly launched a number of studies on the subject.

First came a number of economic impact studies aimed at determining the impact of language engineering and speech technology on global economy and defining the barriers to be overcome.

Second was a series of state-of-the-art studies on specific fields, including a study by Geoffrey Kingscott on machine translation.

Third was the launching of a Language Industry Survey (L.I.S.), a permanent inventory of products and services, language resources, research and teaching programmes.

Acknowledging the need for representative text corpora for all European languages, the Commission decided to participate in the Text Encoding Initiative (T.E.I.) and to establish a Network of European Representative Corpora (N.E.R.C.).

When, at long last, the E.C. decided to allocate a budget for Language Research and Engineering under its R&D Framework programme (22.5 million ECU for LRE between 1991 and 1994), two calls for proposals were launched, in 1991/92 and 1992/93, for specific industrial projects in the field of LRE.

The work of the LRE is monitored by an advisory group called EAGLES (European Advisory Group for Language Engineering Standards) whose members are the leaders of the major European projects in the field (LRE, ESPRIT, and EUREKA).

5. ESPRIT

The ESPRIT programme started in 1984. From the beginning, a large number of projects covered natural-language processing and speech technology.

All projects were co-financed by groups of industrial companies and the E.C. Commission, who contributed around 70 million ECU in the two first phases of ESPRIT. The more costly projects were in the field of speech technology, which is further away from the market than NLP. For a number of projects over fifteen partners from industry and academe volunteered for participation, showing their interest for rapid development of the new technologies.

Language components can also be found in E.C. programmes outside ESPRIT, such as AIM (medical research), DRIVE (transport) and ENS (transnational administrative network).

6. Other European Projects : EUREKA and VERBMOBIL

While the E.C. is mainly funding basic research under the so-called Framework Programme, some of its member states have become aware of the need for pre-competitive research, the results of which can directly be used for industrial applications. These industry-oriented research projects are grouped under the EUREKA programme. Several of them cover language technology subjects and deserve some comments.

Project GENELEX, initiated by GSI-ERLI (France) in 1990, is expected to produce "generic", i.e. multi-purpose dictionaries for a variety of applications. GENELEX partners agreed to make use of the lexical standard developed under the Commission's Multilex project.

Project GRAAL, initiated by Aerospatiale (France) aims to develop parsers for automatic language analysis that can be used in applications such as data base interrogation, automatic text indexing and, possibly, machine translation.

Project EUROLANG is a very ambitious attempt to develop a second-generation MT system for five European languages, based on the ARIANE software developed by Grenoble University and the lexical and terminological resources supplied by Siemens.

VERBMOBIL is a multinational programme jointly developed by

- Carnegie-Mellon University, Pittsburgh, U.S.,
- Automatic Telephone Research, Kyoto, Japan, and
- Siemens and several universities in Germany,

aimed at "face-to-face telephone translation", i.e. making use of speech recognition and MT modules for human communication. It is a three-year programme preceded by a number of studies, which may be prolonged until the end of the century.

7. ECLAT : Language and Technology programme

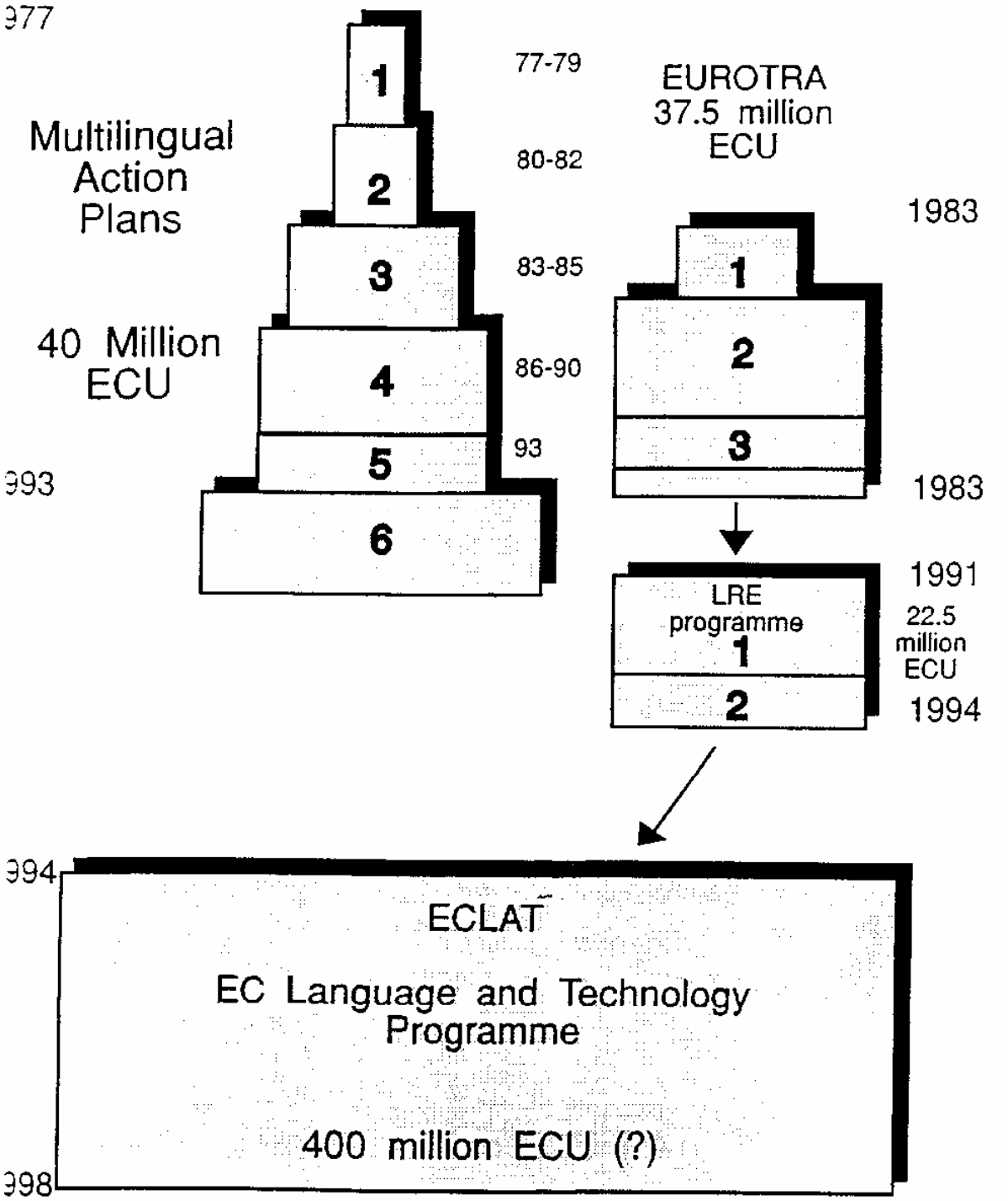
As part of the new Framework Programme for R&D to be approved in 1993 and which is expected to cover the period from 1994 to 1998, the Commission prepares a proposal for a programme on Language and Technology.

This programme has a global, long-term objective, which is the perfect mastery of European languages for information and communication throughout Europe, and it has a specific, short-term objective, which is the creation of a reliable linguistic infrastructure, including standardized, reusable language resources and software tools.

The ECLAT programme will be implemented jointly by the Commission, the national authorities and industrial companies.

The Commission will coordinate and co-finance the development of standards and basic research limited to specific language problems. The national authorities will develop corpora and lexica for their own language(s) and provide for promotion and training. Industry, finally, will have to develop their own tools likely to allow for optimal use of standard resources in their own operations.

EC LANGUAGE PROGRAMMES



MULTILINGUAL ACTION PLAN

OBJECTIVES:	1. Productivity of EC Translation Services 2. Trans-European Information Flow

CONTENT :	Terminology	5%
	Thesauri	5%
	Infrastructure Promotion Training	10%
	Automatic Translation European languages	72%
	MT Japanese and Russian	8%

BUDGET :			
I	1977 - 79	0.7	MECU/year
II	1980-82	1.2	MECU/year
III	1983 - 85	1.2	MECU/year
IV	1986-90	4.0	MECU/year
V	1991-93	5.0	MECU/year

EUROTRA

Programme phases	Planning	Implementation
Conceptual development + Association contracts	1978 - 82	1978 - 82
1. Linguistic and software specifications	1983 - 84	1983 - 84
2. Development of linguistic models, lexical databases and basic software	1985 - 87	1985 - 87
Extension to additional languages	-	1988 - 89
3. Prototype development	1988 - 89	1990 - 92

OBJECTIVES :

1. Development of linguistic models and parsers for all EC languages.
2. Creation of a prototype for an advanced automatic translation system.

EVALUATION :

1. PANNENBORG report 1987 - 88
2. DANZIN report 1989-90
3. OAKLEY report 1992 - 93

LANGUAGE RESEARCH AND ENGINEERING

LRE

ACTION LINES

1. Linguistic modelling
2. Development of computational tools
3. Development of standard, reusable language resources
4. Pilot and demonstration projects
5. Promotion and training

1991-92 PROJECTS

Title	Participating countries	Coordinating body	Subject
LINGUASOFT	UK, IR, GR	The Open Univ. Milton Keynes	Interlingual software
ONOMASTICA	UK, DK, F, D, GR, I, NL, P, E	Edinburgh Univ.	Pronunciation of proper names
TRANSLEARN	GR, UK, P, F,	I L S P, Athens	Corpus-based translation
DELIS	D, F, NL, I, DK, UK (+ Finland)	Univ. Stuttgart	Semantic lexicon-building
SISTA	UK, IR (NL)	Brain Training, Cambridge	Semi-automatic indexing
R G R	NL, UK, D	Stichting Taaltechnologie, Utrecht	Reusable grammars
DISCOURSE	NL, UK, F	Stichting Taaltechnologie, Utrecht	Declarative discourse theory
COBALT	I, UK, F	Quinary Spa, Milano	Knowledge acquisition
EAGLES	I, D, DK, UK, F, ES	Consorzio Pisa Ricerche	Standards advisory group

E.C. LANGUAGE AND TECHNOLOGY

ECLAT
1994-98

OBJECTIVES

1. Global, long-term objective: Total control of E.C. languages in information and communication :
   Political, social and cultural advantages.

2. Specific, short-term objective : Creation of a reliable infrastructure, including standardized, re-usable language resources and software tools :
   Technical and economic advantages.

ACTION LINES

1. Tools for intelligent text handling
2. Intelligent interfacing with data bases
3. Automatic and semi-automatic machine translation
4. Software tools for knowledge acquisition
5. Speech technology : analysis and synthesis

TASK DISTRIBUTION (Subsidiarity principle)

EC institutions:

- **Development of standards**
- Coordination of basic research

National authorities

- **Development of language resources**
- **Promotion and training**

Industrial companies

- **Development of software tools**
- **Pilot applications**

LANGUAGE-BASED
ESPRIT PROJECTS

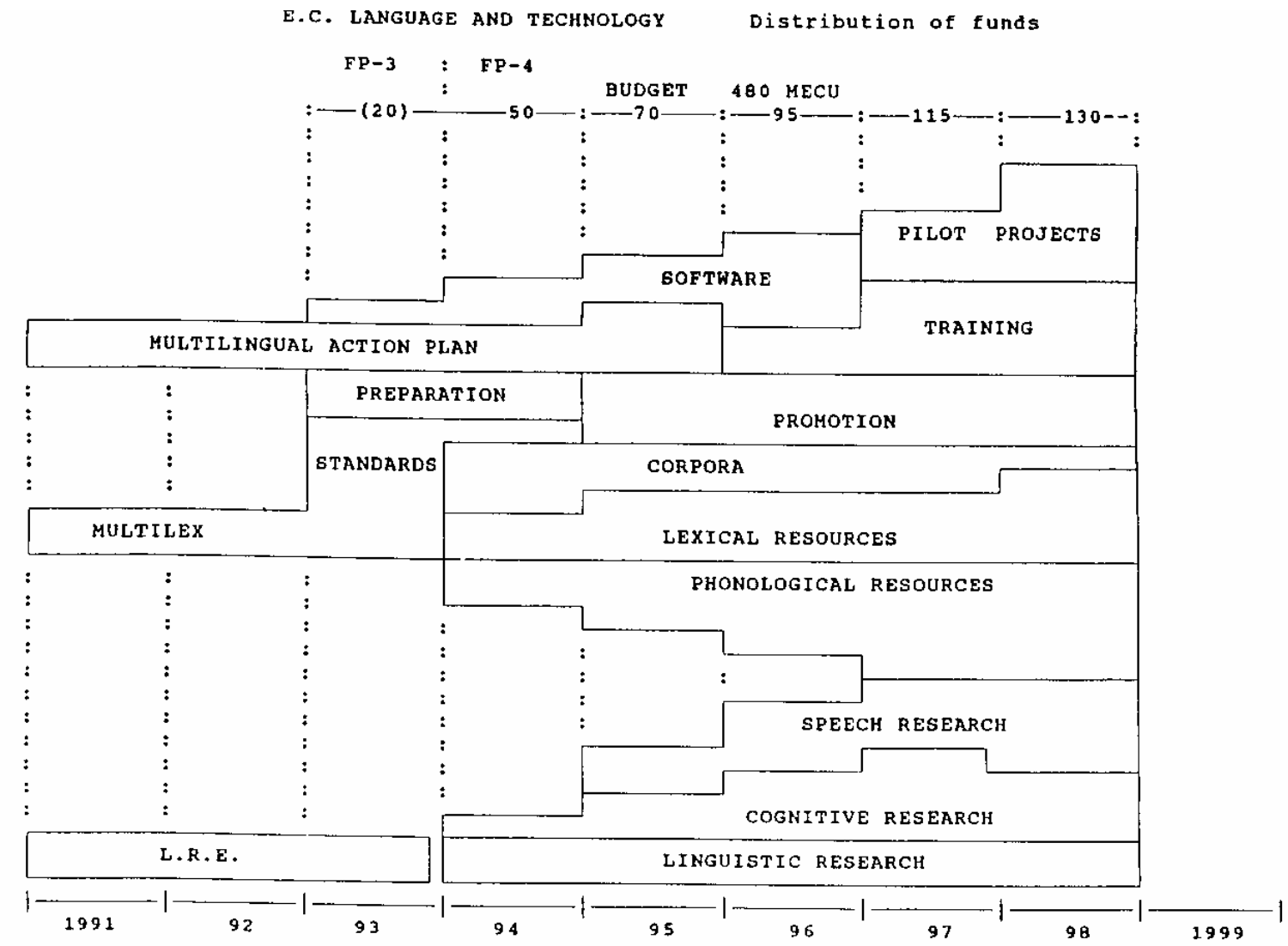
Acronym	Subject	Dates	Main partner(s)
SIP	Speech and image processing	84-89	CSELT + 6
LOKI	Knowledge bases	84-89	BIM + 6
ADKMS	Knowledge management systems	84-89	Nixdorf + 8
ACORD	NLT knowledge bases	85-90	CGE + 6
CFID	Dialogue failure	85-89	St.Patrick (Dublin) + 5
SPIN	Speech office workstation	84-89	CAPSESA + 9
LING-ANALYSIS	Linguistic analysis	85-89	OLIVETTI + 7
SPEECH	Man-machine interface	85-86	Brit. Maritime + 3
DOMESDAY	Data, voice and picture storage	85-88	PHILIPS + 5
IKAROS	Speech recognition	86-89	CGE + 3
MIS	Multilingual information system	87-88	BULL + 3
MULTILINGUA	Speech input-output assessment	87-88	UC London + 4
MULTIWORKS	Multimedia workstation	89-91	OLIVETTI + 10
TWB	Translator's workbench	89-92	TRIUMPH ADLER + 8
ACQUILEX	Lexical knowledge acquisition	89-91	Univ. Pisa + 4
DYANA	Natural language interpretation	89-91	Univ. Edinburgh + 3
ACTS	Connectionist speech recognition	89-91	Med. Res. Counc. + 5
SPRINT	Neurocomputing for speech	89-91	CAP GEMINI + 5
ACCOR	Co-articulation processes	89-91	CNRS + 8
DANDI	Dialogue and discourse	89-91	Univ. Edinburgh + 16
SUNSTAR	Speech understanding	89-94	AEG + 5
ARS	Speech recognition	89-92	CSELT + 9
POLYGLOT	Speech-to-text-to-speech	89-92	Olivetti + 10
SUNDIAL	Speech understanding and dialogue	88-93	LOGICA + 10
MMI 2	Man-machine interface	89-94	BIM + 6
SAM	Speech assessment standard	89-92	UCL + 15
SPELL	Spoken language interactive	90-92	OROS + 4
PAPYRUS	Script recognition	91-93	OLIVETTI + 6
PLUS	Language understanding	90-94	CAP GEMINI + 8
MULTILEX	Multi-purpose lexicon	90-93	TRIUMPH ADLER + 16
EMIR	Multilingual information retrieval	90-93	CEA + 3
NOMOS	Knowledge acquisition	90-93	SOGEI + 7
ROARS	Robust speech recognition	90-93	THOMSON + 3

LANGUAGE-BASED
EUREKA PROJECTS

Title	Topic	Partners	Duration	Budget
GENELEX	Generic, multi-purpose lexicon for 3 European languages (F, I, S)	GSI-ERLI, IBM, LADL, SEMA (F) Lexicon, Pisa Un. Servedi (I), ILTEC (PT) Tecsidel, Barcelona Un. (S)	3 years (91-93)	38 MECU
GRAAL	Re-usable grammar for automatic language analysis	Aérospatiale, EdF, GSI-ERLI, CNRS (F), *Fiat*, IRST, SARITEL (I) ILSP (GR), ILTEC (PT), ISSCO (CH) Lingsoft, NOKIA (SF)	4 years (92-95)	19.8 MECU
EURO-LANG	2nd generation MT system for 5 European languages (D, E, F, I, S)	SONOVISION, MATRA, CAP, LADL, CEGOS, CNET, GETA (F), SIEMENS, KRUPP (D), RANK-XEROX (UK) Pisa, Manchester, Barcelona Univ. Taurus (I)	3 years (92-94)	73.5 MECU

VERBMOBIL PROGRAMME

VERBMOBIL	Speech recognition and translation between 3 languages (E, D, J) for face-to-face dialog	CMU and Stanford Univ (US) ATR, Kyoto Univ. (J) Siemens, Daimler, Telefunken + 9 universities (D)	3 years (92-94)	± 120 MECU ± 120 MECU 120 MECU


E.C. LANGUAGE AND TECHNOLOGY
Distribution of funds
FP-3
FP-4
BUDGET
480 MECU
(20)
50
70
95
115
130
PILOT PROJECTS
SOFTWARE
TRAINING
MULTILINGUAL ACTION PLAN
PREPARATION
PROMOTION
STANDARDS
CORPORA
MULTILEX
LEXICAL RESOURCES
PHONOLOGICAL RESOURCES
SPEECH RESEARCH
COGNITIVE RESEARCH
L.R.E.
LINGUISTIC RESEARCH
1991
92
93
94
95
96
97
98
1999

LRE - 2ND CALL FOR PROPOSALS

PROPOSALS

- **RECEIVED : 88**
- **VALID : 82**

PARTICIPANTS : 400

COUNTRIES : 17

ORGANISATIONS : 320

- **UNIVERSITIES: 55%**
- **COMPANIES & R&D CENTRES : 45%**

OVERALL COST OF PROJECTS : 107 MECU

TOTAL EC CONTRIBUTION : 69 MECU

- **AVERAGE EC CONTRIBUTION : 0.85 MECU**

OVERALL RESOURCES : 13.000 M-M

- **AVERAGE COST / MAN - MONTH : 8.150 ECU**

AVAILABLE EC FUNDS : 8.5 to 9.5 MECU

LRE 1992-93 PROJECTS

Title	Participating Countries	Coordinating Body	Subject
ANTHEM	B - D - LX	RAMIT (Gent)	NLP toolkit for healthcare
GIST	IT - UK - SP - AU	IRST, Trento	Text generator
SIFT	IR - SP - NL	Univ. Limerick	Information selection from text
MULTAC	F - CH - IT - UK - D	CNRS - GRTC	Text handling tools and corpora
FRACAS	UK - NL - D	Univ. Edinburgh	Computational semantics framework
TRANSTERM	F - PT - GR - IT - SF - CH	GIS - ERLI	Acquisition and re-use of terminology
RELATOR	IT - F - DK - UK - D	Univ. Pisa	Network of repositories for linguistic resources
CRISTAL	F - UK - IT - NL	Cap - Gemini	semantic retrieval system
SECC	B - D - F	Siemens (B)	English grammar + style checker
COMPASS	D - F - (UK)	Rank-Xerox (F)	Bilingual on-line dictionaries
EURONET	SP - D - NL - CH	Siemens (SP)	Corpus toolkit
* SQALE	NL - F - D - UK	IZF - TNO (NL)	speech recognizer assessment
* Euro COCOSDA	NL - F - D - UK - IT	UC London	speech methods + resources
* TSNLP	UK - CH - F	Univ. Essex	NLP test suites
* TEMAA	DK - CH - NL - UK - IR	CTS (DK)	NLP evaluation methodology
** RENOS	GR - IT - DK	Data Bank (GR)	Retrieval efficiency

* merged proposals

** liable to budget availability

INTERNATIONAL COOPERATION

involving the E.C. Commission

COUNTRIES INVOLVED	ISSUES COVERED
1. EFTA (European Free Trade Agency) countries; Austria, Sweden, Finland, Norway, Switzerland, Iceland, Liechtenstein; some of the these are expected to join the E.C. before 2000.	° identification of standard language tools and resources ° growing participation in EC programmes (LRE, Esprit, Eureka)
2. EASTERN and CENTRAL EUROPEAN countries: Poland, Hungary, Czech and Slovak Republics, Baltic States, Romania, Bulgaria, Albania, Slovenia, Croatia, Russia, Ukraine, Belorus.	° identification of language tools and resources and linguistic research projects ° financial assistance for joint development projects ° training seminars
3. DEVELOPING COUNTRIES: Mediterranean countries, Latin American countries, India, China, etc.	° training seminars ° financial assistance for joint projects
4. INDUSTRIALIZED COUNTRIES: USA, Canada, Japan	° clearinghouse-type repositories ° exchange agreements ° joint development of standards and evaluation criteria ° coordination of basic research activities

INTERNATIONAL COOPERATION

involving the E.C. Commission

TOPICS COVERED

1. <u>REPRESENTATIVE TEXT CORPORA</u> (EAGLES working group 1)

- **Development of SGML and TEI standards**
- **Creation of the European network of corpora for 9 languages: NERC**
- **Creation of corpora and software tools through LRE - MULTAC**

2. <u>LEXICAL RESOURCES</u> (EAGLES working group 2)

- **Development of a standard through MULTILEX (ESPRIT)**
- **Development of prototype generic lexicon through GENELEX (EUREKA)**
- **Extraction of lexical, terminological and phraseological resources through EQUITEXT**
- **Creation of core grammars through LRE - LSGRAM**

3. <u>SPEECH RESOURCES</u> (EAGLES working group 5)

- **Development of a standard trough SAM (ESPRIT)**
- **Linking speech recognition and synthesis with MT**
- **Creation of specialized resources (names) through LRE - ONOMASTICA**

4. <u>EVALUATION CRITERIA AND METHODOLOGY</u>
(EAGLES working group 3)

- **European workshops on evaluation**
- **Commission's M.T. benchmark for Systran evaluation**

5. <u>AWARENESS AND PROMOTION INITIATIVES</u>

- **Seminars on language technology in the member countries**
- **Participation in IAMT and EAMT**
- **Organization of MT-Summit-V in Luxembourg in 1995**

EC EVALUATION TIMETABLE

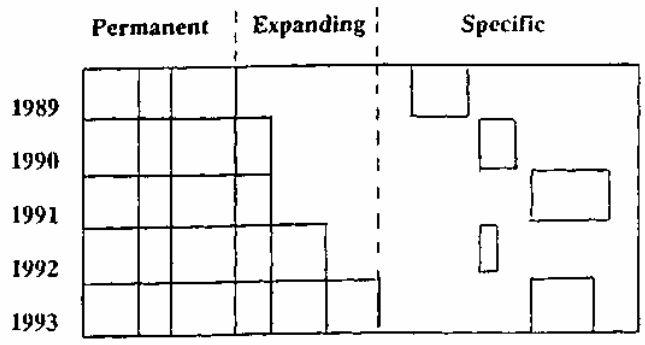
1976 - 77	**Technology Watch**
28 . 2 . 1978	**Workshop on Evaluation**
1978	**Analysis of SYSTRAN and TITUS approach**
1978	**Evaluation of SYSTRAN E→F by BMvD**
1980	**Evaluation of SYSTRAN F→E by BMvD**
from 1981	**Pragmatic, corpus - based assessment of day - to - day improvement / degradation**
1984	**On-site testing of SYSTRAN and LOGOS**
1986	**Comparative assessment of ATLAS, HICATS and SYSTRAN for J → E translation**
1991	**OAKLEY screening audit**
1992	**Monitoring of user reactions**
1993	**Introduction of periodic benchmarking**

EC BENCHMARK ACTION

1. CORPUS BUILDUP

Permanent and Expanding Corpus + Specific Corpora

- size,
- modularity,
- tagging,
- types,
- subjects,
- languages.



2. CORPUS UTILISATION

- for periodic determination of revision rate as a measure for improved cost, quality and rapidity;
- for comparison with other mainframe systems, PC / interactive systems, human translation.

3. INCORPORATION OF BENCHMARKING IN

- development environment
- **production environment**

Assessment of staff, computer time and global cost for corpus buildup, integration and use.

MT - SUMMIT - V in LUXEMBOURG

Proposed site :	**European Center, Luxembourg-Kirchberg**
Proposed dates :	**9-11 July 1995 or**
	27 - 29 June 1995
Management Committee :	**First meeting 21. 9 . 93 , Luxembourg**
Programme Committee :	**First meeting November 93, Luxembourg**
Support :	**Commission subsidy aimed at reduced participation fees.**
	Support by the City of Luxembourg (Reception, dinner)
Conference themes :	**0 New approaches : semantic and example-based**
	0 Integration and portability issues
Travel :	**Air flights into Luxembourg from München, Brussels, London, Paris, Frankfurt, Amsterdam, Zurich, Geneva, Hamburg, Berlin, Copenhagen, Rome, New York, Washington, Strasbourg, Wien.**
	Railway connections : 2/3 hours from Brussels, Paris, Basel and Cologne.
Accommodation :	**Hotel rooms at reduced prices**
Sightseeing events :	**0 City of Luxembourg (Europe's cultural capital in 1995)**
	0 Valley of the Seven Castles
	0 Petite Suisse (Little Switzerland)

Proposal by the E.C. Commission endorsed by the European Association for Machine Translation