COMMISSION OF THE EUROPEAN COMMUNITIES

COM (86) 431 final

Brussels, 25 July 1986

FIRST ANNUAL REPORT FROM THE COMMISSION TO THE COUNCIL AND THE EUROPEAN PARLIAMENT

EUROTRA

SECOND ANNUAL REPORT FROM THE COMMISSION TO THE COUNCIL AND THE EUROPEAN PARLIAMENT,

INCLUDING THE FIRST ACTIVITY REPORT OF THE_EUROTRA ADVISORY COMMITTEE ON PROGRAMME MANAGEMENT (ACPM)

EUROTRA

FIRST ANNUAL REPORT

Submitted by the Commission to the Council and the European Parliament[1]

I. INTRODUCTION

1. According to Article 1 of the Council Decision the Programme started the day of its publication in the Official Journal of the European Communities on the 13th of November 1982.

2. However, the time necessary to unblock the budget, to prepare an over-all programme of work and expenditure scheme and to negotiate the first series of contracts with the universities and research institutes in the Member States puts the effective start up of the work to January 19S3.

3. The report covers the work performed between January 1983 and June 1984.

II. ORGANIZATION

Advisory Committee on Programme Management

4. The Council Decision stipulates that the Commission shall be responsible for the execution of the programme, in particular by means of research contracts and that it shall be assisted by an Advisory Committee on Programe Management (ACPM).

5. The Commission invited the Member States to nominate their delegates in December 1982, and by early March 1983 all delegations were appointed, so that the first meeting could be held on the 7th of April 1983. At this meeting Dr. B.K. Appleyard, Director General of DC XIII was elected Chairman. The list of the Members of the ACPM is given in Annex 2 to this Report.

6. Since then the ACPM met on 27 June 1983, on 19 October 1983, on 7 February 1984 and 5 June 1984. At its second meeting (27 June 1983) the ACPM decided that the plenary sessions of the Committee left too little space for in-depth discussions of scientific and technical matters and created a Technical Subcommittee (TSC) with the mandate to prepare the opinions and advice of the ACPM on scientific matters. The TSC met on 26 September 1983, 18 October 1983, 10 January 1984 and 6 February 1984. The ACPM approved the Terms of Reference of the TSC at its meeting of 19 October 1983, and Prof. A. Zampolli, of the University of Pisa, was elected chairman at the meeting of 10 January 1984.

[1] In accordance with Article 3, second paragraph, of the Council Decision of 4 November 1982 (OJ No. L 317, 13.11.1982, p.19).

7. During its meetings the ACPM and its TSC examined in depth the proposed project organization, the financial planning, and the content and progress of the research work carried out under contract and gave the Commission its advice in accordance with the terms of reference set out in Annex II of the Council Decision. The list of the documents drawn up for the ACPM and the TSC is given in Annex 3 of this Report.

Contracts of Association

8. EUROTRA is a shared-financing R&D programme with an over-all estimated budget of 27 million Ecus, of which 16 million Ecus of Community funds and 11 million Ecus of national contributions. Annex I. point 2(a)2, fourth indent, sets out that the financial and other contributions of the associated parties shall be defined in contracts of association.

9. Unlike other Community R&D programmes, Eurotra aims at the creation of a single system with contributions from all Member States. This creates considerable complications in the balanced attribution of the responsibilities of each Associate, in the definition of the share between the Community and national contributions, taking into account the differences in population and economic power between Member States and in the contract management structure.

10. A standard contract has been drawn up including programmes of work, financial provisions, regulation of intellectual property rights and a common management structure, on which an agreement in principle of all Member States has been reached, and bilateral negotiations to work out details concerning single parties are in progress. It is expected that the contracts of association can be signed by mid-1984,

Project Team

11. The execution of a de-centralized co-operative R&D programme aiming at a unique objective, i.e. the creation of a machine translation system of advanced design including all official languages of the Community, requires a strong scientific and technical coordination of the work. Indeed, Article 2 of the Council Decision sets out that the funds allocated to the Programme include the expenditure on a staff of eight temporary agents.

12. However, the budgetary authorities have not authorized any posts in the years 1983 and 1984, leaving the Programme without a sufficient project team.

13. In order to overcome this difficulty, at least partially, the Commission has negotiated with a number of University Institutes full or part-time secondment of highly qualified scientists to the project.

14. These scientists are assigned tasks originally intended for the project team, in particular, the drawing up of the system specifications and the scientific coordination of work of the research teams in the Member States.

15. However, since these scientists remain located at their own institutes, this situation adds another degree of complication in the decentralized organization scheme.

III FINANCIAL STATEMENT

16. The Council Decision allocates 2 million Ecus for the execution of the first (preparatory) phase of two years (c.f. 2(a) of Annex I), which were fully made available to the budgetary exercises 1982 and 1983.

17. The over-all planning for the preparatory phase foresees the following break-down of the Community expenses :

. Language-specific work in the Member States	**666,000 Ecu**
. Software specifications	**500,000 Ecu**
. Central team (see 13 above)	**400,000 Ecus**
. Operation expenses (equipment, travel, workshops, seminars etc.)	**434,000 Ecus**
Total	2.000.000 Ecus

18. The Community expenditure is to be complemented by approximately 750.000 Ecus of national contributions to the research teams working in the Member States.

19. The delay in the preparation and negotiation of the contracts of association is likely to provoke some reduction of the national contributions in 1984. Measures are being taken to balance possible distortions of the initial planning at the beginning of the second phase of the Programme.

20. In the course of the first year of the Programme, contractual commitments for 975,000 Ecus have been made, respecting to a large extent the initial planning. In early 1984, contracts for 442,000 Ecus have been awarded to continue the specifications work. A list of contracts awarded is given in Annex 4 to this Report.

21. The lack of the national contributions in 1983 and, in some cases, of decisions at national level about the organization of the work, have somewhat slowed down the progress of the work dona by national research teams. It is, however, expected that after signature of the contracts of association the work will be accelerated, and the objectives of the preparatory phase will be achieved.

IV. SCIENTIFIC WORK

22. The objective of the preparatory phase of the Eurotra R&D programme is to draw up system specifications which can be used as the basis for the work of the second phase, i.e. to build a small-scale experimental system with a vocabulary of approximately 2,500 entries in the seven Community languages.

23. It should be noted that before the Council Decision adopting Eurotra, several years of Intensive preliminary design work had been carried out, in part financed from funds of the First and Second Plans of Action for the Improvement of the Transfer of Information between European Languages. This had led to an overall consensus about the architecture and characteristics of the system, on which the specifications could build.

24. The Council decision sets out the following objectives for the preparatory phase :

- definition of the methodology of the work,

- preparation of a detailed programme of linguistic work to be carried out by the participating centres and of the sectors and categories of text* to be covered by the research,

- preparation of more detailed specifications of the linguistic models and strategies for the various components of the system (analysis, transfer, generation),

- preparation of detailed and binding specifications for the Eurotra basic software and the data processing programmes capable of carrying out the various processes : analysis, transfer, generation, monitoring functions and text management,

- preparation of more detailed specifications for the lexical data base.

25. These points will be treated below. It should be kept in mind, however, that this Report covers the first part of the preparatory phase, and is expected to be completed by the end of 1984.

Methodology of Work

26. This point applies mainly to the linguistic work carried out by the national research centres participating in the Programme.

27. The de-centralized co-operative nature of the project imposes a far more explicit spelling out of the underlying assumptions, principles and theories than in any preceding R&D project in machine translation or related fields, which have been without any exception strictly centralized and therefore could rely on the implicit assumptions and idiosyncrasies of a project leader or group.

28. Machine translation is basically an engineering exercise, which takes existing theories where available and adapts them to the practical needs, and complements them with pragmatic solutions, where suitable theories are not available. In order to avoid duplication of effort, i.e. re-invent existing theoretical solutions, the first methodological task consists in identifying such theories. Since linguistics belongs to the domain of humanities, its schools are generally tributary to some philosophical school, and, as in philosophy, their fundamental presuppositions - usually undeclared - may be diametrically opposite to each other.

29. Most of the scientists contributing to Eurotra are located in the ten Member States. There exist considerable differences between linguistic schools between the various countries and also within them. Therefore what within the Eurotra community is called "differences of training, background and experience" constitutes a real problem, because an assertion can be interpreted very differently if related to the one or other underlying theory. This can lead to serious misunderstandings and unending disputes, in particular, if one takes into account that the number of people participating in the project is bound to grow and newcomers may not always be aware of the genesis of the project and of the reasons for some of the choices made in the past.

30. A considerable attention has been therefore given to this problem, and an important part of the central linguistic work (in the framework of the studies ETL-3-B/WL/UK) was dedicated to methodology, the explicit definition of the linguistic framework and the drawing-up of reliable criteria for choice, when several options are open. The final report covering the work carried out in 1983 is available, and work is to continue throughout 1984.

Programme of Linguistic Work

31. At the beginning of the Programme an overall programme of work of the participating centres was prepared and agreed upon both with the research institutes concerned and with the ACPM (c.f. doc. EUROTRA ACPM/11/83).

32. Within the framework of this programme two series of studies were awarded to research institutes in the Member states (ETL-1 and ETL-4) in 1983. The final reports are now available, and have been evaluated.

33. The detailed programme of work for 1984 has been prepared and agreed upon with the Technical Subcommittee of the ACPM and with the research institutes concerned. It is part of the Contracts of Association.

34. A detailed programme of work for the second phase, covering both the research work of the research institutes and the implementation plan of the first version of the system has been prepared by the Commission and is currently being discussed with the Technical Subcommittee of the ACPM and the research institutes, (c.f. doc. TSC/4/84).

35. The opinion delivered by the Technical Subcommittee and approved by the ACPM at its meeting of 19 October 1983 endorses the programme of work, states that the work is on schedule and appreciates its quality (c.f. doc. ACPM/21/83).

Sectors and Categories of Text

36. The work of the second phase, which is to produce a small-scale working system is to be based on a corpus and vocabulary in a limited field, estimated at around 2 500 entries. The choice of the corpus and of the subject field is part of the programme of the preparatory phase,

37. The choice of the corpus has to meet among others the requirement that parallel texts should be available in all seven languages of the Community. This limits the choice virtually to Community documents. After screening of types of documents available, a provisional choice has been to take Communications of the Commission to Council and the relative proposals for Council decisions.

38. The main pre-occupations in the choice of the subject field were on the one hand its present and future importance, and on the other the desirability of keeping the participating scientists as independent as possible from external expertise.

39. Since Eurotra is itself an R&D project in information technology, and the main competences of those participating are linguistics, artificial intelligence, computer science and information science, it appeared as a natural choice to concentrate on information technology or subfields

40. A number of documents, e.g. ESPRIT, are being analyzed with respect to their suitability as becoming part of the corpus in order to come to a definitive choice.

Linguistic Models

41. This work is being carried out in parallel with the methodological work (see above) (in the framework of the ETL-2 and ETL-3 studies and their follow-up) in close co-operation with the research teams in the Member states (in the framework of the ETL1/4 study series). As a matter of fact the main emphasis of the national research teams in 1983 was, and will be in 1984, put into the definition of the linguistic models and strategies and into the verification of their adequacy for their respective languages.

42. As to the linguistic models, both centrally and in the national teams, first priority was given to the determination of the role of transfer and to the definition of the representation which serves as switching point between the three main processes of translation - analysis, transfer and generation - i.e. the Eurotra interface structure.

43. In accordance with the state of the art in linguistic research, there exists a general agreement that an interlingua based system, which is a multilingual translation system on theoretical grounds would be desirable, is not feasible, and that certain portions of the languages, especially the vocabulary, must be treated contrastively, for each language pair separately.

44. This, however unavoidable, choice has the consequence that the number of transfer components, together with the language pairs grows geometrically (n * (n-1) for n languages), whilst analysis and generation can be done monolingually, i.e. once for each language.

45. Although transfer cannot be avoided, it is a general system design principle to make as much as possible in the monolingual components and to reduce the transfer to a bare minimum, ideally to the lexical component.

46. It is the task of the analysis to produce from the source text the interface structure which contains all the information necessary to make transfer and generation in all target languages possible, without knowing into which language the translation will go.

47. Transfer has to accept the interface structure representing the source text and to produce a corresponding interface structure, representing the target text, i.e. containing the transferred information which by definition belongs to the target language, and sufficient additional information computed in analysis to make monolingual generation possible, using exclusively the knowledge of the target language.

48. The preliminary work before the Council decision had produced a fairly precise idea about the type of interface structure required in a multilingual system, and in particular of the depth of analysis necessary, which might be characterised as deep semantic representation with, in addition, a number of indications on more superficial, language-specific phenomena in the source text (which are classified in a generally understood way).

49. The objective of the preparatory phase is to produce a definition ("legislation") of the interface structure which is adequate for all the languages involved, and can be used as the basis for the work of the second phase, keeping in mind that this definition must necessarily be open-ended, and adjustable to needs emerging during the experimental work during the second and also the third phase of the Programme.

50. In the first half of the report period (ETL-1 studies) the researcb teams have examined the existing proposals for the interface structure with respect to two aspects, their consistency and adequacy for the respective languages. The purpose was to uncover inconsistencies and gaps, and, where possible to make proposals for the solution of the problems met. The findings and proposals are contained in the final reports of these studies. To ensure cross-lingual communication, the teams co-operated in subject groups specializing in the three main aspects of the work : structural aspects, taxonomy and strategies.

51. In the second half of the report period (ETL-4 series) three priority areas were investigated across the seven Community languages. These priority areas had emerged in the course of the preceding period, both in the ETL-1 work and in the first interim report of the linguistic specifications (ETL-3). They are : semantic relations, time/tense and modality. The initial proposals were tested against each of the languages, and a number of proposals for improvement of the legislation were made. The cross-language communication was again ensured through subject groups each of which was co-ordinated by a member of the central team.

52. The work remaining for the second year of the preparatory period is on the one hand to clarify some specific aspects of the geometric properties of the interface structure (e.g. conjuncts, quantification, traces etc.) and some semantic categories like singularity/plurality, determination, negation, quantification, type of discourse, scope, speech act etc., which have not yet been treated in sufficient detail.

53. A further area of research, especially for the central team is to prepare a legislation for a number of intermediate representations of linguistic and/or strategic interest.

Linguistic Strategies

54. With respect to linguistic strategies an important principle followed in the software design should be emphasized here. The basic software (see below) does not anticipate any linguistic choice, not even the subdivision of the translation process into analysis, transfer and generation, but offers tools for devising and implementing linguistic strategies.

55. This design, as compared with the architecture of previously conceived systems, shifts many of the responsibilities which were traditionally the domain of the computer scientists into the domain of the linguists. This applies in particular to linguistic strategies.

56. Given the novelty of the software design and of this particular task for the linguists, the work carried out during the report period was mainly exploratory and preparatory. Along with the effort of familiarizing

themselves with the facilities offered, the teams investigated in particular how well-known linguistic strategies implemented in existing systems can be emulated with the new tools.

57. The work on strategies is to continue throughout the preparatory and the second phases of the Programme.

Software Specifications

58. In the framework of the Eurotra R&D programme the basic software has the primary function of a support tool to the linguistic research work to be carried out during the programme period. Second, it must be suitable as a basis for an industrial development after the completion of the programme.

59. The preliminary work carried out on software design before the Council decision had produced a fairly precise picture of the over-all software architecture, taking into account the user requirements -"users", here, are primarily linguists building the machine translation system - and the detailed knowledge of the software underlying a number of existing machine translation, natural language processing and artificial intelligence systems.

60. Details of the software design are given in the documents ACPM/6/83, ACPM/14/83 and the final report of the ETS-1 study. Here, only some of the most important highlights of the system architecture and design philosophy are presented.

61. The software is conceived as a production system enhanced with a powerful control language. This architecture gives the user, i.e. the linguists, the possibility of encoding his expert knowledge in a declarative way, and to specify again in a declarative way the class of data to which this knowledge is to be applied as well as the time and the manner of its application.

62. In order to assure a maximum of flexibility and to give the software a broader scope than just machine translation or Eurotra, the software has been designed as a system generator system. This means that the abstract machine which is implemented covers a very broad class of possible applications, and for each application the "concrete" machine is specified externally through the definition of legal data structures and types and manipulations, and of the syntax and semantics of the user language, and of the associated interpreters.

63. The software specifications foresee the definition of one user environment suitable for the Eurotra programme. This initial definition is, however, open to extensions and/or modifications, when the intensive use of the software by the Eurotra linguists provides feedback and uncovers specific user requirements.

64. The programme of work of the preparatory phase sets out—the preparation of detailed and binding software specifications. These specifications will be the basis for an implementation by a contractor to be selected by an open tenders procedure at the beginning of the second phase. This means that the industrially produced software will become available in 1986 at the earliest.

65. The methodology chosen for the preparation of the software specifications, following advanced software engineering principles. includes, along with the software description, the construction of an operational experimental software assembly using available system specification tools such as FP (functional programming language), Prolog and compiler compilers.

66. The experimental assembly is being constructed in parallel with the drafting of the specifications. The programme of work foresees that the first release will become availablo by mid 1984 and that this release will be tested and refined in the second half of 1984.

67. The primary purpose, in the prospective of software engineering, of producing "runnable specifications" is to verify the correctness and completeness of the specifications themselves first, and later on, to validate the conformity of the system behaviour with the specifications, when the software is implemented industrially.

68. For the Eurotra R&D programme this specification methodology has another important beneficial side effect : the experimental assembly can be used by the linguists as a software tool, whilst awaiting the availability of the industrial implementation.

69. It should be stressed that the experimental assembly imposes limitations on efficiency, the quantity of data treated and on the user comfort. But its availability makes it possible to start linguistic experimentation on the computer two years in advance of the industrial implementation, and this gain of time is likely to become a key factor for the success of the whole of the programme.

70. Both specifications and the experimental assembly are proceeding on schedule. In order to accelerate the procedures for the selection of appropriate contractors for the industrial implementation of the software, it is planned to publish an advance notice of a call for tenders in the second half of 1984.

Lexical Data Bases

71. Lexicography is a linguistic problem, but the organization and use of lexical information in a machine translation system depends crucially on the underlying software tools.

72. In the course of the Report period, work has been carried on both on the linguistic and the software side in parallel, and will continue over the second year.

73. On the linguistic side, investigations are being made, especially in connection with the linguistic specifications, on the types of linguistic information to be included in the dictionaries (e.g. morpho-syntactic, valency, semantic features, case frames, semantic formulae, thesaurus-like classification schemes etc.), and their use in analysis, transfer and generation,

74. On the software side, it should be stressed that the software design does not explicitly foresee the existence of separate entities called "dictionary" or lexical data bases. The software design gives the

possibility of encoding linguistic information in a declarative way, and to use it for the solution of a glvea problem. It is the competence of the linguistic interpretation to consider certain portions of the linguistic information as pertaining to the lexical domain (mono- and multilingual dictionaries). At far as the sotware is concerned, it is the responsibility of the data base management component to maintain this data, and to provide reliable and efficient access aad manipulation mechanisms.

V. SUMMARY

75. The programme of the scientific work is proceeding in conformity with objectives set out in Annex I of the Council decision, it is on schedule, and it can be anticipated that the objectives will be reached by the end of the preparatory phase.

76. As far as the software specifications are concerned, it should be stressed that with the implementation of a useable operational experimental assembly more will be achieved, within the time limits and budgetary allocation, than foreseen in the Council decision.

77. The availability of the experimental assembly before the end of the preparatory phase is a crucial element for the success of the whole of the programme. Waiting for the industrial implementation (i.e. end 1986) would have paralyzed the experimental linguistic work in the Member states, and seriously jeopardized the outcome of the Programme.

78. The delays in the conclusion of the contracts of association which, among others, are the legal basis for the national contributions to the Programme, have somewhat slowed down the definition of the national structures and the scientific work of the national research teams. There exists, however, a reasonable assurance that these contracts will be signed soon and that these delays will have no serious consequences on the outcome of the preparatory phase.

79. The failure of the budgetary authorities and of the administration to allocate the eight temporary posts for the Project Team foreseen in the Council Decision poses serious management and coordination problems. The measures taken to overcome these problems have proven quite successful and the scientific work has not seriously suffered from the consequent additional degree of decentralization. It is, however, anticipated that the execution of the second phase will impose additional requirements on coordination, and the lack of a central team in place in Luxembourg may constitute some risk for the success.