ESTUDIS

EUROTRA. *THE MACHINE TRANSLATION PROJECT OF THE EUROPEAN ECONOMIC COMMUNITIES*

per Peter LAU

The Commission of the European Communities
Luxembourg

I. INTRODUCTION

The author of this article is an employee of the Commission of the European Communities in Luxemburg and is currently a member of the Project Team of the EUROTRA Machine Translation Project.

The ideas presented in the article are the property of a large number of people involved in the project, but the author is solely responsible for the presentation of the ideas in this article, including errors and misrepresentations.

The paper describes the background and the aims of the project, its history from 1978 until today and its future perspectives which also include an extension of the project to cover Spanish and Portuguese. The presentation given is rather general and does not include detailed and specific descriptions neither of the software which has been developed nor of the linguistic specifications on which the work of the participating national groups is based. Readers who take a special interest in these things will find references to more specific papers, reports, books, etc. in the appended reading list.

II. BACKGROUND

The translation services of the institutions of the European Communities (Commission, Parliament, Council, Court of Justice, Court of Auditors, Economic and Social Committee and a few others) employ a joint staff of some 2.000 people. These services provide translations from any

Community language into any other (nine official languages since 1 January 1986). This means that they must cover seventy two language pairs and a gradually broadening range of subject fields, often of a very specialised nature. To do this they need highly qualified (hence well-paid) translators, and with an ever increasing load of translations they seize an ever increasing part of the administration budget of the institutions.

This problem could be solved in various ways. One possibility would be to reduce the number of official languages and agree upon two or three working languages. For political reasons this solution would be totally unacceptable, although it may, in practice, be implemented locally and temporarily in order to ease the burden on the translation services. Another possible solution might be to try to increase the productivity of the human translators. It is well-known, however, that human translators are unable to surpass a certain threshold (although there is no agreement on the value of this threshold: four-seven-ten pages a day?) without a substantial reduction of the quality of the output. Under these constraints the last possible solution seems to be of a technological nature. Part of the work must be taken over by machines, if productivity is to be increased.

In view of this situation it is only natural that the Commission should turn to machine translation (MT) as a possible means of coping with an increasing workload without consuming the entire budget. In 1976 it acquired SYSTRAN which was —and probably still is— the most efficient and highest developed general batch processing system available (other systems like the Canadian METEO and the American systems ALPS and WEIDNER or even LOGOS and METALS are either relatively small, bound to specific language pairs or subject fields or depending on interactive processing, i.e., they need an operator to solve ambiguities found by the system during analysis of the source text).

The development of SYSTRAN, however, started twenty five years ago. In those days computer science was still in its infancy, and computational linguistics was mainly treated by science-fiction writers. It contains only fairly primitive means of linguistic analysis, and it is written in a low level programming language which makes it very difficult for a linguist to understand what happens during the processing of the input text.

Consequently, it was found that the state-of-the-art in linguistics and computer science in the late seventies offered some perspectives for new research on MT. The Commission asked representatives of European institutes working in MT and related fields to study the feasibility of developing an advanced system on a collaborative basis, and after almost five years of preparatory work the Council of Ministers adopted a decision on 4 November 1982 on a research and development programme with the aim of creating a machine translation system of advanced design (EUROTRA) capable of dealing with all the official languages of the Community.

The programme period was divided up into three phases:

1. The first (preparatory) phase in which the organisational infrastructure and the basis of the work on the national languages were to be established. The programme envisaged a co-financed project based on Contracts of Association between the Commission and the Member States, and during this phase the Member States were supposed to set up national groups to carry out the work on the official languages, while the Commission was responsible for the drawing up of linguistic and software specifications to enable the national groups to do development work.

2. The second phase in which a small system capable of treating texts from a limited subject field should be built. This system should cover a vocabulary of around 2.500 lexical entries in each of the languages.

3. The third phase should mainly be used for an expansion of this vocabulary in order for it to cover 20.000 lexical entries at the end of the project mid-1988.

It was acknowledged that a successful realisation of a project of this kind presupposed some linguistic research, especially in semantics. This had become clear from the evidence gathered during the preparatory discussions. According to the Council Decision, research work should concentrate on linguistics, while the software development should be based on already available technologies.

Nonetheless, it was obvious from outset that the collaboration between linguists and computer scientists which is absolutely crucial in a project like EUROTRA, would have to be based on some kind of "software organisation research". It was known that no linguist could give a comprehensive and exhaustive description of his language which would be readily formalizable by a team of computer scientists. On the other hand, it was expected that the interaction between linguists and computer scientists would inspire the linguists to do linguistics in new ways and find solutions at least to some of the problems which would have to be solved if a machine translation project should have a chance of leading to success.

The developmental model which was chosen for this interaction was the so-called rapid prototyping. Instead of making a full and detailed description of the data and the procedures to be implemented before starting the process of implementation, the project should start out on the basis of a tentative problem description, and by means of high level programming languages (Prolog and Lisp) and advanced tools (running under UNIX: YACC, LEX, EMACS and others) a first implementation should be made in the form of running specifications. This implementation should then be used directly by the linguists working in the language groups and the

feed-back from this implementation process would lead to a revision of the specifications and a new implementation.

It was the intention to freeze the specifications at a certain point during the second phase of the project and to let an external contractor provide a full industrial implementation and environment for the extension of the grammars and dictionaries in the third phase.

However, employing the rapid prototyping methodology in a decentralized environment like Eurotra, where all the work on analysis and generation of the individual languages is done by national research teams working at different sites in the Member States, implies a high risk of ending up with a series of incompatible modules.

The decentralized nature of the project does require a high degree of modularity, but in order to make sure that the modules will fit into one final system, a detailed problem definition and a strong scientific framework should be provided, and it was one of the main aims of the preparatory phase to set up a framework on the basis of the problem description which had emerged from the discussions of the five years before the adoption of the Council Decision.

It was generally agreed that Eurotra would have to embody a particular formal theory of translation, if software developers and linguists were to communicate and collaborate in a productive way. The creation of such a theory was never listed among the explicit goals of the project, but it has been a constant concern of the central project bodies, and it is cetain to be one of the major results to be listed for the final evaluation.

III. Aims

The primary objective of the EUROTRA programme is the production of a pre-industrial prototype machine translation system of advanced design and covering all the official languages of the Community. The system is supposed to cover a vocabulary of 20.000 entries in each language and to be able to treat a limited number of subject fields, especially information processing and other kinds of information technology.

The third phase of the project will comprise the drafting of an industrial implementation programme which should lead to the production of a full scale industrial machine translation system offering a sophisticated environment and the possibility of using various system modules independently of one another (e.g. subsystems for single lenguage pairs, specific subject fields, etc.).

In addition to the primary objective there are a number of secondary objectives of considerable importance. EUROTRA is by far the largest European research and development project in computational linguistics. It is the first machine translation project which aims at simultaneous treatment

of 72 language pairs (the nine official languages of the Community), and it is unique in the sense that it builds on the participation of research teams from all Member States and that it is open to the participation of research teams from third countries (at present negotiations are going on with Switzerland).

The project was started in spite of the fact that it was known that the number of computational linguists available in the Member States was extremely reduced, and it is an explicit aim of the project to contribute to the establishment of research centres for computational linguistics and training of computational linguists in Europe.

Moreover, the project aims at promoting scientific cooperation along three different lines. The most obvious kind of cooperation takes place between linguists and computer scientists, but it is essential that this kind of project also furthers the contact between scientific teams in the different Member States, and the third line of cooperation concerns the contact between research institutes and industry.

With the promotion of computational linguistics and scientific cooperation it is hoped that EUROTRA will be of considerable importance, not just for future developments in machine translation, but also for related areas like natural language processing, speech analysis and synthesis, construction and maintenance of large data bases, advanced text processing, etc.

Therefore, the emphasis has been put on the quality of the output rather than on the speed of performance. It is a well-known fact of software engineering that the performance of a system prototype which has been developed along clean theoretical lines may be speeded up by the introduction of short-cuts, limits on search routines and other *dirty tricks*. It is also well-known, that maintenance and updating of systems which have been improved in this way may be extremely difficult. Thus, in order for EUROTRA to be repairable and extensible, it is important that the theoretical framework which has been set up for the project is respected as far as possible.

It would be meaningless to employ the rapid prototyping methodology if every new prototype was open for the introduction of software hacks with the purpose of speeding up the performance.

At the same time, though, it is impossible to do rapid prototyping if the performance of the prototypes is so bad that they are useless as test vehicles for the linguists who are supposed to provide the feedback needed for the extension and improvement process. The solution to this dilemma does not lie in the acceptance of one or the other extreme, but in finding a balance which makes it possible for the linguists, the system designers and the implementers to work and to communicate in a productive way.

The procedure for finding this balance will certainly also be among the major EUROTRA results to be considered in the final evaluation.

IV. HISTORY

It should be clear from the description of the background and the aims of the project that EUROTRA presents a lot of difficult scientific and organisational problems.

The project is supposed to produce scientific results in advanced research fields within a decentralised research scheme where cooperation is hampered not only by geographical distance, but also by differences in training and scientific background.

Linguistics is well-known for its proliferation of different schools which have grown out of fundamentally different philosophical environments and which have considerable difficulties in communicating with each other. Computer science shows more of a unified picture, but still you may find strong disagreements between partisans of different programming and implementation strategies.

Running a machine translation project with finite ressources (27 million ECU) for a finite period of time (5 ½ years) under such circumstances requires a strong management, coordination and organisation. According to the Council Decision the Commission is responsible for the project management, and the Decision foresees the establishment of various consultative, coordinating and steering committees. The real history of the project, however, shows that the scientific and organisational diffiiculties mentioned above were not the only problems which had to be solved, before the project was able to start along the lines set out in the Council Decision.

Economic and political problems in the Member States and the Community Institutions prevented the establishment of national groups and the Project Team of the Commission for shorter or longer periods of time.

The present situation (as of mid-1986 with 3 ½ years of the programme period having elapsed) is such that one Member State still has not signed the Contract of Association, and the Project Team has not reached the size foreseen in the Council Decision.

However, the majority of the national groups were established during 1984 and 1985, and for the corresponding languages it has been possible to conclude the work of the first (preparatory) phase.

The linguistic and software specifications have been developed according to the original plan by a group of scientists mainly drawn from the participating national teams and working under contract with the Commission. During the first months of 1985 the discussions within this group showed a general feeling that the experience gathered through the preparatory phase ought to lead to a general revision of the scientific framework, and the Project Management accepted that such a revision should take place.

By the end of 1985 a new framework had been developed and described in the first version of the EUROTRA Reference Manual. This manual contained the fundamental scientific guidelines for the future work and the specifications on which to base the development work in the language groups. The new framework also inspired a tightening up of the organisational structure and the project planning. At the same time the majority of the national groups had finally become operational, and it was possible to start up the implementation work planned for the second phase at almost full scale.

In relation to the plan of work set out in the Council Decision, however, a delay of one year had been accumulated due to the organisational difficulties encountered during the first three years of the programme period. Furthermore, Spain and Portugal joined the Community at 1 January 1986, and the Council Decision explicitly states that EUROTRA shall include all official Community languages (Greek was included after the Greek accession in 1981). In view of these developments the Commission proposed an amendment of the Council Decision containing an extension of the programme period (1 ½ years), the budget (18 million ECU and the linguistic coverage (including Spanish and Portuguese). This proposal is currently being discussed by the Council and the European Parliament (the two parts of the Community Budget Authority).

Considering the economic development in the Member States and the almost permanent budgetary crisis of the Community during the past three-four years, it would probably have taken a miracle for the EUROTRA programme to proceed on schedule. The scientific and organisational problems alone might have caused delays, and in combination with the economic and political ones they were bound to delay the work. Nonetheless, some important lessons have been learned from the preparatory work, and the project as a whole has shown that scientific cooperation is possible across national frontiers and the borderlines between scientific schools.

The experience gained till now shows that it is possible to produce system specifications in a decentralised environment provided that there is a strong and coherent framework to guide the production. A project of the size and complexity of EUROTRA will always suffer from delays and phase displacements, because so much work is done in parallel, but with a common framework and advanced communication means it is possible to prevent the destructive effects of these flaws.

A decisive coordinating factor in EUROTRA has been the use of an electronic mailing and conferencing system, EUROKOM, which is now used by all participants. With this system it is possible for user groups to enter in a continuous dialogue, to exchange messages and scientific results and to organise their cooperation independently of the geographical distance between the members of the groups.

IV. The present state

The situation as of October 1986 is the following: Three language groups are fully operational, fully equiped (with computers and other necessary facilities), they have completed the work of the first (preparatory) phase, and they carried through all implementation work planned for the first 9 months of the second phase. Two other groups are also operational, and they have terminated the work of the first phase, but, lacking the necessary equipment (especially the computer), they have not been able to do much implementation work. Finally, one language group has just come into existence, the Contract of Association was signed in August.

Two national groups (the Irish and the Luxemburgish) do not participate in the work on the national languages. The Belgian participation has been split up in two parts in such a way that the Dutch language is treated by Belgium ($^1/_3$ of the work) and the Netherlands ($^2/_3$ of the work), while French is being treated by France (ca. 90 %) and Belgium (ca. 10 %). Similar arrangements for Ireland and Luxemburg would have led to very small contributions from these countries because of the size and weight of their «linguistic partners» (Germany and France for Luxemburg and the United Kingdom for Ireland).

Consequently, they have assumed the responsibility for some supplementary tasks of a general nature, i.e. lexicography and terminology (Ireland) and documentation and clearing house functions (Luxemburg).

After some difficult moments during the first phase and the first year of the second phase, the scientific framework and the organisational and management structures have reached a degree of stability and maturity which makes it reasonable to believe that they will survive in their present form until the end of the project.

This is an important achievement in view of the fact that the project will change its size and duration with the amendment of the Council Decision mentioned above. The project history up till now has shown that the process of negotiating and implementing Contracts of Association is far from simple, but with a stable project environment and with an experienced project management which has had to go through this process many times before, it is reasonable to expect that the inclusion of Spain and Portugal will be smoother and easier than previous inclusions, and the strong interest shown so far by the Spanish and Portuguese authorities supports these expectations.

Extending the project to cover two new languages in the middle of the programme period is, of course, a major undertaking. It does not mean, however, that the two new groups will have to start from scratch. They will have access to the scientific framework and to software tools which have already been tested to some degree, and they may profit from the experiences of other language groups as far as possible.

They will find a set of fundamental concepts and ideas which have been developed over a period of more than eight years (the first project preparation group was established in February 1978) and revised in the light of experimental results from implementation work.

V. Basic concepts and ideas

1. EUROTRA is a transfer system based on elaborate *monolingual* analysis and generation modules.

The translation of a text from one language into another may proceed through the establishment of equivalences between textual elements in the two languages. This is what happens in a normal bilingual dictionary, i.e. «horse = caballo». We might, however, adopt a general interlingual relation stating that «*horse, caballo, cheval, Pferd,* etc.» should be represented by the symbol #. In the former case, we need bilingual dictionaries of all pairs of those languages from which or into which we want to translate. In the later case, we just need one dictionary per language establishing the relations between the elements of this language and the interlingual symbols (maybe two dictionaries, if there is no one-to-one relation). The consequences in terms of dictionary requirements are overwhelming: if we need a dictionary for each language pair, the number of dictionaries grow according to the formula n (n−1) for n languages. If we just need two dictionaries per language (one translating into the interlingua and one from the interlingua) the formula is 2n. For n=9, which is the present situation in the Community, this means either seventy two dictionaries (language pair approach) or eighteen dictionaries (interlingual approach).

Unfortunately, nobody has developed a convincing interlingua to be used for translational purposes, and many people even tend to believe that the development of an interlingua is impossible on theoretical grounds. Without entering into this discussion, we can state that no interlingua is available for EUROTRA, which means that the project has to rely on the language pair approach. In order to minimize the time and effort spent on writing bilingual dictionaries for translation proper (the so-called transfer, hence the name transfer system), the system has been designed in such a way that transfer is made as simple as possible, i.e. direct lexical equivalence like «horse = *caballo*», «*caballo* = *cheval*».

Nonetheless, some horses are of a special kind which may prove destructive to lexical transfer. In German, e.g., white horses enjoy the privilege of having a special word to denote them: «*Schimmel.*» Of course, we might make lexical entries like «*schimmel* = *caballo blanco*» in our transfer dictionaries, but once we start this process, we learn like the Sorcerer's Apprentice, that we have appealed to forces which we cannot control. If

«*caballo blanco*» is an entry in a Spanish dictionary, «*base de datos*» will be another, and what about «*armar un lío*» which may well correspond to one word in German or Danish. And if we have come that far, the next step will be to enter syntactic phrases or whole sentences which may not be translated in such a way that the traslation of the whole is a function of the traslation of the parts (idioms, fixed and semi-fixed phrases). This process will never come to an end, because it will always be possible to find new shades of meaning which characterize phrases in a particular context, and so the dictionaries will grow and grow until they reach a size with which no machine and no team of lexicographers can cope.

2. Therefore, in addition to adopting the transfer approach, EUROTRA has introduced fairly sophisticated monolingual analysis and generation modules.

The idea is that by considering big units (sentences) as rule-based collections of small units (words or morphemes), we shall be able to analyse the big units in such a way that the translation of a sentence will be a function of the translations of its parts. This principle is known as compositionality. It is quite obvious that the sentence «John eats the apple» may be translated word-for-word into «*Juan come la manzana*». However, «John kisses Mary» will be «*Juan besa a María*» and we do not want a dictionary entry saying that «Mary» may correspond to «*a María*» if she is the object of the sentence. We want an analytical module for English which finds out that «Mary» is the object of «John kisses Mary», and we want a generation module for Spanish which inserts «*a*» before personal sentential objects, so that we may have a transfer formula which says «Mary = *María*».

By following the compositionality principle we hope that we shall come pretty close to simplifying transfer so much that the majority will be a one-to-one mapping of lexical units. Idioms and fixed phrases will have to be defined as lexical units, so the transfer dictionary will contain e.g. «*armar un lío*» = «kick up a fuss», but the monolingual analysis and generation modules will take care of many problems which would otherwise have had to be solved by the bilingual dictionary.

Furthermore, the division of labour between analysis, transfer and generation fits the modular and decentralized project scheme. Each group may develop its own analysis and generation modules in relative independence. Only the transfer dictionaries require intensive collaboration between all the groups in the project.

3. In order to ensure the compatibility of the analysis and generation modules developed in the language groups, the framework defines a common representational structure which is the output of analysis and the input to generation: the Interface Structure (IS). The IS is a deep syntactic

or semantic representation with some primitive elements (the lexical units) and some structure which contains information about the semantic relations that hold between the primitive elements.

However, given that the input to the translation system (the source text) contains graphemic, lexical, morphological and syntatic types of information, and that all these types must be taken into consideration by the computation of the semantic representation, this computation has been divided into a series of steps starting with the source text and ending up with the IS. Between the source text and the IS a graphemic, a morphological, a configurational and a relational representation have been inserted in order to split up the translational relation between the source text and the target text into smaller subrelations between representations which correspond to traditional linguistic levels of analysis. Splitting up the overall translational relation into small steps and assuming that the principle of compositionality holds for all the subrelations makes it much easier to describe the relation.

Moreover, it provides a new division of labour which contributes to the modularity of the system: Morphology is given a specific place in the system which is different from that of surface syntax, etc. In consequence it is much easier to trace errors and to repair and extend the system.

VI. CONCLUSION

EUROTRA is modular and decentralized not only from the organizational but also from the scientific point of view. Developing a multilingual, modular, extensible and repairable system in a decentralized project environment requires a strong scientific framework and it is extremely important that the participating groups work according to the principles of this framework. It is an achievement in itself that such a framework has been provided for EUROTRA, and that the project has developed an organizational structure which makes it possible to implement the principles of the framework. The establisment of collaborating groups of linguists and computer scientists in nine out of twelve Member States under difficult budgetary conditions and the ensuing promotion of Computational Linguistics in Europe also counts as an important result of the first 3 ½ years of the programme period, and the development of the prototype MT system which is the primary objective of EUROTRA will not only be a significant step towards making MT a natural part of every translation service, it will be a rich source of inspiration and experience for Applied Computational Linguistics in general.

VII. Reading list

Multilingua (Special Issue on EUROTRA) 5-3/1986.
Mouton de Gruyter, Amsterdam, Berlin, New York.

Arnold D. J.: «EUROTRA: A European perspective on MT.»
IEEE Proceedings on Natural Language Processing. 1986.

Johnson R., King M., des Tombe L.: «EUROTRA: A multilingual system under development.»
Computational Linguistics 11, 155-169. 1985.

Somers H.: «Le projet EUROTRA de la Commission des Communautés Européennes.» *La Traduction Automatique: mythe ou réalité* (Actes ITA Premier Colloque, novembre 1985, 17-21). Cergy: ENSEALANGUES / Paris: Gachot.

Des Tombe L., Arnold D., Jaspaert L., Johnson R., Krauwer S., Rosner M., Varile N. and Warwick S.: «A preliminary linguistic framework for EUROTRA.» *In Proceedings of the Conference on Theroretical and Methodological Issues in Machine Translation of Natural Languages.* S. Nirenburg (ed), 283-288. Hamilton NY: Colgate University.

(No author): SYSTRAN et EUROTRA: la traduction automatique à la Commission des Communautés Européennes.» *Contrastes*, hors série A4 «Traduction automatique — aspects européens», 11-42.