I. AN ACTUAL LARGE SCALE PROJECT OF MT

THE EUROPEAN COMMUNITY

Bernard Vauquois

Groupe d'Etudes pour la Traduction Automatique Université Grenoble France

With 6 languages and an expectation up to 9 languages, the Institutions of the European Community have to face an enormous need of translation.

As soon as 1971, D. Berbille, responsible for translations into French at Brussels, was conscious of the predictable gap between the demand for translation and the available means to supply this demand. The general translation services at ECC are employing more than 1,700 persons ; moreover, the other services translate by themselves a part of what they need.

D. Berbille made over 1972 and 1973 a large enquiry about the existing means of all sorts of computerized tools for translation (from simplest automatic dictionaries to fully automatic translation).

In 1976, an experimental step has been started with SYSTRAN for English -> French translation about agriculture and food litterature.

In 1979, this experiment is growing and new pairs of languages are investigated with German and Italian, using SYSTRAN.

In order to find a substitution product, more adequate to multi-lingual translation, the ECC has started a european project in 1978.

The following pages are taken out of the project description of "EUROTRA". It is clearly oriented towards an advanced second generation system

2. Objectives

The aim of the project is to develop an operational multi-lingual machine-aided translation system. The system will be capable of dealing with all present Community languages and of extension to other languages.

In the five-year time span of the project proposed here, it is pro-bably unrealistic to imagine that all pairs of languages currently in use within the Community can be adequately incorporated into the system. However the maximum number of pairs possible should be imple-mented in order to demonstrate the adequacy and generalisability of the system.

Similarly, although the domains of discourse dealt with will still be limited, these restrictions will be the result of the size and cove-rage of the dictionaries, and not inherent in the system itself. New domains may be added by extending the dictionaries.

The design of the system is such that the resultant quality of trans-lation will be substantially higher than that of presently available commercial systems. Development of the system will be done by specia-list groups based in the countries contributing to the project, each group dealing with the language of its own country. Integration of the separate parts of the system into a whole will be carried out by an independent group responsible for the co-ordination, integration and evaluation of system development. Collaboration between contribu-ting groups is an important basic principle of the project, designed to stimulate the growth, throughout Europe, of expertise in text and language processing systems, an area essential to European interests.

The machine-aided translation system proposed will be considered within a larger framework of integrated sub-systems dealing with various aspects of text and language processing in general. It will, for example, be possible to use parts of the translation system as a tool in research and development in information science , quantitative linguistics and lexicography.

計算言語学 18－1 (1979. 7. 11)

At the same time, tools such as dictionaries or term-banks developed primarily for use within other projects (e.g. EURODICAUTOM) should, wherever possible, be used as source information for the translation system.
Development costs should be low enough to ensure cost-effectiveness of the system as a whole, taking into account possible savings in translation costs once the system is operational and potential sales of sub-parts of the system for use by external commercial interests. The project's main objective, therefore, is to produce a system which

a) provides higher quality machine translation on a practical basis
b) is multi-lingual, in that it is capable of translating in parallel from one source language to several target languages
c) is extensible to new language pairs
d) is extensible to new subject areas
e) can be continually improved
f) can be managed co-operatively
g) constitutes a teaching and research tool
h) can be integrated into a suite of text and language processing systems
i) is operational within five years from the starting date of the project.

3. Motivation

The need for good quality machine-aided translation systems to deal with the bulk of routine work has become ever greater in recent years. Within the European Community, multi-lingualism is a basic principle. This means that documents must be translated at present into six languages (between thirty language pairs) and the number of languages will increase as the Community increases. Each new language will dramatically increase the number of language pairs to be dealt with : seven languages involves forty-two language pairs, eight fifty-six, nine seventy-two. Thus an anticipated in--crease in the size of the Community has direct effects on the gra--vity of the translation problem.
It is already difficult to find a sufficient number of adequately qualified translators. Furthermore, even if enough translators were available, the cost of human translation is high and increasing. If a machine can take over some of the routine work, translation costs will decrease. Clearly then, it makes both economic and poli--tical sense to develop machine translation.
That machine-aided translation is both feasible and practicable has already been demonstrated by the existence of commercially available systems. (Appendix 2 gives a survey of present systems.) Indeed, the Community has already acknowledged this point in acquiring one such system.
Unfortunately, however, currently available systems are not really suited to the special needs of the Community. They are essentially bi-lingual, being designed initially to deal with a specific pair of languages. Nor are they multi-lingual in the sense that transla--tion from a single source language to several target languages simultaneously is possible. Similarly, current systems are not designed to allow for the continual integration into the system of new linguistic models, even if such models can demonstrably affect the quality of translation produced. At best, current systems allow only the solution of specific problems susceptible to treatment within the theoretical framework irrevocably embedded within the system. A Community committed to multi-lingualism must, in simple realism, accept that any system adequate to its needs must, from the

time of its initial design, cater for an unlimited number of langua-
-ge pairs and allow for the incorporation of new solutions. To do
otherwise is to design a system which has its own built-in obso-
-lescence : hence the need to develop a new system, rather than to
develop an existing system.
An adequate system must then be multi-lingual,extensible and capable
of capitalizing on new research results in linguistic theory as they
become available. This implies further that the system design must
cater for the use of different linguistic theories. In the years sin-
-ce most of to-day's commercially available systems were designed,
research in linguistics has made enormous progress, especially in
the field of formalisation of linguistic theories. The theoretical
bases currently available for machine translation systems are already
much superior to those available ten or fifteen years ago, and are
often expressed in ways which then had not even been imagined. It is
this rapid improvement in linguistic theory which predicts a corres-
-ponding improvement in the quality of translation produced. Yet
there is no reason to believe that progress in linguistic theory has
come to a halt : indeed, one of the motivations behind this project
is to stimulate further progress. Thus, an adequate system must not
only allow for the use of to-day's advanced theories, it must leave
open the possibility of incorporating to-morrow's.
None of the preceding paragraph is meant to imply that experience
with existing systems is of no value in the design and implementa-
-tion of a new system. On the contrary, it is invaluable.
At the level of software design too, experience with existing systems
has taught us that a new system must be designed right from the start
to be portable and generalisable.
None of the systems currently available commercially has been deve-
-loped within Europe. Given the multi-lingual nature of the European
Community, it is clear that the development of European know-how is
an urgent necessity, as is the training of specialists within this
area. This project is intended to consolidate the expertise which
already exists, and to stimulate the training of specialists.
That expertise does already exist cannot be denied. A great deal of
work has been done within Europe on automatic translation systems,

chiefly aimed at developing the theoretical base necessary for higher
quality translation and at testing the theories developed within (so-
-metimes large) pilot systems. The lack of a European commercial sys-
-tem is explained partly by the paucity of funding for development
work in this area in recent years, which has not however prevented
the development of theoretical work providing a sound basis for newer
and better systems.
Thus, today, a store of expertise is available which should not be
neglected when a more advanced system is to be designed. The system
proposed here, therefore, does not propose new research, but an ap-
-plication of research which already exists. It should not be thought
however, that such a project will block new diversified research. On
the contrary, it should produce as spin-off a great deal of new work,
which in turn will produce valuable results to be incorporated into
the system.
At the moment, European expertise is scattered through the Member
Countries, with different groups tending to concentrate on different
aspects of the translation process. The co-operative nature of the
project described here will bring this scattered expertise together
in a collaborative effort, thus at the same time consolidating and
diffusing the specialist knowledge already available and providing
a sound foundation for the development of an advanced system.
The co-operative nature of the project is also reflected in the fun-
-ding mechanisms proposed, whereby the Community and the Member Coun-
-tries will become partners in the development of the system.
The particular design adopted for the system is also influenced by
the desire to provide the end-user with output tailored to his needs.
Not only will the system be able to deal with all Community languages,
but the user may, by combining the various elements of the system in
different ways, obtain the level of translation he considers ade-
-quate for his particular needs. Similarly, the user may use parts
of the system for purposes quite other than machine translation.
An increase in the number of specialists trained in text and langua-
-ge processing systems can only be achieved by continuing research
on translation

and allied areas. Similarly, continued advances in such systems at the technical level can only come from continued research. This project will stimulate such research whilst at the same time providing specialist tools with which to carry it out. The system will be made available to all interested universities and research institutes within the European Community. Use of the system as an instrument for teaching and as a tool in research can be expected to stimulate work in several different areas. First, there will be a direct effect on both research in linguistics and on the teaching of linguistics. The system will be an extremely useful tool for the development and teaching of grammar, of the theory of syntax and of semantics. Thu it will make a direct contribution to the development of fundamental research vital to the improvement of the theoretical basis of later versions of the system. Secondly, there will clearly be a direct effect on the development of software and hardware specially adapted to text-processing. Then, work on data base systems designed for use in text-processing should benefit, as should all related text-processing areas, such as indexing and abstracting.

The main motivations of the project are therefore :

a) the continuing need for higher quality machine translation,

b) the great increase in demand for translation,

c) the multi-lingual nature of the European Community, and the probability that the Community will grow to include even more language

d) the fact that computer-aided translation has been shown to be practicable,

e) the existence within Europe of the expertise necessary for the development of a system substantially better than any currently available,

f) the beneficial effect on European know-how of consolidating the expertise available whilst simultaneously providing an invaluable tool for the training of future specialists,

g) the benefits to research on text and language processing springing from free availability of a specially designed system.

4. History of Project

On December 23rd, 1976, in view of the pressing need to overcome linguistic barriers within the European Community, the Commission approved an action plan for the improvement of the transfer of information between European languages. The action plan covers the fields of terminological data banks, multi-lingual thesauri and machine-aided translation systems.

The Committee of Experts on the Transfer of Information between European Languages (CETIL) was established to advise the Commission on the implementation of the action plan. It was this Committee which, at its meeting on the 19th September 1977, asked the Commission to organize as soon as possible a meeting of a working group to prepare an outline of a possible European system for machine aided translation.

The Commission therefore invited representatives of thirteen institutions and groups active within Europe on work connected with machine translation to a meeting to co-ordinate work in this area. This meeting was held in Luxembourg on the 15th February 1978. After each group had described briefly its own work and the resources at its disposition, general discussion revealed a consensus of opinion on the fact that not only were there sufficient intellectual resources in Europe for the development of an advanced European system of machine translation, but that all the groups represented were prepared to collaborate in the design and implementation of such a system.

Further discussion established that the most reasonable basis for a European system would be the creation of a new system based on the present state of the art reflected in the pilot systems already operational within Europe. The creation of such a system would be an industrial project in the sense that it should be an operational system, and should not involve undertaking new fundamental diversified research.

The final result of this first meeting was the establishment of a small working group, composed of representatives of the major European groups working in the area, whose task was to draw up a more detailed specification of how a new system might be developed.
Since February 13th the small working group has met a number of times, and has also had the benefit of advice from larger groups of specialists in specific aspects of machine translation systems, such as software support and semantics. Similarly, the Commission team working on the development of SYSTRAN and the members of the Franklin Institute have been in constant contact with the working group and have provided invaluable advice and assistance. The work of the working group has been much facilitated by a remarkable spirit of co-operation and enthusiasm which augurs well for any future collaborative project.
The working group now feels able to present an outline specification of a project to develop a multi-lingual advanced European system for machine translation for the approval of the Community. The present document constitutes that specification.

5. Resumé of System

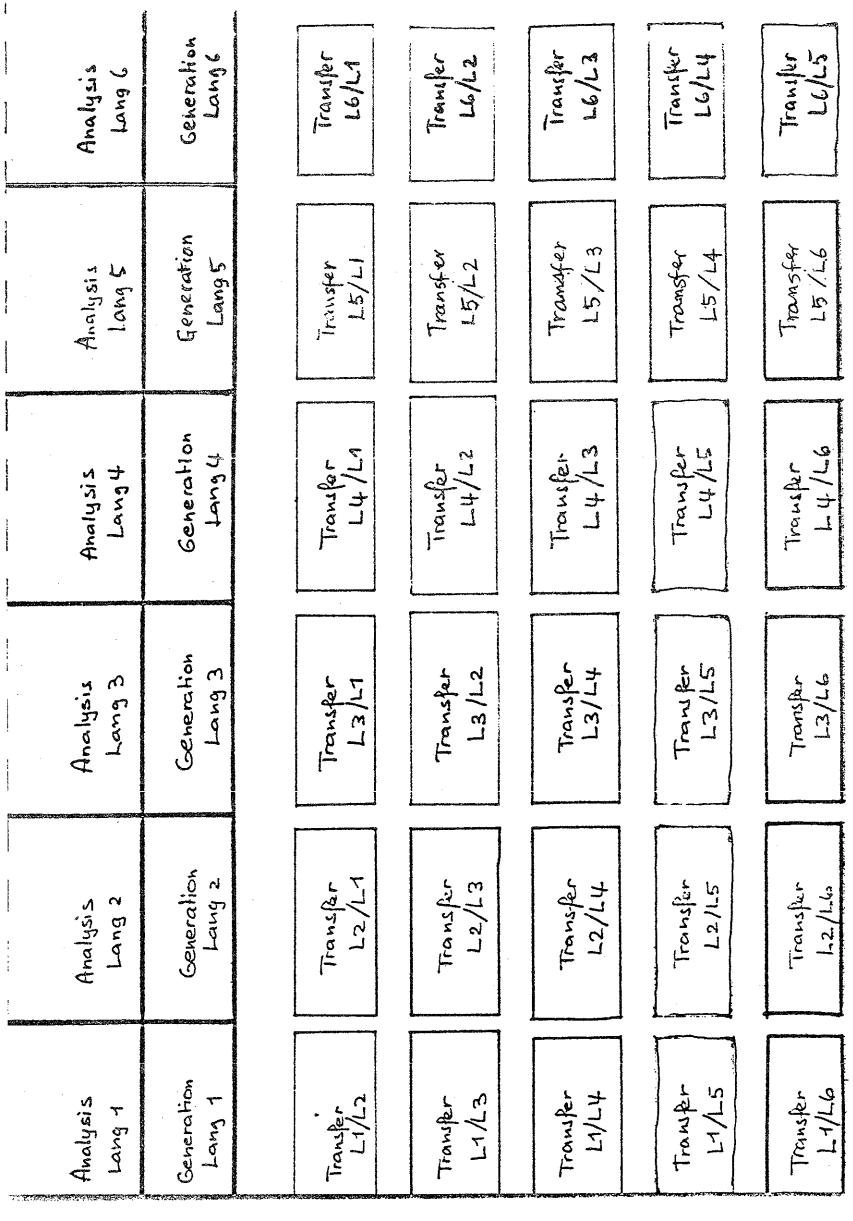
The proposed system will meet a number of criteria. It will be multi-lingual from its conception, provide higher quality translation than existing commercial systems, be extensible to new language pairs, portable between machines, and highly modular to allow for flexible use.
The translation process itself breaks down into three main phases: analysis, transfer and generation. Throughout all three main phases linguistic and computational aspects of the system are kept conceptually and practically distinct, in order to allow for improvement and development of the linguistic aspects of the system whilst retaining the software framework as a basic tool. In this way new language pairs can be added with minimum perturbation of the system, as well as the performance of the pairs already dealt with continually refined.
The attached diagram (p12a) shows how work on the three main phases is split up between the contributing centres. Each contributing centre is responsible for the analysis and generation of its own language, and will develop the transfer components towards other languages in collaboration with the group whose language is the target language involved.
The whole system makes use of one unified data structure. Particular cases of the unified data structure can be used for particular levels of the system, but will be expressed throughout in the same formalism. Thus subset can be used to provide the formal expression of the output from the morphology subsection of the analysis module, for example, whilst a larger subset is used to describe the interface between analysis and transfer. In this way the compatibility between contributing groups is assured and interchange of results at the defined interfaces presents no difficulty. Even before development of the host software for the system is completed, contributing groups will be able to exchange results in this data structure.
The analysis phase as a whole aims at the production of an intermediate representation of the source language text input, in the form of a set of trees with multiple levels of labelling, to account for the morphological, syntactic and logico-semantic information determined during the analysis phase. Additional

labels may provide tactical information to be used during transfer.
The transfer stage takes as input the labelled trees and performs, basically, two kinds of operation upon them. The lexical items of the source language are transformed into the lexical items of the target language, and the structural form of the source text is transformed into a structural form appropriate to the target language. The output from transfer is then another set of labelled trees, again with multiple levels of labelling, but representing this time the underlying structure of the target text, i.e. the translation.
The generation phase takes this underlying structure and generates from it the final output, to give a translated version of the source text.
All parts of the system have access to dictionaries. During analysis and generation the dictionaries are mono-lingual, during transfer they are bi-lingual. The dictionnaries used will contain morphological, syntactic, semantic and statistical information.
The option to use semantic information will be open from the beginning of the analysis phase, and throughout the transfer phase.
No rigid techniques of analysis, transfer or generation will be imposed on the contributing groups. Each group will be left free to develop its own techniques within the broad framework of the software tools provided, with the sole, but stringent, constraint, that the specification of the interface structures must always be scrupulously observed.
The software support provided will include provision of a specially designed very high-level language for the writing of grammar rules, dictionaries and semantic information. A lower level language will include special facilities for character and string handling and for data base manipulation and access for use with the dictionaries. In addition program packages will be provided which allow the user who so wishes to concentrate entirely on the development of grammars according to pre-defined rules, whilst relieving him of the necessity of worrying about the interpretation of the

grammars. It is not however forbidden for a user who wishes to work at a level nearer the machine to write his modules in the medium level language provided and thus to be independent of the program packages supplied.
The system is to be conceived of as forming one part of an integrated text-processing system. Thus particular modules may be used for purposes other than machine translation, and input to the system may come from sources other than direct requests for translation.
The operational system will provide for user-defined level of translation and for user-defined connection of modules. This means that the user will be able to decide for himself what level of translation he requires, or to use particular modules for his own purposes. It will also allow for both inter-active on-line use and for batch processing.

Implementation of System

The implementation of the system is to be carried out in accordance with the scheme shown in the attached diagram. The work to be done by any particular group is shown by consistently using one colour for that group. Where the specification of work to be done is shown in one colour but is enclosed in a rectangle of another colour, then the work is to be done in collaboration by the two groups whose colours are used.
This particular division of work has been chosen to allow maximum flexibility in the linguistic methods used whilst still ensuring the coherence of the system as a whole. It also allows us to profit as much as possible from the expertise currently available in the contributing countries , as well as offering a solid basis for comparison amongst the methods used, and, by ensuring comparability, contributes to the consolidation and development of European expertise in the field. Insistence on simultaneous collaboration and freedom of strategy to be chosen can only benefit the present and future systems, whilst stimulating, through the provision of tools for teaching and research, the growth of European competence in this area.
The assembly and integration of the system is to be carried out by an independent group, who will also be responsible for preparing evaluation tests and later for exploitation of the



6. Co-operation between groups.

For each of the Community languages, it is envisaged that there will be one or more specialist groups collaborating in the project. The internal organisation of the contributing groups will be left to the countries concerned, but one member of each contributing group will represent his group within the central organisational structure.

It is proposed that each contributing group works on the analysis of source text written in its own language and on the generation of the translation into its own language from the intermediate representation produced by the transfer stage. The transfer modules will be implemented by a joint team made up from members of the two teams working on the languages involved in the transfer.

Other possible alternatives have been discussed, but rejected.

First the possibility of one group taking responsability for morphology, another for syntax, another for semantics, another for transfer, another for generation and so on was considered.

At first sight this possibility seems tempting, if only because there exist within our pool of expertise groups which are precisely experts on morphology, on syntactic analysis, on semantics etc.

Nonetheless such a division of labour seems impractical for several reasons: each group, would, by implication, have to have at its disposition experts in all the languages to be treated, and using consultant native speakers as informants is both clumsy and less reliable than being a native speaker. Secondly, such a horizontal division of labour makes final integration of the system considerably more difficult. Thirdly, scheduling becomes much more critical, in the sense that delay or failure in a group working on an earlier module of the system blocks all later modules. The later modules may be developed independently, and, given the rigid definition of interfaces, to some extent tested independently. But no working system is possible without <u>all</u> modules being present.

Then, the tasks to be shared out on this proposal are very unbalanced relative to one another. Syntax analysis, for example, is a much heavier task than morphology. Finally, this suggestion would

automatically imply rigid divisions between the different tasks, even though technical considerations make it preferable to avoid rigid di-
-visions between, for example, morphology and semantics.
Another possibility would be analogous to the solution adopted, ex-
-cept that each group would take responsability for the analysis and generation of a language other than its own, transfer again being done by a joint team. The only advantage of this suggestion would come from acceptance of the argument that non-native speakers are more aware of subtleties and difficulties in a language than native speakers. This is not the place to discuss the validity of this ar-
-gument, since this proposal is open to the same practical objections as the previous proposal, insofar as it involves, once again, each group having constant access to at least one specialist consultant linguist in the language being treated, and relying very heavily on him, a method which would quickly prove unworkeable.
Finally the possibility of creating one single centralised group responsible for the whole project was considered. Whilst this might, undoubtedly, prove the most efficient solution, there is a number of strong arguments against it. It does not appear feasible that a cen-
-tralized project where everything is done by Commission staff could be launched. Along with other difficulties which one could foresee this would generate an unhealthy sense of competition with the national centers, in particular for human resources, which would be especially damaging given that the national centres have contributed heavily to the design of the project.
For all these reasons, then, the best solution seems to be that ini-
-tially proposed, where each group takes responsibility for those parts of the system which involve its own language, and those parts which involve two languages are implemented by joint teams.
The division of labour described above touches only the linguistic parts of the system, which are particular to a specific language. The construction of dictionaries and of software must be considered separately.
The software is common to the whole system. All groups will use the same basic software. Its design and implementation is therefore cri-
-tical to the success of the project as a whole.

Also, the final putting together of the operational system will have to be done with the specially designed software. Thus the time factor is largely dependent on how fast adequate software can be developed. At the same time, although the drafting of software specifications is necessarily a co-operative effort, since the same software will be used by all groups, the actual implementation of software cannot sensibly be done co-operatively, and therefore cannot be integrated into the sche-
-me set out above.
It is proposed therefore that software should first be implemented on one machine, taking care to ensure its later transportability, and afterwards transferred to all machines used by contributing groups. (Note that transportation of the system will involve the re-programming only of the monitors, which are the only machine dependent parts of the system. No re-programming of the system itself is invoved).
As soon as official approval for the project has been obtained ten-
-ders should be invited for the implementation of software, and prio-
-rity given to negotiating a contract as soon as possible.
Which machine is used as host machine for the initial software deve-
-lopment will depend on with whom the contract is made. Offers should be invited both from private contractors and from any interested uni-
-versity groups. It should be noted however that university groups would be expected to work under the same conditions as those demanded of private software houses.
In inviting tenders, it would be advisable to ask if the contractor would be prepared also to contract for the maintenance of the soft-
-ware.
Like software, dictionaries pose a special problem in that little can be done without at least a test dictionary and that the main work of constructing the dictionaries is somewhat independent of the rest of the work.
In the initial stages it will be important that all groups use a dic-
-tionary based on the same corpus, in order that testing and evaluatio of the system components can be standardized. But it is not necessary that this initial test dictionary be very large.
It is therefore proposed that as soon as the project has been accepte the Commission should be asked to supply a test corpus of texts, with a specific limited domain, and that a common test dictionary based on this corpus should be constructed.