[*Machine Translation Summit*, September 17-19, 1987, Hakone Prince Hotel, Japan]

EUROTRA
The EEC R&D programme for the creation of a machine translation system of advanced design

Sergei PERSCHKE, Commission of the European Communities, Luxembourg

1. Introduction

The European Community has twelve member States and nine official and working languages. In addition to the historical, cultural and legal differences which are an obstacle to the achievement of the political objectives of the Community, such as, for instance the creation of a unified internal market by 1992, the languages are not just another obstacle to be overcome, but they constitute an additional cost factor which has a negative influence on the competitivity of the EC as a whole.

Politically, the institutions of the Community are forced to communicate with the institutions and citizens of the members States in their national languages. To do this they have had to create the largest translation and interpretation services in the world, at a considerable cost. There is no point in attempting to reduce the number of languages. It would be the end of the Community. Economically, to reach the customers, the producers of goods and services must speak the language of the consumer. If it is true that the future of the industrialised countries lies in the development of new technologies, the products become increasingly complex, and with the complexity the volume of the associated documentation grows and becomes a considerable cost factor. The documentation must be translated for the product to reach the customer. In spite of the considerable efforts in the area of foreign language teaching in Europe, only a small minority (mostly in the international environment) is able to work in a foreign language without it constituting a handicap. Thus, translation and interpretation is a necessity, and in order to reduce its cost which is enormous, we must rationalize, i.e. automatize. It ought to be remembered that many of the problems encountered in machine translation are identical with those encountered in other applications of natural language processing. Therefore the efforts put into this field will also have an effect in other areas.

2. Historical background

The Community became interested in machine translation in the early sixties, when it started information science research in the context of its Joint Research Centre (JRC). In this period it sponsored among others things the Grenoble project and a number of other European projects. The most significant event of this early period was a participation in the final stage of the Georgetown system. This was further developed in Ispra and was used until 1975 quite successfully for the translation of Russian scientific articles into English and for a current awareness service (translation of the tables of contents of incoming Russian periodicals). However, attempts to start a new R&D project in the JRC framework did not succeed due to the general lack of interest in this domain in Europe (in part due to the ALPAC report).

The situation changed in the mid-seventies when the Community started action programmes in the field of information and documentation and multilingualism, in which particular attention was paid to the question of language barriers, and a number of initiatives were taken :

- development of multilingual thesauri to facilitate access to data bases
- support to the TITUS system (ITF, France)
- development of the multilingual terminological data bank Eurodicautom
- acquisition of user rights for Systran and the development of a number of language pairs (initially F-E, E-F, E-I, and recently a few more).

The use of Systran was intended principally for the internal needs of the Community language services, for the IDST community and for the public sector in the EC member States.

The involvement of the European Community in Machine Translation certainly stimulated the interest both in the public and in the private sector, which eventually led to some kind of renaissance which helped to overcome the disastrous consequences of the ALPAC report. But it showed also the lamentable state of research both in Europe and elsewhere : Systran was basically built on the technology developed by the Georgetown project (Peter Toma had been one of the most prominent figures in Georgetown), and there was no sign of a successor technology for the next generation of machine translation (and natural language processing systems in general), nor was there much professional skill around. For this reason, the Community started to explore the possibility of a European machine translation project which should prepare the future. With the active contribution of the most prestigious figures in this field in Europe (Vauquois, Masterman, Eggers, Zampolli, Wilks, King, Sager and many more), the understanding was reached that a new machine translation project was desirable and feasible, that the Community should take an active role in fostering European cooperation and in ensuring the equal treatment of all European languages.

Under the Chairmanship of Margaret King, the Eurotra coordination group was created in 1978. This group included participants from the competent university institutes of all member States. The Group assisted the Commission in the preparation of a programme proposal which was submitted to Council in June 1980 and which was finally adopted in November 1982.

The duration of the programme was set at five and a half years, and the Community budget was to be 16 million ECU. This was to be complemented with approximately 10 million ECU of direct national contributions as Eurotra is a cost sharing programme.

The programme is subdivided in three phases :

Phase 1 : Preparatory phase (2 years, 2 million ECU) :
preparatory phase intended for the project definition and the setting-up of the organizational structure

Phase 2 : (2 years, 8.5 million ECU) Phase of basic and applied linguistic research, with the objective of building a small corpus-based prototype system with a vocabulary of 2.500 entries.

Phase 3 : (18 months, 5.5 million ECU) Phase of consolidation and assessment of the results, in which the prototype system is to be extended to cover a limited subject field with a vocabulary of 20.000 entries.

Following the accession of Spain and Portugal, which entailed the inclusion of two additional languages in the project, the second phase was extended to 3 years with a budget of 13 million ECU and the third phase was extended to 2 years, to accommodate the additional work load, and some of the delays in the start-up.

The overall objective of the project is the creation of a relatively small prototype system, operational in a limited subject field and for a limited number of text types, which can serve as a basis for an industrial development. The Commission is expected to submit a proposal to Council for this industrial development project before the end of the third phase of the programme.

3. State of the Project

3.1. Organization

Eurotra is a highly decentralized, cooperative venture.

The management and coordination of the project is the responsibility of the Commission which has created a project team charged with :

- management and administration
- planning
- linguistic and software specifications
- monitoring of decentralized work
- testing and integration of the results
- internal assessment

The team has now 11 staff and is being extended to a total a 20. For particular tasks both in the relation to specifications and to implementation work of general interest it can draw on the resources of the national groups and on external experts.

The national groups are in principle responsible for the analysis and synthesis of their own language and for the transfer from the other languages into their own language. Approximately one half of the resources of each group are allocated to the implementation of the various prototype systems in accordance with the general programme of work. The rest of the resources are devoted to basic and applied linguistic research. This can be language-specific or of general interest, as an extension of the central project team. The average size of each team is around 15.

Since the linguistic map of Europe is somewhat complex, the project has to accommodate a number of special cases :

- Belgium shares with the Netherlands and France the work on the Dutch and French languages. In addition it has a special task assigned to it in the field of computerized lexicography;
- Ireland is responsible for the definition of a general approach to terminology for Eurotra.
- Luxembourg fulfils the function of a clearing house for documentation and as a test and reference center for the software.

The permanent assessment and planning of the project is done collectively by the directors of the various teams who together form the "Eurotra Liaison Group". This group, which has decision-making powers, meets approximately once a month. Extensive use is made of electronic mail and conferencing services to maintain communication between the different centres participating in the project.

3.2. Progress of the project

In an ideal world, the organization described above ought to have existed before the actual research work started. However, Eurotra is the first Community project in the field of natural language processing, and all of the organizational structure had to be created virtually ex nihilo. The speed with which this happened depended very much on the political commitment of the member States, the efficiency of the national administrations, and the state of research in the field. There is a time lag of two and a half years between the first and the last country joining the project - and as a consequence, there are quite considerable differences in the state of advancement of work on the various languages. Late starters and late comers (especially Spain and Portugal who joined the EC in 1986) hope to catch up with the most advanced groups by the end of the programme.

During the second phase the implementation work by the language groups has been subdivided in two cycles. The first cycle which ended in February 1987 had as its objective the creation of a very small prototype system with a limited linguistic coverage and a vocabulary of some 500 entries. The full goal was achieved by three groups (D, DK, E) while the other groups obtained only partial results (and are completing the work). In the second cycle which will end in mid-1988 the prototype is being extended in its coverage and vocabulary which will increase to 2.500 entries).

In parallel with the implementation work research is going on to improve the linguistic specifications, especially the interface structure, the abstract translation machine, and the basic software instantiating it.

The preparation of the third phase has also started. Both the goals and the method to reach them are under examination. At the request of the Council of Ministers of the EC and of the European Parliament who discussed Eurotra at length when the extension of the project to Spain and Portugal was deliberated, an assessment of Eurotra by a Panel of independent experts is being carried out. In particular this is to determine the degree of success of the programme to date and to make recommendations for both the third phase and for the post-Eurotra industrial development. The findings of the Panel are expected to be available in October 1987.

4. General system description

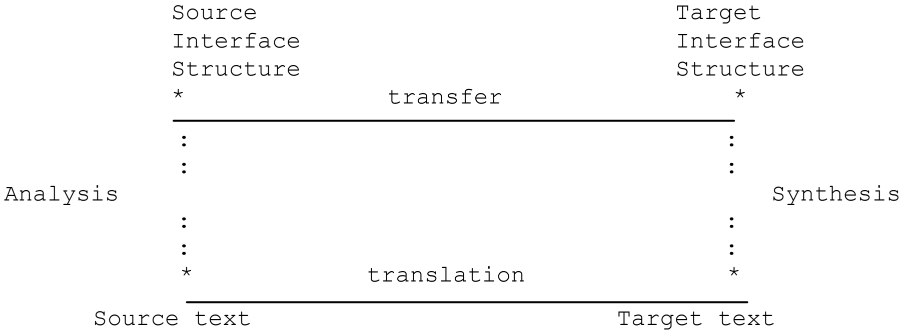
Given the decentralized, cooperative organization of the project, which is also likely to be continued in the industrial system development plan, there exists more demand for an explicit declaration of the theoretical foundations of the system design than in other, more centralized projects.

In Eurotra, this theoretical framework serves two additional purposes : it allows for a more problem-oriented software design, and serves as a guideline for the linguistic specifications of the system.

4.1. The abstract system model

The embryo of the general system design was contained in a very early agreement that Eurotra should be a transfer-based system. The multilinguality of the project, and the geometric progression of the language pairs (n * n-1) made it clear that transfer should be minimized (asymptotically towards naught as a long-term research objective) to make it manageable.

This decision lead to the breakdown of the translation process which can be seen as a complex relation between a source and a target text, into three elements : analysis, transfer, synthesis.



The relationship "translation" is too complex to be given any formal description, which could be expressed in computational terms, and so are "analysis" and "synthesis", but the desire, and need, to simplify the "transfer" to the bare minimum led as a side effect to a relation which could be easily described in formal terms, and which in broad terms could be summarized as compositionality. Thus, we get a pair of representations (the source and the target interface structures) and a simple relationship which characterizes the differences between them.

For the representations we define a formal description device, a generator (based on unification), with an associated user language and computational apparatus, and for the transfer one assumes that the identity relationship holds between source and target objects unless a different relationship is stated explicitly.

The transfer model was then extended to both analysis and synthesis : the relationship between text and interface structure was broken down into a sequence of relationships between intermediate representations for which the principle of compositionality holds. The picture, then looks as follows :

analysis   synthesis
: --------------------------:   : -------------------------- :
source — R1 — R2 — R3 —.... IS_s transfer IS_t...R3 — R2 — R1 — target
text   text

It is up to the linguists to define the right form and number (possibly low) of intermediate representations, and the mappings between them. In addition it is desirable that each representation capture some linguistically meaningful dimension of description (such as "morphology" or "syntax"). The set of intermediate representations in analysis and synthesis need not be the same (since synthesis is not simply the mirror image of analysis) but it is desirable for reasons of economy. It is certainly not possible to have exactly the same generators both for analysis and synthesis :

- one may want to exclude some phenomena from synthesis which must be accepted in analysis (notational variants, re-current errors, sloppy substandard use of language etc.)
- the cases of and conditions for disambiguation in synthesis may differ from those in analysis.

But with those two exceptions the representation and generators can be the same in analysis and synthesis.

From the system design aspect, this has a number of beneficial effects :

- it foresees modularity in the linguistic design
- it allows for a high degree of problem-orientedness and declarativity in the user language
- the existence of only two basic computational devices - generators and translators allows for a large scope for optimization of the implementation
- it serves as a rigorous framework for the linguistic specifications.

4.2. The linguistic model

4.2.1. Interface structures

Given the existential need to minimize transfer, a considerable amount of thought and work has been invested in the definition of the interface structure. To minimize transfer means to maximize the share of the interlingual elements by abstracting away from language-specific phenomena. The method chosen for Eurotra is basically a semantic model-theoretic approach to a number of linguistic phenomena which traditionally are considered as problem cases in MT. These include among others :

- semantic relations (case frame bound complements and modifiers)
- time/aspect/aktionsart
- modality
- quantification/determination/negation
- pronominalization
- scope
- focus
- emphasis etc.

The objective is to create an interface structure which is isomorphic for all languages treated, modulo the lexical component. Since the lexicon is by far the largest component of a translation system, efforts are being made to minimize the lexical transfer, too. This is achieved by factoring out from the transfer lexicon everything which can be perceived as belonging to "terminology". Theoretically, "terminology" is based on language-independent, i.e. interlingual concepts, which have denominations in each language. This approach suits Eurotra, and there is a continuous research going on to maximize the portion of the vocabulary covered which can be treated as terms (E-terms in the Eurotra jargon).

There are two issues which are being distinguished very strictly in the Eurotra design :

a) representational aspects : under this term we include all those elements of information which must be computed in analysis and which are necessary to allow for a correct synthesis after transfer. This information cannot be provided on the sole knowledge of the target language.

b) disambiguation aspects : by this term we mean all those elements of information which are not strictly necessary for the purposes of a) above, but which serve to resolve ambiguities. In fact, ambiguities, both within one language - be it in the analysis or in the synthesis scenario - and across languages, are haphazard, and the information needed to resolve them is also haphazard. It is therefore not possible to devise an a-prioristic theoretical model for the information needed to resolve the ambiguities, neither within one language, nor across languages.

Nevertheless it appears to be desirable to define a common methodology within the project for dealing with such information in a coherent and principled way. It should be noted here that the class of semantic knowledge to be incorporated into Eurotra is of the kind of "common-sense" without subject-field specific models. Thus, for example, "bottle" is described as a container for liquids, "full" as a quantifier of the content and in "bottle full of water", "water" is recognized as the content of the bottle. It is obvious that not all of the ambiguities can be resolved by applying this kind of knowledge and reasoning. Trends in research point clearly towards AI techniques as a line of continuation. It is, however, not clear today, especially in statistical terms, which classes of problems will remain unsolved, which techniques will be needed to solve them, and, most important, whether it is worthwhile to apply them in a translation system.

4.2.2. Intermediate representations

While it is an obvious objective to minimize the distance between interface structures in order to reduce the size of the transfer modules, it is desirable to maximize it between the intermediate representations within one language, in order to reduce their number. It is also desirable to associate as far as possible each of the representations with one of the traditional dimensions of linguistic description, within the constraints imposed by the "compositionality" and "one shotness" constraint of the mapping between representations.

Modulo language-specific peculiarities which may make it necessary to shift the treatment of one or other phenomenon between representations, it appears feasible to define for Eurotra a common set of intermediate representations. Since these representations become necessarily the more language-specific the closer they get to text, they can be defined centrally only in quite general terms and it is the responsibility of each language group to supply the language-specific complements to it.

Very roughly, the following representations are envisaged :

Actual text (AT) which is the unedited source/target file, in our system environment most likely an ASCII file following some coding convention.

Normalized text (ENT) which in general terms gives a structural description of the AT file, separates textual from non-textual parts and describes the textual parts in terms of words, sentences, paragraphs, titles etc.

Morphological structure (EMS) which describes the words as a complex construct over morphemes and deals with the usual phenomena such as inflexion, derivation, composition, prefixes, clitics, contractions etc.

Surface syntax (ECS) which is seen as a surface-oriented configurational type of syntactic description.

Deep syntax (ERS) which is a relational structure, abstracting away from a number of surface-oriented structures (expressing them by features), such as particles, articles, auxiliaries, valency-bound preposition, passives etc.

Interface structure (IS) which abstracts away from a number of syntax-oriented phenomena such as subjects, objects, tenses, voices, etc. (see above).

4.2.3. Dictionary

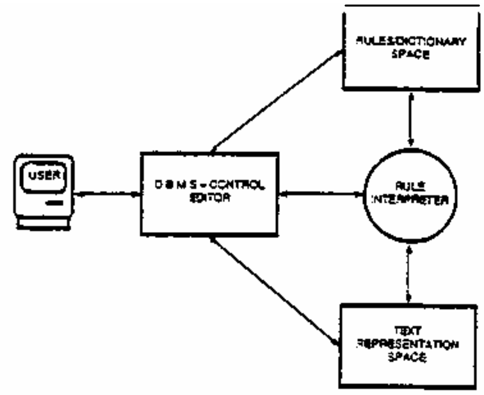
The Eurotra model creates an apparent paradox with regard to the lexicon of each language : considering a generator as a formal object describing a "language", the terminal (or "non-compositional") symbols of this language are the lexicon, and their description a dictionary. Within our model, theoretically, one ends up with as many "monolingual" dictionaries, as there are representations (six per language), and as many "monolingual transfer dictionaries" as there are couples of representations (ten per language), in addition, of course, to the bi-lingual dictionaries.
Theoretically, this picture holds, but for the linguist a view of a unique monolingual dictionary is created. In this view the representation-specific information becomes an ergonomic device by which the linguist can concentrate on one class of problems at a time. Mapping between representations is assumed to be identity as a default, and the linguist has to state only the exceptions.

4.3. The basic software for Eurotra

In Eurotra, but also in a machine translation system in general, the software has to fulfil two distinct functions whose characteristics and requirements are apparently contradictory. It has to serve as a system development and maintenance tool on the one hand, and as a translation programme on the other. During the whole of the project period the only real user of the system is the linguist, who has to fulfil two tasks :

a) create, inspect, modify etc., declarations, grammars, dictionaries etc.
b) test them against text data.

Therefore, to satisfy the two requirements the general system architecture has been conceived as follows:



As can be seen, the central part of the system development environment - a commercially available relational data base management system - has been chosen. The "user language" in which definitions, dictionaries, grammars etc. are written is defined in terms of the standard services offered by the DBMS (SQL, SQL by form, report generators etc.). Generators and translators are standard files of the DBMS, which are optimized with respect to the system development environment.
Conceptually, the rule interpreter is basically a unifier trying to apply rules/dictionary entries, which are perceived as assertions about data in the work space representing text at a given moment of processing.
While the data access and manipulation mechanisms provided by the DBMS are perfectly suitable for the needs of the system development environment, they are insufficient for any reasonable degree of efficiency at run time for the rule interpreter.

For the time being, Prolog is being used as the implementation language for the rule interpreter, and the Data base is downloaded into a form accessible to Prolog. It is foreseen for efficiency reasons to implement the critical parts of the rule interpreter in some procedural language (C). This may make it necessary to modify the design of the data base, in order to adapt it to the needs of the interpreter.

Conclusion

The project is now at the half-way stage and it is too early to make predictions on the system which will be created by 1990. It is, however, already evident now that its very existence is most beneficial in a number of respects :

- The ambitious objectives - a multilingual and extensible system - required a whole series of innovative solutions, and the approach which is envisaged will certainly contain quite a number of scientifically interesting elements
- Eurotra is the first large cross-national R&D project in the field of language processing, and the degree of cooperation already achieved surprises all observers familiar with the scene
- Cross-national cooperation may have created some initial communication and comprehension difficulties, but the involvement of different approaches, schools and cultures in approaching the same problem has certainly been most stimulating and has led to a remarkable cross-fertilization

As one of the consequences - most unusual in research in the humanities - none of the ideas and problem solutions can be attributed to a single person, because they are a result of an intensive exchange of views and debate. Errors and mis-representation are, however, the sole responsibility of the author.