E U R O T R A

Sergei PERSCHKE

Commission of the European Communities
Directorate Generale Telecommunications, Information
Industries and Innovation (DG XIII)
L -2920 Luxembourg

1. INTRODUCTION

Machine Translation is an engineering endeavour. To succeed in this endeavour, it is necessary to combine the scientific knowledge in the relevant theoretical domains (theoretical, computational, descriptive linguistics, lexicology and terminology, formal languages and computer science) into a coherent, operational system which accomplishes a specific task.

As in other endeavours, the transition from theoretical knowledge to an operational system is not achieved in a single step. It is necessary to create a technology, a set of methods, devices, tools, which in turn are used to construct the desired artefact, and this process is usually not linear, but requires several iterations.

In retrospect we can say that the most important achievement of EUROTRA is the creation of the bases for a linguistic technology which can be used to build machine translation systems and other advanced applications involving languages.

In order to appreciate the progress in linguistic technology made by EUROTRA, it would be useful to recall the state of affairs in the early eighties when the programme started.

The disenchantment with machine translation caused by the ALPAC report was not quite over. There had been no public funding of machine translation research in the USA since 1965 and both Japan and the USSR who had publicly endorsed the conclusion of the ALPAC report had just started reconsidering their position.

Canada who had sponsored research of the University of Montreal (TAUM), discontinued the project in 1980 on economic grounds (too expensive to develop a practical system) although scientifically it appeared to be promising.

The United Kingdom had discontinued sponsoring the NPL (National Physical Laboratory) project in 1967.

In Germany the SFB-100 SUSY project at the University of Saarbrücken had matured to a research prototype. But a decision to invest into a product development on its basis was never made.

In France the machine translation project of Grenoble (GETA) started in the early sixties had also produced a research prototype (Ariane-78), and discussions about a national development project which eventually lead to

CALLIOPE were still underway.

The Joint research Centre of the EC had participated in the final development of the Georgetown System and had run a fairly successful translation service since 1965 for its own and external researchers. This service was discontinued in 1975 as no resources for converting the system from the IBM 7090 to the 360 series were available.

In the same year 1975 DG XIII acquired a licence on Systran for purposes of scientific and technical information and for the use by the Commission's translation services. Since, Systran has been continously developed and extended to new language pairs.

It is against this background that the Commission submitted in June 1980 its proposal for EUROTRA to the Council after over two years of reflexion and discussions, and it took another two and a half years of discussions with Council and the European Parliament before the proposal was finally approved.

2. THE EUROTRA PROJECT

The Commission's motivation to propose EUROTRA was first of all its own need to cope with the growing volume of the translations which the Treaties of Rome impose by ruling that all legal acts of the Community must be made available in the official languages of its Member States. The successive enlargements of the Community from initially six Member States to twelve in 1986 brought the number of the official languages up to nine with 72 possible directions of translation.

The availability of an advanced MT system could certainly alleviate the pressure on the translation services by increasing their productivity, but it was also considered that it could be of more general benefit to the economic activities in the Community.

Lastly, however impressive the performance of the operational systems was, it was felt that they were not the ultimate answer to the problem and that a concerted R&D effort was needed to prepare for a next generation of MT which would incorporate the advances made in the meanwhile in linguistics and in computer science.

EUROTRA is a shared cost programme jointly financed by the Member States and the Community in the framework of contracts of association. It is a one-project programme with the active participation of research teams in all twelve Member States. It covers from the outset all official languages of the Community.

The de-centralized cooperative nature of the project has introduced a hitherto unprecedented requirement : in order to succeed, EUROTRA had to create a very strong theoretical framework and a set of common methods and tools (formalism, software, linguistic specifications) and to reach consensus of all participating parties. Much of the energy of the first two phases of the programme was devoted to the definition of this framework and to its implementation. The effort to achieve this objective may appear to be excessive (if one compares with other, highly centralized MT projects), but it produced two extremely important side effects :

- it has created in Europe a network and a community of researchers who "speak the same language";

- it has created the basis for a common linguistic technology and, maybe, has prepared the ground for a quantum leap: the transformation of linguistics into a "hard" scientific discipline.

In the following chapters a short outline of this framework is given

2.1. General System Philosophy

The only way of describing complex processes known to us consists in breaking them down into a number of discrete states and in defining the transition from one state to the next in terms of a relationship between them.

Translation is a complex process, and thus, if we want to describe it, we are forced to apply this method.

If we refer to written text, and not speech, which would add another dimension of complexity, we are actually confronted with two discrete states : the source text and the target text (which is the translation of the source text).

Considering these states, two aspects appear of interest :

a The source and the target text are written in two different languages, but there exists some kind of equivalence relation between them, which we usually describe as both texts having the same meaning. Our own approach to verifying this equivalence consists in reading and understanding the two texts and in comparing what we have understood. The criteria of comparison may widely vary, and with the change of criteria also the judgement whether something is a "good translation", an "acceptable translation" or a translation at all, may change as well.

   Since linguistics and cognitive science are far from knowing what "meaning" means, and still further from being able to give a formal description of the meaning of a given text (and from automatically deriving it from the text), we are left, and shall be for a long time, with a set of mostly undeclared, un-formalized and ill-understood subjective criteria each time we want to judge the success of any endeavour involving translation. Popular wisdom seeing translation as an art and not as a science, or coining proverbs like "traduttore-traditore" actually highlights this problem.

b Translating is a process which takes a source text in one language and produces an equivalent target text in another language (with all the reservations about the equivalence relation itself). Our own approach to translation, in very simplified terms, consists in reading and understanding the source text and in producing from this understanding a text in the target language which satisfies the equivalence criteria which are set by the translator himself (or imposed on him).

   In trying to mechanize the translation process, we meet the same difficulties as in judging a translation : our lack of scientific knowledge about meaning and understanding. In devising an automatic translation system we are therefore unable to simulate our own intellectual activity.

As so often, we have to devise a mechanism which produces an equivalent (or at least similar) result, but the internal working of this artefact in no way allows the conclusion that it is a simulation of the analogous human activity (even if it happened to be so, we couldn't know it).

EUROTRA is not the first attempt to create a machine translation system. Many projects have been started before, and some of them have led to operational systems, which in a few cases are practically usable. Thus, EUROTRA could look back on a rich experience and learn both from the successes and failures of its predecessors.

However, EUROTRA is the first project aiming at the creation of a genuinely multilingual system which adds to the complexity of the undertaking in comparison with its predecessors.

The "obvious" approach to a multilingual system is an interlingua: analysis maps the source text on an interlingual representation, and synthesis generates the equivalent target text from it. The problem with this solution is that the state-of-the-art of linguistics and cognitive science is far from allowing the definition of such an interlingua for the purpose of practical MT.

On the basis of this consideration a more realistic approach was chosen: a transfer-based system, which does not attempt a fully interlingual representation of the texts, but leaves a certain number of language-specific elements which are subject to contrastive, bi-lingual treatment. While analysis and synthesis are monolingual and are done once for each language, the number of transfer components increases geometrically (n* n-1 transfers for n languages). To make this approach at all manageable in practical terms, it is necessary to reduce the size and complexity of each transfer component as much as possible. This need has largely determined the direction and methodology of linguistic research In EUROTRA.

It consists in choosing as a starting point an interface structure which is recognized to be inadequate, but which is linguistically well-understood and assessed. This starting point, which could be characterized as a sentence-based deep syntactic representation is understood as a working hypothesis for an interface structure which should be suitable for minimizing transfer. The cases of complex transfer are candidates for research. Success is measured by the reduction of the need for transfer, and, incidentally, also by the reduction of overgeneration.

2.2. The Stratificational Approach

The transfer based system architecture adds two additional discrete states to the description of the translation process : the source and the target interface structures with a clear and simple mapping between them . However, the mapping between source text and interface structure (analysis) and interface structure and target text (synthesis) is still too complex for a formal description. In order to master these mappings, there is a need to define a certain number of additional states such that the mappings between each pair of these states become simple and formally describable.

There exists an analogy between this approach, which is dictated by technical considerations and linguistic tradition which has sub-divided linguistic

descriptions in a number of more or less autonomous dimensions, such as phonology, graphemics, morphology, surface syntax, deep syntax, semantics etc. In order to enable the project to incorporate the accumulated linguistic knowledge, it is desirable to define the various states of linguistic description In a way that they reflect as far as possible these traditional dimensions. How far this can be achieved depends on the formal properties of the devices which are used to describe the various representations and the mappings between them.

This approach leads to what is known as a stratificational system : it has been taken into consideration In the early days of MT, but it was abandoned mostly on grounds of computational efficiency. It is mostly due to overgeneration in analysis : such a system produces a large number of descriptions of a given text at the early stages of analysis, which then must be all inspected and to a large extend rejected at the later stages.

There exists no obvious answer to this problem now. One could just argue that the simplicity and tidiness of the system design justifies this built-in inefficiency, which will be in any case more than set off by the efficiency of the computers. However, in our opinion, it is not satisfactory to rely on the increasing speed of the hardware : one needs a systematic approach to the problem of the interaction of modules in a sequential system which could eventually lead to a non-sequential architecture.

2.3. Formalism and Software

The stratificational model involves two basic system components which should be appropriate for all dimensions of linguistic descriptions: a generator and a translator.

The generator is based on unification grammars, with a fairly elaborated feature theory and a basically context free parser extended by a number of features aimed at handling discontinuities, ellipsis, unbound dependencies, linear precedence etc. It accepts the output from a translator which is an "unconsolidated" tree. (Neither the dominance relation between a mother node and its daughters, nor the precedence relation between sister nodes are fixed and can be changed through the application of rules). The application of the generator rules produces the definitive representation.

Translators are tree-to-tree transducers. They apply the principle of compositionality and one-off translation, i.e. the left-hand-sides of the rules describe well-formed source structures and the right-hand-sides well-formed target structures. They rely heavily on default translation so that only the difference between source and target structures need to be recorded. The rule interpreter is written in Prolog.

In addition to the kernel system a user environment has been developed for the creation and maintenance of the dictionaries and grammars. This environment is built around a relational DBMS which is interfaced directly with the kernel system at run time.

2.4 Linguistic specifications

Linguistic research leading to specifications is concentrating mainly on the

improvement of the interface structure aiming at a better interlinguality and at the control of overgeneration. The results of this research, once agreed upon and tested are summarized in the EUROTRA reference manual and become binding for the implementation by the national teams.

For the intermediate states a number of representations have been defined which allow to link text to the interface structures. The definition of these representations, unlike the interface structure, has however the status of recommendations, since in view of the diversity of the languages treated no detailed binding specifications can be produced. In general these intermediate representations are defined as :

- AT : the actual text;
- ENT: normalized text;
- EMS: morphological structure;
- ECS: constituent structure (surface syntax);
- ERS: dependency structure (deep syntax);
- IS : interface structure.

Particular attention is given to dictionaries and terminology. In terms of the system architecture each representation has its own dictionary (which represent the terminal elements). Non-compositionality, i.e. the interpretation of a construct of one representation as an atomic object in another (or vice-versa) is treated by translators.

Terminology is of special interest in a multilingual translation system. On the one hand, the quality of translation depends crucially on it, and on the other hand the theoretical concept of terms eliminates the need for transfer and adds an element of interlinguality to a quantitatively considerable part of the system.

Terminology introduces an additional interesting aspect in as far as it represents to a certain extent the knowledge of a subject field including its paradigmatic relations which are different from those dealt with by general (common sense) semantics.

2.5. Linguistic implementations

Actual implementation of the analysis, synthesis and transfer modules is one of the main tasks of the national research teams. The goal of the project is to achieve by the end of 1990 the coverage of a limited subject field (subfields of telecommunications) with an estimated vocabulary of app. 20.000 entries of which about two thirds should be terms. At present for all languages partial implementations exist with a fairly broad grammatical and lexical coverage (5-6000 entries in average). Terminology is not yet integrated as experimental implementation and testing are not concluded. As far as transfer is concerned, coverage is uneven across the language pairs. In addition, the specific working method by which complex transfer is not immediately implemented, but used as input for the improvement of the interface structure, reduces in the present implementation the capability of

transfer with respect to analysis and synthesis. Complex transfer will be implemented systematically for those cases where research is unlikely to produce conclusive results during the lifetime of the project.

Linguistic research is planned till the end of the project in preparation of the next step towards a practical system.

3. FUTURE PLANS

Planning of research and technological development (R&TD) In the Community is intimately tied to the Framework Programme introduced as an institutional instrument by the Single Act.

EUROTRA is part of the Framework Programme and has to follow its evolution in its future plans.

At present, a two-stage scenario is determining our future activities :

(a) In the present Framework Programme, in addition to the current EUROTRA Programme, additional actions are foreseen, which are in part the immediate follow-up and in part extensions of the present programme.

They concern in particular :

- start-up of the industrial development of EUROTRA
- development of methods and tools for the reusability of lexical resources in computerized applications;
- creation of standards for lexical and terminological data.

Considering the time scales (mid-1990 to 1992) and the financial resources earmarked for the actions, the objectives of this programme are necessarily modest. We must consider it as a transition in preparation of the next Framework Programme.

(b) The Commission has started at the end of 1988 the preparation of a revision of the current Framework Programme. The discussions on the general orientations and content of the revision are not yet finished, but there appears to be a consensus in the Commission and the Member States that there should be an in-depth revision, which in practice could be the next Framework Programme.

With regard to MT and language problems in general, consultation with our advisory bodies and research and government circles in the Member States identified the need for a global policy and a strategic R&TD programme in the field of linguistic technologies, and led to a first outline of the scientific and technical objectives of such a programme.