Anatomy of Eurotra : a multi-lingual machine translation system

by

Anne DE ROECK
Projet EUROTRA - GENÈVE

1. History

The history of EUROTRA starts in 1977 in the offices of the Commission of the European communities in Luxemburg. At that time, 6 official languages (D, DK, E, F, I, NL) were used in the EEC, which meant that a very large number of Community documents had to be produced in all of those 6 languages. A small calculation tells us that 6 languages involve 6 x 5 = 30 language pairs and therefore also 30 translations. An enormous amount of work which, combined with short deadlines put (and still puts) the Commission's translation department under considerable pressure. Then there is also the financial side of the question. In 1979 for example, the cost of translation ammounted to 331.8 MUCE (± 13,272 milj BFr.) or 50.75 % of the total Budget (817.6 MUCE or ± 32,704 milj BFr.) of Parliament, Counsil, Court and Commission.

The fact that since then more countries have joined the EEC has not simplified matters : the amount of documents produced has increased as well as the number of language pairs between which translation has to happen.

In 1981, the Communities have 7 official languages. If Spain and Portugal become member states, they will have 9, which means that translation between 9 x 8 = 72 language pairs will have to be assured.

For those reasons the commission decided it might be worth while to invest in a cooperative machine translation project : EUROTRA. The important word is *cooperative*. Participation in the planning is ensured by university centres from all Community countries who try and combine their own views on Machine Translation with those of their colleagues.

2. AIMS

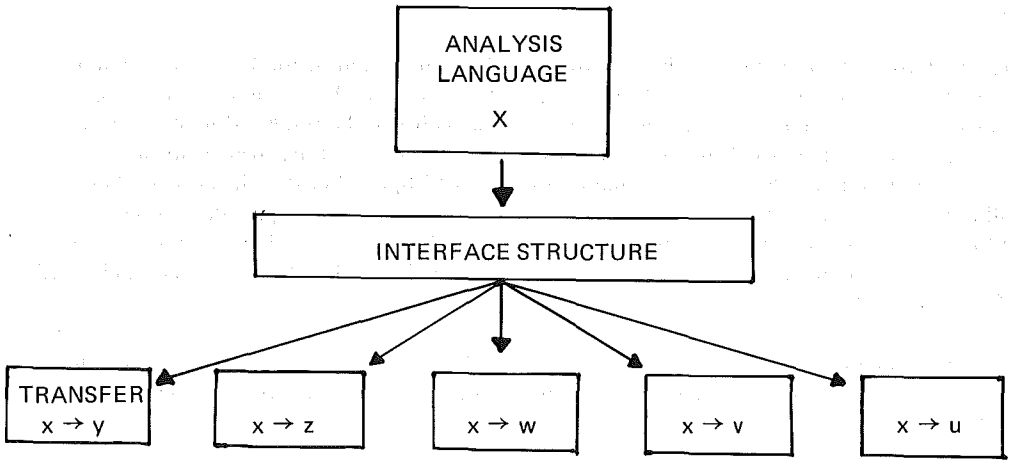
When designing a system like EUROTRA one has to know one's limitations as well as what one is aiming at. The following paragraphs describe what basic ideas underly the conception of the project.

A. Multilinguality

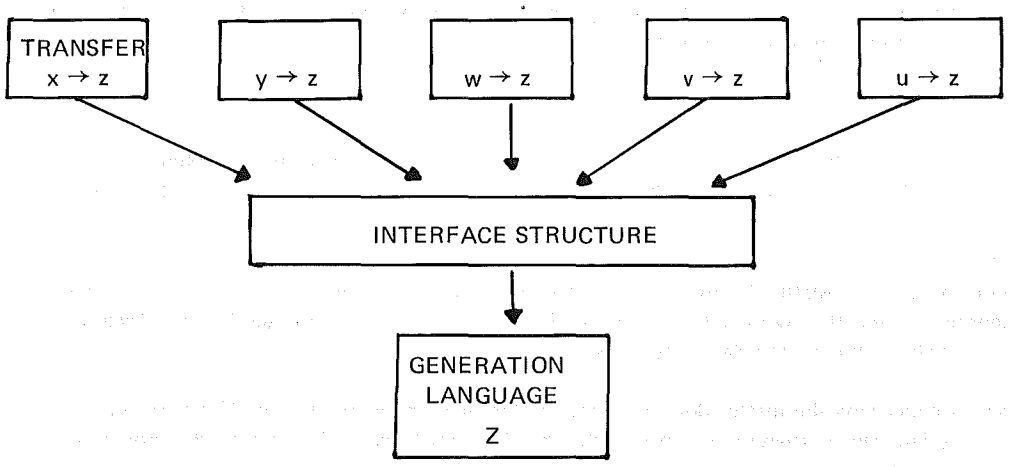
An obvious way of translating by machine is to follow the human translation model of 1 translator per language pair and thus write 1 translation model per required translation (eg Systran, Weidner .. are all essentially bi-lingual translation systems).

This way of conceiving the matter does not decrease the quantity of work (still 42 transl. for 7 lang.) : every language is analysed 6 times, each time different because the analysis depends upon the target language.

EUROTRA tries to avoid this by adapting the criterion of *multilinguality*.



Each input language corresponds to exactly one analysis during which the target language is *not* taken into consideration. The result of analysis is an interface structure that will serve as an input to one of the bilingual transfer modules.



The output from the transfer components is, again, an interface structure which will, for each language, be input to one and only are generation component.

As analysis and generation depend solely on source and target language respectively they need only be done once for each language thus avoiding going several times through similar processes, as happens in bilingual translation systems.

The only step in the whole process that depends on more than 1 language at the same time is transfer; this is mainly because current linguistic research has not yet given us an "interface" structure common to all languages. For reasons of economy the transfer phase should be kept as small as possible.

B. Collaboration

As was already stated before EUROTRA is to be a cooperative project. Research centres spread over the Community countries join efforts in designing the system and the system will be implemented in a similar manner. This set up entails a number of complications.

First of all there is decentralisation for geographical and administrative reasons. The work involving a particular language will be done by national groups in their respective countries, which requires an administrative organisation able to cope with the situation and to guarantee the necessary communication.

But that is not enough. Most of the groups already have experience in computational linguistics if not in machine translation itself and have come to conclusions on the subject depending on the research established in the centre they belong to as well as on features inherent to the language they concentrate on.

For instance, it is very unlikely that the analysis of a highly inflectional language like German will happen following the same strategies as the analysis of English which has nearly no morphology at all and therefore might require semantic information to be brought in at an earlier stage.

EUROTRA should allow each group to follow the strategies they think most efficient for handling the language they deal with. The same goes for linguistic principles that any group might wish to respect.

The EUROTRA formalism has been developed in such on way that linguistic information of all levels (morphological, syntactic, semantic . . .) can be used and interpreted in the light of different linguistic theories, without losing the multilingual aspect which is considered as equally important.

C. Development

EUROTRA is a research and development project. The reason why the Commission allowed for the project relates closely to the opinion that present results in both linguistics and computer

science make machine translation possible.

EUROTRA is based on the results of previous work, profiting as much as possible from former systems (including pilot studies) and avoiding their mistakes. This does not mean that possible new results rising from present or future research should be excluded. The aim of EUROTRA is to leave room for those developments to be incorporated whenever they come about.

Also nothing in the current policy of the Communities gives us cause to believe there will not be new member countries, and therefore new official languages which EUROTRA should be able to incorporate.

The mere size and complexity of the system require that any additions (e.g. of the two types stated above) should be done by augmentation rather than by altering what already exists. This demand has consequences for the whole conception of the project, more especially for the software design.

3. Requirements

Considering the aims EUROTRA puts forward (c.f. 2) a number of constraints on the technical aspect of the system can be noted.

A. Modularity and extensibility

The complexity of the system together with the fact that a large number of modifications is to be expected make a modular set up into a must.

The size of the project, both in linguistic material as in physical (geographical) realisation require that alterations and extensions should be possible without touching what has been established before. Only a modular design can guarantee system integrity without destroying the basic principle of development.

B. Portability

The groups of collaborators will be located in several countries, and in some cases in more than one city within a country. Not all centres have access to the same type of machine. Therefore software specifications must guarantee the system to be portable over several machine types and operating systems.

It is not excluded that a commercial development period will be added after completion, which will stress even more the importance of portability.

C. Homogeneity

The output of each language analysis will be an interface structure (which will also after transfer be input to any generation phase).

The success of the whole operation depends on the interchangeability of this structure : it cannot be allowed that output from an analysis would not be fit for input into transfer or that any generation module would refuse the interface structure coming to if from transfer. The form in which linguistic facts are represented should be rigidly determined and should contain the same criteria for every interface structure resulting from each analysis phase.

The definition of what information interface structures can contain should take into account the kind of data the separate groups are going to use as well as leave enough freedom for different linguistic strategies they might want to apply.

The homogeneity should also be extended to the software design, this to avoid unnecessary complications.

4. Conclusion

EUROTRA has been presented here as it is up to now : a system in its design phrase. On current research results we have every reason to believe it should work.

When it does it will be the first multi lingual translation system based on such a large scale.

On top of that, its development provides a testing ground for the feasibility of linguistic and computational views, helping both linguists and computer scientists by facilitating their research.