[*Terminologie et Traduction*, no.1, 1989]

EUROTRA - GOALS, ORGANIZATION AND FRAMEWORK DESIGN

Jürgen Vollmer

Outline

0. The Setting

The multilingualism of the European Community is certainly a huge burden for the EC institutions and for trade and industry in general. To solve the problem of coping with the enormous amount of translations necessary the Commission has launched the Eurotra multilingual MT project.

Let us throw a glance on this major problem and on the cost to cope with it. The EC now has nine official languages (Spanish, Danish, German, Greek, English, French, Italian, Dutch and Portuguese). This means that 72 language pairs have to be dealt with in the translation services of the different EC institutions. As a consequence the Community now runs the largest translation and interpretation services in the world at a very high cost (between 35 and 65 per cent of the operational expenditure in the various institutions).

But even so, only a limited service can be provided and many documents are not translated into all official languages or not translated at all due to a lack of resources. The enforcement of urgent political measures is constantly delayed because they become effective only after translation and publication in all languages in the Official Journal of the EC.

It is, however, not possible to increase human and financial resources in such a way that this serious problem could be overcome. The Commission had only two options to face this fact:

- abolition of multilingualism by reducing the number of official languages
- innovation in the translation services by developing efficient tools to assist the translators.

The first alternative is unacceptable to most member states, the second, however, is feasible and efforts have been made in this area. The most ambitious project currently carried out is Eurotra.

In November 1983 the Council of the EC took the decision to initiate this project and asked the Commission to execute it in cooperation with the member states.

The technical objective of Eurotra is the creation of a prototype machine translation system capable of dealing with all Community languages. Politically the project can be seen as a research initiative in the field of MT and computational linguistics aiming at the creation of a "critical mass" of expertise in a highly sophisticated field of endeavour in Europe.

The Eurotra project is unique in many ways. This starts already with the way it is financed and managed. The Council decided to allocate 16 million ECU to this project and to split it up into three phases with an overall duration of 5.5 years. This was extended after the accession of Spain and Portugal to the EC to a period of 7 years and a sum of 20.5 million ECU.

During the preparatory phase of two years the organizational arrangements for the project were agreed and the linguistic and software specifications were defined.

During the second phase of 3 years duration two main areas of work are covered:

(i) basic linguistic research including

the development of initial linguistic models for the analysis and generation of each of the official languages and for transfer between them. This work is corpus based and covers a vocabulary of 2500 entries in a limited specialized field;

the preparation of the lexical database for the above-mentioned vocabulary;

the study of suitable linguistic strategies for the execution of analysis, transfer and generation;

(ii) development of the basic software of Eurotra, in particular of a 'rule interpreter' for testing linguistic rules on a computer.

During the third phase of 2 years duration the bulk of the development work will be done and the researchers will concentrate on:

improvement of the software in order to accelerate testing;

revision of the linguistic models and implementation of the prototype;

extension of the text corpus and of the subject field covered;

revision and extension of the lexical coverage to 20.000 entries per language;

evaluation of the system's technical and economic performance ;

preparation of a proposal for the development of an operational system on an industrial scale.

1. The Organization of the Project

As we have already mentioned before the project is of a decentralized and cooperative structure with the bulk of the research and development work being carried out by teams located in all Member States. The main reason for this setup is that it is difficult if not impossible to gather the required expertise in languages, linguistics

and computer science in one place, and that it would lead to a brain drain which is in complete contradiction with one of the goals of the project: spreading expertise in NLP throughout the Community. At this time there are about 160 people working on this project in 20 centers located in all 12 Member States.

Each of those teams, with the exception of Luxembourg and Ireland, is working on the analysis and synthesis of its own language and the transfer from the other languages to its own. Ireland has been given the task of terminology and lexicography for the project, whereas Luxembourg is acting as a documentation center and as software clearing house for Eurotra.

This structure together with the modalities of co-financing the project have been legally laid down in contracts of association between the Communities and the different Member States.

The executive responsibility lies with the Commission of the European Communities which provides a team of 14 officials (to be increased to 22) in charge of the central management of the program, the administration of contracts, the coordination of the national language groups and the supervision of the system design, i.e. linguistic and software specifications.

In its task the Commission is assisted by the Management and Coordination Committee for Linguistic Problems as an advisory body, and by the Common Steering Committee for the execution of the contracts of association. The day to day technical management, planning and internal assessment of the project is carried out by a Liaison Group consisting of the directors of the national centers and the Eurotra project leader Dr. S. Perschke.

The coherence of the different modules is ensured by a common set of tools consisting of the basic software and of linguistic specifications produced and continuously improved, after testing in the national research centers, by a group of scientists seconded to this task from the various Eurotra centers.

2. Basic Requirements for the Eurotra Framework Design

The need for a formal framework in Natural Language Processing is a commonly accepted fact in computational linguistics and with a view to the Eurotra project it is even more important because of the variety in scientific background and training of the researchers involved and

their geographical dispersion. There are several basic requirements to be fulfilled by such a framework if it were to help the project.

In. a situation like ours the metalinguistic framework is the main vehicle to guarantee that all members of the project share the same assumptions and can communicate in an effective way. The framework therefore had to be *formal* and *simple.*

Due to the experimental nature of MT in general, and of multilingual MT in particular, the framework has to provide an adequate research and experimentation tool. It therefore had to be *modular* and *easily modifiable.*

In a large project like ours a certain turnover of staff with a subsequent need for training is a fact of life. The framework therefore had to be *easy to learn and teach.*

One of the characteristics which make Eurotra unique is its truly multilingual nature. This means that the investment required to produce it does not grow geometrically with the number of languages covered, as does the number of language pairs. The adhesion of Spain and Portugal for instance added two new languages to the seven official languages we had but made the number of language pairs jump from 42 to 72.

This requires that the effort necessary to produce the bilingual components (i.e. the transfer modules for the language pairs) should be reduced to a fraction only of the effort necessary to produce the monolingual components (analysis and synthesis for each language covered).

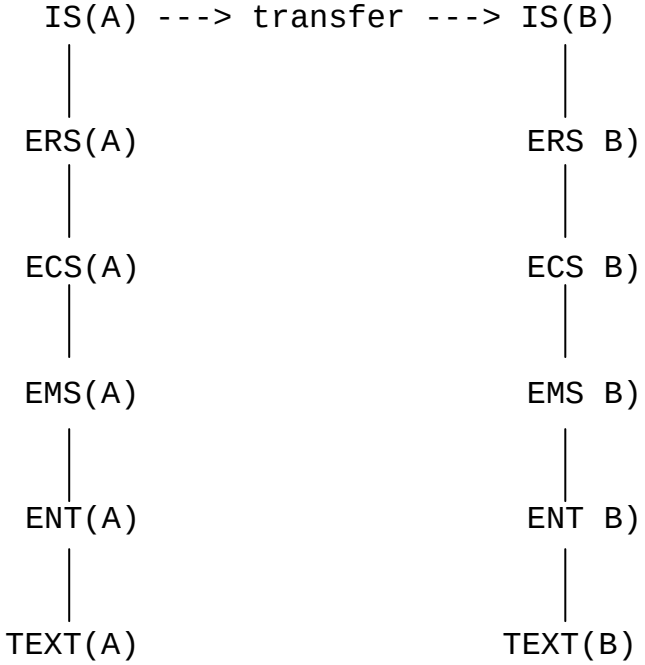
Eurotra claims that the bilingual components can, in principle, be reduced to lexical transfer and it is one of our goals to prove this assumption to be true.

A substantial amount of research is needed to determine the appropriate interface structure, in other words the representation which will serve as input to transfer, in such a way that simple transfer is possible. If this representation is too "interlingual", the quality of translation can be expected to suffer as too much of the information from the source text will be lost. If, on the other hand, the interface representation is not "interlingual" enough, too much source language information will be preserved and the goal of simple transfer will not be achieved.

This leads us to the conclusion that a framework which is adequate for a truly multilingual MT system has to be modular to emphasize experimentation with the interface structure in order to determine its optimal definition without the necessity to change the definition of the remaining representations. The reason for this is obvious, as we cannot afford to throw everything away each time inadequacies in the interface structure are discovered.

We concluded from this that different levels of linguistic description such as morphology, syntax, semantics should be separately defined together with the mappings between them.

Another major reason for computing a representation of a text in order to translate it is the impossibility to systematically relate a text and its translation directly. Similarly the relation between a text and its interface structure is so complex that it is necessary to break this relation up into a sequence of less complex relations. How this is done in Eurotra is shown below:



(Eurotra Stratificational Approach)

where A designates the source language and B the target language, ENT - Eurotra Normalized Text, EMS - Eurotra Morphological Structure, ECS Eurotra Configurational Structure, ERS - Eurotra Relational Structure, and IS - Interface Structure.

In Eurotra terms the description of a level of representation is called a generator and the mapping between two adjacent levels a translator. In our framework analysis and synthesis of a language is achieved through a series of translation steps from the source text to the interface structure; transfer then is just another such translation step and probably the simplest one. The generators are explicitly defined by a set of rules and the translators are simple compositional rules themselves.

The specific type of generators we have chosen in Eurotra was determined by various factors such as the sheer size of the project, its unique decentralized organization and the need to adapt the design during the research phase.

For these reasons the language to describe the grammars had to be easy to learn and teach while being flexible enough to cater for different linguistic assumptions from different schools of thought. In other words our framework had to be formal and simple with an adequate expressive power. It also has to cope with widely differing descriptions of linguistic phenomena occurring in all the significantly different official languages of the Community, where we basically have to deal with three families of languages: Latin, Germanic and Greek.

This required that, also taking into account the research orientation of the project, the grammar formalism be modifiable without making all existing grammar encoding obsolete whenever an addition or modification has to be made.

3. The Generators

Looking back to what has just been said about expressive power and simplicity on the one hand and the requirements of an experimental environment on the other we tried to satisfy the requirements of formality and simplicity by a compromise between generality and constrainedness based on a simple formalism with a declarative semantics: an extended contextfree formalism with an operational semantics based on unification.

The well known classical context-free rewrite rules only admit atomic terminal and non-terminal symbols like in:

```
NP => Det N.
```

The Eurotra formalism, however, allows sets of attribute-value pairs (feature bundles) to occur as terminals and nonterminals, like in:

```
npl7 = (cat=np num=x def=y}  [{cat=det def=y lex=the}
       {cat=n num=x lex=cat}]
```

which could be rewritten as:

```
NP => Det N
num(NP) = num(N)
def(NP) = def(Det)
lex(Det) = the
lex(N) = cat
```

In addition to the standard context-free rules there are also so-called feature rules. These are tree structured patterns applied to complete parse trees. They are essentially used to percolate information in the trees and to enforce co-occurrence restrictions on features.

To give an example we could write the following rules:
np23 = {cat=np} [{cat=det lex=the} {acat=n lex=cat}]
and as feature rules:

```
np-def = {cat=np def=x}  [{cat=det def=x} *$]

np-num = {cat=np num=x}   [*$ {cat=n num=x} ]
```

These three rules would achieve the same effect as the single rule we have just seen before, but they have the obvious advantage that the percolation of number and definiteness has to be stated only once for all noun phrases of the given form.

4. The translators

The translation component of the system translates a grammatical representation of a level to the next one by relating two generators (or grammars) under the constraint of compositionality. This could be formulated as follows:

The translation of a complex object is a (simple) function of the translation of its parts.

An example of a simple translation rule from the German to English transfer grammar is the following:

```
t123 = {sf=gov   cat=v   argl____feat=collective   arg2
       feat=abstr_nonhum lu=verabschieden}
       => adopt
```

Originally, translation was done by defining for every rule of a generator a corresponding set of rules in the generators of the adjacent levels. This led to an undesired increase in the number of translation rules.

Consequently the translation component was simplified by making "normal", straightforward compositional translation a default operation, where only exceptions have to be explicitly stated. This was only possible because the generators' operational semantics could be modified to cater for more powerful information manipulations.

5. How far has Eurotra advanced?

The currently implemented system covers all seven original languages foreseen, that is Danish, German, Greek, English, French, Italian, and Dutch, whereas Spanish and Portuguese are treated according to a different time schedule due to their late arrival

Analysis modules exist for all seven languages, generation modules for only five (no Italian and Dutch generation).

Transfer components are available for the following language pairs:

Danish-English	Danish-German
English-Danish	English-German
English-Greek	French-Greek
German-Danish	German-English
German-Greek	Greek-English

An average grammar has about 400 rules and 600 lexical entries (allowing for ca. 3000 full word forms in moderately inflected languages).

The linguistic phenomena covered include main and relative clauses, all types of noun phrases, simple coordination in noun phrases, all verbal tenses (but excluding modal constructions), possessive, reflexive, relative and indefinite pronouns, all prepositional phrases, adverbs, numerals, and particles.

Research, and experimentation is still ongoing to deal with ellipsis, modality, negation, scope, quantification, time-tense relation and pronoun resolution.

The implementation is based on the original Eurotra framework which has since been modified mainly in the definition of generators and translators as mentioned before to allow for simpler translation components. The grammars are currently being recoded and first reports have shown that it takes one person a few days to recode about half a generator while the translation components get drastically reduced by deletion of rules..

The implementation of the virtual machine is mainly in C-Prolog. A new version of the software including a relational data base management system for the dictionaries (and later the grammars) has been released last fall.

6. Conclusions

The multilingualism of the European Community is a major obstacle to cultural, economic and political integration. Eurotra wil be a tool to overcome communication and trade barriers in a multilingual Community and it will create Europe-wide competence in computational linguistics and NLP. The program also creates the necessary infrastructure for research and development in these areas.

We have tried to demonstrate here that the design of a formal framework for encoding linguistic knowledge had to cater for the experimental nature of advanced NLP applications in general and for the complex organizational structure of a project like Eurotra in particular. This is why we have argued that the concepts of formality, simplicity, modularity, modifiability, constrainedness, expressiveness and learnability are all equally important factors in our particular situation, and that it is only by compromise that a balance between them can be found.

Jürgen VOLLMER
Terminology and Computer
Applications Department (TAI)
Commission of the European Communities
Luxembourg

REFERENCES

(1) *Official Journal of the European Communities* L317 13.11.1982, pp. 19-23

(2) *Official Journal of the European Communities* L342 4.12.1986, p. 59

(3) *Multilingua Eurotra Special Issue,* Vol 3-5, 1986

(4) S. Perschke, *Machine Translation Aims to Unite the EEC Nations,* in: Technology Ireland, October 1986, pp. 17-20

(5) Giovanni B. Varile, Eurotra - *The European R&D Programme for Multilingual Machine Translation,* in: ESPRIT - International Joint Collaborative AI, Special Session at IJCAI-87, Milan, pp. 140-149.