ΜΕΤΑ

Eurotra: the Philosophy Behind it

Ineke Schuurman

See table of contents

Explore this journal

Cite this article

Schuurman, I. (1994). Eurotra: the Philosophy Behind it. *Meta*, *39*(1), 176–183. https://doi.org/10.7202/004059ar

Article abstract

La Communauté européenne a lancé un programme de recherche et développement portant sur le traitement des langues naturelles et plus spécifiquement sur la traduction automatique. Ce programme, Eurotra, s'est achevé à la fin de 1992. La première partie de cet article présente les fondements de ce programme et la deuxième, la philosophie sousjacente au projet.

EUROTRA: THE PHILOSOPHY BEHIND IT

INEKE SCHUURMAN[1]
Katholieke Universiteit, Leuven, Belgium

> The multilingualism of the EC imposes a huge burden on industry and trade as well as on the EC institutions themselves. (EUROTRA 1990: 2)

Résumé

La Communauté européenne a lancé un programme de recherche et développement portant sur le traitement des langues naturelles et plus spécifiquement sur la traduction automatique. Ce programme, Eurotra, s'est achevé à la fin de 1992. La première partie de cet article présente les fondements de ce programme et la deuxième, la philosophie sous jacente au projet.

INTRODUCTION

Some ten years ago, the European Communities launched a R&D Programme focussed on Natural Language Processing, more especially on Machine Translation. This programme, called Eurotra, lasted from 1982, when it was officially approved by the European Parliament, till the end of last year.

In the first part of this paper I will sketch the background to the programme; in the second part the focus will be on the rationale of the project itself.

THE PROGRAMME

The European Communities' language policy

As has been acknowledged by many people, Eurotreans as well as outsiders (*cf.* Maxwell *et al.* 1988), Eurotra was a very ambitious programme in several respects. Not only was it the largest project in the world, both with respect to the number of people and languages involved, it also aimed for translations of high quality, while expertise all over the Community had to be developed:

> [...] it remains one of the most ambitious and most experimental [projects], in that it is attempting to define the foundations of multilingual high-quality translation (Hutchins 1988: 31)

So the Eurotra programme had two basic aims, both related to the language policy of the European Community. A technical one, to develop (a prototype of) a multilingual machine translation system of an advanced nature capable of dealing with all the official languages of the Community; and a political one, creating and disseminating expertise in machine translation, computational linguistics and natural language processing throughout the Community.

No doubt, the first aim was inspired by the fact that the Commission of the European Communities hosts one of the largest translation and interpretation services in the world, owing to EC language policy: in the European Communities nine languages officially spoken in the member states are stated to be of equal importance[2], a consequence of multilingualism in its purest form. The languages and the countries in which they are (the) official languages are Danish (Denmark), Dutch (Belgium, the Netherlands), English

(Ireland, United Kingdom), French (Belgium, France, Luxembourg), German (Belgium, Germany, Luxembourg), Greek (Greece), Italian (Italy), Portuguese (Portugal) and Spanish (Spain).

Having so many official languages has been a fundamental decision. The following quote (Oakley 1993, preface) may give the reader some sense of the background:

> The problems of language are amongst the largest challenges facing the European Community. We are divided by our different languages and the resulting communication failures; we all pay the price and some countries suffer a real penalty behind their minority language barriers. The cost, both in direct economic terms and in the loss of cohesion generated, is very heavy, especially compared to our major competitors in the USA and Japan who have no such internal communication problems. But our languages are of great importance to all of us, epitomising as they do to our past, our history, and our culture. So in a world where much of our differences and individuality has to be surrendered to the greater good of the emerging new Europe, where we have to improve our ability to communicate with each other, it is more than ever important to hold on to and enhance our languages, to cling on to that reminder of our roots in an increasingly shared culture.

This policy has a number of consequences. A practical one concerns communication between residents of the EC with different mother tongues. This continues to be a problem, although many people, especially those living in the smaller countries, learn a foreign language (English and/or French and/or German) at school. But it will always be more difficult to phrase things succinctly in a language that is not one's mother tongue! For business matters, this implies that numerous documents need to be translated, over and over again.

It will be clear that from time to time there is a plea to reduce the number of official EC languages, as all these translations are very expensive. But especially residents of the smaller countries are fiercely opposed to such a step. They fear that this will be the beginning of the end for their mother tongues. And the residents of the larger countries are in favour of a reduction only as long as their own language remains an official EC language. So, politically, such a solution is very controversial.

With respect to the institutions of the European Communities themselves, everyone may address them in his or her own mother tongue in order not to create a language barrier. They will also get answers in their own language. This way, people with, for example, English, French or German as mother tongue, no matter whether they are members of Parliament or ordinary citizens, derive no advantage from the fact that they belong to a major language community, as would have been the case had these three been the official working languages[3]. But this means that at the moment there are 72 (9*8) language pairs, as translations have to be made from each language into every other language.

It is to be expected that in the near future four of the EFTA-countries[4] will join the European Community. These are Austria (German), Finland (Finnish), Norway (Norwegian) and Sweden (Swedish). This would add another three official languages to the nine already present, which brings the total number of language pairs, a figure that is of relevance to the translation services, to 132 (12*11). And, of course, the discussion whether this is a desirable situation or not will flare up once more. Most people will agree that for purely economic reasons a reduction would be good. But note that for these very same reasons the best situation would be to have just one official language. It is quite significant that such a proposal has never been made, as in that case there would be too many losers. And because at the moment all the smaller languages together constitute a majority, they can take a hard line against the bigger ones. For all these reasons, it is to be expected that the workload of the translation services of the Commission of the European Communities will only increase.

And as having nine official languages, as is the case at the moment, also means that all official documents are to be translated into all these languages before they have any official or legal status, it is easy to understand why the Commission needs such a large translation service (at the moment over 1,000 professional translators). And still the enforcement of urgent political measures is constantly delayed because such measures hold only when they have been published in parallel in each official language in the *Official Journal*. It is just impossible (both physically and financially) to translate everything in time. To give you some idea of the cost, this amounts to 35 to 65% of the operational expenditure in the various EC Institutions (EUROTRA 1990: 2). Therefore in that same brochure the following sigh was heaved (EUROTRA 1990: 2):

> Europe's richness in languages is a double-sided coin. On the one hand it is generally acknow- .
> ledged that the variety in European languages and cultures is an asset worth preserving. On
> the other hand there is the enormous price that has to be paid in order to maintain this cultur-
> al richness, that is the huge financial cost that arises when language barriers have to be over-
> come by human translation.

In 1992 the cost of translation to the European Commission itself exceeded 150 Mecu per annum, to which have to be added the hidden costs, in failure of full communication and delays inherent in a system where translation is required but is only available in due course, dwarving the direct costs (Oakley 1992: 11.1).

Note that the costs incurred in the EC institutions will only be the tip of the iceberg (EUROTRA 1990: 2):

> Much more serious is the cost of multilingualism for industry, commerce and services all
> over Europe. This is enormous (although its true volume is unknown).

and one will get an idea of the scale of the problem.

Machine Translation

Some time ago, the heavy translation workload made the Commission think of bringing in a machine translation system; accordingly, in the second half of the seventies they bought Systran. But the fact that this was a system of non-European origin caused some resentment, as by that time at several research centres in Europe people were working on machine translation systems of more advanced design. Furthermore, as adding new language pairs to Systran turned out to be more problematic than expected, the Commission decided to start its own R&D programme, Eurotra. In so doing, it also bore in mind that the Community could not afford to fall behind the US and Japan in the domain of Natural Language Processing.

From 1978 on, researchers from several European universities, the Eurotra Coordination Group, met on a regular basis to prepare their own machine translation project. Several preparatory studies were carried out, financed on a very low level by the Commission. In the very beginning only five languages were involved: Dutch, English, French, German and Italian. After some time Danish and Greek were added. In 1982 the project was approved by the Parliament of the European Communities. However, work could not start before the first contracts of association were signed, which was in 1984. When Portugal and Spain joined the European Communities, Portuguese and Spanish were added (1986).

Dissemination of knowledge

In order to develop a machine translation system for the Commission, the best thing to have done would have been to create a large research unit at one central place. The

solution chosen was rather the opposite, namely to have smaller research units in all member states: Belgium (Leuven, Liège), Denmark (Copenhagen), France (Nancy, Paris), Luxembourg (Luxembourg), Germany (Saarbrücken), Greece (Athens), Ireland (Dublin), Italy (Pisa, Torino), the Netherlands (Utrecht), Portugal (Lisbon), Spain (Barcelona, Madrid), United Kingdom (Essex, Manchester).

This strategy is to be related to the second aim of the Eurotra programme, the dissemination of knowledge in the domain of Computational Linguistics over all member states. In several countries there were no or few centres for computational linguistics before Eurotra started (*e.g.* Belgium, Spain, Greece, Luxembourg, Ireland). Now, in many countries the Eurotra centres have become the main sites for research, development and training in this field. What is more, the training has not been limited to the Eurotra centres themselves, although over 300 staff have been trained. At the universities in the cities where Eurotra centres are located, courses and/or study programmes in computational linguistics have been founded[5]. In most cases (ex-) Eurotra staff are involved. In this way many students have become acquainted with this new area.

Another reason for having a decentralized project was the lack of a special EC research centre for Natural Language Processing to host the project. One language group was established per official language. Such a language group takes care of analysing and generating its own language as well as transfering from other languages. All countries had their own languages groups (on the understanding that the groups in Utrecht (NL) and Leuven (B) together made up the Dutch language group), except for Dublin, Liège and Luxembourg, who had special tasks related to terminology, lexicography and the software environment respectively. Liège carried out some work for the French language group too.

As a consequence of the decentralized structure, tasks of central importance (formulating linguistic specifications, developing the core formalism) had to be carried out by special task groups, whose members were spread all over Europe. The Commission team was too small and also lacked the competence to fulfil all these tasks themselves. Their main task concerned management, especially during the first few years.

It will be clear that such a large, decentralized project needs a fairly heavy management structure with several layers. Therefore, the central, day-to-day coordination was carried out by the Commission, Directorate-General for Telecommunications, Information Industries and Innovation (Luxembourg), assisted by the Directors of the language groups. They met every six weeks at the Commission to discuss the progress of the project and to align the activities in the various centres.

It will hardly come as a surprise that Hutchins (1988) should state that Eurotra was also very ambitious in political logistics.

The outcome

Eurotra was among the first Research and Development Programmes launched by the EC. As is usual in such circumstances, it suffered from a lot of growing pains. One of them is related to the somewhat unusual funding structure, as in this programme part of the funds are provided by the national authorities. And in the early years, up to the start of the so-called Second Framework, it was not related to any of the Commision's R&D programmes. One of the effects of this was that contracts of association had to be negotiated with all member states separately. This led to a situation in which the Eurotra sites entered the programme at different times. The Leuven group, for example, started working in 1984. Utrecht, representing the other two-thirds of the Dutch language group, only joined the project in 1986. In the meantime the small Leuven group had to do all the *Dutch* work on its own.

Nevertheless, the results of Eurotra are quite impressive. Not only is there a fairly small but running prototype in the mainstream version of the Eurotra formalism, there are also several sideline prototypes, like CAT2 and MIMO. There are already some spin-offs like intelligent spelling and grammar checkers. In the years to come it is to be expected that the results of Eurotra will be validated also with respect to machine translation. At the moment there is a large Eureka project, Eurolang, which exploits Eurotra know-how (both by hiring ex-Eurotra staff and by making use of its linguistic backbone). Industry has also shown interest in systems based on Eurotra for translating technical texts (manuals and the like). One such system, PATRANS, is already being worked on (in Copenhagen (DK), launched in 1992).

Quite impressive also are, on the one hand, the language specifications in the Implementation Reports, for example, and, on the other hand, the linguistic specifications in the Reference Manuals. The latter contain grammar rules arranged by subject matter, with examples taken from all the relevant official languages. The former provide an extensive formal description of all the individual languages (including several languages that had never been dealt with before.) Both are considered as being of outstanding value, not only for computational linguistics but also for linguistics in general (Oakley 1993: 7.1-7.2). The Commission intends to make the Reference Manual an official publication shortly. In its present form it is already being used by several projects. Quite a number of these are unrelated to Eurotra, which shows its impact on the NLP world.

THE ACTUAL PROJECT

The design

When in 1978 researchers started thinking of their own European machine translation system, they knew that by definition it should be a system of advanced design in order to be able to cope with the non-trivial demands of the Commission. To mention but two of them: the system should be flexible enough to make it possible to add new languages in a straightforward way, and the translations should be of high quality. The system they came up with was an advanced one, or rather, it was ahead of its time.

In what follows I will concentrate on the philosophy behind the system (for a description of its linguistic aspects, see Copeland *et al.* 1991a; for a description of the formal specifications of the system, see Copeland *et al.* 1991b).

A multilingual system

As at the beginning of Eurotra the number of languages that had to be dealt with was already quite large and was expected to increase even further, the system had to be an extensible, multilingual one. Multilingual, both because of the number of languages involved and because of the demand that accommodating new languages should be possible without too much effort. In such a situation an effective procedure calls for multilinguality, using one and the same analysis of the source language to translate into all the target languages. The use of x bilingual systems was conceived of as inadequate, although this would give one the possibility of adapting the analysis of the source language to the individual target languages. This way translating from German to Dutch would call for an analysis of German, different from that required when translating from German to Greek. Mutatis mutandis the same would hold for generation.

Having to deal with seven languages, as Eurotra had in the beginning, this would require six analyses per language, and six different generations. So the total number of analysis and generation components would amount to eighty-four ((7*6)*2). With nine languages, as in the actual situation, there would be one hundred and forty-four ((9*8)*2).

And with twelve languages, which might be the case in the near future, it would be two hundred and sixty-four ((12*11)*2).

Note that writing such analysis and generation components is by no means a trivial task. It calls for substantial contrastive research, in order to come up with the optimal adaptations: therefore, such a system will be very expensive in terms of manpower. Skilled people with a very good knowledge of both source and target language are needed, not only for writing the transfer components, but for writing the other (analysis and generation) components as well. In fact, in this option the transfer components would be very small, consisting mainly of simple transfer. Most of the bilingual work will have to be done in the other components. Therefore, I have abstracted from the transfer components when mentioning the number of components needed in a bilingual system.

But even in a bilingual system transfer needs to be carried out one way or the other in order to obtain high-quality results (see the paragraph *A transfer-based approach*). Transfer is called simple in case it involves nothing but copying structure and replacing lexical elements from the source language by those of the target language. As the languages to be dealt with are so different (note that some of them are Germanic languages, others Romance languages and that Greek belongs to neither of these groups), it would be rather naive to believe that one could make simple transfer alone.

When translating from Greek to Danish, resemblances (lexical but also structural) will be purely accidental. Even in somewhat related languages, like Dutch and English, one has to deal with non-trivial differences in respect of both lexicon and structure.

(1)   *Hij zwemt graag*
(1')   He likes to swim
(2)   *Ik wil je graag helpen*
(2')   I'll be glad to help you

In these sentences the Dutch adverb *graag* (English *gladly*) is not translated simply as an adverb in English. In the first pair of sentences the semantics of *graag* is slipped into the verb *to like*, in the second pair of sentences the simplex construction in Dutch is translated by a complex one in English (*graag*, vs. *to be glad*). This is called *complex transfer*.

Taking all this into account, a multilingual approach was considered to be more appropriate. As stated above, in such a system one and the same analysis of the source language is used to translate into all the target languages. And all these have but one generation component, which is used with whatever source language. With nine languages there are therefore analysis and nine generation components, one of each for each language. In addition, there will be seventy-two (9*8) transfer components. In transfer the language-dependent rules are handled, if possible by use of simple transfer, but where necessary by use of complex transfer. This component is therefore bilingual. In sum, a system with nine languages calls for ninety (9+9+72) components. In a bilingual system this would have been one hundred and forty-four (see above).

One of the characteristics of a real multilingual system is that it is impossible to tailor source and target language to each other. Instead, one tries to define both analysis and generation components in such a way that language-dependent properties are phrased in a rather abstract, language-independent way, for example in a formal semantic notation.

A transfer-based approach

As mentioned earlier, Eurotra was a very ambitious project, in aiming, among other things, at a multilingual machine translation system with high-quality output. Such requirements make certain approaches less obvious than others. In paragraph *A multilingual system* reasons were given for opting for developing one real multilingual system, instead of a whole series of bilingual ones. But why was a transfer strategy chosen?

At first sight, an interlingua approach seems to be preferable, at least if we leave aside the older versions (those using a standard language as the intermediating language). In progressive interlingua systems there is no need for transfer components at all. And it is true, for large multilingual systems, the *quadratic* (x^n-x) rise in the number of (bilingual) transfer components is in fact not acceptable. So, once more, why is it that Eurotra uses a transfer-based system?

The reason is a very trivial one, an interlingua system was conceived of as unrealistic, at least if one aims for high-quality translation results for an, in principle, unbounded number of languages (as Eurotra does). A high-quality translation is one that is more than a pharaphrase in that it gives more than just the essential message. Also, the way in which this message was phrased is, as far as possible, expressed (*cf.* Hauenschild 1988)[6]. In case of an interlingua system, this calls for very *expensive* analysis and generation components. And then there is the old problem of finding an interlingua that is suitable for any source and target language. Even for Eurotra this was a problem, although one *only* had to find an interlingua suitable for all European languages (as the system was to be used in the EUROPEAN Communities) (*cf.* Hauenschild 1988).

This is just to say that the pure interlingua concept was only something to dream of, and not a realistic alternative for Eurotra. Nevertheless, such a system remains the kind of system one should strive for.

Current interlingua systems are either producing worse translation results when compared with transfer-based systems or they are producing good translation results, but at the cost of their ability to accommodate new languages (*cf.* Tsujii 1988; Van Eynde 1993a). The latter is caused by the fact that in such systems the set of target languages is anticipated and therefore the monolingual components become language-set specific. Such

> attunement opens the door for the inclusion of rules and representations in the monolingual modules which lack linguistic motivation, and this jeopardizes both the construction and the reuasability of the resulting grammars (Van Eynde 1993a:24).

One might have to revise all monolingual grammars in case new languages are added.

There is still another type of system around, which is in fact a mixture of an interlingua-based and a transfer-based system. Wherever possible, an interlingua will be used, the rest being handled through transfer. Such systems can deliver high-quality translations while at the same time new languages can easily be incorporated.

Eurotra is such a system, although it is called a transfer-based system most of the time. And indeed, even in *simple transfer*, the part of the system that comes closest to the interlingua philosophy, language-dependent information that BY CHANCE is identical between the languages involved may also be involved (Allegranza *et al.* 1991). On the other hand, in the areas of tense, aspect, and negation an interlingual semantic approach is made use of. And in certain other areas too, the research results look quite promising (determination, quantification, mood, modality, diathesis) (Allegranza *et al.* 1991; *cf.* also Hauenschild 1988).

Even after some ten years, it turns out that the design of Eurotra is not outdated, keeping in mind the needs it had to meet. In the meantime, some new languages have been added (Spanish and Portuguese), and it worked out!

CONCLUSION

Eurotra, both as a programme and as a project, has been experimental and very ambitious, in that it has attempted to lay the foundations of high-quality multilingual translations and, in so doing, has contributed substantially to the theoretical foundations

of Machine Translation. At the same time, it has promoted computational research all over Europe.

Notes

1. The author has been deputy head of Eurotra-Leuven since 1989. She would like to thank Birgit Bekker, Alex Schoenmakers and Bruno Tersago for their support and helpful comments.
2. To some extent, citizens of certain regions, for example Catalonia (Spain) and Friesland (the Netherlands), are *discriminated* against. Their mother tongues, Catalan and Frisian respectively, are not official languages of the European communities, even though they have official status in the member states themselves. However, in these regions one of the official EC languages is an official language too (In the cases mentioned above, these are Spanish and Dutch).
3. In corridor chats English, French and German will often be used, as these are widely known.
4. The other EFTA (European Free Trade Association) countries are Iceland, Liechtenstein and Switzerland.
5. Except for Luxembourg as there is no university in the Grand Duchy.
6. This being the reason for the firm linguistic (syntactic and in particular semantic) underpinnings of modern high-quality machine translation systems.

REFERENCES

ALLEGRANZA, V., BENNETT, P., DURAND, J., VAN EYNDE, F., HUMPHREYS, L., SCHMIDT, P. and E. STEINER (1991): "Linguistics for Machine Translation: The Eurotra Linguistic Specifications", *Copeland et al.* 1991a, pp. 15-123

COPELAND, C., DURAND, J., KRAUWER, S. and B. MAEGAARD (Eds.) (1991a): *The Eurotra Linguistic Specifications.* Studies in Machine Translation and Natural Language Processing. Vol. 1. Office for Official Publications of the European Communities, Luxembourg

COPELAND, C., DURAND, J., KRAUWER, S. and B. MAEGAARD (Eds.) (1991b): *The Eurotra Formal Specifications.* Studies in Machine Translation and Natural Language Processing. Vol. 2. Office for Official Publications of the European Communities, Luxembourg

EUROTRA (1990): *Eurotra.* Office for Official Publications of the European Communities, Luxembourg

HAUENSCHILD, C. (1988): "Discourse structure - some implications for machine translation", Maxwell et al. 1988, pp. 145-155

HUTCHINS, W. J. (1988): "Recent developments in machine translation", Maxwell et al. 1988, pp. 7-62

MAXWELL, D., SCHUBERT, K. and T. WITKAM (Eds.) (1988): *New Directions in Machine Translation,* Distributed Language Translation Series 4, Foris Publications, Dordrecht-Holland / Providence RI-USA

Oakley report (1993): EUROTRA: Final Review Panel Report. (Final draft). Internal Publication, CEC Luxembourg

TSUJII, J. (1988): "What is a cross-linguistically valid interpretation of discourse?" Maxwell et al. 1988, pp. 157-165

VAN EYNDE, F. (1993a): "Machine translation and linguistic motivation", Van Eynde (1993b), pp. 1-43

VAN EYNDE, F. (Ed.) (1993b): *Linguistic Issues in Machine Translation,* Communication in Artificial Intelligence Series, Pinter Publishers, London and New York