Chapter 14: Projects and systems at the Commission of the European Communities (1976-)

With the accession of more countries to the European Communities, the multilingual policy of the Treaty of Rome was increasing the demands for translations at an alarming rate (Ch.9.1). By 1975 the Commission had already started investigating the possibilities of MT. In that year it had the Systran system demonstrated in Luxembourg (Ch.12.1). The next year a Plan of Action (CEC 1976) was approved which established a coordinated series of studies and projects concerned with the Communities' multilingual problems. Part of the plan was concerned with the development of Systran, part with the development of terminology databanks (with particular reference to EURODICAUTOM, Ch.4.7 and 17.6), part with the development of MT for restricted languages (principally the TITUS system, Ch.17.3), and part with long-term research on a full scale multilingual MT system (the Eurotra system, Ch.14.2 below).

14.1: Systran

The translation service of the Commission of the European Communities began developmental work on Systran systems in February 1976 with the purchase of the English-French version. In early 1978 it acquired the French-English version, which had been under development by the World Translation Center (WTC) in La Jolla since mid-1977. The Commission's contract with WTC included an agreement for substantial development of the systems by staff of the Commission. Indeed most of the developmental work on these two systems took place within the Translation Department at Luxembourg. In 1979 an Italian synthesis program was coupled to the English analysis program, and the development of the English-Italian version was started, again most of it in Luxembourg by staff of the Commission. By the end of 1979 there were thus three language pairs under development, and in March 1981 it was considered that each system was producing reasonable enough output to set up a pilot production service in Luxembourg. Since this date, the number of translations produced has grown steadily and developmental work on the three systems has continued. Encouraged by this achievement, the Commission began in 1982 to work on English-German and French-German systems (Pigott 1983, 1983a).

As we have seen (Ch.12.1) the basic software for the analysis and synthesis programs of Systran is in two parts: the language-independent programs for control and dictionary searching which are provided by the designers, and those specific to particular SL-TL pairs. The latter have been written partly by WTC and partly by programmers at the Commission (Van Slype 1979b). Initially all re-programming had to be done in La Jolla, but later the Commission sponsored Margaret Masterman at the Cambridge Language Research Unit (Ch.5.2) to develop a program for the automatic annotation of the Systran macro-assembler code (Masterman 1980), and thereby facilitating emendation of programs by Commission staff.

Certain components of the computer programs show a high degree of reliability. According to Pigott (1984), morphological analysis and synthesis of French is "100 per cent successful", resolution of grammatical homographs (e.g. deciding whether *light* is a noun, adjective or verb) is over 90% successful, and TL synthesis is unproblematic in general. Although "by far the most complicated part" is SL analysis, Systran has, in Pigott's opinion, "achieved a relatively high level of success".

Since so much of the success of the Systran translation process depends on the quality of the bilingual dictionaries, it has been the enhancement of dictionary entries which has received most attention by those involved with the development of Systran. However, such is the monolithic complexity of Systran's dictionaries that special care has to be taken to ensure that an 'improvement' introduced to deal with one particular problem does not degrade performance in another part of the system (as has been found in the USAF Russian-English system). This is of

course always the danger when changes are introduced piecemeal, but there is little alternative with Systran. Lexical data are typically irregular, and changes had to be introduced by trial and error. Therefore, after Systran became fully operational in the spring of 1981, amendments to systems have been done on test versions and introduced into production versions only after extensive trials.

Probably the most significant innovation has been the considerable expansion of the use of semantic markers and semantic classification. At an early stage of the devilment work, Pigott (1979) had to find out how to use Systran's semantic markers. There were found to be available over 20 'process type' markers for different subject fields: AGPRO (agriculture), ANTEC (analysis), PRAVIA (aviation), PRIO (biology), PRCH (chemistry), PRCR (creative), PREL (electrical), etc. The assignment of these markers to dictionary entries depended largely on subjective judgements of the most appropriate subject field. If texts were restricted to specific fields this might work, but in the context of the Commission's work it was soon found that such subject limitations were impractical, and the use of 'topical glossaries' was dropped. Instead, a much reduced set of general semantic markers was developed which could be assigned more consistently, e.g. DEV (device, tool, instrument), CONTR (container), MATER (material or substance used for production or operation).

Other innovations introduced by the Commission's translator staff working on Systran include a routine for dealing with words not found in the dictionary (Wheeler 1984). In general such words are left untranslated (as in the Russian-English system), but it was felt that something could be done with those having regular endings. The routine enables not only the assignment of the probable semantic marker (e.g. a French word ending in *-meter* would be coded as a device, one ending in *-ologie* or *-isme* as a branch of science), but also the provision of standard TL endings, so that *-ogue*, as in French *radiologue*, would be rendered *-ogist* (*radiologist*). Another innovation concerned the treatment of minutes of meetings (voluminous in the European Communities, and tedious to translate); in French and Italian these are conventionally recorded in the present tense, but in English the custom is to use the past. The routine involved the assignment of a 'typology category', which allowed for automatic tense conversion and the changing of words such as *demain* to *the day after*, and so forth (Wheeler 1984).

Although most innovations were concerned only with lexical problems, some involved syntax. For example, in order to prevent the erroneous translation of *The committee discussed faulty equipment and office management* as *Le comité a étudié l'équipement et l'administration de bureau défectueux*, semantic codes were incorporated which linked the adjective *faulty* and nouns categorised as 'devices' (Pigott 1984). As another example, the phrase *éviter que l'argent soit dépensé* should be translated by the English construction *prevent the money being spent*, and not *that the money be spent*. Rather than introducing additional rules to the analysis routines, it was decided to include instructions in the dictionary entry for *éviter* (Wheeler 1984).

It is admitted that the pragmatic approach of the development team in Luxembourg renders the systems much more complicated. There are clearly dangers in confounding syntax and lexicon procedures (risks which are inherent in all Systran systems) and in devising rules which are so dependent on complex dictionary entries. To some extent, such problems have been eased by the development of tools such as the dictionary concordances, but it is still not known how far complexity can be increased before the limits of adaptation have been reached (Laurian 1984). As we have seen (Ch.12.1) the Russian-English version seems to have now reached that limit.

When the Commission bought the English-French system in February 1976 its dictionary comprised just 6,000 entries; by 1984 there were some 150,000 dictionary entries in each of the three language-pair versions, i.e. English-French, French-English, and English-Italian (Pigott 1984). At the same time the programming rules of analysis and synthesis have also expanded, from some 30,000 lines in the 1976 English-French system to some 100,000 lines in each of the 1984 versions. Obviously, a great deal of effort has been devoted to the development of Systran; in 1983, Pigott (1983a) estimated that "an average of about twelve professional staff (linguists and data

processing experts) have worked full time on the project for the past eight years", with development costs reaching 4.5 million ECU (about four million dollars) in the same period. However, there are good grounds for believing that these costs can be amortised within a few years if the growth in usage continues. After a slow start (1250 pages in 1981, 3150 pages in 1982), "over 40,000 pages of MT were run in 1983 on the various production systems" (Pigott 1984) Most use is made of the French-English and English-Italian systems, "French translators have until now been less enthusiastic ... and generally prefer to use more conventional methods" (Pigott 1983a). As an indication of the impact of MT at the Commission, Pigott (1983a) reports that "in January and February 1983, 50% of the English-Italian workload (293 pages) and 25% of the French-English workload (330 pages) were handled with the assistance of Systran". Since 1982 translation output has increased by the introduction of word processors linked directly to the IBM mainframe on which Systran runs (Pigott 1984).[1]

Rather than fully revised translations, users have the option of a faster service providing lower-quality translations which have received minimal post-editing. The 'rapid post-editing' service is offered for French-English translations of informative texts (working documents, minutes of meetings, reports of activities.) Editing is done on a word processor at a rate of up to 4 or 5 pages an hour. The option is popular with a number of users and, perhaps surprisingly, welcomed with some enthusiasm by CEC translators who find rapid post-editing an interesting challenge (Wagner 1985).

The English-French system at the Commission of the European Communities has been evaluated on two occasions; in October 1976, shortly after delivery (Chaumier et al. 1977), and in June 1978 (Van Slype 1979a, summarised in Van Slype 1979b). In both evaluations comparisons were made between Community documents in the form of human translation (HT) after revision, 'raw' unedited MT, revised MT, and the original English, each in terms of their intelligibility, their fidelity, and types of errors.

On scores of intelligibility (clarity and comprehensibility) unedited MT had improved from 44-47% to 78% (in both evaluations revised MT, at 92-97% in 1976 and 98% in 1978, equalled edited HT (98-99%), and the original texts at 94-99%). As for fidelity, the 1978 evaluation gave a figure of 73% for unedited MT, and a subjective assessment of style rated unedited MT at 76%. Correction rates may have slightly improved: in 1976, only 61-80% of "grammatical agreements (gender, number, elisions, contractions, person, voice, tense, mood) were correctly rendered", while in 1978 the average proportion of words amended was 36% (over half involved the replacement of words). In both evaluation the main source of errors was the dictionary, and Van Slype rightly emphasised that improvements would come with the expansion and improvement of dictionary entries (in 1978 the English-French dictionary contained just 45,000 entries.) Subsequent experience has demonstrated the truth of this judgement.

Van Slype (1983) considered that MT is acceptable to the user if intelligibility exceeds a threshold of 70-75%, and that post-edited MT becomes cheaper than revised HT if the revision rate is below a threshold of 30-40%. Both thresholds have now been reached by the English-French system.

The progressive improvement of the English-French system can be seen from example 'raw' unedited translations. In 1978, the following was produced:

Pendant les dix premiers ans, la politique agricole commune principalement a é'té fondée sur l'organisation commune des marchés agricoles. Par la politique sur des

[1] For later descriptions of Systran in operation at the European Commission see: A.M. Loffler-Laurian: La traduction automatique (Villeneuve d'Ascq (Nord): Presses Universitaires du Septentrion, 1996); special issue on Systran, *Terminologie et Traduction* 1: 1998; A.Petrits et al. 'The Commission's MT system: today and tomorrow', *MT Summit VIII: machine translation in the information age. 18-22 September 2001, Santiago de Compostela, Galicia, Spain. Proceedings*, ed. B.Maegaard (Geneva: EAMT, 2001), pp. 271-275.

marchés et des prix, on a éliminé la fragmentation des marchés agricoles dans la Communauté.

(English original: During its first ten years, the common agricultural policy has been mainly based on the common organisation of the agricultural markets. Through the policy on markets and prices, the fragmentation of agricultural markets within the Community was eliminated.)

A translation from 1983 is the following extract from the minutes of a meeting (illustrating the tense conversion routine mentioned above):

L'attention du groupe de travail est également attirée sur le fait qu'environ 50% de l'argent de traitement de données consacré à l'enquête doit être employé pour des contrôles et des corrections des bandes nationales de données, provoquant des contraintes financières ultérieures en analyses finales des données.

(English original: The Working Group's attention was also drawn to the fact that about 50% of the data processing money devoted to the survey had to be used for controls and corrections of the national data tapes, causing subsequent financial restraints in the final analyses of the data.)

The achievements of the other versions can also be illustrated. Pigott (1983b) gives examples of unedited English-French, English-German and English-Italian versions of the same passage (on the introduction of MT in the Commission):

(English original:

In addition to expansion of the dictionaries and programs for the three language pairs, increasing attention has been paid to the requirements of translators as users, system enhancement now being based largely on feedback from post-editors of machine-translated texts. In the interests of translator acceptance, text processing equipment was installed to eliminate transmission delays and provide appropriate means of post-editing on screen or paper.)

English-French:

En plus de l'expansion des dictionnaires et des programmes pour les trois couples de langues, une attention croissante a été prêtée aux besoins des traducteurs comme utilisateurs, amélioration de système maintenant étant basée en grande partie sur la rétroaction des post-éditeurs des textes traduits par ordinateur. Dans l'intérêt de l'acceptation de traducteur, l'équipement de traitement de textes a été installé pour éliminer des retards de transmission et pour fournir le moyen approprié de post-édition sur l'écran ou le papier.

English-Italian:

Oltre e espansione dei dizionari e dei programmi per le tre coppie linguistiche, l'attenzione crescente è prestata ai requisiti dei traduttori come utenti, potenziamento di sistema adesso che è basato in gran parte sul feed-back dai postredattori dei testi tradotti dall'ordinatore. Negli interessi dell'accettazione del traduttore, l'attrezzatura dell'elaborazione di testi è installata per eliminare i ritardi di transmissione e per fornire i mezzi appropriati di postredigere sullo schermo o sul documento.

English-German (still under development at this time and therefore of poorer quality than the others):

Zusätzlich zu Ausdehnung der Wörterbücher und der Programme für die drei Paare der Sprache ist erhöhende Aufmerksamkeit zu den Bedarf von Übersetzern als Benutzer, Steigerung des Systems gezahlt worden jetzt, auf Rückkopplung von Revisoren maschinell übersetzter Texte grösstenteils stützend. In den Interessen von Annahme des Übersetzers wurde processing Ausrüstung des Texts, um zu beseitigen Verzögerungen

der Übermittlung installiert und zu liefern geeignetes Mittel des Revidierens auf Schutzwand oder Papier.

Finally, two examples are given from the French-English system. The first is taken from the articles of an agreement:

> The application of these methods to the definition of a management and possible valorization policy of waste requires knowledge a variability of the behaviour of the co-products according to their origin (nature of the methods and manufactures) and to their time to production; indeed, if variability is low it will be possible to define general elimination rules and in the contrary case, it will be necessary to organize the catch counts some and the follow-up of waste on the level even the producing factories.

The second is from a technical report:

> The detection of the gamma rays requires their interaction with a matter. It results from this interaction either an electron accompanied by the emission by a photon by lower energy (Compton effect), or a electron-positron pair, dominant phenomenon beyond some MeV. In both cases the produced charged particules take away certain information concerning the direction and the energy of the incidental gamma photon.

Clearly, Systran MT translations are a long way from being 'high quality' products, and there are instances of errors which would appear to be easily resolved, but in general the verdict must be that for those interested in what was intended (the information content) and not in how it was communicated such translations must be of considerable value. There should no longer be anyone who does not believe that MT is here to stay. The volume and quality of translated output in the Commission's translation service is ample proof.

14. 2: Eurotra

It was recognised from the outset that Systran's potential as a multilingual system was limited. It was true that new language pairs could be added within the general framework of Systran, but the fundamental structure of the system was not amenable to full multilinguality, nor open to advances in computational and linguistic techniques. In February 1978 discussions began under the auspices of the Commission among MT experts from European universities including Grenoble, Saarbrücken, Manchester (UMIST) and Pisa (Rolling 1978, King 1982). Development of the software systems and design of the linguistic framework began during 1979, and by 1982 the general structure of the proposed Eurotra system had been agreed among the participating groups (King 1982). The overall director of the project at Luxembourg is Serge Perschke, who had led research at EURATOM on a multilingual MT system (Ch.11.1). The responsibility for coordination and secretarial services is in the hands of the Istituto per gli Studi Semantici e Cognitivi (ISSCO), Geneva, under the direction of Margaret King.

By 1981 it was estimated that there were some 80 researchers involved, mainly from universities, representing each of the eight member states. In November 1982 Eurotra became an independently funded project of the Commission, granted 16 million ECU (about 12 million dollars) by the Council of Ministers for a five and a half year programme of research and development, under the aegis of the Committee for Information and Documentation in Science and Technology, towards "the creation of a machine translation system of advanced design (Eurotra) capable of dealing with all the official languages of the Community. On completion of the programme an operational system prototype should be available in a limited field and for limited categories of text, which would provide the basis for development on an industrial scale in the period following the current programme"." It was envisaged that work would be in three stages: two years of preparatory work, two years of "basic and applied linguistic research" and eighteen months of "stabilisation of the linguistic models and evaluation of results" (*Multilingua* 2 (1) 1983, p.43). Initially the project is being wholly supported by the Commission; in later years, progressively larger proportions of the funding will devolve to national governments.

The specific requirements for Eurotra were, then, that it should be a multilingual system based on the most advanced linguistic and computational techniques, open to progressive improvements and to incorporation of new methods and techniques, that it should be a practical operational system for the use of translation services of the European Communities, and that it should involve all member countries in collaborative research. This ambitious project is still at an early stage, and details of the system are not publicly available (protection of the software from potential commercial exploitation has also played a role); however, brief accounts of the overall philosophy and general design features have been given by King (1981), King (1982), King and Perschke (1982), Laurian (1984), Des Tombe et al. (1985).[2]

Eurotra is probably unique in having been designed as a multilingual research system from its inception. Work on all languages is being pursued simultaneously: programs of analysis, of synthesis and of transfer for Danish, Dutch, English, French, German, Greek and Italian. Spanish and Portuguese will be added now that Spain and Portugal have become full Community members. To achieve multilingual output, the input to transfer and synthesis programs must be specified exactly; analysis programs must deal with all monolingual ambiguity; and, to reduce the potential complexities of bilingual dictionaries and grammars the transfer modules are to be kept as small as possible. As a consequence, transfer 'interface' structures have to be precisely specified in a highly abstract formalism of sufficient richness to carry information required in synthesis. Although approaching interlinguality, transfer elements will not be universals but be common to European languages, i.e. 'Euro-versals'. In many transfer systems, some steps towards TL-like structures are made during transfer; in Eurotra this will not be possible, and consequently, programs for TL synthesis have to be more powerful than envisaged in the past. For example, in translating English *must* (or its equivalent in another SL) the input to the French TL program has to specify *devoir* or *falloir*; but it should do no more. The different syntactic structures required for TL output are to be generated by the French synthesis program alone (finite verb + infinitive for *devoir* vs. impersonal + *que* + subjunctive for *falloir*). They are not to be produced by the syntactic transfer routines (as would be the case in SUSY and METAL) Likewise, the selection of the correct TL prepositional form is to be carried not by routines instigated by the bilingual SL-TL transfer dictionary, but by routines within TL synthesis which take into account only the relevant lexical and structural relationships of the TL.

The requirement that Eurotra should be a practical operational system as early as possible and that it should be as easily extensible as possible has important consequences. Practicality demands fail-safe mechanisms which can ensure that some reasonable translation is produced if analysis programs fail to derive the semantic representations required for transfer. To achieve this, it is desirable that SL structural features at all levels of analysis are retained: morphological information, lexical features, syntactic relations, semantic compatibilities, etc. Extensibility demands flexible incorporation of new grammars and dictionaries, easy amendment of dictionary entries and grammar rules without the risk of unforeseen side effects, and adaptability to new linguistic and computational methods.

Both aims are achieved by the Eurotra project's uniform data structure formalism, the 'chart' concept (Ch.9.15). In the elaboration of the chart formalism the influence of GETA and the contributions of the Saarbrücken group have played major roles. Larger structures (e.g. trees) are built from smaller ones; if there can be more than one structure for a given SL text segment, the chart allows all possible analyses to be expressed. Different methods of analysis can be employed to build different parts of the analysis, e.g. non-deterministic strategies can operate alongside deterministic techniques. Charts can incorporate tree structures of considerable complexity: dependency tree structures on which all levels of linguistic categories and relationships can be represented (as in the GETA representations, Ch.13.3 above). Tree structures are manipulated by

---

[2] For details of later developments see the ten volumes of *Studies in Machine Translation and Natural Language Processing* published by the Commission of the Euroepan Communities from 1991 to 1998.

production rules (tree transduction rules) which are externally controlled by series of (linked) 'subgrammars' (as in GETA). Grammar rules are formulated as specifications of input data structures (i.e. the expected partial analysis already achieved) and of the changes to be carried out. They may be as complex as required, specifying any combination of morphological, syntactic and semantic information, any type of dependency subtree, etc. There are consequently no restrictions on the sequence of analysis; there is no requirement, for example, that morphological analysis must be completed before syntactic analysis, that semantic considerations can be introduced only at particular levels, etc. (King 1982). Methods which are appropriate for SL analysis of one language may not be appropriate for another.

A particular emphasis of Eurotra planning is that the design should allow the easy incorporation of new advances in linguistics and artificial intelligence. At present, the system does not include 'knowledge database' components or inference mechanisms (cf.Ch.15 below). In certain respects, Eurotra is not an experimental project. It is not exploring new techniques as such, but it is attempting to unify the best of current systems – and this means adaptability and extensibility as well as robustness and reliability. There is no pretence that translations will be of high quality; it is hoped that 'raw' output will be usable unedited for many information gathering needs, but for other purposes revision will be necessary. Adequate text editing facilities are therefore essential for the final operational system.

Possibly the most inspiring aspect of Eurotra, but probably the most fraught with problems, is its multinationality: the ideal of harnessing the scattered expertise of European MT research in one grand project. It was thought to be impractical to gather together a team of experts in one place, and it was in any case politically desirable that all member states of the Community should participate in a project which was intended to benefit them all, and that they should all reap the side effects of research and development on advanced computer systems. Individual teams in each country are developing the analysis and synthesis programs for their own language, using the flexibility of the Eurotra framework to develop methods most appropriate to their language and applying any expertise they may have already in particular techniques. The only constraints are that they conform to the basic 'transfer' formalisms: dependency trees bearing at least information on surface structure functions (subjects, objects), case relations, valencies, etc.

Given the political logistics of European research it is perhaps not surprising that there have been difficulties in setting up the groups in individual countries. There was little problem with groups in Denmark, Belgium, the Netherlands, the United Kingdom and the Federal Republic of Germany. In Denmark, the centre is at Copenhagen University; in Belgium it is at the Catholic University of Louvain; in the Netherlands it is at the Institute for Applied Linguistics in Utrecht; in the United Kingdom, research is centred at the University of Manchester Institute of Science and Technology (UMIST) and at the University of Essex; and in Germany the natural choice was the SUSY team at the University of the Saar. In Italy and Greece the establishment of expert teams has been difficult; there are willing individuals in Turin, Milan, Pisa, Bologna, Corfu, Thessalonika and Athens, but the organisational framework is lacking.

Most 'political' problems, however, have arisen in France. Two MT groups have long been active, at Nancy and at Grenoble. Undoubtedly, the most substantial research has been done by GETA at Grenoble and it was natural that the GETA group in Grenoble should expect a leading role in the development of Eurotra. Furthermore, GETA had already established a cooperative project with international aspirations, the 'Leibniz' group.

In a number of respects, the Leibniz group was the forerunner of the Eurotra project. It had been founded in July 1974 by GETA on the initiative of Jean-Marie Zemb (Vauquois 1975, Boitet 1977, Chandioux 1976a). The aim was to promote collaboration between MT research teams in Europe and Canada. The participants included the research groups at Grenoble (GETA), Nancy, Saarbrücken (SUSY), and Montreal (TAUM), and some individual researchers, Yorick Wilks, Antonio Zampolli (Pisa) and Daun Fraga (Campinas, Brazil). Nearly all had at one time or another
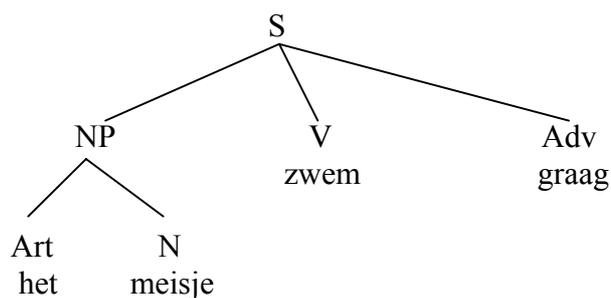
made use of the GETA facilities in Grenoble for research on their own systems (Ch.13.3). At meetings in 1975 agreement was reached on the formalism of a 'pivot language' to function in transfer components: a labelled dependency tree incorporating certain 'interlingual' elements. The computational base was to be provided by the ATEF and CETA systems developed in Grenoble and the REZO parser developed in Montreal.

Having assumed the leading role in European MT research for so many years it was perhaps not surprising, therefore, that GETA was unwilling to compromise by changing to the Eurotra approach. No other group could match its years of research experience; on the other hand, Grenoble had as yet produced no working system, and there were rumours that CNRS was withdrawing its support. It is a difficult dilemma: Eurotra owes much of its philosophy to the past research at Grenoble, but for political reasons, if for no others, it cannot permit one national group to dominate its progress.
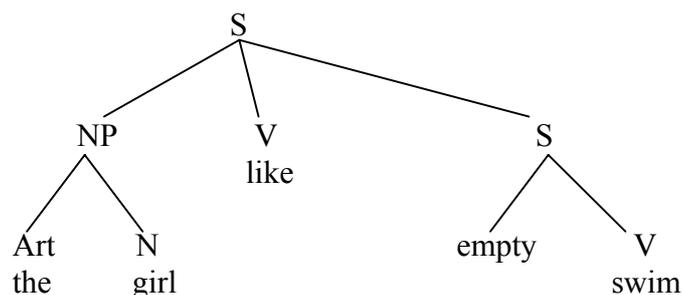
If the Eurotra project fails, then it will be more likely to be for political and organisational reasons than for theoretical impracticality. In MT research there are no longer particular problems of finding systems or procedures or forms of representation which work; there are relatively few problems attached to particular constructions or linguistic phenomena; the problem is to determine the overall model of the system: the techniques exist, but the objectives are often unclear (for example, deciding when a system is producing 'good', 'usable', 'adequate' translations) − i .e. essentially the same questions which have concerned MT researchers from the beginning.

The particular contribution of Eurotra is that it has brought into sharper focus than before the real problems of designing multilingual systems. On the linguistic side may be mentioned in particular the definition of transfer components in the context of a system for languages of common historical origin, i.e. what are the 'Euro-versals' to be (case markers, tense markers), what should be the form of input SL 'transfer' structures and of output TL 'transfer' structures, and in what respects should they differ, if at all.

Some idea of the thinking on transfer rules has been given by Krauwer and Des Tombe (1984). The example given is the translation from the Dutch *het meisje zwemt graag* (equivalent to German *das Mädchen schwimmt gern*) into English *the girl likes to swim*. A fundamental principle is that transfer representations should be (internally) valid for the languages in question, i.e. they should not be distorted for the same of supposed interlinguality. Therefore, assuming the interface representations are respectively:

```
                    S
         _____/ | _____
        NP          V           Adv
      /    \        zwem        graag
    Art     N
    het     meisje
```

and:

```
                 S
        _____/ | _____
       NP         V           S
     /   \        like      /   \
   Art    N             empty    V
   the    girl                   swim
```

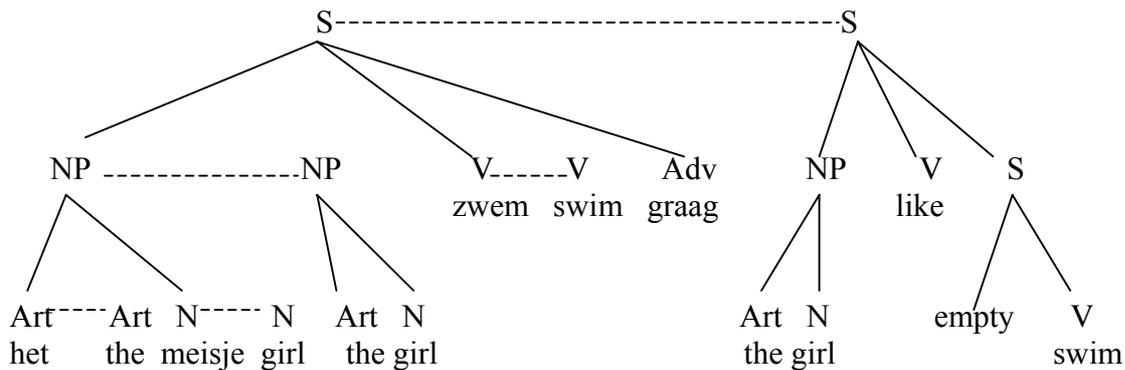then it is clear that the structural changes must be triggered by a lexical transfer rule, i.e. informally (and grossly simplified):

<subject, adverb (graag), X> → <subject, verb (like),''empty' (bound to subject), X>

The rule would operate in this way:

where the TL interface tree is built up a node at a time from SL nodes starting at the bottom and working up to the topmost (NP and S) nodes; the dotted lines indicate the transfer of each node from SL tree to TL tree.

Much of the preliminary linguistic planning (e.g. Arnold et al. 1984, Des Tombe et al. 1985) has been concerned with the specification of 'transfer' and other intermediate structures (morphological and 'surface syntactic' representations) and with the provision of guidelines for individual (national) teams on the treatment of particular types of linguistic structure (e.g. coordination, modification, compound nouns). In general terms, representations are labelled dependency trees comparable to those of GETA. Facilities for lexical decomposition are included, e.g. *unanalysability* as the tree in Fig.30.
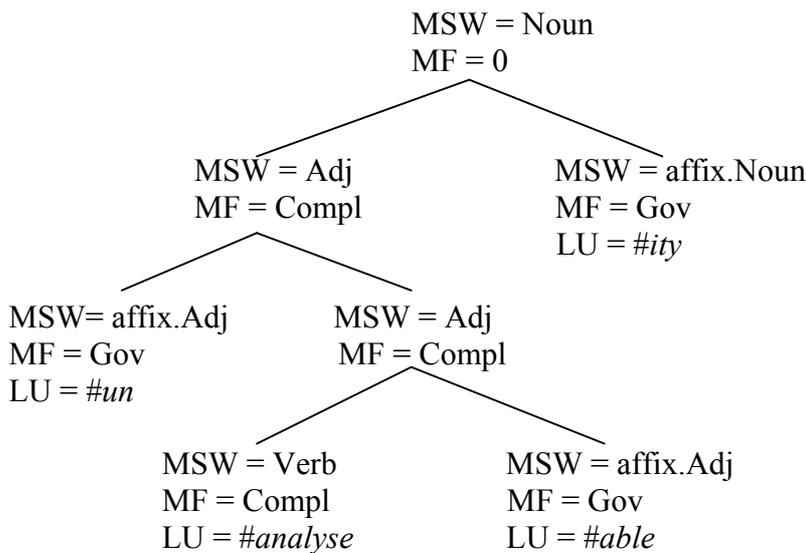

Fig.30: Eurotra lexical structure

Surface syntax trees assume relatively familiar forms, e.g. *John must go,* as in Fig.31. For transfer representations the trees can combine GETA-type mixed 'deep' and 'surface' labelling with lexical decomposition; e.g. for the sentence *He explained his position comprehensibly* the representation in Fig. 32.
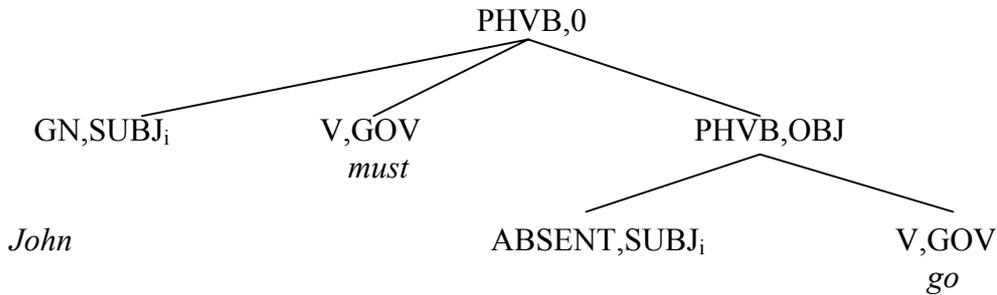

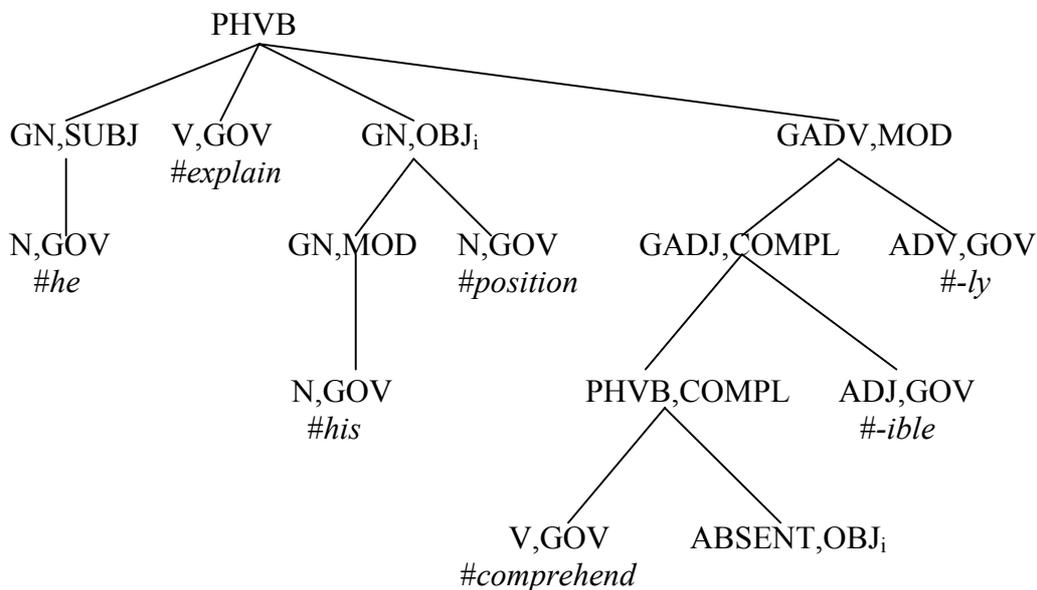Fig.31: Eurotra syntactic representation

Fig.32: Eurotra interface representation

A particular problem for a system as complex as Eurotra is that of 'robustness' (Arnold & Johnson 1984), the construction and development of subgrammars which do not fail, which can deal with illegal input (whether errors or spelling or grammar in SL texts, or invalid intermediate results produced by internal malfunctions of the system), and which ensure 'correct' output, i.e. representations which are not just well-formed but which are valid transformations ('translations') of the input representations. Such questions are linked to the development of software, which may well be the area in which the most substantial advances will have been achieved: the development of procedures for efficient handling of highly complex operations, and the development of higher-level software enabling linguists to specify structures and procedures without needing to know how the system or the programming languages would handle them (even in other systems with a strict separation of linguistic data and algorithmic processes, linguists need to know how to program.)

After five and a half years' research, sometime in 1988, the prototype Eurotra system will be tested on texts, probably in the field of information technology. By that time, it is expected that linguistic and software development will have attained a high measure of completeness and that

dictionaries of some 20,000 lexical items for each of the seven languages will have been compiled. At this point, the potential development of an operational system will be assessed.[3]

Whatever the final outcome of the Eurotra project, there will probably be little doubt that it represents a milestone in MT research. There has been no previous project of such ambitions or of such complexity of organisation. No other project has had longer or more thorough preparations. No other project has brought together the linguistic and computational expertise of so many countries in a joint international research project on such an impressive scale. All expect that Eurotra will contribute immeasurably to the theoretical foundations of MT research for many years into the future.

---

[3] The project ended in 1992. The closing report, dated February 1993, is: *Final evaluation of the results of Eurotra: a specific programme concerning the preparation of development of an operational Eurotra system for machine translation.* Brussels: Commission of the European Communities, 1994. (COM (94) 69 final)