[internal document of the European Commission, written May 1993; reproduced with author's permission]

SYSTRAN DEVELOPMENT AT THE EC COMMISSION

1976 TO 1992

IAN PIGOTT

PREFACE

On more than one occasion, I have heard it said that, if properly presented, the story of Systran could compete for TV audiences with any of the well known survival series such as *Dallas* or *Dynasty*.

Indeed, the characters involved over the years are colourful enough in themselves.

First there was Peter Toma, the idealistic, Hungarian-born inventor of the system who saw machine translation as a contribution towards achieving world peace. In recent years he has set up private university faculties in New Zealand and Argentina in those areas of the globe which are least likely to be affected by nuclear fallout. Then came Sadao Kawasaki, a Tokyo businessman who had made a few millions in cosmetics and saw Systran as a means of overcoming misunderstandings between the Japanese and the rest of the world. There was Helmut Fischer, ostensibly interested in *sanitizing* the German Systran Institut, but in fact more intent on clandestine dealings with the Supreme Soviet. And now, finally, we have Jean Gachot, an industrial valve manufacturer, who is valiantly trying to apply his experience in hydraulics to setting up a network for translation requests from French minitel terminals.

Inside the Commission, we have had plenty of colour too. Peter Wheeler put the same vigour into the early years of French-English development as he did into driving his bright red Jaguar E-type around the highways and byways of the Grand Duchy. Unfortunately, during one of our tougher periods, he was persuaded to join the competition and went on to coordinate German-French development for Logos Corporation. Theo Holtz, second in command of the translation department for several years, looked on Systran as a means of sorting out the sheep from the goats or differentiating between economy class and first class (i.e. human) translation service for Commission users. Then we

had all the division heads, directors and directors-general who either supported or denigrated Systran depending on its political acceptability at the time.

And so we could go on. We could include those who, on purely linguistic grounds, were convinced from the start that Systran could never translate as it had no *grammar* or those who, on purely informatics grounds, knew that Systran was doomed to disappear as it was based on antiquated *Assembler* code or on its *own* lexical data base.

We should also perhaps mention those of our colleagues and consultants who saw a more promising future for machine translation including those who believed that some day the majority of translators could and would become MT post-editors (whether in 10, 20 or in 30 years time).

Finally, I must, I suppose, turn back to my own interest in Systran. I have begun to wonder exactly what made me tick for so long. How could I continue to work on a project which for years on end appeared to have so little chance of success, particularly when closely confronted with hightech approaches on the R&D side (cf. Eurotra - a machine translation system of advanced design)?

In retrospect, I find it difficult to answer the question. All I would say is that when any specific problem occurred, it seemed to me that the computer could usually be programmed to provide the answer. The first example came right at the beginning of my Systran experience, back in February 1976. About 20 pages of text had been "translated" from English into French. One of the output sentences started off *sur l'autre main*, a word-for-word translation of the English o*n the other hand*. I can remember envisaging dictionaries of hundreds of thousands of idioms, containing authentic solutions to this kind of problem. Here, the appropriate translation would have been *d'autre part*.

Today, of course, we have such dictionaries. We also have an infrastructure and a technological environment which make it *acceptable* to use computers to translate. Furthermore, we have a very much greater need for rapid, medium-quality translation and, last but not least, we have a management structure consisting of pragmatists rather than classicists or linguistic purists.

At the time of writing, I have just received statistics showing that almost 1500 texts averaging six pages each are now being translated every month. There is reasonable expectation that in the Commission alone that volume will double over the next half year and that the other EC institutions will soon start to offer the Systran service to their own staff on a free-access basis.

While much remains to be done on further quality improvement, on the incorporation of new language pairs and on upgrading the informatics infrastructure, there now remains no doubt that Systran is here to stay.

Although I sincerely hope that its development will continue on the same steady and proven course, it has become clear that my main contribution, that of fighting for the survival of the system during its more difficult years, is no longer required. It has evolved from childhood through adolescence to adult life and has, I believe, become robust enough to fend for itself in the future.

Perhaps objectives and development strategy will change. Europe is a growing entity with ever more complex communication problems of its own. I hope,

however, that Systran management at the Commission will not pass entirely from the linguist to the professional project leader in the same way as it has in the commercial environment.

While machine translation is very different to human translation, it still requires the same basic skills for its development, integration and use. These include practical experience in translation work, a gift for breaking down translation problems into manageable categories of linguistic phenomena and, last but not least, a feel for what the user really needs in terms of quality, speed, layout and language combinations.

If Systran can continue to be developed along these lines, it will certainly thrive for a number of years to come.

CHAPTER 1

THE BEGINNINGS

The origins of Systran

During the 1950s and 1960s, the U.S. authorities became ever more concerned about technical progress in the Soviet Union. It was widely believed that if only the hundreds of thousands of pages of technical literature could be translated into English, the Americans would be able to foresee developments in the area of space technology and atomic physics, not to speak of the more general area of defence.

Both government and industry therefore began to invest in machine translation R&D. Much of the work was centred around Georgetown University in Washington, DC, where linguists tried to devise methods of coding language knowledge. These efforts eventually led to IBM's Mark 1 and Mark 2 systems which were used by the USAF and to the Georgetown system which was used by the U.S. Atomic Energy Commission and later by the EC's Joint Research Centre in Ispra. Both, of course, were for Russian-English translation. They continued to be used until the late 1960s.

Peter Toma, who had left Hungary after the Second World War, worked for a time as a liaison officer between the U.S. Third Army in Bavaria and the Hungarian Red Cross. He quickly realized the importance of improving communications between Russian and English and decided, as a first step, to master Russian himself.

After moving to California in 1956, he became actively interested in applying his practical knowledge of language to his increasing interest in computer technology, his aim being to produce a pragmatic machine translation system. Toma, unlike most other scientists of his day, did not believe that *linguistics*

could provide a workable solution to the computerization of language. He was convinced that language processing had to be adapted to the capabilities of the computer rather than the other way round.

So it was that he started to apply his hypotheses to a number of MT initiatives in the late fifties and early sixties while operating computers at the Californian Institute of Technology in Pasadena. First came Autotran, then Technotran and finally, with the advent of the IBM 360 system in 1963/64, Systran.

According to Toma, the experts responsible for the devastating ALPAC report in 1965 set out to defend linguistics rather than MT and thus had no qualms in criticizing his pragmatic approach. Although ALPAC put a complete stop to U.S. funding for machine translation research, Toma was lucky enough to convince the Deutsche Forschungsgemeinschaft in Bonn of the potential of his approach. He was thus able to continue development in Germany and by 1967 had an operational prototype running on an IBM 360-30. Final debugging was carried out at the University of Bonn on a 360-50 between September and November 1968.

The USAF Foreign Technology Division immediately recognized Systran's advantages over other MT systems and installed a first version in 1969. The Russian-English system proved a great success and has been used and further developed ever since.

Further support for Systran came from NASA in 1973 with the Apollo-Soyuz project. In order to provide for written communications between Russian and English in both directions, NASA financed the development of a modest English-Russian system.

In 1974, Toma applied the results of the English analysis program to a prototype for English-French, initially tested by the Canadian federal translation office as well as by the Canadian headquarters of both Ford and General Motors. Of these three, General Motors was the only one to continue support.

The Commission's interest

There are two rather different accounts of how Systran came to be adopted by the Commission, one which gives most of the credit to the Commission for being so far-sighted in looking for an MT system in those early years, the other ascribing the move to Toma's own initiative in trying to find a European market for his system.

The most probable seems to be the latter. We do know for certain that Toma came into contact with a number of Commission officials and experts in 1975 at a time when he was organizing demonstrations in Switzerland at the University of Zurich. The Russian-English version which, by then, was fully operational was of little direct interest to the Community but a small English-French prototype had been developed.

Not entirely by coincidence, Herbert Bruderer, initially with the support of the Swiss Air Force, was also completing a world survey of machine translation systems which gave considerable credit to Systran. Furthermore, at that time

DG XIII was promoting the Euronet initiative which was aimed at introducing standardized access procedures for documentary database interrogation throughout the Community. In this connection, multilingual thesauri were being developed and it may have seemed logical to look into MT technology as a means of providing ever richer multilingual facilities.

As a result of all this, the experts working with DG XIII in the area of information science and technology (IDST) approved pilot projects with Systran and Titus which were to start at the beginning of 1976. The choice of these two systems was based on the findings of the Bruderer study, namely the fact that Systran was the only operational full-text system covering translation between any of the Community languages (English into French) while Titus, a limited syntax system developed by the French Textile Institute, was operational between English, French, German and Spanish in all directions. Also in Systran's favour in 1975/76 was the fact that the Commission had an IBM 360 machine which could run the software without modification.

The initial objective, by the end of 1976, was to test the usefulness of Systran by applying it to the English-French translation of a documentary data base of Community interest as well as to investigate its extendibility to another language pair (French-English). As for Titus, the main aim was to extend the extremely rigid writing rules to a slightly more flexible set of syntactic patterns. The contract with Titus was in fact cancelled in mid-stream when it became clear that work was not being carried out on schedule. Systran thus remained the only system to be subjected to proper development, testing and investigation.

Start of practical work

By the end of 1975, Loll Rolling of DG XIII had managed to encourage his superiors (Georges Anderla, director, and Raymond Appleyard, director general), to sign an initial contract with Toma's company, World Translation Center, of La Jolla, California. WTC was to enhance the pilot English-French system in accordance with the Commission's requirements and develop the beginnings of a French-English system.

The contract also provided the Commission with certain user rights (use by the Community Institutions, government agencies of the EC Member States and in connection with Euronet) and property rights (in particular ownership of the dictionaries developed by the Commission).

Toma had recommended that the Commission should rely on translators to coordinate development work. I was fortunate enough to be selected as one of a team of six chosen from the English and French translation departments in Brussels and Luxembourg. Work began in February 1976 for an initial period of two months.

The aim was to develop dictionaries designed to translate abstracts from the Food Science and Technology data base which was chosen on two counts: first, because of the interest in applying Systran to the translation of data bases accessible via Euronet and second, as a result of the Commission's more general interest in the closely related field of agriculture.

For practical reasons, the team was housed in the old computer centre in Luxembourg (now known as the Bâtiment Cube) where the 360 machine was installed. One of Toma's head linguist programmers, Joann Ryan, coordinated the initial work and provided basic on-the-job training in dictionary coding.

Unfortunately, despite what appeared to me to be fairly positive progress, at the end of the two months, all the other translators opted to return to normal translation work as they saw little or no future in the application of Systran to translation at the Commission.

As a result, I was left on my own for the next few months, making every possible effort to enrich the dictionaries and liaise with Toma's staff in California in an attempt to improve the general quality of output. Later in the year, we managed to hire two external linguists under contract to help with the dictionary coding.

These efforts appeared to pay off as the evaluation conducted by Bureau Marcel van Dijk at the end of 1976 was fairly positive, particularly as it recommended a continuation of the development effort.

This of course provided a basis not only for improving the English-French system but also for embarking on the development of new language pairs. It also led to the hiring of more external staff, initially under individual contracts, later under the cover of contracting companies.

When in 1978 and 1979 further evaluations were carried out, there could no longer be any doubt that substantial progress was being made. It was, however, going to take over ten more years before the approach could be implemented for general use within the Commission.

CHAPTER 2

THE SYSTRAN APPROACH

Systran is often referred to as an *all-purpose* system. In other words, it was not designed specifically to deal with any particular domain, document type or sublanguage despite the fact that certain users - in particular the Xerox Corporation - tailored it down to process limited vocabulary and syntax.

The all-purpose approach has both pros and cons. Among the advantages is the innate possibility of applying the same system to widely varying types of translation ranging from minutes of meetings to highly technical research reports or maintenance manuals. The major disadvantage is that a great deal of time and effort has to be devoted to a choice of translation equivalents which will prove the most valid in a wide range of circumstances. Furthermore, on the parsing side, it is difficult to take short cuts, as a correct solution in one environment might prove to be entirely false in another.

It is important to remember that, unlike most of the other early machine translation developments, Systran was never based on any particular linguistic theory. Indeed, Peter Toma states clearly in his patent that his approach is designed to reduce the translation process to the possibilities and limitations of computer logic. This, in itself, does much to explain the down-to-earth Systran approach.

Those who have worked in other MT research environments or in the more general area of computational linguistics are often surprised to discover that Systran is not based on any specific linguistic theory or grammar. Some even go so far as to maintain that any system which does not contain clearly expressed grammar rules for the languages involved in the translation process cannot be expected to produce satisfactory results.

In the following pages, I shall attempt to describe in fairly straightforward terms how Systran actually functions and how it is that, in the absence of linguistic niceties, unparalleled levels of translation quality can indeed be achieved.

The basic Systran components

Like most operational MT systems, Systran can be broken down into three main components: the basic system, the linguistic programs and the dictionaries.

The basic system supports and controls the other two components. It is written in 360 Assembler, some 100,000 lines in all, and unlike the linguistic programs it is rather opaque in that few macro instructions are used.

It is important to remember that from the outset (1968-69), the major goals included speed and effective use of data storage. This is indeed one of the reasons why Systran still continues to execute so quickly and why it consumes relatively little processing capacity.

The basic system contains various programs and routines common to all language pairs. These can be grouped into a number of subsystems including:

- text interface and manipulation programs;
- dictionary update and management;
- translation dictionary creation and access;
- translation control, including dictionary lookup;
- miscellaneous utilities;
- common subroutines.

These items are generally language-independent and include all the main programs that are called from job control level.

Both the Systran production system and the development system are thus almost completely self-contained. In particular, the dictionary access modules form an integral part of the software and do not need to rely on any independent data base system. This has the obvious advantage that the system can be installed in one piece without the need for any third-party software or high-level programming-language facilities.

Experience has shown that the basic system is extremely reliable, only requiring relatively minor modifications over the years. In the production environment, it has proved to be robust, seldom causing interruptions or system failures.

The disadvantage of this approach is the fact that it is indeed based on IBM 360 architecture in the MVS operating system environment and is thus not easily transportable to other platforms such as Unix or DOS. On the other hand, machines capable of running the VM system (e.g. OS/2 or RISC) can also run Systran in a desktop environment with minor conversions.

It would therefore appear that for running a machine translation production facility under which several language combinations and several hundreds of requests per day are to be processed, the existing informatics approach is likely to remain the most reliable and the most efficient for some time to come.

The linguistic programs

While the linguistic programs also rely on the 360 Assembler language, they differ from the basic system in that they are written almost entirely in an easy-to-learn, specially adapted macro language. As most of the macro instructions were specially written for Systran, they have distinct advantages over general-purpose high-level languages such as Fortran or C in that most of the sets of instructions which are commonly used in the translation process are indeed covered by an appropriate macro.

Examples of the types of macro used are CW (current word), CKCAT (check semantic category) and SCANL (scan left within the sentence). It is thus comparatively easy for linguists without any previous knowledge of programming to learn to program Systran linguistic routines. Furthermore, although the system is Assembler-based, most of the linguistic work can be undertaken with only a very superficial knowledge of *low-level* Assembler programming.

The linguistic programs are called sequentially. Analysis of the source language (which is now completely independent of the target) consists of six main passes comprising grammatical homograph resolution, clause boundary analysis and various levels of syntactic dependency establishment. The transfer program is restricted to so-called lexical routines which deal with the bilingual processing of particularly difficult structures. And at the target level, morphology, syntax and word order are computed for the language in question. This may sound very simple but, as we shall see later, it is in fact complex.

The dictionaries

The Systran dictionaries are unique in the field of machine translation on three separate counts: their functionality, their coverage and their size.

Very roughly, they can be classified into two main types:

- those containing one-word entries (often called *Stems*);

- those consisting of multi-word expressions (idioms and *Limited Semantic* or *LS* entries) or of single words in prespecified contexts (*Conditional Limited Semantic* or *CLS* entries).

At source language level, each one-word entry contains morphological, grammatical, syntactic and semantic information as well as an indication of potential part-of-speech homography. Multiword entries range from string expressions of various types to contextual rules which examine the attributes and dependencies of any or all elements. At target level, meanings are assigned on the basis of criteria varying from the most generally useful equivalent for a one word entry to specific translations required for a predefined context based on a string expression or on contextual conditions.

Consideration can also be given to a subject field or document type in assigning so-called *topical glossary* meanings which, at any level of the dictionary, will supersede the general meaning provided the corresponding subject field parameter is used as part of the translation request.

The modular approach

Initially Systran was considered to be of bilingual design, the packages being developed in the form of language pairs. There were however always very clear divisions between analysis, transfer and synthesis.

It was on this basis that we decided in 1986 to develop fully modular components irrespective of the language combinations. For example, the same

English analysis program and source-language dictionary is used for translation into French, Italian, German, Dutch, Spanish, Portuguese and Greek.

This has had a very positive effect on quality improvement as certain language combinations require a far deeper level of analysis and lexical transfer than others. The discipline of having to cater for the case, gender and word order requirements of German has almost accidentally had the effect of improving many of the almost acceptable results obtained, for example, in English-French translations. And the identification of technical noun phrases for one language pair has provided an immediate basis for identifying potential translation problems in the other pairs.

Last but not least, the linguists working on any given source language have a far better awareness of the various phenomena requiring attention.

CHAPTER 3

LINGUISTIC PROBLEMS

Text versus grammar

One of the main surprises for linguists when they are first confronted with machine translation systems is that real texts contain many more phenomena than those to be found in even the most comprehensive grammar books.

Grammars tend to be limited to linguistic descriptions of words and the behaviour of words within a sentence. At the risk of oversimplification, they contain notions of gender, number and case for nouns, tense, person and mood for verbs and explain how syntactic structures can be created in context, for example: subject + verb + direct object + adverb.

Information of this type is, of course, extremely important for describing linguistic behaviour for machine translation, but it is far from complete.

Among the phenomena not sufficiently described in grammars are the behaviour of figures or digits in text, the effect of formatting or page layout on sentence structure, the nature of grammatical homography or the impact of punctuation on meaning.

In machine translation, factors such as these require special attention as they can play a vital role in providing criteria for dependable parsing.

More often than not, we were to find that we were breaking completely new ground when confronted with phenomena of this type. The software therefore had to be developed on the basis of trial and error, usually making use of fairly large and representative text corpora.

Letters, figures and digits

One of the main problems with letters and figures is that they can play completely different roles depending on how they are used in context. The diversity of usage only becomes apparent when a wide variety of document types and page presentations is to be analysed.

Some typical uses of figures include:

- quantities and measurements (10 ships, 3 inch gap);
- date structures (4 July 1992, 4th July, July 4th, July 4, 4.7.92, 4-10 July, 4, 5 & 8 July);
- years (1990, 1990-1992, 1990-92, 1980s, 70s & 80s, '70s and '80s, 1990 statistics, Europe 2000);
- paragraph enumerators: 1., 1.2, 1.2.23, 1a, 1), (2), (35);
- decimals (23.5, 1.2, 12,312.5);

- currencies ($13, £1,234);
- percentages (12%, 12 to 15%, 12% increase).

Not only do these sequences and structures require adequate linguistic descriptions, they also need to be supported by routines that can differentiate between items which take on different roles in context. For example, 2.3 can either be a decimal or a paragraph enumerator affecting not only the analysis of the source text but also the nature of the translation (the decimal 2.3 in English would become 2,3 in French whereas the enumerator would remain the same).

The problem is compounded by the fact that literal figures often appear in enumeration with digits depending on the conventions used by the author (e.g. seven to 12 percent). Furthermore, the Roman numerals (I, V, X, C, M or i, v, x, c, m, etc.) are often difficult to differentiate from letters or initials.

In the case of letters, difficulties can be encountered in differentiating between:

- initials: A.B. Brown;
- abbreviations: U.S. Navy;
- enumerators: I. Introduction;
- Roman numerals: DG I, Appendix iv;
- real words: I am, A long time ago;
- chemical elements, K, I, P;
- countries or nationalities: B, F, D.

Here again, most development work was based on contextual analysis of texts actually submitted for machine translation. As in most other levels of Systran processing, it is only fair to point out that frequency of occurrence often plays a very important role in the decision-making process. In other words, if one of the possible solutions occurs only in a small minority of cases, it is dealt with as an exception and will be obtained only when precise contextual criteria are met.

Formatting

Page layout and document formatting in general are concepts which are of vital important in MT processing for the following reasons:

- Layout often gives clues as to whether sequences of words are to be handled as sentences or rather as headings or titles.

- Conventions such as an introductory clause followed by a number of indents often require special treatment at the linguistic level as they may well constitute a complex sentence rather than a string of individual units of text.

- Re-establishment of page layout at the target level will prove useful to the user, particularly if formatting code can be successfully reincorporated as this will accelerate on-screen post-editing (wrap around, etc.).

Treatment of many of the phenomena which occur in running text was always dependent on a careful analysis of different document types.

Simplistic rules such as "The end of a sentence can be identified by the presence of a full stop" were found to be of little help in practice. Far more important were the correct identification of initials and abbreviations (which often end in a full stop themselves), the analysis of hard carriage returns and the processing of both text and figures contained in various types of table (indexes, statistics, agenda items).

Even today, many translation errors continue to occur as a result of formatting phenomena. Frequently such errors are due to unusual changes in column length or to the use of hard carriage returns when automatic wrap around should have been allowed to occur. In general, however, performance is good. Indeed, when normal typing conventions are used, a reliable level of sentence definition can be ensured.

Grammatical homography

Of all the linguistic problems to be tackled in text analysis for machine translation, the establishment of grammatical homography has certainly proved to be the most difficult.

A grammatical homograph is a word form which can take on two or more part-of-speech values depending on the particular context in which it is used. In English, *light* can behave as a noun (a bright light), a verb (to light a cigarette) or an adjective (a light weight). In French, *en* can be a preposition (en France) or a pronoun (il s'en va).

To a greater or lesser extent, all the source languages with which we have dealt (English, French, German, Spanish) have caused problems in this area but there can be no doubt that English has been particularly difficult to deal with, particularly as in a typical English sentence over 40% of the words are likely to be grammatical homographs.

For example, the homographs in the last clause are *as* (preposition, conjunction), *in* (preposition, adverb), *a* (article, initial), *English* (noun, adjective), *sentence* (noun, verb), *over* (preposition, adverb), *the* (article, adverb), *words* (noun, verb), *likely* (adverb, adjective), *to* (preposition, infinitive particle).

In the absence of reliable information on the behaviour of grammatical homographs in context, extremely complex routines had to be developed from scratch or on a trial-and-error basis.

Initially, the strategy was to attempt to disambiguate each word in the sentence sequentially from left to right, using the result obtained for the first word as part of th e information needed to resolve the second, and so forth. More recently, particularly for English, the approach has been to sort out all the reliable values first and then to go back to those items in the sentence which have proved more difficult to process.

For some part-of-speech sequences (e.g. past tense *vs.* past participle), several thousand lines of codes have had to be programmed for reliable results to be obtained.

The relatively successful resolution of grammatical homographs is certainly one of the main reasons why Systran competes well with other machine translation systems. Indeed, as far as the parsing of English and French is concerned, it still seems to provide better results than any other piece of software.

Punctuation

It is widely recognized that, for certain languages and particularly for English, clear rules of punctuation are sadly lacking.

What is perhaps more difficult to establish is exactly how different punctuation marks affect the parsing of running text in practice.

While the problem is a fairly general one which occurs with most types of punctuation, there have been certain items which have caused particular problems.

Let us first take the colon. The examples of use given in grammar books and dictionaries for languages such as English and French seem to indicate that it is used more often than not within a sentence for introducing a relationship, an enumeration or a subordinate clause. While these uses do indeed occur to some extent in running text, by far the most common use of the colon was found to be that of an end-of-sentence indicator. This information is of course of vital importance in the Systran context.

Once this had been established and the colon's default value was changed from that of an enumerator or clause opener (similar to the comma) to that of a sentence boundary marker, the quality of translation improved considerably.

Similar phenomena were to occur in relation to the question mark, which was often omitted at the end of a direct question, or the slash, which was found to be used frequently as an enumerator between nouns or noun phrases (e.g. *nitrogen/trace element analysis* or *hardware/software requirements*).

* * * * *

The examples of "grammatical" phenomena given above represent but a small but typical subset of the types of problem which have cropped up over the years.

Fortunately, we have usually found that Systran has been able to cope quite well with the incorporation of computerized logic to solve the various problems encountered. What is not so easy, of course, is to retrieve from Systran coherent descriptions of all the various phenomena which have required special attention in the analysis of written text. The rules or algorithms do, of course exist, but they are often spread over different levels of analysis, transfer and synthesis and can depend on various types of dictionary entry.

What has always been of primary importance is performance rather than academic exercises in linguistics.

Those interested in investigating the results of Systran analysis for various text types, page formats or communication environments could of course examine the results and may, in some cases, be able to draw valid conclusions for other language processing ventures.

CHAPTER 4

LANGUAGE PAIRS

In this chapter, I shall try to provide information on two questions which often arise in discussions on Systran. These are:

- How or why were the current 16 language pairs selected?

- What is the relative difficulty of developing different language combinations?

Finally, I shall comment on the potential need for further language pairs and the difficulty of developing them.

Choice of language pairs

The repertoire of languages in the Commission's Systran system is a result of a number of factors ranging from the availability of existing developments to the need for new ones as perceived by internal staff and/or external experts.

It was, of course, clear from the start that English and French would be key languages. It was therefore not surprising that the English-French prototype which was initially delivered to the Commission was not only fully tested but was subsequently extended to cover a variety of subject fields and text types. Similarly, French-English, which was first developed at the Commission's request as a means of demonstrating the flexibility of the Systran approach, was quickly to become a key pair.

Both of these were initially developed with Euronet in mind. As a result, initial development work was based on documentary data bases such as FSTA (Food Science and Technology) and Predicasts (business information) for English and CRIF (metallurgy) and CNRS (research) for French. While these provided a quick and dirty means of building up the dictionaries, they were very different from the type of text normally handled in the Commission services.

The next pair to be developed (1978) was English-Italian. The aim here was twofold:

- to test how easy it was to add a new target language to an existing source language;

- to cater for a widely used target population (native speakers of Italian were said to be more numerous than any other group in the Community), partly as a result of opinions expressed by our advisory committee CETIL.

In fact, English-Italian proved comparatively easy to develop in view of the similarity between French target and Italian target. The English source dictionary could also be used as a model for English-Italian.

Despite the Commission's feeling that some language combinations containing German should have priority, the opinion of the CETIL experts was that the newly adopted Eurotra project would soon provide operational systems for German.

By 1982, however, it had started to become clear that Eurotra would not be operational within the foreseeable future. It was therefore finally decided to embark on German. That year two contracts were signed, one with Toma's company, WTC, for French-German, which required development more or less from scratch (combining the French analysis module with German synthesis), the other with the Systran Institut, for a development based on their existing English-German system.

Fears expressed on the difficulty of developing German proved only too true. Mainly as a result of German word order, but also because of its general syntax, it was not possible to achieve results comparable to those for the existing combinations of romance languages and English. Indeed, even now ten years later, the quality of the English-German and French-German systems is generally far from satisfactory.

In 1984, we were to have two further pairs developed, this time on the recommendation of our experts. These were English-Dutch and French-Dutch, principally aimed at the Belgian market. Interestingly, this was the first time in the history of Systran that a new target module had been developed for use with more than one source language. It is a good example of how we were moving into an ever more modular environment at that time.

Dutch was by no means an easy language to develop but progress was quicker than with German. This was partly as a result of the fact that the Commission had full control from the start whereas the German target module had first been developed in a rather unconventional way for users in Germany.

Even at this stage, little interest was being expressed in Systran by the translation service despite the fact that translators had been involved in pilot projects involving post-editing since 1980. For this reason, DG IX which was responsible for translation, was always hesitant to make proposals on new language pairs as it was felt they would be of little or no use for internal purposes.

Mainly as a result of the enlargement of the Community to Spain and Portugal, English-Spanish and English-Portuguese were added in 1986. These presented no particular problems but, unlike many of the other pairs, were not immediately applicable to any well-defined user population.

The next development was based purely on in-house requirements. By 1987, a number of Italian translators had begun to use the English-Italian system. In general, they were quite pleased with the results. Indeed, they felt that if satisfactory results could be obtained for a combination of two rather different languages (English and Italian), then results were bound to be even better for French into Italian.

As we already had a French analysis module and an Italian synthesis module as well as comprehensive dictionaries for French-English and English-Italian, we were able to create a fairly good French-Italian system almost completely automatically. Indeed, the two languages proved so similar that development sped on and it was not long before the results for this pair were substantially better than for any other. For some subject fields and document types they were almost comparable to human translations.

At the request of the Greek authorities, we began in 1988 to work on English-Greek under an arrangement whereby the development costs were split between the Community and Greece. The reason for this approach was the Commission's stance that this pair was by no means a priority for internal use.

Our development of Spanish as a source language, which began in 1989, was based not only on our understanding that the demand for translation from this widely-spoken language would be fairly high but also on the fact that a Spanish-English system was already under development for the U.S. Air Force. We therefore decided to develop both Spanish-English and Spanish-French in parallel.

Finally, in 1989, we created a French-Spanish system on much the same basis as the French-Italian development.

Language pairs in retrospect

Looking back over the years, we have reason to question our strategy on language pair selection with the proviso that it is, of course, only too easy to be wise after the event.

While the initial developments, English-French and French-English, have now begun to prove useful, it would have been more logical to start with easier combinations such as French-Italian. Results would have been better after relatively short periods of development and the reputation of machine translation - particularly its usefulness for translators - would not have suffered as much as it did.

Other combinations which could have given better results than those obtained in the early years might have included German-Dutch, English-Danish and Italian-French. Many of the structural problems could have been avoided while terminology would have been far easier to research than for combinations of languages belonging to different families.

On the other hand, the actual volume of work for these pairs would certainly have been less than for English-French and French-English. On this count, then, the choice was a sensible one.

Another mistake which was made in the early years was to underestimate the importance of German. Although an extremely difficult language to process, German is one which constantly causes difficulties in the Institutional environment as it is seldom understood well by non-native speakers.

Many user departments, including the translation services, have commented that machine translation would have been far more useful if it could have been developed to translate from the minor languages (Danish, Dutch, Greek, Portuguese) into the major ones (English and French). This strategy has not yet been implemented, mainly because the volumes of text to be translated seem very low. Furthermore, most of the texts to be translated come in from the Member States on paper rather than in machine-readable form and would therefore cause input problems.

If the current in-house production statistics are any indication of the need for translation from minor languages, it can be seen that even from Spanish - which is hardly minor - the throughput is only about 3% as compared to about 50% for French and 43% for English.

Future priorities

After almost 17 years, we have a good idea of the relative time and cost of developing combinations of various families of languages up to a given quality level.

The simplest case is that of developing any combination of romance languages. On the basis of results obtained from French-Italian and French-Spanish, we can estimate that, given the existence of sizeable dictionaries at the source and target levels, French-Portuguese could reach usable quality within about a year with a staff investment of two man-years. The same could be achieved for Spanish-Italian and Spanish-Portuguese if it were considered worthwhile to initiate a development for which the throughput would probably be fairly low.

Next, Italian could be developed as a source language for translation into French (and, if necessary, Spanish and Portuguese). Good results could probably be obtained in less than two calendar years (three to four man-years). Even Portuguese as a source could be developed just as quickly for translation into any romance target language if translation demand so justified.

Given the current reliability of English analysis and the structure of Danish, an operational English-Danish pair could be developed in about one year. Alternatively this pair could be acquired from Gachot's Californian company where systems exist from English into all the Scandinavian languages (as part of the Xerox development).

Only later would I recommend embarking on the analysis of Danish, Dutch or Greek as source languages as the cost could be high (seven to eight man-years for each) and the results could be disappointing. Of course, if these projects were to be cofinanced by the Member States, there would be an added incentive.

CHAPTER 5

DEVELOPMENT CONTRACTS AND STAFFING

Initially, Systran was introduced by Toma and his staff as a package which could be maintained and further developed in-house by a combination of linguists (translators) and computer experts (DG XIII programmers). Toma emphasized from the start that the bulk of the development work would be of a linguistic nature.

While the in-house scenario proved possible in other environments such as the Xerox Corporation and General Motors, the Commission turned out to be a special case for the following reasons:

- Most of the translators originally assigned to the project were not interested in remaining with it.

- While there was some knowledge of IBM 360 Assembler among the programming staff, the general view was that high-level languages should be used instead. There was thus little incentive for individuals to waste much of their time on Systran.

- The project was not considered important enough for allocation of in-house staff resources on a long-term basis.

There was therefore little choice but to rely on outside assistance. This took on a number of different forms over the years.

Freelance contracts

The main priority at the beginning was to build up the Systran dictionaries, first for English-French and soon after for French-English. The corresponding work on the program was handled by means of service contracts with WTC, Toma's company in California.

At that time, procedures existed at the Commission for concluding contracts on an individual basis.

With the help of a consultant, advertisements were placed in papers in Luxembourg and Brussels for linguists with a command of English and French. The response was good and we were able to invite about 15 candidates to an initial training course.

We were fortunate in that about 12 proved suitable for the task. We therefore drew up individual piece-work contracts, remuneration being based on the number of dictionary entries submitted each month.

The results were surprisingly good in terms of both quantity and quality. Indeed, many of the staff volunteered for further training and reached a high-level of proficiency. What is more, the cost of handling this type of freelance work proved to be substantially less than the same degree of development handled under the service contracts which came later.

Service contracts for local maintenance

Owing to the Commission's decision in 1977 to discourage freelance contracts in favour of service contracts carried out by companies, we put out a tender for the work involved.

The company selected for the first year was the German branch of the Franklin Institut.

In 1979, the German Systran Institut, which had been created with Toma's blessing, received the assignment. Their involvement was to last until 1984 when our relationship was terminated as a result of clandestine negotiations with Moscow.

From 1984 to 1990, the maintenance and development work was awarded to Informalux, mainly owing to the fact that they were able to provide adequate computer capacity.

Since 1990, following a new call for tenders, the main contractor has been the Luxembourg company, Telinfo.

Service contracts with WTC

From the beginning, the Commission had insisted on obtaining the entire Systran program in source code.

This proved to be an excellent decision as it was soon to provide us with the possibility of undertaking not only dictionary development, as was the case with most other users, but also to work on improvements at all levels of the system.

Initially, however, work was split roughly into two parts: dictionary enhancement was handled locally in Luxembourg while programming improvements, including the development of new language pairs, were contracted out to Toma.

During the period 1976 to 1985, a total of nine contracts amounting to the comparatively small sum of US $590,000 were concluded with WTC, mostly for development of new language pairs but also for on-going work on existing components, in particular English analysis for which a high level of competence was available in the person of Joann Ryan.

It was only when Toma created the German Systran Institut (see below), to which he gave full responsibility for Europe, including the Commission, that the transatlantic connections began to suffer.

In general, though, the work carried out by WTC corresponded very well to the Commission's requirements, in that new languages pairs could be developed quickly up to a certain basic level before being integrated into the Luxembourg environment. Jeanne Homer of WTC became expert in developing basic target-language software for any language we chose.

Follow-up work in Luxembourg

Once basic developments had been undertaken by WTC, we could rely on the language knowledge of our local contractual staff to make the necessary further improvements in close coordination with the Commission officials (mainly translators) assigned to the project.

Some of the staff have worked under a series of maintenance and development contracts and have gained a high level of expertise and experience in various aspects of the system.

In particular, Giuliana Usuelli has played a vital part both in dictionary coordination and in French analysis, Juan Paez, working on English analysis, has been responsible for a number of rationalizations and innovations, Antonella Moruzzi has adeptly mastered all the levels of systems programming, while David Broman, Pit Urhausen, Alfiero Severini and Alberto Fontaneda have made lasting contributions to the development of the linguistic routines. Pierre Thillen, who started his Systran career as a linguist, has in recent years channelled his unceasing efforts into managing a team which now numbers thirty-five.

Other development contracts

The maintenance and development contracts in Luxembourg combined with the work undertaken by WTC in La Jolla are together responsible for most of the progress achieved over the years.

It should not be forgotten, however, that an important contribution was made by the Canadian company WTCC (World Translation Company of Canada) in 1979 when we acquired their *Systran II* software. This consisted mainly of peripheral programs designed to provide an interface with word processing systems. Apart from the *NATEX* and *SETUP* routines which covered formatting algorithms, it also contained enhancements to the dictionary structures and the various levels of printouts for debugging and development purposes.

WTCC went out of business shortly after this contract was concluded. They will be remembered for their realization that if Systran was to be brought into general use, it would have to be on the basis of user-friendly interfaces with word processing systems and telecommunications. Toma, unfortunately, had not had the foresight to recognize this.

Also in 1979, a contract carried out by Veronica Lawson on patent translation contributed to Systran's development, not only by incorporating the jargon of patent texts but also on extending its ability to deal with long sentences (up to 400 words).

The Commission's team

This chapter would hardly be complete without a word on the Commission's Systran coordination staff.

The team grew more by accident than by design as translators expressed interest in working on the system. Vague agreements between the translation services and DG XIII usually provided a basis for staff to be allowed to work under those responsible for project coordination and management. The precise status of the individuals concerned was unfortunately never made clear.

As I pointed out earlier, of the six translators originally assigned to the project in 1976, I was the only one to remain.

It soon became obvious that the task of coordinating development between California and Luxembourg as well as monitoring specific results was too much for one person, particularly when we began fully-fledged development of a second language pair in 1977.

We were fortunate enough to find another Systran enthusiast in the person of Peter Wheeler, also an English translator, who came to grips very quickly with the workings of the system and the priorities for improvement. He devoted most of his efforts to the French-English development and did much to encourage the involvement of English-speaking translators. In addition, he proved to be a tremendous asset in our efforts to introduce a word processing infrastructure. We were sorry when, for personal reasons, he chose to leave the project after a few years.

Once the English-Italian development began to provide results, Delfina Campanella joined the team. She soon saw the extent to which Systran could be an asset to translators and spent most of her time and effort on taking account of translators' needs. Once she had been able to encourage translators to make use of English-Italian, she concentrated on French-Italian. Here, the results were astoundingly good and showed clearly the extent to which machine translation could be applied as a tool to assist in normal translation work.

The high quality of the English-French system is to be ascribed not only to the initial efforts of Bernard Lavorel but in particular to the intense involvement of Francine Braun who has striven to incorporate a multitude of terms and stylistic improvements.

The difficult task of coordinating work on French and English into German has been handled for almost ten years now by Rosemarie Sauer. Thanks to her continued optimism results for this particularly difficult target language are finally beginning to show real promise. It is to be hoped that she will have the heart to stick with the project until more widely usable results can finally be obtained.

We have also been able to rely in recent years on more general expertise regarding the infrastructure. Here I would mention the contribution of Kees van der Horst and Iain Urquhart who between them are largely responsible for the degree to which it has been possible to make Systran an integral part of the Commission's E-mail and informatics infrastructure.

Last but not least a word of praise for the secretarial and administrative staff who have had to battle with rather less challenging aspects of the project:

- Francine Facchin, who has coordinated our immediate infrastructure needs covering everything from dictionary encoding to OCR work and telecommunications, backed from the start by Monique Kneip;

- Bernard Guille and his predecessors who have been of invaluable help in dealing with the procedural side of contract management and the preparation of the decision-making process under the Multilingual Action Plans;

- all the secretaries who have manipulated our local computers and word processors, sometimes directly for the benefit of Systran itself, sometimes in the

interests of peripheral activities such as conference organization, promotional activities or report writing.

* * * * *

Conclusions

The combination of in-house expertise and external maintenance and development staff has, in my opinion, been one of the main reasons for the project's success.

The main role of the Commission staff has been to define strategy and to coordinate the development work while the development contractor has always been in a position to contribute positively to the manner in which software enhancement has taken place.

The interaction between Commission staff and contractors has been conducted along a number of mutually comprehensive lines:

- frequent contacts between the Commission official responsible for a given language pair and the corresponding staff with the contractor;

- monthly meetings for general coordination of on-going development and future strategy, chaired by the project leader in the presence of key contractual staff and Commission officials;

- dictionary coordination at the time of each major update (i.e. four or five times per year) in order to decide, in the presence of at least one Commission representative, on final modifications before the release of a new system;

- processing of feedback by the Commission coordinators before communicating it to the contractor as a basis for system improvement.

This approach, in contrast to other scenarios adopted by the Commission in similar projects, has clearly borne fruit. Minor improvements may be called for

but, by and large, the mutual trust built up between all participants has had a very positive effect on the general rhythm of development.

It is an approach which could no doubt be applied to other areas where constant interaction between the Commission and its contractual staff is called for.

CHAPTER 6

EXTERNAL USERS

I am often asked why it was that the Commission became so interested in encouraging use of Systran outside the European Institutions.

The simple answer is that initially the system was by and large considered to be of little use for in-house work. The translation services found, quite rightly, that the quality was not sufficiently high to be able to contribute to overall efficiency while other user departments were unable to access the system easily owing to the lack of a comprehensive infrastructure combining E-mail with word processing.

Systran's survival depended to a significant extent on feedback from users, both for system improvement and as a justification for further development funding. It was therefore most fortunate that we were able to find some key organizations who were willing to participate actively in pilot projects.

Two major contributors

Early in 1982, when the English-French and French-English systems were beginning to mature, contracts were signed with Aerospatiale and Kernforschungszentrum Karlsruhe (KfK) as a means of testing Systran's potential in a user environment.

The agreements, which were based on the Commission's right to make use of the system for *government agencies in the Member States*, provided access to Systran in return for feedback aimed at further development.

Aerospatiale contributed actively by providing its own term bank for English-French and French-English as well as by financing specialized development work for the aircraft domain. Major improvements were made but, unfortunately, owing to the difficulty of liaising with the translation services in Aerospatiale's divisions in various parts of France, the only real user was the relatively small head office at Suresnes near Paris.

It is interesting to note that, apparently as a result of directives from the French Ministry of Defence, Aerospatiale was later encouraged to deal with Gachot S.A. for Systran as the latter was able to provide access to a computer inside the hexagon.

KfK became enthusiastic developers and users, their aim being to provide a service for translating French nuclear research reports into English. They employed a trained Systran expert, Vanna Genesio, to coordinate development, making an important contribution to the French-English system during the mid-eighties. They can be regarded as the first major user of the Commission's Systran development, translating several thousands of pages in a fully automated environment which coupled OCR with the Systran software. In this way, it was possible to feed in reports on paper and obtain a raw machine translation of bulky documents (e.g. 500 pages) within hours or days rather than weeks or months.

Also of interest is the fact that KfK believed in the concept of "fully automatic machine translation" from the start. Indeed, Dolf Habermann who coordinated Systran activities there, was convinced that raw Systran output could be brought up to a quality standard that was sufficient for scientific experts interested in monitoring progress in research. KfK's continued use of the system seems to prove that this objective has now been achieved.

NATO

Collaboration with NATO, which began in 1985 on the basis of a three part agreement between Toma, NATO and the Commission, has proved useful on several counts.

First, NATO as an international organization bears a number of similarities with the Commission. In particular, it has a translation service which is responsible for processing documents from a wide variety of subject fields.

Second, many of the translations to be carried out are of political importance. Lack of French versions of documents at a given meeting could, in the worst case, mean that an agenda item would not be discussed.

Third, as the institution is based in Brussels, contacts could be made between representatives of the Commission's translation services and those of NATO.

It is heartening to be able to report that NATO was not only able to make a very positive contribution to the development of the English-French system, thanks to the efforts of Jan Carter who had earlier gained considerable experience in Luxembourg, but was also able to bring the system into day-to-day use. The efforts of Albert Cox to introduce Systran for use by translators were followed with interest by the Commission's own translation hierarchy.

If they can do it, it was thought, why can't we?

NATO are continuing to use the system but it has not yet been brought into generalized use with free access for all staff.

The bureau service venture

In 1983, given the difficulty of introducing Systran into the Commission's own services, it was decided that efforts should be made to provide an infrastructure so that any government organization in the EC Member States could have access to the system.

A call for expressions of interest was therefore put out in the Official Journal. On the basis of the replies received, it was decided to authorize companies with suitable computers to provide service to public organizations in four different Member States.

The companies chosen were:

- ECAT, in Luxembourg, which concluded contracts with the Commission's Esprit environment for two to three years;

- ORDA-B, in Belgium, which was unable to find interested clients in the Belgian public sector;

- CSATA, in Italy, which still runs a modest but successful operation and provides some useful feedback in the area of informatics;

and

- Gachot S.A., in France, which was not only to attract one or two public service users but was soon to buy up most of the Systran companies worldwide including Toma's interests in California and Systran Institut's interests in Luxembourg and Germany.

It is difficult, in retrospect, to judge whether the bureau service venture was of any use to the Commission. It certainly did not have the immediate impact that might have been expected. On the other hand, it was to bring about some difficulties for the Commission a year or two later when Gachot, the new owner of Systran, began further negotiations with the Commission.

Other external users

The most enthusiastic of our other users has been the Deutsche Bundesbahn whose initial interest in French-German has now extended to Systran's

potential application in the European railway networks as a whole. DB has tried repeatedly to devise reliable methods of improving translation into German. While the *text suite* approach provided a limited amount of success, they are still not happy with the general level of quality.

Armin Schmidt, the DB project coordinator, has nevertheless made an important contribution by emphasizing Systran's performance for some of the better developed language pairs. He hopes to be able to obtain support for introducing the system for translation between the member states of the UIC (Union Internationale des Chemins de fer).

The IAEA (International Atomic Energy Agency) in Vienna experimented with Systran for a couple of years starting in 1988 when Geoffrey Byrne-Sutton was responsible for translation.

Also in the nuclear area, the French CEA (Commisariat à l'Energie Atomique) showed some initial interest after concluding a contract in 1985. After a long period of silence, they have once again begun to investigate Systran's potential for translating various types of document.

The least successful users to date have been the University of Pisa (1986) where only experimental use has been made in connection with the University of Florence, the Regione Toscana (1988) and the Regione Piemonte (1986) which have given no direct feedback, and the Bundesstelle für Fernmeldestatistik (1990) who have been slow to initiate operations.

Conclusions

The emphasis placed on external use until now has been mainly for strategic reasons. Lack of user interest in Systran could have meant that the project would have been discontinued at an early date.

Yet the initiative was not altogether in vain. Three organizations in particular, Aerospatiale, KfK and NATO, were able to contribute positively to enhancing the system and providing user reactions.

For the years to come, careful thought should be given to the manner in which Systran can be made available to external users, both as a means to overcome language barriers in an expanding Europe and as a basis for dividing the development burden (including financing) between the Commission and the user community as a whole.

There is good reason to believe that the national authorities in a number of Member States may be in a position to take part in this effort. Indications have also been received from Greece, Belgium, the UK and Spain that joint action could be envisaged.

CHAPTER 7

EVALUATIONS

Evaluations are of course a key component of any project. They can be used for various purposes ranging from a desire to obtain objective information on evolving performance or, for the more politically or strategically minded, as a basis on which to bring about change or reorientation, including more generalized use of the product.

Systran has had its fair share of evaluations. Indeed, much of the criticism which was brought to bear on the project in the early years was a result of the way in which consultants interpreted translators' reactions to the system.

This is not to say that the evaluations did not serve a useful purpose. On the contrary, much good advice on how to proceed came out of the analyses which

were commissioned. The only general criticism which might be made is that the assessments came at rather too frequent intervals and were not always mutually supportive.

Micro- and macro-analyses

Georges Van Slype of the Belgian consultancy Bureau van Dijk was charged with a number of Systran evaluations in the early years.

The main objectives were to assess system performance with reference to the number and category of errors, the efficiency of post-editing, and general characteristics such as readability and intelligibility. The detailed error analysis was sometimes referred to as micro-evaluation whereas the other factors came under the concept of macro-evaluation.

The general conclusion on four evaluations carried out on English-French, French-English and English-Italian between 1976 and 1979 was that development work should be continued and that pilot operations should be organized with public and private translation services.

While technical experts and policy makers welcomed the results, some translators questioned them on the grounds that they could not work normally under test conditions. As time went by and post-editing experience grew, the Van Dijk studies were increasingly seen by translators as a DG XIII exercise to impose Systran on the translation services.

Improvability

Veronica Lawson, an independent translator, and Margaret Masterman of the Cambridge Language Research Unit were involved in a number of studies based primarily on the improvability of Systran and its extension to other document types and subject fields.

These studies, which were carried out between 1978 and 1980, paid special attention to the efficiency of post-editing before and after intensive development as well as to a linguistic appreciation of the system's configuration.

On the basis of the work conducted on patent texts, it was concluded that the system could indeed be successfully extended to other fields.

The system's linguistic structure was said to be logical and comprehensive, making it well suited to future enhancements of various types.

It was pointed out, however, that the Systran programming language was not readily comprehensible to the non-expert and that this gave rise to many unfounded criticisms.

As a result of these evaluations, extensions were made into new subject areas and an algorithm, *Elucid*, was developed to *translate* the Systran programming language into natural language (English). To some extent, this demonstrated the complexity of the programming instructions and attenuated some of the more loudly voiced criticism.

KfK's investigation

During the period 1980 to 1985 when Systran was being developed for French-English translation in the nuclear field, KfK monitored progress on translation quality by preparing statistics on comprehensibility, input errors, common word errors, technical word errors, grammatical errors, word order errors and serious grammatical errors.

Among the conclusions was the affirmation that comprehensibility increased from 75% to 95% of all sentences. The total number of errors in the raw translation of two sample texts (kept secret from the Commission's development team) of about 150 sentences each, fell from over 300 to about 45.

Not only did this prove that the development methodology was efficient but it had a significant impact on attitudes at the Commission as to the possibility of providing a raw machine translation service rather than post-edited translation. KfK's use of Systran was invariably without post-editing.

Comparative assessment of Systran versions

In 1984, the Commission was approached by a large multinational company based in the United States which was investigating the possibility of providing networked machine translation services in Europe and America. They were particularly interested in comparing the raw output from the Commission's system to that obtained from the version developed in the United States.

The main criteria used covered the overall quality of raw output, post-editing speed and the cost of further development.

The Commission's system was found to be better on all counts for four of the five language pairs investigated. (The fifth language pair, English-German, was a comparatively new addition to the Commission's repertoire.)

Although collaboration with the Commission was recommended, no concrete action ensued as the company's board felt that the risk factor in introducing services of this type was too high.

Technical infrastructure

As a result of the increasing difficulties experienced in getting Systran to *work* at the Commission, despite success elsewhere, the Belgian software house Sobemap was charged to undertake a study relating to technical infrastructure.

This led to a comprehensive proposal for integrating Systran in the Commission's Unix-based infrastructure, complete with staffing requirements. The plan was subsequently implemented by DG IX, then in charge of translation, as a top priority. It was mainly as a result of this initiative that the current Systran service has been developed.

Assessment by translators

In 1988, Ivo Dubois, who was then director of translation, coordinated what was intended to be a two-year assessment of the usefulness of Systran to the translation service. It was based on the experience of two translators from each of six language units.

The aim was to establish document types and subject fields which could be efficiently processed by Systran with post-editing.

The experiment in fact lasted only for about a year. Some of the results were quite positive but the translators' attitude remained rather negative.

The decision to use Systran for translating the minutes of the chefs de cabinet was a direct outcome of the assessment. This application continues to be one of the most deeply appreciated services provided as a result of Systran accessibility.

The Oakley evaluation

Nineteen-ninety was to see a number of important changes in regard to the Systran project. On the one hand, Eddy Brackeniers had been appointed director-general of the new Translation Service while Frans de Bruïne had become director of DG XIII-B, the department responsible for Systran development.

A Belgian consultancy, CEGOS, was charged to undertake a general investigation of the Translation Service while a panel chaired by the British research executive, Brian Oakley, was given the job of evaluating the work achieved under the various Multilingual Action Plans with particular reference to Systran.

CEGOS was to list Systran as the service which users would most like to see developed.

The Oakley report provided a long list of detailed recommendations, the most important of which concerned the need to adapt the Systran service to user requirements and to re-engineer the software by rewriting it in a high-level language and by incorporating a relational data base for dictionary management.

While more attention was indeed given to user needs, the perceived priority for re-engineering could not be implemented, mainly for economic reasons.

Overall, the Oakley report together with the CEGOS finding were to be responsible for much greater interest in Systran, leading to intensified promotion of the service for use by all Commission officials.

Future evaluations

As can be seen from the various accounts given in this chapter, evaluation of machine translation is no easy matter.

What is still lacking, is a methodology which can be reliably applied at regular intervals to demonstrate the degree of progress achieved.

Perhaps the main reason why such a methodology has not yet emerged is that machine translation services seem to evolve with user needs, which in turn evolve with general developments in infrastructure (networks, hardware, word processing, aids to document preparation, OCR, etc.), as well as with priorities in translation processing *per se*.

Ultimately, the best judge of the success of any machine translation operation is the user, not just because of what he says about the system but, more importantly, in terms of the degree to which he uses it and benefits from it in his day-to-day work.

Careful monitoring of user reactions combined with statistical information on the number of requests and the volumes of text translated for different language pairs or different document types could thus prove to be the most reliable system of evaluating how successfully the system is implemented with respect to time.

Finally, the development exercise should adapt more quickly than in the past to the type of improvements and changes required by the user. Here, informatics factors such as user friendliness or speed of execution could be as important as the enhancement of translation quality.

CHAPTER 8

VERSIONS AND CONVERSIONS

Systran's image has certainly not benefited from the fact that various versions of the system have been developed to suit the needs of user and marketing groups in different parts of the world.

The trouble started in the late seventies when Toma began to grant *exclusive* licences to a number of organizations and companies. In point of fact, the agreements were rarely exclusive and in some cases were clearly mutually contradictory.

For example, an agreement signed with the Iona Corporation in Japan granted full rights for any language combination involving Japanese while a similar agreement with the Munich Systran Institut covered all language pairs involving German. Who, then, would have the rights for German-Japanese?

Serious problems emerged in Canada when it was discovered that Toma unwittingly had given world rights to WTCC for French-English and English-French depriving Toma's own company of any further business based on these two key systems. The situation was partly rectified in the early 1980's when Sadao Kawasaki of Iona was to buy back these rights and share them equally between Toma's WTC and Iona.

Toma had also granted licences to companies such as General Motors in Canada and the Xerox Corporation in the United States, allowing each not only to use Systran for their own internal needs but also to use it in worldwide service operations. In the course of time, Toma lost his copies of the contracts, creating even more legal confusion.

Source code

One of the main causes of diversity was Toma's release of the Systran *source code* to a number of sites.

Most software suppliers restrict releases to the *object code* which allows users to execute the program but not to modify it. Access to the source code allows them to make changes at any level of the system.

It was as a result of provision of the source code that Toma enabled WTCC in Canada, the Systran Institut in Germany and the Commission itself to work on comprehensive system development and not simply on dictionary additions.

Originally, of course, Toma had intended all those concerned to work in full collaboration and harmony. What in fact happened was that business interests outweighed more noble causes with the result that different, and largely incompatible versions soon began to emerge.

Shortly after signing licences with WTCC and Systran Institut in the early eighties, Toma decided to try to re-establish his reputation as the main system supplier. For this reason, he began to develop the so-called Universal system which contained a number of new features which were not yet available in the Canadian and European versions.

At the same time, the Canadians were adding new features of their own at the peripheral level while the Germans were independently continuing development of certain linguistic algorithms.

The Commission was caught squarely in the middle of all this confusion. Ideally, we would have liked to continue our relationship with Toma in order to benefit from the new features with were being introduced, particularly as many

of these were based on our own suggestions (more powerful dictionaries, more rational taxonomies of semantic codes, increased use of macros at the programming and coding levels). At the same time, we could see the potential benefits of much of the work undertaken by the Canadians.

Unfortunately, Toma's contract with Systran Institut gave them the status of his representative in Europe, providing them with the authority to negotiate directly with the Commission on his behalf. This would have been all very well if Systran Institut had maintained good relations with Toma but the situation quickly deteriorated and both companies soon put an end to all technical exchanges.

In order to advance, the Commission was increasingly forced into undertaking major software developments of its own, based as far as possible on the logic being used in California. This proved to be easier to achieve than may be thought as we were able to draw on the experience of Thomas Pahl of Bonn who had worked on the original Systran system and had maintained contacts with Toma.

Yet despite our good intentions, it was not long before two rather different versions of key language pairs such as English-French began to emerge.

We believed that when Gachot took over Systran interests from both Toma and Systran Institut in the winter of 1985-1986 it would finally be possible to put all the pieces together. Unfortunately, to date this has not been achieved despite quite considerable efforts on our part.

It is for this reason that rather different versions of the key language pairs (English-French, French-English, English-German, German-English) continue to exist to this day.

The Siemens conversion

When the Commission first acquired Systran in 1976, it was installed without difficulty on the IBM 370 machine which was then in use.

From the start, though, it was clear that Systran could not continue to run in-house on an IBM computer as it was already Commission policy to restrict

support to European suppliers. Indeed, it was thought that it would not be too difficult to convert it to run on another platform.

The machine which was chosen was the Siemens BS 2000 which bore some similarities to IBM in that both systems architectures were based on a common root, namely developments undertaken by RCA in the fifties and sixties.

Toma himself put forward a conversion proposal which finally took the form of a contract with his German company, Platonis. The idea was to use a Univac machine in Calif ornia as a basis for a rewrite which was expected to be compatible with Siemens. (The Univac and Siemens operating systems had both stemmed from a common Honeywell system.)

The conversion proved far more difficult than might have been expected. Enormous difficulties were encountered with almost, but not quite compatible system macros resulting in months of unsuccessful tests.

The Univac link was abandoned and Siemens experts were brought in. Only in 1982 did the system begin to run properly, thanks to the involvement of Thomas Pahl who knew both the Systran system and the BS 2000.

Unbelievably, the day on which it was decided that Systran should finally migrate, the Commission ruled that the Siemens machine should be set aside exclusively for work connected with the steel crisis which was receiving a great deal of attention at the time. We therefore had no choice but to revert to an external IBM machine.

The VM version

The U.S. Air Force has always played an important part in Systran developments.

As a high-volume user for translations from Russian into English, the USAF made significant investments in system improvement at various levels.

In the mid-eighties, a need evolved for rapid translations of small quantities of text submitted by users wishing to grasp the meaning of titles of books or headlines of articles in magazines or journals.

While the traditional MVS system was still the most efficient for translating running text, it proved to be comparatively slow for interactive screen-based access. IBM suggested that the problem could be overcome by adapting Systran to run directly under their VM operating system.

The work was carried out with assistance from Toma's Californian office and advice from IBM.

When finally installed at the USAF Foreign Technology Division in Dayton, Ohio, the VM version operated very well but in fact proved less popular than might have been expected. The USAF therefore continued to maintain MVS as the basic system.

There was, however, to be a significant, if unexpected spin-off effect.

PC versions

The USAF's Systran licence bears similarities to the agreement concluded between the Commission and Toma, in that it covered use of the system by U.S. government agencies.

One of their main clients turned out to be the U.S. Army who saw Systran not just as a tool for accessing foreign-language information but as a means to translate English maintenance instructions into other languages.

As tensions began to build up in the Middle East in the late 1980s, the Army showed increasing interest in using Systran's English-Arabic system. The USAF were not particularly interested in running that particular version and did not wish to become involved in maintaining a system specifically for another agency.

For this reason, the Air Force began to investigate the possibility of porting Systran to a desktop environment, believing that other users would be happy to acquire integral hardware-software packages without the need for mainframe tie-ups.

The VM version proved to be an excellent starting point for PC versions. Initially a DOS version was developed but was found to be far too slow. However, when IBM launched the PS/2, speed and efficiency were found to be comparable to those of a mainframe environment.

At least a dozen English-Arabic packages were made available to the Army on this basis and are still operational at one of their bases in Florida.

The Unix conversion project

Systran operations at the Commission have been unduly subject to conversion problems.

In the early eighties, it was decided by those responsible for computer services that the Unix operating system should become the basis for future developments. In parallel with this decision, it was made clear that any systems running on non-standard operating systems, such as Systran's MVS, could no longer be supported.

As one of our main aims was to develop Systran for internal use, this strategy presented serious problems in that the Standards Implementation Committee made it known that unless efforts were made to convert to Unix, further Systran development contracts would be in jeopardy.

In 1985, we therefore undertook a study with Thomas Pahl's company Codework to investigate the effort required to port Systran to a Unix platform.

A full and detailed proposal was put forward for converting the basic system and at least one language pair to Unix. The minimum cost was estimated at about 1 million Ecu and could have increased to around 4 million Ecu for the seven language pairs under development at that time.

This would have absorbed virtually all Systran development funding for about three years and would have provided results which, at best, would have led to a slow-down in execution time.

For these reasons, no conversion was made.

The Oakley follow-up

One of Oakley's main recommendations in 1991 was that Systran should be rewritten in a high-level language such as C++ so as to enable porting to modern desktop environments.

In preparation for work along these lines, various experts were invited to make proposals.

Not surprisingly, the estimate was in the range of 6 million Ecu for the major language pairs. So once again it was decided that we could not afford to pay for the work under existing budgets.

Possible solutions

What has continued to surprise me over recent years is that despite the availability of PC versions of Systran in the United States, the Commission has made no real attempt to obtain the software.

One of the reasons may be that the most efficient version, that on the PS/2, is still considered by the Commission to be non-standard. However, rules can be bent under exceptional circumstances, particularly if cost differences of several orders of magnitude are involved.

Finally, it appears that there are now IBM office computers such as the 9000 series and the RISC machines which could run the Commission's MVS versions with only minor modifications. And now that MVS seems to be on the cards for open-systems approval, we may well see direct PC support for MVS in the not-too-distant future.

CHAPTER 9

SYSTRAN WORLDWIDE

The previous chapter contained some information on how versions of Systran were licensed to different organizations over the years. Here, I shall try to give a fairly up-to-date account of how Systran is being used in various parts of the world.

On the development side, there continue to be three major interests:

- the Gachot group with its head office, Systran S.A., in Soisy near Paris but its main development centre, Systran Translation Systems & Latsec, in La Jolla, California;

- the Iona group, headed by Sadao Kawasaki, which created Systran Corporation, Tokyo, mainly for Japanese-English and English-Japanese translation;

- the Commission itself which, with internal and contractual staffing of some 45 persons, has certainly become the most important development centre.

The U.S. Air Force

Apart from the Commission, the main user of Systran continues to be the USAF's Foreign Technology Division in Dayton, Ohio.

FTD were instrumental in encouraging Toma to complete his first operational version of Systran for Russian-English in the late 1960s and have done much to promote improvements to the programs and dictionaries over the years.

Their main interest continues to be to translate foreign language information into English although there are indications, now that the cold war is over, that they are becoming more interested in opening up their resources to a wider audience, possibly with the need for more comprehensive translation facilities.

Be that as it may, the Russian-English system still continues to be the most widely used. Also installed are French-English, German-English, Spanish-English and Japanese-English. Of these, French-English, developed under Commission contracts, is said to give the highest level of quality.

Most users are involved primarily in information scanning and tend to be happy with unedited raw output. Post-editing by professional translators is however available although the use of semi-automatic, screen-driven editing is usually considered sufficient. Here, the translator is guided to passages in the text where potential errors have been detected by the system. The work is said to be rather frustrating at times as, at this stage in development, most of the potential errors do not in fact require any attention.

The USAF has always shown great interest in the Commission's developments and on several occasions has made proposals on joint developments. Until now, these have been of little direct benefit to the Commission owing to the difficulty of coordinating operations through the USAF Systran contractor, Latsec Inc., in California.

Recently, thanks to the initiative of Dale Bostad of FTD, the USAF has attempted to encourage the Gachot group to release the Russian-English system to the Commission on mutually acceptable terms.

Xerox Corporation

Of all the private companies which have been involved in machine translation, the Xerox Corporation of Rochester, New York, has certainly made the most useful contribution.

Xerox became interested in machine translation in the late seventies when they embarked on a new approach to document preparation based on the use of limited vocabulary and document drafting recommendations.

In this environment, with a total general and technical vocabulary of about 7000 words coupled with short, clearly and unambiguously written sentences, it was found that exceptionally good results could be obtained with Systran.

Experience with the English-French system in 1978-79 paved the way for an ambitious development program which included Italian, Spanish, Portuguese and German as new target languages. These were integrated well on schedule with the result that by 1984, Xerox were using Systran to translate the bulk of their technical and maintenance documentation into these five targets.

Interestingly enough, field engineers (rather than translators) in Europe and South America were encouraged to post-edit the translations and provide feedback to the development centre in Rochester.

By the mid-eighties, Xerox was boasting that thanks to machine translation, they had been able to cut product launch dates to Europe from six months to two weeks. This may be something of an exaggeration, but there are certainly instances of foreign language maintenance manuals being published before the English originals!

In 1989, Xerox decided to expand the repertoire by including the four Scandinavian languages, Danish, Norwegian, Swedish and Finnish. Initial results with these seems to be promising, particularly for Danish.

Side by side with work on linguistic developments, Xerox was able to make good use of its own workstations technology to develop interfaces between publishing software and Systran. Today they have facilities which enable fully formatted, multi-fonted page layouts complete with graphics and photographs to be

processed by Systran and automatically re-established in the various target languages.

Finally, they now offer networked Systran translation services to their clients in Europe and throughout the American continent.

General Motors

The Canadian arm of General Motors began to show interest in Systran in 1974 when they were having to cope with Quebec's French-language laws.

They have generally been very quiet about their involvement, partly, no doubt, because of a licence received from Toma in 1975 which enabled them to develop and market the system on their own account.

One symptom of this is that the name of Systran is never used by General Motors or its subsidiary, E.D.S. However, Stan Sereda, the E.D.S. project manager, speaking at the 1986 World Systran Conference in Luxembourg explained how machine translation had become an integral part of their document production chain.

At that time, while English-French was still the main language pair, some use had already been made of English-Spanish.

General Motors' continued use of Systran was indicated recently by Ed Lipmann of IBM who informed me that the IBM package *Translation Manager II* had been adapted to interface with the Systran version used by E.D.S. of Whitby, Ontario.

Japanese

Certainly the most difficult Systran developments have been English-Japanese and, above all, Japanese-English.

While some European language pairs have presented serious problems, the challenge of dealing with Japanese was enormous.

Apart from technical problems involving character sets and Kanji-Kana conversions, the main problem in translating from Japanese proved to be word-boundary definition. Indeed, while in Indo-European languages sentences are divided into individual words, a Japanese sentence consists of a continuous string of characters. Thus, before translation can start, each character string has to be broken up into word clusters.

Most Japanese machine translation suppliers avoided this issue by insisting on human pre-editors. Kawasaki, of Systran Japan, decided on the opposite course: fully automatic processing.

Now, after some ten years of intense effort, promising results are being obtained. While little real use is being made of Systran in Japan (apart from the translation of IBM computer manuals from English into Japanese), the U.S. Air Force have reported that raw translations from Japanese into English are evolving well.

The Commission itself has had occasion to deal with Systran Japan in the framework of its JapInfo project which consists of accessing pertinent Japanese-language information in Japan and having it translated and post-edited into English. Two machine translation systems were initially used for this work, Fujitsu's Atlas system and Systran. Use of Systran was discontinued some two years ago when we were informed that the staff who had worked on post-editing were to be used principally for further development work.

In August 1992, we were told by Eriko Akazawa of Iona that Systran Japan was now preparing to market its developments worldwide.

Arabic

The real origins of the English-Arabic system are difficult to establish clearly.

Possibly, the first initiative came from Saudi Arabia where a member of the Sindi family showed interest in investing in Systran in the late 1970s. Thereafter, not only Toma but also the founders of the Systran Institut in Munich as well as Jean Gachot became involved.

In the early 1980s, it was Gachot who continued to finance the project which soon began to bear fruit. However, while the Saudi oil interests were ready to invest in practically any seemingly attractive high-tech project in the seventies, by 1983-84, when the Systran English-Arabic system was ripe for demonstrations, far less enthusiasm was shown.

Several potential customers have appeared over the years, most recently in Libya. Apparently, though, Gachot has not yet been able to find a profitable market.

The only real user I know of to date is the U.S. Army in Florida which installed the system on PCs in 1989.

Minitel

When Jean Gachot acquired the Systran system in 1986, he immediately set about providing access to Systran from the French minitel network.

The minitel is a small terminal connected to household and office telephones which enables practically any telephone subscriber in France to send electronic messages.

While the device was not really perfectly suited as an input device for Systran, the service started off on a promising basis, particularly on occasions when Gachot had been interviewed on radio or television or when articles on Systran appeared in the press.

However, rather like the *minitel rose* service, interest in Systran began to wane as users found that the translations they received were not as good as they had expected. In most cases, the reason for substandard output was a result of badly drafted or incorrectly formatted source texts. In other words, what was

lacking was a level of user friendliness which advised the user on how to improve the response.

In principle, though, this type of approach could prove interesting for professional users if it continues to be developed along the right lines.

Future cooperation

In general, Systran users seem to be keen on exchanging ideas and even, in some cases, on exchanging data. This was clearly indicated at the World Systran Conference in 1986.

On the more practical level, however, systems and dictionaries have diverged so much that it is no longer easy to combine developments for a given language pair. This was clearly shown in our efforts to merge our developments with those of Gachot or those of the U.S. Air Force.

But more general contacts between Systran users can provide interesting information on the success of new components, particularly those developed for improving the user interface. Here, for example, the Commission might well benefit from the experience of others in the use of local or client-oriented dictionaries or in interfaces which combine publishing software with access to Systran.

Given the progress made by the Commission on Systran use and integration in recent years, we could contribute much to mutual exchanges of views.

THE EUROTRA CONNECTION

No account of Systran's development would be complete without a word on Eurotra. Indeed, whenever the Commission's role in machine translation is under discussion, someone always expresses surprise at the fact that we are, or have been, involved in two very different projects.

There is, of course, a logical explanation as to how this came about.

The language barrier conference

In 1977, Loll Rolling of DG XIII organized a well represented international conference in Luxembourg on the theme *Overcoming the Language Barrier*.

The main objective was to increase awareness of all that was going on in language processing from term banks to multilingual thesauri and from research to applications. Machine translation was, of course, one of the key topics under discussion, particularly as presentations of the Commission's early experience with Systran were made. Last but not least, Peter Toma himself gave a paper on Systran's prospects.

Another prominent figure at the conference was Bernard Vauquois, head of machine translation research at the University of Grenoble. Vauquois was highly regarded in Europe for the work he had done on applying his linguistically oriented approach to an operational Russian-French system.

As the conference was drawing to a close, Vauquois pointedly asked from the floor why it was that the Commission was investing in American technology rather than in the results of European R&D.

Georges Anderla, the XIII-B director responsible at that time, invited Vauquois to his office the very next day. There it was decided that the Commission would

indeed study the possibility of supporting European research side by side with the further development of Systran.

What happened was that the Saarbrücken R&D group soon learnt of the Commission's interest in Grenoble and they too called for support. A comparative evaluation of the two approaches was made leading to a recommendation that both teams should work together in order to complement each other's technology and experience.

Naturally, it was not long before the British asked to be included too, then came the Danes, the Dutch, the Italians... The result was a proposal called Eurotra.

The Geneva meetings

For a number of years beginning in 1978, Maggie King of ISSCO (Institute for Semantic and Cognitive Studies), Geneva (chosen in the interests of neutrality), chaired monthly Eurotra meetings with Commission officials and national representatives.

The aim was to draft design specifications on which the Council regulation for Eurotra was ultimately to be based.

The stated objective was to *create a European machine translation system of advanced design*. It was generally assumed that it would indeed be possible to develop an operational system based on the results of European research.

The early meetings got off to a good start as they brought together a wide variety of expertise, including representatives of *existing systems* such as Peter Wheeler and myself. There were also a number of pragmatists such as Frank Knowles of Aston, Yorick Wilks of Cambridge and Bruno Zolta of Milan.

Perhaps somewhat inevitably, the more academically oriented linguists, particularly those from countries like Denmark and the Netherlands which had no direct experience of natural language processing, slowly started to gain ground. They, after all, had much to gain from maintaining that current technology left much to be desired and that a completely new approach was necessary.

Within a year or so, emphasis was no longer being placed on key building blocks like dictionaries or syntactic parsing but was devoted increasingly to ways in which one could disambiguate sentences like *He saw the girl with the binoculars* or *Time flies like an arrow*.

It was all very interesting from a theoretical point of view but could hardly be expected to contribute to practical progress.

As a result, several of us decided to leave the linguists to their own deliberations and continue with more down-to-earth developments.

Lack of synergy

I am not alone in regretting the general lack of synergy between Eurotra and Systran when both projects were developing in parallel.

Even if the basic principles of each approach were different, there were some common denominators.

Both systems required a dictionary, both could benefit from corpus analysis, both had to deal with phenomena particular to a given language such as noun decomposition in German. But while those of us working on Systran were always ready to share the results of our efforts, little interest was expressed by the Eurotra team.

Most surprisingly, in cases where we felt Eurotra staff could help *us* in solving some of *our* linguistic problems, we were told that this could only be done under additional contracts with Systran funding! We therefore had no option but to do all the work ourselves.

Conclusions

In order to explain the apparent divergence of goals between Systran and Eurotra, clear explanations must now be given to the general public as well as to those specializing in language processing.

Only in this way will it be possible to explain how it was that the Commission could embark on such widely differing paths for so long.

Nor must we forget to mention the political pressures which were imposed on the Commission from the Member States. These included:

- prestige for the research efforts in each centre with little interest in the overall Community result;

- a desire for Community funding, not only to support the Eurotra effort but to serve as a basis for r esearch of more direct interest to the institute in question;

- a feeling expressed on numerous occasions that each national language required particular treatment because of its own very special grammar and syntax;

- the enlargement of the Community during the design and implementation stages of the project with the result that the six language environment soon expanded to nine languages, leading to 72 potential language pairs rather than the original 30.

Under these circumstances, and given the division of Eurotra project management between the Member States and the Commission on the basis of association agreements, it is not surprising that the Commission's project leader was obliged to give more and more attention to research to the detriment of operational developments.

It can only be hoped that in future the Commission will ensure that all its resources can be developed with a full exchange of experience and expertise in order that a natural osmosis can occur between research and applications.

Finally, the interest in Systran expressed by coordinators of more recent R&D projects in the same area (e.g. *Eurolang* , *Graal*, *Ilodoc*) shows that there is real interest in achieving synergy between applications and research in the area of language processing.

CHAPTER 11

SYSTRAN'S PLACE ON THE WORLD SCENE

Machine translation has come a long way since the Commission first showed interest in the mid-seventies.

Today there are about 20 systems on the market for European languages and as many more for translating in and out of Japanese.

The market is currently split between the older and more complex systems like Logos, Metal and Systran, which were designed and developed to run on mainframe computers, and software packages for PCs which are cheap but lack sophistication.

The mainframe systems

Three systems stand out clearly as being based on more or less the same basic technology. These are:

- the Spanam/Engspan systems developed under Muriel Vasconcellos at the Pan American Health Organization (PAHO) in Washington, DC;

- Logos, developed by Bernard Scott in Middletown, N.Y., but now managed by Bill Hohenstein in Dedham, MA;

- Systran itself.

All were developed initially on IBM 360 computers and all have similar approaches to sequential parsing and integrated dictionary data bases.

The results obtained depend to a large extent on the size and sophistication of the dictionaries for various language pairs which in turn depend on the degree to which the systems have been used in practice.

The two PAHO systems give highly acceptable results between English and Spanish, particularly in the areas of health, food science and agriculture, but have not been extended to other language pairs.

Logos, which was initially marketed for German-English with some success, has in recent years been developed for the Canadian market with the emphasis on English-French and French-English. All in all some ten pairs are under development but, for the time being, German-English appears to be the only one which can compete with Systran in terms of quality.

Logos, in addition, has now developed attractive user-friendly versions for a number of platforms, particularly for Unix-based workstations.

Systran, on the other hand, has the advantage of large, well tested dictionaries and has a far wider repertoire of language pairs than any other system.

A fourth system, Metal, can also be included in this category in that its origins date back to research work in the 1960s at the University of Austin in Texas. This difference is, however, that Metal is rule based, i.e. it has a defined set of linguistic or grammar rules which are accessed at various stages of the parsing program, creating as many valid representations of a sentence as possible before eliminating those found to be less acceptable.

While the system has performed quite well on translating certain types of maintenance manual, further development has become increasingly difficult as additions to the dictionary and, in particular, to the program often leading to unexpected negative side effects which outweigh the benefits of the intended improvements. (By contrast, the sequentially-based logic of the other three systems seems far more suited to continued improvement.)

The future of Metal is somewhat in doubt as Siemens, the current promoter, is not entirely happy with sales and performance and has begun to participate heavily in two Eureka projects.

The PC approach

Since the late seventies when developers in and around Brigham Young University in Provo, Utah, showed interest in what today has become known as workstation technology, we have seen a proliferation of desktop software for machine translation.

First came the Weidner and Alps systems from Provo itself, then the Smart system from New York, Globalink and Linguistic Products from Texas, the d'Agostini system from Udine in Italy, XLT from Montreal and various systems like FB translator, Bilingua and Winger from Germany and Scandinavia.

These systems can cost as little as $300 and can be run on standard PCs. The problem, of course, is that the systems supplied usually have fairly small dictionaries which may well not be suitable for the type of translation to be carried out. There is therefore a considerable burden on the user to update the system for his own purposes, a task which in many cases can be counter-productive.

On the more positive side, many of the features now being introduced in the PC software are very user friendly and are certainly welcomed by some users. And for certain language pairs and subject fields, the PC systems may well produce output which is comparable to that available from the older mainframe systems.

As far as sales are concerned, certain PC packages like Linguistic Products and Globalink seem to be doing quite well. Generally, their marketing prospects seem to be more promising than those of the larger system suppliers who always seem to be in need of further financing.

Japanese systems

All the large Japanese corporations have been confronted with the problem of translation, both from English into Japanese, in order to keep pace with foreign technology, and from Japanese into English and other languages when products are to be launched on the international market.

For this reason, it was no surprise that MITI included machine translation as one of the priorities in its fifth generation computer architecture at the beginning of the eighties. Practically every large Japanese company immediately began to invest heavily in the new technology with the result that over a dozen systems are already in use, nearly all from English into Japanese and a few for Japanese-English.

What is interesting about the Japanese approach is that initially the basic technology used was that developed by Vauquois at the University of Grenoble. This is to be explained in part by the fact that Makoto Nagao, professor of electrical engineering at Kyoto University, had himself had contacts with Vauquois when his interest in natural language processing was growing from term banks to machine translation. Furthermore, the Japanese were strongly influenced by the Commission's Eurotra project which initially also had close ties with Grenoble.

What is strange about the Japanese systems is that the approaches followed by all the various companies therefore have a common origin - Nagao. They are rule-based rather than sequential and make use of the tree-structure approach to linguistic analysis so favoured by Vauquois.

It is only fair to say, that some of the drawbacks of this approach are now becoming apparent with the result that certain developers are beginning to take a much more pragmatic, Systran-like approach to the problem.

Perhaps the most highly developed of the Japanese systems is Fujitsu's Atlas which was originally developed for English-Japanese and Japanese-English but is now being extended to other languages such as German and Spanish. Fujitsu provide access via telecommunications on much the same basis as the Commission provides access to Systran.

One of the more impressive developments is the Toshiba system which integrates in one and the same unit OCR and translation software. An English source text can be fed in the top and a Japanese translation comes out a minute later at the bottom.

Sharp prefers to distribute software packages for use on desktops. I understand that several thousand have already been sold for English-Japanese and that user satisfaction is comparatively high.

While the quality of output between Japanese and English is generally far lower that what machine translation can provide for European languages, use of systems in Japan is far more widespread than elsewhere in the world owing to the difficulty of finding translators. Indeed, those charged with the editing of machine translation output often have quite different qualifications, such as subject field knowledge, but can still do an excellent post-editing job.

The Japanese are continuing to invest heavily and may well attack the European and U.S. markets within the next couple of years with machine translation software for European languages. For the time being, however, they still appear to be directing the bulk of their effort to further improvement of translation in and out of Japanese.

Future market trends

A strange tug-of-war is currently evolving between high-quality mainframe-type systems which can be accessed through the networks and lower quality PC software packages.

From the purely business point of view, it seems probable that the PC approach will continue to grow in terms of market share while sales of the larger mainframe packages will, at best, remain pretty constant.

This is not to say, however, that systems like Systran, Logos and Atlas will not continue to evolve: they certainly will. But their improvement will depend first and foremost on the efforts of the in-house development staff in large organizations such as the Commission, the UN agencies or multinational suppliers like Xerox or Fujitsu.

By the turn of the century, though, we shall probably see a merging of technologies as desktop systems increasingly take on the capacity to run mainframe-type software and as the existing PC software develops to catch up with older developments.

Prospects for Systran

As far as Systran itself is concerned, its chances are twofold:

- owing to continued use by the Commission, the U.S. Air Force and Xerox, its dictionaries will continue to grow and its accessing mechanisms will improve;

- with respect to PC technology, the Systran versions which are presently installed on PS/2 machines in the United States could well become the preferred product for the evolving market.

For the foreseeable future, though, there seems to be no system which could compete with Systran in the Commission environment either for ease of access, updatability or quality and coverage.

Nor do current R&D projects like Graal or Eurolang seem to present real competition in our environment, even if these could well produce interesting results for industries intent on preparing and distributing maintenance manuals in several languages.

We should, however, continue to monitor developments and should not be afraid to share our technology with industry in the interests of providing better tools for overcoming language barriers in the enlarging Community.

CHAPTER 12

IN-HOUSE TECHNICAL INFRASTRUCTURE

The attempt to introduce Systran on an operational basis at the Commission was never just a matter of providing adequate translation quality.

From the start it was obvious that if the system was to work, it would be necessary to link it into a suitable infrastructure for text input, telecommunications and editing.

The Wang years

When we first started to experiment with Systran, all input - whether text or system updates - was prepared on IBM 80-column punch cards. These were fed directly into the mainframe card reader at the computer centre and, a few hours later, produced output on A3-size computer listings.

This was hardly a way in which to introduce the service to users.

We therefore charged the Belgian software house Sobemap with the task of studying and recommending word processing equipment which could interface with Systran.

The final choice was the Wang OIS system which was not only the most widely used facility at the time but also had distinct advantages over other word processing systems in that it covered all the languages of interest and catered for extensive telecommunications possibilities, including interfacing with the IBM 370 mainframe on which Systran was being developed.

We were all very happy with the performance and ease of use of the interlinked Wang terminals. The equipment proved to be very reliable and the English and French translation departments which had been given workstations and printers to enable them to experiment with Systran soon began to use them in connection with human translation too.

This initial experience with word processors was of course to lead to much wider computerization of the translation services in the years to come but it was in fact thanks to Systran that initial experience was gained.

Olivetti and Philips

By about 1984, the Commission had started to introduce its own strategy for office systems. Initially, the idea was to replace typewriters by stand-alone word processors. The two machines selected were the Olivetti ETS and the Philips 5020. Wang was excluded as their machines were designed primarily for networking.

It was not long before serious problems were being experienced in trying to channel Olivetti and Philips documents to Systran. A few attempts were made to benefit from conversion programs written by the European Parliament but the situation was far from satisfactory.

Eventually, as a result of brainstorming between DG XIII and DG IX, Sobemap was once again charged with finding a solution.

The raw Systran service

This time, Sobemap proposed a much more ambitious strategy which provided a basis for access to Systran from any part of the Commission.

The basic idea was that the Systran service should ultimately be available to everyone. The Commission's new informatics architecture was based on the use of Unix servers, dumb terminals and telecommunications. Sobemap therefore recommended that a clear distinction should be made between the development

exercise, which could remain in DG XIII on a Wang/IBM environment, and the service side which would be managed separately.

In the event, Kees van der Horst of the translation department was appointed service manager. By 1987, a pilot operation had begun, linking pilot users to the Commission's Amdahl machine which was upgraded to run the MVS operating system needed for Systran.

It was not always easy to interface the Unix-based Q-office word-processing package with Systran, particularly when it was found that many operators failed to use standard formatting conventions. Progress was however made and the first translations began to come in.

One of the fundamental differences in this approach was that the service was no longer intended primarily for translators but rather for the people in the services who actually wanted translations for themselves.

From menus to INSEM

With the help of Iain Urquhart, who by 1990 was spending all his time on improving the Systran interfaces, menu driven access based on Q-office was introduced. In this way, a user could specify a document name, choose a language pair and, optionally, one or more subject fields before submitting the translation by *uucp*, the Unix telecommunications protocol.

The text would first be sent to the Systran server, an NCR tower in Luxembourg, and from there to the Amdahl machine.

Within twenty minutes or so, the user could then retrieve his translation which would have the same basic name as the original followed by the code of the target language. For example, an English language document called *test* would be returned after translation into French as *test.FR*.

This procedure was satisfactory as long as the user had texts in Q-office format and as long as his department continued to update the interface. However, it was not long before more and more users were requesting access from PCs and more and more new departments were interested in trying out Systran.

To simplify matters over the short term, the Commission's internal electronic messaging system, INSEM, was used as a new gateway into Systran. By sending an E-mail message via the INSEM local server to a Systran mailbox, a translation could be obtained without the need for any special interface. In addition, by coupling the Q-office formatting code to Systran, it now became possible to receive fully-formatted translations. And last but not least, users of MS-Word and WordPerfect could also access Systran from their PCs thanks to the conversion possibilities which had become part of the INSEM service.

Future improvements

These accessing possibilities have provided a basis for processing over 1600 translation requests per month from a wide variety of users and terminals.

While the service generally operates well, there are a few drawbacks:

- turnaround time can sometimes extend to well over half an hour and is seldom less than five minutes;

- documents from PCs in MS-Word or WordPerfect formats are not always correctly converted by the INSEM interface;

- users are not advised when problems arise.

To overcome these, work is currently in hand to make use of the LAN technology now being introduced. Initial tests have shown that the turnaround can be reduced to less than one minute and that status messages can be flashed to the user on line. Furthermore, the user can remain connected to Systran for the entire operation rather than making the two separate connections under the old procedure.

Direct access through the LAN is also expected to make it easier to introduce direct interfaces for the various word processing packages in use in the PC environment.

Access from the outside

At the time of writing, plans exist for making Systran available to the other EC Institutions as well as to users working on Commission projects initiated by other DGs such as DG XI's network for the civil security services.

Here, use is already being made of protocols like X25 or X400 which allow outside users to access the Systran server. Furthermore, many of the other institutions have a technical infrastructure very similar to that of the Commission with the result that they are able to make use of the INSEM procedures.

Improvements to all these interfaces are envisaged over the coming months in parallel with the planned promotion of Systran in the other European Institutions.

CHAPTER 13

VALUE-ADDED SERVICES

Systran is, of course, a machine translation system. In other words, it is designed to process a source-language text so as to provide a full translation in a target language.

While this type of use is obviously of interest to many, some of those working in the field of multilingual documentation or translation seem to require additional levels of service.

There are already indications that Systran may be able to serve as a basis for such applications.

Terminology lists

For certain language pairs, the Systran dictionaries contain a large number of string terms, mainly noun expressions, which in most cases correspond to technical or administrative terminology in subject areas of interest to the Commission.

These dictionaries have been developed mainly on the basis of texts actually submitted for translation and therefore contain terms which could well be those translators most need.

The advantage of using Systran for retrieving terminology is that an alphabetical term list can be produced within minutes for any document in machine readable form, in much the same way as for machine translations. The user can then consult the list on screen or on paper as an aid to normal translation.

Initial tests have shown that output is already quite promising for mature, widely-used systems such as English-French or French-English. Owing to the multitarget approach, the results for other language combinations for which English or French are source languages will no doubt also be of interest.

The following options are now being considered:

- creation of simple, networked access for mature language pairs;

- definition of a further development strategy based on the reactions of pilot users, including possible interlinking with updates for the Eurodicautom term bank or incorporation of other terminological resources;

- extension to missing target languages from English (Danish) or French (Portuguese, Danish, Greek).

A more general assessment of extensions to this facility could be made later in the light of user reactions.

On-line access to CELEX

The CELEX data base contains all the legislative documents of the Community in all nine official languages. It is frequently used by translators who need to have exact translations of the titles of Community regulations or proposals.

Until now, a degree of human intelligence has been needed to convert official references into their CELEX equivalents and subsequently to interrogate the data base for each reference separately.

An interface within Systran now exists which automatically retrieves all references to legislation from any source document in machine readable form and converts the references into the CELEX code.

It should soon be possible to resubmit the list to CELEX automatically so that the translator can receive a full list of appropriate titles in the target language for any document in machine-readable form.

The problems now to be solved are of a technical nature:

- the CELEX batch interface should be made available through the network for input from Systran;

- this will no doubt entail upgrading the Bull telecommunications protocol from the Commission's own MFTS protocol to the LAN-oriented file transfer protocol (FTP);

- user-friendly interfaces, for example in the form of mail-boxes, will also need to be introduced on the Systran Unix server or a similar machine.

Equitext

The Equitext software which is designed to pick out bilingual lists of terms from two different language versions of the same document should be further developed and brought on line for use by translators.

Ideally the system should work in two rather different modes:

- automatic comparison of up to 2000 pages of text in any two of the languages covered (English, French and German) leading to a listing of all noun expressions with frequency counts;

- the possibility for the user to interrogate representative corpora based on pretranslated material in order to be able to retrieve translations of any term together w ith frequency information.

These facilities should not only assist in the further development of Systran, Eurodicautom and related systems, but should provide a more reliable basis for checking the reliability of technical terminology in general.

If the English-French-German approach proves successful, Equitext could be developed for other languages, at least at the target level.

Additional services

Over the next 12 months, use of the Systran software could also be tested for use by translators for purposes of spell checking or even, at a later stage, style checking.

We could also consider using the Systran structures for preparing lists or compendia of pretranslated material which would be accessible on-line on the basis of the *longest match* principle.

CHAPTER 14

USER REACTIONS

Whenever I come across a new machine translation system or computerized aid for translators, I always try to find out what the users have to say about it. Those interested in Systran's use at the Commission may well want to have the same kind of information.

While the reactions of users always appear more authentic if they are communicated directly to the interested party, I shall try to summarize the various ways in which Systran has been received.

Translators

For many years, it appeared to those of us involved in Systran development at the Commission that the primary users of the system would be translators. This was based partly on the feeling that Systran output could not be used without post-editing by translators and partly because other early users such as the USAF or General Motors had also seen the need for translator involvement.

Until the late eighties, then, the translator was considered to be the user.

Translators had, of course, been involved from the very start, not only in assessing the usefulness of the system in formal and unofficial evaluation work, but in the actual development effort.

Not surprisingly, reactions in the early days tended to be fairly negative. The policy at the Commission at the time was to translate all documents up to a high quality standard which meant that any use of Systran involved heavy post-editing.

The most cooperative pilot users turned out to be the English translators in Luxembourg. From 1979 until about 1986, thanks to tie-ups with the Wang word-processing network, several translators were able to try their hand at on-screen revision work.

Rapid post-editing

The general consensus among those who acquired experience in this new art was that if normal quality standards were to be achieved, Systran had little to offer. The post-editing time often exceeded the time the translator would have spent on dictating his translation in the usual way. Furthermore, the rather strange syntax of Systran output often led to unnatural structures in the final edited version. In other words, translators sometimes felt they were being upset by Systran to the detriment of their own, more creative approaches.

As a result, in about 1980, Emma Wagner of the English translation department in Luxembourg came up with the idea of *rapid post-editing*. She felt that Systran could be of assistance if end-users were willing to accept grammatically correct translations even if these were lacking in style. Such editing work could, she believed, be carried out at a rate which was substantially higher than that for normal human translation, namely up to four pages per hour.

A pilot operation along these lines was set up for translating minutes of meetings for two user groups.

The end-users were happy enough. They not only received their translations more quickly than usual but were also able to benefit from a word processing

infrastructure. This meant clean-typed output which could either be immediately photocopied for distribution or further edited by the end-user if modifications were called for.

The only problem with the approach was that the user community did not grow. On the contrary, it tended to diminish in size as the original enthusiasts were replaced.

Finally, in 1989 the innovators of the rapid post-editing approach came to the conclusion that Systran, after all, was not saving them much time and could sometimes have negative repercussions on their work. Since then, they have preferred to use more conventional approaches.

They did, nevertheless, all agree that the introduction of word-processing equipment for Systran had a beneficial effect on translation processing in general.

Systran output as a reference

Once Italian as a target language became available, a number of Italian translators began to show interest in the system.

Their approach turned out to be very different to that of the English. For a start, little interest was shown in on-screen work, the general opinion being that typing should be left to secretaries.

What they did begin to see fairly soon, however, was that Systran was able to provide useful terminology. Before starting to translate technical reports on research or industry, the Italians would therefore often request a Systran translation for reference purposes. Although they used more traditional approaches in their actual translation work, they would use the Systran output for reference.

Of course, the amount of applicable terminology grew from month to month and from year to year with the result that by the early 1990s some Italian translators were able to record substantial time savings, not only in translating from English but also from French into Italian.

Most of the credit for this development should be given to Delfina Campanella who has maintained excellent relations with her Italian colleagues while ensuring that their suggestions for improvement were included in on-going development work.

French and German translators

I have decided to group these two languages as, despite considerable differences in the quality of output, the translators themselves have until now taken largely the same attitude.

By and large, both have taken considerable interest in how Systran is developing, both believe that perhaps some day in the not too distant future Systran may be able to provide a useful service, but both prefer to use more traditional approaches as the Systran output is far from easy to revise.

There have recently been indications that some French translators in Brussels are finally beginning to see some cases in which Systran could be an asset.

Other targets

Few meaningful reactions have been received from translators working on the other target languages.

Dutch target is certainly not generally considered to be up to post-editing standard.

One or two isolated enthusiasts are beginning to appear for Portuguese while, finally, some interest in Spanish is now emerging.

Exceptions to the rule

After many years of experimentation, translators are finding that some documents are beginning to give promising results with Systran.

The best example seems to be the minutes of the weekly chefs de cabinet meeting which need to be translated as quickly as possible from French into English for non French-speaking commissioners.

For almost two years now, these documents have been successfully processed by Systran with post-editing at a rate which cuts translation time down to a few hours.

The final result is considered to be acceptable for both translators and end users.

Much more recently, Italian translators in Brussels have begun to use Systran for the translation of parliamentary questions. Here too the results appear to be very promising.

One of the prerequisites for Systran use still seems to be that the source document should be available in machine-readable form even if OCR technology is progressing well. Indeed, one of the reasons why translators have been ready to work with the chefs de cabinet minutes and the parliamentary questions is that these documents are available from the network.

Conclusions for translators

I do not think it would be an exaggeration to say that for translators Systran has probably arrived before its time.

While most in-house translators welcome any aids which will facilitate their task, the Systran approach which simply provides them with a mechanically produced text for editing is certainly not what they are inclined to ask for.

On the one hand, the difficulty of requesting a Systran translation for texts which are not in machine-readable form or for translators who do not have a terminal in their own office, is not to be underestimated. On the other, many translators seem to feel that what they need most is an improvement in access to terminology or to relevant documentary data bases.

Finally, the art of post-editing is very different to that of revision of human translation.

For all these reasons, it is not surprising that the general reaction from translators has been somewhat sceptical.

Other users

It was Dolf Habermann of KfK, the nuclear research centre in Karlsruhe, who first put forward the idea of using Systran without any participation of translators.

His reasoning was based on the fact that human translation usually cost too much and took too long to be of direct interest to the research community. On the other hand, few German scientists could read French and were thus unable to benefit from many of the research papers published by their colleagues across the Rhine.

Habermann believed that if Systran could be extended to cover the terminology of nuclear physics, the raw translations would be acceptable for information scanning purposes.

After three or four years of intensive development, this objective was indeed reached. Reactions from the user population both at the Karlsruhe centre and from nuclear scientists in Britain were very positive.

The in-house raw Systran service

Reactions of this kind played an important part in convincing the Commission hierarchy that raw Systran output may well provide a useful service inside the Commission.

In addition, a pilot operation with DG XVII, the energy department, in 1985 had clearly shown that in many cases the end users would have preferred to get the raw Systran output within minutes rather than a post-edited version a few days later.

Once it was finally decided to make Systran available through the network, its potential immediately became clear.

The two or three preselected users soon started to tell their friends about what could be obtained and requests from new users started to pour in.

At this stage, the main advantage seemed to be that a quick and dirty translation could be obtained very quickly.

It was only in 1991 at the time of the Oakley evaluation that we started to have more meaningful feedback.

This showed that users appreciated the service for one or more of the following reasons:

- with suitable editing, it helped them to speed up the internal translation work they had always had to do anyway;
- it provided them with a basis for drafting documents in their own language;
- it helped them to scan foreign language information;
- it was sometimes sufficient in the raw state for use at meetings with national experts.

On the more negative side, it was felt by many that the service was not yet suitable for the following reasons:

- specialized terminology was still missing;

- certain language pairs, particularly those involving German, were not up to standard;

- turnaround speed was often too long;

- further improvement of interfaces with various word-processing packages was necessary.

All these factors have been borne in mind in current development of the system and its interfaces.

While it is no easy business to incorporate all the terminology for a given subject field or document type, procedures are now being developed to facilitate the provision of feedback from frequent users.

In addition, careful consideration is being given to improving ease of access and user-friendliness.

Adapting to the customer

Certainly, in a field like machine translation and in an institution like the Commission, if the service does not fit the customer's needs, there is no reason why he should make use of it.

We have seen in industry the extent to which machine translation, as part and parcel of a fully automated document production process, can cut costs and speed up product releases.

But what the customer wants is sometimes hard to deliver, given the way in which documents are often submitted.

Many of the critical reactions we have received may well be justified but those of us working in the development environment are sometimes amazed at what actually happens in practice.

Among the more obvious errors, we see:

- countless spelling mistakes, making it almost impossible for the computer to provide usable results;

- formatting errors, causing sentences to be split up or wrapped together more or less at random;

- long, rambling sentences, full of polysyllabic words but devoid of clear ideas;

- source texts containing several different languages or long lists of names and addresses.

Given all this, it is quite surprising that interest in the Systran service has continued to grow.

In the future, it is to be hoped that additional computerized aids as well as better system documentation and training will improve the way in which the service is used a s well as the criteria on which further developments are based.

Statistics and percentages

One of the most direct forms of user reaction comes from the monthly statistics on Systran use.

For September 1992, it is interesting to note that of the 1700 translations requested, 1500 or 88.2% came from end-user departments, 186 or 10.9% from the Translation Service and 14 or 0.8% from the development team.

The most popular language pairs were French-English (29.2%) and English-French (22.3%), none of the others accounting for more than 8.4% The least

popular systems are those with Spanish as a source language with 1.6% for Spanish-French and just 0.9% for Spanish-English.

The only four departments submitting over 100 requests for the month were the Translation Service (182), Agriculture (180), the Administration (164) and the Statistical Office (147).

All in all 39 departments accessed the system for a total of 535 different users.

These data in themselves are possibly the best indication of how users are reacting to Systran.

CHAPTER 15

PROMOTION

Our Systran interests have been promoted in many different ways and on many different occasions since the earliest versions of the system began to produce translations towards the end of 1976.

Our first publicity

On that occasion, BBC television sent a crew over to Luxembourg to make a film of Systran as it was then operating at our Luxembourg computer centre. We spent a couple of hours with the BBC producer giving some general information

on our work and then, at about two o'clock in the afternoon, we went down to the floor where the mainframe was installed.

A paragraph of text had been prepared on punchcards. The first shot was of an operator feeding the cards into the reading device which buzzed through them.

We were told that the translation would be printed out within half an hour or so. The cameraman set up his equipment next to a high-speed printer where the output was expected to appear.

The minutes and then the hours went by. Three, four, five, six and finally seven o'clock. The crew had to be back in Britain for another session the next morning and had to reach Calais before midnight. Still no output had been forthcoming.

In desperation, they finally filmed some statistical data that was coming off the printer, packed their bags and drove off. I went back to the machine room to see what was happening. Before I could get near the printer, the operator handed me the printout which had come the minute the BBC people had left.

It's what the professionals call the demonstration syndrome. Machines simply do not like to perform for the media.

The sequence finally became part of a highly successful programme called Babel which was retransmitted several times in the United Kingdom and other English-speaking countries. On the TV screen, it was impossible to see that the printout for the Statistical Office was not the authentic Systran output!

I should add as a footnote that our Systran efforts were recognized by other TV stations, notably FR3 which did a 30-minute programme, rebroadcast four times, on Systran developments at the Commission.

Conferences

The first conference organized by the Commission at which Systran was discussed in any detail was the one on Language Barriers in 1977. In 1978, the

Commission subsidized the first of a continuing series of ASLIB conferences on Translating and the Computer which attracted audiences from the translation profession.

At a more expert level, in the early eighties there were several conferences in the United Kingdom and France on machine translation research and development, which soon led to the Expolangues series currently held all over Europe.

On these occasions we were able not only to attend workshops or conference sessions on Systran but were often able to present practical demonstrations of our approach. The most successful events were in Paris, Frankfurt and Lisbon.

We also had the opportunity of demonstrating Systran at the Europe 2000 conference in Strasbourg in 1983, at the Milan fair in 1984 and at Delft in 1985.

Finally, participation in the United Nations inter-agency meetings on translation and documentation as well as NATO's Agard conferences ensured a wide international audience for our papers.

Many of these events were of course reported in the press and it was not long before the Commission had begun to gain a reputation as a world leader in the field of machine translation.

In-house promotion

Promotion of Systran inside the Commission proved to be more difficult to organize than our efforts to inform the outside world of the progress we were making.

There was reticence on the part of the hierarchy for several reasons. First, it was difficult to provide clear evidence that Systran was of assistance to translators or other users, second, for many years the technical infrastructure was hardly suited to networking users into the Systran system and third, it was not clear whether DG XIII or DG IX should take on responsibility for encouraging in-house use.

The situation became clearer in 1988 when the raw Systran experiment was initiated by the translation department. However, it was not until 1991 that the need for a real promotion effort was recognized, partly as a result of the Oakley evaluation and partly on the basis of improvements to the network and to Systran itself.

This initially took the form of the appointment of an official from the Translation Service (Dorothy Senez) to liaise with current and potential users and, in 1992, was followed by the publication of a full-colour brochure in French and English which was distributed to all Commission officials in Brussels and Luxembourg.

In addition, user documentation containing advice on how to get the best out of Systran has now been produced for interested users. It contains recommendations on document drafting and formatting as well as a step-by-step explanation of how to access Systran through the in-house networks.

What next?

Promotional efforts to date, particularly those over the past three or four months, have led to a considerable jump in Systran use. In 1988, monthly throughput averaged less than 50 genuine requests or about 400 pages. By 1991, this had grown to about 300 requests or 1800 pages. Now, after distributing the brochure, we have achieved peaks of 1700 requests from 535 different users representing over 10,000 pages of translation (September 1992).

I would estimate that about half the current throughput is from users who are not just interested in the technology but who have something concrete to gain from it.

What we now need to do is to set up more reliable structures in order to ensure that those who stand to benefit most actually understand how to access the system. This will not only entail stronger public relations and training efforts but, probably most important of all, will require improvements to the infrastructure so that even the non-expert can obtain usable results with a minimum of effort.

Of course, there will be much to learn from the users themselves. Only now are we beginning to receive feedback on what still needs to be done to improve the services offered and on the different types of application the current user community has in mind.

Lastly, we should not be afraid to benefit from the experience of users of machine translation and related services outside the Commission. A great deal is going on in the field and software packages for language processing are now penetrating the market, particularly for PCs.

CHAPTER 16

THE LIGHTER MOMENTS

No account of Systran's evolution at the Commission would be complete without a few words on the more comical translations which the system has come up with from time to time.

In the very early days, practically every translation contained at least one strange mistranslation. Often these originated from proper nouns. I will always remember how the name of one of our most seasoned interpreters, Mrs van Hoof, was translated *Madame sabot de fourgonnette* while the Israeli premier, Begin, repeatedly came out as a verb form:

> U.S. supports Begin proposal.
>
> Les supports américains commencent la proposition.

Experience with French as a source language brought even stranger assertions:

> Nous *avions* envisagé un développement des structures de base.
>
> We *aeroplanes* forecast development of basic structures.

No one had realized that the plural of *avion* (plane) would coincide with the verb form related to *avoir*.

But by far the most astounding translation of all occurred the first time the prepositional phrase *vis à vis* came up in a French text.

Systran's rendering in English, which perhaps represents some inherent, but as yet insufficiently understood philosophy of the machine-to-man relationship, came across loud and clear:

LIVE TO SCREW!

CHAPTER 17

RECOMMENDATIONS

Many of the preceding chapters contain a summary of my own views on the manner in which we could proceed on some of the topics under discussion.

Before finalizing this rather personalized account of Systran's development at the Commission over the past 17 years, I should like to take the opportunity of putting forward a few ideas on priorities for the future as I see them.

For ease of reference, I have divided them into four categories: linguistic development, technical infrastructure, promotion and new applications. The relative importance of each category will, however, in my opinion depend very much on the answer to a much more general question: **Who is our target population?**

Here, I must emphasize my view that **for a number of years to come, the bulk of all the Commission's Systran development efforts should be aimed at satisfying the internal needs of the European Institutions**, starting of course with the Commission itself.

Only later, depending on the more general evolution of the machine translation market, should we consider adapting our Systran version to the needs of outside users.

1. Linguistic development

1.1 Contractual relationship

The close relationship between the Commission's linguists and the maintenance and development work carried out under contract has proved to be a dependable method of tailoring enhancements to user requirements. It could however be improved by creating better channels for obtaining feedback from the user community.

1.2 Quality improvement

Particular emphasis should be given to improving translation quality in all six directions between English, French and German. This could be achieved by intensifying the current development effort on these

languages as well as by bringing the linguistic routines for German more in line with those for the other languages.

1.3 New language pairs

Care should be taken not to reproduce the confusion which has often arisen in other machine translation projects when too many new language combinations have been added too quickly.

On the other hand, of all the nine official languages, Danish is still completely missing from the Commission's repertoire although a Systran English-Danish system already exists! The next logical step could then be to add English-Danish, possibly followed by French-Portuguese which would be easy to develop and could prove useful to translators.

In my opinion, it is too early to add additional source languages such as Greek or Dutch although the existing Russian-English system could be added without the need for development.

1.4 Value-added services

Systran's dictionary and parsing potential should be further developed and integrated into the Commission's infrastructure to provide on-line terminology scans, fully automatic access to CELEX and, at a later stage, guidance in document drafting and layout.

2. Technical infrastructure

2.1 Speed of access

It now usually takes a user from 10 to 30 minutes to obtain a Systran translation. Experiments with the LAN file transfer procedures which are now coming into generalized use has shown that this can be cut down to about 40 seconds, enable the user to obtain a Systran translation in just one session. This approach should be widely introduced to replace the comparatively unreliable and very slow INSEM procedure.

2.2 Interfaces

Much progress has been made on interfacing texts from the Commission environment with Systran. It is now possible to submit Q-office texts with full formatting and to receive a translation in the same format.

However, an increasing number of users are switching from Q-office to Word or WordPerfect. Although they can still obtain Systran output by using the INSEM facilities, the results are nearly always substandard from the point of view of displays and formats. These conversion errors usually have negative effects on linguistic quality too.

What is now needed is more direct interfacing between Word and Systran and between WordPerfect and Systran. In this context, careful consideration should be given to reprocessing through a pivot format, i.e. one which is compatible with both packages.

2.3 Access to new services

Two value added services, word lists and CELEX interrogation, are now being introduced via Systran. The logical solution would be to make these available through additional mailboxes but these might overload the current server.

Steps must therefore be taken to ensure that all Systran services may be accessed efficiently via the most widely available electronic mail possibilities. This could entail the installation of an additional Unix server over the short term.

2.4 Access for other institutions

Further work on networking with the other EC Institutions should be undertaken either specifically for the Systran environment or in the more general context of inter-institutional communications.

While the Economic and Social Committee, the Court of Auditors and, to some extent, the European Parliament can connect to the Systran server, problems remain to be solved, in particular, for the European Investment Bank and the Council of Ministers.

This matter deserves careful consideration although some of the responsibility should be placed with the various institutions involved.

2.5 Re-engineering

Re-engineering, i.e. re-writing Systran in another programming language as proposed by Oakley, does not seem to have any immediate advantages for the EC institutions and would cost a fortune.

The recommendation here is, then, to remain on an IBM-compatible mainframe with the MVS operating system which can offer access to a wide range of users by means of telecommunications.

If PC-type distribution were to become a priority, it would be wise to try to benefit from the technology developed in conjunction with the U.S. Air Force for PC versions or, better still, for PS/2 versions of Systran.

3. Promotion

3.1 In-house priorities

While the distribution of the Systran brochure has obviously had a marked effect, there is reason to suppose that many users may not have appreciated the level of the service offered or the extent to which it could be of benefit to them.

The Systran promotion team should try to establish how far the brochure has in fact created genuine interest and if not, why not.

Owing to the comparatively low cost of promotional material of this type, targeted mailing of user documentation should be considered.

Finally, the public relations efforts which are already taking place should be intensified, particularly with departments which already have an enthusiastic Systran user community.

3.2 Other EC institutions

The recent promotional efforts with the other EC institutions have been generally successful.

Consideration should now be given to preparing promotional material similar to the Commission's brochure for these new potential users.

If possible, Systran coordinators should be appointed in each institution both for promotional purposes and for handling linguistic and informatics problems with the Commission.

3.3 External users

As I have already stated, I do not see generalized promotion of Systran as a priority at this time.

Nevertheless, preparatory measures could be undertaken to assess the extent to which ministries and government departments in the Member States could make use of Systran.

A first step could take the form of a presentation in Luxembourg to representatives from ministries which have already expressed an interest or, failing this, to the research or industry ministries. In this connection, the Greek experience as well as DG XI's initiative with Systran in the area of civil protection could be raised.

3.4 Commission image

Our participation in international events over the past few years has been less frequent than in the past.

The current success with Systran should be widely publicized at such events, not only for general purposes of prestige but more specifically as a basis on which to request financial resources for on-going development work.

In this connection, we should also attempt to be more directly involved in the language-related Eureka projects.

4. New applications

4.1 Language engineering

It is largely thanks to progress on Systran that the Commission is now in a position to participate more actively in the area of language engineering.

Consideration should be given to using Systran's infrastructure for managing exchanges of dictionaries or other lexical or terminological data as well as for participating in corpus-oriented work.

4.2 Evolving technologies

Care should be taken to ensure that Systran keeps pace with evolving technologies, particularly in the field of desktop systems and telecommunications interfaces.

The Commission should therefore keep abreast of on-going developments in the personal computer sector to assess whether Systran can be ported to a stand-alone office-systems environment at reasonable cost (e.g. OS/2 or RISC systems) with a view to providing service to clients who cannot easily make use of the Commission's internal network (EC delegations, NATO, etc.).

In relation to interfaces, we should monitor progress on word processing packages with a view to optimizing the links between the linguistic side of the Systran package and requirements for formatting, page presentation and graphics. Wherever possible, we should attempt to incorporate international standards such as ODA or SGML.

4.3 Other machine translation systems

The Commission should continue to monitor new approaches to machine translation, both with a view to implementing the technology in-house for certain language pairs or certain environments as well as to benefit from new technological approaches in the Systran environment itself.

In this connection, in the coming years significant developments are expected in systems which combine document drafting with machine translation and in techniques for updating local and/or centralized dictionaries.

Last but not least, fax technology coupled with OCR could play an important part in providing a reliable means for submitting texts in hard copy to Systran or other machine translation systems.