[*Terminologie et Traduction*, no.1, 1986]

CURRENT SYSTRAN DEVELOPMENTS AT THE EC COMMISSION

Ian M. Pigott,
Commission of the European Communities

Introduction

In 1976, when the Commission first started to develop Systran for the English-French language pair, a great deal of scepticism was expressed from a wide variety of circles. Most of the translators working on the original development team left in desperation after two or three months while potential users who had the opportunity to see the raw output in those early days almost invariably ridiculed the machine's performance. However, the time and effort put into Systran developments at the Commission are beginning to pay off.

We now have four high-quality language pairs - English-French, French-English, English-Italian and English-German - and are about to release a fifth, French-German. In addition, we are progressing rapidly with the development of French and English into Dutch and have already started work on English into Spanish and Portuguese.

Few would now deny that machine translation has passed from dream to reality.

Today I shall attempt to give an overview of our approach to Systran development and use at the Commission in the hope that more extensive coordination of our efforts with those of other development and user groups will result in even better returns in the future than those we have been able to report until now.

Linguistic approach

As will become increasingly apparent during this conference, many Systran users have extremely specific needs as regards language pairs and subject areas. Here at the Commission, we recognized from the start that for any machine translation system to be successful for us, it would need to have the potential not only of being extensible to many different language combinations but of providing the ability to cover many subject fields and specialized subsectors for any given language pair.

When we acquired the basic package ten years ago, it was obvious that these requirements would call for tremendous development efforts. Whereas other users could no doubt run a successful operation with relatively small dictionaries based on fairly specific types of input document, our terminology requirements alone would have to cover all major fields of interest; and as for document typology, we could expect everything from data base abstracts to official reports or personal communications, all with wide variations in language usage and syntax.

For this reason, it was not long before we decided on what we now refer to as the wide-based approach - that is a development philosophy based on the probability that the precise subject sector or document type of a text for machine translation processing can rarely be successfully defined. We therefore decided to create very general basic dictionaries containing translation equivalents which could, as far as possible, be applied to a wide range of subject fields. However, in order to achieve specialized equivalents in context, we put tremendous efforts into developing a meaningful set of semantic and syntactic codes which could be used as a basis for contextual dictionary entries relying on the structural relationships of the words in a document.

The same philosophy has been applied to the enhancement of the translation software itself at both source and target levels with the result that texts of various types and formats can be successfully processed through the system. In this connection, it has be necessary to pay a great deal of attention to the analysis of text formatting or page representation in the source document so as to arrive at the establishment of meaningful units of text when dealing with tables, indents, columns and other display characteristics. Such an approach also facilitates the reproduction of the original page format in the translated text which is particularly useful for post-editing purposes.

This development philosophy is now starting to pay dividends. At the Commission, for example, it enables us to translate reports from our Secretariat General on all sectors of interest using the same basic system, while outside users report that they are extremely surprised at the results obtained in specialized fields that have never been the subject of specific developments.

The approach is also a great advantage for our increasing user group as it means that improvements made for one user become immediately available to all other users. In other words, if technical terminology or improvements in the translation programs are made for the benefit of, say, a user working in the field of nuclear research, these same improvements will lead to significantly better translations of documents from other users dealing with related fields such as energy, chemistry or physics.

We do, of course, still have a facility for introducing special meanings by subject field - the so-called topical glossary approach - but this we only use as a last resort when the wide-based contextual approach proves to be inadequate. However, even in those areas where topical glossaries have been used, in any given text well below one percent of the meanings will in fact be provided on the basis of a subject field parameter.

Dictionary enhancement

One of the top priorities in improving the standard of raw translations is to ensure that the dictionaries are coded with great care and that potential improvements do not lead to degradations.

This is achieved in two ways. First, all dictionary information from various sources is carefully checked by the linguist responsible for a given language pair and second, the coding techniques used are reviewed by a dictionary coordinator.

These two steps usually lead to the elimination of well over *95% of* entries liable to conflict with existing dictionary equivalents but, owing to the tremendous complexity of language, the final measure of success can only be made on the basis of translation tests.

Thus, before any new release - and on average we make new releases about once every two months for all users - the quality of translation obtained from the potential new release is very carefully compared with that of the previous production version. From time to time, serious errors are discovered at this stage and can be corrected before they can cause any damage. I may say, in this connection, that post-editors often feel that a previously corrected error has begun to reappear. In point of fact, the error is nearly always a result of an anomaly in the source text which has not occurred before.

Another area in which we are beginning to make real headway in dictionary work is at a more technical level, namely in the automation of various dictionary coding and updating procedures. Until recently, dictionary coders had to fill in a great deal of complicated morphological information at the target level, particularly when dealing with irregular verbs, nouns and adjectives. This process has now been almost completely automated and the coder simply has to insert the basic form of the meaning he requires. This not only saves time, it also avoids many of the human errors that traditionally resulted from unnecessary complexity. In addition, it makes it easier for external development staff to participate directly in the dictionary coding effort.

Over the coming year, we hope to extend the automated approach to other areas of dictionary coding. If this is achieved, future development work, particularly as far as the creation of new language pairs is concerned, will be able to proceed much more rapidly than in the past.

And last but not least, we are presently undertaking a comprehensive revision and rationalization of the dictionary coding manuals in order to provide clear and updated information on all the coding features now available. Examples of the new approach to dictionary coding are available and will be discussed at the workshop sessions.

Infrastructure

Over the years we have come to realize the importance of a reliable technical infrastructure for the Commission user environment.

Our aim is to provide translators and secretarial staff with user-friendly methods of inputting source texts, requesting Systran translations and undertaking post-editing, wherever possible directly on screen. Furthermore, we are becoming ever more conscious of the need to provide networking facilities between the Systran environment and the equipment used by requesters of translations in order to facilitate speedy electronic transmission of texts from office to office and from building to building.

Up to now, we have proved conclusively that Systran can be used successfully as long as fully compatible equipment is used throughout the production chain. What we are now trying to do - and in fact this is taking rather longer than was initially expected - is to provide a means of connecting the various types of equipment now in use to one fully generalized network.

This aspect will be dealt with in greater detail later and it will certainly be one of the main subjects to be discussed at the workshop on word processing.

Coordination

Of all Systran users, the Commission is in the somewhat enviable position of having had direct dealings with all groups responsible for European language pairs. Since the early days when we dealt exclusively with Dr Toma's group, World Translation Center, we have acquired enhancements from the Canadian group, World Translation Company of Canada, which did much to provide peripheral software designed to facilitate a user interface between the linguistic software and text processing systems. We have also dealt extensively with Systran Institut, Germany, which for many years represented Dr Toma's interests in Europe, and more recently with Informalux, a subsidiary of Arbed, Luxembourg, which built up a team of Systran experts which is probably now second to none. Finally, we have for many years been able to call on the services of Thomas Pahl, currently a director of Codework GmbH, who has an unmatched wealth of expertise in fundamental Systran programming.

That list of credits may certainly appear very impressive - and from a theoretical point of view it has the makings of enormous assets. However, from a more practical standpoint it has had serious disadvantages in that, owing to the lack of collaboration between Systran development groups over the past five or six years, we have been obliged to pick and choose.

For example, although we have always relied on World Translation Center to provide us with basic developments for new language pairs, we have found it increasingly necessary to undertake a certain amount of local conversion work in order to be able to integrate these developments in our own operational environment. Initially, such adaptations were relatively easy to manage but as the Universal Systran system from California has evolved, we have found that up to three or four man-months are needed to tailor the system to our own requirements and those of our European end users.

I am happy to report today that I now see a real possibility of concrete collaboration between all the current Systran development groups - World Translation Center and Latsec in California, Systran Institut in Germany, Systran Corporation in Tokyo and the Commission itself.

One of the main reasons for my optimism is that Mr Jean Gachot, who is here with us today, has now acquired the La Jolla development groups created by Peter Toma (World Translation Center and Latsec) as well as a majority share in Systran Institut, Germany, and a 50% share in ECAT, Luxembourg. It is to be hoped that this consolidation of the Systran development effort will lead quickly to a more streamlined development and marketing policy than has been possible in recent years.

Indeed, the very fact that so many Systran enthusiasts have come to this conference shows that a true desire to collaborate, to avoid duplication of effort, to pool experience and know-how, now exists on a world scale. It is my earnest hope, that by Friday evening, we shall have some concrete proposals on how to achieve this objective which will now be of direct benefit to the various development and user groups concerned but which also could have an enormous impact on the place of Systran in the MT world.

Certainly, 1f we can finally put an end to local variations in Systran developments and can restore constructive collaboration between developers and users, the future of Systran will be ensured.

Future priorities

One of the main aims of this conference is to establish guidelines for future cooperation on Systran developments. In this connection, I feel it would be useful for me to emphasize those areas which could benefit from such coordinated efforts in the hope that any mention of them in the formal presentations today and tomorrow will be carefully borne in mind for the workshops on Thursday and Friday. To my mind there are five areas which require particular attention:

1. Rationalization of linguistic software

   Until 1978, there was only one Systran system, and although Systran was already being used for several language pairs, all basic development work was centralized in La Jolla, California. Since that time, owing to problems of licensing and user rights, various groups, including the Commission were left no choice but to go it alone.

   It is for this reason that for certain language pairs, in particular English-French, French-English, English-German and German-English, there are now at least two sets of Systran translation programs. As up to 95% of any given system is in fact completely identical to its counterpart, it would now appear to be most important to see that existing discrepancies are eliminated and that future development is coordinated.

2. Dictionary development

   Inconsistencies in dictionary development are even more difficult to resolve than variations in software, given the fact that even within one linguistic approach there may be several sets of largely incompatible dictionaries.

   The single most costly item in developing a high-quality language pair is indeed in the area of dictionary expansion. It therefore seems ridiculous that different development groups working in the same subject area for the same language pair should be developing independent dictionaries. This is a high-priority item, but obviously can only be achieved once the basic translation system has been standardized.

3. User interfaces

   The Systran system itself, developed as it was long before the word processing revolution, leaves much to be desired as far as the user interface is concerned. In some cases, system suppliers such as the Canadian group, the Commission and now Systran Japan have tried to develop sophisticated peripherals which will interact with a wide variety of input and output devices. Moreover, many users - in particular the US Air Force - have introduced peripherals of their own.

As the data processing environment becomes increasingly complex with the advent of optical character reading, an ever wider variety of Input and output devices, with personal computers and micros supporting more and more text processing packages and, equally important, with integrated photocomposition and telecommunication facilities, it is essential that developers and users alike should pool their resources and develop a generalized interface between the translation machine proper and the particular applications it is called upon to serve.

This is a vital requirement for Systran, particularly as we are faced with a proliferation of MT stand-alones which in many cases do not require the level of interfacing any large mainframe application warrants.

4. Post-editing techniques

In all the three aforementioned areas, Systran experts are generally aware of the problems that exist and may well have discussed ways and means of resolving them.

In the area of post-editing, on the other hand, there has been little, if any collaboration between users on how best to use hardware and software facilities to assist the translator or, indeed, on how the translator should be trained to undertake the task in hand. In my experience, most translators have learnt to post-edit by trial and error. I feel the time has now come to recognize post-editors' requirements and, as far as possible, to cater for them in the Systran system or support software.

5. Coordination of user feedback

In most user environments, system improvements are based on feedback from users - who may be translators, terminologists or simply requesters of translations.

Sadly lacking in this context is any clearly prescribed set of guidelines to facilitate the work of reporting MT errors and suggesting ways in which they can be remedied. All too often, the only basis for system improvements is the post-edited text as it comes back from the translator's office.

I am convinced that many of the translators here today will have constructive comments to make on this subject which could appropriately be put forward at the workshop on user requirements.

These then are the five areas where much could be done in improving the performance and development potential of Systran. I sincerely hope that each and every one of them will receive the attention it deserves at this conference.

Conclusions

In conclusion, I should like to stress the following points:

1. After ten years of instensive development, the operational language pairs at the Commission have reached a level of quality and performance suited to an increasing number of document types and subject fields.

2. A well-defined programme of development of new language pairs will bring to at least nine the number of operational systems available at the Commission over the next two years.

3. The various modes in which Systran MT is used deserve examination in order that appropriate post-editing training can be ensured and that adequate support hardware and software can be provided.

In closing, I should once again like to stress the basic message of this paper, and indeed, the message of this conference, which can be summed up in one word, cooperation.

- cooperation between development groups in an effort to provide unified systems and dictionaries for all language pairs; and, just as essential,

- cooperation between user groups in order to ensure that feedback from all applications is taken into account at all levels of development and production.