[*Machine Translation Today: Translating and the Computer 15*. Papers at a conference…18-19 November 1993 (London: Aslib)]

DEVELOPMENTS IN SYSTRAN

September 1994

Dorothy Senez
JECL 6/65 -Translation Service
Commission of the European Communities
200, rue de la Loi
1049 Brussels

ABSTRACT

Systran, the European Commission's multilingual machine translation system, is a fast service which is available to all Commission officials. The computer cannot match the skills of the professional translator, who must continue to be responsible for all texts which are legally binding or which are for publication. But machine translation can deal, in a matter of minutes, with short-lived documents, designed, say, for information or preparatory work, and which are required urgently. It can also give a broad view of a paper in an unfamiliar language, so that an official can decide how much, if any, of it needs to go to translators. In this way, much time can be saved for a translation service which is already facing a relentless increase in the volume of its work and which will have to cope with the new languages of an enlarged European Union. We have set up a post-editing service to correct machine texts for users who cannot do this in their own departments.

Raw machine translation is only one of a number of multilingual services now being made available. The switch to personal computers throughout the Commission, and the greater use of increasingly reliable electronic mail, also means that other forms of help can be given. First, a bridge has been created between Systran and Celex (the multilingual data base containing Community legislation). Secondly, and only in recent months, Eurodicautom (the Commission's multilingual terminology data bank) has been incorporated in the Systran dictionaries. With this link, it will be easy to look up technical terms in a given language and have them returned in one or more other languages.

A survey has shown how officials use Systran and has enabled us to identify their needs. In all these ways, Systran is making excellent progress as a means of rapid communication between the many departments of a multilingual Commission. Our aim is to enhance the quality of Systran, to broaden its application to the languages of the Community and to explain and vigorously promote its use.

USE AND USERS OF RAW MACHINE TRANSLATION

INCREASED USE OF MT

I intend to discuss, mainly, new progress in Systran, much of which has been stimulated by the recommendations of the 1991 Oakley Report. As you know, this was an evaluation of the Commission's Multilingual Action Plan. But first, let me comment on the growth of raw machine translation in the Commission. In 1988 only 4 000 pages of MT were processed. By 1993, this figure had risen to 120 000 pages. The widespread use of machine translation in the Commission is therefore a fairly recent phenomenon. Previously the technical conditions for its use were not met. Documents were not prepared in a form that the machine could read. Direct access to the system was not possible. There was no clear determination by management to promote MT. Not enough effort went into marketing it. On the technical side, these past five years have seen important developments in the Commission's use of computers. Input for MT is text that can be read by a machine. The main difficulty in the past was the cost of getting text to the computer. Now, with electronic files, it is fairly easy to send documents from one department to another by electronic mail. Scanning equipment is also more readily available.

Coming to the past year and a half, between the first six months of 1993 and the corresponding period of 1994, there was a 47% increase in the output of raw machine translation. Much of this may be attributed to an intensive promotion campaign conducted in tandem by the Translation Service and Directorate-General XIII (Telecommunications, Information Market and Exploitation of Research). Informative brochures on machine translation were distributed to all members of staff; posters have been put up in all Commission buildings; MT users can call a help-desk should they have any queries or difficulties; and regular visits are paid to user departments to answer questions about the system and to discuss the possibility of introducing specific terminology.

SYSTRAN AND THE PROFESSIONAL TRANSLATOR

Professional linguists are no doubt wondering about the reactions of in-house translators to this exponential rise in MT. Well, one can hardly say they have been enthusiastic. For, as long as MT was felt to be a substitute for human translation, it was unlikely they would welcome it with open arms. They can be reassured. The computer can never match the skills of the translator. The aim of our promotion campaign was not to seek to impose machine translation on professionals who knew full well that raw machine translation as such was not going to revolutionize their daily work, nor to advertise the translating machine as a competitor, but to allow non-linguist staff in the various departments to help themselves to machine translation as and when they needed it. MT complements, rather than substitutes the service translators provide. A clear distinction is drawn between the product of the machine and the work of the translator. All texts which are legally binding, or

which are meant for publication must remain in the hands of the professionals. Machine translation can be entrusted with only short-lived documents designed, say, for information or preparatory work and which are required urgently.

The Translation Service already faces a relentless increase in the volume of its work. If the output of Systran in 1993 was 120 000 pages, that of the translators was a million pages. And the service will have to cope with the new languages of an enlarged European Union. But it is not only the language service which is under great pressure. Multilingualism is a permanent feature of the daily grind in the various directorates-general. There is this British official. His French-speaking boss requires him to read, even draft, documents in a language other than his own. A memo is needed that day and he would have to wait his turn in the queue for that excellent job from the Translation Service. But help is at hand from Systran, which can give him a rough-and-ready version of his text in an average time of six minutes. MT does fulfil a useful role here as a makeshift solution when deadlines are tight and there are last-minute documents to produce, sometimes in several languages. Then, take the case of an official who is presented with a lengthy document in an unfamiliar language. Systran can quickly give him a broad view of its contents to browse over. According to his needs he may then ask for a professional translation of the whole paper or of only a few paragraphs, or he may discard it altogether. In this way the Translation Service is saved much unnecessary work. I will have something to say, later, about the use of Systran by the translators themselves.

POST-EDITING SERVICE

Now that machine translation is so freely available, it becomes essential to monitor its use and provide appropriate backup measures. Special care has been taken in all publicity material to stress that machine translation should be reserved for ephemeral documents and is to be considered as a stopgap solution to language difficulties encountered in day-to-day work. Inevitably, the unthinkable happens. A raw translation slips through the net and finds its way onto a Director's desk in the form of an official document. To facilitate immediate identification of a raw text a warning message "!!RAW MACHINE TRANSLATION!!" now appears every 300 words or so in the text to lessen the risk of any confusion. It is also recommended that when a post-edited document is used for further distribution, a header is inserted drawing the reader's attention to the fact that he is reading the humanly revised output of a machine.

It is with the express aim of providing a better backup infrastructure for MT users that a post-editing service is being established. The recent survey on the use of MT at the Commission showed that there are a very large number of MT users who correct their texts themselves or who have them corrected by colleagues who are fluent in the target language. Post-editing has been set up to offer additional help to these users, and to those who do not have the ability to post-edit within their own department. These people are now able to call on a rapid post-editing service which relies on a network of freelance translators who have expressed an interest in this

type of work. Not only are MT users relieved of the time-consuming task of correcting their raw translations, but tighter linguistic checks can be kept on the treatment of urgent texts. The quality offered by this service is at a level which is acceptable for purposes other than publication (working documents, internal notes, etc.) There are numerous examples of users who are quite happy to accept a document in less than pristine prose, provided they can have it when they really need it.

This preselection of the right type of text is of first importance. Requests for the post-editing service are examined carefully to ensure that they are in fact suitable for this kind of treatment. In our institution these will mainly be ephemeral documents of a routine, administrative nature, such as technical fiches, preparatory working documents, minutes of meetings, studies, and so forth. Users of Systran have expressed a keen interest in this service, provided we can respect very tight deadlines. Since speed is of the essence, the documents are transmitted to and from requesting departments and freelance post-editors entirely by electronic mail. Requesters make their own assessment of translations. Users of the scheme are well satisfied and deadlines are being met. The language pairs most used, in view of their higher quality, are English-French and French-English.

So far the service has been kept within very modest proportions, the "grapevine" being our main source of custom. At the moment, with no further promotion it is estimated that production for 1994 should be in the region of 3 000 pages. Freelance resources are limited at the present tune and a call for tenders is envisaged to set up a network of post-editors in order to intensify and promote the service.

What do we ask of our freelance post-editors? An MT user is not looking for perfection. He needs information in another language quickly. What the post-editing service offers is not really translation, nor indeed revision. The free-lance post-editor is simply aiming to produce a grammatically correct text without undue attention to stylistic detail. The amount of correction depends on the skill of the post-editor, but above all, on the quality of the raw translation, which can vary considerably with the quality of the source text and its suitability for machine translation. If there is one rule which can be applied, it is that of economy of means. If the text is suitable, the post-editor will choose the most direct route, the simplest solution, and resist the temptation to introduce his own linguistic refinements. The job is not translation itself, but it requires an experienced, fast and efficient translator, who can make a text comprehensible by means of the least number of changes. The skill of the post-editor resides in his ability to judge the seriousness of mistakes and to determine to what extent they need to be corrected.

In addition to this external post-editing service, a limited amount of post-editing is also done within the Translation Service itself. In a number of isolated cases Systran is used as an aid for human translators. Suitable documents are identified

by in-house translators from their own unit's workload. An important spin-off from these post-editing activities within the Translation Service is the feedback that can be channelled to the MT development team. The incorporation of feedback into the Systran dictionaries involves a number of stages, the first of which is the detection of suitable documents and the analysis of Systran translations from the point of view of terminology. Those terms which are deemed to benefit the entire user community are then coded in Systran's general dictionaries with a view to obtaining an acceptable quality in the target language. Specific dictionaries are then created for atypical errors and expressions in relation to the general Systran dictionaries, so as not to affect the overall stability of the system. And the final stage in the process is post-editing proper on PC. The experiment has been judged favourably by the translators involved. Two updates have been made to the system to incorporate the feedback of translator/post-editors and the improvement in the quality of the translation has been judged to be most encouraging.

INFORMATION FOR USERS AND HELP DESK

Contact with users in the operating departments was made easier by the setting up of a machine translation help desk, which also manages the post-editing service. It is used a great deal. Good communication with users of machine translation is a vital part of the operation. Users' expectations of the product of the machine should not be unduly high. They need to be warned about the quality of MT output if they are to avoid disappointment. Moreover, they need to know which type of document lends itself to machine translation. Some attempts have also been made to introduce the notion of "writing for the machine". So far, these have been limited to recommendations regarding simplified syntax in the drafting of texts prior to their submission for MT. Straightforward directives regarding the formatting and transmission of texts have been more successful. Meetings are held with specific departments to explain the importance of feedback and answer any questions raised. The help desk also offers scanning facilities for those departments which are not so equipped. The kinds of problems encountered by users range from lost texts, to being unfamiliar with electronic mail procedures.

THE USERS

When the raw machine translation service was first offered for general use on a help-yourself basis, we had little idea who was using the system. Requests came from machine numbers and cryptic passwords which could not be identified. Identification has since become much easier and the creation of a data base of users has enabled us to keep a close track on who they are. There are about 2,000 users, 20% of whom are in the Translation Service and 80% in the other Commission departments. Certain departments make much more use of MT than others, depending on the type of their work and their specific informatics environment. Of the total number, 30% (about 700) are regular users, that is they have requested at least five translations per month.

Once users had been identified, the next logical step was to ascertain their needs. An in-depth survey on the use of machine translation at the Commission was carried out. Officials were interviewed personally using a questionnaire specifically devised for the purpose. Data from the survey, over a period of twelve months, shows that the system is predominantly used for the translation of short texts (2 or 3 pages) and for correspondence, minutes of meetings, summaries, notes or reports. Interventions on the original text are infrequent and are limited to a spelling check or the formatting of the document. Use of simplified syntax when drafting the original is very rare. On the other hand 90% of users correct the raw versions, in most cases with the help of colleagues whose mother tongue is the target language. Contrary to what we supposed, the vast majority of post-edited texts are not limited to internal diffusion but are destined for a wider audience. It is reasonable to predict that the use of machine translation for browsing purposes (i.e. as a reading tool) would be much greater if the lesser used languages were available as source languages.

To summarise, raw machine translation has three distinct applications within the institution. First, in the operating departments, it is used as a translation tool, particularly for urgent or short documents which cannot be handled in time by the Translation Service. In the Translation Service itself, as a result of the applications currently being developed, interest is shifting towards exploiting the system as a terminology pre-processing tool. Secondly, it is often used in Commission departments as a drafting tool. The author requests a Systran translation when he is required to write in a language other than his native tongue. Finally, to a limited extent, because of the specific language combinations available, it is used as an information tool for browsing purposes: the Systran translation is requested to enable the reader to understand a text written in a language with which he is unfamiliar. He can decide to ask for a translator's help with the whole or only part of the text, or to discard it if it is not relevant to his needs.

TECHNICAL DEVELOPMENTS

The increase in the number of languages, texts, and specialized areas which the Translation Service must cover has led it to seek ever more advanced technical solutions in order to fulfil its mission at the lowest cost. On the one hand, attempts are being made to improve the structure of written communication by rationalizing and standardizing original text. On the other, the memory, speed and capacity of the machine is being placed at the service of officials working with many languages.

INFORMATICS

The Commission has decided to switch from a Unix-based microcomputer network as its primary working environment to a personal computer network. This has meant a lot more work for the informatics experts of Systran. Access to the system had been tuned to the Unix word processing system and the decision to tolerate two word processing packages (WordPerfect and Word for Windows) within the

institution did not simplify matters. Absolute priority was therefore given to extending the range of formats accepted by Systran (WordPerfect and WinWord). The interim period has been trying, with both Unix and PC word processing systems in widespread use throughout the Commission. Systran is particularly sensitive to any invisible codes lurking in the texts after conversion and these can cause problems of analysis. These procedures are still being stabilized and users will have to exercise a little patience before a complete service can be offered, particularly for Greek.

BRIDGES

Following the switch to personal computers at the Commission, a new, improved user interface had to be created. From our contacts with potential customers of machine translation it was clear that lack of familiarity with informatics access procedures to Systran constituted a barrier to its use and a user-friendly interface for Windows was needed. It has been designed to guide the user through the different stages of his request.

At the time various experiments were in hand. They explored ways of making the most of the machine-translation system by exploiting its potential as a text pre-processing tool. It was suggested that the new interface could usefully be extended to make a number of different multilingual tools generally available, both within the Translation Service and throughout the various Commission departments. The new interface, affectionately known as Euramis, would offer three distinct products in addition to raw machine translation:

1. identification by Systran of Celex references (directives, regulations, etc.) in the source text, provision of complete titles in the target
2. Eurodicautom terminology look-up from text
3. Eurodicautom terminology look-up from user's list of terms

1. *Celex bridge*

One example of highly successful synergy between information tools is the creation of a bridge between the MT system and Celex, the multilingual data base containing Community legislation in the nine official languages. A large proportion of the original documents received by the Translation Service's planning units contain references to titles of legislative acts, such as regulations, directives or decisions. Translators spend valuable tune searching in the Celex data base or in the Official Journals to check that the title is correctly expressed in the appropriate target language. Every document in the Celex base has a unique reference number, which is the same for all language versions of that document. A specific algorithm was devised: any references to Community legislation contained in a source document are recognized at the analysis stage of the MT process; the reference number is automatically generated; and a search is made in the relevant target version of the Celex data base. The correct title, along with its publication reference, is then

reproduced at the end of the raw translation. Hence, a routine has been integrated into Systran, which makes it possible to extract automatically from the Celex base the title(s) in the target language corresponding to the reference number mentioned in the source text. The incorporation of this routine has not only proved to be an extremely useful tool for translators. It has also opened the door to other ways of turning the machine translation system to account. Why could the Systran text analyzer not be used as a means of exploiting other types of text pre-processing?

The Informatics Department of the Translation Service had in fact developed two e-mail based servers which provided multilingual services entirely automatically. One server handled raw machine translation requests. The other provided batch look-up of Eurodicautom, the Community's nine-language Terminology Data Bank, looking up lists of terms in a given source language and returning corresponding terminological data in one or more target languages. Both servers were based on common principles and a common software infrastructure. Consequently, it was a relatively simple matter to establish bridges between the servers in order to provide new products.

2. ***Terminology look-up of text***

One approach seemed particularly interesting. The idea was to combine Systran source-text analysis with Eurodicautom terminology look-up. In this way a system was constructed which identifies possible terminology within running text and then provides the relevant Eurodicautom entries in one or other target language. As an experiment, a bridge between the existing Systran and Eurodicautom servers was established. The procedure is quite simple and was developed entirely from existing possibilities. Take an example based on English-French. The English text is first introduced into Systran for basic analysis. The output from Systran is not, however, any kind of translation, but simply a list of English terms which have been recognised in the Systran dictionaries, following syntactical and morphological analysis of the text. The English expressions are then looked up in Eurodicautom and the corresponding French data extracted. The combined English and French Eurodicautom data is returned to the requesting user. In short, bilingual terminology can be generated automatically from an arbitrary text. The limiting factor is the number of source languages that Systran can analyze. However, for each of the four source languages Eurodicautom can provide eight target languages. Consequently a Systran/Eurodicautom hybrid can support a total of 32 language combinations. Automatic terminology look-up can therefore be provided for language combinations such as French-Danish, which do not exist in Systran at all.

Initial tests revealed a number of weaknesses. At first the Systran hit-rate was too low (not enough potential Eurodicautom terminology was recognized). The Eurodicautom hit-rate was too high with too much Eurodicautom data in output. Finally, the presentation of Eurodicautom output needed to be refined. These problems have been substantially reduced by the coding of Eurodicautom terminology within Systran (of which more anon), which has significantly enhanced the initial hit rate. At the same time refinement of the Eurodicautom batch programs will reduce the amount of irrelevant data provided.

3. **Eurodicautom batch queries**

The third option is terminology batch queries. Lists of terms in a given source language are submitted to the Eurodicautom server which returns the output in one or several target languages. All current language combinations (72) are supported. For those with a more specialized interest in terminology, the interface will offer filters for queries, enabling the user to determine the amount of information that is required, such as definitions, references and so on. Subject fields can be indicated and the scope of the answers can be controlled by selecting the desired level of match of text items.

The embryonic development of the new graphical interface, Euramis, continues. It has been available for testing among a restricted population since the summer. The prime concern of the developers is that it should be easy to use. Eventually, when it is on general release, it should enable the uninitiated and, hopefully, even the computer-shy to visualize the various multilingual products offered by the Translation Service in an integrated package.

IMPORTATION OF EURODICAUTOM

Returning to Eurodicautom. Here were two rich and extensive sources of terminology, Systran and Eurodicautom, sitting side by side and functioning independently. Surely this was pointless and wasteful. Why not enrich Systran with the resources of the Community's terminology data bank? But there were daunting technical problems in the way of bringing the two together.

The main obstacles to the success of the operation were the three fundamental differences between Eurodicautom and Systran dictionaries:

EURODICAUTOM	SYSTRAN	SOLUTION
no grammatical information available	basic grammatical information needed	• default rules (for word class, gender etc.) on the basis of existing dictionaries • automatic detection of the principal word in multi-word units
several solutions for one term in the same subject field	choice of one solution necessary	• automatic detection of "best possible" equivalent • remaining solutions will be stored in comment lines
domain-oriented subject fields	user-oriented topical glossaries	• for the time being: filter between Systran topical glossaries and corresponding Eurodicautom subject fields[1] • eventually, automatic detection of subject field by means of statistical methods

Comparative tests of Systran dictionaries with and without the Eurodicautom entries are being carried out to measure qualitative improvements in the translation of technical texts. Further tests are being run at various levels relating to the accuracy of linguistic strategies concerning the different language pairs and to general strategies concerning dictionary look-up. The procedures will have to be refined as errors occur and improvements will be part of normal development work. The informatics procedures for updating the dictionaries are being adapted to the new structures and types of information. During the work on the importation of Eurodicautom into Systran, important feedback has been forwarded to the Eurodicautom team concerning such things as missing codes, spelling mistakes.

The main benefits to be anticipated are, on the one hand, improved quality of Systran translations for all texts of a technical nature, particularly in those fields insufficiently covered by Systran, and, on the other hand, the development of Systran as a terminological pre-processing service.

[1] *The Eurodicautom subject field classification is more detailed than the Systran topical glossaries (e.g. one topical glossary for Biology, Medicine, Chemistry and Environment, with four corresponding subject fields in Eurodicautom).*

THE FUTURE

LANGUAGE POLICY

Let us look to the future, first of all in regard to language policy. If we consider the breakdown of production for the various language pairs, French-English and English-French are by far the most used, reflecting the higher quality of these linguistic combinations in Systran. However, the user survey revealed that nearly 90% of users wish for an improvement in the quality of German as a source language to enable rapid reading. The priority given to the other source languages, in decreasing order, is Dutch, Danish, Greek, Italian and Portuguese.

What are our priorities for future language development in machine translation? Internal communication is in fact covered by the French-English and the English-French pairs, where quality has reached an acceptable level, provided the right type of text is submitted for processing. Our immediate concern is to improve German in the system, both as a source and as a target language. German is a language in which not all Commission officials are proficient and in which there is a great deal of written communication. In the longer term, the strategy is to reverse the pattern of development of language pairs from non-vehicular source languages into the main languages of communication within the institution, and indeed tolerance of machine quality is highest for these combinations. Machine translation should be made available from the lesser-known source languages into the working language they most resemble (from Italian, Portuguese and Greek into French; and from Dutch and Danish into English). Hence, browsing requirements can be met and Commission officials who are not fluent in one of those languages can obtain rudimentary translations of documents written in less widely known languages. As we have seen, they may, in fact, use the raw machine translation as a means of selecting specific sections of a document for further, human translation.

In short, the priorities for language development are threefold:

a) consolidation of the three basic pairs of the system between the three working languages of the institution, and hence, priority given to German as a source and target language.
b) development or acquisition of language pairs with non-vehicular source languages into one of the working languages (i.e. the one with which it has the most affinity).
c) promotion of co-financing for non-priority languages.

LINGUISTIC DEVELOPMENT

General linguistic development work is always based on feedback sent by users via the Systran promotion team in Brussels. This systematic work on "live texts" is necessary to improve the linguistic content of existing language pairs. One aspect of linguistic development involves the introduction of terminology specific to individual

departments. In addition to feedback from regular clients the list of not-found words generated automatically at the end of each translation is encoded. The second aspect of this development work, the improvement of programmes, is based on systematic detection of errors in the analysis or synthesis programmes. Errors are archived according to their type, frequency and importance. Development is based on a strict hierarchy. The most serious and most frequent errors are dealt with first. Development work varies according to the "age" of the couple. The youngest couple, German-French, for example, contains enormous gaps, both at the level of the dictionaries and at the level of the programmes, whereas work on French-English, the Darby and Joan of the Commission's system, is concentrated mainly on well-targeted texts in fields for which specific needs have been identified. The rate of progress of a language pair is also dependent on the number of people in charge of its development. It is therefore necessary to spread as best as possible the available resources according to the priorities which have been fixed. In recent months priority has been given to the importation of Eurodicautom into Systran but once this operation is complete there will be monthly "low-risk" updates. An acceleration in the rate of updates can increase the level of satisfaction of regular users thanks to a more rapid adaptation of the system to their needs. Every effort is made to maintain close collaboration with users.

With the number of MT requests constantly on the increase, some organizational measures had to be taken. Every text processed by Systran is now classified according to the type of document, the domain (according to the Eurodicautom classification system), the requesting server and the number of pages. A corpus has thus been created which is updated daily. This corpus constitutes a basis for evaluating the development of the system. Tests can be carried out at any moment on a specific type or domain.

Some new aspects are being handled by the development team. Until recently it was not possible to "teach" the system how to translate sentences which occur regularly in repetitive texts. A programme has now been implemented, in a pilot version, which recognizes fixed sentences and integrates them into the Systran output, replacing them by their pre-defined translations. Certain types of variation can be handled, but the system is not as powerful as "standard" translation memories in that it does not treat fuzzy matches to the same extent and translation equivalents have to be established sentence by sentence. In other words, there is no automatic alignment for whole documents. Its advantage lies in that it is already fully integrated in Systran. Our ultimate aim, however, is to introduce an existing powerful translation memory into the translation process. For this reason, the pilot version is being used by a limited population within the Translation Service.

FUTURE NEEDS

The perceived needs within the Commission for machine translation fall into three categories:

1. For documents coming into the Commission. The main need is for a system that will allow browsing, in the sense that a reader should be able to follow

the general argument of a text with sufficient confidence to know whether it merits more accurate or in-depth treatment.

2. For documents circulating within the Commission. A machine translation system is required for understanding, as an aid to drafting and as an aid for the preparation of working documents in the three working languages (French, English and German).
3. For the preparation of documents to be disseminated outside the Commission. Here the need is perhaps not for machine translation as such, rather for tools which would facilitate the preparation of high quality, publishable translations.

It may well be that these different usages might imply the use of different types of systems.

CONCLUSIONS

I should like to conclude with the following remarks. First of all, potential users of MT must be able to draw a clear distinction between the product of a machine and the work of a human translator. They must know which type of text can safely be entrusted to the machine. The product of the machine is useful for ephemeral texts of an informative nature and preference should be given to simple texts where certain stylistic rules have been respected. If all these conditions are met, the machine will be able to cover an ever greater share of the requirements of multilingual communication within an institution such as the Commission. Machine translation has to be presented as a rapid communication tool for use by staff in the various operational departments and not only - as in the past - as a tool for translators.