THE USE OF MACHINE TRANSLATION IN THE COMMISSION

Dorothy Senez
Translation Service, European Commission, 200 rue de la Loi, 1049 Brussels, Belgium
E-mail: DOROTHY.SENEZ@SDT.CEC.BE

ABSTRACT

Systran, the European Commission's machine translation (MT) system is an extremely rapid service which is available to all Commission officials via the internal electronic mail network. The widespread use of MT in the Commission is fairly recent. In 1988 only 4 000 pages were processed by MT. In 1994 this figure had risen to 140 000 pages as a result of a promotion campaign and improvements to the Commission's computer infrastructure. An in-depth study of our use of MT reveals that there are about 2,500 users, 20% of whom are in the Translation Service and 80% in the other Commission departments. The majority of users are non-linguist staff who help themselves to machine translation as and when they need it. Users' expectations of the product of the machine should not be unduly high. MT can be entrusted with only short-lived documents required urgently for information or preparatory work. The professional translator must continue to be responsible for all texts which are legally binding or which are for publication. Raw machine translation has three distinct applications within the operational departments of the institution. First, for translation, particularly for short or urgent documents. Secondly, for drafting, when an author is required to write in a language other than his native tongue. Finally, to a limited extent, because of the specific language combinations available, for information: 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 translato's help with the whole or only part of the text, or to discard it if it is not relevant to his needs.

Now that machine translation is so freely available, it becomes essential to monitor its use and provide appropriate backup. MT users can call a help desk should they have any queries or difficulties. A post-editing service has been set up to correct machine texts for users who cannot do this in their own departments. The quality offered by this service is at a level which is acceptable for purposes other than publication.

The dictionaries are continually being expanded and have been enhanced recently by the importation of a significant number of entries from the Eurodicautom terminological database. Other important technical developments have been the construction of a bridge between the machine translation system and CELEX, the database containing Community legislation, and the creation of a SYSTRAN/Eurodicautom hybrid enabling terminology to be extracted from a text. Raw machine translation is only one of a number of multilingual services now being made available. In the Translation Service itself, as a result of the applications currently being developed, interest is shifting towards exploiting machine translation as a terminology tool in the pre-processing of texts.

INCREASED USE OF MACHINE TRANSLATION

The SYSTRAN machine translation system was created in the USA in the mid-sixties and it was introduced to the Commission in 1976 when work started on the pilot English-French version. However, it is only recently that the use of MT by the Commission became widespread. In 1988 only 4 000 pages of MT were processed; in 1994, the number had risen to 140 000. Currently, production averages over 18 000 pages per month, and requests come from the majority of departments. In the Commission today an expanding MT system emerges less as an aid to translators, and much more as a means of communication for staff. It has been notoriously difficult to provide clear evidence that MT speeds up the work of in-house translators. The generalized use of MT may therefore be seen by them as a threat. Where it exists, the fear is unwarranted. The Translation Service will continue to wield the sole responsibility for all documents which are legally binding or are destined for publication. And if the output of Systran in 1994 was 140 000 pages, that of the professional translators was over a million.

THE OAKLEY REPORT AND ITS FOLLOW-UP

An important milestone in the history of machine translation in our institution was the commissioning of an evaluation of the SYSTRAN system by a panel of outside experts chaired by Dr Brian Oakley. The main findings, published in November 1991, were that, in view of its many language pairs and powerful dictionaries, SYSTRAN was the most suitable system for the Commission over the short term. In the medium term, however, it was recommended that it should be adapted to modern programming practices. The Oakley report concluded that SYSTRAN, though not very well known, met a real need. The most pressing task was to identify user requirements and adapt SYSTRAN accordingly. Development had now to be geared first and foremost to users' requirements. We had to promote and market better, to develop back-up services. We set up a help desk for MT users to call should they have any queries or difficulties. The help desk also manages an experimental post-editing service and offers scanning facilities for those departments which are not so equipped. Work was to be concentrated on such matters as promotion, unproved interfaces and technical infrastructure.

Promotion

Much of the increased use of MT may be attributed to an energetic promotion campaign conducted in tandem by the Translation Service and Directorate-General XIII (Telecommunications, Information Market and Exploitation of Research). We produced a brochure and distributed it throughout the Commission. We also visited departments which were using MT to answer questions and to discuss the possibility of introducing specific terminology . Our advice was that one should not expect too much of MT. But if it was "far from perfect" it was very fast; any member of the Commission staff could help himself to MT in a matter of 5 minutes. It was also free. Staff who were not linguists could call on it as often as they needed it. MT provided a makeshift solution when deadlines were tight and some sort of translation was demanded at the last minute and perhaps in several languages. A hard-pressed official might well be satisfied with translations in less than deathless prose provided he could have them immediately. Users had to discipline themselves in the matter of syntax. Spelling mistakes, errors in format,

long rambling sentences would make for poor results. The machine could be entrusted only with short-lived documents designed, say, for information and preparatory work and which were needed urgently. All texts which were legally binding or were to be published had to be referred to professional translators. By giving us feedback, users could help to improve the system and ensure that their own needs were met. The more the system was used the better it would become. If simple texts were presented to it and certain rules of style were followed, the machine would play its part in meeting the Commission's linguistic problems - new languages, new areas of specialization and ever more paperwork.

Technical infrastructure

The increase in the use of MT was not due solely to our promotion efforts. The past five years have seen important developments in the Commission's computer infrastructure. Previously the technical facilities were inadequate. Direct access to the system was not possible. Processing time in the server and in SYSTRAN itself is negligible. Total turn-around time, which is primarily a function of the speed at which texts can be transported round the telecommunications network, has now been reduced to a matter of minutes thanks to increasingly reliable electronic mail.

New interface

We set out to achieve better integration of MT into the users' informatics environment. 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. The new interface for Windows has been called EURAMIS and has been designed to guide the user through the different stages of his request. It is at present being tested by the Translation Service and is shortly to be distributed throughout the Commission departments. Machine translation will then be an integral part of the standard PC menu, just a tool among others such as a word processing package or an e-mail option. With the generalized distribution of EURAMIS we can no doubt expect to see yet another sharp rise in the use of MT within the institution.

THE USERS

The Oakley report states “Future investment should start from the users’ requirements”. Accordingly, we set ourselves the task of finding out who our users were and what they were doing with their raw MT. Users are widely dispersed between two sites in Luxembourg and 40 different buildings in Brussels. There are research centres in a number of Member States. Most users work on personal computers connected by LAN to UNIX communications servers. Provision is centralized, on a client-server basis via local area network. This approach views MT primarily as part of a communications system rather than part of a document handling one.

Once users had been identified, we studied their needs. In a survey, regular users over a period of several months were interviewed by means of a questionnaire specially devised for the purpose. There are about 2,500 users, 20% of whom are in the Translation Service and 80% in the other Commission departments. Certain end-user 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) can be considered regular users, that is they have requested at least five translations per month. Even the most enthusiastic among them, however, are relatively infrequent users. Few people are likely to need a translation of any kind more than a few times a week and the statistics bear this out. A high level demand, therefore, amounts to an aggregate of a small number of individual demands. The data showed that the system is predominantly used for the translation of short texts of 2-3 pages, correspondence, minutes of meetings, memos, and such like. An overworked Translation Service cannot always produce such documents in time.

APPLICATIONS

Broadly speaking, the MT system at the Commission is being used first and foremost as a translation tool, not by professional translators, but by non-linguists in the operational departments, and as an aid in the drafting of working documents in the three main working languages (French, English and German). It's use as an information or reading tool, widely recognized as the classic market for machine translation, is limited. There are two mam reasons for this. The first is the lack of coverage of the less frequently used language pairs as source languages. 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 lesser used languages were available as source languages. The second reason is the lack of adequate scanning facilities.

A distinction can usefully be drawn between three main categories of application: browsing, "translation", and drafting. End users in the operational departments are concerned with all three categories but in-house translators fall under the second category only.

1. **Browsers** do not know the source language, and use MT as an aid to understanding it. A raw translation is required for information purposes 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. Faced with a document in an unfamiliar language, MT output can quickly give him a broad view of its contents and, according to his needs, he can decide to ask for a professional translation of all or part of the text, or he may discard it altogether.

2. **"Translators"** know both source and target languages and use MT as an aid to translation into their own language. The large majority of these users are administrators required to produce their own translations in the course of their normal work. Translation is not their primary activity. Experience has shown that there users' requirements are not stringent. Professional translators working in-house using raw output with the aim of producing a "faultless" final product have very high quality requirements.

3. **Authors** know the source language well. They use MT as an aid to translating from that language into another which they may not know well or not know at all. In some cases the user cannot assess the quality of the output. This specific use of MT is potentially dangerous and needs to be carefully identified and monitored. To avoid any risk of confusion and to facilitate immediate identification of machine output, a decision was taken to incorporate a warning message "!!RAW MACHINE TRANSLATION!!", which appears automatically every 300 words or so in the machine text returned to the requester. A Commission official is often required to draft documents in his director's language rather than his own. He will naturally express his thoughts more readily in his mother tongue, and he can usually get the help of a colleague whose language is the target one to correct his final version. The survey showed, reassuringly enough, that 90% of users do correct the raw versions, either themselves or with the help of colleagues who are fluent in the target language. Contrary to what we supposed, the vast majority of texts post-edited in this way are not limited to internal diffusion but are destined for a wider audience.

RAPID POST-EDITING SERVICE

Even satisfied users find post-editing a burden and, as we have seen, the survey revealed a need to keep tighter linguistic checks on the treatment of urgent texts. Since machine translation is so freely available it became essential to provide a better back-up infrastructure, particularly for users in groups 2 and 3. To this end an experimental post-editing service has been established, offering additional help to those users who do not have the ability to post-edit within their own department. The post-editing service relies on a network of freelance translators who link with our department entirely by electronic mail, since speed is of the essence. The post-editor is asked to remove the actual errors introduced by the machine and to carve a grammatically correct and workable text. The amount of correction depends on the skill of the post-editor in judging the seriousness of mistakes and in determining to what extent they need to be corrected. He must resist the temptation to introduce his own linguistic refinements. The job is not translation itself, nor is it really revision, but it does require an experienced, fast and efficient translator, who can make a text comprehensible by means of the least number of changes. The texts we process are mainly routine, administrative documents (working documents, internal notes, etc.) and the quality offered by this service is at a level which is acceptable for purposes other than publication.

Requests for rapid post-editing of MT output come to us directly from end-user departments and are examined carefully to make sure they are suitable for this type of treatment. This prior selection of the right type of text is of first importance. Users of SYSTRAN have expressed a keen interest, provided we can respect very tight deadlines. They make their own assessment of translations and are well satisfied with the service. The language pairs most used, in view of their higher quality, are English-French and French-English. So far the volume of work has been kept within very modest proportions. Freelance resources are limited at the present time and a call for tenders is envisaged to set up a network of post-editors in order to intensify and promote the service.

LINGUISTIC DEVELOPMENTS

Every effort is made to enlist the cooperation of users. General linguistic development work is based on their feedback. This systematic work on "live texts", involves the introduction of terminology specific to individual departments and is necessary to improve the linguistic content of existing language pairs. It varies according to the maturity of the language pair involved and is limited by the resources available. Central dictionary updates for "lower-risk" terminology entries have been reduced to monthly cycles, and this acceleration in the rate of updates will increase the level of satisfaction of regular users thanks to a more rapid adaptation of the system to their needs. The more complex procedures required for the systematic detection of errors in the analysis or synthesis programmes, are still carried out on a three-monthly basis.

In isolated cases, MT is used as an aid for the human translator within the Translation Service itself. Suitable documents of a repetitive nature are identified by a small team of in-house volunteers from their own unit's workload. Most of this type of work is being done in the French, English, Italian and Spanish units. An important spin-off from these targeted post-editing activities within the Translation Service is the feedback that can be channelled to the SYSTRAN development team, who introduce terminology specific to recurring documents of a repetitive nature. 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. In the case of German target, the current challenge is to fill in the serious gaps and tackle some basic difficulties which until now have prevented German from becoming viable.

LANGUAGE PAIRS

Although the system was initially considered to be of bilingual design, a more modular approach was developed in the mid-eighties. This enabled analysis in one language to be applicable to others. Now, the development team in Luxembourg is working on 17 language pairs:

seven from English (into French, Italian, German, Dutch, Spanish, Portuguese and Greek),
five from French (into English, German, Dutch, Italian and Spanish),
two from German (into English and French)
two from Spanish (into English and French),
one from Greek (into French).

All but one of these (Greek-French) are available for use by any Commission official via the internal e-mail network. Machine output for French-English and English-French is satisfactory, provided the right type of text is submitted for processing. The quality of the other pairs varies considerably, depending on the tune spent on development and on the syntactic and lexical affinity of the languages concerned.

The statistics show that French-English and English-French are by far the most requested combinations, and this reflects the sustained efforts of development over the years. However, the user survey revealed that nearly 90% of users wish for an improvement in the quality of German in the system. German is a language in which not all Commission officials are proficient and in which there is a great deal of written communication. The Germanic languages have proved more difficult to develop. However, in response to the wishes of our users, extra resources have been allocated to language pairs involving German, and encouraging progress is being made.

In the longer term, the strategy is to reverse the pattern of development of language pairs from lesser-known source languages into the mam languages of communication within the institution, because machine output is more readily tolerated if it is required only for the purpose of gathering information. Machine translation should be made available from 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, rudimentary translations of documents written in less widely known languages can be obtained for browsing. The survey revealed a special need for machine translation from Russian into English. And naturally there is a potential requirement not only for Scandinavian but also for Eastern European languages as sources.

Continued development of existing language pairs will depend on the availability of resources. The better developed pairs still require on-going enhancement. They are showing particularly good results when development is targeted to specific texts in restricted domains. But the statistics clearly show that a number of the language pairs currently on offer at the Commission are not yielding good returns on our investment.

In short, the priorities for language development for the Commission's needs are: 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; development or acquisition of language pairs with non-vehicular source languages into one of the institution's working languages. Interest has been expressed in repeating with other Member States the highly successful experience with the Greek Government in the co-financing of the English-Greek combination.

ENHANCEMENTS TO THE SYSTEM

Importation of Eurodicautom

The most important enhancement made to the system recently has been the importation of data from Eurodicautom into the SYSTRAN dictionaries. It was wasteful for two rich and extensive sources of terminology, SYSTRAN and Eurodicautom, to be sitting side by side and functioning independently. Why not enrich SYSTRAN with the resources of the Community's terminological database? To achieve this a second, external dictionary had to be grafted onto the existing MT dictionaries. But the main obstacles to the success of the operation were the fundamental differences between the Eurodicautom and SYSTRAN dictionaries. The scope of a database differs from that of an MT dictionary. A correct equivalent in one can be an inconsistency for the other. Eurodicautom is a descriptive data base, offering various alternatives for a given term, stating its sources, adding definitions and synonyms. SYSTRAN dictionaries on the other hand have a one-to-one structure,

allowing one single translation per word or string of words in a given subject field. Moreover, there is a lack of basic grammatical information in Eurodicautom, and this is essential to SYSTRAN. Finally, the Eurodicautom subject field classification is more detailed than the SYSTRAN topical glossaries. Despite the many difficulties encountered the project has been successfully completed. Comparative tests of SYSTRAN dictionaries with and without the Eurodicautom entries have been carried out. The main benefit to be anticipated is improved quality of SYSTRAN translations for all texts of a technical nature, particularly in those fields insufficiently covered by SYSTRAN. But, even more significantly, the Eurodicautom experience has made clear that SYSTRAN can be adapted to other needs.

Replace function

At the request of a number of our in-house post-editors, it is now possible to teach the system to translate entire sentences which occur regularly in repetitive texts. In a pilot programme, fixed sentences are recognized and integrated into the SYSTRAN output, replacing them by their pre-defined translation. 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.

NEW DEVELOPMENTS

As I mentioned earlier, use of the machine translation system as such by in-house translators is extremely limited. However, by turning the MT system to account, SYSTRAN can offer help to these translators by relieving them of some the time-consuming work involved in checking references and terminology. Tools which automate these often tedious, but necessary tasks would certainly be appreciated. Hence, interest within the Translation Service is shifting towards exploiting the system's potential as a database accessing tool and as a terminology tool.

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 database containing Community legislation in the nine official languages. A large proportion of the documents coming into the Translation Service contain references to titles of legislative acts, and translators spend valuable time checking 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 whereby any references to Community legislation contained in a source document is 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 returned to the requester. Hence, a routine has been integrated into SYSTRAN, which makes it possible to extract titles automatically from the CELEX base. This has not only proved to be an extremely useful tool for translators, it has also opened the door

to new ways of exploiting the SYSTRAN text analyzer in the field of text pre-processing.

Eurodicautom look-up from text

Two separate e-mail servers had been developed at the Commission providing multilingual services entirely automatically. One server handled raw machine translation requests to SYSTRAN. The other provided batch look-up of Eurodicautom, 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. 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 a given text and then provides the relevant Eurodicautom entries in one or other target language. The procedure is quite simple and was developed entirely from existing possibilities. The 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 terms which have been recognized in the SYSTRAN dictionaries, following syntactical and morphological analysis of the text. The list of source expressions is then looked up in Eurodicautom and the corresponding data extracted in the target language(s). In short, terminology lists 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 data in output and the presentation of the output needed to be refined. For those with a more specialized interest in terminology, the user can determine the amount of information that is required, such as definitions or references. Subject fields can be indicated and the scope of the answers can be controlled by selecting the desired level of match of text items.

EURAMIS INTERFACE

So we really have come a long way in the last twenty years. In the beginning, input was prepared in IBM 80-column punch cards which were fed directly into the mainframe card reader and produced output on A3-size computer listings - a few hours later. Nowadays raw machine translation is only one of a number of multilingual services being made available to all Commission officials via the new EURAMIS interface. At this stage of development, EURAMIS offers 4 possibilities:

a SYSTRAN raw machine translation (now with Eurodicautom entries incorporated)

a link to the CELEX data base in 9 languages (which is also available by default along with a SYSTRAN translation)

translation by Eurodicautom of the terms contained in a text following analysis by the SYSTRAN system (32 language pairs)

translation by Eurodicautom of a list of terms created by a user (72 language pairs).

This interface has been available on an experimental basis to the Translation Service for a couple of months now and is already proving its usefulness. After final testing and modifications it is destined for wider distribution to the other Commission departments. All Commission officials will be able to benefit from a much more user-friendly access to machine translation. Already it is foreseen that this interface will gradually be equipped with further applications, including a vast translation memory.

CONCLUSIONS

To summarize, it would appear that machine translation is making excellent progress as a means of rapid communication in a multilingual institution such as the Commission and that there is a continuing requirement for a large-scale, robust, batch-processing system such as SYSTRAN. Much of the potential market is still to be tapped, and yet it is already is clear from the increase in growth that a demand for machine translation exists. MT has progressed over the years to become an operational system with definite applications within the institution, provided potential users draw a clear distinction between the product of a machine and the work of a human translator. Our goals are to enhance the quality of satisfactory language pairs and to extend the MT service to other Community languages in a service-oriented framework. With the forthcoming generalized distribution of the EURAMIS interface to all end-user departments, the overall demand for MT is likely to increase, perhaps even very sharply, and we must be prepared to meet that demand.

References

Oakley, B. et al (1991) Evaluation of the Commission's Multilingual Action Plan 1976-1991. Final Report. Luxembourg, internal document

Paesmans H. (1994) The Commission as User of Language Technology, Translation Service, Luxembourg, internal document

Paesmans H. (1994) The Translator's Tools, Translation Service, Luxembourg, internal document

Petrits, A (1994) The Current State of the Commission's SYSTRAN MT System, Translation Service, Luxembourg, internal document

Pigott, I (1992) SYSTRAN Development at the EC Commission 1976 to 1992, internal document

Senez, D (1995) Developments in SYSTRAN, Aslib Proceedings, vol 47, no 4, April 1995

Sixth Action Plan for the improvement of the Transfer of Information between Languages 1994-1995, Commission of the European Communities.

Telindus T.A. (1994) Final Report on the Importation of Eurodicautom Data into SYSTRAN Dictionaries August 1993 - December 1994, Luxembourg

Urquhart I. (1992) SYSTRAN - Delivering Machine Translation to Users, Translation Service, internal document