MT Summit VIII
Santiago de Compostela (Spain), 18 – 22 September 2001

THE COMMISSION'S MT SYSTEM: TODAY AND TOMORROW

Paper Type: U

Angeliki Petrits, Francine Braun-Chen, Jesús Manuel Martínez García, Cameron Ross, Rosemarie Sauer, Angelo Torquati, Alain Reichling

European Commission
Translation Service
Development of Multilingual Computer Tools
200, rue de la Loi
B-1049 Brussels
Belgium
angelique.petrits@cec.eu.int, francine.braun-chen@cec.eu.int, cameron.ross@cec.eu.int, jesus-manuel.martinez-garcia@cec.eu.int, rosemarie.sauer-stipperger@cec.eu.int, angelo.torquati@cec.eu.int, alain.reichling@cec.eu.int

Abstract

This paper presents a snapshot of how the Commission's MT system (EC SYSTRAN) is used today and a glimpse of how that picture will change tomorrow. It looks in turn at: the origins of the system; how it is accessed; who requests MT and why; how users can influence the quality of output; the Rapid Post-editing Service; and the latest usage statistics, which augur well for the future. The paper closes with a look at that future, touching on the move to a new computer platform and plans for new language pairs, concluding that after twenty-five years of development, MT has become an integral part of the Commission's working environment.

Keywords
Machine Translation, Users, Post-editing, Future, European Commission.

GENERAL BACKGROUND

History

The Commission has been involved in MT since 1976, when it acquired certain rights for the development and use of SYSTRAN from Peter Toma's World Translation Center in the United States. Until 1998, the Commission's version of the system – or *EC* SYSTRAN – was funded by the Directorate-General for the Information Society under a series of framework research programmes. However, when the system started to be used widely in-house, and therefore became an operational concern, control was passed over to the Translation Service.

Development commenced with English and French as both source and target, since they are the principal working languages in the EU institutions. English-Italian and English/French-German were added in 1978 and 1982 respectively, and other language pairs were gradually included as new Member States joined the EU.

Apart from including the main EU working languages, the principal development criteria were, on the one hand, the potential translation quality expected from a specific pair, and on the other, the participation of Member States. For instance, it was decided to develop French-Spanish because a satisfactory translation quality could be obtained within a short time. In contrast, English-Greek has been developed thanks to a co-financing agreement between the Greek government and the Commission.

Today the Commission's MT system provides 18 language pairs (see Figure 1 below) which are accessible not only to in-house staff, but also to the other EU institutions and some public-sector bodies in the Member States. Development has always been based on internal translation needs, and EC SYSTRAN has therefore been tailored to translate Commission documents.

Figure 1: Commission language pairs

Access

EC SYSTRAN is currently installed on a mainframe at the Commission's Data Centre in Luxembourg.

Access to the system can be obtained in three different ways:

1. By e-mail. Users send their source document as an attached file to a mailbox, specifying their requirements in terms of target language(s), text type and domain terminology, and get the translation back in the same way. The snag is remembering what terminology options are available!

2. Through a user-friendly interface on the EU Intranet, which displays all options offered by the system. Users can send documents to MT either by typing in a text box or by attaching up to four files, provided that they concern the same language combination. The translation is returned to their personal mailbox.

3. Through the EURAMIS (*European Advanced Multilingual Information Service*) interface, which on top of MT, offers different translation tools such as access to the Eurodicautom terminological database, CELEX (legislative database), and translation memory. Combination of these tools is also possible. For instance, the user can ask for *TM+MT*: translation memory retrieval with MT for the parts of the document which could not be found in the memory. However, this interface is accessible only to the Commission's Translation Service.

In addition to selecting the source and target language(s), users can choose between a variety of terminological options:

1. Domains

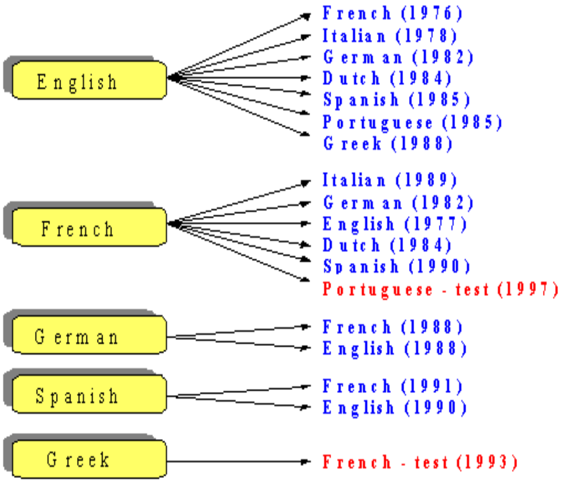
There are 36 subject fields (*agriculture*, *law*, *transport*, etc.), from which the user can select up to three. When activated, domain-related terminology will take precedence over any more general translation which might also be possible. For instance, the translation of the French noun *coeur* would be *heart* in the general dictionary, but *core* when the *energy* domain is activated.

2. User codes

User codes are not for general use. They are allocated to individual users who have asked for specific terminology entries in the system. These special requests are coded in customised dictionaries to ensure that they do not interfere with the main dictionaries. There are 17 user codes.

3. Text type

Users are asked to specify whether the document they intend to submit to MT is: the *minutes* of a meeting (in order to obtain a change of tense between e.g. English and French and thus respect the different language conventions); a *letter* (so that courtesy closes are correctly replaced); or an *instruction manual* (where the imperative mode will be used instead of the infinitive – French only).



Finally, users can ask for their documents to be treated as *confidential*: normally all documents sent to MT are automatically stored on the server and become part of the development corpus, but confidential texts are deleted after processing. In addition, Commission MT managers receive a *side-by-side* print-out of each document, which displays the source text and its translation in opposite columns. The *side-by-side* also provides information on how the source text was analysed and what expressions and lexical routines were accessed in order to produce the translation.

Improvement

Users of the Web interface have an additional option, which is to send feedback to the MT Help Desk. All comments concerning terminology or phrasing are then submitted to the development team for processing (the dictionaries can be updated every 24 hours if necessary).

For the introduction of new entries into the system, a corpus-based approach has been adopted. This means that the coding of dictionary entries is based on the frequency of occurrence of words and phrases in context. The most common meaning is coded as the default entry, with exceptions being covered by domains or contextual rules.

For example, the default translation for the English verb *to work* is *fonctionner*, and not *travailler*, because *to work* mostly appears with the former meaning in Community texts; the translation *travailler* has nevertheless been introduced in the form of a contextual rule which is triggered when the subject of *to work* is human.

The Commission's development corpus is composed of texts sent by users (unless they are confidential) and contains some 100 000 pages. This can be exploited in several ways – for example by means of *KWICs* (*Key Words in Context)*, a facility for illustrating the different contexts in which a word / expression appears. The lexicographic approach used is thus very pragmatic.

USE OF EC SYSTRAN

User groups

For statistical purposes, users are allocated to one of three groups: Commission (as funder of the service, the most crucial group), other EU institutions and bodies, and the public sector in the Member States. Within each group, there are two types of user: administrators and translators.

1. Administrators use MT for three main reasons:

a) For browsing texts written in a language they do not know. The quality of the translation may not be high, but the speed is remarkable: the computer can translate 2 000 pages per hour. Users can then decide if they wish to submit their texts (or part of them) for human translation, for rapid post-editing (see below), or whether the information provided in the raw translation is sufficient.

b) For the fast translation of urgently needed texts which often have a standardised structure and terminology (minutes of meetings, reports, etc.). A reasonably high translation quality can be obtained after correction by someone who has the target language as his/her mother tongue. The texts can then be distributed for internal use.

c) For drafting in a language other than their mother tongue or main language. Some officials prefer to write a text in their own language first, request a machine translation and then correct the output.

2. Translators, on the other hand, use MT almost exclusively as a basis for providing a polished translation. A series of practical experiments conducted with translators who regularly use MT have shown that *in the right circumstances* (language pair, text type, domain, style), savings of up to a third could be achieved in translation time.

MT and human translation have for too long been worlds apart. On the one side, MT was too often conceived and presented as a real alternative to human translation. On the other side, translators naturally resented this and tended to reject even the possibility of a machine achieving anything close to a translation.

In the European Commission, these two worlds came together from the very beginning. One reason why our MT system is regarded by the outside world as being amongst the most robust and dependable is the sustained commitment the Commission has shown since 1976. A good example of that commitment is the rapid development of French-Portuguese thanks to the feedback of a handful of enthusiastic Portuguese translators - a development which was itself inspired by the work of their Spanish colleagues, who have taken some giant MT strides over the last few years.

In the Translation Service, MT is evolving into a *productivity tool*, as translators embrace the technology in their daily work, often in combination with the extensive translation memories now available to all.

That said, many translators still have deep concerns about the impact the technology may have on the quality of translations. The more experienced practitioners of this new art of MT editing - or *post-editing* - thus have the additional task of proving such fears to be ill-founded.

And there may indeed also be a case for MT as a *quality tool*. Well trained as they are to deal with the intricacies of a text, translators can have trouble with tedious and repetitive work involving scores of figures, references, tables and formats, all of which can lead to endless revision work to the detriment of the time invested in the core quality - *linguistic* quality - of the translation. As texts of this kind are a regular feature of work in most large organisations, raw MT, coupled with translation memory whenever possible, can have a liberating effect on translators, allowing them to concentrate on the real problems of the text and letting the machine take care of the rest.

But in order to achieve this result, a translator needs to go beyond a passive approach to a technology that can be deceptively easy for the user: if MT output is available at the push of a button, MT *post-editing* requires skills that can only be acquired through practice, in real working conditions. It also requires of the translator a certain amount of perseverance, since the first translations based on MT output will probably take more time than translations starting with the traditional blank page.

The experience of translators working daily with MT suggests that the technology may even have some addictive properties. At least, there are some who, in their own words, "can't live without MT"! For one example (in Spanish) of a dedicated "MTer", see: http://europa.eu.int/comm/translation/bulletins/puntoycoma/48/pyc485.htm.

The use of raw MT output by professional translators was probably not foreseen by the pioneers of this technology, who were on the contrary intent on providing a cheap and fast alternative to human translation. It demands of translators new attitudes, new skills, new ways of personal and collective organisation, and could have a major impact on the future of the profession.

User guidelines

A seminar on MT forms part of the training programme for new Commission officials. The MT system, its potential and limits, is presented to them together with specific guidelines about how to achieve the best possible results. Here is an overview of the main areas where users have to pay particular attention:

1. Spelling errors are not recognised by EC SYSTRAN. Misspellings will not only remain untranslated, but will adversely affect the translation of the whole sentence – *The Commission vice president* will be translated as *Le président du vice de la Commission* just because the hyphen between *vice* and *president* is missing. Users are thus encouraged to use spelling checkers before submitting their document to machine translation.

2. Syntax should be uncomplicated and clear. Elliptical sentences should be avoided as they are prone to multiple interpretations. For instance, the sentence *Bush blocked Parliament proposal* will be translated as *Le buisson a bloqué la proposition du Parlement*, whereas the same sentence phrased slightly differently (*Mr Bush*, or *President Bush*, instead of only *Bush*) will be correctly translated.

3. Finally, it is important to stick to a simple format. For example, avoid inserting hard line breaks in the middle of a sentence, because the system will identify any text after a break as a new sentence, and the whole analysis will collapse.

Quality of the raw output

Users are also warned that the quality of the raw machine translation varies considerably and depends on four different factors:

1. Type of source text

EC SYSTRAN has been developed for the past 25 years to meet specific Commission needs. Therefore, internal documents with EC jargon, run-of-the-mill reports or minutes of meetings will be much better translated than a creative article or a piece of literature.

2. Quality of source text

In most cases, Commission officials have to work in a language which is not their own. As a result, original documents are often of poorer linguistic quality. Whereas human translators can circumvent this problem by using their imagination or by contacting the author, the machine has no such fall-back. The raw output of a poor original will be poor too.

3. Time spent on development

In general, the more standardised a text is, the higher the quality of translation will be. The more time spent on the development of the system, the lower the number of errors will ultimately be. For instance, the combination of English and French gives one of the most satisfactory results because the Commission has invested on these pairs for many years. On the other hand, language pairs with German and Dutch need substantial development to reach the same level.

4. Affinity of languages combined

Language combinations of the same family will result in a better translation quality in a shorter time. French-Spanish, one of the relatively recent language pairs, can easily compete with the quality of English-French. Consequently, it is one of the most popular amongst translators. Adaptations have been made on the basis of feedback from Spanish translators who use the system regularly in order to produce a final translation. For the same reasons, French-Portuguese gave very encouraging results after only six months of development.

Post-editing Service

The Translation Service offers an external Rapid Post-editing Service for requesters who need to translate internal documents with very tight deadlines. In such a case, officials can send their texts to the MT Help Desk in Brussels, which will in turn translate them with EC SYSTRAN and pass the results on to freelance translators for correction. Emphasis is on speed and accuracy rather than style or in-house jargon.

In the case of documents intended for external distribution, however, Commission administrators should always ask the Translation Service for a fully polished "human" product.

User statistics for 2000

The Commission accounted for 77% of the total pages machine-translated in 2000, of which almost half was requested by translators. The remaining 23% were shared evenly between other EU institutions and public-sector bodies.

The MT statistics over the last 10 years (and especially from the mid-90s onwards) reveal that the number of users is steadily increasing, with a fivefold increase in demand since the system became generally available by e-mail.

MT Requests rose across the board last year, the total climbing by over 23% from 78 894 in 1999 to 97 199 in 2000. The increase for the Commission was 20% and for the Translation Service 12%.

Demand in terms of pages amounted to 546 248. Here too, advances were made in most areas: demand amongst Commission administrators rose by 7%, Parliament registered a growth of almost 100% compared to 1999, and public-sector usage increased by more than 140%. Demand also rose for the other EU institutions and bodies.

The main exception was the Translation Service itself, where in spite of more requests, the MT page count - and therefore average document length - fell; perhaps the generalisation and expansion of translation memories within the Service plays a role here.

Tentative figures for the first 5 months of 2001, however, suggest that both the Service's requests *and* pages are again increasing, and that **overall** demand for the system could reach a new high.

No matter how the statistics are interpreted, the clearest trend is the steady growth in Commission MT usage since the beginning of the decade, when the system was first made generally available. This trend is paralleled by a growing demand from other institutions and external users, pointing to an increasing awareness of the Commission's MT system amongst other EU bodies and the Member States, an awareness which is in turn reflected by the growing public use of MT technology on the Web.

IMMEDIATE FUTURE

Maintenance of the system

The wide and growing use of MT within the European Commission can be seen as reflecting the enormous *potential demand* for the technology in a professional environment; but should not be taken as a measure of the *satisfaction* of the Commission's officials with the linguistic output of the system. Much remains to be done, and enhancement of the 18 existing language pairs is being pursued. All feedback received by users will be incorporated as quickly as possible.

Migration project

For 25 years, EC SYSTRAN has operated with the old IBM Assembler computer language. In 1997, however, the Commission's Data Centre announced that the mainframe computer which supported Assembler would be phased out within 5 years.

This left the Translation Service the choice of finding a modern emulator for the Amdahl or rewriting EC SYSTRAN's programs in a more recent computer language; one feasibility study later, the Service decided to convert, or *migrate,* EC SYSTRAN's basic programs to *C.*

A two-year migration project was launched in late 1998, but in spite of the knowledge gained from the commercial version (which had been migrated several years before), the work proved complex and time-consuming. There are still a few months of running-in ahead before the new system can fully enter production mode, but thereafter users should begin to enjoy the benefits of a modern computer platform.

New language pairs

Since 1999, the Translation Service has been participating in the development of new language pairs under the MLIS (Multilingual Information Society) programme managed by the Directorate-General for the Information Society.

Three projects are under way for the development or improvement of language pairs involving combinations of English and French with Greek, Portuguese and Dutch. Partners in the project include the Greek, Portuguese and Flemish/Dutch governments as well as SYSTRAN Luxembourg S.A.

By the end of these projects, six of the existing language pairs will be substantially improved: English-Dutch/Greek/Portuguese, French-Dutch/Portuguese and Greek-French. Moreover, six new language pairs will be created from scratch: Portuguese-English/French, Dutch-English/French, Greek-English and French-Greek; these should at least be of sufficient quality for browsing purposes.

Efforts are also being made to introduce MT for the Nordic and Eastern European languages. Several proposals have been made concerning collaboration on Danish, but as yet nothing has come to fruition. As for Swedish, the government is conducting a review of the MT industry before considering any cooperation. On the other hand, MLIS *is* co-financing Finnish-English/English-Finnish developments involving the translation and parsing technology of Kielikone and Conexor respectively. Furthermore, looking ahead to the enlargement of the EU, projects have started for English into Hungarian/Polish and Polish/Hungarian into French (under the HLT, *or Human Language Technologies* programme).

Other projects

The Commission will be considering plans for the (semi-)automatic coding of new entries in MT dictionaries on the basis of corpora and glossaries. It also hopes to allow MT users to enter their own terms in a private dictionary by means of a personal coding interface. The trick will be to ensure that those terms take precedence over the translation provided by the main MT dictionaries. The possibility of translating EC Web pages dynamically also features among future projects.

At the time of writing, a call for tenders for standard maintenance of the existing system is in preparation, the current maintenance contract being due to end in December 2001. As a complement, the Commission is considering a call for an expression of interest for teleservices. These would mainly concern new language pairs but might also involve domains which EC SYSTRAN current language pairs do not cover so well.

Finally, as part of the IDA programme (Interchange of Data between Administrations), the Translation Service will be conducting a feasibility study on potential MT needs in the Member States. The study will consist of three parts:

(1) a survey concerning the principal needs of European public administrations in the field of MT;
(2) a definition of the infrastructure necessary: a) to coordinate access to the Commission's MT system and b) to carry out the technical, linguistic and terminological developments requested; questions include means of access (Web, batch, e-mail, multiple sites, etc.), confidentiality, and efficient integration of linguistic and terminological resources;
(3) assessment of the financial and human resources needed to complete the work.

CONCLUSION

After twenty-five years of development, machine translation has now become a helpful option for some of the everyday translation needs in the Commission's administrative departments. It can also be used by translators as an effective support tool, although the picture varies according to the language pair (depending on language affinity, length of development, and amount of feedback received).

When user guidelines are correctly observed, MT, despite its inherent limits, can be of substantial help to the language-frustrated official. Of course, machine translation is just a tool among others – it is not aimed at replacing human translators, nor can it be a solution to all translation needs. But in the complex multilingual environment that is the Commission, it *can* rescue translators from some dull work and facilitate communication between time-challenged administrators.