The current state of the Commission's SYSTRAN MT system

Angeliki Petrits

European Commission SYSTRAN development team, DG XIII, Luxembourg

Introduction

SYSTRAN is an international MT system that has been bought, developed and is used by the European Commission for internal purposes. In its current state, the Commission's system contains 16 language pairs, where English, French and recently German and Spanish play key roles.

These pairs provide translation from English into French, Italian, German, Dutch, Spanish, Portuguese and Greek; from French into English, German, Dutch, Italian and Spanish; from German into English and French; and from Spanish into English and French. The development of Greek into French was started in November 1993, and a pilot system will be operational in 1997.

The SYSTRAN system presented at the Cranfield Conference in 1984 by Ian Pigott and Peter Wheeler was something less mature than the system you will find today. At that time, specific linguistic problems related to source language analysis and target language generation required tackling. Their correct handling was fundamental to the overall development of the system and was long a preoccupation of the project team.

Many of these problems have now been partially or totally resolved, which has led to a substantial improvement in translation quality. Main changes in these ten years also include an increase in the language pairs under development and in the amount of pages translated by the system as well as the availability of SYSTRAN to all EC officials, made possible thanks to the computerisation of the Commission departments.

It had been said ten years ago that with SYSTRAN we adopted a **pragmatic** approach to development. As a result, interest has also concentrated on analysing and satisfying user requirements. This opened new perspectives for the enhancement of the SYSTRAN MT system by exploiting existing linguistic resources, such as the CELEX multilingual legal data base and the EURODICAUTOM electronic dictionary. During these ten years, SYSTRAN has shifted from a pure development mode towards a user-oriented policy.

An indication that SYSTRAN is internationally recognised with world-wide applications is that the Commission is not its only user. Among others one should mention the other EU Institutions (the European Parliament, the Council, the Court of Auditors, etc.), some national authorities (like the Deutsche Bahnen and the Kernforschung Centrum Karlsruhe in Germany, Aérospatiale, the CNRS and the CEA in France and the Universities of Pisa and Florence as well as the CSATA Technopole in Bari, Italy) and some international organisations, such as NATO in Brussels. The US Air Force, the Xerox Corp, General Motors, Bull and France Telecom also have a SYSTRAN licence.

1 Technical characteristics

Before tracing the chronology of developments within the Commission's system, it is worth recalling briefly the technical characteristics which are partly responsible for the success that SYSTRAN enjoys today.

The system functions thanks to:

- **basic programs** written in **assembler** and **linguistic program**s written in **SPL** (SYSTRAN Programming Language);

- **dictionaries** which**,** according to the philosophy of the system, play a very important role because they contain much morphological, semantic and syntactic information for each entry.

There are two types of dictionaries:

- a **basic one-word dictionary (STEM),** the entries of which are words and idioms coded independently of their linguistic context, and which are the same for all target languages combined with the same source; it is a *multi-target* dictionary;

- **a contextual or expressions dictionary (IDLS)** (Idiom/Limited Semantics) containing expressions and lexicographic and syntactic rules which determine the translation of a single word or expression (noun phrase, verb phrase etc.) according to the context; this concerns a given target language.

To illustrate better the roles of the two dictionaries, let us consider the following example:

Delors ***package*** has been translated by SYSTRAN as ***emballage*** *Delors* instead of ***paquet*** *Delors*.

In order to correct this, one has to write a rule (in the IDLS dictionary) stating that when the word *package* is modified by the word *Delors*, the translation *paquet* instead of *emballage* (the translation contained in the STEM dictionary) should be provided.

Both dictionaries and programs are updated every three months, and the new version of the system is then installed on the Commission's mainframe.

Apart from the above basic dictionaries, there are also specialised dictionaries, called Topical Glossaries (TGs). These are dictionaries containing terms from a topical field (agriculture, finance, minutes etc.) that the system will consult at the user's request. Usually, TGs are used in the case of a conflict between the general translation of a term and its translation in a specialised field.

At the Commission, we have adopted a **corpus-based approach**; this means that the coding of dictionary entries is based on the frequencies of occurrence of words / phrases in context. The most common meaning is coded as the default entry in the STEM dictionary. The exceptions are covered by the IDLS (in the form of contextual rules) and by TGs.

e.g. The translation given in the STEM dictionary for the English verb *to work* is *fonctionner* and not *travailler* because *work* appears in this sense in most EU texts. The translation *travailler* has been introduced in the IDLS dictionary with the code HUSUB (human subject).

The dictionaries are not reversible, owing to the fact that the source language dictionary contains far more information than the target ones. It has nonetheless proved possible to use the target dictionaries as a bridge for the rapid development of new bilingual dictionaries.

A new element is the enrichment of the SYSTRAN dictionaries with EURODICAUTOM entries. EURODICAUTOM, the Commission's terminological data base covering 9 languages, is being integrated into SYSTRAN (See *figure 10 - 1* for the numbers of entries in the SYSTRAN dictionaries after importing the EURODICAUTOM data).

Tests are being carried out to measure the amount of progress obtained. However, it is predicted that better translation quality will be achieved in technical texts, whereas in general ones it is likely that SYSTRAN will provide better translation, the reason being that EURODICAUTOM is highly specialised.

Systran Dictionaries before and after importing Eurodicautom data

language pair	original entries	new entries
DE - EN	145.000	249.000
DE - FR	65.000	262.000
EN - DE	81.000	350.000
EN - EL	46.000	220.000
EN - ES	59.000	259.000
EN - FR	130.000	440.000
EN - IT	110.000	275.000
EN - NL	47.000	294.000
EN - PT	43.000	227.000
ES - EN	32.000	258.000
ES - FR	27.000	262.000
FR - DE	94.000	293.000
FR - EN	147.000	430.000
FR - ES	45.000	265.000
FR - IT	40.000	286.000
FR - NL	43.000	353.000
Total	**1.154.000**	**4.723.000**

Figure 10 - 1

2 History of development

The SYSTRAN MT system was invented in the USA in the late fifties by the Hungarian Peter Toma. After having moved to California in 1956, Toma put his convictions into practice for the development of an operational MT system. Unlike most of his contemporary colleagues, he did not believe that linguistics could provide a satisfactory solution to the formalisation of language, but was convinced that the analysis of language had to fit into the existing computer philosophy.

The first operational language pair developed in 1970 was Russian-English for military reasons. In 1973, NASA financed the development of English-Russian for the Apollo-Soyuz project, while in 1974 Toma's group applied the results of the English analysis to an English-French prototype.

The Commission bought SYSTRAN in 1976 from WTC (World Translation Center), Toma's company in La Jolla, California. Development started with the English - French language pair and was extended to the other 15 pairs available today.

The development of SYSTRAN and the introduction of new language pairs did not take place overnight. The Translation Service of the European Commission is the largest in the world, with some 1200 translators and terminologists translating about a million pages per year.

The increase in the number of texts and languages, and in the technicality of the subjects handled, has led the Translation Service to seek ever more advanced technical solutions to enable it to fulfil its mission.

SYSTRAN was chosen by the European Commission because at the time it was the only operational fully automatic system available for English-French, and it was with this language pair that SYSTRAN took its first faltering steps there in 1976, in the shape of a pilot project. On the basis of what were considered encouraging results, a French-English version of the system was created the following year, and in 1980 an English-Italian version was introduced.

YEAR	LANGUAGE PAIRS	STEM	EXPR
1976	English-French	69.000	61.000
1977	French-English	57.000	90.000
1978	English-Italian	59.000	51.000
1982	English-German	54.000	27.000
1982	French-German	45.000	49.000
1984	English-Dutch	38.000	9.000
1984	French-Dutch	25.000	18.000
1985	English-Spanish	46.000	13.000
1985	English-Portuguese	37.000	6.000
1988	English-Greek	39.000	7.000
1988	German-English	122.000	23.000
1988	German-French	54.000	11.000
1989	French-Italian	29.000	11.000
1990	French-Spanish	33.000	12.000
1990	Spanish-English	27.000	5.000
1991	Spanish-French	26.000	1.000
1993	Greek-French	n.a.	n.a.
	TOTAL	760.000	394.000

Note: n.a.: not available

Figure 10 - 2

The dictionaries were gradually extended to cover the main areas of European Commission activity. Moreover, with the 1980s came the arrival of new language pairs. There are now a total of 17 pairs under development, of which 16 (the exception being Greek-French) are available to staff throughout the institution.

In *figure 10 - 2* you can see the year in which development started for each language pair and the current number of entries in both dictionaries, the one-word dictionary and the expressions dictionary

One might wonder why the Commission has chosen to develop these particular 17 language pairs instead of some others from the 72 we would have by combining the 9 official EU languages.

The criteria for developing one language pair instead of another have been based on three issues:

a. **On the internal needs** of the Commission. Since English and French are the main working languages, it was natural for us to start with these two languages as both source and target and to add later Italian and German.

b. **On the translation quality** expected from related languages as well as the availability of trained linguists. As a result, there has been a strong preference for developing language pairs involving combinations of either Romance (French-Italian) or Germanic languages (English-German), because we could predict that with a minor effort satisfactory results could be obtained, which has proved to be true.

c. **On the budgetary restrictions** imposed on us. Financially, it was impossible to develop 72 language pairs, which would have been ideal. So, when a member state was willing to co-finance the development of a particular language pair, the Commission would give priority to that development. This was the case with Greece in 1988 for the development of English-Greek and recently of Greek-French. The product acquired after this development can be used by the Commission and the Greek authorities with no limitation in time. As a result, a SYSTRAN office was created in Athens last September in order to provide translation services to the Greek public sector.

The future policy of SYSTRAN is based on an assessment of the system carried out in 1991 at the European Commission by a panel of external experts. The main conclusions of the evaluation group were that:

- SYSTRAN remains the machine translation system best suited to the needs of European Commission users, owing to its large dictionaries and the number of language pairs on offer;

- SYSTRAN should be re-engineered and incorporated within an integrated document handling environment; however, in the medium term, the service should be enhanced by the addition of new language pairs;

- a post-editing service should be set up and the European Commission should continue to support natural language research in the field of Machine Translation.

3 How SYSTRAN is currently being used and enhanced

Since 1990, all 16 language pairs have been available via electronic mail to all Commission officials who are equipped with either a terminal or a PC connected to the network.

SYSTRAN is easy to use. Users simply send their documents via electronic mail to a special mailbox (M152) stating their requirements in language combinations (see *figure 10 - 3*).

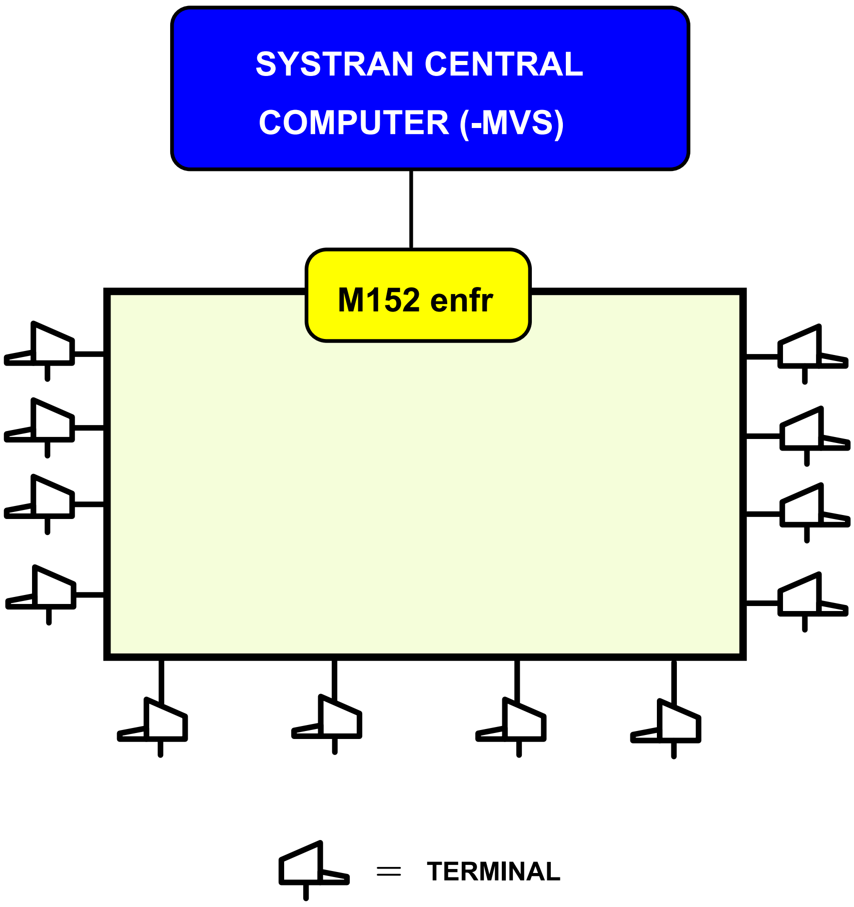


Figure *10* -3

e.g. **M152 enfr** (*English/Français*) for a translation from English into French, **ende** (*English/Deutsch*) for a translation from English into German, etc.

Normally, the translation is returned in the same way within half an hour, depending on the number of requests submitted. Sometimes end-users receive their translation in only a few minutes.

By the end of the year, MT will be only part of a new interface called EURAMIS providing different tools for translators. This new, user-friendly interface will be available on all PCs connected to the network and will be supported by Windows.

At first, it will offer the user four possibilities:

- a SYSTRAN raw machine translation (with the integrated EURODICAUTOM data base)

- a link to the EU nine-language legal data base, CELEX, whereby references to legal texts contained in documents for translation are extracted automatically in the target language (a service already available by default together with SYSTRAN translations);

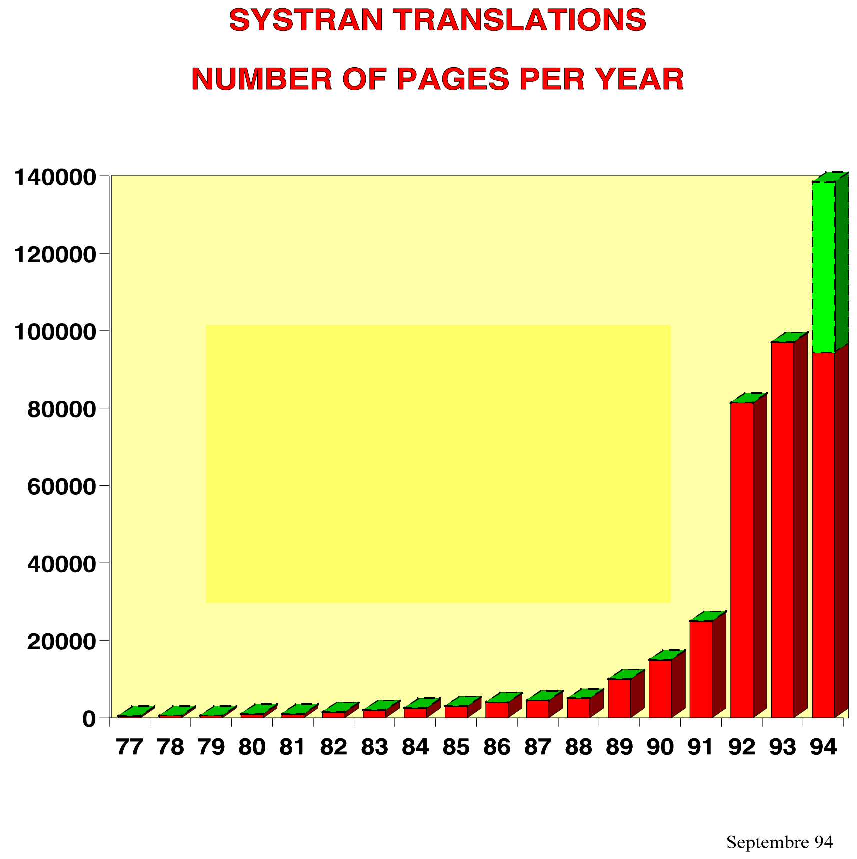
- translation by EURODICAUTOM of a list of terms (72 language pairs);

- translation by EURODICAUTOM (through SYSTRAN) of the terms contained in a text (32 language pairs). SYSTRAN analyses the text and extracts the terms which are subsequently submitted to translation by EURODICAUTOM.

It is foreseen that this interface will be gradually equipped with further tools, e.g. a translation memory, available not only to the Translation Service but to all EC officials.

The availability of SYSTRAN through the Commission's network has been widely advertised in recent years. Posters have been put up in different Commission buildings; special brochures have been distributed to all officials; current and potential users have received a leaflet with guidelines for drafting documents suitable for MT; help-desks in Brussels and Luxembourg have been set up in order to assist SYSTRAN users. Our promotion team in Brussels is encouraging users to provide feedback of their translations to the development team. It has also set up a post-editing service on an experimental basis to help end-users in a hurry to improve the raw output SYSTRAN provides.

The computerisation of the Commission services together with the SYSTRAN publicity campaign has led to a considerable increase in the pages translated by the system in recent years (see *figure 10 - 4*). It is clear that in the last three years the number of pages has climbed considerably. It now ranges from 10,000 to 15,000 pages monthly, whereas not that long ago it stood at only 1,500 pages. It is expected that by the end of 1994 the figure will be around 140,000 pages per year.



SYSTRAN TRANSLATIONS

NUMBER OF PAGES PER YEAR

Septembre 94

The lighter shade represents the projected increase at the end of 1994.

Figure 10 - 4

The language pairs most requested (see *figure 10 - 5*) are by far English-French and French-English, as these two are the hard-core part of the system and can provide translation of an acceptable quality.

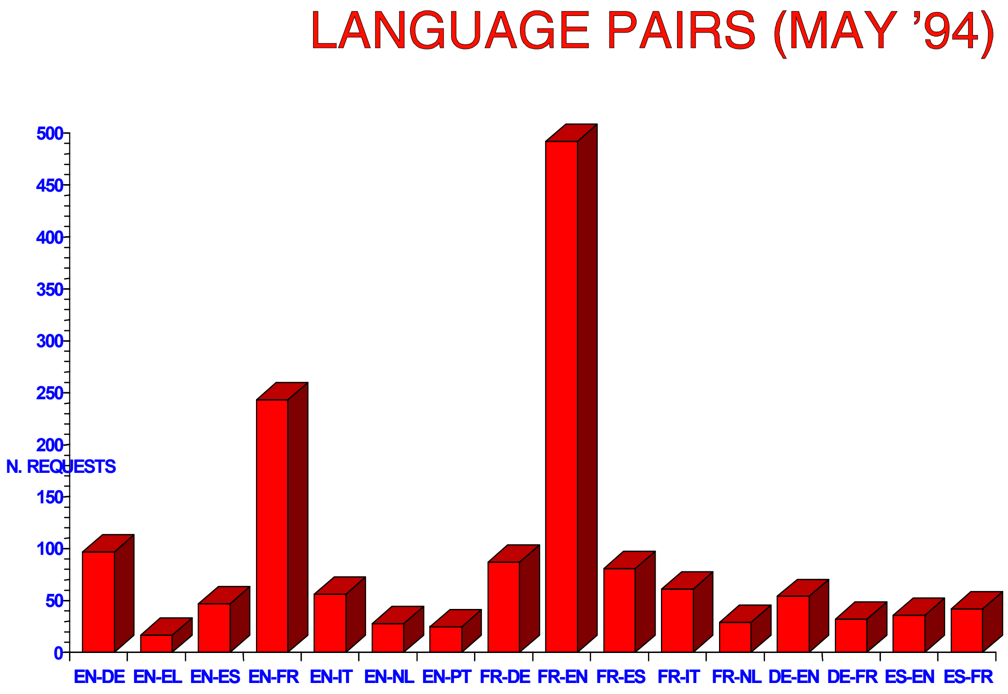


Figure 10 - 5

But, where does the SYSTRAN development team intervene in order to enhance the system and correct the errors? Let me say a few words about our way of working .

Everyone in the SYSTRAN development team is responsible for one or more language pairs. We receive automatically every text submitted to MT in the Commission for the language combination we are in charge of, unless the text has been characterised as *confidential* by the user. This means that the texts submitted to MT and their raw output are transmitted at the same time to the requester and linguist of the development team.

All texts received are then classified according to the user, his department, the type of document, the subject field and the number of pages. This information is entered in a data base and can be extracted at any time for tests or statistical purposes.

Among the development team's tasks is:

- the review of SYSTRAN translations in order to identify error types requiring attention, and their correction;

- the supervision and the correction of dictionary updates;

- contacts with in-house and external users with a view to establishing development priorities;

- participation in promotion strategies.

Sometimes the user provides feedback by sending the post-edited version of the translation to the SYSTRAN team. In that case, priority is given to the correction of these errors and the introduction of user-specific terminology.

5 Applications of the system

SYSTRAN is used inside the Commission for three purposes:

- For the **fast translation** of short repetitive texts with standardised structure and terminology (mail, minutes of meetings, parliamentary questions, reports etc.), where a high translation quality can be obtained; after fast post-editing these texts can be distributed for internal use; SYSTRAN is usually less suitable for long texts or for documents which are binding on the institutions.

- For **browsing** of texts written in a language the user does not know; here the quality of the translation may not be high, but the speed is remarkable: SYSTRAN translates 2, 000 pages per hour; users can then decide if they wish to submit their texts or part of them to "human" translation, or if the information provided in the raw translation satisfies their needs.

- For **drafting** purposes; users write a text in their mother tongue and request a SYSTRAN translation; this enables them to have a document drafted in something other than their native or main language.

The quality of each language pair varies considerably, depending both on the time spent on the development of each pair and the syntactic and lexical affinity of the languages concerned. The most satisfactory results are offered by pairs involving English, French, Spanish and Italian, while efforts are being concentrated on improving translation both into and from German.

The more standardised a text is, the higher the quality of the translation will be (I mainly mean EU texts). The more time spent on the further development of the system, the lower the number of errors will ultimately be.

SYSTRAN does not aim at replacing the human translator. It cannot compete with him, since the computer does not have the experience and the knowledge of the world that only humans can acquire. It is simply a complementary, surprisingly fast tool, that can save translators or end-users some dull work. 80% of SYSTRAN customers are not translators. They seek an urgent solution to a translation need that cannot be satisfied by the Translation Service.

Based on the users' needs, the Translation Service in co-operation with DG XIII has established a number of priorities for future development. These priorities include the improvement of German, both as source and target, since it is a working language which is largely unknown outside Germany, and development of language pairs from non-vehicular source languages (Danish, Portuguese, Dutch and Greek) into the working languages they each more resemble (English or French) for browsing purposes. It is a field where tolerance of machine translation quality is the highest.

6 Conclusion

After eighteen years of marginalised but continuous development, SYSTRAN has now become an effective answer to some of the pragmatic translation needs in the operational departments of the Commission.

In its early days it was not welcomed with open arms by the translators, as it was seen as an *unfair* competitor. However, the computerisation of the Commission services together with a general change of attitude, has radically altered the image of SYSTRAN. Now that it is available through the network to all Commission staff, it is no longer considered a substitute for human translation but as a product which complements the service translators provide.

Its further development is part of a Commission policy which has to cope with an ever-growing workload and with the duty to preserve multilingualism in Europe.

References

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

Paesmans, H. (1994), *The Translator's tools*, European Commission, Translation Service, internal document.

Pigott, I. (1992), *SYSTRAN development at the EC Commission, 1976 to 1992,* European Commission, internal document.

Senez, D. (1994), *Developments in SYSTRAN,* ASLIB Proceedings.

Toma, P. (1986), *SYSTRAN's contribution to mankind*, in **Terminologie et Traduction** 1, ISSN 0256 7873.