[From: IFTT’89, April 26-28, Oiso Prince Hotel, Japan]

SYSTRAN AT THE COMMISSION OF THE EUROPEAN COMMUNITIES

Ian M. Pigott
Commission of the European Communities

General

Since 1976, the Commission has been developing Systran, originally created in the United States, for several European language pairs. Currently a team of 35 officials and contractual staff are working on the development of 12 versions: six from English (into French, Italian, German, Dutch, Spanish and Portuguese), four from French (into English, German, Dutch and Italian) and two from German (into English and French).

We have found that the Latin languages and English generally cause fewer problems in MT development than the Germanic languages. As a result, the quality of English-French or French-Italian translations is considerably better than that of the German-French or French-Dutch versions. Nevertheless, we are making very encouraging progress on all the language combinations.

Linguistic development

The Commission's versions of Systran have become ever more modular over the years. Source language analysis is now completely independent of target generation, even as far as the basic dictionary is concerned. This means that the results of English analysis, for example, are now applied to translations into six different target languages. At the target level, too, the same synthesis programs are used with minor variations in conjunction with different source languages.

The only truly bilingual part of the system is at the transfer stage where powerful contextual dictionary coding is applied to identify the correct translation of words or structures in context. However, even here, *is* has been found that contextual rules suitable for one language from a given group (e.g. a Latin language) can often be used with little or no modification for translation into another language of the same group.

Owing to Systran's highly developed syntactic and semantic features, the classical problems of grammatical homography and polysemy have been largely overcome for the well-developed language pairs. Quality improvement is thus based increasingly on fine-tuning transfer into the target language by the incorporation of the terminology required for a particular subject field and/or document type.

Currently, Systran's one-word source dictionaries for English and French contain about 80,000 entries, thus covering most words encountered in written text. However, the contextual dictionaries, which currently have between 20,000 and 80,000 entries per language pair, are being constantly expanded. We estimate that contextual dictionaries of about 200,000 entries will ultimately be developed for the major language pairs.

Technical infrastructure

At the Commission, Systran runs on an Amdahl mainframe (IBM compatible) under the MVS operating system. Translations are processed at a rate of about 500,000 words per hour. Access via the internal network is routed through a Unix server which schedules translation requests from various departments of the Commission in Brussels and Luxembourg. User-friendly menu systems have been installed and are now available on over 400 user workstations.

Most users now have local equipment (generally Unix or MS-DOS) which they use for document drafting as well as any post-editing. We are making increased use of optical character reading for input, particularly when documents are not available in machine readable form.

Post-editing

In certain cases, the user is satisfied with a raw translation without revision, especially if he has a document translated for his own information, but in most cases post-editing is required.

Commission translators have tackled this problem in several ways: some make use of machine translations as a basis for preparing a final version by traditional methods, others carry out post-editing either on paper or directly on screen. Each method has its advantages and its disadvantages and depends, inter alia, on the individual's experience.

In order to evaluate Systran's potential for Commission requirements, a group of translators has recently been set up to examine the results obtained with various types of documents, bearing in mind the quality required by the users, the time limits imposed and the purpose of the translation. In parallel, we intend to encourage generalized access to Systran for raw translations which can be particularly useful when traditional translation services cannot not keep to tight deadlines.

Users

The Commission's main aim in developing Systran is to promote internal use but the user rights which we have acquired cover public sector bodies in the Member States of the European Community.

The most important external users are currently NATO in Brussels, the Nuclear Research Centre in Karlsruhe (FRG) and the German railways. Over the years, these organizations have contributed much to extending the Systran dictionaries for a number of specialized subject fields as well as to general improvement of the quality of the translations.

The Gachot group which owns the rights for the private sector has an agreement with the Commission on coordinated development covering all the Community languages. Gachot S.A. of Soisy-sous-Montmorency near Paris is thus in a position to give access to Systran to all interested parties, either via the French Minitel terminals, or on networks better suited to the requirements of machine translation.

Finally, the Commission is preparing a major project on the translation of patents into English, French, German and Spanish. We soon hope to conclude an agreement with the European Patent Office as a basis for integrating terminology relating to new technologies.

The future

Over the course of the next twelve months, the Commission will probably begin the development of a fourth source language, Spanish, initially for translations into English. We may also incorporate Greek as a target language from English at the request of the Greek authorities.

At the more general level, we are currently investigating the development of an additional level of semantic coding as a means of enhancing analysis and of providing additional information for contextual coding. For German as a source language, we intend to develop powerful features for dividing compound nouns into their component parts which can then be parsed on their own syntactic and semantic attributes.

Now that our in-house technical infrastructure has become more stable, we shall also be looking at the possibility of providing rich document formatting at the target level. Until now, we have had no option but to use revisable print-image output - based solely on blanks and carriage returns - as a wide variety of peripheral terminals (word processors, PCs, etc.) have been connected into Systran without proper interfacing.

Conclusions

The quality of Systran is already relatively good for French-English, French-Italian, English-French, English-Italian and English-Spanish. Between now and 1990, we hope to reach similar levels for the other language pairs under development.

The Commission's top priority is development for internal needs but it is in a position to sign contracts with public bodies within the European Community.

Quality improvement in the future will be based first and foremost on the requirements of users in various subject areas in collaboration with the group of in-house translators who have been assigned to the project. Such improvement should be facilitated by the incorporation of more powerful semantic features in the Systran software itself.