[*Machine Translation Summit*, September 17-19, 1987, Hakone Prince Hotel, Japan]

SYSTRAN: A MACHINE TRANSLATION SYSTEM TO MEET USER NEEDS

Systran Corporation

Joann P. Ryan

It is fitting that we should open this session on commercial machine translation systems with a presentation on Systran, a pioneer among MT systems, whose users and developers are best characterized by their shared persistent belief in the proposition that machine translation is feasible -- and not only feasible, but practical and useful. It is very encouraging to look around me today and see how many of you have indicated by your presence at this conference that you share this belief. Because Systran developers have always had as a goal the production of useful translations -- and because Systran has been a production system since 1969 -- I would like to focus my talk on the five qualities that make a system useful: speed, reasonable cost, accuracy, accessibility, and flexibility.

As a brief introduction to Systran, I would like to quote from Prof. Juan Sager's closing remarks at the World Systran Conference held in February, 1986, in which he stated that Systran "is used more widely and by a greater range of users for a larger diversity of purposes than any other system currently in use." This statement alone bears witness to the fact that Systran has already met the above 5 criteria for a wide variety of users, ranging from large organizations, like the U.S. Air Force, NASA, the Commission of the European Communities (CEC), Xerox Corporation and General Motors of Canada, who sponsored so much of the early system development, to the many new users who today access Systran through on-line terminals and service bureaus.

Systran was developed by LATSEC, Inc. and World Translation Center, Inc. in La Jolla, California; today there are additional development groups in Luxembourg, Paris and Tokyo. Systran now offers 15 operational language pairs. These include English into 8 languages: French, German, Italian, Spanish, Portuguese, Russian, Japanese and Dutch; French into English, German and Dutch; and Russian, German, Japanese and Spanish into English. English-Arabic is under development, while pilot systems exist for 6 other language pairs: German into French, Spanish and Italian; and Chinese, Portuguese and Italian into English. The subject fields covered are already too numerous to list here, while document types range from abstracts, technical reports, journal articles, and service manuals to minutes of meetings and newspaper articles, with the range of linguistic styles increasing dramatically as many new users gain access to Systran.

Speed: Translations 4 to Many Times Faster Depending on End Use

In speaking about the features that make Systran the most widely-used MT system in the world today, I could begin by quoting such statistics as the fact that Systran, because its basic software is written in IBM 360/370 assembly language, can translate up to 2 million words per CPU hour on the FACOM M-380, which is used in Japan for translations between English and Japanese. Clearly 2 million words per hour is very impressive, especially by comparison with human output of approximately 2,000 words per day or even with the second fastest Japanese system, which translates at 60,000 words per hour on the same machine. But, although impressive, such statistics are largely irrelevant, because what makes any system of translation useful is how fast it can get a finished document into the hands of an end user in the form in which the end user wants it.

Therefore, we might find it more interesting to examine the statement of a Systran user at the European Commission, who says that a final copy of a post-

edited Systran translation of 25 pages can be completed to good standards of accuracy and style in a day and a half, thus reducing the total time for the document to go through the necessary translation channels from 28 days to 3 days. His estimate that net translation rates are approximately 4 times faster than that of conventional human translation is comparable to the Xerox Corporation's estimate that they are able to produce a finished service manual in any of 5 target languages approximately 4 to 5 times faster using Systran. The advantages of this relatively rapid turnaround to administrators who need a translated document as input to their decision making and to companies whose competitive edge depends on their ability to rapidly produce fully-translated technical documentation for a product in all targeted markets are obvious, especially when one considers that the documentation for a new airplane may run to 30,000 pages.

These figures are based on a method of "rapid post-editing" developed by the CEC and on full post-editing of a document written in Xerox's controlled input language. But when we see that users are beginning to recognize the value of raw MT for information purposes, Systran's potential translation speed of 500 words per second becomes increasingly relevant. Since 1986, the U.S. Air Force has been offering users of its translation service on-line access to raw Systran translations from Russian, French and German into English in a variety of scientific and technical fields. (Up until that time users received raw or minimally post-edited text but had no direct access to Systran.) Since 1980, the Nuclear Research Center of Karlsruhe (KfK) has been providing scientists with unedited Systran translations of research papers on fast breeder technology. Gachot S.A., a service bureau in France, offers the possibility of direct access to Systran on an experimental basis to the 4.5 million users of the Minitel network, with a response time of under one minute. Beginning next year, Systran Corporation of Tokyo plans to offer database users on-line access to English-Japanese translations from one of the largest international databases, thus for the first time offering Japanese users rapid access to the information explosion in English scientific and technical literature.

The increasing acceptability of raw MT output, the rapidly growing use of optical character recognition (OCR) equipment for input, the ongoing development of Systran linguistic programs and dictionaries, and the increased experience of post-editors in dealing with MT output are all factors that will ensure that the true speed of Systran MT, the rate of production of a useful document, will continue to increase along with machine speeds.

Cost: Systran Now Affordable by Small Users

Systran MT is no longer prohibitively expensive. Originally only a few large organizations could afford to support the cost of development of both linguistic and basic software programs, let alone the large dictionaries necessary to produce good-quality translations. Today, thanks to the support of these early users, many of the Systran dictionaries are already highly developed in a variety of subject fields, so that translations in these fields produce relatively few not-found-words (NFW's). The Air Force's Russian-English dictionaries contain over 200,000 single words and over 200,000 expressions in 20 different subject fields, so that true NFW's are quite rare. 4 of the language pairs used by the CEC currently have dictionaries of between 100,000 and 200,000 words. The KfK, which uses the CEC French-English dictionary to produce raw translations in the field of nuclear technology, reports approximately 1 NFW per page.

A virtual explosion in the size of the Japanese dictionaries from the current size of 50,000 basic entries to over a half million words and expressions is planned over the next two years. Systran Corporation of Tokyo is in the process of incorporating a large (250,000-word) database of scientific and technical terminology into its Japanese dictionaries. Meanwhile, a group of medical doctors in Japan is compiling a data base of 250,000 medical terms in order to standardize current medical terminology, and Systran Corporation will also incorporate all 250,000 of these terms into its Japanese dictionaries in order to offer rapid translation of medical documents to users in Japan and abroad.

The availability of such large dictionaries in a variety of subject fields

means that Systran translations are now affordable by the smaller user. Thus the CEC has taken the step of making Commission Systran available to 4 translation service bureaus in Luxembourg, France, Belgium and Italy. Such service bureaus make it possible for smaller users to obtain fully post-edited documents at a cost comparable to that of human translation, as well as the previously unavailable options of raw or partially edited translations at an even lower cost.

Accuracy: Sophisticated Tools for Semanto-Syntactic Analysis

The question that MT developers are probably asked most often is: "What level of accuracy does your system produce?" It is useless to even attempt to define a general criterion of accuracy for a production system like Systran that is used to translate so many different types of documents for so many different purposes. Each user must define his own measure of accuracy, as most large users, such as the CEC and the U.S. Air Force, have already done. KfK, using its own criterion of comprehensibility, determined that the number of understandable sentences in unedited French-English output increased from 75% to 95% between 1980 and 1985, a period during which the dictionaries and linguistic programs were being improved with the help of feedback from KfK. During the same period the number of incomprehensible sentences dropped from 6% to 1%.

One component of accuracy that an MT system has been shown to provide is consistency of terminology, which can be difficult to achieve when many human translators work on different parts of the same document. As technical documents grow more specialized, there may often be only a few experts who can supply the proper terminology, as was the case with the addition of space terminology to the Systran Russian-English and English-Russian dictionaries in 1973 for the Apollo-Soyuz project. The expertise of these specialists was incorporated into the Systran dictionaries, as MT was the only feasible way to supply accurate translations of the highly specialized documents needed by Russian and American engineers in time for the joint meetings crucial to the success of this project.

Many tools are available to the linguist for improving the accuracy of Systran output. One of the most important tools is the easy-to-learn macro language used by Systran linguistic programmers, which eliminates the "language barrier" between the linguist and the programmer by making it easy for linguists to gain the programming skills needed to write their own programs. Another is the extended lexical expression coding facilities, which allow an advanced dictionary coder to create a lexical expression that is equivalent to a small program.

Systran's accuracy is increased by the fact that the amount of information that can be stored for a word or expression is virtually unlimited. In the internal computer representation of a sentence, a fixed length of 160 bytes is available for each word, with some of those bytes containing pointers to variable-length areas for storage of additional syntactic and semantic information and meanings. It is important to note that both syntactic and semantic information is available to Systran linguistic programs at every stage of source language analysis and target language generation. Systran's semantic classification system consists of a set of 500 hierarchically-structured semantic categories. Since no system of semantic categorization can hope to give a complete definition of the world, this system is updated from time to time after careful study of areas where gaps exist.

There is hardly time for a detailed discussion of all the linguistic programs in Systran so I will mention only two types of programs that are crucial to ambiguity resolution and thus important contributors to output accuracy. The first type is called a homograph routine; its function is to resolve part-of-speech ambiguity, such as the noun-verb ambiguity, of which the word "test" is only one of the innumerable examples encountered in English. Every Systran source language analysis contains a set of these routines; they are an especially important element in analyzing the English language, in which at least half the words in every sentence are homographs, a percentage which increases as the style becomes more telegraphic. It is probably safe to say that Systran has the most highly developed homograph routines of any MT system in operation today.

The second type of linguistic program is called a lexical routine; it is one

of the principal means of resolving the word-sense ambiguity which frequently continues to exist after the part-of-speech ambiguity has been resolved. Lexical routines are also used to enhance the modularity of the system by combining all the rules for performing a certain type of linguistic transformation into one easily updatable program module. In the Systran Universal System developed by World Translation Center in La Jolla, most lexical routines have the added advantage of having an analysis component which is independent of the target language. In this system, lexical routines can also be called (with different entry points, if necessary) from any point in the system, including from a dictionary expression, thus giving added flexibility to the system.

It may seem strange that lexical routines can be called from a dictionary expression, but Systran dictionaries are nothing like conventional dictionaries. Almost any type of rule used in a linguistic program can also be used in a lexical expression, which means that these dictionaries offer many powerful tools for ambiguity resolution. Thus dictionary expressions are often used to resolve homograph ambiguities and even to parse certain word groups, as well as to select target language meanings. For that reason, each expression is given a point of entry that allows it to be matched at the most appropriate stage of sentence analysis.

Systran dictionaries also offer the possibility of a default word-sense resolution if none of the other methods of ambiguity resolution has provided a solution. This is the subject field code, or "topical glossary" code, which is used extensively by the U.S. Air Force with its wide range of subject fields. Currently UP to 30 such codes are available, and more may be added.

Accessibility: Increased Availability via Service Bureaus and On-Line Terminals

Increasing the accessibility of Systran means making Systran easily available to anyone who wants to use it. To functionaries at the European Commission, that means making available user-friendly word processors, with appropriate telecommunications links, to all translators and clerical staff, as well as to all senior staff members who have a need for translations, and then providing them with training in everything from how to select the documents that are best suited for MT to techniques of rapid post-editing.

Increasing the accessibility of Systran also means making Systran available, via the new service bureaus, to a large number of users who cannot afford the expense of an on-site installation. Much of the attraction of these services lies in the fact that they are able to offer users a choice of Systran packages: raw MT output, which is becoming increasingly popular; partially post-edited output, using the rapid post-editing method developed at the CEC; and fully post-edited output packaged into a finished document. The U.S. Air Force has its own brand of partial post-editing, facilitated by the use of a software program which examines the internal results of the Systran translation and calls the post-editor's attention to the roughly 20% of the document that may need post-editing. The results of this partial post-edit are so satisfactory that only about 1% of the MT output is returned for full post-editing.

To some, increasing the accessibility of Systran means something even more revolutionary -- making Systran available to individuals around the world, when just a few years ago it was accessible only to a few large organizations. The U.S. Air Force has taken a great leap forward in this area by making Systran directly accessible to researchers via 800 on-line terminals, while Gachot S.A. has introduced Systran on an experimental basis to a potential 4.5 million users of Minitel in France. The Systran Corporation's plan to provide on-line access to English-Japanese translations from a major international database is a recent exciting development in this area. Demand for MT is certain to increase dramatically as users of these new services recognize the advantages of having a translator available 24 hours a day, and this in turn is certain to spark a new wave of linguistic development as Systran faces the challenges of an ever-increasing variety of linguistic styles and terminology.

Flexibility: An Open-Ended, Easily Extendable System

Systran's flexibility is perhaps best summed UP by the statement which I

quoted earlier: the observation that Systran "is used more widely and by a greater number of users for a larger diversity of purposes than any other system currently in use." The diverse needs of Systran users, from the Xerox Corporation with its controlled input and need for 100% accurate translation of copier service manuals from English into 5 target languages, to the European Commission, with its ultimate need for 72 language pairs and perhaps the greatest diversity of document types, have always been the impetus to new developments in Systran, and we are grateful that these needs have been so wide-ranging as to make Systran as versatile a tool as it is today.

An important element of Systran flexibility is the modularity of its programs. It is this modularity which makes the system easily extendable to new language pairs. The source language analysis programs, which represent the largest component of the linguistic programs, are the same for every language pair with a common source language. Thus the amount of linguistic programming required to add a generation component for a new target language to an existing source language analysis is kept to a minimum, thanks to such useful features as multi-target lexical routines and source-language-independent target language generation modules. The development of the Systran multi-target dictionary in 1983 extended the same modularity in source language analysis to the Systran dictionaries, allowing one source language dictionary to be used as the basis for multiple language pairs. This concept was first implemented for the Xerox Corporation with its need for a multi-target dictionary from English into 5 target languages.

The flexibility of Systran architecture is indicated by the ease with which linguistic developments in one language analysis can be transferred to the analysis of another language, even a very different language. Thus Russian analysis contributed much to the development of Japanese analysis, as well as to the analysis of more closely related languages, and the French analysis programs were based on the English analysis programs, which in turn facilitated the creation of Spanish, Portuguese and Italian analysis programs.

As a final comment on system flexibility, it should be added that Systran displays an encouraging robustness, which allows it to meet the demands of not-found-words and various types of ill-formed input without flinching. It may be argued that some badly mangled "sentences" are best left untranslated, but the important thing is that the whole translation process does not come to an abrupt halt when such input is encountered. The system has also shown itself to be updatable on a regular basis, with rigorous testing ensuring that each such update produces a measurable improvement in translation with no serious degradation.

Conclusion: Toward the Goal of Better World Communication

It has been inspiring to all of us who have worked with Systran over the years -- and we hope to all other MT developers -- to see that Systran has been able to satisfy the needs of its sponsors and users at the same time as it has been helping to bring MT closer to the goal of fully-automatic high-quality machine translation. Systran has also helped to build a bridge between MT and human translators, because input from translators in the various user organizations has played an important role at every stage of system development. Systran and other early MT systems have also contributed to the development of artificial intelligence and will continue to do so as, for example, existing MT dictionaries are used to supply expert systems with large knowledge bases: in fact, MT systems are themselves expert systems, and MT system developers in turn have much to learn from the latest advances in AI. Conferences such as this give us all an opportunity to gain valuable insights from each other and to mutually reinforce our shared belief that it is possible to use computers to improve human communication. In this way we are all helping to fulfill one of the major goals of Systran developers -- world peace through better world communication.

REFERENCES

1. Proceedings of the World Systran Conference, Commission of the European Communities, Luxembourg, 1986.
2. Nirenburg, S. (ed). Machine Translation, Cambridge University Press, 1987.