[*Translating and the computer 13: The Theory and Practice of Machine Translation – A Marriage of Convenience?* Papers presented at a conference … 28-29 November 1991 (London: Aslib, 1992)]

SYSTRAN: OR THE REALITY OF MACHINE TRANSLATION

François Secheresse

Systran

INTRODUCTION

Let me first tell you who I am. My name is François Secheresse and I obtained 10 years ago a Master's degree in foreign languages applied to technical translation. I have been working in the field of translation since then. My first job was as a technical translator for GEC Alsthom. After two years in this job, I became head of the transfer of technology translation department. As head of this department GEC Alsthom's general management asked me to do a survey on the various computer-aided translation tools. The result of this two year survey was the creation of an interactive computer aided translation department. This enabled us to increase out productivity by 250% after a period of 18 months and we built up a dictionary containing 20,000 entries.

Having personally checked very often the benefit and gain of time offered by these kind of tools, I decided to start marketing them. I therefore now am in charge of the Marketing and Sales department of SYSTRAN. But I am not going to advertise for SYSTRAN! This paper was written in French and then translated by a human translator. Why not by Systran? The reason is very simple, such a text can in no way be fully machine translated. I want to remain realistic in my presentation of Systran, and both the advantages and the drawbacks will be explained.

TRANSLATION TOOLS

In my opinion, one can divide translation tools into three main groups. The first group includes any kind of automatic translation tools, thanks to which no human intervention is necessary during the translation process. This is the case of SYSTRAN for example. The second group includes computer-aided translation tools such as ALPS, TOVNA, GLOBAL LINK, and PC TRANSLATOR. The third main group includes electronic dictionaries with an online access, such as TERMEX or the COLLINS ONLINE, for example.

As you have seen, the market now offers many tools corresponding to many applications, many documents aiming at many targets. None of them can reasonably claim to help translate any kind of document and to solve any kind of problem raised in the field of translation. But some being very complementary to others, the user might succeed in translating very different documents, provided however he has many of these tools.

Moreover, it is very important to stress the differences existing between the various ways machine translation may be seen throughout the world. In Japan or Canada for example, investments made are far beyond those made in Europe. As a matter of fact, Europe remains highly sceptical. In France some firms are members of CIGREF (Club Informatique des Grandes Enterprises Françaises) and all are very interested in machine translation, but this interest hardly goes beyond mere intellectual interest, and in fact these new tools need financial investments to be further developed.

SYSTRAN SYSTEM

The Systran system was developed in the United States at the beginning of the sixties. First the English-Arab and then all of the existing language pairs were bought by a French group in 1986: the Gachot Group.

Systran users are mainly big groups working in the chemical, computer, energy, aeronautics and electronics sectors. The Systran system is not integrated into the user's PC. In fact the host computer is in France (in Soisy) and is linked to all of Europe.

There are three ways of using the Systran system nowadays: (1) via a telematics service centre, (2) you can have access to our Express-Translation system which I will explain below, and (3) you also have the possibility offered by Systran to rent one or more language pairs to install them directly on your mainframe. The only condition is to be equipped with an IBM computer or a compatible computer that may serve as a host computer operating a VM or MVS system.

Express-Translation

The person wishing to use our Express Translation system must have a PC fitted with a modem so as to send documents via the French Transpac network, or the English PSS network for example. Express-Translation opens access to all of Systran's technical dictionaries, about 20 in all. The result obtained by Systran with the various language pairs varies according to the stage of development of the parsing, and also according to the number of entries in the dictionaries.

But how is a document processed via Express-Translation? The first step is to turn the paper document into a magnetic file. There are two ways of doing this: with a scanner if the paper document is of good quality, or by retyping the whole document. The best is of course to have the document on a floppy because it is rather frustrating to take minutes retyping or scanning a document when its translation will take only a few seconds!

Then comes the pre-editing of the document. It consists of three main steps. First is the spell check. Second is what we call the maximization, which is the optimization of syntax in order to perform the translation. The aim being to guarantee that the document sent to Systran has a structure that is compatible with what the system is awaiting. If a sentence is too long in a French source text for example, the maximisation software will display a message to inform the user that it would be better to make two sentences. Third is the possible prevention of the translating process for some terms, some sentences, even some complete paragraphs. There also comes a conversion step. We use our Format-Text software to convert a Word or WordPerfect file into an ASCII file. All format codes are put aside and the ASCII file is then translated.

After having selected a language pair, we tell the system which specific dictionaries will have to be reviewed first. Among Systran's 20 technical dictionaries, let us take the example of computer, mechanical engineering and electronic dictionaries. The system will systematically review the general dictionary. Then it will review all remaining dictionaries alphabetically. But before selecting the three technical dictionaries we may inform the system that the first dictionaries to view are the personalized ones. These have been prepared either by Systran's linguistics department for a specific user, or by the user himself on his own PC computer.

Once the translation is complete, which takes about one minute per page (including transmission), we re-convert the ASCII document thanks to our Format-text software and re-

insert each and every format code. The final document is then identical to what the original document was. The last step is the post-editing: the document is reviewed, both in terms of terminology and syntax.

What about the quality of documents translated by Systran? Everyone interested in machine translation has one systematic question: what quality does our system offer? I personally think that this final quality depends on the final use which is going to be made of the document. I find it rather difficult to speak of quality in absolute terms. Even if their needs vary, most users tend to label translations as either good or bad. A few years ago, Systran could have been compared to a black box preventing the user from any kind of intervention.

SYSTRAN IMPROVEMENTS

In the first place, we improved the system by enabling the user to ask Systran's linguistics department to integrate a specific terminology adapted to certain major industrial sectors, such as electronics or energy. In the second place, our linguistics department created specific dictionaries appropriate to the home terminology of end-users. The only problem is that the answering time has been lengthened and the user has to wait from 10 to 15 days before getting an answer to the asked list of terms. We therefore had to offer the end-user the possibility of creating by himself sort of an instantaneous coding. The result of it was what we call the CDS (Customer Specific Dictionary), a coding tool making possible the creation of a specific dictionary directly on the user's computer.

To come back to the problem of machine translation quality, there seem to be today three different levels of quality. First of all machine translation can provide rough material for internal use. Of course, there still remain a few terminological and syntactic approximations. But this is not a problem since the end-user now knows what it is about and will be able to determine whether he needs it revised or not. The second quality level nears human translation quality and you can obtain it by post-editing the translation. The very question that remains is: Is it possible to have, after that post-editing, the same quality as you would have with a real human translation and is there any gain in time? My answer will clearly be yes, provided however that the original document is adapted to machine translation. My answer will be no if the document's syntax is not compatible with what machine translation systems await. As a matter of fact, no machine translation system is universal. It would be difficult to believe that the possible gain in time is always very great whatever the type of text and topic. We must admit that for some documents the time spent rechecking and post-editing one page will be more or less the same as the time spent for simply translating this page. Machine translation cannot be asked to offer immediately a significant gain in time and productivity. One has to teach the system first so as to improve quality and optimize it.

The third level is located in between the two other extremes. This intermediary quality goes far beyond the normal quality of a machine translation while remaining below the quality of a human one. Any kind of approximation or error, be they syntactic or terminological, must be corrected but when the sentence is clearly understandable and correct and even if there is a smarter way to say it, it can remain as it came out of the computer. Most of the time, this last type of correction is highly disliked by all translators since they have to lower their standards. They find it highly frustrating to leave, for example, an active sentence when it should be a passive one.

I now am going to review the parameters which help optimize the automatic translation. As I said no system will automatically know the syntax of each and every document or the terminology of each and every technical field. Let us image we are beginning the work with a firm specializing in optical fibres. We know perfectly that our system is not specialized in this terminology. We therefore are going to prepare a specific dictionary, both for the optical fibres technology and for the home vocabulary of the firm.

The first step is a corpus analysis. We are going to study about 1,000 pages representing the company's documents. We will note how frequently some terms occur, we will draw up lists of these most frequent key words and then we will prepare a dictionary dedicated to all the documents coming from this firm. Terminological approximations will thus be highly reduced and the post-editor will gain a lot of time. But this is not enough and most of the work consists in restructuring the document. We therefore have to concentrate on the syntax. The best solution would be to work with the technical writers who created the documents, since they must understand that the way the document was written has a decisive influence on the way the translation system will react. The aim is not to change the terminological richness of the original document but to try to prevent any kind of complex syntactic structure, thus preventing mistakes and ambiguities. Once this is done, the automatic translation's quality will be improved, thus facilitating post-editing. Our experience has shown in some of our customers a strong desire to play the game of machine translation according to these rules, thus leading to a real gain in time. Therefore, nobody can say that machine translation is not profitable. Machine translation cannot be profitable in two days but really is profitable in the long run.

THE FUTURE FOR SYSTRAN

Now what are the prospect and tendencies for Systran? Our main target is to have for all language pairs the same quality as that we have obtained for the Russian-English couple. We are perfectly aware that our parsing of the source English language is not sufficient today, and this has a real influence on the quality of the translation. We have the know-how but once again all this depends on the budget. The lack of finances prevents improvements from being as rapid as we would like them to be. At the beginning of 1991, we launched a database enabling us to take into account the feedback of all our end-users. This database helps us improve the system according to the users' needs. A users' club is also now being created.

We have signed agreements with some firms to use the Systran system at the group's level. In the framework of our partnership with Rank Xerox, we have prepared an advanced solution Docutran, which is an interface opening access to the Systran host computer from the 6085 or from the Globalview workstation. This offers a considerable gain in time since the end-user receives his translation in the same format and under the same presentation, be it from the text point of view or from the graphic point of view. The computer-aided publishing that usually takes place immediately after the translation of the technical text is thus eliminated. Docutran is the very first system that offers the possibility of effectively integrating the translation process into the document's production process. Another project of Systran is to interface Ventura and Systran. We are now creating a terminological database called Multidic which you can have access to via our French Minitel. This database will be a great help to documentation and translation centres.

It now also has become possible in Europe to install Systran on site, as it used to be the case in the United States during the last few years. Any firm with significant translation

volume now is in a position to ask for a license to use the language couples it needs. We also intend to rewrite all of Systran's programmes in C language. Most of them are now written in assembler. This rewriting in C will offer the possibility to use Systran on Sun workstations operating the Unix system. This will also help solve the problem of secrecy encountered by some major groups. As a matter of fact, companies dealing with classified documents prefer a translation on a home computer rather than a connection to an outside mainframe.

The mentality is changing as regards machine translation. Systran will be available at the City of Sciences in La Villette for a period of seven months, as part of the Machines to Communicate exposition. This change in mentality, this ever greater interest in machine translation can, in my opinion, be explained by the combination of two factors: the permanent technical improvements and the much more objective and realistic way machine translation is being approached. Today, we are perfectly aware of what the contingencies are, and of what the difficulties encountered represent. The aim is no longer to replace human translators by machines, but to stress the highly obvious complementarity existing between them, the machine constituting a mere helper to the human translator.

I think this convergence of these two factors may lead us to envisage very interesting cooperation with, for example, the universities of Lille, Angers or Paris VII. The Paris VII university asked me last year to teach to students already having made five years of studies after Baccalaureate. The lecture I gave was about translation tools. I was delighted by this first experiment and intend to do the same in 1992. The French university at last has decided to open its doors to some new techniques, because it has realized that students learning translation today would be the translators of tomorrow. They therefore need to know the tools they will have to work with.

Terminology has an ever larger place in the process of traditional translating and machine translating. I think we are heading towards a generation of translators who are going to handle systems like Systran in a natural and effective way. They will have the luck of avoiding all the repetitive tasks often linked to translation and devote their time to their professions' most interesting aspect. Future translators will help create real partnerships between the people that design these systems and the people that use them, provided however that these tools help gain enough time, and provided the documents obtained are of sufficient quality to be revised. As I have already said, there are many parameters to be taken into account if we want to obtain a good quality translation and nobody can reasonably claim that any machine translation system will lead to a considerable gain in time and productivity if its quality is not sufficient.

As far as I am concerned the quality of machine translation depends on the systems' performance of course but also on the end-users' motivation and on their readiness to change their work habits. If I dare repeat myself, there is no universal system. We may have excellent opportunities with Systran and other existing systems provided however that no mistake is made at targeting the documents. The game has got precise regulations and rules. If you respect them, then after some time, you will be able to raise your productivity considerably.

As you see, when used properly, Systran might help you increase your productivity considerably. But I dare say some documents cannot be translated by our system. A firm wishing to see all of its technical documents machine translated would undoubtedly have to possess different types of systems because of the various problems to be encountered.

AUTHOR

François Secheresse, Systran SA, 26 bis, avenue de Paris, B 14, 95230 Soisy sous Montmorency, France