Jean Gachot

Resurrecting

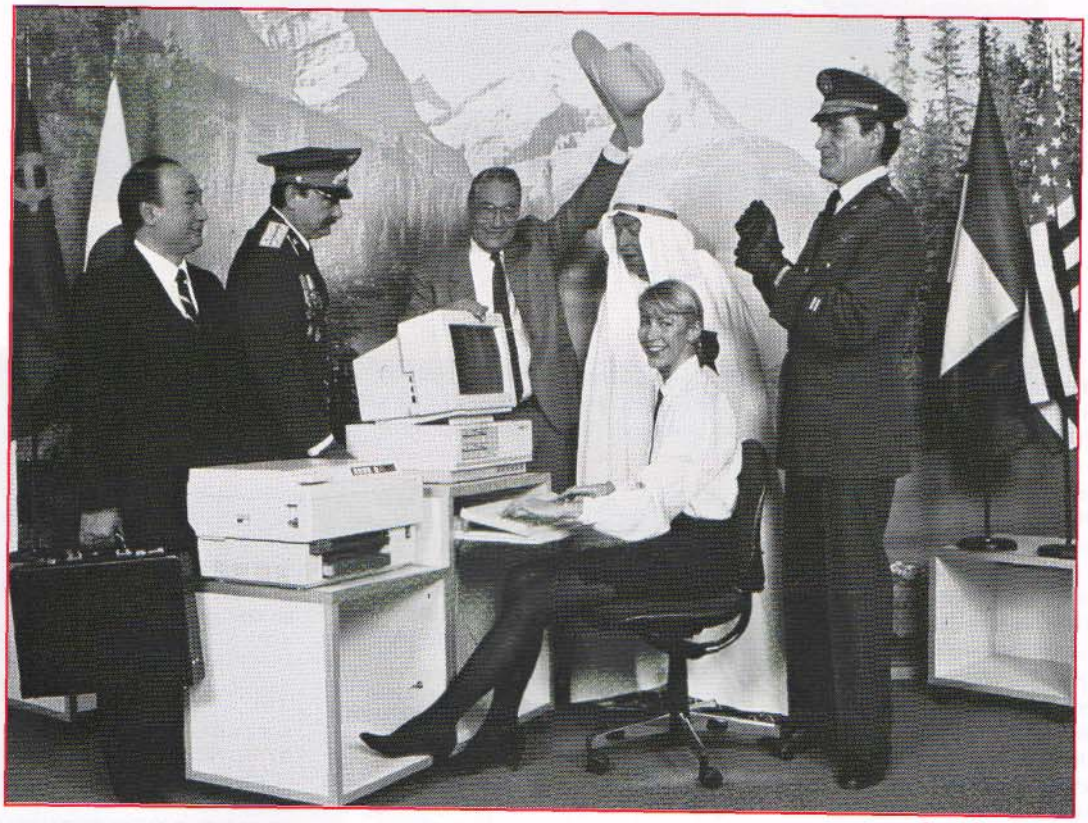

Photo: Georges Bosio

Systran

Nobody, except John Smart, has ever made money from Machine Translation. Now industrial valve magnate Jean Gachot has bought all the rights to Systran, the venerable MT workhorse. Will helping kids with their homework help him make machine translation profitable?

by Andrew Joscelyne

The story of how the managing director of a group of companies, whose interests run from industrial plumbing taps through plastics to software development, became the owner of the largest automatic translation system in the world began in the early 1970's when Jean Gachot happened to meet Peter Toma at Davos during a management symposium.

Toma, the inventor and first purveyor of Systran, had been forced to finance his ongoing research into his offspring by selling off various operating rights. The European Community, for example, bought the rights to use certain European language pairs in the public sector. Systran Institut GmbH, with branches in Germany and Luxembourg, had bought the rights to the Arabic synthesis program.

By the late 70s, Systran had fragmented into American, Canadian and European branches. Development of the whole shebang was more or less immobilized, since no single user wanted to invest money in a system others might benefit from without paying for. The lack of Systran development left the market open to new challengers, with corporations and national governments stepping in to develop second generation MT systems like Eurotra, Rosetta, Calliope and others.

Over the last 20 years, some $40m has been pumped into improving Systran. Add another $13m invested by the EC on their language pairs, plus other investments made by Xerox and the US Air Force) and you have a round figure of some $65m in investments.

Gachot collects the pieces

In 1982, from his eyrie at Soisy to the north-west of Paris, Jean Gachot cast his wary eye over the geo-economic scene and reckoned that owning rights to the Arabic synthesis module (the part of the translation program which generates Arabic target language output) might offer future commercial rewards. So, he signed an agreement with Systran Institut to finance the development of this relatively longshot module.

Gachot is a man who likes to look into the future, and make a profit while he's at it. Back in the 60's his industrial valve company had been among the first to computerize various commercial and technical tasks. After installing a big number-cruncher, he developed a programming environment in French called Ulysse, still marketed today by a company called Steria. So when the unhurried Systran Institut proved too slow delivering the goods, and he discovered that he had spare computing capacity on his hands, Gachot, "a man in a hurry to get results," as one of his employees put it, set up a special unit at Soisy, putting Dr. Sami Trabulsi, his present Systran head, in charge. And bit by bit, Gachot set about collecting all the various pieces of Systran. Or almost all of them.

In 1984, he acquired the rights to introduce new words into the basic dictionaries (see sidebar for a brief sketch), then a year later replied to the EC tender by making Soisy a service center for Systran in France. In 1986, he bought into the Systran Institut, then signed an agreement to buy WTC and LATSEC in La Jolla, the companies Toma himself had originally set up. The falling dollar will doubtless help Gachot pay off the cool $5m he has to fork out for them. Also in 1986, Gachot acquired a further 39.4% of Systran Institut, a figure scheduled to rise to 76% by 1990, for a total cost of $1.7m.

And so, by the end of last year, the man in the stetson was heading an international group of Systran operators, with his son in charge of the US branch, and a new Gachot subsidiary, Systran International, ready to run the to-be-unified Systran operations. One fish, however, escaped the Gachot net.

"Unfortunately," Jean Gachot admits ruefully, "we have not been able to buy the IONA company holding rights to the Japanese programs. A pity, since Japan is by far the largest translation market." Indeed, available figures show that last year the translation market in Japan was worth some $4.8 billion, almost double that of the US, next down the list. Working agreements and odd spinoffs will have to compensate for the big one that got away.

Putting them together again

Malicious tongues have murmured that Gachot has, in fact, bought up a megadinosaur from MT's paleolithic age.

Systran might not have been cobbled together from the leftover punch-cards dumped in Georgetown garbage cans, as one wag has put it, but it is certainly true that if the program were written today it wouldn't be done in pure assembly language, what with skilled assembly language programmers becoming increasingly difficult to find.

Gachot, naturally, sees things differently. "Over the last 20 years, some $40m has been pumped into improving the system. Add another $13m invested by the EC on their language-pairs, plus other investments made by Xerox Corp. and the US Air Force), and you have a round figure of some $65m in investments."

An outside analyst, however, believed these figures exaggerated. "The value of the dollar may be low now, but that doesn't mean you can go back and recalculate the dollar figures. The EC spent perhaps $10 million, while the other organizations may have invested 5 or 6 million, not more."

The Gachot Group, in inheriting the sum total of local improvements, can now proceed to what Toma might only have dreamed of – the grand unification of current systems.

While 98% of the code is the same, divergence among the different versions has occurred. Improvements, for example, were implemented simultaneously, but differently at different sites, especially with powerful new dictionary features. Thus, while the results were, for all practical purposes, the same, the code is different. In addition, the Systran Institut's inability to carry out agreements with Toma and the EC forced the EC to simulate various improvements on its own. And then there's the matter of programmer's personalities.

"Programmers in La Jolla, for example," one observer close to Systran commented, "have found it especially difficult to accept improvements not made inhouse."

So the first task of the new dispensation has been to incorporate the US multi-target system parameters into the best results of the European developments. This entails everything from dictionary entry unification, to creating a series of practical macro-instructions.

"But in some cases, changing a single byte can create a change in the whole system," explained Sami Trabulsi, Gachot's Systran technical director.

To ease intercontinental communications in this effort, a $10,000-a-month direct satellite connection has been established between La Jolla and Paris, as well as an extension to the computerless Luxembourg subsidiary of Systran Institut. At the beginning of this year, the whole Systran works was in fact rerouted onto the US system for direct access to all dictionaries. Result: a massive increase in the speed of translation.

The effects of Gachot's strategy can already be seen in his relations with EC Systran users. Now that Gachot has bought the operating rights for the EC, longstanding problems in Systran's European translating environment are apparently about to disappear. In fact, Gachot's new deal is based on the old WTC-EC agreement which includes nothing about continued co-operation. But the EC demanded that a unification requirement be written into the new agreement. The two parties duly signed, agreeing to work together in the dictionary-making field. Now, the EC's 32 staff handling lexicographical problems via Informalux, the service company, are helping Gachot co-workers unify the semantic and syntactic codes that have hitherto diverged.

Quality output

Exploiting the Soisy IBM 4381 and the La Jolla IBM 4361, both totally dedicated to word-crunching, the Systran teams hope to hit the legendary 96% quality output in 12 language pairs by the 1990s. According to a report published last November, the epoch-making Russian-English pair at the La Jolla base has already reached this MT quality control Nirvana for the translation of technical texts.

Jean Gachot feels he needs to invest only another $10m in order

to reach this target – small change when compared to the investments plowed into Systran over the past two decades. Of course, unification of the variants includes aligning all relevant pairs on a single English source, and the consequent increase of all dictionaries to a 450,000 word base. Owning the whole works, though, will speed this up without having to fall back on client comments filtering through to the labs as has usually been the case.

The Systran team will nevertheless have their work cut out for them if they wish to meet their deadlines: an in-house LATSEC report on quality control for language pairs, which was circulating at the end of 1987 showed that even if French-English/English-French reached an 86.3% overall quality (most errors being "meaning problems"), English-German and English-Russian are still only at 67.3%. Dictionary updating for some of these pairs will certainly reduce the "Not found words" error, but various other linguistic programming problems remain. This quality target is important for Gachot's long-term commercial strategy, since he sees automatic translation not merely as an industrial language service but as a potential benefit for the whole industrial family.

Feedback suggests, however, that a lot of schoolkids access Systran to help them with their language homework, even though the lack of idiomatic usage and the abundance of plain errors on the current system must be getting them bad marks.

Commercializing Systran

Gachot's "brainwave," as Trabulsi puts it, has been to commercialize his big bag of tricks by means of what he calls "teletranslation" – exploiting, on the one hand, the growing value-added telecom networks to reach a market of industrial users who would normally be unable to afford a monster like Systran, and on the other, reaching the general public through videotex terminals.

At first, WTCC (Canada) and EC markets were serviced by offering custom service to individual clients, i.e., catering to individual front-end sensitivity by using specific terminals, using a unique entry format and developing specialized post-editing software. Later, WTCC developed plans to use independent wordprocessing equipment, and the EC began to introduce the possibility of interfacing with anyone with standard telecommunications protocols. Gachot has decided to continue and expand this "distribution of computing power to a large variety of users," offering his dedicated mainframe to any client willing to pay the subscription. The users buy translation credits, access any of the available language pairs in Systran from their PC using their own entry format, and receive raw output for post-editing on their own terminals after a short delay.

To offer this service, Gachot engineers have had to develop interfaces to handle different WP formats such as WordPerfect and Wordstar 2000, whether under DOS or other operating environments. However, you can't access Systran on a Macintosh, although it's promised by the enf of the year.

A second feature of this teletranslation service bringing Systran to the people is the creation of client-specific filter dictionaries. These miniglossaries, of 4000 words maximum, do not emulate the "personal dictionaries" available on lightweight CAT systems, but resemble house-style search modules. If, for example, a company always uses "KG" in documents for the name of the boss, the filter dictionary willprevent Systran translating it as "kilogram."

New clients

An example of a major company interested in Gachot's on-line teletranslation agreement is Aerospatiale, with their annual 120,000 pages of English-to-French technical documentation and other texts. Each division (aircraft, helicopters, tactical missiles, satellite technology, etc.) has its own terminology requirements. They access Systran directly for large quantities of translation, or call up Aerospatiale's Suresnes central information bureau via an IBM network for the small stuff. Systran output is, of course, raw language. The problem for any large text manufacturer like Aerospatiale is that translation as such is only one link in the chain of document production.

As documentation chief Oleg Lavroff puts it, "A system like Systran is 'heavy' and must somehow be blended into the process without unbalancing the whole operation."

The problem, as they say, is under study.

Aerospatiale also needs speedy translations of articles, works-in-progress and congress reports for their research engineers. Originally, Aerospatiale thought Systran would be the ideal solution. At present, however, as Lavroff explains, "Systran output needs considerable improvement to be understandable to a hurried reader without radical post-editing."

Gachot responds that if all terminological inputs are codified and the current aleatory syntactic problems are cleared up, post- editing should take only 4 minutes per page. First, the inhouse OCR will process texts at 30 seconds a page – say a minute allowing for corrections. Then with current translation time, using the TRANSPAC 1200 baud telecom network, an engineer could have a 6 page (1550 words, 10750 characters) French version of an urgent English report in just over half an hour.

If the translation credits system is slowly gathering momentum for the industrial user, the general public videotex translation market is Jean Gachot's pride and joy. Pay him a visit, and he will immediately switch on his minitel and start tapping the MITRAD code to access his own brain-child. He claims an average of some 1400 connections per day generating around 190,000 words translated per month (a 76% increase on the first six months of 1987).

"This is perhaps Gachot's main achievement," Ian Piggot, EC Systran head commented. "Until now, the whole field of machine translation was regarded with skepticism. Through the minitel, Gachot has actually brought MT into the average home."

When his photo and the story of Systran appeared in the Figaro weekend magazine this January, it apparently stimulated a record 500 hours of connections over that weekend alone. Goodness knows what users made of what they got for their $10 an hour: you really need to be bilingual to appreciate the wayward nature of some of Systran's output. (Feedback suggests that a lot of schoolkids access Systran to help them with their language homework.)

Here's an example of the kind of quality now on tap. The following bit of LT advertising:

Language Technology is the only magazine in the world focusing on the people, the companies and the products that are filling the new technological toolbox.

Translating back by Systran from its French translation resulted in:

The technology of language is the only store in the world focusing itself on the persons, the companies and the products which fill the new one toolbox technological.

Yeah, well, LT may open a shop, but the text still needs a post-editor. It is not yet the complete automatic translator its publicity boasts of. Although, in fairness, the quality of translation is directly related to the quality of the

Gachot's "brainwave" has been to commercialize his big bag of tricks by means of what he calls "teletranslation" – exploiting, on the one hand, the growing value-added telecom networks to reach a market of industrial users who would normally be unable to afford a monster like Systran, and on the other, reaching the general public through videotex terminals.

source text, which often suffers from appalling spelling and grammatical mistakes.

Ian Piggot again: "The main problem with minitel is not the quality of the system or the dictionaries, but the quality of the input itself. The only answer is for some sort of online editing system, on the order of John Smart's Max."

Still, consultation of the range of dictionaries (slang, science, industries, aeronautics) is on the increase, and the impatient Gachot is goading the sluggish Harrap's into more agreements for electronic dictionary compilation. Soon Aerospatiale's own vast glossary will be available on minitel – no doubt downpayment for translation credits – but, oddly, competing with the already accessible Graissin aeroglossary.

But the man in the stetson's motto is "ever onward." He intends to include a macro-control key on his minitel service which will allow users to look up a word from dictionary to dictionary at the touch of a single key, allowing translators speedy cross-checks. Other plans include a multilingual juridical database as well as loading the vast database on Russian held on paper at present by Le Monde newspaper.

The Future

Gachot visited Japan and persuaded Nippon Telegraph and Telephone to buy his translation service for their local videotex network. Apparently there is, as the saying goes, a "vast market" for the speedy translation of commercial documents from other European languages into English for immediate understanding by English-speaking Japanese, for which a continually increasing battery of language-pairs is on tap.

Systran engineers can now develop new Minitel-accessible language pairs in only a year: Portuguese-English and Italian-English are the latest to become available.

And yet, despite this exponential growth in data-access, the first generation Minitel terminal is still ill-adapted to the speedy typing of long texts, and analog telephone lines are too slow for active information exchange. More and more users are therefore downloading relevant screen pages onto their PCs and then working directly from their own files. Where Minitel does come into its own, as most owners know, is in electric message services, whereby conversational exchanges can be made using a Minitelese writing style in the most perfect anonymity. Mmm, could be a market there, you imagine vaguely.

Re-enter Gachot: coming soon on your screens will be a new international interlingual telemessage service – "electronic chatting" as he calls it, specially designed for the handful of New Yorkers with a Minitel. You contact your transatlantic correspondent, tap in "Bonjour," wait 15 seconds and your greeting will arrive on his screen translated by the redoubtable Systran into English. And so on. Actual translation time, which was about 5 minutes for a sentence a year ago — is now down to around 10 seconds and dropping every month, so the illusion of real time conversation will be almost complete.

The next step will, of course, be to provide an on-line translation service for those late-night sexy singles chats for horny tele-cruisers – usually known in France as "Minitel Rose." No sooner had Jean Gachot whipped out his pocket micro-recorder and murmured in the new idea to himself on the run, than his loyal team began working on it.

Almost endearingly, Gachot mentions that his son Denis is combing all the teach-yourself English manuals for everyday expressions to be plugged into the MITRAD service offering the lingo as she is spoke. Given the quality of today's translation, though, some of that would-be smoldering fantasy might prove a bit of a turn-off after it has been groped by Systran.

Japan?

Back on the ground after these manic flights into the Disneyland of interstellar MT for keyboard wizards of all ages, there is the nasty little question of competition from other CAT software and the general direction that translation tooling is taking.

Sami Trabulsi foresees no threat from the ongoing Eurotra program, or any eventual spinoff of France's Calliope program. "First generation systems like Systran," he says, "are far from saying their last word, especially when you realize that Systran has more instructions of assembler per language-pair than any other current system."

Moreover, he argues, the second generation systems now just beginning to emerge from the research labs lack dictionary power, while Systran has built up its vast dictionaries over a long period of practical translation work. He believes, in fact, that there could be "very important exchanges" between Systran engineers and the newer programs. But he fears that after the year 2000, "if we don't look out in Europe, we risk being completely invaded by the third generation systems under development in Japan".

However impressed Jean Gachot himself is by the massive Japanese investments in communications soft- and hardware, after the Hakone MT Summit in October 1987, he felt strengthened in his own resolve to push on with the updating of his 30-year-old Systran.

"The only way to acquire a significant part of the world translation market" he says,"is to increase the quality of translation to the percentage we have determined."

The relatively modest investments this family outfit is prepared to sink into the system seem to be worthwhile because "they are not subject to depreciation with time, and we need not fear any competing technological or scientific breakthroughs." Rapidly perfecting his English-Spanish pair, currently at 87% quality, will, he believes, secure him the lucrative South American translation market. And no doubt recent news of a hushhush contract from the Northern half of the other side of the Atlantic will help make his dreams of a multilingual Minitel Rose (what a tongue she has!) come true. If Systran ain't a winner, he'll eat his stetson.

Andrew Joscelyne is LT's Paris-based Contributing Editor.

Systran claims to be a "universal translation system" able to take any text — i.e., a string of characters separated by blanks into words — discover its words, find their meanings and then reconstitute the same in another language. To do this Systran needs a vast database of word information coupled to a set of inference engines that works them over in search of what the linguistic programs need to analyze or synthesize a language.

The database is composed of two types of dictionaries: the STEM dictionaries offer complete morpho-grammatical and semantic-functional information on each word, while the battery of contextual LIMITED SEMANTICS dictionaries allow word-tracking to the most refined level — there are idiomatic, homographic nominal group, parsing and contextual meaning dictionaries.

The programs which "understand" the source text — analyzing it sentence by sentence via an interpreter, and which "produce" the target version — and synthesize it, are written in the Systran derivative of assembler, which can manipulate specifically linguistic objects. When they are working well, they allow the system to reach a 500,000 word-per-hour rate. With one language pair containing around 30 megabytes of info, together with management and control systems containing 100,000 lines of code, plus the 120,000 lines of macro-instructions of linguistic programming, Systran boasts more instructions per pair than any other MT system.