Dr Peter Toma

Toma explains why he sold Systran rights

Further report on the Luxembourg conference

The World Systran Conference, one of the most important events in machine translation to be held in recent years, took place in Luxembourg in February, shortly before the March issue of Language Monthly went to press. A brief report appeared in that issue. A more detailed report of some of the contributions is given below.

The World Systran conference began with a contribution from **Dr Peter Toma**, creator of Systran and often called the "Father of Machine Translation", who described how he had come to develop the original system, and how important a motive was the concept of world peace.

"I witnessed during the second world war how language barriers inhibited the quest for peace", he told the audience. "At the end of the war it was obvious that we had entered an era dominated by sophisticated weaponry capable of mass destruction and I felt even more deeply that I had to devote my energy to the elimination of conflict-causing factors. As a first step to overcome the language problem, I felt that I should know as many languages as possible".

After explaining how he learned Russian using Linguaphone records borrowed from a refugee in post-war Bavaria, Dr Toma described how he later found himself at the California Institute of Technology when they acquired their first computer. After getting acquainted with the logical operations of the machine, he became fascinated by the "obvious practicality" of using its capabilities to translate languages automatically. He already had a daytime job, and yet needed plenty of computer time to prepare and test out algorithms. As the computer, a Datatron 205, had a drum which had to be shut down in the evening and restarted in the morning, and as there were frequent problems in restarting it, he offered to watch the drum all night, and take the necessary steps should something go wrong, if he could be allowed to use the computer at night to test and debug his programs.

"Of course," he said, "such an arrangement necessitated keeping rather unusual hours. A typical day for me was as follows: my regular work between 8 a.m. and 4.30 p.m., with a short lunch break; eat an early dinner at 5 p.m. and sleep between 6 and 10.30; at 11 p.m. take control of the computer room until 7 a.m. the following morning, subsequently eat breakfast, shower and back to work again at 8. This went on for many months...

"Many of the algorithms which I thought out and tested during those long nights are working in Systran today, although before I devoted myself exclusively to the design and development of Systran I also created several other workable systems in the field of machine translation like the Serna system at Georgetown, and subsequently the Autotran and Technotran systems.

"Systran was really born with the IBM 360 computers in 1963/64. My purpose in designing, developing and implementing it was to have a system which takes full advantage of the latest hardware to overcome language barriers on a large scale. My wish that this system serve mankind from an idealistic point of view always genuinely surpassed monetary considerations. My friends used to tell me: 'Peter, you gave away the farm'."

"I have been criticised", Dr Toma added, "for signing the first contract with the European Commission for a relatively small amount and making available not only the English-French module but for the same small sum the total basic Systran software ten years ago. I was told repeatedly that I should have asked for a higher sum, continuous royalties, etc. My consideration was, and is, Systran's contribution to the Commission activities and the fact that it would promote better understanding between member countries and even between countries outside the Commission."

Dr Toma said that he saw Systran as making a contribution to mankind in two ways. One was that by eliminating language barriers on a large scale it definitely contributes to lessening the tension between nations, thereby helping to prevent confrontations. But now this was not enough. Because of the dangers of nuclear war, Systran had to do more.

"Therefore I have sold World Translation Center Inc., with all the corresponding rights, and decided to use the revenue from the sale to begin developing two great Systran-supported projects which will eliminate the ever increasing danger mankind is facing today. The first project is an International symposium on conflict resolution to be held in Dunedin, New Zealand, on October 28 of this year, the second a private university on the South Island devoted to eliminating conflicts and preparing in a special manner a new generation of statesmen, politicians and diplomats from students, selected because of their talents, from many countries."

After describing in some detail his proposals for the symposium and the university (those interested may contact Dr Toma at PO Box 917, Chief Post Office, Dunedin, New Zealand), he recalled the 1960s setback to machine translation of the ALPAC report.

"Twenty years ago even members of the Academy of Science belonged to the group of sceptics as far as machine translation was concerned. In 1963 and 1964 experts were called before a special committee of the Academy to testify that machine translation could not be done. The dates of the hearing were carefully selected to fall at a time when I was in Europe. The ALPAC report was a devastating blow to machine translation particularly in the United States."

In 1965, however, the German Science Foundation (Deutsche Forschungsgemeinschaft) called a one-day meeting of top German linguists to discuss the Systran approach. Dr Toma gave a presentation and answered questions, and the majority decided that the Systran concept was the right one. This led to the first contract to develop the system on a larger scale.

"Until that time," Dr Toma continued, and even during the following two years I wrote all my programs myself in 360 assembly languages. The first system was debugged and implemented on an IBM 360-30. The first 32 bytes in the analysis area (that is all I had at the time) and almost all the bits in those bytes still carry the names I gave them in 1965-67."

It was the attitude of the German Science Foundation which persuaded the United States Air Force to make an open bid for practical machine translation. Systran tendered against IBM and the Bunder Ramo Corporation, and won the contract.

The next and probably most significant step in Systran's development was the continuous sponsorship by the Foreign Technology Division of the United States Air Force at Wright Patterson Air Force base. The air force, explained Dr Toma, had no suitable computer at that time and in southern California there was only one installation, at the Systems Development Corporation in Santa Monica. But the first IBM 360-50 in Europe had been installed at the University of Bonn, and Professor Unger allowed Dr Toma to use the night shift for debugging Systran during the months of September, October and November 1968, and with the assistance of programmers from the university, Systran was implemented for the first time on a larger computer. Early in 1969 Systran was installed at the Wright Patterson base.

The installation of Systran ten years ago at the European Commission had contributed considerably to the system's success. Other early users who participated in the development of new language pairs as well as dictionaries included NASA, General Motors and the Xerox Corporation.

Mr Ian Pigott, Systran project leader at the European Commission, spoke to the conference about current developments at the Commission. When the basic package was acquired ten years ago, he recalled, it was obvious that because of the wide-range of subject matter likely to be dealt with, it would be necessary to create very general basic dictionaries. "However, in order to achieve specialised equivalents in context, we put tremendous efforts into developing a meaningful set of semantic and syntactic codes which could be used as a basis for contextual dictionary entries relying on the structural relationships of the words in a document."

This development philosophy was now starting to pay dividends, with outside users reporting their surprise at the results obtained in specialised fields that had never been the subject of specific developments. It also meant that improvements made for one user could become immediately available to all other users. While there was still a facility for introducing special meanings by subject field – the so-called topical glossary approach – even when such glossaries were used in any given text well below one per cent of the meanings would in fact have been provided on the basis of a subject field parameter.

Mr Ian Pigott described how dictionary coding and enhancement was carried out, and the growing automation of this work, which will enable future development to be carried out more rapidly.

Describing some of the details of coordination between Systran users, Mr Pigott welcomed the desire for closer collaboration shown by the conference. He called for a rationalisation of the software development, where various groups had been working on their own, for more coordination in dictionary development, for better interfaces with user applications such as word processing, more collaboration on post-editing techniques, and more coordination of user feedback.

The translator's viewpoint was given by **Andrew Evans**, of the Commission, who described experience of the technique called rapid post-editing. A good post-editor needed good translation skills, dexterity in using a word processor, common sense, decisiveness, an ability to work quickly, and an ability he called a "nose for a wrong 'un", that is the innocuous looking translation where the machine translation had got hold of the wrong end of the stick entirely. The rapid post-editor had a limit of 30 minutes a page, and a target of 20 minutes a page.

Mr Loll Rolling, head of information transfer for the European Communities, criticised the policy under which Systran could not be put into wider application using the equipment which had been tried and found to work. He is believed to have been referring in part to an imposed "buy European" policy dictating the use of certain makes of computer and data processing equipment rather than the Wang equipment which the translators and MT specialists have come to prefer.

Loll Rolling

"Vous avez appris", he said, "que les utilisateurs ont été parfaitement satisfaits des services rapides et efficaces rendus par Systran et demandent l'extension de ces services vers d'autres couples de langues.

"Il est d'autant plus regrettable ... que la Direction Informatique de la Commission a choisi de ne pas autoriser la mise en route de l'application Systran sur l'équipement qui a fait ses preuves à la Commission, à Luxembourg et à Bruxelles."

After quoting comments made by Members of the European Parliament Mr Rolling added: "Je pense qu'il est en effet essentiel que l'informatique soit au service des fonctionnaires, traducteurs et autres, et non l'inverse.

"Cela veut dire qu'il faut créer une infrastructure conviviale qui soit compatible avec les besoins des traducteurs et avec les caractéristiques de Systran, et non adapter Systran à une *stratégie informatique* conçue par quelques 'spécialistes' qui ont choisi d'ignorer les caractéristiques du système et les besoins des traducteurs et des usagers."

Mr Rolling suggested that instead of the 2,800 pages a year which the Commission managed to produce with difficulty at the moment, it would be possible, given an adequate infrastructure, to produce 280,000 pages. It had been estimated, he said, that the organisations present at the conference, using Systran, translated between them more than 200,000 pages a year.

Mr José Mendez, of the Belgian commercial translation company Mendez S.A., described the use of Systran in a bureau service environment. A bureau's single largest potential market, he said, was that of multilingual documentation production. As far as technical brochures were concerned, translation represented on average a mere seven per cent of the total production costs.

Using normal (i.e. non-MT) translating methods, Mendez's experience was that production time divided roughly into 35% translation, 15% client revision, 20% each typesetting and studio artwork, 10% printing and distribution. In recent years stiff competition and rapid communications have forced commercial companies operating internationally to launch a new product simultaneously throughout the world market, along with the relevant back-up material.

"Delays in marketing a new product in one particular market could cost the company irreparable damage, especially if its market lead over competition in that particular area is marginal. Speed and accuracy are the name of the game".

To guarantee a constant standard and improve overall quality, the translation industry had been forced to adopt a more scientific approach. This had required an investment in the preparatory stage to avoid unnecessary delays and increase accuracy as well as to integrate machine-aided translation into the working environment, thus improving output.

An integrated system, incorporating image and text processing and machine aided translation, producing a colour page layout, could improve a company's overall productivity by a factor of three. But the initial investment was high.

"At Mendez", he said, "we are in the process of developing just such a system. It is too early to give a precise cost analysis of the benefits of Systran to the service bureau but the advantages are demonstrated by the fact that our freelance translators, using Systran as an additional translating tool, actually approached us, proposing a reduction in their fee because the system rendered their task so much simpler and faster!"

The complete service – pre-editing, raw translation and post-editing – was currently being sold by Mendez at the same cost as traditional translation methods. This had meant a slight reduction in the profit margin but a considerable improvement in productivity and accuracy of output.

Clients were now beginning to differentiate the levels of translation required. If only a quick rendering of sense were needed, very often a client would be satisfied with a raw machine translation. If the document being processed were to form part of a sales speech it required fast post-editing. If it were destined for a prestige sales brochure it would require additional "stylistic gloss".

Mendez took a major step into office automation three years ago when they set up Dataset SA in collaboration with the Antwerp-based printers Graphica. They first came into contact with the Systran system through their extensive contacts with, and work undertaken for, the Commission of the European Communities. Last January they signed an agreement with Orda-B to become Belgium's official Systran service centre. The transition to Systran went smoothly as Mendez had already invested heavily in office automation equipment, including the Wang OIS system.

Mr Mendez concluded by looking at the two camps into which the MT industry was divided. On the one hand there was the user-friendly cheaper desktop systems, offering a wide range of basic software/dictionary packages; this was easier to sell in the short-term but will lead to what he called "software migration" and a disunited user base. On the other hand there was the Systran type of larger, more mature packages, with centralised benefits for all users, and with the added advantage of constant expansion and diversification of common data dictionaries.

"I firmly believe", he said, "that this latter, unified approach to MT will have the best overall results. I also believe that both the private and public sectors should collaborate closely with service bureaux offering Systran (even create a Systran user association), thus encouraging more generalised access and diversified input to MT, until the day is reached when sophisticated voice analysis technology has put a computerised simultaneous interpreting system into the pocket of the man in the street."

Mr Leonard Siebenaler, of the European Centre for Automatic Translation (ECAT), spoke about the use of Systran for translating documents for the European Communities ESPRIT programme, which seeks to foster new technology developments. The expected resistance and scepticism towards machine translation, he said, was rarely encountered. The experience of many of the companies and organisations participating in the programme was that translation services currently are not used because they are expensive, inaccessible or inconvenient.

Mr Mendez's remarks on the necessity for immediate multilingual document support for product development for companies with an international market were echoed by **Mr Leonard Rudorfer** of Dornier GmbH.

Ich glaube, daß in Zukunft, wenn die einzelnen Sprachpaare noch ausgereifter sind, kein größeres Unternehmen, das seine eigene Produkte in alle Länder der Welt verkauft, ohne die maschinelle Rohübersetzung auskommen kann. Die Produkte werden technisch immer komplizierter und der Nutzer kann ohne die technische Dokumentation die Produkte/Systeme weder betreiben noch warten (instandhalten).

Aus diesem Grunde beinhaltet die Anschaffung eines Produktes/Systemes (Hardware) automatisch die Mitlieferung der notwendigen Dokumentation in der jeweiligen Landessprache.

Es wird immer der Anbieter den Auftrag erhalten, der in **kürzester** Zeit in der Lage ist, neben der Hardware auch die technische Dokumentation in der jeweiligen Landessprache mitzuliefern.

....Abschließend möchte ich noch sagen, daß ich mir die Zukunft ohne die Hilfe der maschinellen Rohübersetzung nicht vorstellen kann.

A major contribution was given by **Mr Dale Bostad**, of the United States Air Force Foreign Technology Division, Dayton, Ohio, who spoke about the use of Systran by the air force, and explained the editing software now developed. Between 50,000 and 60,000 pages of Russian are translated by the system every year, as well as several thousand pages of French. The standard product was what was called partially-edited MT; only about 20% of a given text was edited, and what was to be edited was now determined by a specially created software program, known as EDITSYS.

The program itself is a module that allows us to go in and test at the bit/byte level the final analysis area of sentences. Virtually all of the linguistic macros in the system can be used for testing. When a given test condition is met the program generates a full-width line of a certain character in front of the condition and this line is interspersed in the text and displayed on the screen. As an editor scrolls through the translated text he halts whenever a flag line appears and makes an editing decision. If no editing is required he erases the flag line with two keystrokes; otherwise he corrects the error.

Post-editing is limited to the immediate environment around the flag. A skilled editor can edit 15-20 pages of Russian-English translation per hour using this technique.

Flags are generated by EDITSYS to check the following situations:

NOT FOUND WORDS – All legitimate not-found words or words incorrectly input are flagged. True not-found words are now relatively rare, since the dictionaries contain 200,000 individual entries.

ACRONYMS – All acronyms are checked to see if their expansions are correct. Thousands of acronyms are expanded in the dictionaries, but those of three characters or less require close scrutiny.

REARRANGEMENT – Byte 144 indicating rearrangement is flagged. Approximately 20% of sentences from Russian are rearranged with an accuracy rate of 90%. One sentence out of ten must be edited when words or phrases are moved into incorrect slots.

CONTIGUOUS SLASHED ENTRIES – There are several thousand slashed entries in the Russian-English system and when slashed words in English occur next to each other, smooth reading of the text is impeded. The most frequent occurences are adjective + adjective, adjective + noun, and noun + noun.

SPURIOUS "GOOD" TERMS – These are words that have been typed in or scanned in incorrectly but which match up against the dictionary. Examples are BOLE instead of BOLEE, S04 instead of SL04, and BIT6 instead of BYT6.

UNCERTAINTY CODE – Byte 57,04 is tested. This uncertainty code is turned on in certain homograph routines at the point where the logic becomes tenuous, there is no statistical evidence for one dictionary default over another, and in fact resolution is a toss-up.

PROBLEM WORDS – There is a flag generated for certain problem words (about 40 in number) which the system has not been able to resolve with sufficient accuracy. This category is fluid; as routines or expressions are developed for these words they are no longer flagged. Of course, new conditions or words also arise which require flagging.

The air force's next step will be the development of interactive raw machine translation for researchers, whereby they run translation procedures at their own terminals without the mediation of any translators. The Russian system was projected to be on-line in March, the French and German systems by the midsummer of 1986. All that the researcher or his secretary would have to do is to type in the material to be translated on a terminal, make some menu selections, and the raw MT would be displayed on his screen in a matter of seconds. It is expected that this will replace oral translation screening, serve to pinpoint items to be translated, and also encourage researchers to get more involved in MT.

Two speakers, **Eriko Akazawa** of the Systran Corporation of Japan, and **Mr S. Trabulsi**, who is working on the development of an Arabic system for Gachot SA, France, described some of the lexical and syntactical problems that have to be tackled in translating Japanese or Arabic. Mr Trabulsi put in a strong plea for developments to be made with a view to maintaining the essential unity of the system.

En effet, nous sommes, le groupe Gachot avec la Systran Corporation au Japon, responsables aujourd'hui de la commercialisation du système dans le monde entier hors la Commission Européenne. Et nous sommes donc les premiers intéressés par l'unicité du système. Car si le développement du Systran a été long et coûteux, imaginez ce que serait le développement de deux Systrans.

Dr Peter Walker dealt in some detail with the environment of translation in the European Communities. He had found that in practice a user had to count on five days delay between receipt of a one- or two-page document and the availability of a finished copy in three languages, and approximately four weeks delay for a 25-page document, unless special arrangements were made in advance. Rapid post-edited MT avoided some of these delays, particularly on longer documents when work could be done some four times more quickly than with conventional translation. If the translator/revisor was available and prepared to work directly on to the visual display unit,

Dr Peter Walker

then a clean-typed copy of the post-edited translation of 25 pages can be completed to good standards of accuracy and style in a day and a half.

Dr Walker suggested that as many errors in MT arose from the dubious nature of the input, the benefits of pre-editing ought to be further explored.

The pros and cons of topical glossary coding in Systran led to an energetic debate in the discussion session chaired by **Professor Juan Sager**. Professor Sager, who is chairman of the advisory committee which for the last nine years has monitored the development of Systran in the Commission, also summed up the results of the conference. He said that the step by step approach of introducing Systran had done the general cause of MT more good than harm. Systran was being used, if not by as many people as some would like or believe the system capable of, certainly in more different ways than anyone imagined ten years ago.

He described the European Communities and the Commission in particular as a "powerhouse of multilingual activity".

In a badly run, hermetic, inward-looking service this filter of translation and to a lesser extent interpreting distorts the outflowing information and we have had warnings of Eurospeak, Euroenglish etc. In a well-run service, which I know these Institutions to be, there is a lively awareness of the need for translation services to have their antennae tuned to all manifestations of language so that the texts produced in the Institutions – of which some 90% can be said to be the result of translation – are all as fresh and genuine as any texts of comparable pragmatic impact written in the language of the member states.

After attacking the attempt to give figures for the speed of machine translation, or percentages of accuracy, Professor Sager went on to suggest a principle which should govern all natural language databases and processing programs, which he proposed could be called the *Principle of Improvement through Monitored Use*. He gave as a definition, "As natural languages constantly undergo changes, databases and programs must be considered to be dynamic, requiring constant adaptation to changing usage". It was therefore the users, including translators, who collectively determined the input, content and output of the natural language database and processing systems. The affirmation of this principle by various speakers had been the first major result of the conference.

"As tools, machine translation systems must adapt to user demand; as tools to be used on a changing linguistic substance, systems must be continuously monitored for their accurate reflection of current usage."

Another major subject had been the problems associated with fitting machine translation systems into the "paperless office", which only confirmed the view that translation is simply a special case of information and documentation studies.

"Automation provides the stimulus of integrating translation into a general communication theory and only in this wider environment can we fully benefit from work done elsewhere and connect machine translation into the mainstream of the information market."

A third major theme had been the diversity of Systran, in applications and language coverage.

Professor Sager looked at the advantages and possible drawbacks of moves to unification of software, procedures and development. He concluded by looking at what machine translation could offer compared with human translation.

The greatest merit of machine translation is that it has spearheaded the industrial revolution in translation, which until recently was entirely based on individual human production. Translation was "hand-crafted" with all that this entails in quality, quantity and cost and slowness of production.

Optical character recognition and other data capturing devices, rapid and cheap printing, are all peripheral to translation itself and are used independently of machine translation to speed up the process and to reduce costs. Combined with various levels of dictionary look-up, interactive or pre-translation, they become integral constituents of a translator's work station.

Machine translation offers for the first time an alternative to purely human manufacture and thereby a choice to the customer, varying from the cheap and nasty of the dime store to the Tiffany and Cartier of translation. As we are talking about a transformation process the cost of this work should stand in some relation to the value of the original product and future use of the transformed product. This was not posible with a purely human service. There was always a basic cost and as the quality could not be consistently controlled, the results could not be related.

Now we can produce translations by diverse means in greater quantity, at greater speed, as well as at lower cost, so that the customer has a genume choice. Customer education in the use of the new product and the choices of products and methods available, let alone the training of machine operators are major problems in a constantly evolving industry and market.