Japan now leads in new MT systems, says Rolling

Conference learns of latest developments in France

LOLL ROLLING, head of the Multilingual Action Programme for the Commission of the European Communities in Luxembourg, told the eighth annual international Translating and the Computer conference that as far as new machine translation (MT) systems was concerned, Japan now appears to be well in the lead.

"Whereas developments in Europe and the United States appear once again to have suffered from managerial and administrative problems, virtually all the top Japanese computer manufacturers now have ambitious MT development projects", he said.

MT survey
Mr Rolling was surveying developments in machine translation since last year's conference. The subject had been reported more fully than ever, and its reputation as a respectable field of science had been restored by the publication of John Hutchins's book *Machine translation - past, present, future.*

Systran had continued to improve as a result of the increasing amount of feedback received from an ever wider number of users. The Commission of the European Communities had recently concluded user agreements with a number of international organisations, and Gachot, which had acquired many of the world Systran rights, had coupled it to the French Minitel network with its two million subscribers.

Logos had continued to thrive in Germany and Switzerland with its German-English and English-German systems, and had already had limited success with English-French in Canada. With increased modularity of the source and target language components, development of German-French and English-Spanish systems was now in progress. The German-French development had received substantial support from the Walloon administration in Belgium and could thus be expected to advance fairly rapidly.

Smart, another American system, was developing a number of new language combinations and work had begun on French-English and Spanish-English. Work had already begun on developing two "dialects" of Portuguese as target languages from English for Portugal and Brazil, and there were future plans for systems from English into Greek and Turkish. The basic software had also been improved, in particular by introducing a facility to scan to the previous and subsequent sentences when undertaking source language analysis.

Weidner had continued to progress in Japan, where literally thousands of Japanese-English packages had been sold, and had opened up bureau services in Canada and the United Kingdom as well as in the United States. Weidner, he said, had also become the first MT system to become involved in Scandinavian languages with the development of an English-Norwegian system in collaboration with the University of Bergen. Mr Rolling recalled ALPS's success in concluding a major contract with IBM.

Compatibility
On a more general front, Mr Rolling went on, 1986 had seen increased concern with compatibility by practically all system suppliers. As, however, there were still no fully acceptable standards for telecommunications between European languages, the problem of compatibility would probably get worse before it got better, particularly as most new PC word processing packages contain ad hoc character sets.

Returning to the question of managerial problems, Mr Rolling commented: "I can say that many problems are due to the fact that managers, who are well aware of the cost of initial investment, generally largely underestimate the cost of after-sales service and the complexity of organising feedback and updating. Those who buy and sell systems for millions of dollars or pounds take tremendous risks: to lose their millions, and to damage the reputation of MT".

Left to right - Professor Juan Sager, Lana Castellano, Brigitte Linschoft-Stiller and Loll Rolling.

GETA systems
A description of MT systems developed using the methodology and software of the French system, GETA, was given by PROFESSOR CHRISTIAN BOITET of the University of Grenoble.

He defined four phases in the translation process: *1* acquisition of the document and terminological preparation; *2* rough translation; *3* revision, if any, sometimes done in several passes (for technical documents, a technical revision by a - possibly monolingual - specialist in the field was often required); *4* output of the final document.

Professor Boitet then set out the main points of GETA's linguistic methodology, and mentioned differences with other systems.

The first point was that *translation units are longer than sentences*, and typically two or three paragraphs long. All other real-size translation systems, he said, translate sentence by sentence.

The systems rely on the *transfer approach*. Analysis and generation (synthesis) phases are strictly monolingual, while transfer (lexical and structural) is bilingual.

Also, the systems *use only linguistic knowledge*, and do not as yet make any use of the explicit representation of discourse separated from the linguistic knowledge. In other word, they behave like compilers of artificial languages, which translate programs without being able to recognise the functions computed by these programs.

Multilevel interface structures or decorated tree structures, an idea originated by the late Professor Bernard Vauquois, who played a major part in the GETA's development, represent units of translation at various levels of linguistic interpretation, ranging from lexical properties to semantic and logical relations. *Lexical units* are used to represent whole derivational families, thus allowing easy inter-class paraphrasing in generation. "To the best of our knowledge", said Professor Boitet, "no other group or firm uses yet this organisation of the lexicon". *Heuristics* are used in linguistic programming, as well as declarative/combinatorial techniques. *Structural Correspondence Static Grammars* had been introduced by Professor Vauquois in 1983 to specify and document the dynamic grammars written in the various *Specialised Languages for Linguistic Programming (SLLPs)*.

In software linguists and lexicographers use an interactive and integrated programming environment hiding all ancillary tasks. The SLLPs were used for the linguistic programming, based on production systems and incorporating very high-level data and control structures (decorated trees, recursion, parallelism, non-determinism, heuristic functions), with built-in checks of possible sources of undecidability (infinite loops). The generalised use of *transducers* rather than *analysers* led to the possibility of implementing *fail-soft techniques*.

Professor Boitet then went on to describe the systems which had been developed using this methodology, the Russian-French system, which was a real-size operational prototype, and a German-French system done as a feasibility study. There were also two prototypes systems based on Ariane-78, a language- and theory-independent software environment for building multilingual MT systems, for English-Malay and English-Thai.

National project
The French Machine Aided Translation National Project, started in November 1983, was now nearing the end of its third phase, scheduled for February 1987. Financing of the project had come 50% from public and 50% from private sources. For the first development it has been decided to build a French-English system tailored to aviation manuals of the kind produced by Sonovision, one of the private sector companies involved in developing the system (see story in *Language Monthly* for February 1986).

The core of the architecture of the lingware and the software for the project, said Professor Boitet, comes from previous work done at GETA, but new tools and techniques had been added. These included the connection of a translator work station, and the use of what were termed *structural correspondence static grammars (SCSGs)* to describe the correspondence between the strings of a natural language and the corresponding interface structures.

"Special care has been taken to describe a reasonable core grammar and to study in detail the particularities of the typology at hand. As any sublanguage, it offers grammatical constructions which would be judged ungrammatical in other contexts. These SCSGs have then been used as reference and documentation while writing the very large dynamic grammars."

A lexical database had been developed with a base for each language and a base for each transfer pair. The information attached to the terms is one in terms of "static properties", which means that it is the same for analysis as for generation. All MT dictionaries are now generated from the database. A division is made between general and terminological terms, for which different "indexing forms" have been prepared. Terminology is simpler. The questions asked to the indexers still require some linguistic training, but less than with the previous method.

CALLIOPE lexicons
Professor Boitet gave further details on the work done to create grammars and dictionaries for CALLIOPE-AERO, the project for translating aeronautical manuals from French into English, and for CALLIOPE-INFO, for translating computer manuals from English into French.

"The size of grammars and dictionaries is obviously heavily dependent on the application at hand," he said. In the case of CALLIOPE-AERO, the typology of the manuals included all normal syntactic constructions, with the exception of interrogative clauses, relative clauses introduced by *Don't*, and imperative forms of verbs (replaced by the infinitive form) and a lot of special phenomena.

A preliminary study of the corpus for the lexicon had led to an estimate that 6,000 general terms and 15,000 terminological terms would be necessary for the system to be usable. The first part was almost complete, while the second may just be complete at the end of the project.

The dictionaries comprise now around 8,000 lexical units in the running system, with more in the lexical database, or about 12,000 terms, in both languages. As far as the grammars were concerned, there were about 175 rules for morphological analysis, 600 for structural analysis, 90 for structural transfer, 200 for syntactic generation and 20 for morphological generation.

"If we compare this," said Professor Boitet, "with the size of a compiler for some programming language, written in metalanguages such as LEX and TACC, we see that the lingware engineering effort required to create and maintain such an MT system exceeds by far what is required for a compiler. This is made even worse by the fact that natural language is not fixed by decree, but changes, and is not defined by our grammars, but only approximated. Contrary to the case of a compiler, the grammars and dictionaries of an MT system must be easily modifiable, by linguists and not by computer scientists. Hence modularity in the SLLPs and conviviality of the programming environment are essential."

In the CALLIOPE-INFO system, ambiguity "boards" (*planches*),(or two-dimensional representations of rules in an SCSG) are being constructed for English, as they have been for French. They are useful for analysis, where they help design the disambiguation (dynamic) rules. The dynamic grammars for the analysis of English and the generation of French were offsprings of those developed by GETA, in-house or in collaborative work. Indexing of the terminology was again being done by Sonovision, and the aim was to attain 6,000 specialised terms for the first version, expected around mid-1987.

Japanese projects

The considerable efforts being made in Japan to further machine translation were described by PETER WHITELOCK, who until recently was project leader of the English-Japanese MT system at the Centre for Computational Linguistics, University of Manchester Institute of Science and Technology (UMIST). He revealed that the overall translation market in Japan had been estimated at around one trillion yen, and that spending on machine translation was around 35 to 40 million yen, estimated to rise to a figure equivalent to one million pounds sterling by 1990. "The interest and investment made in Japan is probably greater than that of the rest of the world put together", he said.

After listing the various systems now available for Japanese-English and English-Japanese, their prices and their dictionary size, Mr Whitelock described the National Dictionary Project, a nine-year programme aiming to create a basic dictionary of some 200,000 words. He mentioned the Advanced Telecommunications Research project for automated telephony (reported on in the November 1985 issue of *Language Monthly*). He did not know whether the Japanese would have "a marketable product at the end", but the project was certainly ambitious, incorporating as it did voice recognition technology. A German-Japanese project was being conducted at the University of Stuttgart.

Muriel Vasconcellos (left) with Jean Datta

In mentioning a project at UMIST to enable monolingual English people to enter Kanji characters, Mr Whitelock mentioned the possibility of some machine translation work being processed by what he called *paratranslators*, "non linguistically naive" persons who could learn to handle the work without necessarily having a knowledge of the source language. Inevitably, the term *paratranslator*, and the thought of non-linguists taking over some of the translation work, reverberated in the subsequent discussion.

Post-editing

Techniques of post-editing machine translation were detailed by DR MURIEL VASCONCELLOS, head of the SPANAM machine translation project at the Pan American Health Organisation in Washington. Translators working on the project, she said, were expected to produce some 4,000 words a day of translation, and contribute to the maintenance of the English-Spanish machine translation dictionaries. They must work on screen, which meant that on joining they needed to learn word processing techniques, some of which she described. The work on machine translation at the University of Saarbrucken, which had led to the development of the SUSY MT system, was described by KARL-HEINZ FREIGANG. He also described the work there on an automated translator's workstation, and the incorporation of computational linguistics, machine translation and machine-aided translation techniques and experiences in the teaching of those following translation and interpreting courses.

JEAN DATTA, of the United Nations Industrial Development Organisation in Vienna, gave a cogent description of how a large organisation ought to prepare for the introduction of machine translation. It should be done over a period of years, as there were problems of changing existing habits. Controlling input language was of major importance, and no opportunity should be lost of bringing influence to bear on this matter. The input had to be improved, at the least in circumscribed factual areas. Similarly output could be streamlined, from an analysis of applications. There could be a place for what she called "no-frills" output, i.e. unrevised MT output. Time and guidance were needed to develop MT post-editing skills, therefore it was important not to put "MT out in front" in introducing computerisation. The whole approach, she said, was the "gradual layered introduction of the new technology"; a gnat could be swallowed whole but an elephant needed to be tackled a slice at a time. She showed charts showing the introduction of the various phases.

Translation practice
The findings of the survey into translation practice in Europe conducted by the Digital Equipment Corporation in conjunction with *Language Monthly* were presented by DAVID SMITH, Digital's Translation Programme Manager. The report, which was also produced as a 36-page printed booklet in time for the conference, is based on an analysis of 253 completed questionnaires received from 18 countries, the main ones being the United Kingdom, Netherlands, France, Federal Republic of Germany, Italy and Belgium, and dealt with such matters as length of experience, tasks performed, languages handled, percentage of time spent on various tasks, equipment used and degree of satisfaction, and attitudes to computer usage and new technology. A more detailed examination of the report will be given in a future issue of *Language Monthly*.

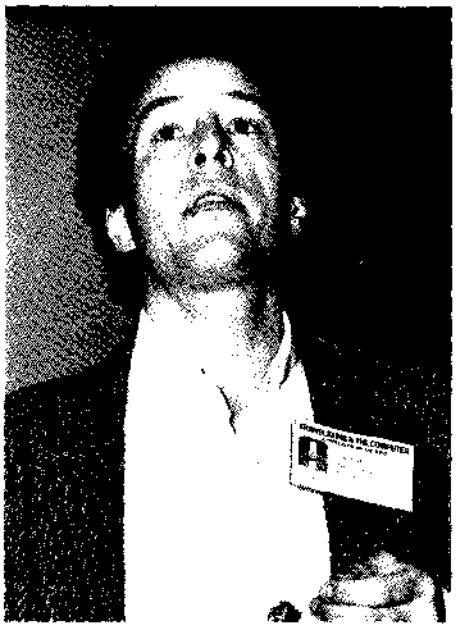
Anyone interested in obtaining a copy of the report should apply to Linda Hempel, Translation Programme, International Engineering, Digital Equipment Corporation, Reading, England.

PETER FENWICK, an information technology consultant with a particular interest in character sets, demonstrated the new version of the extended ASCII character set shortly to be issued, and showed examples of other character sets used in communication, including the Japanese and Chinese sets.

PATRICK CHAFFEY, of the University of Oslo, described the production of the ADNOM glossary of standardised English terms for Norwegian institutions and occupational designations, and gave some indications as to how similar procedures could be used in other countries. PETER ARTHERN, head of the English translation division of the Council of the European Communities, Brussels, spoke about the theory and practice of revision. He mentioned how little seemed to have been written on this subject, and described a method he had evolved for evaluating the work of revisers in his division. This could be visually represented on a chart, and reduced to a mathematical formula. CATRIONA PICKEN, vice chairman of the Intitute of Translation and Interpreting'(ITI), described the development of the Institute.

Continuing training
A major survey of the needs for continuing training for the language professions had been conducted earlier in 1986 by Bradford University with the assistance of *Language Monthly* and of the ITI, and the results were presented to the conference by TONY HARTLEY. It was interesting to note the convergence of some of his findings with those of the Digital survey into translation practice, noted above.

There had been over 600 replies from the UK and Europe, with UK



Tony Hartley

translators representing approximately 55 per cent of the total. The use of word processors now appeared to be widespread, with two out of three responses indicating they used one. "UK national and local government organisations would seem," he said amid laughter, "however, to be a notable exception to this general rule".

"For the great majority of users the computer is made to function exclusively as an enhanced typewriter; the facilities it offers for running other translation-oriented packages or for data communications appear under-exploited. For example, whereas only one in ten UK respondents routinely manages terminology in a computerised database, the figure is twice as high for continental Europe as a whole, and reaches one in three for private sector translators in Germany. Similarly, only one UK respondent in 25 regularly accesses commercial databases, while one in seven of their European counterparts does so." However, such findings might be attributable in part to the preponderance of freelances in the UK sample and of in-house translators in the European set.

Given the already extensive usage of word processors, said Mr Hartley, the very high demand indicated for further familiarisation seemed somewhat surprising. He speculated that word processing might be a fairly new venture for many translators and they were working with relatively unsophisticated and inflexible systems. There was undoubtedly a call for introductory courses in word

David Smith and David Tyldesley of the Digital Equipment Corporation

processing, but also there was an expressed need for non-partisan advice on the comparative merits of the many systems available. Virtually all employers of translators voiced a desire to learn more about machine assisted translation, about 50% in the UK% and 60% in the rest of Europe, with particular interest shown in France and the Federal Republic of Germany. The desire for post-experience training rose to about 65% and 70% for computerised terminology management and on-line data base interrogation respectively, with the same high level of interest in the principles and practice of terminology work.

"The survey invites the conclusion that a fundamental competence in information retrieval which in itself has little to do with high technology is under-developed. It is the sudden proliferation of available information which has exposed in the training of many translators a weakness which they have recognised and wish to remedy".

Appropriately, in view of these remarks, another speaker at the conference, PAUL BURTON, an information scientist, was to give a paper on principles of information retrieval for personal information systems for translators.

On machine aids to translation, Mr Hartley said that the survey showed that even young translators with a specialist qualification are as numerous as older translators in wishing for in-service training. "The implication is clearly that the establishments at which they gained their qualifications are failing to move with the times".

Desire for more subject knowledge had emerged strongly, even though it had not been mentioned specifically in the questionnaire. There was a demand for familiarisation with the conventions of technical writing in English and of differences between British and United States English.

Among his conclusions Tony Hartley found that translators, though busy, were prepared to attend training sessions for three or more days at a time, and were anxious to meet other translators in their fields. There was an opportunity, and a need, for universities to put on integrative courses.

The Translating and the Computer conferences are organised by Aslib, the Association for Information Management, in association with the Institute of Translation and Interpreting, and are held in London every November.

View of the audience during the opening speeches

Jo Mendez, of Brussels (left) with Stephen Collins (Stockholm)

A view of the crowded exhibition hall