[From: *Lebende Sprachen* 24, 1979, pp.103-107]

Professor JUAN C. SAGER
Head, Department of European Studies and Modern Languages
The University of Manchester Institute of Science and Technology

The Computer and Multilingualism at the European Commission

(Reflections on the CEC's Plan of Action for the Improvement of Transfer of Information between European Languages)

The political reality of the European Communities requires a multilingual regime. This is, on the one hand, necessitated by the powers of the institutions which extend into each member country and affect even individuals in certain aspects of their lives: on the other hand, the absolute equality of all member states and the possibility of individuals to be in direct contact with the Community institutions requires parity of status of all official languages. Finally, the provision of the Treaty of Rome which establishes the novel legal principle of a single *original* document written in different languages is at the same time, though probably unwittingly, a defence and a recognition of the singular cultural value represented by the linguistic diversity of Europe. It may have been the historical accident of an initially small community of only four languages (i. e. 12 language pairs) that permitted the adoption of this principle; with a larger number of languages the idea might have been too daunting to contemplate. As it was, the Commission took in its stride the accession of Denmark, Ireland and Great Britain (an addition of 18 language pairs) even though this increased its linguistic staff to a third of the total staff complement and linguistic and linguistic services now absorb approximately half the budget of the

European Parliament. The addition of Spanish, Portuguese and Greek, which would increase the language pairs in translation to 72, does, however, give cause for reflection. The cost of a multilingual regime has therefore been subjected to a serious re-appraisal.

When in 1976 the Commission once again took stock of the cost of its linguistic services, it discovered that it was faced with a demand for translation rising at some 10% per year in what is probably the largest translation service in the world. Conscious of the possibilities of computer assistance in information handling, it consequently asked the European Parliament to authorise a three-year Plan of Action for the improvement of the transfer of information between European languages with a budget of some three million units of account. The specific areas in which development was thought desirable were:

— automatic pre-translation of unprocessed texts drafted in natural language,
— automatic translation of texts drafted in limited syntax,
— terminology banks,
— multilingual thesauri,
— technical infrastructure,
— assessment of applied research,
— encouragement of multilingualism.

The emphasis of the Plan clearly is on cost reduction but without any concession to lower quality of translation.

The plan was approved in December 1976 and the Directorate General XIII (Scientific and Technical Information and Information Management) was entrusted with its implementation. The same directorate is responsible for the development of EURONET, a network of various documentation data bases linked by the post-offices of the member states which will provide access (preferably multilingual) to scientific and technical information.

The first actions within the plan were to commission a state of the art report on multilingual systems, an evaluation of existing documentation thesauri for their multilingual potential and the development of an English-French translation version of SYSTRAN the machine translation system invented by Dr. Toma and his World Translation Centre in California. At the same time a contract was drawn up for the development of an improved version of TITUS the machine translation system for abstracts. This had been developed for the Institut Textile de France but had unfortunately to be abandoned because of insufficient progress within the time stipulated in the original contract.

The first public activity of the Plan of Action was the organisation of a four day Congress on Multilingual Systems and Networks under the title 'Overcoming the Language Barrier', held in Luxembourg in May 1977. This impressive display of applied linguistics research and development which brought together some 700 participants to listen to a great number of papers on Teaching and the Use of Languages in the European Community, Multilingual Terminology, Multilingual Thesauri and Documentation Systems, and Human and Machine-Aided Translation. The purpose of the Congress was partly to establish links between the practical work in the Commission and relevant research and development, and partly to give those responsible for the Action Plan a clearer picture of existing and developing systems and methods so that this plan could be based on the best knowledge available.

The Action Plan made provision for an advisory committee to assist the Commission with planning and executing the programme. This committee (CETIL)* consists of experts from the member states of the Community and from the major Community institutions with the following mandate:

CETIL

— constitutes a forum for the exchange of information on the situation in the Member States and on Community level: knowledge of language in various branches of activity and levels of qualification, language teaching policies, translation activities, language policies with regard to scientific and technical publication, ongoing or planned research activities:
— supervises the Commission's action programme by evaluating priorities and analysing results;
— makes recommendations concerning the orientation of research and development in the field of multilingualism.

The first meeting of CETIL took place in September 1977 and it has been meeting at three monthly intervals since. These regular meetings are concerned with receiving progress reports on the various projects sponsored by the plan, to discuss new initiatives and to evaluate work carried out under contract by outside consultants. The Committee also sponsors a number of workshops and other meetings dedicated to particular topics and to which a wider range of specialists is invited.

In its first 18 months of existence the Committee has been mainly concerned with three areas of activities which variously affect the translation process, text-processing, terminology banks and machine-aided translation, all of which involve computerised methods of information transfer. These are discussed here not in chronological order but in the sequence of increased machine involvement in assisting the human translator.

Text-Processing

The cost of translation does not only lie in the salary of the highly skilled and specialised professional translator but a high proportion of the cost arises from the typing of drafts and final versions and the additional administrative circuit a document has to pass through from an originating or requesting department to the translation department and back again. The larger the organisation the more complex is such a circuit, and any effort to rationalise translation cost must concern itself with this range of problems which are in a way quite fortuitous and incidental to translation. The complexity of the process is also determined by the status of the original and the translation. The original may only be a draft and lose any significance after translation or it may co-exist as a document of equal status with the translation, which is the case for the majority of Community documents which by law must be translated into the official Community languages and enjoy equal status.

Automatic word-processing and text editing are by now well known office and publication techniques, and it is recognised that they can produce considerable savings in staff and materials cost and administrative time. Their monolingual use is well established and even in multilingual enterprises there is considerable experience of their use in such organisations as Siemens AG and the Bundessprachenamt of the Federal Republic of Germany. The constant innovation in this field, especially since the advent of microprocessors and the resultant diversity of systems and facilities on offer does however require careful study before adopting a particular procedure as there is as yet little if any compatibility between the systems on offer. The advantages of text-processing for translation are considerable especially where multiple translation and multilingual publication is required. Once stored in the memory a text can be recalled in parts or in its entirety for the various purposes of translation and revision, selective publication or re-use as a draft for other work. As revision seldom affects more than a certain proportion of a text, those sections which are unaffected by revision

can be re-used and left in the store and only those words or sentences which are deleted, replaced, or transposed have then to be processed. Equally an originating department at present often retypes a translation it receives from a translation department to fit it into an existing document, to make it conform to a certain format or simply to provide it with the appropriate letterhead. All these operations can be performed by a machine at a fraction of the typing cost. Joint output by several translators working on one longer document can also be assembled by machine into a unit without typographic seams and text units repeated in many translations, such as common rubrics in contracts and other legal documents can be stored and inserted automatically at the appropriate place.

Such machines can also be operated directly by translators or revisors who can thus test stylistically complex translations by putting alternative versions directly onto a display screen for comparison. In fact, a whole new range of translation and revision techniques is now available which are likely to produce new attitudes and working habits.

Two studies produced as part of the Action Plan provide some interesting data on the cost efficiency of these machines. It was shown that the complex administrative procedures associated with logging documents in their circuit through several departments can be simplified significantly and speeded up if all documents sent for translation are initially in machine readable form. The retyping effort of revised typescripts can almost be halved as only 55 % of all lines of text are affected by revision at a rate of less than two corrections per line. For an organisation processing several hundred thousand pages of translations annually considerable savings can thus be made. The time saved in typing and document handling is given as 39 minutes per document or 40 % of total time for a translation into one language, and a correspondingly higher saving (43%) for translation into three languages. The saving of typing cost and time is even more significant if account is taken of the facts that the multilingual typing pools in Brussels and Luxembourg employ non-native speakers who may experience difficulties with audio-typing of dictated translations and that the diversity of subject matter does not allow a typist to specialise.

Even though there are persuasive arguments in favour of introducing text-processing, the actual implementation of such a decision for as large an organisation as the European Community institutions requires considerable time and study, as it would not only affect the translation process but every other aspect of work. Besides introducing completely new filing systems and document circuits, staff at all levels would have to be retrained and considerable initial investment would be required at a time when constant innovation may make a system obsolete by the time it is in full use. It is therefore not surprising that the Commission is still considering the whole question of text-processing and that no immediate decision can be expected. It is hoped however that a number of pilot operations can be started in order to provide information useful to a full implementation.

Terminology Banks

The Commission is one of the pioneers in computer dictionaries and its EURODICAUTOM system is well known. The benefits of storing the vocabularies of special languages in automatic memories for regular updating and diverse output are widely recognised as can be seen by the ever increasing number of terminological data banks all over the world and especially in Europe. The Action Plan is, therefore, not concerned with the introduction of such a tool but with its expansion and use.

The present volume of the dictionary is approximately 150,000 term units. In theory these should be available in all six Community languages thus giving a total of nearly a million terms. In practice however the present total is of the order of half a million terms and the languages are very unevenly represented; of 100 terms units 98 % have a French equivalent, 92% English, 43% German, 32% Italian, 25% Dutch and 16% Danish equivalents. Great efforts have been made recently to achieve a greater equality of coverage. During 1978 the growth for English and French was virtually nil whereas entries for German increased by 10%, for Italian and Dutch by 25% and for Danish by 50% respectively. This imbalance is being corrected vigorously but it also reflects the situation of the availability of reliable terminology. The Action Plan has permitted the purchase of some 150,000 new terms from various sources but staff resources at present only permit the processing and incorporation of some 1,000 terms a month and it is not easy to recruit qualified terminologists.

A fair amount of staff time has also recently been devoted to improvements in the system and the depuration of terminological holdings, which is, of course, part of the regular maintenance required for any large term bank. During 1978 almost as many terms have been cancelled as have been added to the collection. This work consists of eliminating redundant entries and merging items which relate to the same concept but are classified or presented differently in the various subject or language subdivisions of the bank.

The term bank is as yet too small to provide a satisfactory coverage of all subjects and languages relevant to the translation departments. A massive annual input of some 50,000 term units, corresponding to 300,000 terms in all six languages, is therefore planned for the next three years, after which a normal rate of input and updating can be resumed. It has also recently been decided to make EURODICAUTOM available to EURONET users and this is another justification for a speedy expansion as a high rate of negative replies would soon discourage any potential users. On the other hand the benefits of a supranational agreed terminology are of considerable advantage to encourage and improve multilingual communications among governments and other organisations inside the Community.

In November 1978 the Commission organised a workshop on EURODICAUTOM in order to receive advice on its development and use. The experts invited commented favourably on the present state of development and recommended rapid expansion. It was felt that the system was sufficiently developed to be made available to all translators in the Commission. Widespread use, it was thought, was the best means of ensuring that the format of the data and the types of access would conform to real user requirements rather than an idealised user imagined by the designers. The main emphasis of the system at the moment is on on-line use. Experience with other data-banks suggest that batch processing of enquiries within a few hours and computer output on microfiches are equally important and that hard copy print-out in particular provides useful feedback from users both for expansion and improvement of the data.

In contrast to most other term banks which concentrate on term pairs and their definition EURODICAUTOM is organised on the principle of terms in context. This approach is justified by its designers as being appropriate both to the polysemous nature of most of the terms required for Commission translations belonging, as they do, mainly to the social sciences as well as to the relatively low degree of subject specialisation of most Commission translators which arises from the considerable diversity of subject matter being translated. It is undoubtedly true that a terminology bank should be designed to reflect the nature of its holdings and

the user needs it serves. Nevertheless, the experts at the workshop considered that the extremely comprehensive and complex information at present provided in response to a simple enquiry for a translation equivalent should be broken down in such a way that a translator can, possibly in the first instance, obtain abbreviated results, e. g. a single term pair with subject, source and quality codes, and context, definitions etc. only as a secondary operation.

Now that EURODICAUTOM is entering into full use a number of these suggestions based on practical experience with operational term banks are likely to be examined more closely or may indeed be brought up by users themselves in the new management committee which will also include translators.

Machine-aided Translation

After the initial euphoria and the subsequent disappointment of a decade ago the subject of automatic translation is discussed with a great deal of caution. People are now more modest in their claims and expectations for involving machines directly in the translation process. Two major lessons have been learnt: Documents requiring translation are so diverse in nature that no one system is ever likely to be suitable for all manner of texts; this opens the way for the concurrent development of several systems with different types of objective. In addition, fully automated systems are rightly considered utopian; all systems now being developed with the aim of producing equivalents to regular human translation consider human intervention before, during or after machine translation as an essential part of the process.

Raw, unedited machine output does not as yet resemble human translation but can, nevertheless, satisfy certain information needs which are at present only catered for by much more costly human translation or not at all. It can provide enough information about a document to give the reader a general idea of its content and to allow him to decide whether the document is relevant for his purpose and whether the whole or any particular section requires a full human translation. This is a function we generally associate with information retrieval procedures rather than translation and as a service it is at present not available in Europe. Simple system can be specifically designed for this purpose; this service can also be provided by the raw output of more sophisticated systems which involve post-editing for full translation. There is a clear need for such a service especially for the lesser known languages like Chinese or Arabic, which are incidentally also the languages from which translations are more costly and less readily available. A Russian-English version of Systran is at present being used by the US Air Force for this purpose and the possibilities of such a use are at present being explored by the Commission for the three Systran language pairs it has acquired.

Machine translation of pre-edited texts, i.e. texts written in a limited syntax and vocabulary is a goal which is not being very widely pursued, probably because its application is limited to controlled writing situations. It has, however, considerable potential for individualised systems of translations of manuals, abstracts, minutes, weather reports, i.e. all situations in which professional writers can be asked to write originals specifically for machine processing. The TITUS II system developed for the Institut Textile de France translates abstracts simultaneously into several languages. The Commission, as stated earlier, was prepared to assist in the development of an improved version of TITUS but had to abandon its sponsorship because the developers encountered some difficulties which prevented them from fulfilling their contract. It is reported that TITUS III is now nearing completion and it is hoped that the progress is such that a use beyond the translation of abstracts in the field of textile technology can be considered. There is no doubt that major savings in translation cost can be achieved if certain multilingual texts are written in a limited syntax. Routine information like weather and other staple reports, tenders for contracts, instructions, manuals, contracts and even some minutes and memos are not less readable for being so concise and unambiguous that a machine can translate them into any number of languages. In fact, much monolingual technical writing is already heavily restricted for the sake of greater clarity. In public administration such a step requires a substantial change of attitude to writing which will probably come about only under the extreme pressure of cost-efficiency arguments. The more widely known and developed forms of machine translation are based on the principle of virtually un-edited input but require considerable post-editing of the output. The argument about this type of machine translation is no longer whether it can be done — this has been proved — but whether the post-editing effort is tolerable — tolerance being expressed not only in terms of the time involved in converting machine output into a form acceptable by a reader as equivalent to human translation, but also in terms of the psychological strain upon a highly skilled professional when confronted with a constant flow of errors which will lead to frustration or negligence in his work.

At its first meeting, the Advisory committee received the results of the first evaluation of the English-French version of SYSTRAN which the Commission had purchased in 1976.. A summary of the results is published in the Proceedings of the May 1977 Congress. Even with a relatively high revision rate, the report argued that "the cost of creating the English-French SYSTRAN system could be fully recovered within one year, if the total workload of the CEC in this field, i.e. approx. 20 million words per year, was covered by SYSTRAN". This is, of course, a relative figure as the cost of developing a system can be borne by one or several purchasers and there are additional costs involved in the maintenance of a system and in the expansion of the dictionaries. The results of the evaluation were so encouraging that the Commission decided to pay for the development of an English-French and an English-Italian version which have recently been delivered and are being tested. It also commissioned some improvements to the English-French version and this modified version has been the subject of a second evaluation, the results of which will be widely publicised. The second evaluation showed a considerable increase in the intelligibility of the output and an admittedly small user enquiry was, on the whole, also favourable to a limited information use of the raw output. Surprisingly, however, the amount of postediting time required was higher than for the first evaluation. This was largely attributable to the degree of difficulty of the texts processed, and was confirmed by the fact that human translation, used for comparison, also proved more time consuming than for the first evaluation. This unexpected fact shows the great difficulties that exist in balancing the variables in any text evaluation. There are no absolute criteria for measuring the difficulty of a text in the same way as the various criteria which can and have been applied to evaluations such as intelligibility, fidelity and acceptability are ultimately subjective.

It is, therefore, not surprising that a social workshop organised by the Commission on text evaluation criteria did not produce a consistent methodology or a single set of reliable criteria for evaluation. One additional merit of the various projects carried out under the Action Plan is that the consistent set of evaluations being performed on the various SYSTRAN versions will significantly contribute to our understanding of what is involved in text evaluation.

The problems associated with introducing a machine-aided translation circuit were tested in a pilot-operation of SYSTRAN over two months in 1978. This experiment showed the organisational difficulties in a large organisation which have to be overcome so that a document can be processed with the speed that is one of the major attractions of machine translation.

Another major hurdle for the widespread use of machine-aided translation is the compilation of the massive dictionaries required for a wide subject coverage.

The size of the dictionaries, the slowness of the circuit and the lack of text encoding facilities are, however, not the only obstacles to a full implementation of machine-aided translation in the Commission. So far the translation departments still find the amount of post-editing required for SYSTRAN unacceptable and cannot therefore commit themselves to even a small scale use of the system. Another improvement contract which will produce results late in 1979 may alter this situation. In the meantime some necessary ancillary software will also be available and the organisational framework will be improved.

It is obvious that the complexities of introducing machine-aided translation can only be faced by a very large organisation like the Commission. The results of this work will serve as a model for other institutions and the Commission is prepared to share its experience with the governments of member states who are also given the opportunity of using any system developed by the Commission. In this way, the enormous effort involved in developing new translation techniques will benefit all the Community.

Nevertheless, it is generally recognised that SYSTRAN has a strictly limited development potential which is at present being investigated for the Action Plan by the Cambridge Language Research Unit. The considerable size of its dictionaries and its single language pair structure also make SYSTRAN rather cumbersome for multilingual use. For this reason the Action Plan is sponsoring discussions among machine translation research groups in European universities with a view to formulating specifications for a more advanced European translation system which would be developed cooperatively and financed by the Commission and member states. Consultations are well advanced and concrete proposals for this project are expected shortly. The experience gained with SYSTRAN will be of great benefit to this new project particularly with regard to user requirements and reactions. The new system is intended to be modular, so that • its various parts can also be used independently for other forms of language processing, and multilingual in the sense that substantial language specific analysis/synthesis modules are used in conjunction with a common transfer strategy.

Such an ambitious scheme will take several years to develop, but there is now a general consensus that since the ALPAC report of 1966, which recommended postponement of machine translation research, sufficient additional knowledge has been gained to produce practically useful translation systems.

Conclusion:

- The Action Plan has so far concentrated on machine assistance in three phases of the transfer process. Even in these areas considerable innovation is likely with progress in micro-computer technology and the paperless office. Multilingual communication cannot be treated as several parallel sets of monolingual situations, and improvements can only be achieved if its fundamentally different nature at the level of the participants, the texts, and the transmission is fully recognised. There are, therefore, many other areas in the multilingual communication process which may profitably be investigated in order to reduce cost and to improve efficiency.

One of these is the nature of texts; this would involve the study of existing conventions in texts and the possibility of conventionalising further certain text modules, the introduction of multilingual forms or text patterns for largely repetitive operations, the setting up of levels of quality criteria for translation so as to differentiate according to the importance of the text to be translated, and the development of quick, reduced information circuits in the form of summaries.

Another area is the development of multilingual comprehension ability among a greater number of people working with or for the Community institutions. Whilst it is preferable and more efficient for people to speak and write their own language the ability to understand a foreign language in its spoken and especially in its written form can be acquired without undue effort. The development of comprehension skills would make communication more efficient and incidentally reduce the need for some translation and interpreting. Considerable experience is available in the teaching of languages for a limited range of use and limited fields of communication, and special courses can readily be revised for particular purposes.

Most of these developments will not only benefit the European Community institutions, but EURONET and national governments will have access to the substantial expertise developed as a result of this Plan of Action and its tangible results in the form of studies and systems ready for implementation.

Notes

• Comite d'experts pour le transfer de 1'information entre langues europeennes.

[2] The author is chairman of the Advisory Committee assisting the European Commission with the Action Plan. Most of the information presented here has been obtained from unpublished documents; the opinions expressed represent the author's personal views.

*

[2] The following publications provide further information on the work of the Action Plan and the European Commission's translation services.

AFTERM (1977), *Terminologies 76, Colloque international,* Paris 1976 (contains a description of EURODICAUTOM).

Aslib (1979), *Proceedings of the ASLIB Symposium on Translating and the Computer,* North Holland. (Contains articles on EURODICAUTOM, SYSTRAN and the Action Plan in general.)

BRUDERER, H. E. (1978), *Handbuch der maschinellen und maschinenunterstützten Sprachübersetzung. Automatische ÜBERSETZUNG natürlicher Sprachen und mehrsprachige Terminologiedatenbanken.* Verlag Dokumentation, München.

COMMISSION OF THE EUROPEAN COMMUNITIES (1977), *Overcoming the Language Barrier, Third European Congress on Information Systems and Networks,* Luxembourg 1977, *2* vols., München. (Contains a summary of the 1st Systran evaluation.)

HUTCHINS, W. J. (1978), *Progress in Documentation, Machine Translation and Machine-aided Translation,* in: *Journal of Documentation,* 34, *2,* pp. 119—159. (A very good survey of machine translation.)

ROLLING, L. (1978), *The Facts about Automatic Translation, Proceedings of the FID Symposium,* Edinburgh.

SAGER, J. C. — JOHNSON, R. L. (1978). *Terminology, the State of the Art, AILA Bulletin, 22,* 1, pp. 1—12.