COMPUTER - AIDED TRANSLATION TODAY AND TOMORROW

by Loll Rolling
Commission of the European Communities
Luxembourg

presented at the
FID Congress Edinburgh
September 1978

COMPUTER - AIDED TRANSLATION TODAY AND TOMORROW

1. Characteristics and structure of existing translation systems.

It is the firm belief of the European authorities that the language barriers between E.C. Member States can only be overcome by a consistently balanced effort towards diversified language teaching and the development of cost-effective multilingual tools. Adoption of a single language for Community communication is a political, technical, and economic absurdity.

The European Community now has six official languages, and its policy is to give equal status to all six in order to maximize communication and coordination between the Member States. This means that all official documents must be translated into the five other languages, and that the Commission's translation services, which now include a total of more than 1300 linguists, had to translate 538.000 pages in 1976. In order to reduce this workload, the Commission has been developing a 6-longuage terminology data bank, which is now operational, and has decided to break new ground in the field of automatic translation.

A computer-aided translation system can be broken down, generally, into several components (or modules). Input of text can be performed by the classical keypunching method, or by magnetic encoding, or by optical character recognition equipment. If the systems operator is in charge of the input, he will incorporate pagination instructions; if, on the contrary, he receives a text already stored on magnetic tape, an interface program is required for conversion to the specific format of the translation system.

The minimum components of the actual translation system are: a source language analysis program, a lexical transfer program, and a target language generation and synthesis program. The programming language is generally either Fortran or Assembler.

In addition, bilingual dictionaries, covering the various subject fields, are required for every language pair. They should aim at covering all the words likely to occur in the texts to be translated and any multi-word expression the meaning of which differs from the combined meanings of its component words.

For languages with a large number of flexions (different endings due to conjugation of verbs and declension of nouns) the dictionary will contain only stems, but the dictionary lookup will be preceded by a morphological analysis.

A dictionary is generally produced by frequency analysis of a representative text corpus in the subject field to be covered, followed by the coding of single words and expressions, in accordance with a number of rules which are different for every system. It is hoped that a common dictionary format can be adopted in the future by the initiators of systems under development, so that dictionaries developed for one system can be used in other systems.

A study produced by Herbert Bruderer showed that 9 machine translation systems were operational somewhere or other in 1976.

Of these, three were free-text systems with dictionaries large enough for actual translation work, three had only experiment-size dictionary samples covering a number of languages, and three were of the so-called limited-syntax type. These correspond to the situation where the owner of a translation system is himself the producer of the text and can see to it that the items to be translated are composed of a limited number of terms and a limited number of well defined syntactical structures. In one case, the limited-syntax situation is produced by a combination of transliteration and pre-editing of the source text.

Of the larger free-text translation systems, all three were U.S. baaed and initially aimed at translating from Russian into English. The system now in operation at the Oak Ridge National Laboratory is still limited to this language couple; the Systran system, developed by Dr. Peter Toma, extends into English-branch, English-Russian, English-Spanish and French-English, and the Logos system, which was widely used for the translation of English into Vietnamese, now tentatively covers a number of languages including French, Spanish and German.

Systran, which has been operating for US government agencies for a number of years, recently made its breakthrough towards large-scale application, both in Canada and in the European Community.

The Commission acquired the Systran system which it developed, during 1976, for the English-French language pair, with a dictionary in the field of food science and technology. The system is now being extended to cover French-English and English-Italian, and German will be introduced as a fourth language in 1979. Now that the economic viability of Systran, complemented by human post-editing, has been demonstrated, a number of requests have been reaching the Commission for pilot operations using Systran in various environments.

This is especially meaningful as Euronet, the European information network: will be going into operation at the end of 1978. It is likely that a number of suppliers will wish to make their data bases available via Euronet in languages other than English.

While the Commission is thus probing the market situation and demonstrating the actual demand for low-cost machine-aided translation, it is also aware of a need for high-quality, fully automatic translation in response to the shortage of highly qualified human translators rather than with a view to saving money.

In order to open the way for the advent of such a system, it has initiated a number of studies aiming at the creation of an efficient infrastructure for automatic translation, including methods and equipment for low-cost, error-proof text recording and an appealing man/machine interface for on-line post-editing,

Another study is aimed at saving the efforts invested into Systran machine dictionaries by achieving compatibility between these and the dictionaries of future systems.

2. Quality evaluation

A number of methods and criteria have been proposed, throughout the recent history of machine translation, for the evaluation of the quality of the product.

A workshop was held at the European Commission on 28th February 1978, on this problem. An impressive number of papers was contributed, and the outcome of the lengthy discussion was that one should distinguish between global or macro-evaluation, with intelligibility and revision time as the main criteria, and analytic or micro-evaluation.

Correctness or fidelity might have been a more absolute criterion of quality, but it is difficult to quantify the relative importance of various types . of deficiencies, and the users of machine translation are more interested in the usability of the product.

Intel1igibility ratings by a group of independent evaluators give a clue as to the actual usability of the raw product.

Revision time might have been a good criterion but it was shown to depend heavily on the revisor's background and goodwill.

Micro-evaluation consists in determining the number of errors of various types occurring in the product: it supplies the indispensable feedback for continuous improvement of system software and dictionaries.

Revision rate, which is the percentage rate of words involved in the post-editing process, is a measure of usability as well as of cost.

These are the measurable criteria, but inevitably different users will be prepared to pay different amounts for different translation qualities.

Quality evaluations of the English -> French Systran system were performed by an independent consultant in November 1976 and in July 1978.

The text samples, of over 10 000 words, included scientific journal papers, abstracts, and Commission documents in the field of agricultural economics and food technology.

The results were the following:

Criteria	First evaluation Nov, 1976	Second evaluation July 1978
Intelligibility		
- of original text	97 %	99 %
- of revised human translation	98 %	98 %
- of unrevised machine translation	45% *	7.8 % *
- of revised machine translation	96 %	98 %
Revision rate of machine translation		
- by professional translators	-	42 %
- by professional evaluators	29 %	31 %

*) These figures correspond to machine translation based on the currently existing dictionaries. If the dictionaries are updated prior to processing, the figures become 75% and 90 %, respectively.

3. Cost evaluation

The cost factors of human translation are the following:

a) translation, in writing or by dictation into a recorder;
b) typing of the translated text;
c) post-editing, by the translator himself and/or an independent revisor;
d) typing of the edited text.

Factors (a) and (c) implicitly include the investment constituted by the translators' and revisors' professional training. Any lack of such training must be compensated by time (and money) spent on referring to specialized dictionaries or consulting specialized terminologists during the translation and revision periods.

The cost factors of machine translation, on the other hand, can be divided into investment factors and operational factors.

Investment factors are:

a) creation of the system software;
b) creation of the bilingual dictionary covering the subject field.

Operational factors are:

c) pre-editing, by clerical staff;
d) text input by keypunching, magnetic encoding, or optical character recognition; or, alternatively; conversion of text existing on magnetic tape into the required format by an interface program;
e) translation and printing by computer;
f) post-editing, by linguistic staff;
g) typewriting (or photocomposition, or computer printing) of the edited text..

A first cost evaluation of the Commission's Systran system was carried out in November 1976.

The cost was established per translated word, for Systran and for human translation within the CEC as well as by free-lance translators. The results showed that

- unrevised machine translation is considerably less expensive than any human translation;
- revised machine translation is less expensive than revised human translation produced by the CEC services;
- revised machine translation is more expensive than unrevised human translation produced by free-lance translators;
- revised machine translation breaks even with unrevised human translation by free-lance translators, if the texts are available in machine-readable form, doing away with the need for keypunching.

Comparing the evolution of cost and quality as the system evolves toward
- enlarged dictionaries
- sophisticated homograph routines

it becomes evident that every gain in quality must be paid by an increase in cost.

The level of quality that one can get corresponds to the amount of money that one is willing to spend, and a high degree of perfection corresponds to a level of cost that is much higher than that of human-only translation.

4. Infrastructure

The actual utilisation of a translation system requires a complex technical infrastructure.

A computer with a large central memory must be available, not only for the translation operation, but also for dictionary buildup and updating. Personnel must be available for systems maintenance, pre-editing, keypunching, post-editing, dictionary coding and updating.

Pre-editing aims at correcting errors in the source text and eliminating keypunching errors, introducing pagination instructions (identification of paragraphs, forumulae, proper names and tables) and possibly even resolving basic ambiguities in the source text.

Post-editing aims at correcting translation errors and raising the stylistic quality to a level acceptable to the end user.

5. The European contribution

While the European institutions were thus developing and evaluating Systran, they did not lose sight of the progress made in the field of computational linguistics in Europe.

In a number of meetings convened by the Commission during the first months of 1978 the representatives of a number of European universities, including Grenoble and Saarbrücken, Pisa and Manchester, agreed to pool their resources and to develop a single high-quality European translation system under the responsibility of the Commission. The planning phase will be terminated and the actual development of the software and the linguistic modules will start early in 1979. The goal is to have an experimental system by 1982, and a fully operational system, covering at least the four major European languages, in 1984.

In the European system dictionaries and linguistic systems (syntactic analysis and generation routines) will be independent from the software components, so that the operation of the system will no longer require complex combinations of competences. Emphasis will be on the portability of the system between various makes of computer, and on the ease of updating, allowing the linguists in charge to take into account the outcome of the latest linguistic research.

The high cost of the system development should be more than offset by the high quality output achieved by sophisticated parsers developed in the universities.

6. Enlarging the market

The activity of the human free-lance translator used to lie almost entirely in the field of literature. Only recently, with the growing activity of the international organizations, more translators were needed for economic and technical texts.

The market for computer-aided translation has little in common with the traditional market.

It can be described as follows:

(a) Translation departments of large national and international authorities. The interest of national authorities tends to limit itself to language couples

involving their official language, except for bi- or tri-lingual countries such as Canada, Belgium, or Switzerland. International bodies require a wider spectrum of language knowledge, so that the cost of both human and computer-aided translation tends to be higher. This applies to the European Community institutions and to the United Nations' agencies.

(b) Scientific and technical data bases of a bibliographical nature.
The creation of Euronet will make English-language data bases accessible all over Europe, and the users in non-English-speaking countries will welcome the availability of machine-translated titles and abstracts in their own languages, Some will even want to use translation systems made available through Euronet for the raw translation of machine-readable text.
Cooperative information systems the input of which is produced in the languages of the contributing countries are another type of future users of machine translation.

(c) Publishers of periodicals of all types will have to examine the cost-effectiveness of editing their journals in cover-to-cover translation in other parts of the world.

(d) Industrial companies will make extensive use of machine translation for the preparation of multilingual sales brochures and operation manuals for their export trade.

It is expected that in the years to come the demand for translation will remain higher than the combined offer by trained human translators and the computerized services.

On the one hand, the political and economic scene is becoming increasingly multilingual; on the other side, the availability of very low-cost, rapidly available raw translation will lead to the development of an entirely new market.

In the medium term, however, a new profession is expected to emerge: In parallel with the classical translator, there will be a body of so-called editors specialized in the pre- and post-editing of the raw products of machine translation.

The outlook is one of a booming development with advantages for both the trained linguist, the publisher, the system operator and the user of translations.

Great expectations!