THE FACTS ABOUT AUTOMATIC TRANSLATION

Loll ROLLING
Commission of the European Communities
Luxembourg

ABSTRACT

There is an immense gap between the claims of the theoretical linguists as to what the computer must ultimately do in the field of text analysis and translation, and the hard economic facts facing the system designer. A state-of-the-art survey shows the various alternatives proposed in the past, and the state of progress achieved today. The European Community, after having the Systran system adapted to European languages, is investigating the feasibility of a more advanced system that will be developed in cooperation with national authorities and European universities. Methods and criteria for the evaluation of cost, time and quality are defined so as to allow for impartial and comparative assessment of the products of various translation systems. An assessment is made of the future market and mode of exploitation of computer-aided translation systems.

AUTOMATIC TRANSLATION

1. DEFINITION AND OBJECTIVES

Translation is the transposition of information from one natural language into another. That is what everybody agrees upon. But even the translators themselves disagree as to the objective and mode of implementation of this activity.

The more technically-minded aim at precision and fidelity to the original, whereas the literary-minded among them aim at "passing the message" on to the reader, even if the syntactical structure and the choice of words are not identical in the source and target languages. In the first case, the "philosophy" of the source language is preserved, whereas in the second it is superseded by that of the target language. The first activity is a technical one, the second has artistic as well as technical components.

It is obvious, therefore, that the incorporation of mechanical aids into the translation process can help to solve the problems of the technical translator but not those of the person who translates novels or even poetry.

Initially, as long as the quality of the machine product is not entirely satisfactory, it will be necessary to supplement machine translation with either a pre-editing or a post-editing process, or both.

Pre-editing aims at correcting errors in the source text and eliminating keypunching errors, introducing pagination instructions (identification of paragraphs, formulae, 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.

2. STATE-OF-THE-ART

I shall not go into the (short) history of machine translation. Much has been said and written about the Georgetown project, the ALPAC report and the Leibniz group. I shall only describe the present situation and attempt to produce a typology or classification of the systems that are in existence today.

This typology is based on a study produced by Herbert Bruderer in 1977. According to Bruderer, 9 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. based 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, now extends into English-French, English-Russian, English-Spanish and French-English, and the Logos system 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.

This shows where the actual need for translation is located:

First: There is obviously a need for translation from Russian into English in the United States, stemming partly from USSR-US cooperation in space research and, more generally, from the need for the U.S. to be fully informed about any progress in Soviet science and technology.

Second: There is a need for translating large volumes of text from English into French in Canada, due to the politico-linguistic situation of the Province of Quebec.

Third: There is an equally large need for translation from and into the main European languages in the European institutions; this need derives directly from the cooperative structure of the European Community.

Fourth: There is a large potential market for the translation of technical manuals by companies in the export trade.

A detailed market study will be carried out by a C.E.C. consultant in 1979. It is to investigate linguistic ability in various fields of interest within the E.C. countries and determine the need for translation in the following areas:

- national and international authorities which have large translation services in operation;

- cooperative information systems the input of which is produced in the languages of the contributing countries;

- bibliographic data bases to be made available through international networks to various language groups;

- publications, which could increase their circulation by cover-to-cover translation into other European languages.

3. COMPONENTS; PROGRAMS AND DICTIONARIES

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.

Dictionary buildup is in fact one of the main cost factors, as will be shown in Section 5.

4. IMPLEMENTATION

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.

With a new machine translation system, the most delicate problems, however, arise from the reactions of experienced high-quality translators and revisors to the raw output.

The post-editing of machine-translated texts in fact differs considerably from the revision of human translations, and a transition period is required to allow staff to adapt to a new working environment.

5. COST AND PERFORMANCE 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 refering 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 nicely split 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 cost and performance evaluation of the Systran English - French translation system was performed by an independent consultant in November 1976.

The text sample of 10,000 words included three types of texts, namely abstracts, scientific journal papers, and internal CEC reports in the field of food science and technology.

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 translation;
- 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.

The evaluation study also showed that considerable cost reduction can be achieved in the future by:

- improving the quality of the initial product by feedback routines and reducing the revision requirements;
- streamlining the input and output mechanisms.

The evaluator finally expressed the opinion that the amounts invested in the creation of machine translation systems and of the corresponding dictionaries can be written off within relatively short periods of time if the volume of text to be translated is relatively large. For example, 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.

6. 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.

Intelligibility ratings by a group of independant evaluators give a clue as to the actual usability of the raw-product. Revision time might have been a good quality 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 revision process, is also a good quality measure. These are the measurable criteria, but inevitably different users will be prepared to pay different amounts for different translation qualities.

7. THE ROAD TO UTOPIA?

The emergence of the digital computer in the Fifties had raised great expectations as to the ability of the new machines to solve the language problem by fully automatic word-to-word translation.

Great were the disillusions when it was shown that the problems involved were so intricate that it would take a new generation of computers and legions of linguists to solve them.

Today again, we are faced with a crowd of foolish claims and unreasonable demands as to the miracles that must be performed by machine translation:
The following claims were made in a recent paper by one author (my comments are in brackets):

- Translation must reflect the cultural background of the target language, not the source language (thus a description of a baseball game must come out in French as the description of a soccer game?).
- Deficiencies of the original must be corrected in the translation process (including incorrect syntax, spelling errors, and absence of punctuation?).
- Translation must do away with apparent ambiguities of the source text (even those intentionally left by the author?).
- A system must be able to distinguish proper names from hitherto unknown words (have you heard of Sebastien Mouche, the inventor of the "bateau-mouche"?).
- A system must be given sufficient knowledge of the world (!) to avoid mistranslation of nonsensical statements.
- Human translators do not make errors. (The latter is the official position of many translation services, and is largely shared by people in my own institution).

If we want to succeed, i.e. provide our institutions with a cost-effective instrument to overcome our language barriers, we must do away with these unreasonable demands and concentrate on the essential that is feasible now.

Bar-Hillel, who was well known for his skeptical attitudes, suggested that automatic translation will have to rely on strategies rather than a theory, and that the issue may well be an economic and not a scientific one.

AUTOMATIC TRANSLATION

8. EUROPEAN PLANNING

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-language terminology data bank, which is now operational, and has decided to break new ground in the field of automatic translation.

As a first step, after an investigation of existing systems, 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.

9. A LOOK INTO THE EIGHTIES

The Bruderer study, cited in Section 2, also mentions a number of systems that are in experimental phase, but almost operational. These include the AVIATION System of TAUM (Montreal) and the Brigham-Young University system in Provo (USA), but also the remarkable

European realisations of the Grenoble and Saarbrücken Universities. The EURATOM system, located at Ispra, was discontinued in 1975.

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.

*

To conclude, a statement of policy:
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.