Seventh Cranfield International
Conference on Mechanised Information
Storage and Retrieval Systems

July, 1979

THE SECOND BIRTH OF MACHINE TRANSLATION
a timely event for data base suppliers and operators

by
Loll N. Rolling
Commission of the European Communities
LUXEMBOURG

THE SECOND BIRTH OF MACHINE TRANSLATION, a timely event for data base suppliers and users.

L.N. ROLLING

COMMISSION OF THE EUROPEAN COMMUNITIES

Definition and objectives of computer-aided translation and a history of its evolution from the word-by-word translation of the Fifties to the artificial-intelligence approach of the Eighties.

The marketplace of machine translation extends from the international institutions to trade and industry and, last not least, the data base industry.

Two possibilities are identified, including the translation of global data base input by the supplier, and the on-the-spot translation of items retrieved on-line by the user.

Data are supplied on the cost and time required as well as the quality level that can be attained today.

Translation is the transfer of information from one natural language to another.

The human translator reads the source text, comprehending the message, and tries to re-formulate the message into a target-language statement. In order to do this properly, he must be fluent in the two languages, but he must also know the cultural environment of the two languages. He must not only know the special terminologies and idioms in each language separately, but he must also have a good feeling for equivalent terminologies and idiomatic expressions.

The very first attempts at computerizing the translation process did not take these requirements into account. The promoters, who were computer people and not linguists, attempted to perform word-by-word translation comparable to what is provided today by pocket translators.

The failure of these first attempts led to the recognition that there was more to it than just word-to-word equivalences, and efforts were started to develop transformational grammars and, based on these, to perform actual syntactic analysis of the source texts. But these efforts came too late, because, in the light of the shabby results of the first attempts, the US government had decided (and others sponsors followed) to discontinue the financial support for computer-aided translation projects. The immediate consequence was that M.T. went underground.

On the one hand, military intelligence bodies secretly financed the development of Russian-English translation systems in order to get access to the Russian scientific literature. The astronauts of the Apollo-Soyouz project benefited from this effort.

On the other hand, the activities of the higher-level linguistics institutes which had been aimed at machine translation, resulted in a booming development of theoretical linguistics and comparative analyses of structure and terminology, mainly of ancient and medieval literature.

*

Today, text processing by computers has become commonplace, and the networks of Lockheed and SDC have shown that every potential user can be given access to the information he needs.

But Euronet-Diane is now operational and may well become a source of frustration for those who do not speak the major languages used in the data bases offered : English, of course, and French and German to a lesser extent.

Thus the scene was set for the second birth of machine translation, and the European Commission felt that it had a major responsibility in this development and that it would not be forgiven a failure this time.

Worldwide inventories of multilingual systems were prepared by Herbert Bruderer in Bern and Georges Van Slype in Brussels.

A prospective study of the present translation market and its possible extension for various types of text showed that a large need existed

- in international institutions, European and worlwide ;
- in multi-lingual countries such as Canada, Belgium and Switzerland ;
- in co-operative information systems, the input of which is prepared in the languages of the contributing countries ; in multilingual information systems, and in bibliographic data bases to be made available to various language groups ;
- in technical manuals and promotional brochures for industrial companies in the export trade, and
- in scientific publications.

The survey also shows that there is a need for both high-quality translation and "quick-and-dirty" raw translations. Whereas existing systems could immediately supply the latter, high quality could be obtained only via human post-editing, which is time-consuming and expensive.

The Commission therefore decided to do four things in parallel :

- develop multilingual thesauri, on the basis of its past experience in this field, and using a software package specially developed for thesaurus maintenance,
- compile a terminology bank to assist the human translator,
- use and improve an existing translation system to cope with its immediate needs and
- develop a sophisticated new system based on the experience accumulated by the linguists over the last decade.

Systran is the existing system, Eurotra the system to be developed in the next few years.

The characteristics of Systran and Eurotra are shown in table 1.

Systran was developed by World Translation Center in the U.S. for use by the European Commission and for a Canadian company who produces English-French translations for the automobile industry in Canada.

Translations can be used unedited, for the comprehension of the subject content of documents in the information gathering process.

It has to be post-edited, either manually or using text-processing equipment, in the information dissemination process.

Language couples now covered by Systran at the European Commission include English-French, French-English and English-Italian.

Systran is operational on IBM and Siemens computers.

Whereas data processing and linguistics are strongly interconnected in Systran, requiring very competent linguist-analysts, Eurotra will attempt to separate the two, allowing for system improvement by linguists alone, without any involvement of analysts.

Whereas Systran makes use of language-couple-specific devices, Eurotra will be designed to allow simultaneous translation from one source-language to all target languages.

Finally, Eurotra will incorporate the latest findings of research into artificial intelligence, emulating the human translator's knowledge-of-the-world.

Dictionary formats must be compatible between Systran and Eurotra, so that terminological information can be transferred between the systems without extra cost.

The Commission has acquired the right to use Systran within the European institutions, within the government agencies of the Member States, and for use in Euronet.

Table 2 shows a comparison of the cost elements of human translation compared with the present cost of Systran, as determined by an official evaluation, and the expected cost of Eurotra.

Human translation involves translation proper, in writing or by dictation into a recorder, typing of the translated text, post-editing by the translator himself or an independent revisor, and typing of the revised text.

The post-editing effort is likely to decrease from Systran to Eurotra, whereas the translation cost will increase with growing system sophistication.

Today's computer-aided translation process includes text imput by either keypunching or mag-tape encoding. The translation programme identifies any word not found in the system dictionaries.

The raw output undergoes post-editing by human linguists.

To-morrow's translation system will comprise an interface programme for the conversion of text received on magnetic tape into the required format.

The source text can be pre-edited, and the raw translation will be post-edited using sophisticated text processing machinery.

*

An analysis of translation modes, applicable to the text components involved in multilingual retrieval, shows which operations are technically and economically feasible.

"Free indexing terms" which are increasingly used as subject-content indicators in publications, but are of no use for multilingual retrieval, were not taken into account.

Controlled-language abstracts, involving the use of restricted syntax and controlled vocabularies, were included although they are used in a limited number of retrieval systems only.

There are two alternatives for translation in the scientific and technical data base field : batch-mode translation of the entire data base during the input phase, and translation of the retrieved items only in the output phase.

Batch-mode translation is inexpensive per item translated and does not require permanent access to the computer ; il allows installation of the same data base in different language versions on different computers in different countries ; the output can be used directly by the end-user.

On-line translation requires a lower overall expenditure, but the organisational setup, providing permanent access to both the data base and the translation system, is very complex.

Manual translation (i.e. translation by trained linguists) is more expensive than computer-aided translation and can therefore be applied to retrieved items only, the only exception being the case of international secondary journals, where the input, prepared in a large array of languages, must be translated into one common language for publication.

The translation of retrieved titles only does not justify the organisational effort involved in on-line translation.

The translation of full documents, on the other hand, cannot be scheduled prior to retrieval, in view of the high cost required for translating items that may never be asked for.

The most promising and therefore most frequently occurring combinations will therefore be the following :

(a) cover-to-cover translation of free-language abstracts and descriptors by a combination of computer-aided translation (for the abstracts) and a multilingual thesaurus (for the descriptors).

(b) item-by-item translation of retrieved source-language abstracts into the user's language, which may be followed by the translation of full documents obtained through a back-up mechanism.

These appear to be the best alternatives for an aptimal exploitation of the available multilingual tools.

*

Regarding the availability of the tools developed, the Commission's policy is as follows ;

. Multilingual thesauri, developed using Commission funds, are freely

available, and Astute, the software package used by the Commission for the compilation, maintenance and printing of thesauri is available on request.

- Eurodicautom, the Commission's multilingual terminology data bank, will be made accessible via Euronet.

- The Systran system will be made available to data base suppliers or operators ; its use will be limited to the translation of data bases connected to Euronet ; a number of requests have been received already.

- The Eurotra system will be made available not only for exploitation in the Euronet environment, but may be equally well employed for linguistic research and teaching in the institutions of the Member States.

*

* *

TABLE 1

M. T. SYSTEMS CHARACTERISTICS

	SYSTRAN	EUROTRA
Development period	1975 - 79 (E. C.)	1980 - 84
Developed by	World Translation Center (Dr. P. TOMA)	C.E.C. + six European Universities
Quality objectives	Intelligibility > 90% (now 78%) Revision rate < 20% (now 33%)	Intelligibility 99% Revision rate < 10%
Fields of application	unedited : INFORM. RETRIEVAL edited : INDUSTRIAL USES	unedited : INTERNAT. COMMUN. edited : PUBLICATION
Languages (to be) covered by C.E.C. development	E→F, F→E, E→I, ...	E, F, D, I, NL, DK in all combinations (60)
D. P. support	IBM and SIEMENS	Full portability
Dictionary format	C O M P A T I B L E	

T A B L E 2

C U R R E N T T R A N S L A T I O N C O S T

Investments not included

HUMAN TRANSLATION

TRANSLATION handwriting or dictating	Typing if dic.	POST - EDITING by translator or revisor	TYPING	100 %

SYSTRAN

INPUT	MACHINE TRANS-LATION	POST - EDITING	TYPING	70 %

EUROTRA

INPUT	MACHINE TRANSLATION	POST EDITING	40 %

(Alternative presentation of table 2)

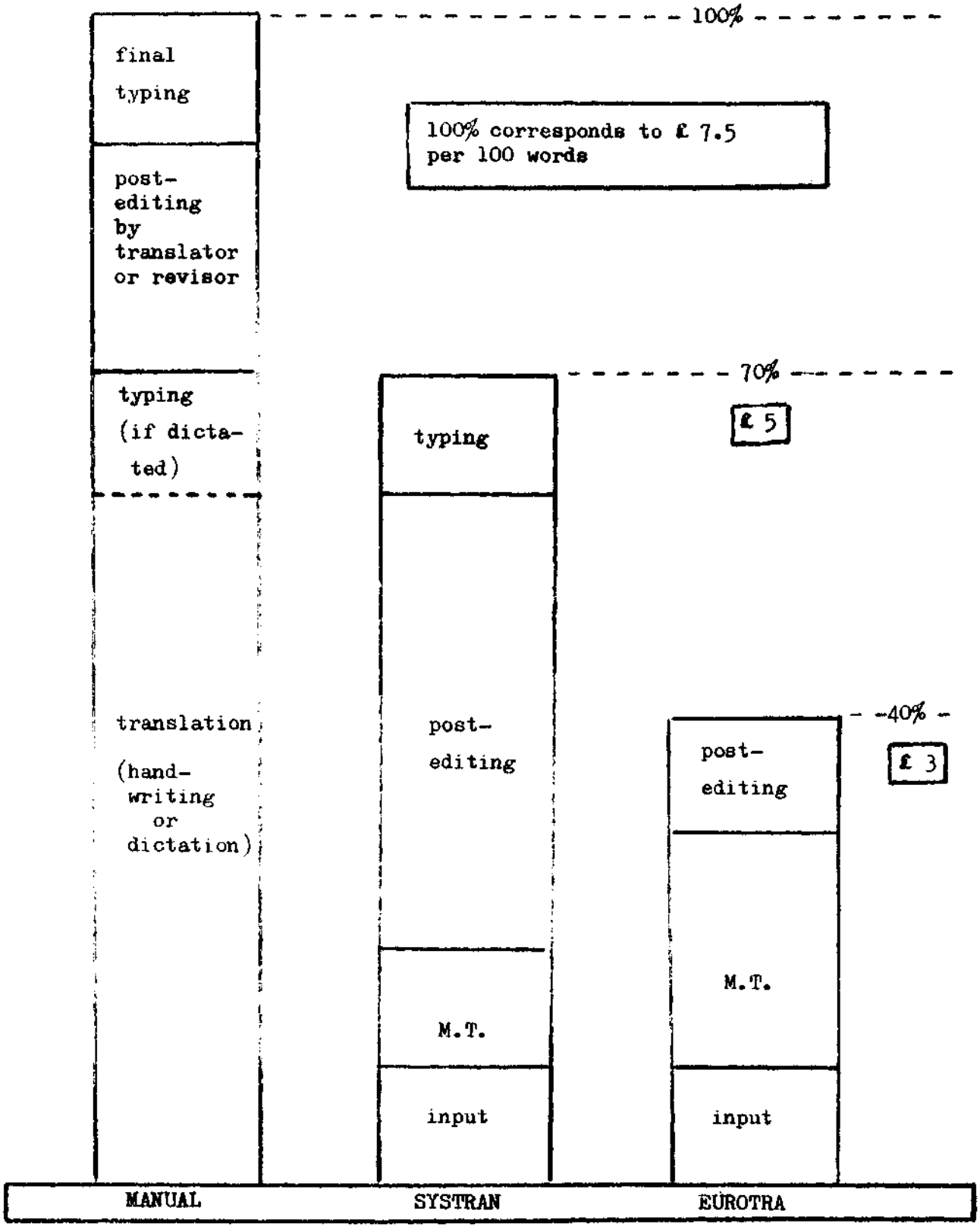

Table 2 : Current translation cost