

Published in cooperation with
the Board on Science and Technology
for International Development,
Office of International Affairs,
National Research Council

***MICROCOMPUTER
APPLICATIONS
IN EDUCATION AND
TRAINING FOR
DEVELOPING
COUNTRIES***

Proceedings of
a Meeting
on the Use of
Microcomputers for
Developing Countries

Westview Press / Boulder and London

1987

The Contribution of Machine Translation: Present Status and Future Perspectives

MURIEL VASCONCELLOS

THE BROAD PERSPECTIVE

The reality of the translation of natural language by computer opens up new perspectives for the exchange of knowledge among cultures. The costly overhead associated with human translation—in terms of both economics and time delays—can now be greatly reduced thanks to recent advances in machine translation (MT). Rough, or “raw,” output for purposes of information only is already available at dramatic savings, and polished translations are produced in as little as one-third or even one-quarter the time it takes to complete a human translation at approximately half the expense. These savings create the very real possibility of equal accelerations in the transfer of information: for the same amount of money, volumes can be doubled, trebled, or quadrupled while delivery speeds improve in the same proportions. Machine translation, because of its increased flexibility, is already being used for the processing of information that was not being translated previously. It is not an exaggeration to say that wherever there is a need for the transfer of knowledge between cultures with different languages, there is a potential application for this new and powerful technology.

Microcomputers, to some degree, are already extending the reach of MT, bringing it within the capabilities of the small investor and even of the single individual working independently. The microcomputer is used both as a word processor for text that is being translated on a larger computer and as the host itself of the MT system. The low cost and ready accessibility of the microcomputer, combined with the savings to be effected through the use of machine translation, hold promise for quantum increases in communication across language barriers.

The Present Status of Machine Translation

The concept of machine translation actually dates back more than half a century (Zarechnak, 1979). Two fairly detailed processes were proposed simultaneously in France and the Soviet Union in 1933, but they were far ahead of the technology of the time. For the two decades following the invention of the digital computer in 1946, scientists tried a multitude of approaches in their search to automate the principles of translation, while suggesting that computers would soon be capable of translation without any need for human intervention. The linguistic assumptions that fueled this effort have since been proven to be somewhat simplistic. As research progressed, the real complexity of language became more apparent; the problem was that at that time linguistic formulations, programming techniques, and computers were not yet equal to the task. Still, the need and the interest remained, and isolated initiatives, including commercial ones, continued to be pursued.

Improvements in programming techniques were gradually matched by increased knowledge of the syntactic and semantic rules that are involved in the analysis of natural language. By the 1970s, computers had acquired the speed and efficiency that made it possible to handle the enormous and deeply coded dictionaries that are needed for machine translation. Miniaturization in turn, brought personalization and made it possible for the linguist to be his own programmer. Word processing provided the routine availability of text in machine-readable form.

The advent of word processing was a watershed for production machine translation. Not only did it facilitate input, it provided the translator with an efficient tool for postediting. The importance of having fully automatic high-quality (FAHQ) output was

no longer as great, and the translator was now incorporated into the process. By 1980, systems were in place at a number of sites around the world. In 1983, Ian Piggott reported at the annual Aslib meeting in London that 400,000 pages had been run in production environments over the preceding 12 months (Piggott, 1985:98), corresponding to the output of approximately 230 full-time translators working in the traditional mode. Four systems (ALPS, Logos, Weidner, and Smart) were by then available in bureau service, and one of these, Weidner, was being offered on microcomputers. Logos was on the Wang OIA/140. Since then, a new system, Textus, has been developed by the inventor of Systran, and is available on an IBM PC.

THE EXPERIENCE OF SPANAM AND ENGSPAN*

Milestones

The Pan American Health Organization (PAHO), regional office for the Americas of the World Health Organization (WHO), entered the machine translation picture in 1976, just at the threshold of the linguistics and technological advances that were to make MT a more feasible concept. Since that date, PAHO has developed two in-house mainframe systems using its own resources. The first system to be undertaken was SPANAM, which has been translating from Spanish into English since early 1980 (sample output is shown in Figure 26.1). ENGSPAN, which translates from English into Spanish, was developed with partial support from the U.S. Agency for International Development (AID) and became operational in 1984. (Sample output is shown in Figure 26.2.)

SPANAM began to provide machine translation to internal users at PAHO in 1980, shortly after a network of Wang word processors was introduced at its headquarters in Washington, D.C. Since that time, word processing documents, usually produced in the secretariat for other purposes, have been submitted from the Wang to the IBM mainframe, where the translation algorithm is run against the system's large dictionaries, which are stored on permanently mounted disks. After the translation program has

*SPANAM and ENGSPAN are trademarks of the Pan American Health Organization.

LA ESTRATEGIA DEL ILRAD

Las dos enfermedades mencionadas son causadas por parásitos que son transmitidos por insectos. La mosca tsetse transmite los tripanosomas y la teileriosis es transmitida por las garrapatas. En ambos casos las relaciones entre parásitos, huéspedes y vectores son complejas y sutiles, y por tanto la intervención es difícil. Además, en ambos casos, otros animales salvajes y domésticos sirven también como huéspedes de los parásitos, creando así reservas de infección prácticamente inaccesibles a las medidas de control.

THE STRATEGY OF ILRAD

The two diseases mentioned are caused by parasites that are transmitted by insects. The tsetse fly transmits the trypanosomes and theileriosis is transmitted by the ticks. In both cases the relations between parasites, hosts and vectors are complex and subtle, and accordingly the intervention is difficult. In addition, in both cases, other wild and domestic animals serve as well as hosts of the parasites, creating thus reservoirs of infection practically inaccessible to the measures of control.

FIGURE 26.1 Sample translation by SPANAM.

Among the 20 known virus diseases of the potato, the two OK
that significantly affect production- potato virus Y and
potato leaf roll virus -have received the most attention at
CIP. Good resistance to virtually all major viruses is now NO
available and in the process of incorporation into national
breeding populations.

CIP researchers are also screening germplasm for OK
resistance to the insect vectors that transmit viral
diseases. The principal vectors are the green peach aphid OK
and the potato aphid; others include leafhoppers, leaf OK
miners, potato tuberworm moths, and weevils. On certain OK TU
hybrid potato plants, CIP researchers have noted glandular
foliar hairs with sticky tips that trap insects -mainly
aphids, but also flea beetles and mites-- reducing epidemic
infestations.

Entre las 20 enfermedades conocidas víricas de la papa,
los dos que significativamente afectan la producción- virus
Y de papa y virus de enrollamiento de la hoja -han recibido
la mayoría de atención en el CIP. La resistencia buena
a virtualmente todos virus principales es ahora disponible
y en el proceso de incorporación en poblaciones nacionales
de mejoramiento.

Los investigadores del CIP también están examinando el
germoplasma para resistencia a los insectos vectores que
transmiten enfermedades víricas. Los vectores principales
son el áfido verde del durazno y el áfido de papa; los
otros incluyen saltarillas, minadores de hojas, polillas de
papa, y gorgojos. En ciertas plantas de papa híbrida,
investigadores del CIP han notado pelos foliculares
glandulares con puntas pegajosas que atrapan insectos.
principalmente áfidos, pero también pulgillas y ácaros -
reduciendo infestaciones epidémicas.

FIGURE 26.2 Sample translation by ENGSPAN.

run, the resulting document is returned to the Wang for postediting on the word processing screen (Figure 26.3). SPANAM has always been what Lawson (1982) would call a "try-anything" type of system: the vocabulary and syntax of the input are entirely free, and the text is not pre-edited at any point. As of the end of September 1985, it had produced a total of 2.5 million words (about 8,500 pages) for more than 90 requesting units under some 830 separate work orders. The service is provided to offices and programs not only at the headquarters in Washington but also in the field and to WHO in Geneva.

ENGSPAN, the counterpart system, became fully operational in August 1985. It operates in the same mode and has already produced more than a half-million words of Spanish text, most of it for publication. It is linguistically more sophisticated than SPANAM in a number of respects: its English analysis uses an augmented transition network (ATN) to parse the entire sentence based on 175 types of syntactic and semantic information; idiomatic expressions are highly context-sensitive; and special rules for Spanish, based on an additional 101 criteria, operate at the synthesis stage to generate structures that do not exist in English—for example, certain uses of the subjunctive.

Over the next year, the plan is to provide SPANAM with a parser for Spanish similar to ENGSPAN's parser for English, as well as with many of the other features that have been developed for the newer system over the last two years.

Cost Effectiveness

Even though SPANAM is slated to undergo improvement, its practical efficiency has already been demonstrated repeatedly. One of its early applications, carried out over two months starting in January 1981, showed how much money could be saved. The task called for the translation of 101,296 words of Spanish-language contributions to PAHO's large biennial budget document, which by regulation must be published in both languages. This was felt to be a particularly appropriate application since much of the retyping and proofreading that have traditionally been involved could be reduced or eliminated with MT. Also, the transfer of numerics would be guaranteed to be accurate. The results exceeded expectations.

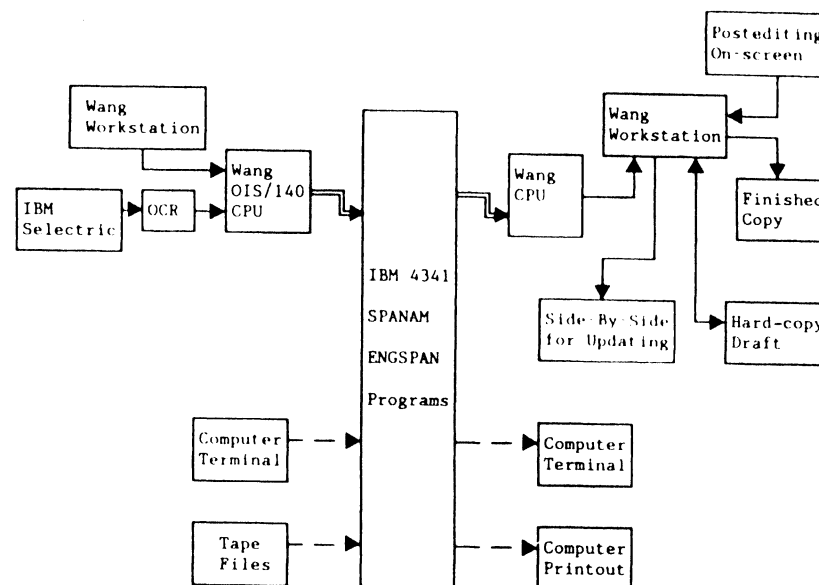


FIGURE 26.3 The PAHO Machine Translation System, 1980.

It turned out that SPANAM cost 61 percent less than contract translation for the same number of words at the then prevailing rate of \$55 per thousand words (Table 26.1). There was a monetary savings of \$5,078.48. In terms of time, a job that would traditionally have taken 66 days was accomplished in 36, for a savings of 45 percent. This calculation takes into account all factors of overhead for both modes, including a hypothetical cost for machine time, for which in fact there is no charge at PAHO.

The effectiveness of SPANAM, and now ENGSPAN, can also be measured in terms of daily output per posteditor. The overall figure for both systems averages between 6,000 and 6,500 words in an eight-hour day. On one occasion, it was possible to postedit 11,376 words of SPANAM output in eight hours. It should be kept in mind that this is finished, machine-readable text that does not have to be recopied. These figures compare with the traditional standard in the United Nations agencies of 2,000 words a day per translator, to which must be added the cost of transcription. At the latter rates, translation would remain a luxury for the elite.

TABLE 26.1 Example of Cost Savings with SPANAM

| | Amount in US \$ | Staff-days |
|---|--------------------|------------|
| TRADITIONAL PROCEDURES: | | |
| Contract translation at \$55 per 1000 words; Processing; Cross-checking; Keying of translation; Proofreading, adjustments: | \$8,296.18 | 65.75 |
| SPANAM: | | |
| Postediting of 101,296 words; Supervision; Submission, Retrieval, formatting of text; Proofreading, adjustments; Machine-time:* | \$3,217.70 | 36.25 |
| SAVINGS ACHIEVED: | \$5,078.48 | 29.50 |

*A hypothetical figure, since PAHO does not have a job accounting system for charging the use of the computer.

It has been our experience that the contract translators, who earn twice as much at human translation as at postediting MT, will in most cases prefer to do a machine translation. Since the contractors have the option of going either route, and since it is in their interest to make as much money as they can, it must be assumed that at the very minimum they are able to work at least twice as fast using the machine output. They have also commented that they prefer to use machine translation because they feel more rested at the end of the day.

The Working Environment

The posteditors of SPANAM and ENGSPAN are also responsible for enhancing the system dictionaries. They are trained in applying the codes that are required in order to activate appropriate decisions by the parser and to trigger special, context-sensitive equivalents in the target. As they postedit, they follow a side-by-side printout that shows the source language on the left, the target on the right, and a series of diagnostic flags that appear in the center column between the two texts. These flags will indicate, among other things, words that were not found in the source dictionary, sentences that are longer than 70 words, and, in the

case of ENGSPAN, the status of the English analysis. The last will show whether there was a complete parse, a partial parse, or no parse—in which case there will still be a translation, but it will be based on only a phrase-by-phrase analysis rather than on the actions taken by the ATN parser. The output will also have an indication of translational equivalents that are found in the specialized microglossaries and of terms that appear in PAHO's own database of standardized terms (WHOTERM), or of terms that carry a high rating for "reliability."

All this information demonstrates to the translators ways in which the glosses and the coding in the dictionaries can be improved. As they work along at the screen, they use the hard copy to mark all the changes that they will be making in the dictionaries later. The benefit of working in this way is that all the contextual cues are fresh in the translator's mind at the moment the notes are made. This saves much time later on in deciding how the entries should be coded. The translator also jots down problems related to the translation algorithm itself, and these are communicated to the computational linguists, who are usually able to correct them promptly.

As of September 1985, the SPANAM dictionary had a total of 61,252 source entries (94 percent base forms, 6 percent full forms), and ENGSPAN had 45,185. The relatively large size of these dictionaries means that more than 99 percent of the words from a randomly selected text in any of our health-related fields will find a match (about 99.85 percent for SPANAM and 99.53 percent for ENGSPAN). Under these circumstances, all the analysis and synthesis that the system is capable of will be performed. Smaller dictionaries would give many more "not-found" words, and this would mean that the analysis could easily break down.

Our experience has shown that the combination of production, terminology retrieval, dictionary and program maintenance, and advanced development makes for a highly effective environment in which each function supports—and actually potentiates—the others.

FUTURE PROSPECTS FOR SPANAM AND ENGSPAN

The future of machine translation at PAHO is unfolding in three major directions: (1) the porting of ENGSPAN, and eventually SPANAM, to other mainframe sites, especially where it can serve the community of Third World nations; (2) the development of additional language pairs; and (3) the adaptation of the different systems to the microcomputer environment.

The mainframe version is soon to begin operation on the computers of the U.S. Agency for International Development, where it will process texts that are submitted either at headquarters in Washington or, via telecommunication, from field offices around the world. Similarly, field offices of PAHO, and perhaps national ministries of health, will soon be linked in a network through which the MT systems can be used. At a later date, it will be important to locate ENGSPAN and SPANAM at computer centers that serve the sister organizations of the United Nations system. And finally, serious inquiries are already being received from other areas of the public sector and from the private sector, where there is also interest in obtaining these systems.

For additional language pairs, the first priority is English into Portuguese, an official language of PAHO. With the English analysis and Spanish synthesis that have been developed for ENGSPAN, we already have about three-fourths of the system in place, and ENGPORT, as we now call it, could become a reality, with only a small investment, in less than 18 months. Another combination that would be relatively easy to develop is English into Haitian Creole. WHO is committed to publishing more information in vernacular languages. The application of MT to such an effort could help to boost both literacy and employment among people whose mother tongue does not have a strong written tradition.

The third direction that lies ahead is adaptation to a microcomputer. Some of the factors involved in this step are discussed in the next section.

Machine Translation and Microcomputers

It is certainly inevitable that machine translation will migrate increasingly to the microcomputer environment. Already, the microcomputer is a word processor at both the input and the

output end of MT. To some extent, it also provides access to lexical databases, both the translator's own and large ones for which the microcomputer can serve as an on-line terminal. The latter technology—being exploited for example with SUSANNAH, a microcomputer environment that has been developed for translators by the University of Saarbrücken—is bound to be maximized as telecommunication networking expands and as windowing technology becomes more widely available.

What is of interest now is the microcomputer as host to the MT system itself. All the major commercial vendors are moving in this direction, and one of them, Weidner, has had a microcomputer product on the market since late 1983. From the beginning, Weidner ran on a minicomputer, and Logos has been running on a minicomputer and a Wang OIS/140 for nearly two years. Another minicomputer product is METAL, developed at the University of Texas and launched by Siemens in 1984. Textus, the only system so far to make its initial debut on a microcomputer, has been on the market for a short time.

There is no doubt that it is now technically feasible to port a major MT system to a microcomputer. However, there are limitations. The processing of translation is slower by several orders of magnitude. Certain complex operations may have to be sacrificed. For example, Textus does not have the powerful routines for the handling of idioms that are characteristic of its mainframe predecessor, Systran. Parsing, depending on the technology used, may take up large amounts of CPU and slow down a system considerably. In order to fit within the strict confines of space allowed, the programs may have to employ extensive overlays. Another consideration is that the programming languages which offer the best utilization of microcomputer capabilities—for example "C"—are not the ones used originally to write the mainframe versions. In the case of PL/1, the language in which SPANAM and ENGSPAN are written, the compiler for the microcomputer has dispensed with some of the functions that are heavily used in our systems. Workarounds are possible, but in the long run it may be better eventually to rewrite the code to maximize the features offered by a language specifically geared to the microcomputer.

As far as the dictionaries are concerned, the latest hard and not-so-hard disks are now capable of providing up to about 30 MB of on-line access. This space is barely adequate for dictionaries such as ENGSPAN's with 40,000 entries or more, and a dictionary

that is much smaller will limit the overall power of any MT system. The disk must be of a medium that is easily updatable. It is possible—indeed more than likely—that disk capacity will increase substantially in the very near future. The main problem is to find a fast and efficient way to perform the dictionary lookup, which with ENGSPAN may occur as many as five times for a single word.

If a microcomputer is being used for production, it cannot be used at the same time for postediting. Thus, if any degree of intensive production is envisioned, it may be necessary to choose between running the jobs at night or having one unit devoted solely to the translation program itself. In either case, it may be desirable to have several additional units networked for purposes of postediting.

On the positive side, the microcomputer is excellent for user-friendly on-line updating of the MT dictionaries. Weidner has devised interactive facilities for a microcomputer, and Logos has a sophisticated on-line program that runs on a minicomputer. Atamiri, a new system yet to be fully developed, is programmed to update dictionaries interactively in several languages at the same time.

The foregoing considerations all seem insignificant, however, they become extremely relevant when they are weighed against the heretofore undreamed-of advantage of having a microcomputer MT system available at a minute fraction of mainframe cost for use in small operations or by single individuals.

CONCLUSION

There is no doubt that machine translation has a major contribution to make in bridging the isolation gap between cultures with different languages. In the Americas, for example, there is need to greatly accelerate the exchange of experience between the countries that speak Spanish and Portuguese, on the one hand, and those that speak English, especially in the Caribbean. Throughout the world there is knowledge to be exchanged, and this exchange has been held back by the slowness and high cost of traditional human translation.

In the time that it takes for word processing to become the universal mode for the capture of human language, machine translation systems will become even more clever than they are today.

We will truly see the dissolving of language barriers and the world-wide exchange of information that has been the dream of so many.

REFERENCES

- Automatic Language Processing Advisory Committee. *Language and Machines: Computers in Translation and Linguistics: A Report by the Automatic Language Processing Advisory Committee (ALPAC)*. Washington, D.C.: National Academy of Sciences, Division of Behavioral Sciences. National Research Council Publication 1416, 1966.
- Lawson, Veronica, ed. *Practical Experience of Machine Translation*. Amsterdam, New York: North-Holland, 1982. (Especially her Introduction; also: Machine translation and people, pp. v-viii and 3-9.)
- Lawson, Veronica. Machine translation. In: *The Translator's Handbook*, ed. by Catriona Picken. London: Aslib, pp. 81-88, 1983.
- Lawson Veronica, ed. *Tools for the Trade: Translating and the Computer 5*. London: Aslib, 1985.
- Piggott, Ian. Recent developments in practical machine translation. In: *Tools for the Trade*, ed. by Veronica Lawson, 1985.
- Vasconcellos, Muriel, and Marjorie Leon. SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization. *Computational Linguistics* 11 (2/3):122-136, 1985.
- Zarechnak, Michael. The history of machine translation. In: *Machine Translation*, by B. Henisz-Bostert, R. Ross MacDonald, and M. Zarechnak. The Hague: pp. 3-87, 1979.