OVERCOMING LANGUAGE BARRIERS: THE HUMAN/MACHINE RELATIONSHIP

Report of the Fourth Annual Conference

of the

Center for Research and Documentation

on World Language Problems

New York, December 13-14, 1985

Edited by

Humphrey Tonkin

and

Karen Johnson-Weiner

Center for Research and Documentation on World Language Problems
New York 1986

Muriel Vasconcellos (Chief, Terminology and Machine Translation Program, Pan American Health Organization)

MACHINE AIDS TO TRANSLATION: A HOLISTIC SCENARIO FOR MAXIMIZING THE TECHNOLOGY

Abstract

A realistic appreciation of what machines can and cannot do for the translation process will depend ultimately on an integrated view of all the forces that limit or enhance the effectiveness of the technology, the pace of technological development itself, progress in our knowledge about language, and the evolution of social attitudes both toward translation and within the translation environment.

A flexible and holistic approach has guided the implementation of machine translation at the Pan American Health Organization over the past six years. SPANAM, the system developed in-house for the translation of Spanish into English, has been operational since January 1980. Experience with SPANAM provided the basis for development of a system from English into Spanish, ENGSPAN, which became fully operational in August 1985 and has already produced half a million words.

1. The Human-Machine Continuum

Machine aids, which cover a wide range of types, may be usefully thought of as points along a continuum ranging from no automation to full automation. At one end, the person generates the text, which must then be captured on paper, and at the other end, a text must somehow be fed into a machine, which can generate an output automatically. Starting at the human end, the engineering challenge is to find ways of capturing thought on paper--in other words, of bridging the problem of output. In the service of this goal we have seen first the contribution of typewriters, then dictating machines, and now word processors. At the other end of the spectrum there is the concept of fully automated machine translation, in which no human being intervenes. Here the mechanical obstacle is the input. The text must be in machine-readable form before it can be processed by a computer. Only after it has been input can we concern ourselves with teaching the algorithm to address the complexities of language, and with processing the output and all the principles that this entails.

By looking at a continuum, we can think in terms of intermediate stages at which degrees of human/machine interaction support one or the other mode. In the advance from the left-hand side toward automation, humans have gone beyond mere output devices to embrace, in addition, the computerized database containing information about terminology. In addition, the human translator now uses the word processor to organize an on-line file of special and frequently used terms.

Moving in the other direction, from the machine toward the human, there are ways in which translators can interact with machine-generated text, and writers and linguists can customize the input text so that the job for the machine is made easier.

What is important is that, regardless of which is the initial end of the continuum, it is possible to integrate contributions coming from the other direction. In translation services today the environment may be seen not as a fixed set of tools but rather as a dynamic process in which the advantages of technology are maximized at all points along the way, wherever they are most appropriate.

2. Our Relationship to the State of the Art

A realistic appreciation of what machines can and cannot do for the translation process will depend ultimately on an integrated view of all the forces that limit or enhance the effectiveness of the technology: the pace of technological development itself, progress in our knowledge about language, and the evolution of social attitudes both toward translation and within the translation environment.

The idea of applying machines to the translation process is an old and persistent theme which antedates by far the development of the computer as we know it. It had already been on people's minds for some time, in fact, when inventors in France and Russia (George Artsruni and P.P. Trojanskij) independently announced the development of prototype machine translation systems in 1933--more than half a century ago (Zarechnak 1979). The concept was not to gain impetus, however, until the digital computer became a reality.

The ENIAC, the world's first electronic digital computer, had scarcely been unveiled in 1946 when discussions began that same year between Warren Weaver, of the Rockefeller Foundation, and Professor A.D. Booth, of Birkbeck College, London University, about the use of this type of machine for the translation of natural

language (ibid.). For the next two decades the hope was fostered that computers would ultimately be capable of producing a fully automatic translation of high quality. It was based on the belief that there are linguistic universals which can be subjected to logical analysis, and that by taking cues from the linguistic environment it is possible to assign a specific meaning to each word in a written text (Warren Weaver 1949, reprinted in Locke and Booth 1955).

This belief, whether realistic or not, was still ahead of the technology of its time. The development of machine translation was to be held up for many years by limitations in the design and capacity of computers and also by the limitations of programming languages. It was not until the mid-1970s that computers would be large enough and fast enough, and their programming languages clever enough, to manipulate the enormous deeply coded dictionaries and the complex types of rules that are required for the processing of natural language.

Moreover, there was the tedious and costly step of keying in the text so that it would be in machine-readable form. Once the machine had done its job, postediting was a manual task, and a revised text would have to be retyped for presentation to the requester. Either step alone was antieconomic; the two together made machine translation far more expensive than its traditional predecessor.

In the interim, while computers were evolving toward their potential, the field of linguistics began to look for ways to make predictable statements about the behavior of language. As research progressed, the complexities became more apparent, but at the same time solutions often followed. The effort of pushing against the frontiers of what was possible soon brought major advances in the understanding and expression of syntactic rules. The linguist's capabilities were expanded not only by these new approaches to linguistic knowledge but also by the development of higher-order programming languages such as PL/1, SNOBOL, and later LISP, C, PROLOG, ADA, and others that are suitable for the processing of natural language. The linguist found that she could be her own programmer. This possibility vastly facilitated the development of linguistic algorithms and gave impetus to the subdiscipline of computational linguistics.

At the same time, the computer was also making its contribution, at the other end of the spectrum, to the existing translation environment. The capacity to store large volumes of data made it, feasible in the 1960s to consider for the first time the development of lexical databases that would pool the contents of different dictionaries and glossaries, capture the fruits of ongoing terminological research, facilitate the updating of dictionaries, and disseminate the latest word on neologisms and decisions about competing terms.

But the possibility of large-scale storage and manipulation by the computer did not yet mean that the benefits of the data that it was processing were available to the average individual translator. There were still many hurdles to overcome in the mechanics of input and output. Because of these bottlenecks, for a long time there were basically only two choices, either human-generated translation, still with little help from machines, or else machine translation with little human interaction. In other words, the options were still concentrated at the two ends of the human-machine continuum, and at each end the situation was fraught with inefficiencies. If the translation was generated by a person, most of the steps were manual and traditional except for use of the dictating machine. A notable exception, in a few translation services, was the possibility of consulting a lexical database. But this procedure could be complicated and result in frustration. On the other hand, for translations generated by the computer, there were still the almost insurmountable

problems of input and of reprocessing the result--problems for which no new solutions had appeared since the 1960s.

This whole picture was to change, however, at the beginning of the 1980s. By that time increasing miniaturization and personalization of the computer had brought the widespread use of word processing technology and with it a number of major contributions to the translation process. For human-generated translation it represented a quantum advance in outputting human thought, with many creative possibilities for reducing keystrokes, and it made for more efficient preparation of records for lexical databases, including individualized ones. For machine translation, in turn, it brought the routine availability of text in machine-readable form and also an easy and effective capability for the postediting of output.

So, as it can be seen, the state of affairs is always in flux--first there is the need, then the technology meets it part way. In time we learn the strengths and disadvantages of the latest innovations, and then we move on to a reformulation of the need. This cycle is tempered by increased intellectual understanding of the translation process itself and by personal and social adaptation to the changing world in which we work.

3. The Pan American Health Organization: Example of a Holistic Approach

Flexible response to the technology has been the leitmotif in our implementation of machine translation at the Pan American Health Organization over the past six years (Vasconcellos 1985, Vasconcellos and León 1985).

The Pan American Health Organization (PAHO), Regional Office for the Americas of the World Health Organization, entered the machine translation picture in 1976, just at the threshold of the linguistic and technological advances that were to make MT a more feasible concept. Since that date PAHO has developed two in-house mainframe systems using the Organization's own resources. The first system to be undertaken was SPANAM,[1] which has been translating from Spanish into English since early 1980 (sample output in Fig. 1). In the course of generating some 3 million words of production text, it has undergone a number of adaptations in response to what we have learned as we have been implementing the system--and also to changing needs and circumstances. This experience gave us the capability of developing an even better and more sophisticated system from English into Spanish, ENGSPAN,[2] which in the past year has already produced more than half a million words (sample output in Fig. 2). This activity had partial support from the U.S. Agency for International Development (AID).[3]

The texts to be translated come primarily from the corpus of documents routinely prepared on the Wang word processor for other purposes. In some cases, text can also be input to the Wang by means of optical character recognition. We now have in our shop a DEST multilingual model, Turbofont 223, which reads, directly into the Wang system, five of the popular typescript faces in Spanish and French as well as in English. Typeset documents cannot be read, nor can some faces of typescript. Of course, every character that the OCR misses results in a not-found word for the MT dictionaries and their output. There are also input problems with the documents prepared directly on the Wang. Some of them contain a high proportion of typographical errors, and others have to be reformatted. So we have had to face the fact that there is a difference between the ideal vs. the real availability of text for MT--and we are coming to grips with it.

The word-processing documents are telecommunicated to the IBM mainframe (currently an IBM 4381), where they are translated and returned to the Wang for postediting on-screen.

Postediting is facilitated by a series of customized macros at the level of the word processor which are designed to deal with pragmatic distinctions that the systems cannot handle, and the postediting process itself is the subject of ongoing linguistic analysis (Vasconcellos 1986). Today all our translators who are not revisers have postediting included in their job descriptions. As they work, they jot down suggestions for the dictionaries on a side-by-side printout, and then later they enter the appropriate updates themselves. The translators also provide feedback to the computational linguists for improvements to the algorithm, and they sometimes suggest operational enhancements as well.

SPANAM's dictionaries have some 61,000 source entries (94 percent base forms, 6 percent full forms), and ENGSPAN's have about 45,000, with 47,000 in the target (statistics as of December 1985). Not-found words are rare--less than 1 percent in either case if we do not include typographical errors in the input, repeated occurrences of the same not-found word, or alphanumeric combinations that do not affect the text. Still, there is a continuing need for work on the dictionaries, both to refine and deepen the coding of existing entries and also to add idioms which will trigger variant translations that are specific for given contexts.

In order to save repetitious research, approved and reliable terms are specially marked in the output. The criteria for these markings come from internationally approved sources. We also have a database, WHOTERM, which is limited to technical terminology for certain biomedical fields. It resides on the Wang. Terms that are in WHOTERM are flagged in the machine output as well, but the mark is different so that the translator will know that a complete terminological record is available on the Wang station itself. These two sets of flags amount to an automatic system for retrieving technical terminology (to the extent that we can vouch for it) in the place where it occurs in the text. This obviates some of the frustrations that are ordinarily inherent in the consultation of lexical databases.

4. Future Directions

At PAHO we are aware of the need for advancement on several fronts. With the technology that we now have, we will be working on a number of tasks. For both SPANAM and ENGSPAN, we want to introduce more flagging of technical terminology, and we also want to continue to add idioms and variant translations in the dictionaries and microglossaries--especially, at this time, in the field of agriculture. SPANAM is soon to be the subject of more sophisticated analysis and synthesis, bringing it to the level of ENGSPAN, which will also undergo further enhancement. There will be continued linguistic analysis of the postediting process. Word processing can be further maximized both by providing the translators with specialized training in advanced functions and by continuing to develop the power of our macros. In the near future we also plan to port ENGSPAN to a microcomputer and to develop an interactive on-line program for updating the dictionaries.

On the larger horizon, the technology in general is moving toward advances that will make it possible for individual translators to have access to large, centralized lexical databases. Also, on-line access to the translator's "shoebox" file of

special terms is becoming more generalized. Windowing technology is making it possible to view different files at the same time: the input text, the output text, and files from different dictionaries.

It can be expected that input and output will continue to become easier with improvements both in word processing and optical character recognition. More sophisticated cursor manipulation will make it possible to speed up postediting and also retrieval from other files.

And most important, all these advantages will be coming to us in small packages that will fit on our desktops and which even the free-lance translator can afford.

At the social level, it is already happening that the widespread use of this technology is giving it power and the impetus to grow. In particular, the increased use of MT by professional translators will lead to the fine-tuning of postediting techniques. And to changes in attitude about translated text and its purposes. With faster turnaround, and greater flexibility regarding the quality of output, it is safe to say that the demand for translation will increase substantially--as indeed it must, if we are to respond to the need for cross-language communication in our modern world.

NOTES

[1]SPANAM and ENGSPAN are trademarks of the Pan American Health Organization.

[2]The computational and linguistic development of ENGSPAN has been carried out by Marjorie Leόn, senior computational linguist on the project, and Lee Ann Schwartz, computational linguist.

[3]Grant DPE-5542-G-SS-3048-00 awarded to the Pan American Health Organization under letter dated 3 August 1983.

REFERENCES

Locke, W.N., and A.D. Booth, eds. (1955) Machine Translation of Languages. New York: Wiley.

Vasconcellos, M. (1985) 'Management of the Machine Translation Environment.' In V. Lawson, ed. Tools for the Trade. London: Aslib. 115-129.

Vasconcellos, M. (1986) 'Functional Considerations in the Postediting of Machine-Translated Output: Dealing with V(S)O versus SVO.' Computers and Translation, 1, 1. 3-17.

Vasconcellos, M., and M. Leόn. (1985) 'SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization.' Computational Linguistics 11, 2/3. 122-136.

Zarechnak, M. (1979) 'The History of Machine Translation.' In B. Henisz-Dostert, R. Ross Macdonald, and M. Zarechnak, Machine Translation. The Hague: Mouton. 3-87.