

MACHINE TRANSLATION
AT THE PAN AMERICAN HEALTH ORGANIZATION:
A REVIEW OF HIGHLIGHTS AND INSIGHTS

Muriel Vasconcellos
Pan American Health Organization

0. Introduction

SPANAM, working from Spanish into English, has been providing machine translation to internal users at the Pan American Health Organization (PAHO) since early 1980. Operations are done in batch mode. The vocabulary and syntax of the input are entirely free, and the text is not preedited at any point. According to the categories of Lawson (1982), it qualifies as a "try-anything"-type system. As of the end of September 1983, a total of 1,350,366 words had been machine-translated for 62 users under 425 separate work orders. The service reaches beyond headquarters in Washington, D.C., to include programs in the field and at the World Health Organization in Geneva.

The present report will review some of the major highlights in the history of this activity, bringing out a few of the lessons learned and insights gained along the way; it will summarize its current status; and it will mention some improvements that are scheduled for the future. The project's evolution is best understood by bearing in mind that for the past three years there has been combined effort along multiple fronts: production for users, terminology work, dictionary-building, enhancement of the current translation program for Spanish into English (SPANAM), and, especially in the past year, development of machine translation in the other direction, i.e. English into Spanish (ENGSPAN). We believe very strongly that the project's viability comes from the opportunity we have had to work on different aspects of the system at the same time.

1. Background

1.1 Early development of MT at PAHO

PAHO is the specialized agency that deals with health matters within the Inter-American system. Its secretariat also serves as the World Health Organization's regional office for the Western Hemisphere. The official languages are English, French, Portuguese, and Spanish. In terms of volume of translation required, over the years the

pattern has been that approximately 55% of all translation is into Spanish, 34% into English, 10% into Portuguese, and 1% into French. In the mid-1970's PAHO began to think about MT as a tool for dealing more efficiently with its multilingual needs.

Following a feasibility study, a team of consultants was contracted in 1976 to begin work on an in-house MT system. The approach decided on was originally quite similar to that developed at Georgetown University in the late 1950's and early 1960's (GAT--Zarechnak 1979).

For the PAHO system, it was agreed from the start that postediting would be part of the process. Preediting, on the other hand, was never seriously contemplated; the Administration wanted a system that would articulate with the routine flow of text within the secretariat. Also, the system was to run on the regular mainframe computer (then an IBM 360 with a disk operating system) without taking up much space in core or impairing any other operations that might be running at the same time. These were the main considerations in mind when work began on the project.

Since at first the major concern was with setting up the architecture itself of the system, including its extensive supporting software, the decision was made to start with the less difficult of the language combinations that would be needed, namely Spanish into English. Over the next three years, from 1976 to 1979, the basic program for Spanish-to-English translation was written, a full range of supporting software was developed, and the dictionaries were built to a level of some 48,000 source entries with target glosses as appropriate.

The year 1979 brought a turning-point for machine translation at PAHO. Momentum was gained on two fronts. First, a full-time computational linguist was assigned to the project's regular staff. And second, an interface was established between the IBM mainframe computer, where the programs and dictionaries reside, and the Organization's regular word processing system (at the time a Wang WPS 30). This meant that it was no longer necessary to have a text specially keyboarded for purposes of machine translation. Before then, any text to be translated had to be input to the computer on punched cards. This slowed down the process considerably and precluded any serious thought of production for actual users. By the end of 1979, however, a conversion program had been written which could successfully cope with

(2)

any text prepared in a normal format using standard typing conventions. From then on, any Spanish text on the Wang system, regardless of the purpose for which it had originally been entered, was available for machine translation.

1.2 SPANAM Becomes Viable

In the beginning, of course, production was not the full-scale enterprise that it is today. News of the system's availability was spread by word of mouth and through an active program of briefings and demonstrations (demonstrations, by the way, always done on random text). The system was never imposed on users. Where we felt that a particular application was especially appropriate and that we had the capacity to deal with it, we would establish contact with that office and offer the services of SPANAM. On the whole, however, the users have come to us.

Our first major project was the 1981 edition of the Organization's biennial budget document, a large volume more than half of which is submitted in Spanish from different offices in the field. This application was felt to be particularly appropriate 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 (Table 1). In our evaluation, the first step was to add up all the costs--postediting (by a junior translator hired on contract), supervision, operation of the system, and final proofreading and adjustments, as well as a hypothetical charge for computer time. We looked at both the dollar cost and the total investment in terms of staff-days. The expenditures came to US\$3,218 for 101,296 words of translation, and time spent on the project amounted to a total of 36 staff-days. Then came the comparison: had the same amount of text been translated and processed in the traditional way, the corresponding figures would have been \$8,296.18 and 65.75 staff-days. There was a monetary saving of \$5,078.48, or 61%, and the staff-days were reduced by 29.5, or 45%. The users were greatly satisfied with the experience and called on us again in 1983 to do the document for the current biennium.

In the two years between, SPANAM translated texts in a wide range of fields and for varying purposes. Particularly, we have been asked to translate documentation for meetings, which is routinely prepared on the word processor in both

Spanish and English. Other types of text have included international agreements, reports of short-term consultants, summaries and protocols for the international data bases on cancer, scientific abstracts, volumes of proceedings, training manuals, lists of supplies, and material for regularly recurring publications such as the news bulletin of the U.S.-Mexico Border Health Association, the Epidemiological Bulletin, and a newsletter entitled Disaster Preparedness in the Americas.

1.3 Optical Character Recognition

In spring 1981 the capability of preparing machine-readable text was increased through an interface between the word-processing system and an optical character reader (OCR) on the premises, Compuscan's model Alphaword II. This equipment had been installed initially for the transmission of Telex messages. In theory, the interface with the word processor permitted all IBM Selectric typewriters to become input devices to the word processor. At first it was expected that this innovation would provide a large source of machine-readable text for MT. However, use of the OCR for word processing (i.e. non-Telex) applications was subject to a series of impediments from the start, with the result that it did not become generalized, and now it is only a rare option. Since this capability is of such great interest for machine translation, some of the deterring factors will be mentioned. First there was the requirement for special preparation of the text, including use of the OCR-B type face, which encountered resistance in the secretariat. Moreover, the product's software did not read diacritics, which meant that Spanish texts with accents in their normal position on top of the letter could not be processed. PAHO developed a program for the conversion of two input characters into a single accented character on the Wang, but there were persistent difficulties in mounting it. Moreover, there was the delicate nature of the OCR equipment, coupled with the requirement that the operator be present at all times in order to intervene whenever a character could not be read. And finally, in our particular environment there were scheduling problems because of the conflict with Telex transmissions as well as the need to cope with a switching mechanism for sharing the TC line with another Wang device. These factors combined to make for a "critical mass" of resistance to its use. It is hoped that a second OCR will soon be purchased which will be used exclusively for input to the word-processing

system. It will need to be sturdier and operator-independent, have multilingual software, and not require special preparation of the text. In the meantime, OCR is still an option at PAHO, and we do machine-translate texts from the field which have been prepared using the OCR-B element.

2. Current Status

1.2 Outline of the System: SPANAM

All machine translation at PAHO is run in batch mode. For normal production, the configuration is batch via remote job entry (RJE)--i.e. from the word processor (now a Wang OIS 140) to the IBM mainframe (a model 4341 running on DOS/VSE) and back to the word processor. It is also possible to send files on tape directly to the computer and, for test or demonstration purposes, to key in a text at the computer terminal.

Turnaround using the Wang in this mode--including transmission time plus clock time while the translation is run on the IBM--is quite rapid. Over the years the clock time has improved steadily even though the program and the dictionary record have become increasingly complex (Table 2). A major jump in 1982 came from a switchover from ISAM to VSAM. More recent improvements have been due to specific efforts to make the lookup faster. The statistics in CPU time, which have been available since 1981, show a range of from 2,600 to 3,200 words a minute. The foregoing figures are equivalent to 42,000 words (168 pages) per hour for clock time and 192,000 words (768 pages) per hour for CPU time. With turnaround of this order, we are quite satisfied with our batch operation, and there is very little incentive for us to experiment with an interactive mode for production translation.

When the job is received at the IBM mainframe, the first thing that happens is that a conversion program interprets the Wang characters and changes them to a representation which matches the representation used in the dictionaries. Mainly, this entails distinguishing facultative punctuation from that which is pertinent to a given dictionary entry. For example, periods after abbreviations should be read as part of the word, whereas sentence punctuation should not. Hyphens used to break words at the end of a line are deleted. Also, a letter plus orthographic accent is interpreted as two characters. Once the conversion is completed, the text is stored on disk and the translation program itself is called

into operation.

The translation is done by sentences. The program picks up one sentence at a time, and within that sentence each word is looked up individually. There is no attempt to sort the text for lookup. The first step in the lookup is an initial check against a small dictionary of high-frequency words whose entire entries have been read into core. Then the rest of the source words are matched against key items, either full forms or stems, in the large Spanish source dictionary that resides permanently on disk. After that a second pass is made in order to identify idioms. When a match is made, whether of a single word or its idiom replacement, the corresponding entry is copied into a workspace where operations are to be performed on the sentence. The English target dictionary, which is also kept on disk, but in a separate place, is not consulted until much later, just before the synthesis.

The grammatical work of the program is performed through a series of modules. The analysis of the source text focuses on contrastive situations that are encountered particularly in the transfer from a Romance language to English--there is no independent "interlingua." A series of modules deal with: the disambiguation of part-of-speech homographs, prepositional government, interpretation of pronouns and articles, and manipulation of the verb string. Within each of these modules local parsing routines provide the information needed in order to make the appropriate decisions. After these modules have been exercised, a set of patterns are introduced for the rearrangement of noun phrases. Once all these steps have been performed, the appropriate gloss with its accompanying codes are picked up from the main target dictionary or its microglossaries, and the appropriate target forms are synthesized. A few other minor routines are applied to the resulting text, and this then reconverted and transmitted back to the Wang.

As for space requirements, the program uses about 210 K of core, not including VSAM overhead, and the resident dictionaries take up a total of 17.4 megabytes on disk. The workspace and patterns occupy another 1.1 MB, and there are also a few additional files and libraries for which disk space is required. The rest of the allocated area is for the systems being developed from English into Spanish and Portuguese (Table 3).

2.2 The Dictionaries

Initially, SPANAM's dictionary development was done according to the Georgetown methodology--i.e. using twin-text concordances of running text already extant in the two languages. For this purpose, 40,000 words of text were chosen from different PAHO publications, some of them technical and others general. The resulting corpus served as the basis for the preparation of hand-coded entries specifically addressed to texts of the kind and in the subject areas that PAHO deals with. However, once the system became operational, the corpus was largely abandoned and focus was shifted to actual production.

Production text has clear advantages as a basis for research and development: mainly, it is current, and it prepares the system so that it can provide increasingly better service for the user. The twin texts that had been used, on the other hand, tended to have a sort of "sanitized" prose--often a mixture of original Spanish and original English--and always, of course, one side was a translation rather than naturally occurring language. Today the large Spanish stem dictionary stands at 56,000 entries. Of these, about 16,000 are "analytical" entries--i.e. deeply hand-coded. The growth of the dictionaries is traced in Table 4.

In both the SPANAM and ENGSPAN, stems or canonical forms are entered in preference to full forms whenever possible. This means that nouns are in the singular, adjectives are in the masculine singular, and verbs are listed without any inflectional endings. Full forms are retained for words that are part-of-speech homographs, for nouns and adjectives that participate in certain types of idioms, and for a few of the most highly irregular verbs. In SPANAM, full forms represent about 6% of the total source dictionary. About 26,000 of the entries correspond to general vocabulary, and 30,000--more than half--are specialized terms in the fields that PAHO works in. This latter is the side of the dictionary that grows the fastest. New entries are constantly being added based on the results of production jobs. The updates to be made are noted in the course of postediting.

With a dictionary of this size, half of it focused on user-specific terminology, the incidence of not-found words is minimal. Most of the not-found words are proper names, abbreviations or acronyms, or nonce-formations. Nevertheless, the program allows for a certain amount of gap analysis, so that failure to find a word does not disrupt the

work of the grammatical modules. The not-found words are simply copied into the translation in the same form that they had in the original text. They are also flagged in the source text. In the absence of any morphology that the program may be looking for, the default assumption is that the not-found words are nouns. This works for the proper names, acronyms, Latin names of species, disease entities, and, in general, much of the more rarely used terminology in the biomedical fields.

Naturally there are still some problems. Even though a word is found, it could be a homograph for which not all the possible alternatives have been provided for in the dictionary. And, of course, there is the question of polysemous forms--for us, words of the same part of speech which, by extension of their semantic field, take on different meanings in different contexts. These are dealt with through microglossaries and idioms. There are now several specialized microglossaries that contain variant translations corresponding to particular disciplines. Different users supply their preferred vocabulary, and when a word or term conflicts with a gloss in the main dictionary which we would prefer to maintain, we enter the new term in the microglossary so that it will be elicited only when translations are run in the particular subfield. An example might be medios de cultivo, which can mean either 'means of cultivation' in a text on agriculture or 'culture media' in a text on laboratory procedures. Figure 1 shows test sentences for which this feature was specified.

In addition, idiomatic treatment may be required, even though the words have been found and disambiguated correctly--either to disambiguate the different uses of a single word or to assign a new meaning to an entire construction. The maximum potential length of an idiom is 25 words. Currently there are about 3,000 idioms in the Spanish source dictionary (included in the total of 56,000 entries). In the future we are planning to incorporate into SPANAM different types of idioms that have been developed for ENGSPAN. This flexibility will enable us to introduce a larger number of idioms. We are aware that idioms contribute importantly to the intelligibility of the output.

Another dictionary-based feature allows the user to request that terminology coded as "reliable" be specially marked in the output. This is also shown in Figure 1. Marks appear at the beginning and end of a word or phrase that has been assigned

a reliability code of 3 or higher on a scale of 0 to 5. The codes are assigned according to strict standards that have been worked out by common agreement with WHO Headquarters in Geneva and the participating WHO regional offices. Naturally, it is possible that a particular gloss could be reliable for one context but not for another; however, since the posteditors or reviewers are expected to be trained and experienced in translation, they will be able to judge whether the context is appropriate, and the fact of knowing that a term has been researched and agreed on can often be valuable information for them. An example of standardized terminology is the list of nonproprietary names of drugs: the Spanish and English versions were read into the SPANAM dictionaries directly from a copy of the tape that is used in Geneva to maintain the master international list. When these names appear in the translation with the reliability mark, the posteditor knows that they do not need to be checked any further; contrariwise, a drug name that does not appear with the mark will be either a trade name or a misspelling in the source text.

Before long it will be possible to consult a new data base, WHOTERM (Ahlroth and Lowe 1983), which will be resident on the word processor and will provide definitions and other data for terms that bear appropriate flags in the MT output. This large set of files of technical terminology is being developed by WHO in Geneva and will soon be installed at PAHO in Washington.

For purposes of consultation, random entries can be retrieved from the SPANAM and ENGSPAN dictionaries using either the word processor as RJE or the computer terminal. Updating of the dictionaries can also be done in both modes. In all cases the software is user-friendly.

With updating, many defaults are entered automatically by the program unless the user specifies otherwise, and the descriptors and their codes are mnemonic. Entries can be made quickly--even analytical entries with deep coding. The list of fields for possible codes in the dictionary entry continues to increase as enhancements are added to the program. An update can consist of any number of entries and any combination of functions. It is submitted as a batch job regardless of the input mode. The result can be verified in different ways: (1) by reviewing the printout produced by the update program, which gives diagnostics when an item is not entered; (2) by requesting a copy, again either on the computer terminal or the Wang, of

the particular records that are affected; and/or (3) by running test sentences. Another input device for dictionary updates is the OCR: in 1982, when we had a group of five student interns from Georgetown University, the updates were prepared on IBM Selectric typewriters using the OCR format. This made it possible for their work to be reviewed by the computational linguist before it was incorporated in the dictionaries.

2.3 Production

The use of SPANAM has risen steadily:

| | <u>Words</u> | <u>Pages</u> |
|----------|----------------|--------------|
| 1980 | 90,153 | 361 |
| 1981 | 325,333 | 1,301 |
| 1982 | 449,013 | 1,796 |
| Sep 1983 | <u>485,867</u> | <u>1,942</u> |
| Total | 1,350,366 | 5,400 |

As mentioned earlier, it was always expected that the output would have to be postedited. Needs for postediting cover a wide range, depending on the use to which the text is to be put: there are requests for raw output and others for a polished product for publication, with midpoints between the extremes. The specific requirements are discussed with the user at the time the job is brought in, and assistance is given in filling out the job request form (Figure 2).

The degree of postediting also varies depending on the quality of the machine's product. Quality of the raw output is governed to quite a large extent by the amount of dictionary work that has already been done in the particular subject field. The genre of discourse is also an important factor. The system turns out its best performance on long technical documents and reports. Speeches sometimes translate surprisingly well, other times not so smoothly. We do letters and memoranda, although this type of application is not encouraged, and we have even done scripts for educational films. And finally, we have found that another significant factor is the variation in syntactic and presentational styles between different authors, regardless of the subject area or the genre.

Most of the postediting is done by one of our own staff working on-screen. Sometimes, however, we have delivered raw, or nearly raw, output to

editors or technical writers who have wanted a rough draft to work from. The Director's Annual Report has been done in this way. We also delivered an almost-raw job to a seasoned professional translator working on contract for the Organization, and his reaction was highly enthusiastic. However, our experience has been that it is considerably easier to postedit on-screen and to deal with users who will be doing the same--mainly because we do not have secretarial staff and the entry of hand-written corrections from hard copy constitutes an extra step that we would prefer to avoid.

The average output is about 6,500 words a day for one posteditor, who has other duties as well, such as dealing with the users, transmitting texts for translation, tracking down terminology, keeping records and statistics, and maintaining the diskette storage system. Still, this figure is about triple the volume usually considered average for traditional translation in the international organizations--namely 2,000 words a day. Moreover, a final machine-readable copy is produced. Occasionally, as on several days last August and September, the volume has reached 10,000 words a day. Thus it is conservative to estimate that the gain in terms of time and cost is at least three-fold.

Output is delivered either on diskette or by informing the user that the translation is available on the word processing system. The document bears the words MACHINE TRANSLATION on each header page, and the last page announces that THE FOREGOING TEXT IS A POSTEDITED MACHINE TRANSLATION.

Since not-found words are flagged by the system, we can also provide spelling-check service for our users. When the Spanish version is to be distributed and it has misspelled words, we supply the user with a list of these words so that they can be corrected.

Many of the users report that they are pleasantly surprised at the prompt turnaround and are also satisfied with the translation.

The success of SPANAM is owed at least as much to skillful and rapid postediting as it is to quality of the machine output. This latter factor makes all the difference in whether a product is usable or not. There are special skills to be acquired which greatly enhance the effectiveness of the postediting: one learns the difficulties to expect, how to correct them the quickest way

possible on the word processor, and how to fix a text without extensively rearranging it. Not necessarily is there a direct correlation between quality of the machine output and the extent of postediting required. The amount of postediting will depend on the needs of the user and, even more importantly, on the ability of the posteditor to make few but strategic changes. In our environment we have found that time spent on postediting is a more meaningful measure than the number of errors that the system generates--and that this time can be cut dramatically by skilled use of features on the word processor such features as global search, global replace, selective replace, word-switching, etc. Even such a mechanical factor as the capability of quickly positioning the cursor quickly can make a significant difference. SPANAM has, in addition to these features, a series of string manipulations that are specifically designed for dealing with English MT output--for example, use of a single glossary key to search for and delete the, of, or there; to delete an unwanted comma or insert one before and; to change that to who, that to which, or its to their, etc. This capability is constantly being upgraded, as we realize it is important not only for speeding up the work but also for reducing the annoyance factor for the posteditor.

2.4 Development of ENGSPAN

In view of the growing demand for information in the Spanish-speaking countries of the Americas, especially from machine-readable data bases, as well as the current heavy load of human translation, work began about a year ago on the system from English into Spanish, ENGSPAN. We are happy to say that this activity recently received a supporting grant from the U.S. Agency for International Development (AID), which will cover the period August 1983 through July 1985.

At the start of the grant period, the English source dictionary had approximately 40,000 entries, most of them already tied to appropriate equivalents in the Spanish target dictionary. These two ENGSPAN dictionaries had been created by reversing the SPANAM dictionaries and culling out duplicate or clearly inappropriate glosses--about 26%. The algorithm included: (1) a lemmatization module, (2) procedures for looking up single words and phrases, (3) routines for resolving a limited number of homograph types, (4) a module for recognizing and synthesizing simple noun phrases, and (5) a complete procedure for the synthesis of

inflected Spanish verb forms in all tenses and moods of the 1st and 3rd persons singular and plural. In short, the architecture was in place which made it possible to produce machine output consisting of Spanish words.

Since the analysis of English requires more extensive parsing, and hence more exhaustive coding, than that of Spanish, the dictionary record has gradually been revised and expanded.

There is currently a working corpus of 50,000 running words made up of texts in the field of public health. Test translations are already giving promising results on a 9,000-word segment. A seven-phase strategy has been adopted for the accelerated development of ENGSPAN under the grant from AID, and work is well under way on the first of these phases--namely analysis and disambiguation of the English noun phrase. Parsing is now possible for many types of ambiguous noun phrases and sentences. Already as part of this phase, semantic coding is being introduced.

3. Agenda for the Future

3.1 Improvements to SPANAM

As advances are made in ENGSPAN, it is planned to capture any improvements that might have relevance for SPANAM. In particular, we look forward to the possibility of having expanded parsing strategies that deal with embedding, gap analysis, semantic units whose components can be analyzed for purposes of parsing, and dictionary-based lexical routines capable of handling discontinuous elements and classes of elements. These changes will involve extensive deep coding of existing dictionary entries as well as the addition of new entries.

Correlation SPANAM with WHOTERM, so that WHOTERM entries are flagged in the output, is another activity that is planned.

3.2 The Agenda for ENGSPAN

The program for the accelerated development of ENGSPAN, as approved by AID, calls for seven phases of activity in connection with the algorithm and five phases in relation to the dictionaries.

Work on the algorithm will involve, basically, the

development and introduction of new codes for dealing with noun phrases, the verb string, prepositional phrases, adverbs, and nonfinite verb forms. At the end of the first year intensive study will begin on clause-level parsing, clause relationships, and special problems of discourse analysis. We are not striving for perfection; we plan to attack the problems that are statistically most frequent under each of these headings. Our goal for number-person-gender agreement is 60% by the end of the first year and 80% by the end of the second year.

Dictionary-building will be undertaken in tandem with the foregoing development of the algorithm. The noun-phrase analysis will affect the codes of nouns, determiners, numeratives, and adjectives, and the verb string will trigger features of selectional restriction and strict subcategorization. Discontinuous idioms will be introduced, as described above. And finally, attention will be given to the selection of specialized terminological glosses in the target area of discourse.

During the last six months of the project an evaluative study of the system software will look into the possibility of its being adapted to a mini- or microcomputer. Our goal is for ENGSPAN to function as part of the system of health information in the countries of the Americas.

When ENGSPAN is developed to an operational level, we hope and expect that it will be of valuable service to the Organization in fulfilling its mission to share information and technology with its member countries. Because our larger and long-term objective is to convey information fast, at low cost, and in a form and volume designed to reach strategic readerships and provide them with benefits, in the form of knowledge, that might not have been available to them otherwise.

ACKNOWLEDGMENTS

Many individuals have been involved in the development of SPANAM, and it would be impossible to list them all. Special recognition, however, must go to Marjorie Leon, the staff computational linguist since 1979, and to Allen B. Tucker, Jr., consultant to the project for five years, who was responsible, among other things, for giving SPANAM its user-friendly orientation. From the computational linguistics program at Georgetown University the project has had consulting support from Michael Zarechnak, Leonard Shaefer, and R.

Ross Macdonald, head of that program, whose death on 16 June 1983 was a very sad loss.

In PAHO, the project is indebted to Luis Larrea Alba, Jr., Chief of General Services, for his support throughout the years, and to Dr. Charles L. Williams, Jr., Deputy Director of the Organization until his retirement in 1979, whose vision was responsible for bringing machine translation to PAHO.

REFERENCES

Ahlroth, E., and D. Armstrong Lowe. The WHO Terminology Information System: Interim Report. [Geneva: World Health Organization, 1983.] HBI/ISS/83.1. (offset)

Lawson, Veronica. Machine translation and people. In her: Practical Experience of Machine Translation. Amsterdam, New York: North-Holland, 1982. p. 5.

Tucker, Allen B., Muriel Vasconcellos, and Marjorie Leon. PAHO Machine Translation System: Introduction and Users' Manual. Washington, D.C.: Pan American Health Organization, July 1980.

Zarechnak, Michael. The history of machine translation. In: Machine Translation, by B. Henisz-Dostert, R. Ross Macdonald, and Michael Zarechnak. The Hague: Mouton, 1979. pp. 29-30,32,134ff.

Table 1

Machine translation of PAHO budget document, January 1981.

| | US\$ amount | Staff-days |
|---|-----------------|-------------|
| Postediting of 101,296 words by junior translator on contract, 200 hr at \$8.00 | 1,600.00 | 25.0 |
| Supervision, 10 hr at \$20.73 | 207.30 | 1.25 |
| Submission, retrieval, and formatting of text, 40 hr at \$16.81 | 672.40 | 5.0 |
| Proofreading and adjustments for style, 40 hr at \$10.95 | 438.00 | 5.0 |
| Hypothetical charge for machine time, \$580/CPU hr | <u>300.00</u> | <u>-</u> |
| Total: | 3,217.70 | 36.25 |
| The same 101,296 words translated by the procedures used in the past would have entailed: | | |
| Contract translation at \$55.00 per 1,000 words | 5,571.28 | 33.0 |
| Processing of translation, 10 hr at \$9.95 | 99.50 | 1.25 |
| Cross-checking of translation against original text, 112 hr at \$10.95 | 1,226.40 | 14.0 |
| Keying of translation onto Wang, 80 hr at \$9.05 | 724.00 | 10.0 |
| Final proofreading and corrections, 60 hr at \$10.95 | <u>675.00</u> | <u>7.5</u> |
| Total: | 8,296.18 | 65.75 |
| SAVINGS EFFECTED: | <u>5,078.48</u> | <u>29.5</u> |

Table 2
Translation speeds, SPANAM, 1979-1983.

| Year | Best clock time | | | Average CPU time | |
|------|-----------------|--------|---------|------------------|---------|
| | wpm | wph | pages/h | wpm | wph |
| 1979 | 160 | 9,600 | 38 | Not available | |
| 1980 | 176 | 10,560 | 42 | Not available | |
| 1981 | 192 | 11,520 | 46 | 3,184 | 191,000 |
| 1982 | 580* | 34,800 | 139 | 2,600 | 156,000 |
| 1983 | 700 | 42,000 | 168 | 2,880 | 172,800 |

*Reflects change to VSAM lookup.

Table 3

Space requirements, PAHO Machine Translation System, December 1982.

| <u>Core utilized for translation run:</u> | | | <u>Files:</u> | |
|---|-----------------|---------------------|---------------------------|-------------------------------|
| | | | <u>Current</u> | <u>Projected</u> |
| SPANAM | Size parameter | 210 K | | |
| | System overhead | 180 K | | |
| ENGSPAN | Size parameter | 220 K | | |
| | System overhead | 180 K | | |
| <u>Work space on disk:</u> | | | <u>VSAM:</u> | |
| MIS text | 120 tracks | | | |
| | | <u>Total 1.0 MB</u> | English source dictionary | 6.5 MB 7.5 MB |
| | | | Spanish source dictionary | 8.9 MB 9.5 MB |
| | | | English target dictionary | 8.5 MB 9.5 MB |
| | | | Spanish target dictionary | 6.7 MB 7.7 MB |
| | | | <u>Other:</u> | |
| | | | English patterns | 0.1 MB 0.1 MB |
| | | | Spanish patterns | 0.1 MB 0.1 MB |
| | | | POURCE test dictionary | 0.6 MB 0.6 MB |
| | | | PORGET test dictionary | 0.6 MB 0.6 MB |
| | | | ESOURCE test dictionary | 0.6 MB 0.6 MB |
| | | | PTARGE test dictionary | <u>0.6 MB</u> <u>0.6 MB</u> |
| | | | <u>Total:</u> | <u>33.2 MB</u> <u>36.8 MB</u> |

1 track is about 8,000 characters (8 K).
 1 cylinder is 96 K; 1 megabyte (MB) has 10.4 cylinders.
 1 MB corresponds to about 400 pages of running text.

Table 4
 Size of dictionaries, PAHO Machine Translation System,
 1976-1983.

| Year | SPANAM | | ENGSPAN | |
|------|---------|---------------------|---------------------|---------|
| | Spanish | English | English | Spanish |
| 1976 | 4,000 | 3,500 | | |
| 1977 | 7,836 | 7,341 | | |
| 1978 | 38,506 | 38,376 | | |
| 1979 | 48,289 | 53,303 | | |
| 1980 | 50,912 | 55,792 | | |
| 1981 | 53,785 | 51,187 ¹ | 44,411 ² | 44,998 |
| 1982 | 54,383 | 52,223 | 40,107 | 41,358 |
| 1983 | 56,247 | 53,326 ³ | 40,772 | 42,116 |

¹7,000 unmatched target entries were deleted by a special-purpose program.

²Upon reversal of dictionaries, 4,500 duplicate source entries and corresponding target records were deleted by a special-purpose program after selection of the desired gloss.

³1,000 irregular verb forms were deleted by a special-purpose program.

Otras toxicidades

Adriamicina. depresión medular, vómitos, estomatitis,
flebitis por extravasación, erupción cutánea.

Ciclofosfamida: cistitis hemorrágica, alopecia,
disminución de la función gonadal, inmunosupresión.

Utilizaron varios medios de cultivo.

Other `toxic effects`

`Adriamycin`: `depressed bone marrow`, vomiting,
`stomatitis`, `phlebitis` by `extravasation`, skin
eruption.

`Cyclophosphamide`: hemorrhagic `cystitis`, `alopecia`,
reduction of the `gonadal` function, `immunosuppression`.

They utilized several *means of cultivation.

G2

Figure 1. Examples of reliability coding and glossary selection.

The test sentences above demonstrate two optional features of SPANAM. (1) If the user so requests, words or phrases having a reliability code of 3 or above can be flagged with a special symbol. This tells the user that these terms have come from an authoritative source. The terms can, of course, be changed, as can their reliability code, on the basis of appropriate information. (2) Microglossaries make it possible to specify vocabulary from a given area of discourse or a dictionary provided by a particular user. In the example, the default translation of medios de cultivo is 'culture media' (biomedical terminology is default for SPANAM), but specification of Glossary 2 (Agriculture) produces, instead, 'means of cultivation.'

PR _____

REQUISITION FOR MACHINE TRANSLATION
(SPANISH-ENGLISH)

| | | | | |
|-------------------------------|---|--------------------------------|---|---------------------|
| Office code | Person responsible | Phone | Room | Date of requisition |
| Title or description of text: | | | | |
| Current form of text: | | | | |
| <input type="checkbox"/> | Wang | Document no. _____ | Disk no. _____ | |
| <input type="checkbox"/> | OCR | (Recopied by ATP into _____ M) | | |
| <input type="checkbox"/> | Other medium | Identification no. _____ | | |
| Date required: | | Comments: | | |
| Intended use of translation: | | | | |
| <input type="checkbox"/> | Information only | | | |
| <input type="checkbox"/> | Draft to be postedited by requesting office | | | |
| <input type="checkbox"/> | Internal distribution | | | |
| <input type="checkbox"/> | Other Explanation: _____ | | | |
| Format of output: | | Medium of output: | | |
| <input type="checkbox"/> | English only, double-spaced | <input type="checkbox"/> | On OIS/140 system disk | |
| <input type="checkbox"/> | English only, single-spaced | <input type="checkbox"/> | Hard copy | |
| <input type="checkbox"/> | Other: _____ | <input type="checkbox"/> | Diskette for other Wang system No. _____ | |

For ATP use only

| | |
|------------------|--|
| Date received: | Wang Document Nos. |
| | Side-by-side: _____ T English only: _____ M |
| Date dispatched: | Statistics |
| | No. of words: _____ Time: _____ |

Figure 2