MACHINE TRANSLATION AT THE

PAN AMERICAN HEALTH ORGANIZATION

Muriel Vasconcellos
Pan American Health Organization

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.

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.

The decision was made to start with 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 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

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.

2. Current Status

2.1 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 product 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.

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.

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 existing 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. Today the large Spanish stem dictionary stands at 56,000 entries. Of these, about 16,000 are "analytical" entries--i.e. deeply hand-coded.

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.

turally there are still some problems. Even
ough a word is found, it could be a
mograph for which not all the possible
ternatives have been provided for in the
ctionary. And, of course, there is the
estion of polysemous forms--for us, words of
e same part of speech which, by extension of
eir semantic field, take on different
anings in different contexts. These are
alt with through microglossaries and
ioms. There are now several specialized
croglossaries that contain variant
anslations corresponding to particular
sciplines. Different users supply their
eferred vocabulary, and when a word or term
nflicts with a gloss in the main dictionary
ich we would prefer to maintain, we enter
e new term in the microglossary so that it
ll be elicited only when translations are
n in the particular subfield. An example
ght be medios de cultivo, which can mean
ther 'means of cultivation' in a text on
riculture or 'culture media' in a text on
boratory procedures.

addition, idiomatic treatment may be
quired, even though the words have been
und and disambiguated correctly--either to
sambiguate the different uses of a single
rd or to assign a new meaning to an entire
nstruction. The maximum potential length of
idiom is 25 words. Currently there are
out 3,000 idioms in the Spanish source
ctionary (included in the total of 56,000
tries). In the future we are planning to
corporate into SPANAM different types of
ioms that have been developed for ENGSPAN.
is flexibility will enable us to introduce a
rge number of idioms. We are aware that
ioms contribute importantly to the
telligibility of the output.

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.

2.3 Production

The use of SPANAM has risen steadily:

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

The degree of postediting 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 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. 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 TEST IS A POSTEDITED MACHINE 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. SPANAM has 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 to 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 of 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. 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.

ACKNOWLEDGEMENTS

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 P. 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 budget document, OD169, January 1981.

All Spanish submissions received by ABU during the months December-February were machine-translated, postedited, and returned tc ABU within the scheduled period. A total of 101,296 words were processed. For the experimental period, it has been estimated that the following costs were incurred:

	$ amount	Man-days
Postediting by junior translator contract (does not include training in MT procedures) 200 hours at $8.00	1,600.00	25.0
Supervision by Coordinating Terminologist, 10 hours at $20.73	207.30	1.25
Submission, retrieval, and formating of text by Dictionary Officer, 40 hours at $16.81	672.40	5.0
Final proofreading and adjustments for style by G-7 staff, 40 hours at $10.95	438.00	5.0
If machine time were to have been charged at a commercial rate ($580/CPU hr), the cost would have been appr. $3/1,000 words	300.00	
	3,217.70	36.25

The same 101,296 words done according to the procedures used in the past would have entailed:

	$ amount	Man-days
Contract translation at $55/1,000 words	5,571.28	33.0
Processing of translation by ATS, G-6 staff, 10 hours at $9.95	99.50	1.25
Cross-checking of translation against original text by G-7 staff, 112 hours at $10.95	1,226.40	14.0
Keying of translation onto Wang, G-5 staff, 80 hours at $9.05	724.00	10.0
Final proofreading and corrections, G-7 staff, 60 hours at $10.95	675.00	7.5
	8,296.18	65.75
SAVINGS EFFECTED:	5,078.48	29.5

Table 2
Translation speeds, SPANAM, 1979-1983

	Best clock time			Average CPU time	
Year	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.

Core utilized for translation run:			Files:		
				Current	Projected
SPANAM	Size parameter	210 K			
	System overhead	180 K	VSAM:		
			English source dictionary	6.5 MB	7.5 MB
ENGSPAN	Size parameter	220 K	Spanish source dictionary	8.9 MB	9.5 MB
	System overhead	180 K	English target dictionary	8.5 MB	9.5 MB
			Spanish target dictionary	6.7 MB	7.7 MB
Work space on disk					
			Other:		
MTS text	120 tracks				
		Total 1.0 MB	English patterns	0.1 MB	0.1 MB
			Spanish patterns	0.1 MB	0.1 MB
Program libraries			POURCE test dictionary	0.6 MB	0.6 MB
			PORGET test dictionary	0.6 MB	0.6 MB
Librarian Master	120 tracks		ESOURCE test dictionary	0.6 MB	0.6 MB
Core Image Library	60 cylinders		PTARGE test dictionary	0.6 MB	0.6 MB
Relocatable Library	15 cylinders				
		Total 8.5 MB	Total:	33.2 MB	36.8 MB

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.

	SPANAM		ENGSPAN	
Year	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

[1] 7,000 unmatched target entries were deleted by a special-purpose program.

[2] Upon 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.

[3] 1,000 irregular verb forms were deleted by a special-purpose program.