

Management of the machine translation environment: interaction of functions at the Pan American Health Organization

Muriel Vasconcellos

Chief, Terminology and Machine Translation, Pan American Health Organization, Washington, DC, USA

Spanish-English machine translation at the Pan American Health Organization (WHO regional office) has been fully operational since early 1980. The environment supports, at the same time: production, terminology retrieval, dictionary and program maintenance, and advanced development of a new system from English into Spanish. The interaction of these activities strengthens all of them mutually.

INTRODUCTION

At the Pan American Health Organization (PAHO) we feel that a multifaceted working environment has contributed importantly to the progress of our work in machine translation. Our activity combines, at the same time, production for users, terminology work, dictionary development, enhancement of the current translation programme and development of a second system. Each of the components receives input and support from all the others. We are confident that this approach has been a major factor in the viability that we enjoy today.

EVOLUTION OF THE ENVIRONMENT

PAHO is the specialised international agency in the Americas that deals with public health, and as such it has a statutory role both within the Inter-American system and as part of the UN family, in which it serves as the regional office of the World Health Organization (WHO). The official languages are Spanish, English, Portuguese and French. The volume

of human translation over the past five years has averaged 57 per cent into Spanish, 32 per cent into English, 9.4 per cent into Portuguese, and 1.6 per cent into French.

In the mid-1970s the administrators at PAHO decided to look into machine translation as a means of reducing costs. Quantum advances in the speed, storage capacity, and efficiency of digital computers had made it seem reasonable to reconsider the possibility of mobilising them in the service of translation.

A mainframe computer, then an IBM 360 with a disk operating system, was already in place at PAHO. Based on the results of a feasibility study, it was decided in 1975 to undertake work on a machine translation system that would run on this installation on a time-sharing basis.

From the outset it was recognised that post-editing would be a necessity. This was a trade-off for the fact that the system would have to be able to deal with free syntax, with any vocabulary normally used in the Organization, and, ultimately, with as many different fields and genres of discourse as possible. No consideration was given to a mode of operation that would require pre-editing. The intention was to have a system that would mesh with the routine flow of text within the secretariat.

With these criteria in mind, a team of consultants was contracted in 1976 to develop a system specially tailored to PAHO's needs. Of the two priority combinations, English-Spanish and Spanish-English, the latter was chosen as the first area of concentration. This combination requires fewer parsing strategies in order to produce manageable output, and at the time priority had to be given above all to setting up the architecture of the system and its extensive supporting software.

The next three years were devoted to mounting this architecture and to writing the basic algorithm for translation from Spanish to English. At the end of that period there were twelve PL/1 programs in place performing a variety of tasks, including dictionary update, retrieval, and maintenance. It was also a dictionary-intensive period. In the beginning, the Georgetown methodology (1) was used for dictionary development: hand-coded entries were tied to glosses derived from twin-text concordances of a 40,000-word corpus of PAHO-specific running text. This approach yielded some 8,000 source entries with target equivalents. In order to test the system, however, it was decided to augment this core with multilingual lists of technical terms that were more superficially coded. By 1979 the combined dictionaries came to a level of some 48,000 entries. More than half the total corresponded to terms in the health and biomedical fields, the remainder being general vocabulary.

Toward the end of this initial period, work with the

dictionaries was greatly facilitated by the development of mnemonic, user-friendly software for updating, for side-by-side printing, and for the retrieval of individual records. These were the first collaborative undertakings in which PAHO staff provided feedback and 'wish lists' of features that would be desirable.

The translation algorithm by that time could produce primitive output. There were basic routines for disambiguating part-of-speech homographs, which provided for the possibility of a source word being any combination of noun, verb, or adjective. Idioms could be looked up as units as long as they were fixed strings. Noun phrases were recognised and rearranged in target order. Partial groundwork had been laid for prepositional government. A few lexical routines had been written directly into the program. Rudimentary operations could be performed on the verb string in the third person of the present tense, although it was necessary to have all verb inflections in full form in the source dictionary, and subject pronouns absent from the original Spanish text were inserted in specific environments. It was a fully impacted system, and the programs were not yet modular.

At the time the only mode of input was punched cards. For this reason more than any other, production had not yet been seriously considered. But the picture was to change dramatically at the end of 1979. In November of that year a full-time computational linguist was assigned to the project's regular staff, and shortly thereafter a telecommunication interface was established between the mainframe computer and the Organization's word processing system (then a Wang WPS 30). Thus the word processor was enabled as a remote job entry terminal for sending batch translation jobs to the computer and receiving them back again. It was no longer necessary to have a text specially keyboarded for machine translation. A conversion program was written which interprets for purposes of MT any text prepared in a normal layout using standard typing conventions. Mainly, it recognises format and distinguishes facultative punctuation (capitalisation, full stops, and hyphens) from forms permanently stored in the dictionary. From the time this program was installed, any Spanish text keyed on the Wang system, regardless of the purpose for which it had originally been entered, was available for machine translation. The word processing interface also gave us a powerful tool at the output end. Thanks to the string manipulation features available on the Wang, post-editing on-screen became an easy task from the mechanical standpoint.

It was this combination - the availability of a staff computational linguist in-house and the possibility of sending

and receiving text on the word processor - that provided the stimulus for going into production.

COLLABORATIVE ENHANCEMENT OF SPANAM

Regular use of the system gave it identity, and soon it was baptised Spanam - 'Span' for Spanish, and 'am' for the American Region of WHO. For another year the PAHO staff worked side-by-side with one of the consultants. Inspired by the process of ongoing production, PAHO began to specify the improvements in the algorithm that would have the greatest impact on translation. In response to our recommendations, the following improvements were made: verb synthesis was expanded to include all tenses; verb string manipulation was improved; features were added which permitted the disambiguation of pronouns; idioms were made inflectable; prepositional government was extended in both directions and to various parts of speech; homograph routines were expanded; the noun-phrase patterns were revamped; and the program was modified so as to take orthographic accents (which had not been included up to then) into account. Also during this period a start was made on reorganising the program into a modular structure. This would make it possible to carry on with production while improvements were being made in specific areas of the system.

Gradually the computational linguist became familiar with the system software. By mid-1980 she had completed the first major improvement done independently by PAHO in-house staff, namely the morphological lookup for verbs. Without this development, large-scale production would never have been feasible. Before, when verbs had to be entered in their full form, it often happened that the main verb of a sentence was not found in the dictionary, with the result that the analysis routines were disrupted. After the installation of verb morphology, the incidence of not-found words dropped to less than one per cent, and these in general are not crucial to the structure of the sentence. They are apt to be proper names, abbreviations, Latin terms, and nonc-formations, and in certain environments the system assumes that they are nouns. More recently, several features have been introduced for gap analysis: hyphenated words can now be broken down and dealt with in terms of their components, and the program utilises information from certain prefixes and suffixes. Other improvements added in 1980 included additional work on verb synthesis (in particular on Spanish verb forms occurring in association with the particle *se*, for which a number of treatments are now available) and extension of the maximum length of a source

dictionary entry from five words to twenty-five. On a more general level, further streamlining was done to the program, particularly with a view to making the modules watertight. The structure, as it now stands, is shown in Figure 1.

THE ROAD TO FULL-SCALE PRODUCTION

Starting in mid-1980, production began to steadily gain momentum. People on the staff would hear about Spanam either by word of mouth or from our programme of demonstrations (always on random text), which continues to this day. As our facilities improved, we would establish contact with offices in PAHO where we felt that a particular application might be especially appropriate. For the most part, however, it is the users who have come to us.

In our first major project, the Organization's biennial budget document, we were able to demonstrate a saving in the cost of its translation of 61 per cent and a reduction in staff-days of 45 per cent. The success of this project attracted other users and launched us on our way.

By early 1981 another option became available which potentially would also facilitate production. PAHO's optical character reader, a Compuscan Alphaword II which until that time had been reserved for the transmission of telexes, was interfaced with the word processor. Thus our full configuration includes the OCR (Figure 2). Also, the Wang was upgraded to an OIS/140.

In the last two and a half years we have processed texts in a wide range of fields and for various purposes. Our actual daily average per post-editor, with other duties included, comes to about 6,500 words, and we have been able to post-edit as much as 11,000 words in a day. New software put into use at the end of September 1983 eliminates several housekeeping tasks which previously represented a time overhead of about 20 per cent. This means that the post-editor is able to devote full time to the text, and, with other recent improvements, it should be possible to bring our daily average closer to a consistent level of 8,000 to 10,000 words.

We are constantly developing new techniques and devices for speeding up the process of post-editing. At the cerebral level, we have amassed a bag of tricks for making fewer and more strategic changes in the text. Research time has been cut down by the introduction of reliability marks on all preferred terminology that is found by the dictionary. And at the mechanical end, on the word processor we have designed a series of string functions specially for dealing with English MT output. We try to develop anything that might reduce the work of post-editing, at whatever level the

job can be done most efficiently. This focus, we feel, is much more cost-effective than an exclusive preoccupation with errors that may be generated by the algorithm.

The finished product is delivered by informing the user that the translation is available on the word processing system. Each page of the header bears the words POST-EDITED MACHINE TRANSLATION, and at the end of the document there is a message that reads: THE FOREGOING TEXT IS A POST-EDITED MACHINE TRANSLATION.

Usually our office assumes responsibility for the post-editing. Starting with the budget document, which was a large project, it became evident that we would need someone on the staff with experience in translation who would post-edit and manage the flow of production. The position was created and it has been filled by a trained translator. During slack periods, this person also works on the dictionaries. Sometimes we have given raw, or nearly raw, output to editors or technical writers who were interested only in having a rough draft to work from - and even, on occasion, to other professional translators.*

We do find that it is far more efficient to post-edit on-screen, and for this reason we prefer to deal with users who will be doing the same. The entry of hand-written corrections from hard copy constitutes an extra step which we would prefer to avoid.

Often the only hard copy that we see is the side-by-side output (Spanish on the left, English on the right), printed either on the mainframe computer or the Wang (Figure 3), which we use for guidance purposes during post-editing. As we work, we make note on this copy of any changes that may be needed in the dictionary - changes in the glosses, candidates for micro-glossary treatment, idioms to be introduced, etc. By marking the changes to be made at the time of post-editing, we are able to accelerate the dictionary work. This is an important point at which the functions of our team intersect.

GROWTH OF THE SPANAM DICTIONARIES

The combined experience of development and production enabled us to build the Spanam source dictionary to a level of more than 56,000 entries as of September 1983. Of this total, 94 per cent are bases or stems ('split forms') and 6 per cent are full forms. Although the incidence of not-found words in the output is minimal, we continue constantly

* In December 1984 the Director of PAHO announced a merger of the human and machine translation services.

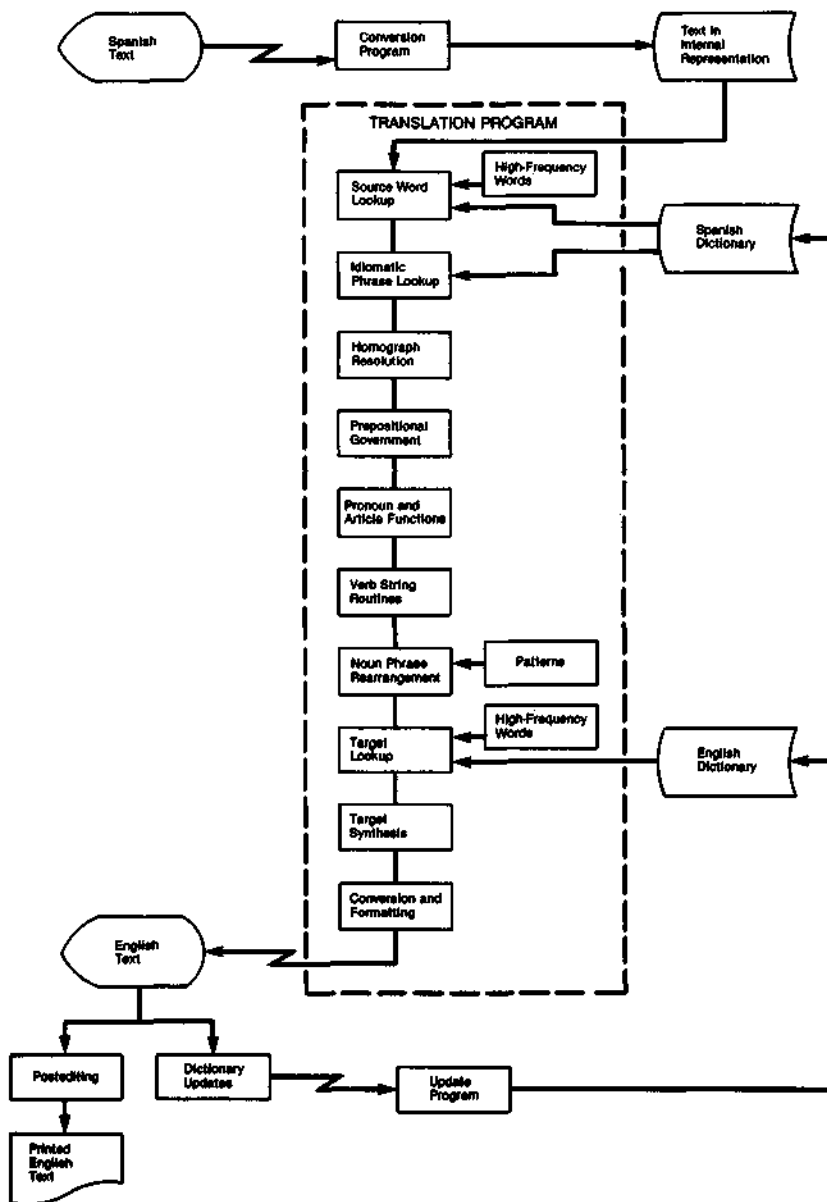


Figure 1. The PAHO Machine Translation System, 1980-

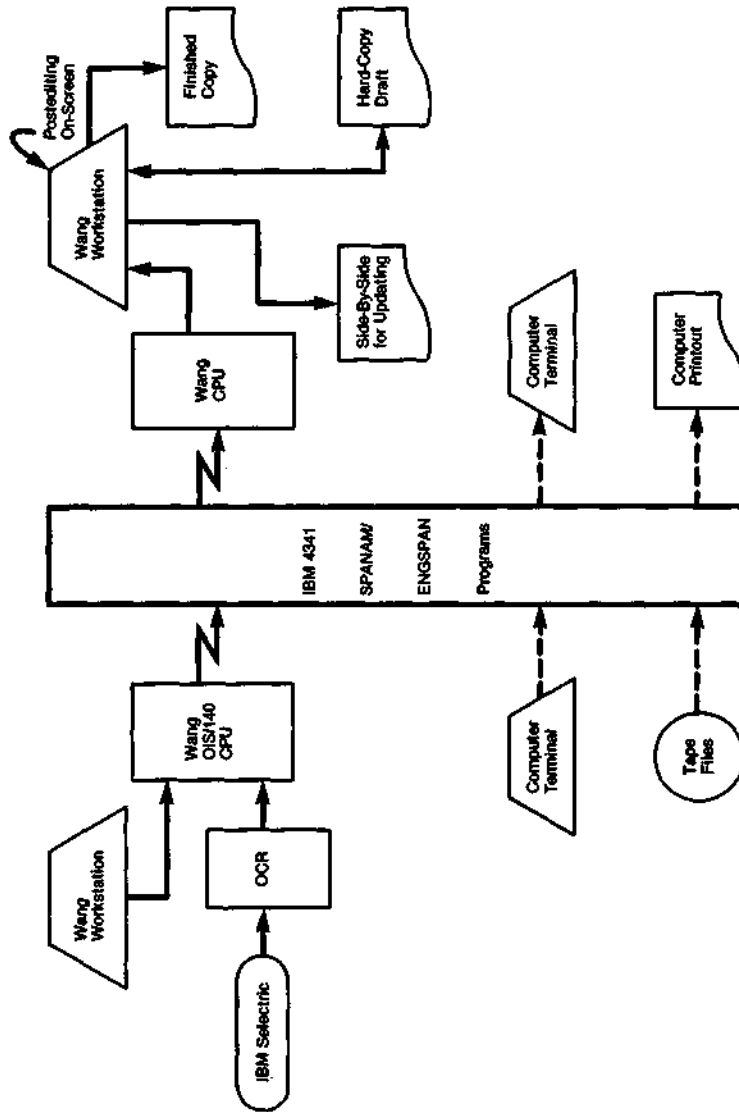


Figure 2. Configuration of Spanam/Engspan, Pan American Health Organization, 1981-

to add new entries based on the results of production and demonstrations. Even if words are found, it is often necessary to improve on their coding, add new homograph possibilities, or incorporate them within an idiom.

Micro-glossaries have been added in order to deal with terms that have different meanings in different disciplines. The micro-glossaries also make it possible for us to attend to the wishes of users who provide us with feedback on their preferred terminology when this differs from glosses that we need to have in the main dictionary. Another feature is the possibility of specifying that preferred and reliable terminology be so marked in the output. These features are an outgrowth of the fact that our office is also responsible for the co-ordination of terminology at PAHO.

Within the year we expect to have installed on the Wang OIS/140 a database of biomedical terminology, Whoterm, which has been developed for the Organization's internal use by WHO/Geneva.(2) To the extent feasible, entries from Whoterm will be incorporated into the MT dictionaries. Thus, when a reliability mark appears in the output, the post-editor will be able to check its definition while remaining at the same workstation.

Entries from the MT dictionaries can be consulted on the word processor as well as at the computer terminal. In effect, the retrieval is a selective print using the same software that prints the hard copies which we use for purposes of consultation and development.

Updating of the dictionaries is also done both at the word processor and the computer terminal, as well as with typewriter and paper, for submission via OCR. The update program is user-friendly in more than one sense: descriptors are mnemonic and, in addition, many defaults are built in. As a result, updating goes quite fast. The update run is always submitted as a batch job, regardless of the mode of input.

Again, as with our other activities, work on the dictionary involves the rest of the system as well. On the one hand, experience from production suggests what is needed, and on the other, problems that appear at first to be at the level of the dictionary may turn out to require adjustments in the algorithm. Or a series of examples, after being worked with in context, may inspire a long-sought solution or a new approach.

FURTHER LESSONS FOR SPANAM: ENGSPAN AS THE TEACHER

Our growth with Spanam prepared us for the challenge of building a system from English into Spanish. The

PAHO NTS 70680 09/19/80 *HDR998999999	SPANISH TO ENGLISH TRANSLATION EXTENSION OF COVERAGE
---	---

La extensión de la cobertura de los servicios de salud a la población no atendida o subatendida de los países de la Región fue la meta central del Plan Decenal y probablemente la de mayor envergadura y trascendencia. Casi todos los países formularon el propósito de extender la cobertura aunque podía apreciarse una diversidad de enfoques para su abordaje, lo cual es comprensible en vista de las diferentes políticas nacionales que habían actuado en la configuración de los sistemas de salud de cada uno de los países. En general, la extensión de la cobertura se podría lograr mediante la expansión de los llamados servicios básicos de salud con servicios mínimos integrales, organizados de acuerdo con el tamaño de los agrupamientos de población y su concentración o dispersión.

La información de que se disponía al iniciarse la década hizo suponer que la población que residía en localidades de 20,000 y más habitantes tenía prácticamente una cobertura del 100% con servicios de salud; que la población que vivía en localidades de 2,000 a 20,000 habitantes estaba cubierta en un 90% y que la población que vivía en localidades de menos de 2,000 habitantes tenía una cobertura que apenas llegaba al 20% con servicios mínimos de salud. La atención se dirigió inmediatamente a la consideración de la forma en que esta última población podía ser mejor atendida. En la mayor parte de los países se dió relieve entonces a la organización de los sistemas de servicios de salud ampliando el número de unidades de atención elemental, entrelazándolos por medio de un sistema de referencia para dar acceso a toda la población a una atención de nivel de complejidad que el caso requiriera.

Como se mencionara en la evaluación inicial, la información que brindaron los países no fue suficiente para acrecentar el conocimiento de que ya se disponía con respecto a la situación de la cobertura.

Figure 3. Sample output of Spanam

PAGE 1

The extension of the coverage of the health services to the underserved or not served population of the countries of the Region was the central goal of the Ten-year Plan and probably that of greater scope and transcendence. Almost all the countries formulated the purpose of extending the coverage although could be appreciated a diversity of approaches for its attack, which is understandable in view of the different national policies that had acted in the configuration of the health systems of each one of the countries. In general, the extension of the coverage could be achieved through the expansion of the called basic health services with integral minimum services, organized in accordance with the size of the groups of population and its concentration or dispersion. The information that was had upon being initiated the decade made to suppose that the population that resided in localities of 20,000 and more inhabitants had practically a coverage of the 100% with health services; That the population that lived in localities of 2,000 to 20,000 inhabitants was covered in a 90% and that the population that lived in localities of less than 2,000 inhabitants had a coverage that scarcely arrived at the 20% with minimum health services. The attention was directed immediately to the consideration of the form in which this last population could be better attended. In most of the countries was given emphasis then to the organization of the systems of health services expanding the number of units of elementary attention, linking them by means of a system of reference in order to give access to the whole population to an attention of level of complexity that the case required.

As there was mentioned in the initial evaluation, the information that gave the countries was not sufficient in order to increase the knowledge that already was had with respect to the situation of the coverage.

Figure 3. (continued)

importance of a combined working mode that would permit the upgrading of Spanam while at the same time starting on the new system, Engspan, was emphasised in evaluations of the system in early 1981. Two separate evaluations were done by Professors Ross Macdonald and Michael Zarechnak of Georgetown University. Both Dr Macdonald and Dr Zarechnak pointed out that insights gained in each endeavour would contribute to the other. And this in fact has proven to be the case.

The chief priorities isolated by the consultants entailed the need to deepen the dictionary coding and to expand the parsing routines. Their recommendations were essential for a system from English into Spanish, but they also applied to improvement of the existing system. Spanam was to continue operating, but features of the new system, as they became available, could be incorporated. Thus the dictionary record, which is common to the two systems, has been expanded to permit extensive possibilities for syntactic and semantic coding that had not existed before. Where there had originally been 82 fields there are now 211. It is an evolutionary process. The new fields are added by turning bytes from the original record into bits. Accordingly, there is much space left to draw on, and further reorganisation is anticipated. The new codes (Table 1) are available to both systems. Also, a greatly improved method for the handling of discontinuous idioms will soon be available. And most important, the parsing of Spanam will benefit from the expanded strategies being developed for Engspan through the use of augmented transition networks based on a slot-and-filler-type grammar. (3)

Spanam/Engspan operate today using a total of sixteen PL/1 programs - six major ones (Table 2) and ten external procedures. They all run on the mainframe, now an IBM 4341. Of thirteen programs delivered to PAHO in 1980, eight are still in use, although they have evolved considerably with the changing environment.

MOMENTUM FOR ENGSPAN

The development of Engspan actually began in late 1981, and as of September 1983 the English source dictionary had approximately 40,000 entries, most of them already tied to appropriate equivalents in the Spanish target. The algorithm has a lemmatisation module, a lookup for single words and idioms, routines for resolving a limited number of homograph types, a module for recognising and synthesising noun phrases, and a complete procedure for the synthesis of inflected Spanish verb forms in all tenses and moods for the first and third persons. There is a working corpus of

50,000 running words. Test translations already give promising results.

Engspan recently received a mandate in the form of a grant from the US Agency for International Development (US AID) for the two-year period beginning August 1983. A second full-time computational linguist has been contracted. Work is focusing on selected aspects of noun phrase analysis, verb selectional restrictions, and clause-level parsing, which are expected to produce the highest yield in terms of impact on translation. The dictionary work - mainly in-depth coding of existing entries - is to accompany the new developments as required.

Toward the end of the grant period, an evaluative study will address the possibility of Engspan being adapted to a mini- or a microcomputer.

CONCLUSION

At all points the work of Spanam and Engspan is closely interrelated, and work goes on simultaneously in every area. The job that lies ahead, I believe, can best be tackled in an environment, such as the one described here, which brings all the phases together under a single roof - post-editing by professional translators, terminology work, dictionary-building, and system refinement. A high degree of interaction with the output is an important factor in the further enhancement of operational systems. This is particularly so in the new era that we are entering on in machine translation. As Professor Yorick Wilks has said, 'there is now no place left for the endlessly diverting question of whether MT is possible or not; it is clearly so'.(4) The current challenge is to whittle away at the remaining inefficiencies in the day-to-day working environment, attacking them at whatever level is most effective. The need now is for a sustained problem-solving effort, always creative and taking advantage of technological innovation as it becomes available and linguistic insights as they become known.

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 León, our staff computational linguist since 1979. From the computational linguistics programme at Georgetown University we have had consulting support from Michael Zarechnak, Leonard Schaefer, and R. Ross Macdonald, head of that programme,

whose sudden death on 16 June 1983 was a very sad blow.

Within the structure at PAHO, we are indebted to our supervisor, Luis Larrea Alba, Jr, Chief of General Services, for his support since the beginning, and to Dr Charles L. Williams, Jr, Deputy Director of the Organization until his retirement in 1979, without whose vision machine translation would never have come to PAHO.

REFERENCES

- (1) ZARECHNAK, M. The history of machine translation. In: HENISZ-DOSTERT, B., ROSS MACDONALD, R. and ZARECHNAK, M. Machine translation. The Hague: Mouton, 1979.
- (2) AHLROTH, E. and ARMSTRONG LOWE, D. The WHO Terminology Information System: interim report. [Geneva: World Health Organization, 1983.] HBI/ISS/83.1.(mimeo).
- (3) WINOGRAD, T. Language as a cognitive process. Reading (Massachusetts) and Menlo Park (California): Addison-Wesley, 1983.
- (4) WILKS, Y. Concluding remarks. In: LAWSON, V. (ed.). Practical experience of machine translation. Proceedings of the third 'Translating and the Computer' conference, London, 5-6 November 1981. Amsterdam, New York: North-Holland, 1982, p. 189.

AUTHOR

Muriel Vasconcellos, Chief, Terminology and Machine Translation, Pan American Health Organization, 525 Twenty-Third Street, NW, Washington, DC 20037, USA.

Table 1. Pan American Health Organization: evolution of the dictionary record, 1976-1983

Remaining fields still in use from original dictionary record designed in 1976		New fields introduced 1981-1983	
Item (word or phrase)	30 bytes	Semantic unit codes	6 bits
Noun synthesis codes	6 bytes	Ambiguity types	16 bits
Reliability code	1 byte	Relation types	8 bits
Lexical number	12 bytes	Affix codes	8 bits
Article government	1 byte	Verb synthesis features	16 bits
Primary part of speech	2 bytes	Verb string parse features	10 bits
Gender	1 byte	Person	6 bits
Number	1 byte	Syntactic features	16 bits
Tense	1 byte	Selectional preferences	16 bits
Parset codes	3 bytes	Semantic features	22 bits
Position	1 byte	Microglossaries	16 bits
Capitalization	1 byte		
Prepositional government	72 bytes	Total:	144 bits
SE particle	2 bytes		(18 bytes)
Source	2 bytes		
Length	2 bytes		
Unassigned	3 bytes		
Total:	142 bytes		
GRAND TOTAL:			<u>160 bytes</u>
Size of each fixed-length record:	160 bytes (unchanged)		
Number of fields in each record:	82 (1976-1980)		
	154 (1981)		
	211 (1982-present)		

Table 2. Pan American Health Organization: machine translation software in use as of September 1983

Name of program	Description	Status
SPANAM	Logic for the translation process itself, Spanish-English,	Begun in 1976; undergoes constant enhancement.
ENGSPAN	Logic for translation. English-Spanish.	Begun in 1982; under development by PAHO staff.
WANGMTS	Prepares text for processing by SPANAM, interprets format, facultative capitalization and punctuation.	Installed late 1979; underwent much enhancement; occasional changes still being made.
UPDATE	Permits dictionary additions, deletions, changes using mnemonics in free form or record image.	Installed 1978; underwent much change 1978-1979; now modified when new dictionary codes are introduced or dictionary format is altered.
DPRINT	Prints dictionary in side-by-side mnemonic format or record image.	Installed 1979; changed in tandem with UPDATE.
MTSCONC	Gives keyword in context; new version arranges words by part of speech.	New program written in 1982 owing to reorganization of dictionary.