Machine translation systems

Edited by
JONATHAN SLOCUM
Microelectronics and Computer Technology Corporation
Austin, Texas
SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization

Muriel Vasconcellos and Marjorie León

1 Project History and Current Status

1.1 Overview

Machine translation has been supporting the activities of the Pan American Health Organization (PAHO) since 1980. SPANAM, the system that translates from Spanish into English, began to produce texts for requesting offices within the Organization in 1980, and the English-Spanish system, ENGSPAN, became operational in 1984. As of December 1985, SPANAM had produced 2,614,779 words of translation (10,459 pages) in response to 872 job orders, while ENGSPAN had translated 612,973 words (2,452 pages), of which 594,293 words (2,377 pages) were production texts, requested under 146 orders. The translation programs run on PAHO's mainframe computer (now an IBM 4381 running under DOS/VSE/SP), which is used for many other purposes as well. A version of the program is also available for the IBM 3081 (OS/MVS). Texts are submitted and retrieved using the ordinary word-processing workstation (Wang OIS/140) as a remote job-entry terminal. Production is in batch mode only. The input texts come from the regular flow of documentation in the Organization, and there are no restrictions as to field of discourse or type of syntax. Specially trained translators, working at the word-processing screen, produce polished output of standard professional quality at a rate between two and three times as fast as traditional translation (4,000-10,000 words a day for post-editing vs. 1,500-3,000 for human

SPANAM and ENGSPAN are trademarks of the Pan American Health Organization.
translation). The output is ready for delivery to the requesting office with no further preparation required.

The translation programs and their supporting software are currently written in PL/I. SPANAM's speed on the mainframe has been registered at 1,500 words per minute in clock time (495,000 words an hour in CPU time), while that of ENGSPAN has reached 836 wpm in clock time (102,000 wph CPU). They run with size parameters of 215K and 800K, respectively. As of December 1985, the dictionaries for SPANAM had 61,282 source entries and 58,485 target entries, while those of ENGSPAN had 45,614 and 47,545, respectively. These reside on permanently mounted disks and occupy from 7 to 10MB each.

SPANAM and ENGSPAN use essentially the same modular system architecture. ENGSPAN is more advanced linguistically, however, since it has an ATN parser and separate lexical and syntactic transfer modules. The overall policy is to regularly upgrade SPANAM as breakthroughs become available in the more sophisticated ENGSPAN. In this way it has been possible to maintain ongoing production with SPANAM while its capabilities are gradually enhanced and expanded. Because of this dynamic mode of development, information about the theoretical status of either SPANAM or ENGSPAN is necessarily short-lived.

1.2 Early History: 1976-1979

The Pan American Health Organization, with headquarters in Washington, D.C., is the specialized international agency in the Americas that has responsibility for action in the field of public health. It comes under the umbrellas of both the Inter-American System and the United Nations family, serving in the latter instance as Regional Office of the World Health Organization. In addition to its headquarters staff of 533 in Washington, PAHO has a field staff of 657 that supports both the operational programs in its 10 Pan American centers and 29 other offices in the field serving the 38 member countries.

Business may be conducted in any of the four official languages: Spanish, English, Portuguese, and French. The translation demand is greatest into Spanish, which over the years has corresponded to more
than half the total workload (average 57 percent), and, after that, into English. The demand for Portuguese is considerably smaller, and there is only an occasional requirement for French.

In 1975 the Organization's administrators undertook a feasibility study and determined that MT might be a means of reducing the expenditure for translation. There was already a mainframe computer, then an IBM 360, at the headquarters site, and the decision was made to develop an MT system that would run on this installation on a time-sharing basis. Work was to focus on the Spanish-English and English-Spanish combinations. The effort was to be supported under the Organization's regular budget.

The intention from the outset was that MT should articulate with the routine flow of text in PAHO. Post-editing was considered to be unavoidable, since the system would have to deal with free syntax, with any vocabulary normally used in the Organization, and, in time, with a large range of subjects and different genres of discourse. No serious thought was given to a mode of operation that would require pre-editing.

Initial efforts began in 1976. A team of three part-time consultants worked under contract for the Organization for two years, and one of these consultants remained with the project on a part-time basis for a third year. In the beginning the translation strategy reflected some of the principles that had been applied at Georgetown University in the late 1950's and early 1960's in the course of work on the Russian-English system. It should be stressed, however, that the project's theoretical orientation has evolved significantly, and that the algorithm now reflects up-to-date linguistic and computational approaches.

The consultants chose to address the Spanish-English direction first because they recognized that results could be available earlier than if they had started with English as the source. Parallel efforts were concentrated on the architecture of the system itself and the extensive supporting software. The period 1976-1978 saw the mounting of this architecture and the writing of a basic algorithm for the translation of Spanish into English. At the end of three years the Spanish-English algorithm was in place, as well as a set of PL/I support programs that performed a variety of related tasks. The Spanish
source dictionary had been built to a level of 48,000 entries, with corresponding English glosses in the separate target dictionary. Work on the dictionaries was supported by mnemonic, user-friendly software developed in 1978-1979 to facilitate the operations of updating, side-by-side printing, and retrieval of individual records. A corpus of about 50,000 words had been translated from Spanish into English.

Human resources during the period 1976-1978 consisted of the three part-time consultants together with PAHO's contribution in the form of dictionary manpower (total 24 staff-months in the three-year period) and, starting in 1977, half-time participation of the staff terminologist, who assumed the responsibility of coordination. A computational linguist was recruited and assigned to the project at the end of 1979.

The year 1980 was a turning-point for MT at PAHO. Advances came together which made it possible to move into a production mode. With the computational linguist on board, the programs were greatly improved, and the operational problems of text input, the most serious impediment to production, were resolved. An interface established between the IBM mainframe and the Organization's word-processing facility (then a Wang System 30) enabled MT to take its place in the text-processing chain and tap into a large body of text that had been made machine-readable for other purposes. A program was written to convert the Wang text for processing on the IBM mainframe, and thereafter any Spanish text that had been keyed into the word-processor, regardless of the purpose for which it was entered, was available for machine translation.

1.3 Operational Phase: 1980-present

Today the machine translation activity functions as a component of the Organization's newly merged Language Services, and the human translators in both languages post-edit machine output directly on the screen. However, progress toward this end has been gradual.

In 1981, the increasingly steady flow of requests for translation justified the assignment of a full-time post-editor. The rising demand was matched by an expanded volume of texts prepared origi-
nally in machine-readable form. Whereas word processing had previously been restricted to special services provided by a typing pool, the installation of word-processing hardware (Wang OIS/140) throughout the headquarters building brought all the program units into the text-processing chain. Furthermore, an optical character reader (then a Compuscan Alphaword II) was interfaced with the word-processing system; thus, existing typewriters could also be used as input devices, and this meant that texts could be prepared in the field, and scanned in Washington for subsequent input to SPANAM.

With accelerated production, improvements to SPANAM followed in tandem. From the beginning it was the policy, and continues to be so today with both SPANAM and ENGSPAN, that the output from production serves not only to meet the purpose for which it was requested but also to provide feedback for further development of the algorithm and dictionaries. As post-editing proceeds, the translator makes note of recurring problems at all levels on a side-by-side version of the output (for every translation two outputs are generated, the side-by-side hard copy and the word-processing document on the screen). The messages noted by the translators serve as a basis both for updating the dictionaries and for making enhancements, as feasible, in the algorithm. The capture of this information at the time of post-editing saves much work later on.

This mode of operation has brought the dictionaries to their current size. The totals correspond largely to base, or stem, entries, rather than fully inflected forms. For example, in SPANAM's Spanish source dictionary of 61,282 entries, 94% were stems and 6% were full forms. For both systems the incidence of not-found words in random text is well under 1 percent – limited usually to proper names, scientific names, new acronyms, and nonce formations. Through coordination with the terminology side of the program, the glosses have been increasingly tailored to the specific requirements of PAHO. In addition, microglossaries have been established for various users, so that specialized glosses can be elicited.

Further details about the working environment are given in Sections 2 and 6 below.
1.4 SPANAM/ENGSPAN Development: 1981-present

In early 1981 a long-range strategy was decided on for the continued improvement of SPANAM and the development of a parallel system from English into Spanish. Two consultants from Georgetown University, Professors R. Ross Macdonald and Michael Zarechnak, undertook separate evaluations of SPANAM at that time. Their recommendations led to the adoption of a combined working mode in which improvements were to be introduced in SPANAM according to a predetermined schedule while at the same time development began on the other system, ENGSPAN. Recognizing that each language combination imposed a different set of linguistic priorities, the consultants nevertheless emphasized that greatly expanded parsing was needed in both cases, especially in the analysis of English as a source language. Such parsing, in turn, called for revision of the dictionary record in order to allow for a broader range of syntactic and semantic coding. It was felt that the basic modular architecture of SPANAM, as well as the dictionary record in its essential format, should be used for ENGSPAN as well. A common architecture for the two systems meant that they could continue to share the same supporting software. Thus, improvements could migrate readily from one system to the other; it would be easy for them to cross-fertilize.

Having adopted this approach to development, with each side to benefit systematically from the work being done on the other, the project addressed its attention in 1981 to the enhancements that had been recommended for SPANAM. Then, as the SPANAM effort tapered off, time was devoted increasingly to ENGSPAN. By the end of 1982 the ENGSPAN program and dictionaries (about 40,000 source entries, most of them with acceptable glosses in the Spanish target) were in place.

Translation from English into Spanish has special importance for public health in the developing countries, and this fact provided the incentive for seeking extrabudgetary support from the U.S. Agency for International Development (AID). In August 1983 AID gave the Organization a two-year grant for the accelerated development of ENGSPAN. This funding made it possible to have a second computational linguist
for the grant period, as well as consultants and part-time dictionary assistants who undertook specific tasks within the approved plan of work.

With the added manpower, the project made significant progress on the English-Spanish algorithm. Particular focus was placed on the development of a parser using an augmented transition network (ATN), which as of April 1984 was integrated into the rest of the ENGSPAN program. The dictionary record was modified without any increase in its overall size, so that it can now accommodate 211 fields, as compared with 82 in the 1980 version of SPANAM. Introduction of the new syntactic and semantic codes in the English dictionary was completed in 1985.

2 Application Environment

2.1 Pre-editing Policy

As indicated above, it has always been expected that the output of SPANAM and ENGSPAN would have to be post-edited. There was no application of MT at PAHO for which a customized language would be feasible. Nor was it to be expected in the near future that either of its source languages would be the front end of other language pairs. In these circumstances, there were few advantages to be gained from pre-treatment of the input. Since post-editing was inevitable, it was felt that an additional pre-editing step would be uneconomic: the advantages would not be sufficient to offset the added cost of a second human pass. Moreover, in order for pre-editing to be worthwhile, the process would have to draw on a high degree of linguistic sophistication, and adequate manpower for this purpose was scarce.

Thus, pre-editing in the linguistic sense has been ruled out for SPANAM and ENGSPAN. In theory, a document can be sent for execution by SPANAM without being seen by any human eyes. If the operator has keyed in the original Spanish document using normal in-house typing conventions, no adjustments should be required. With inexperienced operators, the precaution is taken to check the format, particularly the line-spacing and page width, since deviations from the
standard at that level can disrupt the work of the algorithm. OCR input also needs to be reviewed for accuracy.

Production texts are usually run only once. For long projects, particularly with ENGSPAN, the translation is run in several parts, additions or changes being made to the dictionaries after each section is processed. Demonstrations are always performed on random text.

2.2 Post-editing Policy

2.2.1 General Position

It is a firm rule at PAHO that post-editing is done by professional translators who have been specially trained in the techniques of handling machine output. The strictness of this rule is balanced against a flexible standard for the degree of post-editing required.

The decision as to the extensiveness of post-editing takes into account:

- the purpose of the translation,
- the user's own resources for editing,
- the time frame, and
- structural linguistic considerations in the text itself.

A text may be needed for information only, for publication, or for a variety of uses between these two extremes. If it is to be edited by the requesting office, only the most glaring problems are dealt with by the translator. On the other hand, if it is to be published without much further review, the translator devotes careful attention to the quality of the text. These factors are determined in consultation with the user at the time the job is submitted. As to time constraints, it may happen that the work has to be delivered under considerable pressure: information-only translations of 20-25 pages may have to be delivered within a couple of hours, and once a 40-page proposal for funding was delivered in polished form the same day it was requested. With translation for publication, however, longer periods are negotiated.
Contrary to what might be deduced from the nature of PAHO's mission, SPANAM and ENGSPAN are asked to cope with a wide range of subject areas and types of text. There have been: documents for meetings, international agreements, technical and administrative reports, proposals for funding, summaries and protocols for international data bases, journal articles and abstracts, published proceedings of scientific meetings, training manuals, letters, lists of equipment, material for newsletters - even film scripts. They are certainly what Lawson (1982:5) would call "try-anything" systems.

The fact that SPANAM and ENGSPAN have such highly varied applications makes it particularly necessary to use trained professional translators as post-editors. Whereas Martin Kay (1982:74) suggests that the person who interprets machine output "would not have to be a translator and could quite possibly be drawn from a much larger segment of the labor pool," experience in the PAHO environment suggests that this conclusion would be valid only for technical experts working on a text for information purposes - a circumstance that is rare at PAHO.

Only an experienced translator will be aware of the words whose variable meanings are dependent on extralinguistic context. For example, proyecto in Spanish can be translated as "project," "proposal," or "draft," and the choice depends on full knowledge of the situation to which the text refers. Esperar can be translated as "hope" or "expect," and the difference is significant in English - sometimes even crucial. Such ambiguities require the attention of a translator with training, experience, good knowledge of the subject-matter vocabulary in both languages, and a technical understanding of what is meant by the text. Only a person with this combined background is in a position to make the choices that will fully reflect the intention of the original author. Another area in which the translator's role is important is in interpretation of the degree of intensity associated with relative terms. For example, trascendente in Spanish can have much less force than its English cognate, and the entire tone of a message may be over- or underdrawn depending on the interpretation given to a key term of this nature. Indeed, it has been the experience of SPANAM that users, even technical experts, can misinterpret
the glosses appearing in the machine output and assign an altogether incorrect meaning in the process of "correcting" the text. The role of the experienced translator is not to be underestimated.

It was these considerations that led PAHO to form a combined language service, effective as of November 1984, under which all translators would ultimately be expected to devote part of their time to the post-editing of MT output. Fortunately, at the time of the merger there were six translator vacancies, and for four of these posts arrangements were made to include post-editing, and also dictionary maintenance, in the job description and in the announcement that listed the duties of the staff being recruited. This policy will be continued with future vacancies. Thus the incoming translators are aware from the outset that their duties will include post-editing.

2.2.2 Linguistic Strategies

In addition to experience in the interpretation of nuances, the post-editor needs a strong linguistic background in order to master the particular strategies that have proven to be effective in the recasting of certain unwieldy constructions that frequently recur – with SPANAM, for example, the result of translating a verb that was in sentence-initial position in Spanish. For the translator untrained in post-editing, the most time-consuming task is the recasting of such constructions. For problems of this kind a series of "quick-fix" post-editing expedients (QFP) have been developed. At the same time, there is a series of word-processing aids that help to speed up the physical process of editing and to deal with pragmatic decisions in the output which are not handled by syntactic rules.

To follow through with the example mentioned above, certain maneuvers are suggested as being useful with fronted verb constructions in Spanish, which occur frequently and present difficulties for the SVO pattern typically required for English. The purpose of the QFP is to minimize the number of steps that are required in order to make the sentence work. Since it was a V(S)O construction that triggered the problem in the first place, any solution that avoids reordering will necessarily depart from one-on-one syntactic fit. In other words, in
the example of the fronted verb, one might try to see if the opening phrase, which will be a discourse adjunct, a cognitive adjunct (terms from Halliday 1967), or the main verb itself, could be nominalized so that it can serve as the subject of the sentence in English. Such an approach manages to preserve in the theme position (Halliday 1967) the cognitive material which had been thematic in the source text, usually with a parallel effect on the focus position as well (Vasconcellos 1985b). For this reason, the result is often quite satisfactory, even compared with a translation that is syntactically more “faithful” – see Section 6 below, and (Vasconcellos 1986). The examples below compare QFPs with solutions that were actually proposed by translators new to post-editing (THT = traditional human translation).

In example (1) the semantic content of the fronted verb is reworked into a noun phrase that can serve as the subject of the sentence. Time is saved by leaving the rheme of the sentence untouched; only a few characters, highlighted inside the box, were changed. Moreover, additional speed was gained by making changes from left to right, in the same direction in which the text is being reviewed.

(1) Durante 1983 se inició ya la transformación paulatina de estos planteamientos en acciones.

MT: During 1983 [already] the gradual transformation of these proposals into actions.

THT: During 1983 these proposals already began to be gradually transformed into actions. (62 keystrokes)

QFP: During 1983 [progress began toward] the gradual transformation of these proposals into actions. (27 keystrokes)

In example (2), on the other hand, the adjunct itself is nominalized, again with a significant saving of time and keystrokes:
En este estudio se buscará contestar dos preguntas fundamentales:

MT: In this study it will be sought to answer two fundamental questions:

THT: In this study answers to two fundamental questions will be sought: (53 keystrokes)

QFP: This study seeks to answer two fundamental questions: (14 keystrokes)

Use of this approach, wherever feasible, as well as ones that have been devised for other troublesome differences between the source and target languages, adds up to substantial economy, with apparently little or no deterioration in the quality of the translation (see Section 6 below). However, knowing when and how to make such changes requires considerable skill. This is one more reason why the post-editor should have a strong background in translation and, if possible, in linguistics as well.

It has always been emphasized in the post-editing of SPANAM and ENGSPAN that editorial changes should be kept to the minimum that is needed in order to make the output intelligible and acceptable for its intended purpose.

2.2.3 Word-processing Strategies

The post-editors work directly on-screen. Experience has shown that post-editing on hard copy, with the changes entered by a “word-processing operator,” is not a highly efficient mode. Accordingly, attention has also been given to speeding up the post-edit by automating as many of the recurring operations as possible.

The SEARCH-and-REPLACE function on the word processor is heavily used in post-editing. In addition, SPANAM and ENGSPAN have a series of special macros that have been developed for purposes of MT. Besides a full set of possible word switches (1 × 1, 1 × 2, 2 × 1, 2 × 2, 1 × 3, 3 × 1, 3 × 3, etc.), there are macros that deal with the particular character strings that have to be changed most often in the output for
pragmatic reasons. For example, with a single "glossary" keystroke it is possible to perform the following editorial operations in the English output:

SEARCH-and-DELETE:

the, of, there, to, in order to

SEARCH-and-REPLACE:

from replaces of, for/of, for/by, in order to V/for Ving, a/the, which/that, who/that, every/each, among/between, such as/as, some of the/some

The inventory can be changed or expanded at will.

2.2.4 Other Time-savers

From the discussion above it can be seen that speed in post-editing is achieved by a combination of strategies. Some of the points made may appear at first sight to be unimportant, but yet they can add up to a significant difference. One example of an apparently trivial factor is the method of positioning the cursor under the string to be modified. Delays at this point can add up to a surprising proportion of total time spent on post-editing, since they will occur with every change that is made. Informal experiments suggest that the most efficient approach for positioning the cursor is to always use the SEARCH key. (The "mouse" and light pencil remain to be fully investigated.) The slowest method, unfortunately, seems to be the one that is most often used, namely simple manual striking of the directional keys. Since people tend to rely on the directional keys unless otherwise trained, this point is emphasized with the post-editors who work on SPANAM and ENGSPAN.

The staff of the project are constantly on the lookout for new ways of saving time. All tasks are streamlined as much as possible. A series of programs has been developed on the word processor for automating the housekeeping support that has to be done apart from post-editing, and recently some of this work was made even more efficient by passing it on to the mainframe computer.
2.3 Post-editing vis-à-vis Other Aspects of the System

In the PAHO MT environment there is a close link between post-editing and the other aspects of the system. The translators are trained to update the dictionaries, and required changes in the dictionaries are proposed at the time of post-editing. This saves going through the text a second time, and it also captures ideas about the glosses and coding of a particular item when the translator has the entire text fresh in mind. This person, if adequately trained in updating, is in an excellent position to decide how to deal with the specific constructions that tend to recur in production translations.

The translators also alert the computational linguist to areas where the algorithm needs improvement.

2.4 Integration into a Complete Translation Environment

As of November 1985 the human and machine translation activities at PAHO, together with simultaneous interpretation, were merged into a single Language Services Unit. There is now centralized screening of incoming jobs. In addition, requestors are free to specifically request a machine translation if they wish. The triage makes it possible to maximize the effectiveness and efficiency of the respective modes. There is now a more rational utilization of the manpower available at any given time, with the staff being assigned to different duties depending on both needs and skills. Also, it is hoped to be able to ultimately reduce a given person’s day in front of the word-processing screen from eight hours to six through the rotation of assignments. To date, however, there have been no complaints of eyestrain or other discomfort with the VDU.

In the area of management, the SPANAM/ENGSPAN programs on the mainframe computer are also helping with labor-intensive operations for which the human translation service is responsible: it is already performing automatic word counts, and spelling-check systems are being developed to reflect the Organization’s highly technical vocabulary.

SPANAM/ENGSPAN can also help to lighten the load in the human translation process by retrieving technical and scientific terms in con-
text. Considered from the standpoint of term retrieval, MT makes for an efficient lexical data base. With an ordinary LDB, the translator has to go to the terminal (which is not usually for his/her use alone), sign on, and initiate a search. After all the mechanical steps have been performed, there is still the possibility that the term is not in the data base at all, and that the effort will have been wasted. When this happens repeatedly over time, frustration builds up. With SPANAM/ENGSPAN, on the other hand, not only does the translator know immediately what translation has been assigned to the term, but he also knows its degree of reliability and whether or not a full terminological entry is available in the WHOTERM data base, which can be accessed from the same workstation. The status of a term is indicated by small superscript symbols which can be requested at the time the text is sent for translation.

WHOTERM's files normally contain: a definition of each term in English, translational equivalents in up to four languages besides English, synonyms if there are any, a reliability code for the primary term in each language, scope notes, and a subject code. In addition to the general file, it has files with: names of organizational entities, full equivalents for abbreviations, scientific names of pathogens, generic names of drugs in three languages, and chemical names of pesticides with trade names cross-referenced to them (Ahlroth and Armstrong-Lowe, 1983).

SPANAM/ENGSPAN can also aid the human translation process with the system of microglossaries for specialized subject areas (see Section 4.3 below). When a text is known to deal with a certain subject, the translator can request a corresponding microglossary which will contain alternative glosses. One or more of these microglossaries can be specified at the time the job is submitted. The post-editing translator can also have a microglossary of his/her own in which can be stored special glosses that he/she prefers to use.

It is possible for SPANAM/ENGSPAN to provide "slashed entries," i.e., alternative choices in the output entry such as project/proposal/draft, hope/expect, time/weather, etc., although this is not the regular policy. These alternatives can be stored in a microglossary. In the output, the desired translation is saved and the rest is deleted with a
single keystroke.

If the translator provides feedback in the form of suggested or requested changes in the dictionaries, the updating can be done immediately. Some of the requesting offices have developed the habit of providing regular feedback, and this means that their translations become increasingly tailored to their specific requirements.

While there is no doubt that SPANAM/ENGSPAN reach their maximum efficiency when post-edited on-screen, at the same time studies are being done on ways in which a translator can dictate the changes orally and then have a word-processing operator enter them, working from the audio tape.

The human translation service stands to benefit, also, from the macros that have been developed on the word processor in editing a text and from the word count and other support programs.

3 General Translation Approach

Since the bulk of the Organization's translation work involves only Spanish and English, the machine translation system was developed specifically for this pair of languages. No consideration was given to using the interlingua approach. The broad range of subject areas to be dealt with made it impractical to consider using a knowledge-based approach or one based on a representation of the meaning of the text. Although the systems are currently language-specific, significant portions of the algorithm could be adapted for use in a system involving Portuguese or French, the other official languages of the Organization. Because SPANAM and ENGSPAN were developed separately, they reflect different theoretical orientations and utilize different computational techniques. At the same time, they have many features in common.

SPANAM was originally designed as a direct translation system. The translation is produced through a series of operations which analyze the Spanish source string, transform the surface structure to produce a syntactic frame for the English target string, substitute the English glosses indicated by the results of the analysis, insert and/or delete certain grammatical morphemes, and synthesize the re-
quired endings on the English words. The principal stages involved in the translation algorithm are: morphological analysis and single-word lookup, gap analysis, multi-word unit lookup, homograph resolution, subject identification, treatment of prepositions, object pronoun movement, verb string analysis, subject insertion, "do"-insertion, noun phrase rearrangement, target lookup, target synthesis.

ENGSPAN is a lexical and syntactic transfer system based on the slot-and-filler approach to language structure. It performs a separate analysis of the English source string, applies transfer routines based on the contrastive analysis of English and Spanish, and then synthesizes the Spanish target string. The principal stages of this algorithm are: morphological analysis and single-word lookup, gap analysis, substitution and analysis unit lookup, sentence-level parse, lexical transfer, target lookup, syntactic transfer, and target synthesis. The program includes a backup module for homograph resolution, and backup strategies for verb string and noun phrase analysis, that are invoked if the sentence-level parse is unsuccessful.

4 Linguistic Techniques

4.1 Morphological Analysis

SPANAM's morphological lookup procedure makes it possible to find most Spanish words in their stem forms. The algorithm recognizes plural and feminine endings for nouns, pronouns, determiners, quantifiers, and adjectives; person, number, and tense endings for verbs; and derivational endings such as -menter/-ly. Bound clitic pronouns are separated from verb forms and any accent mark related to the presence of the clitic is removed. Another subroutine adds missing accent marks when the source word is written with an initial capital or in all capital letters. The components of compounds formed with hyphens or slashes are looked up as separate words. A few prefixes are also removed from words without a hyphen.

ENGSPAN's morphological analysis procedure, known as LEMMA, is called if the full-form is not found in the dictionary and the word consists of at least four alphabetic characters. This procedure checks for the presence of a number of different endings and contractions.
Each time an ending is removed, the new form of the word is looked up. LEMMA uses morphological and spelling rules and short lists of exceptions in order to determine when to remove or add a final -e, when the word ends in a double consonant, etc. If a lemmatized form of the word is found in the dictionary, its record is checked to make sure that its part of speech corresponds to the ending that was removed. If LEMMA exhausts all its possibilities, the word is checked against a list of prefixes and rules for British spelling. If these efforts are unsuccessful, a dummy record is created for the word and a gap analysis routine is called. Not-found words are initially considered to be common nouns (and proper nouns if capitalized) and given the possibility of also functioning as verbs and adjectives. Information from both LEMMA and derivational suffixes is used in order to confirm or reassign the main part of speech, as well as to confirm, remove, or add possibilities for ambiguities.

The lookup strategy used in both SPANAM and ENGS\^{}P\^{}\^{}AN keeps down the size of the dictionary while at the same time allowing a good deal of flexibility. The dictionary coder has the option of entering a word in its full form, in one or more of its inflected forms, or in its stem form. With irregular forms and homographs, the full form must be used. For example, in the Spanish source dictionary the only entries for the verb esperar are the stem esper and the verb/noun homograph espera. The English source dictionary contains an entry for expect and unexpected, but not for expects, expected, expecting, or unexpectedly.

4.2 Homograph Resolution

SP\^{}\^{}ANAM deals with part-of-speech homographs at several different stages of the program. Ambiguities that can be resolved by morphological clues or capitalization are handled by the lookup procedure. Proper names are also identified at this stage. One-character words are distinguished from letters of the alphabet after the lookup has been completed. The homograph resolution module handles other types of homographs by examining the surrounding context.

The possible parts of speech for a word are indicated in the dictionary record in a series of bit fields which include: verb, noun, adjec-
tive, pronoun, determiner, numerative, preposition, modifier, adverb, conjunction, auxiliary, and prefix. Any combination of two or more bits may be coded. Other sequences of bit codes are used to distinguish between different types of pronouns, adverbs, and conjunctions: relative, interrogative, nominal, adverbial, connector, interruptive, compound, and coordinate.

The use of multiple-word substitution units reduces the number of lexical ambiguities which must be resolved by the algorithm. Analysis units may also be used to selectively specify the part of speech of any or all of the words covered by the unit.

ENgspan's front-line approach to homograph resolution is embodied in the ATN parser which is described below in Section 5.3. The English words can be coded for the same possible parts of speech as in SPANAM. Determination of the function of each word depends on the path taken through the network. The sequence of parts of speech which leads to the first successful parse is used as the basis for the transfer stage.

There are three ways in which lexical information from the dictionary is used to help the parser arrive at the correct analysis. Substitution units compress idioms into one record with a single part of speech. Analysis units can be used to indicate that a group of words can be expected to occur in collocation with a particular function. This information may be overridden, whenever necessary, by the parser. An individual word may also be coded to indicate which of its possible parts of speech is statistically most frequent. Again, the final decision is made by the parser based on the results of the sentence-level analysis.

4.3 Polysemy

SPANAM/ENGSPAN have two principal tools for dealing with polysemy: microglossaries and transfer units. Substitution units and analysis units are also used when common collocations are involved.

A microglossary is a subset of dictionary entries which can be selected for a particular subject area, discourse register, or user. Microglossary entries may be used in both the source and the target
dictionaries. Glosses pertaining to the subject area of international public health form part of the main dictionary. Microglossaries are in use for special translations of terms in the fields of law, finance, sanitary engineering, agriculture, computer science, and biomedical research. The system may have up to 99 microglossaries with any number of entries in each one. The microglossaries to be consulted during the translation of a particular text are specified at run time. The existence of a specific microglossary entry is indicated in the main record for the word. Thus no time is wasted looking for special entries under every word. More than one microglossary may be activated for the same translation, in which event they are listed and consulted in order of priority.

The transfer unit is a lexical transfer rule which is stored in the source dictionary. The existence of a transfer unit is indicated in the record corresponding to the individual source word. A transfer unit specifies a condition to be tested and an action to be performed. Examples of conditions are:

- The subject of this verb has X feature(s) or is word W;
- The object of this verb (or preposition) has X feature(s) or is word W;
- This word modifies a word with X feature(s) or modifies word W;
- This word is modified by a word with X feature(s) or is modified by word W;
- This verb has N object(s);
- This verb governs preposition P and the object of P has X feature(s) or is word W;
- This verb has a complement of type T.

Transfer units are explicitly ordered in the dictionary. The action may either select an alternative translation, insert a word such as a preposition, or delete one or more words. The action also indicates
whether or not additional transfer entries should be sought for the same word.

4.4 Syntactic and Semantic Features

The dictionary record for each lexical item (including substitution units) contains bit fields for storing information about its syntactic and semantic features. These features are used in all stages of the translation. For example, verbs and deverbal nouns are specified as occurring with one or more of the following possibilities: no object, one object, two objects, complement, locative, marked infinitive, unmarked infinitive, declarative clause, imperative clause, interrogative clause, gerund, adjunct, bound preposition, object followed by marked infinitive, object followed by clause, and object followed by bound preposition. Subject and object preferences can be specified as $\pm$Human, $\pm$Animate, and $\pm$Concrete. Other fields can be made available for case frames at such time as they may be introduced. Features which can be coded for nouns include count, bulk, concrete, human, animate, feminine, proper, collective, device, location, time, quantity, scale, color, nationality, material, apposition (noun, infinitive, or clause), body part, condition, and treatment. Adjectives are coded for many of the same features mentioned above. In addition, they can be coded as inflectable, optionally inflectable, comparable, general, temporary, resultative, and tough-movement. Adverbs can be coded as time, place, manner, motive, interruptive, and connector. One of the references used in developing the coding scheme for the English entries was (Sager, 1981).

4.5 Annotated Surface Structure Nodes

In ENGSPAN, the structure produced by the parser consists of a graph of nodes corresponding to each clause and phrase. Each node has a table indicating its constituents, their roles, and their locations. If the constituent is a lexical item, the location is a word number; if it is a phrase or a clause, the location is the pointer to the appropriate node. Each node is annotated with features applicable to the type of
phrase or clause involved. These features include Type, Mood, Person, Number, Tense, Aspect, and Voice.

Both the ATN formalism and the structural representation used in ENGSPAN draw heavily on the presentation of ATN parsers and systemic grammar in Winograd (1983). Winograd’s discussion, in turn, is based on the work of Woods (1969 and 1971) and Kaplan (1973). Of course, the ATN parser has necessarily had to be adapted to the needs and computational environment of the PAHO project.

4.6 Spanish Verb Synthesis

The procedure for the synthesis of Spanish verb forms is based on principles of generative morphology and phonology. The program synthesizes all regular and most of the irregular verbs, in all tenses and moods except the future subjunctive, and in all persons except the second person plural. The verb is entered in the target dictionary in its stem form. Binary codes are used to specify the conjugation class and the 11 exception features that govern the synthesis of irregular forms. Only one dictionary entry is needed for each verb. A small number of highly irregular stems and full forms (74 in all) are listed in a table. The majority of “stem-changing” verbs require no special synthesis coding. The procedure consists of a series of morphological spellout rules; raising, lowering, diphthongization, and deletion rules based on phonological processes; stress assignment rules; and orthographic rules to handle predictable spelling changes.

Spanish verbs are also coded to permit the synthesis of the nominalization corresponding to "the action or process of." The verb synthesis procedure is invoked to produce nominalized forms which correspond to the first or third person present indicative or present subjunctive, or to the past participle or infinitival form of the verb. Other forms are produced by a combination of stem alterations and derivational affixes. Irregular forms require a separate target dictionary entry.
5 Computational Techniques

5.1 Dictionaries

The SPANAM/ENGSPAN dictionaries are VSAM files stored on a permanently mounted disk. The source and target dictionaries are separate files. The basic record has a fixed length of 160 bytes. The source entry is linked to its target gloss by means of a 12-digit lexical number (LEX). The first six digits of the LEX are the unique identification number which is assigned to each pair when it is added to the dictionary. The second half of the LEX is used to specify alternate target glosses associated with the same source entry. The main or default target gloss for each pair has zeroes in these positions.

5.1.1 Single-word Entries

The key for a source entry is the lexical item itself, which may be up to 30 characters in length. The source dictionary is arranged alphabetically. The key for a target entry is the LEX, and the target dictionary is arranged in numerical order.

Words may be entered in the source dictionary either with or without inflectional endings. Most nouns are entered only in the singular and adjectives only in the masculine singular. Verbs are entered as stems. Full-form entries are required for auxiliary verbs, words with highly irregular morphology, and homographs.

Several source items may be linked to the same target gloss by assigning them the same LEX. For example, irregular forms of the same verb or alternative spellings of a word require only one entry in the target dictionary. Likewise, more than one target gloss can be linked to the same source word through the lexical number. Each alternate gloss is distinguished by coding in the second half of the LEX. Two positions are used to designate terms belonging to microglossaries, two for glosses corresponding to different parts of speech, and two for context-sensitive glosses which are triggered by transfer units.
5.1.2 Multiple-word Entries

The dictionaries contain four types of multiple-word entries: substitution units (SU), analysis units (AU), delayed substitution units (DSU), and transfer units (TU). The key for a multiple-word entry in the source dictionary is a string consisting of the first six digits of the LEX for each word in the unit. With an SU or an AU, the words must occur consecutively in the sentence in order for the unit to be activated. A DSU or a TU can cover either a continuous or discontinuous string.

The basic SU contains from two to five words. A different record structure is used for longer entries, such as names of organizations and titles of publications. When an SU is retrieved, the dictionary records corresponding to the individual words are replaced with one record corresponding to the entire sequence. The gloss for the unit is also found in a single entry in the target dictionary. This type of unit is used in order to obtain the correct translation of names of organizations, titles of publications, slogans, etc., and is an efficient way of handling some fixed idioms, phrasal prepositions, and certain technical terminology. An SU record has the same format as a single-word entry. In addition, it contains a character string which indicates the part of speech of each of its members. This information can be used by the parser if it is unable to parse the sentence using the single part of speech specified for the unit. Examples of phrases which are entered as SUs are: by leaps and bounds, International Drinking Water Supply and Sanitation Decade, and Health for All by the Year 2000.

The AU, which also contains from two to five words, has several functions. At the very least, it alerts the analysis routines to the possible presence of a common phrase and provides information on its length and function. It can also be used to resolve the part-of-speech ambiguity of any of its members. Finally, it can specify an alternate translation for one or more of its parts. The AU is an entry in the source dictionary but has no counterpart in the target dictionary. The record for each source word is retained in the representation of the sentence, but the last two digits of its lexical number are modified if a translation other than the main gloss is desired. When the target
lookup is performed, the gloss for each word is retrieved separately. This ensures that the rules for analysis and synthesis of conjoined modifiers will be able to access information about the individual words of the phrase. It also makes it possible for the parser to determine whether or not the individual words are being used as a unit in the given context. Examples of phrases entered as AUs are drinking water and patient care. The algorithm is still able to correctly analyze sequences such as the children have been drinking water with a high fluoride content and it is essential that the patient care for himself.

The DSU is used to handle lexical items such as phrasal verbs which are likely to occur as noncontiguous words in the input. The existence of a DSU is indicated in the source record of the first word of the unit. The unit is retrieved from the dictionary during the sentence-level parse. The decision of whether or not to accept the unit is based on both syntactic and semantic requirements of the parser. If the unit is accepted, it replaces the individual records and causes a different target gloss to be retrieved. Examples of DSUs are look up, put on, and carry out.

The TU is used to specify an alternate translation of a word or words which depends on the occurrence of a specific word or set of features in one of its arguments or in a specified environment. These entries are stored in the source dictionary only and are retrieved after the analysis has been completed. If the conditions specified in the transfer entry are met, the corresponding lexical numbers are modified so that the desired target gloss is selected during the target lookup. For example, if the object of know is coded as +Human, the verb is translated as conocer instead of saber. If female and male modify a noun coded as –Human, they are translated as hembra and macho instead of mujer and hombre.

5.2 Grammar Rules

SPANAM until recently used two basic types of grammar rules: pattern matching and transformations. Pattern matching is used for the recognition and reordering of noun phrases. The grammatical patterns are stored in a file which can be updated without recompiling
the program. The patterns are applied by searching for the longest match first. Transformations are used to identify and synthesize the verb phrase and clitic pronouns. The rules are expressed in PL/I code and are grouped in modules according to the part of speech of the head word. Each group of rules is tested once for each sentence. The structural description of each rule is compared with the input string. The description may require a match of parts of speech, syntactic features, or specific lexical items. If a match is found, the rule is applied. The rule may permute, add, delete, or substitute lexical items or features associated with them.

ENGSPAN's analysis component takes the form of an augmented transition network. The network configuration indicates possible sequences of constituents. The rules governing the acceptability of any specific input string are contained in the conditions attached to the various arcs of the network. The building of the nodes of the structural representation and the assignment of features and roles is determined by actions associated with each arc. The conditions and actions are contained in separate modules which are part of the compiled program. The configuration of states and arcs is specified in a file which is updated on-line. The contents of this file also determine which of the conditions and actions are actually attached to specific arcs for a particular run.

As of December 1985, the ATN grammar had 11 networks: sentence, clause, noun phrase, verb phrase, sentence nominalization, hyphenated compound, prepositional phrase, comparative phrase, adverbial phrase, relative clause, and dependent clause. Each network consists of a set of states connected by arcs. Four types of arcs are used: CATEGORY arcs, which can be taken if the part of speech matches that of the input word; JUMP arcs, which can be taken without matching a word of the input; SEEK arcs, which initiate recursive calls to a network; and SEND arcs, which return control to the calling network after the successful parsing of a constituent. In all, there were 150 conditions and 60 actions as of December 1985.
5.3 Parsing Algorithm

The ENGSPAN algorithm performs a top-down, left-to-right sequential parse using a combination of chronological and explicit backtracking. The parser stops after completing the first successful parse. The path taken through the network depends on the ordering of the arcs at each state, the structural information already determined by the parser, and the codes contained in the dictionary record for each lexical item and multiple-word entry. Also available to the parser is information regarding sentence punctuation, capitalization, parenthetical material, etc., which has been gathered by an earlier procedure. The algorithm processes the words of the input string one at a time, moving from left to right. At each state, all arcs are tested to determine whether they may be taken for the current word. The possible arcs are placed on a pushdown stack and the top arc on the stack is taken. The parser continues through the input string as long as it can find an arc which it is allowed to take. If no arc is found for the current word, the parser backtracks. Which of the alternative arcs is taken off the stack depends on the situation that caused the parser to backtrack. If the end of the string is reached and the algorithm is at a final state in the network, the parse is successful. If no path can be found through the network, the parse fails.

Long-distance dependencies, such as those involved in relative clauses and WH-questions, are parsed using a hold list. When the parser encounters a noun phrase followed by a relative pronoun, a copy of the phrase is placed on the hold list. When a question is being parsed, the questioned element is placed on the list. When a gap is detected in the relative clause or interrogative sentence, the phrase on the hold list is used to fill the appropriate slot.

Whenever backtracking is required, a well-formed phrase list is used to save a copy of the phrases that have been completed but are about to be modified or rejected. For all SEEK arcs, the parser checks to see if a phrase of the appropriate type is already on the well-formed phrase list. If there are several phrases on the list that begin with the same word, the longest phrase is tried first. A new phrase is parsed only if there is nothing on the well-formed phrase list that satisfies
the SEEK arc. In this way, large amounts of reparsing are avoided.

Conjoining is currently being handled by a configuration of arcs at the end of each subnetwork which allows additional phrases of the same type to be parsed recursively. When partial phrases are conjoined, the end of the subnetwork is reached by traversing one or more JUMP arcs.

The ATN parsing algorithm has been developed in an independent PL/I program, using the ENGSPAN input and dictionary lookup modules.\(^1\) It is also totally compatible with SPANAM. The network grammar is read in at run time, making it possible to experiment with different network configurations without recompiling the program. Each time an enhanced version of the parser has been tested and debugged, it replaces the working version in the ENGSPAN program. The parser is being incorporated in a similar way into SPANAM as well.

Figure 1 shows the relationships between the major components of the translation algorithm and the data structures that it utilizes. A more detailed description of the ATN grammar and parser is found in a report submitted to the U.S. Agency for International Development (León and Schwartz, 1986).

5.4 Safety Net

In the event of an unsuccessful parse, ENGSPAN must still produce a Spanish text. A “safety-net” strategy has been built into the translation algorithm in order to enable the transfer component to operate with incomplete information from the analysis component. The quality of the translation obtained using the safety-net procedures is often quite acceptable. For many sentences, it may be comparable in accuracy and grammaticality to a translation based on a successful parse.

\(^1\)A major portion of the parsing routines was developed by Lee Ann Schwartz
Figure 1. Major components of the translation algorithm.
The parser contributes to the safety net by keeping track of the longest path through the network whenever it is forced to backtrack. The amount of work done by the parser is limited by a parameter which can be specified at run time. If the limit is reached before the end of the input string, the parser stops and returns a partial parse. When a partial parse is produced, the surface structure nodes that are still active and have been assigned a headword are used as if they were part of a complete parse. Part-of-speech disambiguation for a word that does not belong to an active surface structure node is accomplished in one of two ways. If the word was parsed as part of the longest path, it is assigned the part of speech specified on the category arc originally traversed by the parser. A backup homograph resolution procedure determines the part of speech for words that were never accepted by the parser, i.e., those that are not included in the longest path and all words within parentheses or dashes. The disambiguation procedure looks at a maximum of two tokens to the left and right, compares the syntactic environment with the coding of the word in question, and selects the most probable part of speech.

Other aspects of the safety net are incorporated in the procedures for lexical and syntactic transfer.

5.5 Transfer Component

5.5.1 Surface Structure Markup

When the parse is successful, the next step is to inspect the surface structure produced by the parser and derive additional information needed for translation. For example, noun phrases are assigned the semantic roles of agent, object, experiencer, or beneficiary. When a subject has been raised out of an embedded clause, the two roles of that noun phrase are indicated. For complements and reduced relative clauses, the modification relationships that were established by the parser are associated with specific lexical items. The result of this structural markup is a table which contains information about both the syntactic and semantic roles of the principal constituents of the sentence.
5.5.2 Lexical Transfer

The selection of alternate target glosses based on sentence context is governed by the AUs and TUs which are described in 4.3 above. Both these types of units offer the possibility of changing the LEX that is used for retrieval. AUs are applied before the sentence is parsed, while TUs are applied just prior to target dictionary lookup. Most types of TUs depend on the information produced by the parser, but the safety-net strategy allows some units to take effect even when the sentence cannot be parsed.

Lexical transfer itself is accomplished by retrieving from the Spanish target dictionary the record corresponding to the LEX of the English entry. This record is then examined to determine whether or not an alternate entry should be retrieved. Codes in the main target entry indicate the existence of additional records for microglossary entries, irregular plural nouns, irregular verb nominalizations, and adjectival translations for premodifying nouns and present participles.

5.5.3 Syntactic Transfer

The strategy for syntactic transfer is to utilize the surface structure of the source sentence, performing whatever operations are necessary in order to generate a grammatically correct and stylistically acceptable sentence in the target language. Whenever possible, the linear order of the phrasal units is preserved in order to maintain the overall cohesion of the text. Correct Spanish syntax is achieved through the manipulation of function words. Topicalization is also used to handle certain instances of preposed noun phrases. One situation that requires the postposing of the subject is that of a clause in which the finite verb is the final element.

The transfer component includes three different types of syntactic transfer rules. The first type is triggered by a lexical item in the English source text or the presence of a particular surface structure in English. The application of these rules generally results in a modification of the LEX used for retrieval of the Spanish gloss and therefore must be performed before the target lookup. The second type is a general syntactic transformation that converts a surface structure of
English to a surface structure of Spanish. These transformations are independent of the lexical items involved and can be performed either before or after target lookup. In ENGSPAN, they are performed prior to the target lookup. The third type of syntactic transfer rule depends on the characteristics of the lexical item in the Spanish target text and therefore must be applied after target lookup.

Examples of source-triggered rules are the transformation of *there + be* into the Spanish *haber* construction, and introduction of a first person subject in a sentence such as *Let us begin*. General rules include the ordering of the elements of the verb string, the insertion and ordering of clitic pronouns, the suppression of subject pronouns, and the nominalization of gerundive phrases that function as the object of a preposition. Examples of rules that require access to information about the target lexical items are subject-verb agreement, the ordering of the elements of the noun phrase, the triggering of the subjunctive mood and imperfect tense, and the transformation that reverses the syntactic roles of the subject and object with verbs such as *gustar* and *faltar*.

Syntactic transfer works best when the input string has been successfully parsed. Some rules, such as those that deal with subject-raising and subject postposition, can only be applied if the necessary information is available from the parse. For other rules, however, the transfer algorithm contains sufficient rules to make up for the lack of a complete parse. For example, verb strings with a limited amount of interruptive material can usually be restructured correctly. Most noun phrases, including many with multiple premodifiers and conjoined structures, can also be identified and rearranged into Spanish word order.

### 5.6 Synthesis Component

When the transfer component has completed its work, each target lexical item is associated with a group of features which specify the inflectional endings to be synthesized. Additional codes indicate whether the item is to receive a derivational prefix or suffix and whether the item is to be preceded by function words such as the
definite article, personal *a*, or the preposition *de*.

Verb synthesis is performed first, since the resulting form may serve as the input to the other synthesis procedures. The next task is the generation of the surface forms of the reflexive, direct object, and indirect object pronouns. Then all elements of the noun phrase are inflected for gender and number. This step is followed by prefixation and the insertion of function words. The final steps in synthesis are the introduction of phonologically determined variants and the adjustment of capitalization and diacritics.

6 Practical Experience

6.1 System Maintenance

6.1.1 Dictionaries: SPANAM

As indicated earlier, the Spanish source dictionary had 61,282 entries and the English target 58,485 as of December 1985. The program for updating the SPANAM dictionaries is user-friendly. Many default codes are entered by the update program automatically. Even though there are now 211 possible fields in which codes can be entered, as opposed to the original 82, almost all of them can be specified using mnemonic descriptors and code names.

Today, updating is done largely on the basis of production text. Every job reveals ways in which the dictionaries can be improved, either with new glosses for individual words or idiomatic phrases, especially with technical terminology, or deeper coding of existing entries. The steady, ongoing development of the dictionaries (Table 1) has ensured both a decrease in unfound words, with advantages for program effectiveness, and closer correspondence to the type of language that is used in the Organization, leaving less work for the post-editor.
Table 1. Dictionary Sizes, PAHO MT Systems, 1976-1985

<table>
<thead>
<tr>
<th>Year</th>
<th>SPANAM</th>
<th>ENSPAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>1976</td>
<td>4,000</td>
<td>3,500</td>
</tr>
<tr>
<td>1977</td>
<td>7,836</td>
<td>7,341</td>
</tr>
<tr>
<td>1978</td>
<td>38,506</td>
<td>38,376</td>
</tr>
<tr>
<td>1979</td>
<td>48,289</td>
<td>53,303</td>
</tr>
<tr>
<td>1980</td>
<td>50,921</td>
<td>55,792</td>
</tr>
<tr>
<td>1981</td>
<td>53,785</td>
<td>51,187(^1)</td>
</tr>
<tr>
<td>1982</td>
<td>54,383</td>
<td>52,223</td>
</tr>
<tr>
<td>1983</td>
<td>56,247</td>
<td>53,326(^3)</td>
</tr>
<tr>
<td>1984</td>
<td>61,240</td>
<td>58,440</td>
</tr>
<tr>
<td>1985</td>
<td>61,282</td>
<td>58,485</td>
</tr>
</tbody>
</table>

Table 2. Translation Speeds, SPANAM, 1979-1985

<table>
<thead>
<tr>
<th>Year</th>
<th>Best clock time</th>
<th>Average CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>wpm</td>
<td>wph</td>
</tr>
<tr>
<td>1979</td>
<td>160</td>
<td>9,600</td>
</tr>
<tr>
<td>1980</td>
<td>176</td>
<td>10,560</td>
</tr>
<tr>
<td>1981</td>
<td>192</td>
<td>11,520</td>
</tr>
<tr>
<td>1982</td>
<td>580(^4)</td>
<td>34,800</td>
</tr>
<tr>
<td>1983</td>
<td>700</td>
<td>42,000</td>
</tr>
<tr>
<td>1984</td>
<td>710</td>
<td>42,600</td>
</tr>
<tr>
<td>1985</td>
<td>1,500</td>
<td>90,000</td>
</tr>
</tbody>
</table>

\(^1\)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.

\(^3\)1,000 irregular verb forms were deleted by a special-purpose program.

\(^4\)Reflects change to VSAM lookup.
As indicated earlier, the translator notes the changes that are needed at the time of post-editing, and later updates the dictionaries. An hour is reserved for this work at the end of the day. If there is no production pending, the translator may spend extra time on dictionary work; on the other hand, if there is pressure, the dictionaries may have to be postponed for a while. Because of this integration of dictionary-building into the work of the translator, the post-editing cost is no longer an element that can be clearly identified.

6.1.2 Dictionaries: ENGSPAN

ENGSPAN has the same user-friendly software as SPANAM for updating its dictionaries. As of December 1985 the English source dictionary had 45,614 entries and the Spanish target had 47,545.

The AID support provided for two half-time dictionary assistants, one a linguist of English mother tongue and one a translator of Spanish mother tongue. A new deeply coded source entry costs from $.60 to $1.00; Spanish target glosses that require research are about the same. Simple changes in existing entries average about $.25 each. The semantic coding of medical terms was accomplished semi-automatically by using the SEARCH-AND-REPLACE function on the word processor, macros, and the Sort utility to identify and group terms that were morphologically related.

6.1.3 Other System Maintenance

The SPANAM and ENGSPAN programs, as well as the support software, are maintained by the two staff computational linguists, who carry out these tasks in addition to development work.

Hardware and system support are provided by the Pan American Health Organization. There is no separate charge for utilization of the computer, either for time or storage space. The programs execute in a partition of 1,024K and utilize 1MB on disk for work space, 8.5MB for program libraries, and 33.2MB for dictionaries and other permanently mounted files.
6.2 Testing

6.2.1 SPANAM

The output of SPANAM is subject to daily scrutiny by the project staff. In addition, frequent demonstrations are given for visitors using random text as input, and these results are examined as well.

An experimental version of the program is used to test and debug system enhancements resulting from production feedback and from the research and development being done for ENGSPAN. Before the experimental version of the program replaces the production version, a control test is performed by translating the same text with both programs and comparing the output using the Document Compare software available on the Wang OIS/140. The CPU time, throughput time, and number of disk I/O operations for both versions are also compared.

6.2.2 ENGSPAN

An experimental corpus of over 50,000 words was selected at the beginning of the project. Sentences are chosen from this corpus for the testing of specific program modules. Following every major enhancement of the algorithm or dictionary, the corpus texts are re-translated and the results compared with previous translations. Production texts, which no longer undergo any previous review by the project staff, are also used to supplement the original corpus. After the translation has been post-edited and delivered to the requesting office, the dictionary is updated by the addition of new words and missing codes, and the translation is then rerun. Problem sentences from these random texts are retained for use in subsequent development tasks.

6.3 Measurement Criteria

6.3.1 Speed: CPU/Throughput Time

The speed of SPANAM has steadily improved over the years (Table 2). When major changes are made in the program, care is always
taken to make sure that they do not cause any significant degradation of speed. With the new 4381 computer, which executes 5 million instructions per second, CPU time has been measured at 8,250 words per minute and clock time at 1,500 words per minute. The best throughput speeds are obtained at night, when there are fewer users working on the computer. During the day, turnaround at peak periods can be considerably slower. The speed is adequate for the current load of production.

With ENGSPAN, the CPU and clock times are 1,700 and 836 words per minute, respectively. These speeds tend to decrease as the coverage of the ATN grammar is expanded.

6.3.2 Comparison with Human Translation Speed

With a trained post-editor, throughput time using SPANAM or ENGSPAN is never slower than that of a human translator. The range is from one and a half to four times as fast, with the average falling between two and three times as fast. The output of a translator working as post-editor ranges from 4,000 to 10,000 words a day, depending on all the factors mentioned earlier, as well as the actual difficulty of the text. On the other hand, human translators working in the international organizations commonly produce around 2,000 words a day, with some services reporting an average of 1,500 and others an average of 2,500. It is possible to reach 3,000 or even higher, but usually not on a regular basis. It should be kept in mind that these rates do not take into account the time of the staff who transcribe the translators' dictation. Freelance translators report higher rates (Leonard, 1985). Given the variability of both sets of figures, it would be difficult to make any hard-and-fast comparisons. However, for the same person using both modes it might be possible to draw some conclusions: one translator who post-edits for SPANAM reports that she consistently produces about three times as much output with MT as she does in the traditional way.
6.3.3 Quality: Correctness

No systematic analysis of errors produced by SPANAM or ENG-SPAN has been done. Three consultants were engaged under different contracts to evaluate the overall status of the project: Professors Yorick Wilks (1978), R. Ross Macdonald (1981), and Michael Zarechnak (1981). While they commented on general characteristics of the output, they were more concerned with underlying processes that might produce the errors than the errors as such. Referring to the quality of the output, Professor Macdonald (1981:7) reported:

The current output is rather good. If a human being had written it, perhaps the output would be considered to be defective in many respects. When it is known, however, that it was produced by a machine, the basis of judgment shifts, and the output seems really very presentable. Any person of good will can understand this output, and I assume that no misleading translations have been discovered that would vitiate the intent of any article.

At the end of the period of the AID grant, the status of ENGSPAN was also evaluated. Its practical implementation was assessed by Ruhl Information Management. The output of the parser was examined by Professor Naomi Sager, who reported as follows:

An ATN English grammar of considerable coverage has been developed ... The number and complexity of constructions handled by the grammar has increased to the point where relatively few types of input sentences are not parsed adequately for subsequent translation ... While some grammar development remains to be done, it is impressive how much of the input material can be parsed with relatively few failures.

The Spanish synthesis was the subject of an evaluation by Professor William Cressey.
6.3.4 Postediting Effort vs. Human Translation Quality

The question of effort required for post-editing is inextricably tied up with standards for human translation. Both these issues are colored by subjective criteria. In Section 2.2 above there was a discussion of linguistic strategies for reducing the time spent on post-editing. The "quick-fix" post-edit definitely takes less time than a "syntactic" post-edit.

In an effort to see how translators would handle some of the same sentences that had been fixed up quickly in post-editing, a set of 17 Spanish source sentences was given to 12 trained translators who were asked to provide spontaneous human versions in English (Vasconcellos 1986). In the results, not one sentence was translated twice in the same way; apart from lexical differences, there were a variety of combinations and permutations in the ordering of the various phrasal elements. However, when the respondents were subsequently shown the "quick-fix" alternatives, they agreed that the latter were at least as good, and sometimes even better, than what they themselves had proposed (probably because there was greater cohesion in the presentation of the semantic components – Vasconcellos 1985b). This exercise underscored the difficulty of measuring the quality of a translation.

6.4 Cost-Effectiveness

Because of all the variables involved, including in particular the purpose of the translation, it is usually rather difficult to make clear-cut comparisons between MT production and traditional translation at PAHO. However, with one large project, carried out in 1980, such a comparison was possible. About half the document, i.e., the part that was originally in Spanish, was fed to SPANAM, while the remaining half was farmed out for translation from English into Spanish by human translators who worked in the traditional mode. For 101,296 words of machine translation, the cost was $3,218, including a hypothetical cost for machine time, and 36 staff-days were devoted to the activity. Had the same number of words been farmed out to human translators, the cost would have been $8,296 and the number of staff- and contract-
days (based on an output of 2,000 words a day) would have amounted to 65.75. Hence there was a monetary saving of $5,078, or 61 percent, and the staff-days were reduced by 29.5, or 45 percent.

Sometimes it is hard to know whether or not SPANAM and ENG-SPAN are translating texts that would otherwise have been submitted for human translation. Quite possibly the user is less hesitant to request a machine translation than a human one. B. Dostert (1979) has reported such a phenomenon in a survey of 58 users of MT.

6.5 Subjective Factors

The SPANAM/ENGSPAN staff have come to the understanding that in the end an MT system will stand or fall depending on the human environment in which it is placed, and that some of the most important factors cannot be measured. In the broad sense, these include: long-term commitment, positive attitudes, innovative responses, creative problem-solving. At the more specific level they include also the real availability of input in machine-readable form, a cooperative spirit among the staff who must share the oversaturated word processing equipment, willingness on the part of the translator to use the word processor for long periods, resourceful post-editing, and a host of other factors of nonresistance that are seldom taken into account.

In addition, when human translators are enlisted as post-editors, they must have a positive attitude toward the capabilities of MT, and, for true gains in productivity, they must be willing to use the keyboard and to become adept with the special editing features that have been developed for the word processor.

In dealing with the output, there must be flexibility in regard to quality. For example, if the rapporteur of a meeting has an hour in which to write up what her speakers said, and she can't understand the Spanish without a translation, there must be a "can-do" type of staff that will produce a document that can be worked from. The need met is the true criterion.
7 Discussion of the Approach Adopted

Macdonald (1979:130-145) pointed out that MT systems tend to polarize toward either an empirical or a theoretical approach. Development of the empirical system proceeds "on the basis of actual experience with appropriate texts" (1981:1), whereas the theoretical approach begins by postulating the adequacy of some particular model of language description which it is hoped will be able to cover all contingencies (1979:130, 1981:1).

Each position has its advantages and its disadvantages. In an empirical approach, a big advantage is that the system concentrates on only that which is of immediate usefulness for the task at hand. Yet as the system expands, "it becomes difficult to add in new operations without reworking some of what has already been done" (1981:1). The theoretical system, on the other hand, is able to proceed with less disruption, but the disadvantage is that "it is extremely difficult to predict as to which of all the complexities will actually arise; complexities may be foreseen and planned for in the system which do not actually appear in the type of texts to be translated" (1981:1).

Rather than advocating one extreme or the other, Macdonald believed that benefit was to be gained from both; ideally, he felt, the two positions should be combined in a melded system (1979:143):

On the whole, the best approach is a compromise between the two extremes, a basically empirical approach in which, however, the researchers strive for an overall perspective (1981:1) …

The preliminary research on the empirical system will serve the purpose of establishing the nature and extent of the problem of machine translation clearly and definitively. When the nature of the problem has been recognized as fully as possible, a rigorous and elegant solution of the problem can be devised (1979:145).

This view, which came through strongly in the 1981 recommendations of Macdonald and Zarechnak, is what has guided the development of SPANAM/ ENGSPAN in the last four years. While SPANAM
began as a purely empirical system, ENGSPAN is a melded system. The flexibility of the basic SPANAM/ENGSPAN architecture has made it possible to introduce a theoretical focus while still preserving that part of the working system which could be used both in the interim and as the safety net in the event of failed parses.

Rather than weighing the techniques of one MT system against those of another, it is felt that a more fruitful approach is to establish certain "positive criteria of relevance" (inspired by Halliday, 1979):

- Who are the system's users?
- In what environment is it being implemented?
- What purpose does it serve?
- On what basis is its use to be justified?

The strength of a system lies ultimately in its capacity to be relevant for its users, its environment, its purpose, its justification. Judgment from this standpoint is believed to be more effective in the long run than evaluation of the relative success or failure of a given theory.

By these standards, SPANAM and ENGSPAN have proven themselves over the years of ongoing production: through the accessibility of their programs, the user-friendliness of their dictionaries, their rapid throughput time, the broad range of texts for which they produce usable translations, and the savings they have effected in terms of both time and money.

Acknowledgments

The authors wish to express special acknowledgment to their co-worker Lee Ann Schwartz. As the second computational linguist working on ENGSPAN, Lee has given unstinting support to the project and its objectives.

Growth of the dictionaries is attributable in great part to the contributions of Rebecca Naidis, who served as post-editor of SPANAM, and Susana Santangelo, who has worked on ENGSPAN. Other assistance on the dictionaries has been given by Catherine Howe, David
Landeck, José Luiz Meurer, Patricia Schmid, and Lucretia Vanderwende. The workflow is coordinated by Clara Inés Rubiano.

The occasional consultants to the project since 1981 have been R. Ross Macdonald, Michael Zarechnak, Leonard Shaefer, William Cressey, Mary Jane Ruhl, Naomi Sager, and Veronica Lawson.

Within PAHO, the project has had continuing support and encouragement over the years from Luis Larrea Alba, Jr., Chief of the Department of General Services.

Finally, both PAHO and the project are most appreciative of the financial support received from the U.S. Agency for International Development.
References


Appendix: Sample ENGSPAN translations

Sentences that were parsed completely are flagged with an “OK.” Partial parses are indicated with a “PP,” and those sentences that could not be parsed by the current version of the program are marked “NO.” “TU” means that a transfer rule has been applied, and “SD” means that a word was not found in the source dictionary.

**Sample Input**

We request that each potential participant send biographical information and a 100-word abstract of the demonstration or paper, so that we can select those who will make the greatest contribution to a useful exchange of information at the symposium.

The task of hiring and assigning staff to perform the work is one which must be completed prior to training.

Laboratory studies have shown marijuana to impair perceptual and perceptual-motor functions important to driving.

**ENGSPAN Output**

OK TU Solicitamos que cada participante potencial envíe información biográfica y un resumen analítico de 100 palabras de la demostración o documento, para que podamos seleccionar los que harán la contribución mayor a un intercambio útil de información en el simposio.

OK La tarea de contratación y asignación de personal para realizar el trabajo es una que se debe completar anterior a adiestramiento.

OK TU Estudios de laboratorio han revelado que marihuana deteriora funciones de percepción y perceptomotrices importantes a conducción.
The mothers who lived within a 5-mile radius were asked to bring their children to the vaccination center.

Often the cold chain is thought to refer only to the refrigeration of vaccine.

The relationship between dietary fat and mammary carcinogenesis in experimental models will be presented and, finally, the possible relationship between dietary fat and hormones will be discussed.

A privately organized by publicly funded foundation is responsible for primary health care in the interior of the country where 10% of the population lives.

The National Program for Drinking Water Supply has the goal to expand to full coverage the already existing distribution system.

A las madres que vivían dentro de un radio de 5 millas se les pidió que trajeran a sus hijos al centro de vacunación.

A menudo la cadena de frío se piensa que se refiere solo-mente a la refrigeración de vacuna.

Se presentará la relación entre grasa en la alimentación y carcinogénesis mamaria en modelos experimentales y, finalmente, se tratará la relación posible entre grasa en la alimentación y hormonas.

Una fundación privadamen-te organizada pero públi-camente financiada es res-ponsable de atención pri-maría de salud en el interi-or del país donde vive un 10% de la población.

El Programa Nacional para Abastecimiento de Agua Potable tiene la meta para ampliar a cobertura total el sistema de distribución ya existente.
Other analyses will test the epidemiologic association between 1) ethylene oxide exposure and leukemia and 2) *PAH exposure in the cola hydrogenation process and cancer of the respiratory system, urogenital system, and the skin.

Otros análisis examinarán la asociación epidemiológica entre 1) la exposición de óxidos de etileno y leucemia y 2) la exposición de PAH en el proceso de hidrogenación de carbón y cáncer del aparato respiratorio, aparato urogenital, y la piel.