SPANAM AND ENGSPAN: MACHINE TRANSLATION AT THE PAN AMERICAN HEALTH ORGANIZATION

Muriel Vasconcellos and Marjorie León[1]
Pan American Health Organization
525 Twenty-third Street, N.W.
Washington, D.C. 20037

1 PROJECT HISTORY AND CURRENT STATUS

1.1 OVERVIEW

Spanish-English machine translation (SPANAM) has been operational at the Pan American Health Organization (PAHO) since 1980. As of May 1984, the system's services had been requested by 87 users under 572 job orders, and the project's total output corresponded to 7,040 pages (1.76 million words) that had actually been used in the service of PAHO's activities. The translation program runs on an IBM mainframe computer (4341 DOS/VSE), which is used for many other purposes as well. 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. A trained full-time post-editor, working at the screen, produces 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 versus 1,500-3,000 for human translation). The post-edited output is ready for delivery to the user with no further preparation required.

The SPANAM program is written in PL/I. It is executed on the mainframe at speeds as high as 700 words per minute in clock time (172,800 words an hour in CPU time), and it runs with a size parameter of 215 K. Its source and target dictionaries (60,150 and 57,315 entries, respectively, as of May 1984) are on permanently mounted disks and occupy about 9 MB each.

While SPANAM continues to build its reputation as a work horse, at the same time development is well advanced on a parallel system that translates from English into Spanish, ENGSPAN. Also written in PL/I, ENGSPAN uses essentially the same modular system architecture that has been developed for SPANAM, but it is conceived on the basis of up-to-date linguistic theory leading to rule-based strategies for the parsing of syntactic and semantic information. 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 UN family, serving in the latter instance as Regional Office of the World Health Organization. In addition to its headquarters staff of 546 in Washington, PAHO has a field staff of 652 that supports both the operational programs in its 10 Pan American centers, located in eight different countries, and its 30 representational offices, in 28 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, 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

[1] Respectively, Chief, Terminology and Machine Translation Program, PAHO, and Senior Computational Linguist, Machine Translation Project, PAHO.

0362-613X/85/020122-136$03.00

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 work was begun in 1976. A team of three part-time consultants worked for the Organization for two years, and one of these consultants remained with the project for a third year. In the beginning the approach drew upon a number of the principles that had evolved at Georgetown University in the late 1950s and early 1960s in the course of work on the Russian-English system known as GAT (Georgetown Automatic Translation, described in Zarechnak 1979).

The first language combination to be addressed was Spanish-English. The consultants had opted for this direction recognizing that results could be available earlier than if they had started with English as the source. Parallel efforts were concentrated on the architecture itself of the system 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 eight other 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 (at that time the verbs required full-form 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. Efforts from English into Spanish had produced one page of text.

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 of 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 full-time computational linguist was recruited and assigned to the project in 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. To begin with, the computational linguist took full charge of the system software, replacing the consultants. Up to that time production had not been feasible because there was no morphological analysis of verbs: the failure to find a high percentage of inflected verbs had meant that many sentences in random text were barred from even the most rudimentary analysis. Thus the first order of business was to develop the needed morphological lookup.

At the same time, the operational problems of text input, another major impediment to production, were also 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 conversion program was written which handled the differences in representation of characters, solved ambiguities of punctuation, and made certain decisions about the format. From the time this program was installed, any Spanish text that had been keyed onto the word processor, regardless of the purpose for which it was entered, was available for machine translation.

1.3 OPERATIONAL PHASE: 1980-PRESENT

As production gained momentum, the MT staff was increased by the assignment of a full-time post-editor and by greater participation of the terminologist as head of the project.

Over the next two years, the sources of machine-readable text for SPANAM increased at a steady pace. The use of word processing at PAHO expanded, and, in addition, another mode of input became possible through optical character recognition (OCR). Whereas word processing had previously been restricted to special services provided by a typing pool, after the installation of word-processing hardware throughout the headquarters building (Wang OIS/140), all program units eventually came to participate in the text-processing chain. Furthermore, the optical character reader (a Compuscan Alphaword II), previously used only for Telex transmission, was interfaced with the word-processing system; this meant that existing typewriters could also be used as input devices, and therefore that texts could be prepared in the field and machine-read in Washington.

With accelerated production, improvements to SPANAM have followed in tandem. From the beginning it has been the policy, and continues to be so today, 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, note is made of recurring problems at all levels. Capture of this information at the time of post-editing saves much work later on. The messages written by the post-editor on the side-by-side text serve as a basis both for updating the dictionaries and for making enhancements, as feasible, in the algorithm.

In this way the Spanish source dictionary had grown to a total of 60,120 entries as of May 1984. Of this total, 94% were bases or stems and 6% were full forms, all with corresponding entries in the English target. Since

1981 the incidence of not-found words in random text has been well under 1% – 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 translations can be elicited.

In its four years of operation, SPANAM has become not only wiser but more efficient as well. The program's speed of run time has increased from 160 words per minute to over 700 wpm. Yet the algorithm, even though it has sustained a major reorganization into modular structure and regularly undergoes enhancement, remains approximately the same size (2,085 statements as of May 1984).

Further details about the working environment of SPANAM are given in sections 2 and 6.

1.4 SYSTEM DEVELOPMENT SPANAM/ENGSPAN: 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 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 also be used for ENGSPAN. 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.[2] This funding has made it possible to have a second computational linguist for the grant period, as well as consultants and part-time dictionary assistants who have undertaken specific tasks within the approved plan of work.

With the added manpower, the project has made significant progress on the English-Spanish algorithm. Particular focus has been 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 has been 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. Deep syntactic and semantic coding has been introduced for dictionary entries corresponding to a sizable proportion of the experimental corpus of 50,000 running words.

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. Since post-editing was inevitable, it was felt that a pre-editing step would be anti-economic: the advantages to be gained 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 would be scarce: the pre-editor would need to be well prepared not only in translation skills but also in knowledge of the algorithm, or at least a number of its capabilities and limitations.

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 whatsoever are required. With inexperienced operators, however, and with texts read automatically by the OCR, 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.

Production texts are run only once. Demonstrations are always performed on random text.

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

2.2 POST-EDITING POLICY

2.2.1 GENERAL POSITION

The multifaceted approach to post-editing is an important feature of SPANAM. On one level, consideration is given to the user's needs and capabilities and to the purpose of the translation. At the same time, specific linguistic strategies developed by the project are often used in order to minimize the recasting of certain unwieldy constructions which frequently recur – mainly the result of verbs in sentence-initial position in Spanish. Finally, 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 SPANAM output which are not handled by syntactic rules.

The degree of post-editing is determined by:

1. the purpose of the translation,
2. the user's own resources for editing,
3. the time frame, and
4. 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 post-editor. On the other hand, if it is to be published without much further review, the post-editor devotes careful attention to the quality of the text. These factors are determined in a conference 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 is 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.

In an open, "try-anything" (Lawson 1982:5) system such as SPANAM, with its highly varied applications, experience has led to the conclusion that post-editing requires a trained professional translator. 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", the SPANAM experience suggests that this conclusion would be valid only for technical experts working on a text for information purposes only. Even in such cases, the technicians at PAHO are encouraged to request a more careful translation of passages that are of particular interest.

Only an experienced translator will be aware of the words whose variable meanings are dependent on extralinguistic context. For example *proyecto* in Spanish can mean 'project', 'proposal', or 'draft', and the choice depends on full knowledge of the situation to which the text refers. *Esperar* can mean 'hope' or 'expect', and the distinction is essential 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.

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 proved to be effective in the processing of SPANAM output. For the inexperienced post-editor, the most time-consuming task is the recasting that is deemed to be "required" when machine-translated constructions turn out to be ungrammatical or intolerably awkward in the English output. SPANAM has addressed this problem by developing a series of "quick-fix" post-editing expedients (QFP) for dealing with the typical problems of Spanish-to-English translation. For example, certain maneuvers are suggested as being useful in the case of fronted verb constructions in Spanish, which occur frequently and present difficulties for the standard SVO pattern in English. The purpose of the QFP is to minimize the number of steps 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 might 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 also Vasconcellos, in preparation). The examples below compare QFPs with solutions that were actually proposed by translator-post-editors (traditional human translation – THT).

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 there was initiated 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.

(2) 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 the foregoing approach, wherever feasible, 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 both in translation and, if possible, in linguistics as well.

It is always emphasized in SPANAM that editorial changes should be kept to the minimum needed in order to make the output intelligible and acceptable for its intended purpose.

2.2.3 WORD-PROCESSING STRATEGIES

The SPANAM 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 has a set of special aids developed for the 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 routines that deal with the character strings that most often have to be changed in SPANAM output. For example, only a single "glossary" keystroke is needed to perform the following editorial operations:

SEARCH-and-DELETE:

the, of, there, to, in order to

SEARCH-and-REPLACE:

from/of, for/of, for/by, in order to –/for –ing, 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 on the surface to be almost trivial, 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 the light pencil appear to be less effective. 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.

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 have 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. Printing is kept to a minimum; finished production is delivered to the user either on a diskette or by a telephone call notifying the office that the job is available on the system.

2.3 POST-EDITING VIS-À-VIS OTHER ASPECTS OF THE SYSTEM

In the SPANAM environment there is a close link between post-editing and the other aspects of the system. The staff post-editor has been trained to update the dictionaries, and currently almost all the dictionary work on SPANAM is done by this person. Required changes to the dictionaries are proposed at the time of post-editing. Hence there is no need to go through the text a second time. Also, glosses and other solutions seem to come to mind most readily when the whole text is actually being worked on. The post-editor, if adequately trained in updating, is in the best position to see what dictionary changes are necessary in order to deal with the specific constructions that tend to recur in production translations.

The post-editor also alerts the computational linguist to areas where the algorithm needs improvement.

2.4 Integration of the System into a Complete Translation Environment

SPANAM/ENGSPAN, the terminology activity, and the traditional human translation unit are in the process of being merged into a single program of language services at PAHO. While human and machine translation even now coordinate the workload to a certain extent, as of May 1984 there was not yet any centralized screening of incoming jobs. It is expected that such a triage will make it possible to maximize the effectiveness and efficiency of the respective services.

Combination of the two activities will also make for a more rational utilization of the manpower available at any given time, with the staff being assigned to a variety of different duties, depending on both needs and skills. It is also a goal to reduce a given person's day in front of the screen from eight hours to six through the rotation of assignments.

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 for both English and Spanish.

In terms of the linguistic work of the human translator, SPANAM/ENGSPAN can help to lighten the load in a number of ways. To begin with, technical and scientific terms are retrieved in context, which means that MT is a sort of very efficient lexical data base. With an ordinary LDB, the translator has to go to the terminal (which is not usually at his desk for his use alone), sign on, and initiate a search. After he has performed all the mechanical steps, there is still the possibility that the term is not in the data base at all, and his effort will have been wasted. When this happens repeatedly over time, a cumulative 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, he is also told its degree of reliability and whether or not it is in the WHOTERM data base. The status of the terms is indicated by small superscript symbols which can be requested at the time the text is sent for translation.

WHOTERM is also on the word-processing system. Its general file contains: a definition of each term in English, translation 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 & Lowe 1983).

SPANAM also aids the translator with its 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 alternate glosses. One or more of these microglossaries can be specified at the time the job is submitted. The translator can also have a microglossary of his own in which he can store special terms he prefers to use.

It is also possible for SPANAM to provide alternate 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 undesired translation is eliminated by striking a single glossary key.

If the translator provides feedback in the form of suggested or requested changes in the dictionaries, the updating can be done immediately. Some of SPANAM's users 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 his changes so that they can be entered by a word-processing operator working from a tape.

The human translation service stands to benefit, also, from the sophisticated facilities that have been developed on the word processor for editing and housekeeping support.

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 precluded the use of 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 required 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, transfer unit lookup, target lookup, syntactic transfer, and target synthesis. The program includes backup modules for homograph resolution, verb string analysis, and noun phrase analysis, which are called in 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 *-mente/-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, including *-'s, -s', -s, -ly, -ed, -ing, -er, -est,* and *-n't.* 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 with the ending which was removed. If LEMMA exhausts all its possibilities, the word is checked against a small list of prefixes *(re-, non-, un-, sub-,* and *pre-).* If one of these prefixes can be removed, another lookup is performed. If this final lookup is unsuccessful, a dummy record is created for the word and a gap analysis routine is called. "Not-found" words are initially considered to be nouns 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 ENGSPAN keeps down the size of the dictionary while 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 word *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

SPANAM deals with 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, adjective, 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, 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, described 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 a 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 sub-dictionary of target glosses which can be set up for a particular subject area, discourse register, or specific user. 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, statistics, and scientific 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 target record containing the principal gloss for the word. Thus, no time is wasted looking for special translations of every word. More than one microglossary may be activated for the same translation, in which case they are listed and consulted in order of priority.

The transfer unit is a rule that is stored in the source dictionary and is retrieved after the analysis of the sentence has been completed. The existence of a transfer unit is indicated in the record corresponding to the individual source word. A transfer unit contains a condition to be tested and an action to be performed. Examples of conditions are:

- Subject of this verb has X feature(s) or is word W.
- 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 has N object(s).
- Context N word(s) to left/right contains word with X feature(s).

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 that are used to store information about its syntactic and semantic features. These features are used in both the analysis and transfer stages of the translation. For example, verbs and deverbal nouns are specified as occurring with one or more of the following codes: no object, one object, two objects, complement, no passive, locative, marked infinitive, unmarked infinitive, declarative clause, imperative clause, interrogative clause, gerund, adjunct, bound preposition, and object followed by bound preposition. Subject and object preferences can be specified as ±Human, ±Animate, and ±Concrete. Other fields are reserved for case frames. 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, 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, General, Temporary condition, Positive connotation, and Negative connotation. 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 Naomi Sager's *Natural Language Information Processing* (1981).

4.5 ANNOTATED SURFACE STRUCTURE NODES

In ENGSPAN, the structure produced by the parser consists of a graph containing nodes corresponding to each clause and phrase. Each node contains a list of 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 (1970, 1973) 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 which govern the synthesis of the 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.

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 assigned to each pair when it is added to the dictionary. The second half of the LEX is used to specify alternative 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 to 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. In this case, each alternative 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. In the case of 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 essential 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 by the unit. Examples of phrases 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 alternative 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, up to now, uses 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 phrases 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.

As indicated earlier, ENGSPAN's grammar rules are expressed in the form of an ATN. 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 May 1984 the ATN grammar had seven networks: sentence, clause, noun phrase, verb phrase, sentence nominalization, hyphenated compound, and prepositional phrase. 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.

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 that 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 which 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 by 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.

In the event of an unsuccessful parse, ENGSPAN is still expected to produce a translation. The longest successful path is always saved, and information from this "partial parse" can be used by the synthesis routines.

Local routines are used to analyze the remainder of the input string. These routines function as a "safety net". They resolve homograph ambiguities and analyze verb strings and noun strings, adding as much information as they can to the structural description of the sentence as a whole.

The ATN parsing algorithm is being developed in an independent PL/I program, using the ENGSPAN input and dictionary lookup modules.[3] It is also totally compatible with SPANAM. The network grammar is read in at runtime, 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 diagram in **Figure 1** shows the relationship between the parser and the "safety net" routines in ENGSPAN. The parser is to be incorporated in a similar way into SPANAM as well.

A complete description of the ATN grammar and parser will be found in the report to be submitted to the U.S. Agency for International Development at the end of the grant period (October 1985).

6 Practical Experience

6.1 System Maintenance

6.1.1 Dictionaries: SPANAM

As of May 1984, the Spanish source dictionary had 60,150 entries and the English target had 57,315. 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 almost exclusively on the basis of production text. Every job reveals ways in which the dictionaries can be improved, with either new glosses for individual words or idiomatic phrases, especially in the case of technical terminology, or deeper coding of existing entries. The steady, ongoing development of the dictionaries **(Table 1)** has ensured both a decrease in not-found words, with advantages for program effectiveness, and closer correspondence to the type of language used in the Organization, leaving less work for the post-editor.

As indicated earlier, it is the post-editor who notes the changes needed at the time of post-editing and who later updates the dictionaries. An hour is reserved for this work at the end of the day. When production permits, the post-editor may spend extra time on dictionary work; if there is pressure, the work may have to be postponed for a while. Because of the integration of dictionary-building into the work of the post-editor, the cost is no longer an element that can be clearly identified.

Table 1. Size of dictionaries, PAHO Machine Translation System, 1976-1984.

	SPANAM		ENGSPAN	
Year	Spanish	English	English	Spanish
1976	4,000	3,500		
1977	7,836	7,341		
1978	38,506	38,376		
1979	48,289	53,303		
1980	50,912	55,792		
1981	53,785	51,187 [1]	44,411 [2]	44,998
1982	54,383	52,223	40,107	41,358
1983	56,247	53,326 [3]	40,772	42,116
May 1984	60,150	57,315	41,210	42,638

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

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

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

6.1.2 Dictionaries: ENGSPAN

ENGSPAN has the same user-friendly software as SPANAM for updating its dictionaries. As of May 1984, the English source dictionary had 41,210 entries and the Spanish target had 42,638.

The AID project has provision 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 $0.60 to $1.00; Spanish target glosses that require research are about the same. Simple changes in existing entries average about $0.25 each.

6.1.3 Other System Maintenance

The SPANAM programs are maintained by the staff computational linguist, who carries out these tasks in addition to development work on ENGSPAN.

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 project has a permanently assigned partition of 512 K in core, as well as 1 MB on disk for work space, 8.5 MB for program libraries, and 33.2 MB for dictionaries and other permanently mounted files.

[3] A major portion of the parsing routines have been developed by Lee Ann Schwartz, who has participated in this activity on a full-time basis since August 1983.

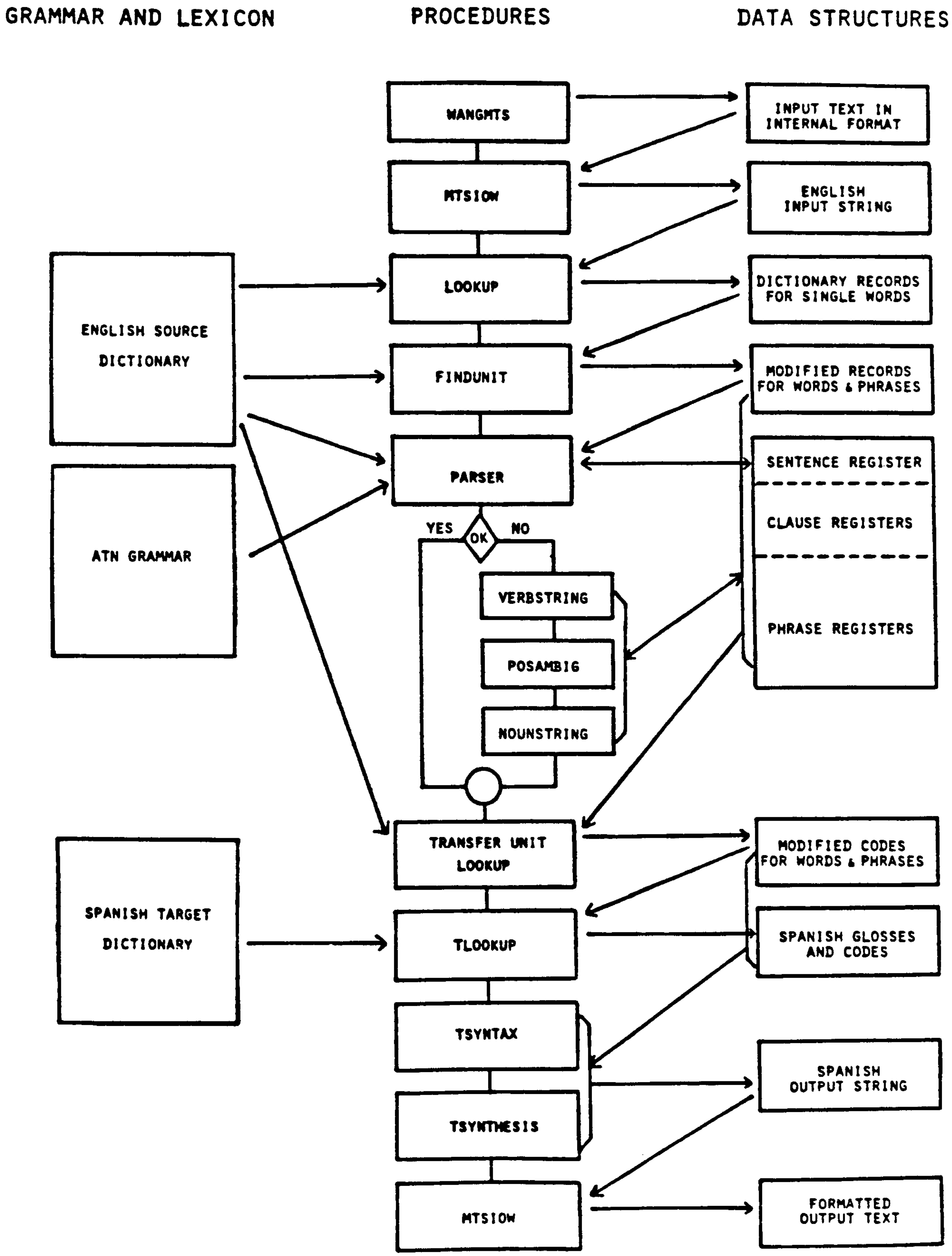

Figure 1. The relationship between the parser and the "safety net" routines in ENGSPAN.

6.2 TESTING

6.2.1 SPANAM

The output of SPANAM is subject to daily scrutiny by the project staff. In addition, weekly demonstrations are given for visitors, using random text as input.

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. This utility program compares the output character by character. The CPU time, throughput time, and number of disk I/O's 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 retranslated and the results are compared with previous translations. The system is also tested using new texts which are translated without previous review by the project staff. After the first translation run, the dictionary is updated in order to add not-found words and missing codes, and then the translation is 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

SPANAM's speeds, both CPU and throughput time, have steadily improved over the years **(Table 2)**. The best throughput time 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.

When major changes are made in the program, care is always taken to make sure that they do not cause any extensive degradation of speed.

The CPU and throughput times for ENGSPAN are 1,400 and 400 words per minute, respectively. These speeds are expected to decrease as the coverage of the ATN grammar is expanded.

Table 2. Translation speeds, SPANAM, 1979-1984.

	Best clock time			Average CPU time	
Year	wpm	wph	pages/h	wpm	wph
1979	160	9,600	38	Not available	
1980	176	10,560	42	Not available	
1981	192	11,520	46	3,184	191,000
1982	580 *	34,800	139	2,600	156,000
1983	700	42,000	168	2,880	172,800
1984	710	42,600	170	2,982	178,920

*Reflects change to VSAM lookup.

6.3.2 COMPARISON WITH HUMAN TRANSLATION SPEED

With a trained post-editor, SPANAM's output 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 SPANAM output ranges from 4,000 to 10,000 words a day per post-editor, depending on all the factors mentioned above, as well as the sheer difficulty of the text. On the other hand, human translators working in the international organizations commonly produce around 2,000 words a day, although some services report 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. Free lances report much higher rates. 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 error analysis has ever been done of SPANAM or ENGSPAN. Three consultants were engaged under different contracts to evaluate the overall status of the project: Professors Yorick Wilks (1978), 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.

6.3.4 Quality: Post-Editing Effort/Human Translation Quality

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

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, in preparation). No one sentence was translated twice in the same way; apart from lexical differences, there was 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, or in some cases even better, than what they themselves had proposed – probably because there was greater cohesion in the presentation of the semantic components (Vasconcellos 1985b, in preparation). 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 SPANAM production and traditional translation at PAHO. On one large project, however, such a comparison was possible because about half the original text was in Spanish and the other half was in English. The former was done on SPANAM and the latter was farmed out to 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 in the traditional mode, the cost would have been $8,196 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 (61%) and the staff-days were reduced by 29.5 (45%).

It is safe to say that the economy effected with SPANAM is sufficient to cover the salary of the post-editor and perhaps another salary at the same level as well. It may yet be some time, however, before the early investment in the project is fully recovered.

Sometimes it is hard to know whether or not SPANAM is translating text 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 this phenomenon in a survey of 58 users of MT.

In the past PAHO has farmed out a large percentage of its translation load. As SPANAM and ENGSPAN increasingly reduce that direct cost, there is clear evidence of savings.

6.5 Subjective Factors

The SPANAM 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 post-editor 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, if human translators are to be 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 and compromise 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 Approach Adopted

Macdonald (1979:130-145) has 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 (1981:1). In an empirical approach, the main advantage is that the system is more compact in that it concentrates on only that which is of immediate usefulness for the task at hand. The main disadvantage is that, 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 advocate 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 three years. While SPANAM began as a purely empirical system, ENGSPAN is now well on its way to being 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 a "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":

- 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 will lie 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 has proved itself through its four years of ongoing production: through the accessibility of its programs, the user-friendliness of its dictionaries, its rapid throughput time, the broad range of text for which it produces a usable translation, and the savings it has effected in terms of both time and money. ENGSPAN is soon to follow suit.

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 largely to the contributions of Rebecca Naidis, 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 occasional consultants to the project since 1981 have been Ross Macdonald, Michael Zarechnak, Leonard Shaefer, and William Cressey.

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 support received from the U.S. Agency for International Development.

References

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

Dostert, Bozena 1979 Users' Evaluation of Machine Translation: Georgetown MT System. Undated typewritten report later expanded into article of the same title in Henisz-Dostert et al.: 149-244.

Halliday, M.A.K. 1967 Notes on Transitivity and Theme in English. Part 2. *Journal of Linguistics* 3:199-244.

Halliday, M.A.K. 1979 *Explorations in the Function of Language.* Elsevier North-Holland, Amsterdam, New York, Oxford.

Henisz-Dostert, B.; Macdonald, R. Ross; and Zarechnak, Michael 1979 *Machine Translation.* Mouton, The Hague.

Kaplan, Ronald M. 1973 A General Syntactic Parser. In: Rustin, Randall: 193-241.

Kay, Martin 1982 Machine Translation. *AJCL* 8(2): 74-77.

Lawson, Veronica 1982 Machine Translation and People. *Practical Experience of Machine Translation.* North Holland, Amsterdam, New York, Oxford: 3-9.

Macdonald, R. Ross 1979 The Problem of Machine Translation. In: Henisz-Dostert et al.: 89-145.

Macdonald, R. Ross 1981 A Study of the PAHO Machine Translation System. Report of short-term consultant under Personal Services Contract APO-09435(WU1) [typescript, 13 p.]. Pan American Health Organization, Washington, D.C.

Rustin, Randall, Ed. 1971 *Natural Language Processing.* Algorithmics Press, New York.

Sager, Naomi 1981 *Natural Language Information Processing: A Computer Grammar of English and its Applications.* Addison-Wesley, Reading, Massachusetts.

Vasconcellos, Muriel 1984 Machine Translation at the Pan American Health Organization: A Review of Highlights and Insights. *Newsletter of the British Computer Society, Natural Language Translations Specialist Group* 14.

Vasconcellos, Muriel 1985a Management of the Machine Translation Enviornment: Interaction of Functions at the Pan American Health Organization. In: Lawson, Veronica, Ed., *Translating and the Computer 5: Tools for the Trade.* Aslib, London: 115-129.

Vasconcellos, Muriel 1985b Theme and Focus: Cross-Language Comparison via Translations from Extended Discourse. Ph.D. dissertation, Georgetown University, Washington, D.C.

Vasconcellos, Muriel In preparation Functional Considerations in the Postediting of Machine-Translated Output. (To appear in Computers and Translation.)

Winograd, Terry 1983 *Language as a Cognitive Process. Volume 1. Syntax.* Addison-Wesley, Reading, Massachusetts.

Woods, W.A. 1969 Augmented Transition Networks for Natural Language Analysis. Report No. CS-1 to the National Science Foundation, Cambridge, Massachusetts.

Woods, W.A. 1971 An Experimental Parsing System for Transition Network Grammars. In: Rustin, Randall: 111-154.

Zarechnak, Miachel 1979 The History of Machine Translation. In: Henisz-Dostert et al.: 1-87 [in particular, 29-30, 32, 134 ff.].