[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

DEVELOPMENT OF ENGLISH-SPANISH MACHINE TRANSLATION

Marjorie León
Pan American Health Organization

The Pan American Health Organization (PAHO) has been involved in the field of machine translation (MT) since 1976. Its Spanish-English machine translation system (SPANAM) became operational in 1980. The SPANAM system is described in Tucker, Vasconcellos, and León (1980) and Vasconcellos (in press). In 1982, work began on the development of the counterpart English-Spanish system (ENGSPAN). An experimental version of the translation program was in place by October of that year. In August 1983, PAHO was awarded a research grant by the U.S. Agency for International Development (AID) to provide additional support for the ENGSPAN project. This paper describes the approach which is being used for the design and implementation of ENGSPAN.

Current Working Environment

The MT project is staffed by four full-time employees. The head of the project is responsible for the management of both production translation and software development, as well as for the coordination of terminology. A posteditor handles all Spanish-English translation and updates the SPANAM dictionaries. The author is responsible for the design and implementation of ENGSPAN and the maintenance of the SPANAM programs and support software. A second computational linguist, funded through the AID grant, is also working on ENGSPAN. Outside consultants evaluated the feasibility of the project and have also participated in some phases of the development work.

The MT system is installed on an IBM 4341 mainframe computer operating under DOS/VSE. The project is assigned a partition of 512K, although our largest program only requires a total of 400K. The programs are written in PL/1. The dictionaries are VSAM files stored on a permanently-mounted disk. Both translation and dictionary updating is done in batch mode. For production translations, the source text is transmitted from the word processor (Wang OIS 140) to the mainframe via telecommunications, and the output is returned to the word processor for postediting. Dictionary updates, tests, and demonstrations can be submitted from either the word processor or the computer terminal. Program development is done from the terminal.

The turnaround time depends on the level of use of the computer at the time the job is submitted. Under optimum conditions, SPANAM can process about 700 words per minute of elapsed time. The CPU time ranges from 2,600 to 3,200 words per minute. Translations and dictionary updates can be submitted at any time during the day. Of course, longer jobs running during off-peak hours are the most efficient. The time required for postediting depends on the purpose for which the translation was requested. A polished translation can usually be produced at a rate of about 800-1000 words per hour.

As an organization, PAHO is involved in three different aspects of MT. It is the software developer, the user of the system, and the end-user of the translation. The system developers (linguists and programmers) and the system users (posteditors and dictionary coders) are members of the same team. In fact, everyone on the project staff has some experience in postediting, dictionary coding, and programming. This working environment makes the development staff keenly aware of the needs and desires of those using the system—both from personal experience and from listening to daily feedback. In turn, the posteditor has an understanding of how the algorithm works and can appreciate the relative complexity of the problems encountered.

While the developers are mainly concerned with the linguistic content of the programs, the operational environment is also kept in mind. A recent case in point involved the format in which the side-by-side output was received on the word processor. Before postediting could begin, a time-consuming glossary had to be run in order to remove the source text, unwanted format lines, spaces, and hard carriage returns. This problem was solved by expanding the output module to create a second file containing only the target translation with the necessary Wang control characters and format lines and no unwanted carriage returns. The same translation run can now produce both a target-only document on the word processor and a side-by-side document either on the Wang or the IBM (terminal and/or printer). An extra step in the production cycle was eliminated and the turnaround time improved.

Development Goals

ENGSPAN is being designed to produce Spanish translations of English texts. It is language-pair specific, but not subject-area specific. The input will not be restricted to any particular sublanguage or discipline, nor can it require pre-editing or the use of restricted syntax. The algorithm is being designed with expository text (both technical and general) in mind, but provisions will also be made for other types of text whenever possible.

Our goal is to produce high-quality raw output which requires only a limited amount of postediting to produce a finished translation. While the quality of the raw output is our main concern, ease of operation is also an important consideration. Dictionary updating should be mnemonic and the user should be required to supply only those codes which cannot be computed from other information already available to the system. The procedures for submitting translations, dictionary updates, dictionary backups, etc. should also be simple. Finally, the system should be efficient in its use of storage space and processing time.

When we reach a satisfactory level of quality, ease of operation, and efficiency, we plan to adapt the system to run on a microcomputer. This will make low-cost machine translation available to the PAHO Country Offices and Pan American Centers and to other cooperating institutions in the Member Countries.

The Experimental Corpus

An important part of our development strategy is the use of an experimental corpus. The corpus contains over 50,000 running words, taken from texts by different authors and dealing with a variety of health-related topics. It is large enough that it contains examples of a wide range of syntactic and semantic phenomena, yet at the same time it provides us with objective data on the relative frequency of occurrence of different types of constructions. We intend to concentrate our efforts on the types of syntax found most frequently in the corpus.

The Structure of the Dictionaries

The system uses separate files for the source and target dictionaries. The records in both files have 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.

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. Host 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 words with highly irregular morphology and for homographs (words which can function as more than one part of speech).

Several source items may be linked to the same target gloss by assigning it the same LEX. For example, irregular forms of the same verb or alternate 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 alternate gloss is distinguished by coding in the second half of the LEX. Two positions are used to designate terms belonging to microglossaries by subject area, two for glosses corresponding to different parts of speech, and two for context-sensitive glosses for polysemous words.

The dictionaries contain two types of multi-word entries: substitution units (SU) and analysis units (AU). The key for a multi-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 both cases, the words must occur consecutively in the sentence in order for the unit to be activated.

The basic SU contains from two to five words. This limit of five words was expanded to a maximum of 25 words through a process of nesting one or more such units in a long semantic unit (LSU) which is retrieved on a second pass through the phrase lookup module. When an SU or LSU 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. An SU record has the same format as a single-word entry and may contain all the same codes. In addition, it may contain a character string which indicates the part of speech of each of its members.

The analysis unit is limited to five words. The AU 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.

Reversal of the SPANAM Dictionaries

At the time we began work on ENGSPAN, the SPANAM dictionaries were stored as ISAM files. They contained approximately 54,000 pairs of entries, including 13,000 single words and 3,000 phrases which had been hand-coded by the MT staff, 9,000 general vocabulary items, and 29,000 medical terms. We also had very user-friendly programs for updating and displaying the dictionaries. In order to take advantage of this considerable investment of time and money, it was decided to use the same record format and to write a program to reverse the dictionaries.

Each dictionary was copied to tape, skipping the records for multi-word entries, inflected forms, auxiliary verbs, prepositions, and items coded as deprecated terms. The Spanish records were sorted into numerical order by LEX and the English records into alphabetical order by the lexical item. The new files were checked for duplicate keys. Whenever more than one record with the same LEX was encountered, the set of records was examined and reordered according to criteria based on the part of speech (verb, noun, adjective, other), reliability code (highest to lowest), and source code (PAHO term, medical term, general term). When the dictionaries were reloaded, the first record became the main entry for the word. The key of each subsequent record was made unique by concatenating an asterisk on the end of the word or adding 1 to the last digit of the LEX.

When the reversed dictionaries were printed out in side-by-side format, multiple source and target entries were grouped together. The dictionary coder then reviewed these entries to determine whether the first entry in each set was the most appropriate entry for the ENGSPAN system and to identify those entries which should be treated as homographs. After the necessary adjustments were made, the extra entries on each side were deleted automatically. Figure 1 shows a page from the newly reversed dictionary, prior to any human intervention.

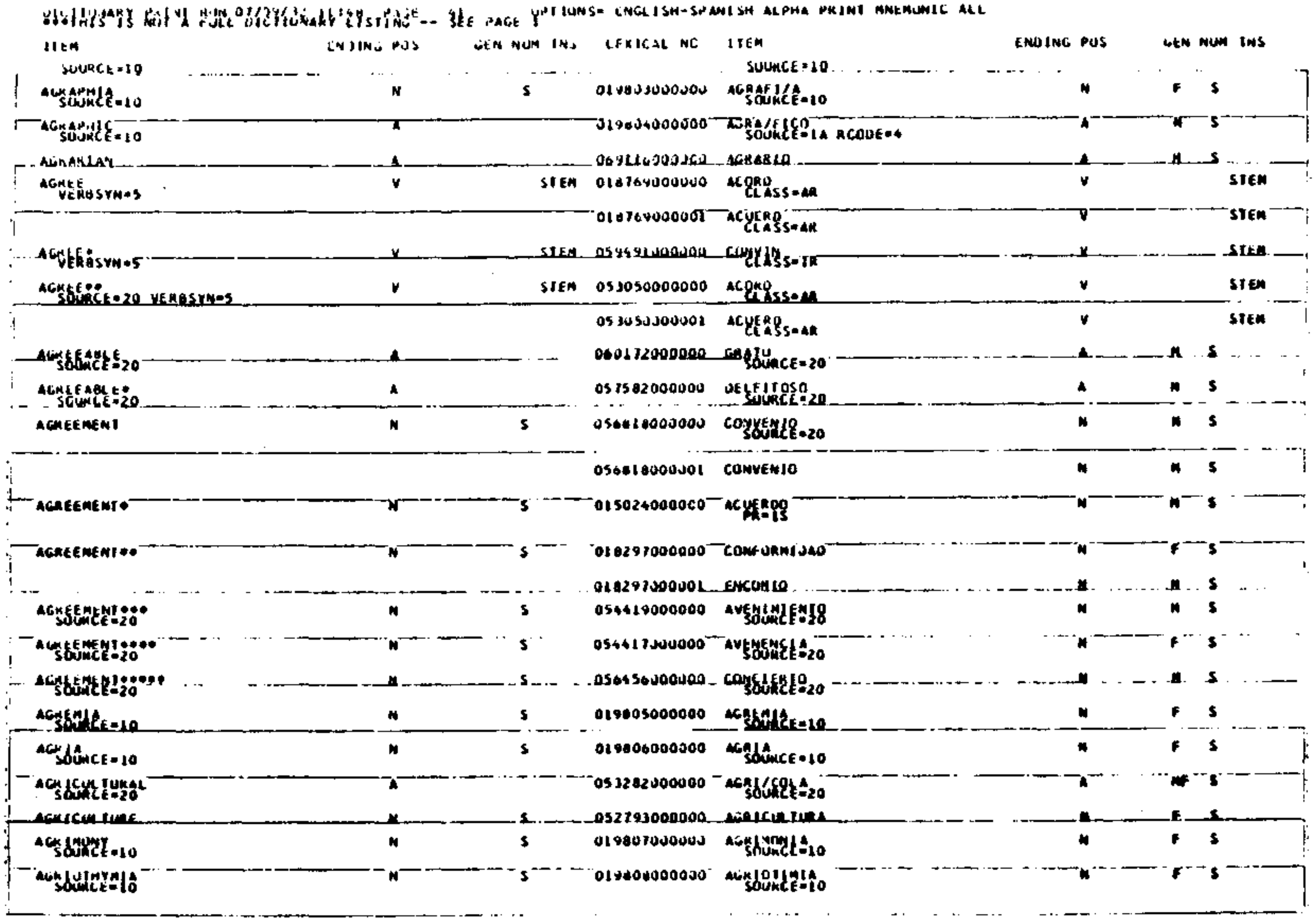
The reversal program produced a total of 44,404 English source entries, including 4,725 duplicates. After the duplicates were removed, and new entries were made for the auxiliary verbs and prepositions, the dictionaries contained approximately 40,000 pairs. Although some glosses still need to be improved, most of the codes for part of speech, gender, and number are correct.

Dictionary Lookup

The dictionary reversal provided us with a large source dictionary consisting mainly of uninflected English words. Our next task was to devise a lookup strategy which could find either the canonical form or an inflected form of a word. A lemmatization procedure (LEMMA), written by the late Dr. R. Ross Macdonald of Georgetown University, was adapted for use with the system.

The dictionary lookup consists of a series of steps which are performed until a match is found for the input word. First, a high-frequency table is checked. Then the full-form is looked up in the main dictionary. If the word is not found, LEMMA is called. 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 again. LEMMA makes use of morphological and spelling rules and short lists of exceptions in order to determine when to remove or add a final -e, whether 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. This routine uses the information provided by LEMMA and a table of other derivational suffixes in order to determine the possible parts of speech of a not-found word.

This lookup strategy facilitates working with random text. It also helps to keep the dictionary smaller. The dictionary coder has the option of entering a word with all its affixes or entering something less than the full form. When dealing with irregular forms and homographs, the full form must be used. For example, the dictionary must contain "meet," "met," and "meeting," but the forms "meets" and "meetings" are not required. Although the word "unwittingly" could be found as "wit," it would be difficult to generate a satisfactory Spanish translation for the adverb based on the gloss for the noun. Thus the dictionary should contain both "unwitting" and "wit" but does not need to have an entry for "unwittingly."



```
DICTIONARY PRINT RUN 07/29/7[illegible] [illegible] PAGE 41   OPTIONS= ENGLISH-SPANISH ALPHA PRINT MNEMONIC ALL
***THIS IS NOT A FULL DICTIONARY LISTING -- SEE PAGE 1
ITEM                      ENDING POS   GEN NUM INS  LEXICAL NO    ITEM                         ENDING POS   GEN NUM INS
   SOURCE=10                                                         SOURCE=10
AGRAPHIA                      N             S       019803000000  AGRAFI/A                         N         F   S
   SOURCE=10                                                         SOURCE=10
AGRAPHIC                      A                     019804000000  AGRA/FICO                        A         M   S
   SOURCE=10                                                         SOURCE=1A RCODE=4
AGRARIAN                      A                     069116000000  AGRARIO                          A         M   S
AGREE                         V            STEM     018769000000  ACORD                            V             STEM
   VERBSYN=5                                                         CLASS=AR
                                                    018769000001  ACUERD                           V             STEM
                                                                     CLASS=AR
AGREE*                        V            STEM     059591000000  CONVIN                           V             STEM
   VERBSYN=5                                                         CLASS=IR
AGREE**                       V            STEM     053050000000  ACORD                            V             STEM
   SOURCE=20 VERBSYN=5                                               CLASS=AR
                                                    053050000001  ACUERD                           V             STEM
                                                                     CLASS=AR
AGREEABLE                     A                     060172000000  GRATO                            A         M   S
   SOURCE=20                                                         SOURCE=20
AGREEABLE*                    A                     057582000000  DELEITOSO                        A         M   S
   SOURCE=20                                                         SOURCE=20
AGREEMENT                     N             S       056818000000  CONVENIO                         N         M   S
                                                                     SOURCE=20
                                                    056818000001  CONVENIO                         N         M   S
AGREEMENT*                    N             S       015024000000  ACUERDO                          N         M   S
                                                                     PR=1S
AGREEMENT**                   N             S       018297000000  CONFORMIDAD                      N         F   S
                                                    018297000001  ENCOMIO                          N         M   S
AGREEMENT***                  N             S       054419000000  AVENIMIENTO                      N         M   S
   SOURCE=20                                                         SOURCE=20
AGREEMENT****                 N             S       054417000000  AVENENCIA                        N         F   S
   SOURCE=20                                                         SOURCE=20
AGREEMENT*****                N             S       056456000000  CONCIERTO                        N         M   S
   SOURCE=20                                                         SOURCE=20
AGREMIA                       N             S       019805000000  AGREMIA                          N         F   S
   SOURCE=10                                                         SOURCE=10
AGRIA                         N             S       019806000000  AGRIA                            N         F   S
   SOURCE=10                                                         SOURCE=10
AGRICULTURAL                  A                     053282000000  AGRI/COLA                        A         MF  S
   SOURCE=20                                                         SOURCE=20
AGRICULTURE                   N             S       052793000000  AGRICULTURA                      N         F   S
AGRIMONY                      N             S       019807000000  AGRIMONIA                        N         F   S
   SOURCE=10                                                         SOURCE=10
AGRIOTHYMIA                   N             S       019808000000  AGRIOTIMIA                       N         F   S
   SOURCE=10                                                         SOURCE=10
```

Figure 1. Display of reversed dictionary entries, prior to manual coding.

Concordances by Dictionary Code

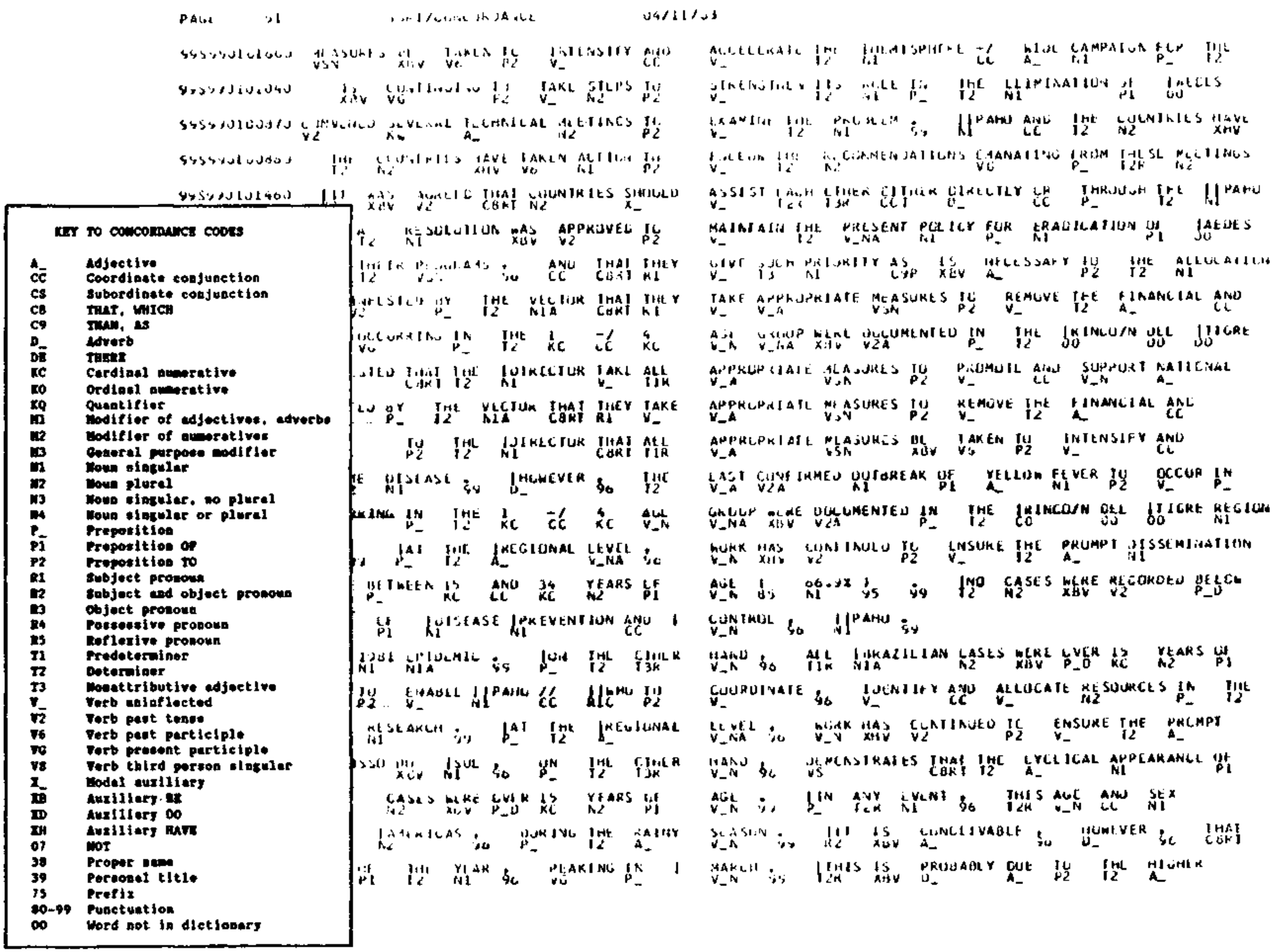
An original program (MTSCODE) was written to produce a KWIC concordance based on sequences of dictionary codes. It was devised as a tool for examining large portions of the English corpus and identifying the common syntactic patterns. Any document on the word processor can be used as the input text. The program uses the input and lookup procedures which were developed for the translation program. Therefore, it does not require full-form dictionary entries and can be run quite successfully on random text. By specifying different options at run time, the user can have the KWIC records sorted by left or right context; by dictionary codes, words, reversed words, or lexical numbers; and in alphabetical or reverse alphabetical order. Frequency counts and lists of words that are missing from the dictionary can also be obtained.

MTSCODE has proved to be a valuable tool for monitoring the part-of-speech and homograph coding in the newly reversed dictionaries. It is also helpful for studying the environments of various types of homographs. Since, the MTSCODE output is a display of the principal codes available to the analysis procedures, it is assisting us in formalizing our syntactic rules. Figure 2 is an example of one type of output produced by MTSCODE.

Expansion of the Dictionary Coding

The depth of coding inherited from the SPANAM dictionaries was not sufficient for the analysis of English. Indeed, the need for deeper coding has been one of the stumbling blocks to the further enhancement of the Spanish-English algorithm. As originally designed, the dictionary record consisted of 160 bytes, which were used to store information in character format in a total of 82 fixed fields. Many of these fields contained binary information—the presence or absence of a particular feature-—signalled by the characters "0" (zero) and "1" (one). Many of the new codes to be introduced also lent themselves to a binary treatment. Instead of increasing the size of the record to accommodate the new codes, it was decided to use the existing space more efficiently by subdividing certain bytes into bit fields. A total of 18 bytes were converted to bit fields, which yielded 144 fields for binary codes.

Some of the new bit fields are used to store information about the syntactic and semantic features of verbs, nouns, and adjectives. For example, verbs and deverbal nouns are specified as occurring with one or more of the following coda: 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. Noun features include count, bulk, concrete, human, animate, feminine, proper, collective, locative, time, body part, condition, and treatment. The need for additional noun features and the exact specifications of adjective features is being determined as work progresses on the translation algorithm. One of the references being used for the coding of English entries is Naomi Sager's description of the Linguistic String Parser (1981).



```
PAGE    51            [illegible]/CONCORDANCE           04/11/83

9999901010660  MEASURES BE   TAKEN TO   INTENSIFY AND      ACCELERATE THE  HEMISPHERE -/   WIDE CAMPAIGN FOR  THE
               V5N      XBV  V6    P2   V_        CC       V_         T2   N1         CC   A_   N1       P_   T2

9999901010040     IS   CONTINUING TO    TAKE STEPS TO      STRENGTHEN ITS  ROLE IN    THE  ELIMINATION OF   AEDES
                  XBV  VG         P2    V_   N2    P2      V_         T2   N1   P_    T2   N1          P1   00

9999901000870  CONVENED SEVERAL TECHNICAL MEETINGS TO      EXAMINE THE  PROBLEM .   PAHO AND  THE  COUNTRIES HAVE
               V2       KQ      A_        N2       P2      V_      T2   N1      99  N1   CC   T2   N2        XHV

9999901000860     THE  COUNTRIES HAVE TAKEN ACTION TO      FOLLOW THE  RECOMMENDATIONS EMANATING FROM THESE MEETINGS
                  T2   N2        XHV  V6    N1     P2      V_     T2   N2              VG        P_   T2R   N2

9999901001460  IT   WAS  AGREED THAT COUNTRIES SHOULD      ASSIST EACH OTHER EITHER DIRECTLY OR   THROUGH THE  PAHO
                    XBV  V2     CBRT N2        X_          V_     T2   T3R   CCT    D_       CC   P_      T2   N1

               A    RESOLUTION WAS  APPROVED TO            MAINTAIN THE  PRESENT POLICY FOR  ERADICATION OF   AEDES
               T2   N1         XBV  V2       P2            V_       T2   V_NA    N1     P_   N1          P1   00

               THEIR PROGRAMS ,   AND  THAT THEY           GIVE SUCH PRIORITY AS   IS   NECESSARY TO   THE  ALLOCATION
               T2    V5       96  CC   CBRT R1             V_   T3   N1       C9P  XBV  A_        P2   T2   N1

               NFESTED BY   THE  VECTOR THAT THEY          TAKE APPROPRIATE MEASURES TO   REMOVE THE  FINANCIAL AND
               V2      P_   T2   N1A    CBRT R1            V_   V_A         V5N      P2   V_     T2   A_        CC

               OCCURRING IN   THE  1   -/  4               AGE  GROUP WERE DOCUMENTED IN   THE  RINCO/N DEL  TIGRE
               VG        P_   T2   KC  CC  KC              V_N  V_NA  XBV  V2A        P_   T2   00      00   00

               STED THAT THE  DIRECTOR TAKE ALL            APPROPRIATE MEASURES TO   PROMOTE AND  SUPPORT NATIONAL
                    CBRT T2   N1       V_   T1R            V_A         V5N      P2   V_      CC   V_N     A_

               ED BY   THE  VECTOR THAT THEY TAKE          APPROPRIATE MEASURES TO   REMOVE THE  FINANCIAL AND
                  P_   T2   N1A    CBRT R1   V_            V_A         V5N      P2   V_     T2   A_        CC

                  TO   THE  DIRECTOR THAT ALL              APPROPRIATE MEASURES BE   TAKEN TO   INTENSIFY AND
                  P2   T2   N1       CBRT T1R              V_A         V5N      XBV  V6    P2   V_        CC

               E  DISEASE .   HOWEVER ,   THE              LAST CONFIRMED OUTBREAK OF   YELLOW FEVER TO   OCCUR IN
                  N1      99  D_      96  T2               V_A  V2A       N1       P1   A_     N1    P2   V_    P_

               RING IN   THE  1   -/  4   AGE              GROUP WERE DOCUMENTED IN   THE  RINCO/N DEL  TIGRE REGION
                    P_   T2   KC  CC  KC  V_N              V_NA  XBV  V2A        P_   T2   00      00   00    N1

                  AT   THE  REGIONAL LEVEL ,               WORK HAS  CONTINUED TO   ENSURE THE  PROMPT DISSEMINATION
                  P_   T2   A_       V_NA  96              V_N  XHV  V2        P2   V_     T2   A_     N1

               BETWEEN 15   AND  34   YEARS OF             AGE  (   66.9% )   .    NO   CASES WERE RECORDED BELOW
               P_      KC   CC   KC   N2    P1             V_N  85  N1    95  99   T2   N2    XBV  V2       P_D

                  OF   DISEASE PREVENTION AND  (           CONTROL ,   PAHO .
                  P1   N1      N1         CC               V_N     96  N1   99

               1981 EPIDEMIC .   ON   THE  OTHER           HAND ,   ALL  BRAZILIAN CASES WERE OVER 15   YEARS OF
               N1   N1A      99  P_   T2   T3R             V_N  96  T1R  N1A       N2    XBV  P_D  KC   N2    P1

               TO   ENABLE PAHO //  WHO  TO                COORDINATE ,   IDENTIFY AND  ALLOCATE RESOURCES IN   THE
               P2   V_     N1   CC  ALC  P2                V_         96  V_       CC   V_       N2        P_   T2

               RESEARCH .   AT   THE  REGIONAL             LEVEL ,   WORK HAS  CONTINUED TO   ENSURE THE  PROMPT
               N1       99  P_   T2   A_                   V_NA  96  V_N  XHV  V2        P2   V_     T2   A_

               SSO DO   SOL ,   ON   THE  OTHER            HAND ,   DEMONSTRATES THAT THE  CYCLICAL APPEARANCE OF
                   XCV  N1  96  P_   T2   T3R              V_N  96  V5           CBRT T2   A_       N1         P1

                  CASES WERE OVER 15   YEARS OF            AGE .   IN   ANY  EVENT ,   THIS AGE  AND  SEX
                  N2    XBV  P_D  KC   N2    P1            V_N 99  P_   T2R  N1    96  T2R  V_N  CC   N1

                  AMERICAS ,   DURING THE  RAINY           SEASON .   IT   IS   CONCEIVABLE ,   HOWEVER ,   THAT
                  N2       96  P_     T2   A_              V_N    99  R2   XBV  A_          96  D_      96  CBRT

               OF   THE  YEAR ,   PEAKING IN   (           MARCH .   THIS IS   PROBABLY DUE  TO   THE  HIGHER
               P1   T2   N1   96  VG      P_               V_N   99  T2R  XBV  D_       A_   P2   T2   A_
```

KEY TO CONCORDANCE CODES

A_	Adjective
CC	Coordinate conjunction
CS	Subordinate conjunction
C8	THAT, WHICH
C9	THAN, AS
D_	Adverb
DE	THERE
KC	Cardinal numerative
KO	Ordinal numerative
KQ	Quantifier
M1	Modifier of adjectives, adverbs
M2	Modifier of numeratives
M3	General purpose modifier
N1	Noun singular
N2	Noun plural
N3	Noun singular, no plural
N4	Noun singular or plural
P_	Preposition
P1	Preposition OF
P2	Preposition TO
R1	Subject pronoun
R2	Subject and object pronoun
R3	Object pronoun
R4	Possessive pronoun
R5	Reflexive pronoun
T1	Predeterminer
T2	Determiner
T3	Nonattributive adjective
V_	Verb uninflected
V2	Verb past tense
V6	Verb past participle
VG	Verb present participle
V8	Verb third person singular
X_	Modal auxiliary
XB	Auxiliary BE
XD	Auxiliary DO
XH	Auxiliary HAVE
07	NOT
38	Proper name
39	Personal title
75	Prefix
80-99	Punctuation
00	Word not in dictionary

Figure 2. Example of concordance by part-of-speech codes.

The conversion to the new record format was accomplished by means of a special-purpose program which rearranged the existing fields and codes. The new codes are being introduced manually. Mnemonic descriptors were added to the dictionary update and display programs so that the dictionary coders do not have to work with binary representation. The PL/1 code is also quite easy to read, since each bit is referred to by a mnemonic identifier.

Another modification of the coding system involved the part of speech codes, which were expanded to permit the subclassification of determiners, numeratives, adjectives, pronouns, modifiers, and conjunctions. The number of possible homograph types was also increased. Words are coded as homographs if they are expected to occur as more than one part of speech in the type of text for which the system is designed. Thus, while the number of homographs in the machine dictionaries is not limited to actual occurrences in the corpus, neither does it include all possible uses of every word.

An attempt is being made to find the optimum degree of specificity in coding that will produce the desired quality of output without overburdening the algorithm or the dictionary coder. New codes are being introduced gradually as they are needed in order to obtain a correct translation. Additional fields can be created or the use of existing ones changed, as necessary.

Putting ENGSPAN Together

The first version of ENGSPAN was created by combining the existing input and output modules with the new source lookup procedure. Since we have been producing some type of Spanish output from the outset, we have been constantly reminded of the requirements for target synthesis. We will not fall into the trap of spending all our time trying to analyze English and have no Spanish to show for it. We are also able to get the reactions of native Spanish speakers whenever we have output that is presentable enough to show to them.

Table 1 contains a list of the support software and other program modules originally developed for SPANAM which are also used for ENGSPAN. Table 2 lists the new program modules which were written for ENGSPAN during 1982 and 1983. Each new module has produced a noticeable improvement in the output, but many important areas remain to be addressed. We have already begun developing a general parsing algorithm and new types of dictionary entries for triggering context-sensitive glosses. Several different approaches are being considered for improving the treatment of prepositions and adjuncts. Special attention will be given to the synthesis of clitic pronouns, the use of the definite article, and the requirement for the subjunctive mood in Spanish. A long-range task is the development of knowledge structures and means of representing the semantic content of sentences and larger chunks of text.

Some of ENGSPAN's new modules are described below.

Table 1. Software common to SPANAM and ENGSPAN.

Name	Function
UPDATE	Adds, changes, and deletes dictionary entries using mnemonic descriptors
DPRINT	Prints out the dictionary entries
WANGMTS	Converts text transmitted from the word processor into the format required by the translation program
MTSINIT	Initializes the high-frequency dictionaries
MTSIOW	Reads in one sentence at a time for translation and formats the output

Table 2. Software written specifically for ENGSPAN.

Name	Function
LEMMA	Removes inflectional endings from words during the lookup procedure
LOOKUP	Looks up individual words in the English dictionary and does gap analysis for not-found words
FINDUNIT	Looks up substitution units and analysis units in the English dictionary
VERBSTRING	Analyzes simple finite and nonfinite verb strings
POSAMBIG	Resolves certain types of homographs
NOUNSTRING	Rearranges noun phrases and determines the need for gender and number agreement for certain sequences of modifiers
TLOOKUP	Looks up glosses in the Spanish dictionary
NOUNSYN	Synthesizes Spanish determiners, numeratives, adjectives, and nouns
VERBSYN	Synthesizes Spanish verbs

VERBSTRING

This module is a combined analysis and transfer routine which was written as a temporary procedure for handling the most frequent types of verb strings until a more general parsing algorithm could be developed. It identifies verb phrases in the source text, resolves homographs involving auxiliaries and main verbs, attempts to determine the subject of each finite verb, and introduces codes that will eventually trigger the synthesis of the proper Spanish inflections. It rearranges auxiliaries, adverbs, and "not"; deletes the pronoun "it" when it occurs as the subject; and deletes the auxiliary "do" when it occurs in questions. It triggers constructions using "haber" when the verb phrase is preceded by "there." English passives are rendered using "se" and the finite form of the verb unless the agent is expressed. The subjunctive mood and the imperfect tense are specified in certain contexts. There are several rules which select between "ser" and "estar."

POSAMBIG

This module attempts to determine the part of speech of words that are coded as homographs and have not already been resolved as verbs. It does so by examining the left and right context of each word. For each homograph type there is a default decision which is used when the context does not meet any of the criteria specified in the algorithm. Additional homograph types need to be added to this module, and some of the existing criteria need to be improved. The function of this module will eventually be performed by the parsing algorithm.

NOUNSTRING

A pattern matching procedure is used for the recognition of noun phrases. The parts of speech of the words are matched with a set of patterns which may begin with an adjective, adverbial modifier, or noun. The routine triggers the agreement of adjectives, determiners, and numeratives in premodifying position and the agreement of past participles in postmodifying position. It also specifies the word order within the target phrase. If a noun premodifier is moved to the right of the head noun, the preposition "de" is inserted. The definite article is inserted before some types of noun phrases if there is no determiner or numerative. A total of 19 noun phrase patterns are currently being tested. The results are being compared with the desired translation of the noun phrases found in the corpus in order to determine the additional types of coding and analysis which are needed.

VERBSYN

The procedure for the synthesis of Spanish verb forms is based on principles of generative morphology and phonology. The program synthesizes regular and 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 11 exception features which govern the synthesis of irregular forms. Only one dictionary entry is needed for each verb. A small number of highly irregular stems and endings are listed in the program itself. The majority of verbs require no synthesis coding except for the conjugation class. 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.

NOUNSYN

This procedure performs the synthesis of feminine and plural endings for determiners, numeratives, adjectives, and nouns. The algorithm contains rules for forming all regular plurals and handling many irregular forms. The majority of Spanish nouns and adjectives require no special synthesis coding in the dictionary entry. If the gloss consists of more than one word, synthesis will be performed on the first word in the default situation. The item may be coded for synthesis of every word or only specific words.

Development of the ATN Parser

The analysis procedures described above are based entirely on the recognition of local syntactic patterns. They break down whenever long distance relationships are involved. From the beginning of the project we knew that we would have to expand the horizons of our analysis routines. The main thrust of our current work is the development of an augmented transition network (ATN) parser, similar to the one described by Winograd (1983). The ATN was selected because it is compatible with our existing architecture, which has a strong syntactic orientation. It provides an effective means of dealing with homographs and allows for the selective use of semantic coding. ATN parser is being designed to provide us with the information we need for Spanish synthesis. At present, we are working only at the sentence level. Eventually, we plan to save certain types of information about previous sentences.

The current version of the parser has four networks: sentence, noun phrase, verb phrase, and prepositional phrase. It also has a special procedure for handling conjoining within the 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 indicate recursive calls to a network; and send arcs, which indicate successful completion of processing in a network.

An augmented transition network allows conditions and actions to be associated with the arcs. If there is a condition on an arc, it must be satisfied before the arc can be taken. If an action is specified, it is performed whenever the arc is taken. The use of conditions provides a mechanism for introducing into the grammar a degree of sensitivity to the left context and to semantic criteria. The actions are used to store the intermediate and final results of the analysis in registers which are available both to the parser and to the synthesis routines.

The algorithm performs a sequential parse with chronological backtracking. The order in which the arcs are tested is specified by the linguist, and the parser stops after completing the first successful parse. The algorithm processes the words of the input string one at a time, moving from left to right. All possible arcs that may be taken for a word at the current state are placed on a pushdown stack. The parser tests each arc on the stack until it finds one that matches the current word. It 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 and tests the alternative arcs which were saved on the stack. 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.

In the event of an unsuccessful parse, ENGSPAN is still expected to produce some kind of a translation. We are experimenting with several strategies for recovering information from a failed parse. For example, whenever backtracking takes place, information regarding the longest successful path is saved. It may be possible to resume the parse at another point in the input string. We are also investigating ways of making the parser more efficient, such as saving well-formed substrings and doing explicit rather than chronological backtracking.

The ATN parsing algorithm is being developed in an independent PL/1 program, using the ENGSPAN input and dictionary lookup modules. The network is read in at runtime, making it possible to experiment with different network configurations without recompiling the program. The next step will be to link the two programs so that ENGSPAN's synthesis modules can access the sentence, clause, and phrase registers created by the parser. If the parse is not successful, ENGSPAN's local disambiguation and analysis routines will be used to fill in as much missing information as possible in order to obtain a default translation. The diagram in Figure 3 shows how the ENGSPAN model will look when the parser has been incorporated.

Multi-Word Dictionary Entries

The strategy regarding the use of multi-word dictionary entries is under review in light of the requirements of the ATN parser and the analysis of conjoined phrases. There is a need to change the way the substitution unit is used and to design several new types of dictionary entries.

The substitution unit should not be used if the parser needs to access the syntactic and semantic codes for each word. This is the case whenever there is a relatively high probability that the phrase may be part of a conjoined structure. For example, the phrase "tertiary care" can be expected

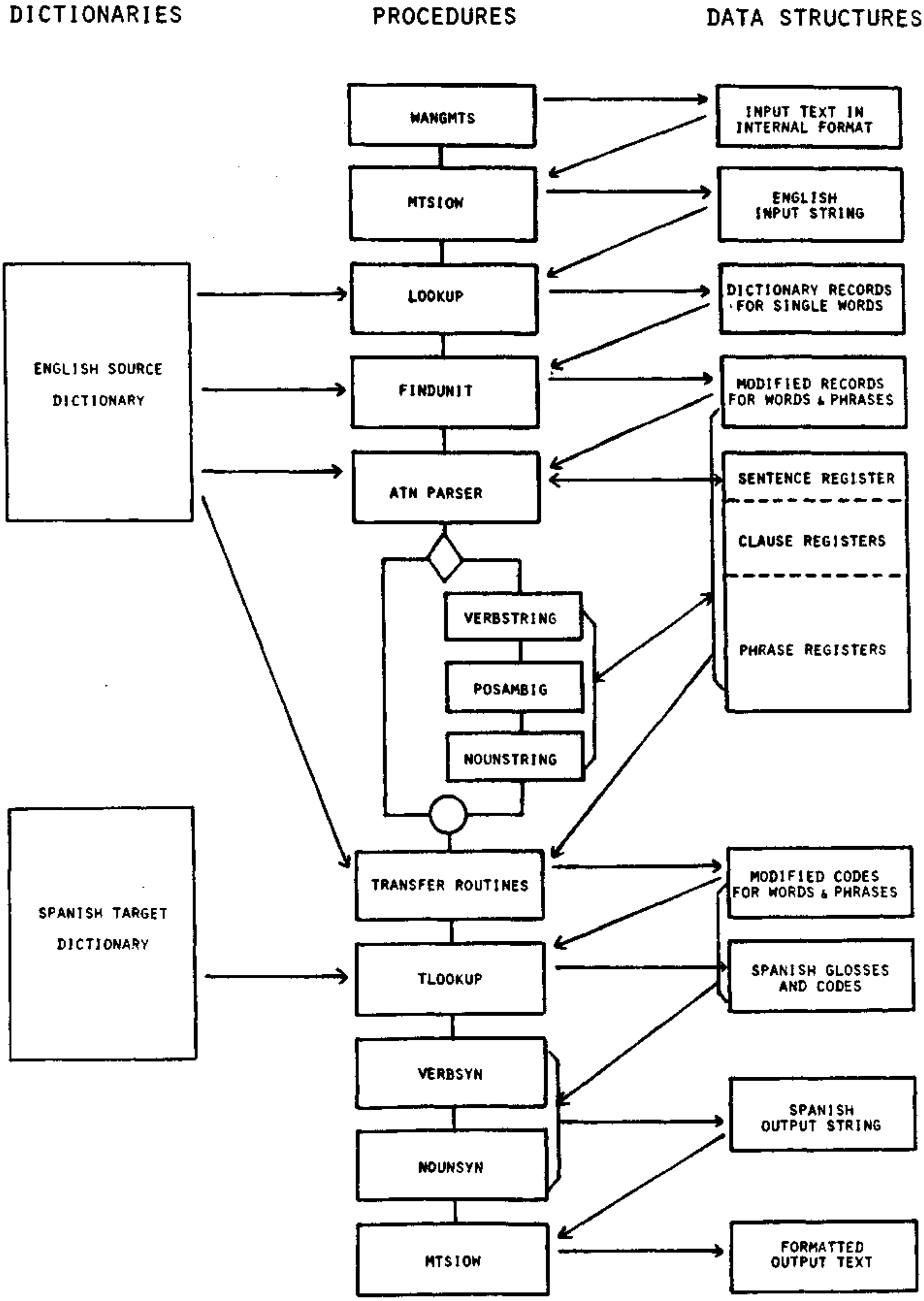


Figure 3. The ENGSPAN model.

to occur as "primary, secondary, and tertiary care." It is also necessary when the same sequence of lexical items can occur with different functions, such as "drug control" and "the use of this drug controls the symptoms." If the parser is to do its job, the number of phrases which can be handled as SUs turns out to be relatively small. These include phrasal prepositions such as "in lieu of," expressions such as "by leaps and bounds," the names of organizations, meetings, and documents, and the names of chemical substances. Many sequences which were formally entered as SUs can be better handled as analysis units.

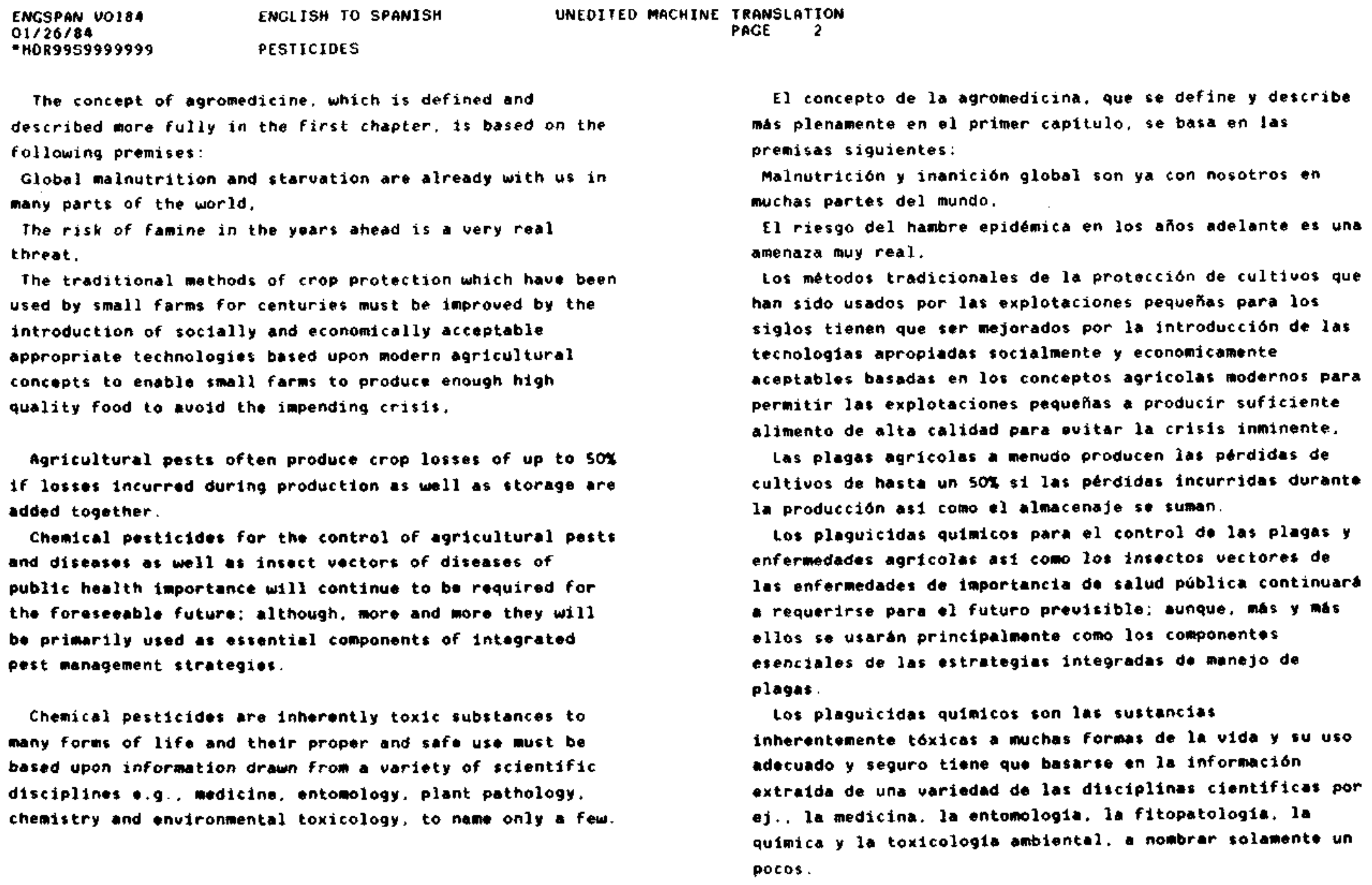
With the reduced use of the SU, the nesting of SUs in order to handle sequences of more than 5 words is no longer feasible, and a new method of handling long units is needed. It is planned to use a variable-length record in the same dictionary. Procedures must be developed to make it as easy as possible for the dictionary coder to add, change, and delete the new type of entry. The implementation of this change will require modifications in the ENGSPAN, DPRINT, and UPDATE programs.

Another type of dictionary entry is being developed to handle lexical items such as phrasal verbs which are likely to occur as noncontiguous words in the input. This type of entry will be used when it may be necessary to replace the individual source dictionary records with another record containing the syntactic and semantic features of the multi-word lexical item. The entry will be retrieved from the dictionary during the parsing of the sentence; the parser will determine whether or not the individual records should be replaced by the multi-word entry.

Still another type of dictionary entry is being developed to specify an alternate translation of a word which depends on the occurrence of a specific word or set of features in one of its arguments. This entry will be used by a transfer procedure which is called after the parse has been completed. The procedure will access the structural information produced by the parser in order to locate the argument in question. If the argument meets the conditions specified in the transfer entry, the alternate translation will be selected.

Sample Output and Dictionary Entries

Figure 4 contains a page of unedited English-Spanish machine translation produced by ENGSPAN in January 1984. The output is in word-processing format. This sample is provided for the purpose of demonstrating that ENGSPAN is working, but that there is still a lot more work to do. Figure 5 shows the dictionary entries for some of the words in the sample text. We have also included, as Figure 6, the raw output which was obtained for the same page of text before any dictionary updating had been done. It is presented in the working format produced on the computer printer. It provides an indication of the results that could be expected for random input text at this time.



ENGSPAN V0184 ENGLISH TO SPANISH UNEDITED MACHINE TRANSLATION
01/26/84 PAGE 2
*HDR9959999999 PESTICIDES

The concept of agromedicine, which is defined and described more fully in the first chapter, is based on the following premises:

Global malnutrition and starvation are already with us in many parts of the world,

The risk of famine in the years ahead is a very real threat,

The traditional methods of crop protection which have been used by small farms for centuries must be improved by the introduction of socially and economically acceptable appropriate technologies based upon modern agricultural concepts to enable small farms to produce enough high quality food to avoid the impending crisis,

Agricultural pests often produce crop losses of up to 50% if losses incurred during production as well as storage are added together.

Chemical pesticides for the control of agricultural pests and diseases as well as insect vectors of diseases of public health importance will continue to be required for the foreseeable future; although, more and more they will be primarily used as essential components of integrated pest management strategies.

Chemical pesticides are inherently toxic substances to many forms of life and their proper and safe use must be based upon information drawn from a variety of scientific disciplines e.g., medicine, entomology, plant pathology, chemistry and environmental toxicology, to name only a few.

El concepto de la agromedicina, que se define y describe más plenamente en el primer capítulo, se basa en las premisas siguientes:

Malnutrición y inanición global son ya con nosotros en muchas partes del mundo,

El riesgo del hambre epidémica en los años adelante es una amenaza muy real,

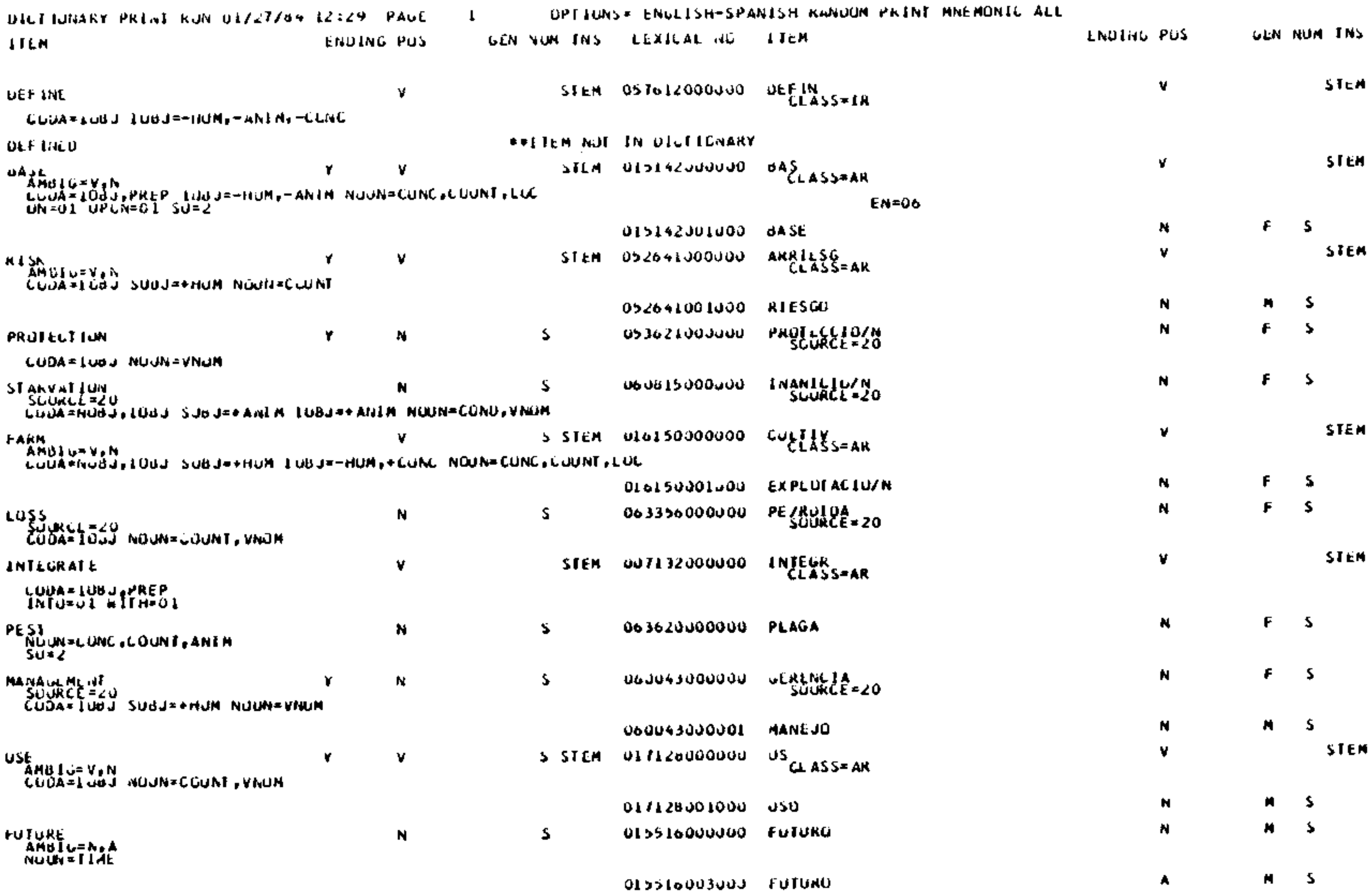
Los métodos tradicionales de la protección de cultivos que han sido usados por las explotaciones pequeñas para los siglos tienen que ser mejorados por la introducción de las tecnologías apropiadas socialmente y economicamente aceptables basadas en los conceptos agrícolas modernos para permitir las explotaciones pequeñas a producir suficiente alimento de alta calidad para evitar la crisis inminente,

Las plagas agrícolas a menudo producen las pérdidas de cultivos de hasta un 50% si las pérdidas incurridas durante la producción así como el almacenaje se suman.

Los plaguicidas químicos para el control de las plagas y enfermedades agrícolas así como los insectos vectores de las enfermedades de importancia de salud pública continuará a requerirse para el futuro previsible; aunque, más y más ellos se usarán principalmente como los componentes esenciales de las estrategias integradas de manejo de plagas.

Los plaguicidas químicos son las sustancias inherentemente tóxicas a muchas formas de la vida y su uso adecuado y seguro tiene que basarse en la información extraída de una variedad de las disciplinas científicas por ej., la medicina, la entomología, la fitopatología, la química y la toxicología ambiental, a nombrar solamente un pocos.

Figure 4. Unedited English-Spanish machine translation produced by ENGSPAN in January 1984.



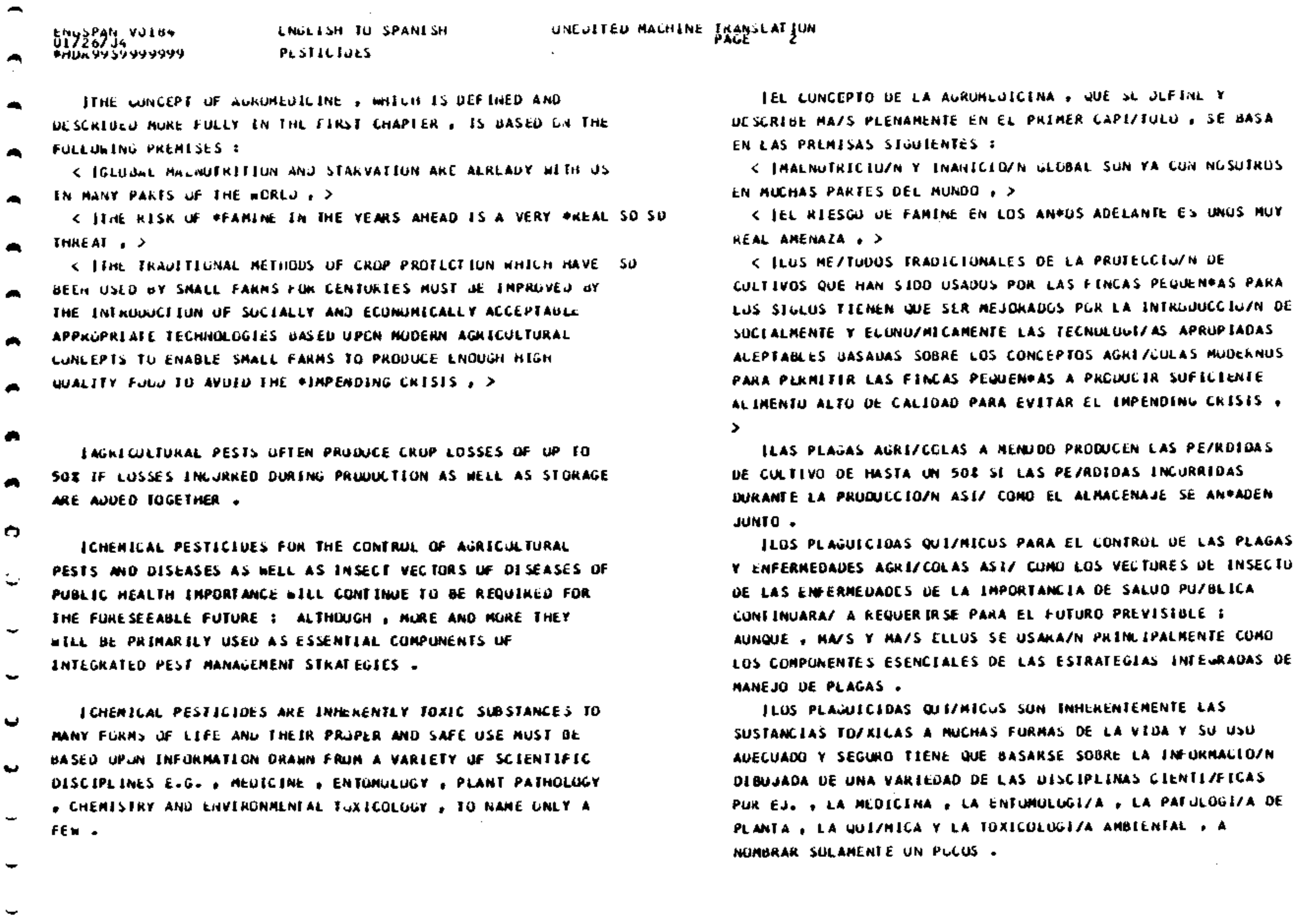
```
DICTIONARY PRINT RUN 01/27/84 12:29  PAGE    1       OPTIONS= ENGLISH-SPANISH RANDOM PRINT MNEMONIC ALL
ITEM                           ENDING POS      GEN NUM INS   LEXICAL NO   ITEM                           ENDING POS      GEN NUM INS

DEFINE                                V               STEM  057612000000  DEFIN                                 V               STEM
                                                                            CLASS=IR
  CODA=IOBJ IOBJ=-HUM,-ANIM,-CONC
DEFINED                                          **ITEM NOT IN DICTIONARY
BASE                           Y      V               STEM  015142000000  BAS                                   V               STEM
  AMBIG=V,N                                                                 CLASS=AR
  CODA=IOBJ,PREP IOBJ=-HUM,-ANIM NOUN=CONC,COUNT,LOC
  ON=01 UPON=01 SU=2                                                                EN=06

                                                            015142001000  BASE                                  N         F   S
RISK                           Y      V               STEM  052641000000  ARRIESG                               V               STEM
  AMBIG=V,N                                                                 CLASS=AR
  CODA=IOBJ SUBJ=+HUM NOUN=COUNT

                                                            052641001000  RIESGO                                N         M   S
PROTECTION                     Y      N             S       053621000000  PROTECCIO/N                           N         F   S
                                                                            SOURCE=20
  CODA=IOBJ NOUN=VNOM
STARVATION                            N             S       060815000000  INANICIO/N                            N         F   S
  SOURCE=20                                                                 SOURCE=20
  CODA=NOBJ,IOBJ SUBJ=+ANIM IOBJ=+ANIM NOUN=COND,VNOM
FARM                                  V             S STEM  016150000000  CULTIV                                V               STEM
  AMBIG=V,N                                                                 CLASS=AR
  CODA=NOBJ,IOBJ SUBJ=+HUM IOBJ=-HUM,+CONC NOUN=CONC,COUNT,LOC
                                                            016150001000  EXPLOTACIO/N                          N         F   S
LOSS                                  N             S       063356000000  PE/RDIDA                              N         F   S
  SOURCE=20                                                                 SOURCE=20
  CODA=IOBJ NOUN=COUNT,VNOM
INTEGRATE                             V               STEM  007132000000  INTEGR                                V               STEM
                                                                            CLASS=AR
  CODA=IOBJ,PREP
  INTO=01 WITH=01
PEST                                  N             S       063620000000  PLAGA                                 N         F   S
  NOUN=CONC,COUNT,ANIM
  SU=2
MANAGEMENT                     Y      N             S       060043000000  GERENCIA                              N         F   S
  SOURCE=20                                                                 SOURCE=20
  CODA=IOBJ SUBJ=+HUM NOUN=VNOM
                                                            060043000001  MANEJO                                N         M   S
USE                            Y      V             S STEM  017128000000  US                                    V               STEM
  AMBIG=V,N                                                                 CLASS=AR
  CODA=IOBJ NOUN=COUNT,VNOM
                                                            017128001000  USO                                   N         M   S
FUTURE                                N             S       015516000000  FUTURO                                N         M   S
  AMBIG=N,A
  NOUN=TIME
                                                            015516003000  FUTURO                                A         M   S
```

Figure 5. Mnemonic display of sample dictionary entries.



```
ENGSPAN V0184          ENGLISH TO SPANISH          UNEDITED MACHINE TRANSLATION
01/26/84                                                            PAGE    2
*HDR9999999999         PESTICIDES

   |THE CONCEPT OF AGROMEDICINE , WHICH IS DEFINED AND
DESCRIBED MORE FULLY IN THE FIRST CHAPTER , IS BASED ON THE
FOLLOWING PREMISES :
  < |GLOBAL MALNUTRITION AND STARVATION ARE ALREADY WITH US
IN MANY PARTS OF THE WORLD , >
  < |THE RISK OF *FAMINE IN THE YEARS AHEAD IS A VERY *REAL SO SD
THREAT , >
  < |THE TRADITIONAL METHODS OF CROP PROTECTION WHICH HAVE  SD
BEEN USED BY SMALL FARMS FOR CENTURIES MUST BE IMPROVED BY
THE INTRODUCTION OF SOCIALLY AND ECONOMICALLY ACCEPTABLE
APPROPRIATE TECHNOLOGIES BASED UPON MODERN AGRICULTURAL
CONCEPTS TO ENABLE SMALL FARMS TO PRODUCE ENOUGH HIGH
QUALITY FOOD TO AVOID THE *IMPENDING CRISIS , >


   |AGRICULTURAL PESTS OFTEN PRODUCE CROP LOSSES OF UP TO
50% IF LOSSES INCURRED DURING PRODUCTION AS WELL AS STORAGE
ARE ADDED TOGETHER .

   |CHEMICAL PESTICIDES FOR THE CONTROL OF AGRICULTURAL
PESTS AND DISEASES AS WELL AS INSECT VECTORS OF DISEASES OF
PUBLIC HEALTH IMPORTANCE WILL CONTINUE TO BE REQUIRED FOR
THE FORESEEABLE FUTURE :  ALTHOUGH , MORE AND MORE THEY
WILL BE PRIMARILY USED AS ESSENTIAL COMPONENTS OF
INTEGRATED PEST MANAGEMENT STRATEGIES .

   |CHEMICAL PESTICIDES ARE INHERENTLY TOXIC SUBSTANCES TO
MANY FORMS OF LIFE AND THEIR PROPER AND SAFE USE MUST BE
BASED UPON INFORMATION DRAWN FROM A VARIETY OF SCIENTIFIC
DISCIPLINES E.G. , MEDICINE , ENTOMOLOGY , PLANT PATHOLOGY
, CHEMISTRY AND ENVIRONMENTAL TOXICOLOGY , TO NAME ONLY A
FEW .

   |EL CONCEPTO DE LA AGROMEDICINA , QUE SE DEFINE Y
DESCRIBE MA/S PLENAMENTE EN EL PRIMER CAPI/TULO , SE BASA
EN LAS PREMISAS SIGUIENTES :
  < |MALNUTRICIO/N Y INANICIO/N GLOBAL SON YA CON NOSOTROS
EN MUCHAS PARTES DEL MUNDO , >
  < |EL RIESGO DE FAMINE EN LOS AN*OS ADELANTE ES UNOS MUY
REAL AMENAZA , >
  < |LOS ME/TODOS TRADICIONALES DE LA PROTECCIO/N DE
CULTIVOS QUE HAN SIDO USADOS POR LAS FINCAS PEQUEN*AS PARA
LOS SIGLOS TIENEN QUE SER MEJORADOS POR LA INTRODUCCIO/N DE
SOCIALMENTE Y ECONO/MICAMENTE LAS TECNOLOGI/AS APROPIADAS
ACEPTABLES BASADAS SOBRE LOS CONCEPTOS AGRI/COLAS MODERNOS
PARA PERMITIR LAS FINCAS PEQUEN*AS A PRODUCIR SUFICIENTE
ALIMENTO ALTO DE CALIDAD PARA EVITAR EL IMPENDING CRISIS ,
>
   |LAS PLAGAS AGRI/COLAS A MENUDO PRODUCEN LAS PE/RDIDAS
DE CULTIVO DE HASTA UN 50% SI LAS PE/RDIDAS INCURRIDAS
DURANTE LA PRODUCCIO/N ASI/ COMO EL ALMACENAJE SE AN*ADEN
JUNTO .
   |LOS PLAGUICIDAS QUI/MICOS PARA EL CONTROL DE LAS PLAGAS
Y ENFERMEDADES AGRI/COLAS ASI/ COMO LOS VECTORES DE INSECTO
DE LAS ENFERMEDADES DE LA IMPORTANCIA DE SALUD PU/BLICA
CONTINUARA/ A REQUERIRSE PARA EL FUTURO PREVISIBLE :
AUNQUE , MA/S Y MA/S ELLOS SE USARA/N PRINCIPALMENTE COMO
LOS COMPONENTES ESENCIALES DE LAS ESTRATEGIAS INTEGRADAS DE
MANEJO DE PLAGAS .
   |LOS PLAGUICIDAS QUI/MICOS SON INHERENTEMENTE LAS
SUSTANCIAS TO/XICAS A MUCHAS FORMAS DE LA VIDA Y SU USO
ADECUADO Y SEGURO TIENE QUE BASARSE SOBRE LA INFORMACIO/N
DIBUJADA DE UNA VARIEDAD DE LAS DISCIPLINAS CIENTI/FICAS
POR EJ. , LA MEDICINA , LA ENTOMOLOGI/A , LA PATOLOGI/A DE
PLANTA , LA QUI/MICA Y LA TOXICOLOGI/A AMBIENTAL , A
NOMBRAR SOLAMENTE UN POCOS .
```

Figure 6. A sample of raw output before dictionary updating.

Putting ENGSPAN to Work

We plan to have the new version of ENGSPAN ready for pilot production by the end of 1984. The output will probably require a substantial amount of postediting, but we expect to be able to show a cost advantage over manual translation.

Acknowledgments

I would like to express my appreciation to Dr. Douglas Clarke and the Cranfield Institute of Technology for making it possible for me to participate in this Conference.

Many individuals have contributed to the design and implementation of ENGSPAN. These include the head of the MT project, Muriel Vasconcellos; my fellow computational linguist, Lee Schwartz; and our consultants R. Ross Macdonald, Michael Zarechnak, and Leonard Shaefer. I would also like to acknowledge Luis Larrea Alba, Jr., Chief of General Services, PAHO, whose support helped make ENGSPAN a reality.

REFERENCES

Sager, Naomi. Natural language information processing: A computer grammar of *English and its applications*. Reading, Massachusetts: Addison-Wesley, 1981.

Tucker, Allen B., Muriel Vasconcellos, and Marjorie León. *PAHO machine translation system: Introduction and users' manual*. Washington, D.C.: Pan American Health Organization, July 1980.

Vasconcellos, Muriel. "Management of the machine translation environment: Interaction of functions at the Pan American Health Organization." To be published in Translating and the Computer 5: Tools for the Trade, edited by Veronica Lawson. Proceeding of the Aslib Conference (London, 10-11 November 1983). In press.

Vasconcellos, Muriel. "Machine translation at the Pan American Health Organization: A review of highlights and insights." To be published by The British Computer Society, Natural Language Translations Specialist Group, *Newsletter.* In press.

Winograd, Terry. *Language as a cognitive process. Volume 1: Syntax*. Reading, Massachusetts: Addison-Wesley, 1983.