

TERMINOLOGY WORK AND AUTOMATIC TRANSLATION SYSTEMS:
A CASE STUDY AT THE PAN AMERICAN HEALTH ORGANIZATION

by Marjorie León, Susana Santangelo, and Muriel Vasconcellos*

Introduction

The ENGSPANTM machine translation system, developed in-house by the Pan American Health Organization (PAHO) with partial assistance from the U.S. Agency for International Development,¹ uses an IBM mainframe to produce batch translations from English into Spanish (Vasconcellos and León 1985). The system has been in full-scale operation since 1985 and has been used in the translation of more than a million words of text. An older system, SPANAM, translates from Spanish into English.

While ENGSPAN's basic design permits it to handle any technical text written in normal English syntax, its typical fare consists of documents in fields related to public health. The system dictionaries as of June 1987 contained 49,931 entries in the English source with 52,412 corresponding translations and alternate translations in Spanish. About half this total corresponded to terminology in biomedicine and public health.

PAHO also has texts in the field of agriculture, given its interest in nutrition and the food chain, and ENGSPAN is currently being used in a pilot outplacement at two agricultural research centers, the International Center for Tropical Agriculture (CIAT) in Cali, Colombia, and the International Rice Research Institute in the Philippines. To equip the system to handle texts in agriculture, it was decided to incorporate into the ENGSPAN dictionaries the descriptors from the AGROVOC Thesaurus.

AGROVOC is a multilingual thesaurus which was prepared under an agreement between the Food and Agriculture Organization of the United Nations (FAO) and the Commission of the European Communities (CEC). It has been implemented in two information systems coordinated by FAO: the Current Agricultural Research Information System (CARIS) and the International Information System for the Agricultural Sciences and Technology (AGRIS). In both systems, document references may be indexed or retrieved in any of three languages: English, French, or Spanish.

Proposed new descriptors are screened by the AGRIS Coordinating Center and reviewed twice a year by a working group of experts in the different languages. In July 1987, AGROVOC contained some 9,700 descriptors and about 7,000 non-descriptors in each language.

*Respectively, senior computational linguist, Spanish linguist, and chief, Terminology and Machine Translation Program, Pan American Health Organization, 525 Twenty-third Street, N.W., Washington, D.C. 20037.

¹Grant DPE-5542-G-SS-3048-00, awarded to the Pan American Health Organization under letter dated 3 August 1983.

Objective

The aim of the exercise was to incorporate the 8,697 AGROVOC descriptors in the ENGSPAN system dictionaries using existing software tools to automate the process wherever possible. The terms were to be fully coded for all their syntactic and semantic requirements, and conflicts with existing ENGSPAN terminology would be rationalized.

It was expected that a sizable proportion of the terms would already be present in ENGSPAN's dictionaries. However, given the difference in purpose between a thesaurus of indexing terms and a dictionary for the translation of running text, it was also recognized that some of the AGROVOC Spanish equivalents might not be appropriate as translational glosses.

It was hoped that the experience would yield a methodology that could be applied to the incorporation of other specialized lexicons in the future. It was also hoped that it would be possible to develop a procedure that would permit porting of the fully coded AGROVOC records to SPANAM.

The ENGSPAN Dictionaries

The ENGSPAN dictionaries comprise two indexed files--the English source dictionary, which is arranged alphabetically, and the Spanish target, whose records are tied to the source through a unique identification number. They are organized into: a high-frequency dictionary of basic function words, the main dictionary of general and public-health terminology, and specialized microglossaries that can override the main dictionary (up to 99 possible microglossaries).

New source-target pairs are added to the main dictionary; the microglossaries are used only in the event of a conflict. The terms in the microglossaries can be either subject-specific or user-specific; always, their purpose is to provide a translation different from the one that the main dictionary would normally give. There are 11 microglossaries currently in place, representing such fields as research medicine, finance, atomic energy, and agriculture. One or more microglossaries may be requested at run-time. They are invoked in the order in which they are specified.

In addition to a term being handled as a microglossary entry, there are two other ways in which a dictionary record may be associated with a given subject area. The SOURCE field is used to indicate the terminological source from which the entry was obtained, and the VOC field is used to indicate that an entry is part of the preferred vocabulary of a specific user. An entry may be coded for only one source, but it may be coded for up to eight user vocabularies. These two fields are used to retrieve lists of terms from the dictionary, while a microglossary is a means of obtaining an alternate gloss during the translation of a text.

Methodology

A tape of 8,697 AGROVOC descriptors and their Spanish equivalents was kindly supplied to PAHO by the AGRIS Processing Unit in Vienna in January

1986.² It soon became apparent, however, that it would not be possible to perform an automatic merge of the tape file and the ENGSPAN dictionaries. The discerning eye of the linguist was needed to resolve many different types of problems. The following set of consecutive AGROVOC entries illustrates some of the considerations that had to be taken into account in order to accomplish the task at hand:

FLAVOUR	AROMA
FLAVOUR ENHANCERS	REFORZADORES DE AROMA
FLAVOURED MILKS	LECHE AROMATIZADA
FLAVOURING	AROMATIZACION
FLAVOURING CROPS	PLANTAS DE CONDIMENTO
FLAVOURINGS	AROMATIZANTES

- AGROVOC use British spelling (FLAVOUR); ENGSPAN uses American spelling (flavor).
- Many AGROVOC descriptors appear in the plural; ENGSPAN entries are usually in the singular.
- The AGROVOC descriptors do not distinguish between upper and lower case; this distinction needs to be made in ENGSPAN for acronyms, proper names, and taxonomic names.
- The Spanish equivalents in AGROVOC do not include the acute accent, diaeresis, or the tilde; these diacritics are required in ENGSPAN.
- Some terms which are countable nouns in AGROVOC (MILKS and FLAVOURINGS) would usually be treated as an uncountable noun (MILK) or a verb nominalization (FLAVOURING) in ENGSPAN.
- Terms used only as nouns (FLAVOUR and MILK) in AGROVOC must be coded for other parts of speech (verb and noun) in ENGSPAN.
- Many AGROVOC terms contain inflected forms of the verb (FLAVOURED and FLAVOURING) which may or may not be needed in the ENGSPAN dictionary.
- Some AGROVOC Spanish equivalents (AROMA) conflict with existing ENGSPAN glosses (SABOR).
- Within AGROVOC, the same word requires different glosses in different contexts (FLAVOURING: AROMATIZACION, DE CONDIMENTO, AROMATIZANTE and CROPS: PLANTAS, CULTIVOS).

After an initial examination of the thesaurus, the descriptors were transferred from the tape to Wang word-processing documents containing about 1,000 terms per document. The documents were submitted for machine translation. In this way, each lexical item was processed by ENGSPAN's lookup procedure, which contains logic to identify British spelling, plural nouns, and inflected verbs, and those lexical items that were components of multiple-word descriptors were translated in context. The linguist then compared the ENGSPAN output with the Spanish equivalents in AGROVOC.

²Thanks are extended to Dr. Helga Schmid and Mr. Robert Portegies-Swart for their collaboration in this undertaking.

On the basis of this review, the terms were separated into the following groups:

- A Single words that were found in the ENGSPAN dictionary and whose gloss matched the AGROVOC equivalent. The source and target entries for these words needed only to be updated to introduce the code VOC=AGRIS.
- B Single words that were found in the ENGSPAN dictionary with the correct part of speech, but the gloss did not match the AGROVOC equivalent. This conflict was resolved by adding an alternate gloss to the agricultural microglossary or by selecting one of the Spanish terms over the other based on terminological or institutional criteria.
- C Single words that were found in the ENGSPAN dictionary but the coding of the source entry needed to be modified to include a new part of speech or new syntactic or semantic features. The source entry was modified and new target entries were added as required.
- D Single words that needed to be added to the ENGSPAN dictionary. The new English word was fully coded for its syntactic and semantic characteristics and added to the source dictionary. The necessary glosses were added to the target dictionary.
- E Multiple-word descriptors with matching translations. No action was taken for these descriptors.
- F Multiple-word descriptors with non-matching translations. These terms were set aside until all work was completed on the single words.

Once the terms had been grouped as described above, procedures were developed for handling the words in group D. For the geographical names, the genus names, and a number of other term-types, the Spanish linguist wrote a text-recall macro (Wang "glossary") to supply all the needed codes. The purpose of the macros was twofold: to avoid having to enter the same codes repeatedly, and to be sure that a full and consistent set of codes was entered in each case. When a term-type was encountered that had a macro, the complete coding for the source and the target entries was entered with two keystrokes. The macros covered such types as:

genus	Acacia - Acacia
geographical location	Abruzzi - Abruzos
material	bagasse - bagazo
device	baler - prensa enfardadora
countable animal	billygoat - macho cabrío
countable/uncountable animal	bluefish - anchoa de banco
uncountable concrete noun	bacon - tocino

After all the single words had been added to the dictionaries, the new entries were verified by retranslating the lists of descriptors. At the same time, the translations of the multiple-word descriptors were rechecked. When discrepancies were found, the linguist had to decide which of ENGSPAN's three types of collocations--the substitution unit (SU), the analysis unit (AU), or the transfer unit (TU)--was best suited to the circumstance. The solutions were designed to cover the most possible cases.

The criteria for selecting among the three types of collocations are:

- SU There is a high degree of certainty that the lexical items will always represent the collocation; it is not likely that other lexical items will interrupt the phrase or be conjoined with one of its elements.
- AU The lexical items are likely to occur contiguously but may belong to a syntactic construction other than that of the collocation; there may be external modification relationships.
- TU The lexical items may not be contiguous; the input string must be analyzed syntactically before the existence of the collocation can be assumed; a rule can be formulated that will apply to a class of lexical items.

Results

Of the 8,697 descriptors, 5,140 were single words and 3,557 were multiple-word expressions. Of the 5,140 single words, 2,601, most of them scientific names, had identical glosses in AGROVOC. Of the 3,557 multiple-word terms, 1,004 had identical equivalents. These terms were almost all strings of genus-plus-species.

Even though a total of 3,605 descriptors (41.5%) required no translation, some of these terms needed to be added to the ENGSPAN dictionary in order to provide the translation program with information about their syntactic and semantic characteristics. Genus names were added and coded so that the species names can be identified without separate dictionary entries for each one. The names of families, orders, classes, and phyla were also added.

In order to provide full syntactic and semantic coding for unfamiliar terms, it was necessary in some cases to consult the hierarchy of the thesaurus itself, Webster's Third New International Dictionary, the Encyclopedia Britannica, and other reference sources. Some terms could not be identified and were set aside until further research could be done.

A total of 1,794 single source words and 2,280 target glosses have been added to the ENGSPAN dictionaries based on the AGROVOC thesaurus. Microglossary entries have been used to resolve conflicts for 138 glosses. Other target glosses correspond to alternate parts of speech or glosses triggered by analysis or transfer units. About 200 conflicts pertaining to single words remain to be resolved and about 700 words are yet to be coded. The multiple-word terms are being dealt with gradually as time permits.

Discussion

Despite a considerable overlap in some portions of agricultural and medical terminology, the incorporation of the AGROVOC descriptors into PAHO's machine translation dictionaries was a complex task. Its complexity was largely due to the desire to ensure that the new dictionary entries would function optimally during the translation process.

Each time a discrepancy between AGROVOC and ENGSPAN was found, the

linguist had to decide among several alternative ways of resolving it. In some instances, the PAHO term prevailed (ANTHRAX: CARBUNCO, not CARBUNCLO). In other cases, the existing gloss was replaced by the AGROVOC term (BIRD: AVE, not PAJARO). In a few cases, the AGROVOC term was disregarded because it was not appropriate for translation (ELEMENT: ELEMENTO, not ELEMENTO QUIMICO). For the most part, however, both alternatives are needed in different situations:

BROADCASTING	RADIODIFUSIO/N and APLICACIO/N A VOLEO
GAME	JUEGO and ANIMALES DE CAZA
NECK	CUELLO and PESCUETO
RUST	O/XIDO and ROYA
DRINKER	BEBEDOR and BEBEDERO
HILL	COLINA (noun) and APORCAR (verb)

The examples given above are candidates for the agricultural micro-glossary. A microglossary entry in the source dictionary can influence the analysis of the input text, while a microglossary entry in the target dictionary is brought into play only during the selection of the target gloss. For example, the distinction between DRINKER (Human) and DRINKER (Device) would be a source entry, whereas the distinction between CUELLO (Human body part) and PESCUETO (Animal body part) would be a target entry. In the case of HILL, which is only a noun in the main dictionary, a source entry would also be used to indicate that it can function as a verb in an agricultural text.

The microglossary strategy, however, is not sufficiently sensitive to provide the desired target gloss in all situations. When multiple equivalents are needed in the same subject area (e.g. SCALE: ESCALA, ESCAMA, BALANZA), one of the alternatives is chosen for the main target gloss and the others may be triggered in specific contexts using the SU, AU, or TU.

When dealing with multiple-word descriptors, the linguist will often have a choice among several combinations of collocations and microglossary entries. The following list contains examples of each type of multiple-word unit and of phrases that do not require special dictionary entries:

<u>English term</u>	<u>Type of unit</u>	<u>Spanish gloss</u>
FISH	(Main entry)	PESCAR (verb), PEZ (noun)
FISH	TU	PESCADO
FISH CULTURE	SU	PISCICULTURA
FISH DETECTION	None	DETECCIO/N DE PECES
FISH EXTRACTS	TU	EXTRACTOS DE PESCADO
FISH FEEDING	None	ALIMENTACIO/N DE PECES
FISH LARVAE	SU	ALEVINES
FISH OILS	Same TU	ACEITES DE PESCADO
FISH POISONING	SU	PESCA CON VENENO
FISH PONDS	AU	ESTANQUES PISCI/COLAS
FISH WASTES	AU	DESECHOS DE(L) PESCADO

Although most genus-plus-species strings are handled correctly when only the genus is found in the dictionary, some of these names had to

treated as collocations because of conflicts with other uses of the same lexical items. For example, both elements of Eucalyptus dives have Spanish glosses in the target dictionary (EUCALIPTO and BUCEAR).

Conclusions

The incorporation of specialized vocabularies into the existing machine translation dictionaries requires a substantial amount of effort. In order to be effective, the work must be done by an individual who is familiar with in-house terminology and translation requirements and the capabilities of the MT system. The MT program itself, its supporting software, and word-processing capabilities can be used to speed up some parts of the work, but human decisions have to be made at almost every turn. Clear criteria should be established at the outset of the project in order to save time.

Many of the problems encountered in this case study will probably be repeated when other lexicons are incorporated into ENGSPAN. Yet each lexicon will also present some special problems of its own. The first step in incorporating any specialized vocabulary should be a thorough study of its content and structure. The time required to devise a plan of action will be time well spent. The result of this painstaking work will be a machine translation system with increased subject-area coverage and better quality output in general.

REFERENCE

- Vasconcellos, Muriel, and Marjorie León. 1985. "SPANAM y ENGSPAN: Machine Translation at the Pan American Health Organization". Computational Linguistics 11 (2-3): 122-136.