[*Translating and the Computer 8*. Proceedings of a conference… 13-14 November 1986, ed. Catriona Picken (London: Aslib, 1987)]

Post-editing on-screen: machine translation from Spanish into English

Dr Muriel Vasconcellos

Chief, Terminology and Machine Translation Program,
Pan American Health Organization, Washington, D.C., USA

BACKGROUND

The end of 1986 marked seven years of ongoing machine translation (MT) production at the Pan American Health Organization (PAHO) in the Spanish-English combination. SPANAM, the Spanish-English MT software developed in-house, and its newer English-Spanish partner ENGSPAN (Vasconcellos and León 1985), are free-syntax 'try-anything' systems (Lawson, 1982) designed to be challenged by many types of discourse. And, in fact, in the course of their service to users they have been exercised on a wide range of text types: scientific narrative, technical specifications, instruction manuals, questionnaires, political rhetoric, even film scripts. Early in the project's history it became evident that, contrary to what might be expected from PAHO's mission as a public health agency in the United Nations and Inter-American systems, SPANAM and ENG-SPAN would be enlisted to translate a broad variety of texts.

As of December 1986 SPANAM had provided a total of 3,271,218 words (13,085 pages) of translation under 1,022 job orders to requesting offices within PAHO. These figures do not include translations run for purposes of linguistic development or demonstration. ENGSPAN, for its part, operational since 1985, had produced, in addition to experimental and demonstration text, 1,197,819 words (4,791 pages) of output under 264 job requests. In conjunction with these translations, specialised dictionaries have been built up in a number of subject fields. At the end of 1986 SPANAM's total dictionary data set contained 64,151 entries in the Spanish source plus their respective glosses and codes in the English target, while ENGSPAN's had 49,359 entries in the English source and 51,722 in the Spanish target.

Nearly all the revision of SPANAM and ENGSPAN output is done by professional translators, specially trained in post-editing techniques, who work at the program site. The program is now part of a larger language services component that includes traditional translation as well, and for the last two years the translators/post-editors have combined their work on MT with other linguistic assignments. The degree of post-editing to be done is decided by a series of factors, assessed on the basis of an initial consultation with a representative from the requesting office: the purpose of the translation, the time frame, and linguistic considerations relative to the text itself—nature of the vocabulary, discourse type, the author's clarity of expression. While we pride ourselves on flexibility in response to different situations, in actual practice it turns out that most of our production is delivered in the form of fully polished translations.

The present paper reports on PAHO's experience with the post-editing of English output over the seven-year period and offers a tentative typology of the strategies that have been developed.

GENERAL APPROACH TO POST-EDITING

The strategies used at PAHO to facilitate translator interaction with MT cover a spectrum that might be visualised as ranging from purely mechanical devices, at one end, through a series of increasingly sophisticated applications of linguistic knowledge all the way to direct involvement in the MT system itself. These different strategies may be plotted along a continuum that progresses from the reactive to the proactive (Figure 1).

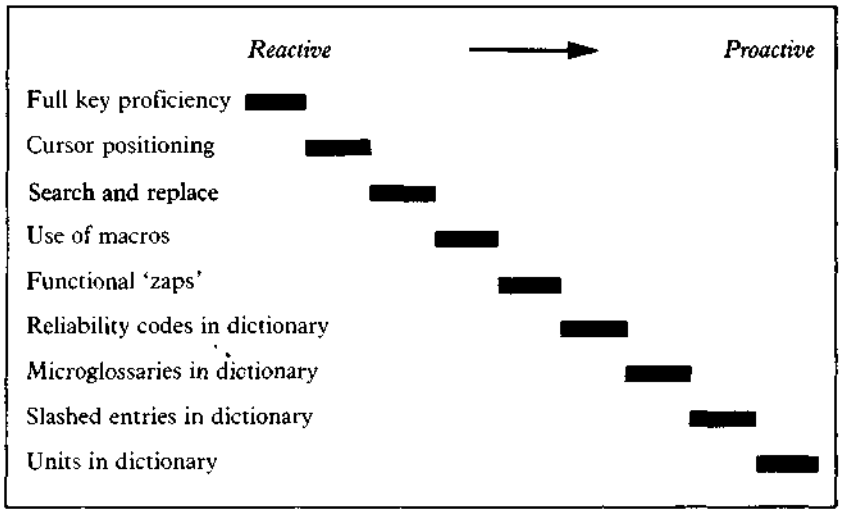

Figure 1. Strategies for the translator/post-editor

Since each step draws on mastery of the preceding steps, it is essential that the SPANAM and ENGSPAN post-editors work directly on-screen. At all levels there are advantages to be gained from this mode of operation. To begin with, corrections are entered more quickly than if they were written by hand on hard copy. If a given change is to be made several times, it need only be performed once and then introduced automatically (either globally or selectively) throughout the text. The effect of this search-and-replace activity is cumulative: the text becomes 'cleaner' as work progresses, and as it more closely approximates its final form the translator/post-editor is confronted with a screen that looks increasingly like the finished product. This helps to economise on re-review, and possibly on reprinting. In terms of the overall text-processing chain, time and money are saved because the final text is immediately machine-readable, for whatever application it may be intended. There is no need to wait for the intercession of an operator—much less a transcriber.

But even more important than these direct advantages, the post-editor's command of the on-screen mode is also the foundation upon which the more sophisticated strategies are based.

At the top of the scale are the strategies by which the translator participates in dictionary development.

The following summary begins with those strategies in which the translator/post-editor reacts to the text as it is presented by the machine and moves progressively up to those that involve the translator in a proactive role of actual contribution to the nature of the machine output.

THE PROGRESSIVE LEVELS: REACTIVE

Considered in terms of creative intervention required of the post-editor, the levels in this group of strategies may be seen progressively as follows:

Full key proficiency

As a prerequisite to everything else that follows, the translator/post-editor needs to be fully proficient with the standard keyboard and also have a thoroughly internalised command of all the functions available through the special word-processing keys, manipulating them with reflex response. The post-editor must not only know the advantages of the different functions—copy, move, search, replace, etc.—but also be able to invoke them creatively without stopping to think.

Efficiency in cursor positioning

As trivial as it may seem, one of the most important factors in moving rapidly through a text is the ability to position the cursor quickly. Constant

and efficient use of the search function can cut time in front of the screen by a significant proportion. Rather than groping with the directional keys, the SPANAM post-editor learns to search for a unique string of characters in the vicinity of the next correction to be made. In order to specify a string that is unique, it is necessary to know about legal and illegal combinations of letters in the particular written language and also, from the field of data processing, something about how format characters are treated in the word processing package that is being used.

Use of search and replace

Effective combination of the search and replace functions is the back-bone of post-editing. Words that are not found, or that need to be changed, can be replaced automatically either on a global basis or selectively, case by case. Thus *underdeveloped* could be replaced by *developing*, or any other fashionable euphemism, in a one-time command. On the other hand, a word that may need to be changed only in certain contexts—for example, *shall* for *will*—would be the subject of a selective replace. It is important to realise that this set of functions can also be used on pieces of words, pieces of words in combination with blanks, and, in the case of some word processing software, even format characters. Hence there are a multitude of avenues that the post-editor can explore that will economise both on keystrokes and on manual searching time. Effective use of these functions requires that the post-editor should combine full key proficiency with a high degree of skill in positioning the cursor.

Use of macros

Moving up the linguistic ladder, we find a series of macros (known as ‘glossaries’ in Wang word processing software) that have been developed by users of SPANAM and ENGSPAN to deal with recurrent situations in the output. Our experience has shown that these aids speed up the process enormously.

The macros are recalled with two keystrokes: the recall key itself plus an arbitrarily assigned keyname. In each case the task that they perform is accomplished at speeds much faster than if they were done manually. Two basic types of macros are used in post-editing: ones that simply move chunks of text without reference to the words being manipulated, and ones that address particular constructions. Within the latter category there are some that offer options associated with particular words and others that deal with given types of construction in general.

Moves

Single words and groups of words can be moved: a pair can be switched (1x1), one word can be moved to the right of two (1x2), two to the right of

one (2 x 1), two to the right of two (2x2), etc. In addition to the combinations just mentioned, the following sets are also available: 1x3,3x1,3x3, 2x3, 3x2, 1x4, 4x1,2x4, 4x2, 3x4, and 4x3. With practice the post-editor becomes adept at recognising particular linguistic constructions that lend themselves to switches of this kind, and grouping of the segments becomes second nature.

Linguistic constructions: specific words

Some of the macros search for specific words and deal with typical situations that they evoke—for example, *among* can be introduced as an alternative to *between.* Others, such as *most* instead of the default *more* (from Spanish *más),* and *than* for *that* (from Spanish *que),* are provisional and soon will not be necessary, since we are in the process of implementing a parser in SPANAM whose results will provide the necessary information for making these decisions computationally.

Linguistic constructions: general

SPANAM also has macros that are structurally based, and these require more sophisticated knowledge on the part of the post-editor. They respond to frequent situations that come up in association with noun phrases and verb phrases—choices that depend on context and cannot be easily programmed into the translation algorithm.

Noun phrases. With noun phrases one of the most recurrent situations arises from the mismatching distribution of determiners in Spanish and English—especially the use of the definite or indefinite article or no article at all. Our post-editors are familiarised with the myriad criteria, many of them pragmatic or functional-informational and particular to the text, that govern the use of articles in English (MacWhinney, 1984). The changes can be quickly made by means of macros. There is one macro, for instance, that simply deletes the next occurrence of *the* in the output translation. In the following example two keystrokes delete the article, and no time is spent positioning the cursor:

(1) S: La teileriosis es transmitida por las garrapatas.
M: Theileriasis is transmitted by the ticks.
P: Theileriasis is transmitted by ticks.

Another macro changes *the* to *a.*

When one or more prepositional phrases postmodify a noun phrase, there are functional reasons for varying the specification of definiteness at the level of the noun phrase, and sometimes at the level of the clause as well. The construction

* *the* N *of the* N

assigns the status of definiteness to two elements, while

*Ø N of Ø N

does not assign it to either. These output configurations can be quickly changed to:

the N of N

or

Ø N *of the* N

The ultimate choice will depend on a number of factors, including the particular cohesive threads of the text.

An even more complex problem with the noun phrase, which intersects with the issue of determiners, has to do with deciding whether or not the head of a post-modifying prepositional phrase in Spanish (typically *de* N_2) should premodify the head noun (N_1) in the English target. This decision, a challenge to the human translator as well, takes into account a multiplicity of considerations—for example, whether or not the phrase is a common collocation in English; whether the discourse is formal or designed to convey a sense of 'shop talk' to those 'in the know'; whether the head noun is an action, state, or process; and, perhaps most important, how the information is distributed with regard to the preceding and subsequent text (Halliday, 1967-68, Quirk *et al*, 1972, Halliday and Hasan, 1976, Vasconcellos, 1985, 1986).

For the SPANAM output there are macros that change a noun phrase rendered as

(the) N_1 *of (the)* N_2

into:

N_2N_1, or, if wanted, the genitive N_2's N_1

For example, in an article on malaria a reference to the cost of research might be dealt with in SPANAM as follows:

(2) S: . . .la inversión en las investigaciones de malaria. . .
M: . . .the investment in the research of malaria. . .
P: . . .the investment in malaria research. . .

(3) S: . . .la inversión de la Argentina en las investigaciones de malaria. . .
M: . . .the investment of Argentina in the research of malaria. . .
P: . . .Argentina 's investment in malaria research. . .

However, the decision to use these macros is not taken lightly. Often the post-editor finds that it is more accurate and more functionally cohesive to

leave the phrase in its extended form. In a related example it can be seen that post-modifying material is lost if the inversion is performed:

(4) S: . . .la inversión en las investigaciones de malaria y otras enfermedades transmitidas por artrópodos. . .
M: . . .the investment in the research [of] malaria and other arthropod-borne diseases. . .
No: *. . .the investment in malaria research and other arthropod-borne diseases. . .
Yes: . . .the investment in research [on] malaria and other arthropod-borne diseases. . .

The SPANAM translator/post-editors are provided with a set of templates and a list of considerations—syntactic, semantic, functional-informational, and pragmatic—to be kept in mind as they deal with noun phrases. They are expected to have a good reason before changing the form N_1 *de* N_2 into N_2N_1. At the same time, however, they are encouraged to reject and modify strings of the following type:

**(theØ)* N *of (theØ)* N *of (theØ)* N

They understand that there is a constraint in English against repeating *of,* which suggests the same case relationship for both or all the nouns, whereas this problem does not exist for *de* in Spanish. With such strings the translators often leave the construction in its extended form and simply vary the preposition that establishes the relationship of one or another term. Such a solution has the advantage of avoiding possible ambiguity in expression of the semantic relationship as well as being informationally more faithful to the original text:

(5) S: . . .el desarrollo de programas de educación nutricional. . .
M: . . .the development of programs [of] nutritional education. . .
P: . . .the development of programs [in] nutritional education. . .

Also, it happens that changing the preposition, rather than inverting the phrase, makes for a *faster* post-edit.

For all these possibilities there are macros at the service of the SPANAM post-editor.

Verb phrases. Verb phrases can also be dealt with using macros. In purpose clauses, for example, the choice between *for* V-*ing* vs. *(in order) to* Vinf is difficult to predict by algorithm, given the combination of semantic and pragmatic criteria that need to be taken into account. The SPANAM post-editor, using a single macro, can switch the default *in order to* Vinf to either *to* Vinf or V-*ing*:

(6) S: . . .el procedimiento para registrar los hogares. . .
M: the procedure in order to register the households. . .
P: . . .the procedure for registering the households. . .

The fronted verb in Spanish presents a major challenge for the translator working into English. This subject is dealt with in detail in the next section. Once again, macros are used in conjunction with solutions to the problem.

Functional treatment of linguistic constructions

There are strong linguistic reasons for keeping the pieces of information in the text, usually expressed as noun phrases, in the same order in which they were presented in the original language. SPANAM's translator/post-editors are sensitive to the information structure, in which the *given* information in a message is presented at the outset and *new* information is introduced gradually, leading up to a *focus of newest* information. The new information in one message becomes given information in the next, serving as a hook on which to attach the upcoming communication. These links form a cohesive pattern within a text and should therefore be left in their original position if at all possible. Information structure is universal, whereas syntactic structure is specific to a given language. It is reasonable, therefore, to assume that the former should be overriding and that syntactic structure, when it differs from that of the original language, may have to be changed in order for a translation to be faithful to the full meaning of the message.

With these concerns in mind, the SPANAM post-editors develop skill in finding solutions that leave the major pieces of information in the respective positions in which they are presented by the machine output.

Prepositions. It often happens that a simple change in preposition will suffice to preserve the order of the text:

(7) S: La diferenciación histológica del RMS del sarcoma de Ewing, neuroblastoma y linfoma no Hodgkin puede ser difícil.
M: The histological differentiation of the RMS of Ewing's sarcoma, neuroblastoma and non-Hodgkin's sarcoma can be difficult.
QF: The histological differentiation of RMS from Ewing's sarcoma, neuroblastoma and non-Hodgkin's sarcoma can be difficult.

This solution has been labelled QF for 'quick fix'. Not only is it quick, it is informationally faithful. On the other hand, in the version below the difficulty of the diagnosis is relegated to the status of *given* information rather than being the point of the message.

No. *It can be difficult to make a histological differentiation between Ewing's sarcoma, neuroblastoma and non-Hodgkin's sarcoma.

V(S)O vs. SVO. In translation from Spanish into English it often happens that the Spanish verb-(subject)-object construction (V(S)O), used in Romance languages with so-called 'presentational' verbs, has to be matched up against the more rigid requirement for subject-verb-object (SVO) in English. Inversion of the sentence, which is necessary if the same *syntactic* structure is to be followed, violates the *information* structure of the original text and, if the post-editor is not careful, can also break up associations with elements that belong together.

(8) S: Se estudiarán todos los pacientes diagnosticados como portadores de LMA que ingresan en la sala de adultos o en la de pediatría del Instituto de Hematología durante el período de 3 años.

M: There will be studied all the patients diagnosed as bearers of AML who enter the adult clinic or in that of pediatrics of the Institute of Hematology during the period of 3 years.

No: ?All the patients diagnosed as bearers of AML who are admitted to the adult or pediatric clinic at the Institute of Hematology during the 3-year period will be studied.

The post-edit above constitutes a communicative misfire because in the original information structure the long noun phrase describing *patients* was intended to be in the position of *new* information rather than *given* information (Halliday, 1967-68, Vasconcellos, 1986). Moreover, the extensive string of post-modifiers makes for too much of a separation between the head noun and its verb. Communication is impaired.

The following alternative introduces the added problem that the time phrase now refers to the wrong antecedent:

No: *All the patients diagnosed as bearers of AML who are admitted to the adult or pediatric clinic at the Institute of Hematology will be studied during the 3-year period.

These difficulties can be avoided, and the information structure preserved, by a simple 'zap' changing the fronted verb to a noun phrase that can serve as subject of the sentence. This 'zap' is facilitated by a macro that seeks out and deletes the upcoming occurrence of *there.*

QF: Studies will be done on all the patients diagnosed as bearers of AML who enter the adult or pediatric clinic of the Institute of Hematology during the 3-year period.

Other examples:

(9) S: **En 1972 se formuló el Plan Decenal de Salud para las Américas.**
M: In 1972 there was formulated the Ten-Year Health Plan for the Americas.
QF: The year 1972 saw formulation of the Ten-Year Health Plan for the Americas.

(10) S: Para su ejecución se ha considerado dos etapas: . . .
M: For its execution there has been considered two stages: . . .
QF: Its execution has been conceived in two stages: . . .

The fronted verb may also be associated with *it*, and there is a macro for dealing with that construction as well.

It is interesting to note that in the functional approach the post-editor works from left to right, backtracking as little as possible. This not only saves time but is more consistent with the natural production of text. It should be emphasised that the 'quick-fix' solutions are used because they are functionally more faithful. They do not correspond to a relaxation of traditional standards, but rather to an improvement in translation style, based on recent advances in knowledge about the structure of discourse and its importance for communication. It is coincidental, and a plus for MT in general, that they are also expedient devices for the post-editing of machine output.

THE PROGRESSIVE LEVELS: PROACTIVE

The translator/post-editors who work with SPANAM are encouraged to become involved in building the dictionaries. In this way everybody wins: the translator gains a sense of control over the output, and this motivation ensures that the updates are done effectively; at the same time, the post-editing translator, working directly with the context, is in the best position to propose appropriate glosses and codes. If suggestions are noted on the side-by-side version of the output at the time of post-editing, time is saved later because the text does not have to be re-reviewed when the updates are actually entered. Dictionary updating can be performed more effectively by the translating post-editor than by a dictionary coder working from less contact with the text. For all these reasons, updating is regarded here as an extension of post-editing on-screen.

Reliability codes

The beginning translator/post-editor, especially if new to the Organization and its text types, often wonders whether or not certain words and phrases

should be changed. There seems to be an idea that the output, produced by a mere machine, is bound to be highly fallible. Although this perception changes as the translator gains experience with the system and with the subject-matter, in any event it is always valuable for users of the system, whether they are post-editors or requesting offices, to know when the terms produced are reliable and do not require further research. SPANAM and ENGSPAN have a means of coding reliable terms so that they are flagged in the output (Figure 2). Translators are encouraged to enter these codes when the appropriate information is available. They are motivated to do the necessary work, knowing that their effort will be captured and hence their own task lightened as post-editor. Also, future colleagues will be spared the duplication of research. As the volume of reliability codes increases, post-editing time will ultimately be shortened as translators are faced with fewer decisions about whether or not to change the output.

Microglossaries

One of the first things that the translator/post-editor notices about the machine output is that, within the same part of speech, many words have different meanings depending on the context—in other words, that they are polysemous. With SPANAM and ENGSPAN there are several ways of dealing with polysemy. The easiest is to introduce an alternate translation in a microglossary geared to a specific subject area. For example:

Spanish	*English*
núcleo	nucleus (biomed.) core (atom.)
cultivo	culture (biomed.) cultivation (agr.)
medios de cultivo	culture media means of cultivation

Microglossaries can be used to reflect to the specific vocabulary of a given organisation. When a term comes from a microglossary, this fact is signalled by a special flag in the side-by-side version of the output.

Slashed entries

Also as a means of dealing with polysemy, the translator can have a personal microglossary that produces more than one translation for a given term. For example:

Spanish	*English*
equipo	equipment/team

```
SPANAM V0981   SPANISH TO ENGLISH   UNEDITED MACHINE TRANSLATION   GLOSSARY=12
09/03/81                                                            PAGE    1
*HDR99S9999999          NO TITLE
```

Otras toxicidades		Other "toxic effects"
Adriamicina. depresión medular, vómitos, estomatitis, flebitis por extravasación, erupción cutánea.		"Adriamycin": "depressed bone marrow", vomiting, "stomatitis", "phlebitis" by "extravasation", skin eruption.
Ciclofosfamida: cistitis hemorrágica, alopecia, disminución de la función gonadal, inmunosupresión.		"Cyclophosphamide": hemorrhagic "cystitis", "alopecia", reduction of the "gonadal" function, "immunosuppression".
Utilizaron varios medios de cultivo.	G2	They utilized several *means of cultivation.

```
EOJ MTS40002                                  DATE 09/03/81,CLOCK 13/35/29.DURATION 00/01/47
```

The test sentences above demonstrate two optional features of SPANAM. (1) If the user so requests, words or phrases having a reliability code of 3 or above can be highlighted with a special symbol. This tells the user that these terms have come from an authoritative source. The terms can, of course, be changed and upgraded to a higher level of reliability when appropriate information is furnished. (2) Micro-glossaries make it possible to specify vocabulary from a given area of discourse or a dictionary provided by a particular user. In the example, the default translation of *medios de cultivo* is 'culture media' (biomedical terminology is the default option in SPANAM), but specification of Glossary 2 (Agriculture) produces, instead, 'means of cultivation'.

Figure 2. Examples of reliability coding and glossary selection

medio	means/environment
tiempo	time/weather
proyecto	project/proposal/draft
su	its/their/his

A macro can be enlisted to isolate the desired gloss. Slashed entries should be used with caution, however, since too many of them will disrupt the cohesive flow of the text.

Units

Perhaps the most powerful and permanent way of dealing with polysemy is for the translator to enter the collocation in the dictionary as a unit. There is a choice of: *substitution units* (SUs), which are hard-wired strings of contiguous words; *analysis units* (AUs), which apply to contiguous words but, depending on the context, need not necessarily trigger the special translation; and *transfer units* (TUs), which apply to noncontiguous collocations and invoke special translations in the target depending on a wide variety of criteria (Vasconcellos and Leon, 1985, León and Schwartz, 1986).

CONCLUSION

Naturally it takes time for the post-editing translator to acquire all the skills that have been cited in this summary. Our experience has shown that it is a process of gradual growth and that some translators who produce excellent post-edits do not in fact command the whole gamut of resources that are available to them. Between SPANAM and ENGSPAN we have had 15 post-editors who have worked directly on-screen. Not all of them have stayed with us, but of the ones who have, it is probably safe to say that they are unanimous in the following conclusions:

— Post-editing skills are developed gradually, and initial judgments are bound to be reversed. The level of comfort is greatly increased at the end of 100,000 words—the equivalent of a month of full-time post-editing.
— Post-editing gets to be more relaxing, and more fun, than translating from scratch. For the same number of words, post-editors are less fatigued at the end of the day.

Some of the SPANAM and ENGSPAN post-editors have produced as many as 10,000 words of polished, camera-ready copy in eight hours.

Those who get involved in development of the dictionaries have found a new dimension in their profession. They enjoy their work.

REFERENCES

Halliday, M.A.K. (1967), Notes on transitivity and theme in English. Parts 1 and 2. *Journal of Linguistics* 3, 37-81, 199-244.

Halliday, M.A.K. (1968), Notes on transitivity and theme in English. Part 3. *Journal of Linguistics* 4, 179 - 215.

Halliday, M.A.K. and Hasan Ruqaiya (1976), *Cohesion in English*. London: Longman.

Lawson, Veronica (1982), Machine translation and people. In her *Practical experience of machine translation.* Amsterdam, New York, Oxford: North-Holland, pp. 3-9.

León, Marjorie and Schwartz, Lee A. (1986), Integrated Development of English-Spanish Machine Translation: From Pilot to Full Operational Capability. Technical Report, Grant DPE-5543-G-SS-3048-00 from the U.S. Agency for International Development to the Pan American Health Organization. October 1986.

MacWhinney, Brian (1984), Grammatical devices for sharing points. In Schiefelbusch, R. and Pickar, J. (eds.) *Acquisition of communicative competence.* Baltimore: University Park Press.

Quirk, Randolph, *et al.* (1972), *A grammar of contemporary English.* New York: Harcourt, Chap. 14 and App. 2.

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

Vasconcellos, Muriel (1986), Functional considerations in the post-editing of machine-translated output. I. Dealing with V(S)O versus SVO. *Computers and Translation* 1(1), 21 -38.

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

AUTHOR

Dr Muriel Vasconcellos, Chief, Terminology and Machine Translation Program, Pan American Health Organization, Washington, D.C., USA.