APPLICATION OF MT IN AN INTEGRATED ENVIRONMENT: UPDATE ON THE PAN AMERICAN HEALTH ORGANIZATION

Muriel Vasconcellos, Pan American Health Organization

Keywords: Machine Translation, Automatic Translation, Computer-Assisted Translation

Abstract: Integration of activities and roles has been the guiding principle in the development and implementation of machine translation at the Pan American Health Organization (PAHO). PAHO has developed its own in-house systems for automatic translation from Spanish into English and English into Spanish.

In the production of MT output to meet the ongoing needs of the Organization, under way since 1980, PAHO has integrated the functions of post-editing, terminology work, dictionary development, enhancement of the current translation programs, and further system development. All post-editing is done by professional translators, who work at the site alongside the computational linguists and are encouraged to acquire an understanding of the algorithm so that they can provide relevant feedback for its improvement. In the newly merged language service, all translators postedit, and this duty is included formally in their post descriptions, as is their assignment to work on the MT dictionaries. These same translators also work in the traditional mode.

1. OVERVIEW

It is now a full decade since the Pan American Health Organization, Regional Office of the World Health Organization, first began its activities in machine translation. MT was intended to address the internal translation load, which averages about 57 per cent into Spanish and 32 per cent into English, with the other two official languages accounting for only a small share--Portuguese, 9.4 per cent, and French only 1.6 per cent.

Development of a system from Spanish into English began in 1976. This combination became operational under the name of SPANAM™ in 1980 and since then has generated some 3 million words of production text for at least 100 end users within the Organization (sample output in Fig. 1). A full-time in-house computational linguist was hired in 1979, and over the years, in response to the experience of "learning by doing" and to shifting needs and circumstances, a series of major changes were introduced in the programs (Refs. 1, 2). This experience gave PAHO the capability of developing an even better and more sophisticated system from English into Spanish, ENGSPAN™, for which partial support was given by the U.S. Agency for International Development (AID).[1] ENGSPAN became operational in

mid-1984 and has already produced, as of June 1986, more than 900,000 words of output (sample in Fig. 2). The system has been installed at AID and will serve its missions throughout the world wherever there is a need for translation into Spanish.

2. SYSTEM CONFIGURATION

In the internal operation at PAHO, which will be the subject of the rest of this paper, the input for SPANAM and ENGSPAN comes primarily from the Organization's normal text processing chain--i.e. documents that are prepared on the Wang word processor for other purposes. Texts are also input for machine translation using optical character recognition (DEST multilingual model, Turbofont 223).

The word-processing documents are telecommunicated to PAHO's mainframe computer (currently an IBM 4381). The Wang OIS/140 serves as a remote job entry terminal (RJE). On the mainframe there is a conversion program, and for each language combination there is a translation program and a pair of dictionaries. As of June 1986 SPANAM's dictionaries had approximately 62,000 paired entries (94 per cent base forms, 6 per cent full forms), and ENGSPAN's had 47,000. There are also programs resident on the mainframe for updating the dictionaries, for printing or viewing lexical entries, and for other housekeeping tasks. Once the text has been translated, it is automatically sent back to the Wang for postediting.

3. AN INTEGRATED WORKING MODE

The integration of functions is an important characteristic of the working environment at PAHO. Many functions combine in the processing of MT output to meet the ongoing needs of the Organization: postediting, terminology work, dictionary development, enhancement of the current translation programs, and further system development. All postediting is done by professional translators, who work at the site alongside the computational linguists and are encouraged to acquire an understanding of the algorithms so that they can provide relevant feedback for improvement of the respective systems.

Recently the human and machine translation services at PAHO were merged. In the new structure, postediting is included in the post descriptions of all entry-level translators, as is the task of dictionary building. The translators work both in the traditional mode and as post-editors. When postediting, they key in their revisions directly at the word-processing screen. The text on the screen is supplemented by a side-by-side hard copy printout which displays the source text on the left, the target text on the right, and a column of diagnostic flags in the center. The information from the side-by-side helps to explain errors in the output. As the translators work, they jot down suggestions for improving the dictionaries--new entries, alternative glosses, deeper syntactic or semantic coding, or flags to signal the reliability of a term--and later they enter the appropriate updates themselves.

With both SPANAM and ENGSPAN the dictionaries are large enough so that not-found words are rare--less than 1 per cent in either case (not counting typographical errors in the input, repeated occurrences of the same not-

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

found word, or alphanumeric combinations that do not affect the text). Still, there is a continuing need for work on the dictionaries, both to refine and deepen the coding of existing entries and also to add idioms that will trigger variant translations for particular contexts.

In order to save repetitious research, approved and reliable terms are specially flagged in the output. The criteria come from internationally approved sources. There is also a data base, WHOTERM, which resides on the Wang and is limited to technical terminology in certain biomedical fields. Terms that are in WHOTERM are signaled in the output as well, but the mark is different so that the translator will know that a complete terminological record is available on the Wang station itself. These two sets of flags amount to an automatic system for retrieving technical terminology in the place where it occurs in the text, obviating some of the frustrations that are ordinarily inherent in the consultation of lexical data bases. The coding of these flags is another area in which the translator contributes to the dictionaries.

In addition to providing information on the status of dictionary entries, the side-by-side printout also alerts the translator to sentences that were only partially parsed or, in some cases, not parsed at all. Unparsed material is analyzed by phrase-level routines, and such output needs to receive extra scrutiny. Based on their review of the diagnostic data, translators sometimes suggest areas in which the algorithm can be improved.

Postediting is facilitated by a series of customized macros at the level of the word processor which are designed to deal with pragmatic distinctions that the MT systems do not yet handle. They can be easily modified. The postediting process itself is also the subject of ongoing linguistic analysis (Ref. 3).

4. Conclusion

While the responsibilities of the different members of the staff give them each a primary focus of concern--for the translators, production; for the two computational linguists, system development, and for the terminologist, coordination of the dictionaries--each activity receives input and support from the others, and all are mutually reinforcing.

The high degree of integration, as described above, makes the MT environment at PAHO rather unusual. We consider that the broad skills of each of our staff, martialed in support of all aspects of the MT systems, has given us maximum returns on what has been a relatively small investment.

REFERENCES

1. Vasconcellos, Muriel, and Marjorie León. SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization. Compputational Linguistics 11 (2/3):122-136, 1985.

2. Vasconcellos, Muriel. Management of the machine translation environment: Interaction of Functions at the Pan American Health Organization. In: Tools for the Trade: Translating and the Computer 5, ed. by Veronica Lawson. London: Aslib, 1985. pp. 115-129.

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

ILRAD
LABORATORIO INTERNACIONAL DE INVESTIGACIONES SOBRE ENFERMEDADES ANIMALES

El Laboratorio Internacional de Investigaciones sobre Enfermedades Animales (ILRAD) se fundó en 1974 con el objeto de ayudar al desarrollo de controles eficaces de dos importantes enfermedades que afectan a la ganadería: la tripanosomiasis y la teileriosis. En conjunto, estas dos enfermedades afectan a la producción ganadera de extensas zonas en unos 50 países en desarrollo de Africa, América Central y del Sur, el Oriente Medio, el subcontinente indio y Asia. Las pérdidas totales que causan, de recursos humanos y económicos, son incalculables, no sólo en materia de leche y carne, sino también de cuero, lana, fertilizantes, tracción animal y otros subproductos animales, y en posibles recursos de capital. Cientos de millones de personas, entre ellos algunos de los más pobres del mundo, resultan gravemente afectados. El ganado G5
rumiante transforma estos elementos a partir de vegetación que el hombre no puede comer, a menudo en terrenos que no puede utilizar para cultivos. En otras zonas, es conveniente la producción integrada de ganado y cereales.

LA ESTRATEGIA DEL ILRAD

Las dos enfermedades mencionadas son causadas por parásitos que son transmitidos por insectos. La mosca tsetsé transmite los tripanosomas y la teileriosis es transmitida por las garrapatas. En ambos casos las relaciones entre parásitos, huéspedes y vectores son complejas y sutiles, y por tanto la intervención es difícil. Además, en ambos casos, otros animales salvajes y domésticos sirven también como huéspedes de los parásitos, creando así reservas de infección prácticamente inaccesibles a las medidas de control.

ILRAD
INTERNATIONAL LABORATORY OF RESEARCH ON ANIMAL DISEASES

The International Laboratory of Research on Animal Diseases (ILRAD) was founded in 1974 for the purpose of helping the development of effective controls of two important diseases that affect livestock raising: trypanosomiasis and theileriosis. Together, these two diseases affect the livestock production of extensive areas in some 50 developing countries of Africa, Central and South America, the Middle East, the Indian subcontinent and Asia. The total losses that cause, of human resources and economic, are untold, not only concerning milk and meat, but also of hide, wool, fertilizers, animal power and other animal by-products, and in possible capital resources. Hundreds of million people, among them some of the poorer of the world, result critically affected. Ruminant livestock transforms these elements starting from vegetation that man cannot eat, often in lands that cannot utilize for *crops. In other areas, it is desirable the integrated production of livestock and grains.

THE STRATEGY OF ILRAD

The two diseases mentioned are caused by parasites that are transmitted by insects. The tsetse fly transmits the trypanosomes and theileriosis is transmitted by the ticks. In both cases the relations between parasites, hosts and vectors are complex and subtle, and accordingly the intervention is difficult. In addition, in both cases, other wild and domestic animals serve as well as hosts of the parasites, creating thus reservoirs of infection practically inaccessible to the measures of control.

Fig. 1. Sample Output of SPANAM TM.

V1085 ENGSPAN (TM) IS A COPYRIGHTED WORK OF THE PAN AMERICAN HEALTH ORGANIZATION ENGLISH TO SPANISH MICRO=7
10/29/85 PAGE 1
*HDR99S9999999 CIP PROGRAMS UNEDITED MACHINE TRANSLATION

*Centro *Internacional de la *Papa	SD SD SD	Centro Internacional de la Papa
ACTIVITIES AND ACHIEVEMENTS	OK	ACTIVIDADES Y LOGROS
CIP's breeding program is producing improved potato	OK	El programa de mejoramiento del CIP está produciendo
genotypes that are insensitive to day length; tolerant of	OK	genotipos mejorados de papa que son insensibles a duración
such environmental stresses as heat, cold, drought, and		del fotoperíodo; tolerante de dichas tensiones ambientales
soil salinity; high in energy content; and capable of high	OK OK	como calor, frío, sequía, y salinidad de suelo; alto en
yields under the various agroclimatic and agronomic regimes		contenido de energía; y capaz de rendimientos altos bajo
that are typical of the lowland tropics.		los diversos regímenes agroclimáticos y agronómicos que son
		típicos de las tierras tropicales bajas.
Resistance to the pests and diseases that attack potatoes	PP	La resistencia a las plagas y enfermedades que atacan
in the tropics is a prime requisite in improved varieties.		papas en las zonas tórridas es un requisito básico en
Breeding lines at CIP are screened for resistance to a	OK	variedades mejoradas. Las líneas de mejoramiento en el CIP
number of fungal diseases: late blight and bacterial wilt	OK	se examinan para resistencia a varias enfermedades causadas
are commonly associated, constituting a major impediment to		por hongos: el tizón tardío y pudrición bacterial
potato production in the tropics. Combined resistance is	OK	comúnmente se asocian, constituyendo un impedimento
now incorporated in CIP varieties.		principal a producción de papa en las zonas tórridas. La
		resistencia combinada ahora se incorpora en variedades del
		CIP.
Among the 20 known virus diseases of the potato, the two	OK	Entre las 20 enfermedades conocidas víricas de la papa,
that significantly affect production--potato virus Y and		los dos que significativamente afectan la producción--virus
potato leaf roll virus--have received the most attention at		Y de papa y virus de enrollamiento de la hoja--han recibido
CIP. Good resistance to virtually all major viruses is now	NO	la mayoría de atención en el CIP. La resistencia buena
available and in the process of incorporation into national		a virtualmente todos virus principales es ahora disponible
breeding populations.		y en el proceso de incorporación en poblaciones nacionales
		de mejoramiento.
CIP researchers are also screening germplasm for	OK	Los investigadores del CIP también están examinando el
resistance to the insect vectors that transmit viral		germoplasma para resistencia a los insectos vectores que
diseases. The principal vectors are the green peach aphid	OK	transmiten enfermedades víricas. Los vectores principales
and the potato aphid; others include leafhoppers, leaf	OK	son el áfido verde del durazno y el áfido de papa; los
miners, potato tuberworm moths, and weevils. On certain	OK TU	otros incluyen saltarillas, minadores de hojas, polillas de
hybrid potato plants, CIP researchers have noted glandular		papa, y gorgojos. En ciertas plantas de papa híbrida,
foliar hairs with sticky tips that trap insects--mainly		investigadores del CIP han notado pelos foliculares
aphids, but also flea beetles and mites--reducing epidemic		glandulares con puntas pegajosas que atrapan insectos--
infestations.		principalmente áfidos, pero también pulgillas y ácaros--
		reduciendo infestaciones epidémicas.

Fig. 2. Sample Output of ENGSPAN™.