SPANAM® and ENGSPAN® for Windows® 2000: An MT Pioneer Keeps up with Technology

Marjorie León

Pan American Health Organization
525 23rd St., N.W.
Washington, DC 20037
leonmarj@paho.org

Abstract

The Pan American Health Organization (PAHO) is proud to present the latest release of its fully automatic Spanish-to-English and English-to-Spanish machine translation systems. SPANAM and ENGSPAN have been ported to the 32-bit Windows platform. The bilingual graphical user interface provides easy access to all the features of the system. The translation engine can be accessed in three different ways: file translation from the desktop or word processing application, sentence translation from within the dictionary update module, or cut-and-paste translation using an ActiveX component. Any user can view all of PAHO's dictionary entries (words, expressions, and rules), and dictionary coders can add new entries of every type and modify all but a small number of protected records. The system is designed to be used by translation professionals in an institutional setting. Administrative utilities include job accounting, dictionary update log, terminology import and export, and dictionary merge. Users can view and print side-by-side listings of source and target texts, lists of not-found words, and the parse of any sentence.

Keywords

Machine translation; Spanish; English; electronic dictionaries; natural language processing

System Requirements

The PAHO MT system is a 32-bit Windows application. It runs under Windows 95/98/NT and Windows 2000. Both standalone and LAN versions are available. The full installation requires 100 MB of hard disk storage. A Pentium processor and a minimum of 32MB RAM are required. The software is protected with a hardware security device.

User Interface

The language of the graphical user interface can be selected at runtime. All menus, dialogs, and help files are available in both English and Spanish, including ToolTips and context-sensitive help.

Types of Translation

PAHO's MT system is intended principally for translation of technical and scientific texts. It is also used at PAHO to translate administrative documents, training manuals, and official correspondence.

Language combinations

The system includes two translation programs: SPANAM which translates from Spanish to English (U.S.) and ENGSPAN which translates from either U.S. or British English to Spanish.

File formats

The following document formats are accepted and preserved by the system: MS Word 97/2000 and WordPerfect (saved in Rich Text Format), HTML, SGML, and text files.

Runtime options

The user can select specialized vocabularies (microglossaries) and special grammars. The user can have the program flag those target glosses that are coded as highly reliable and select other special processing features at runtime.

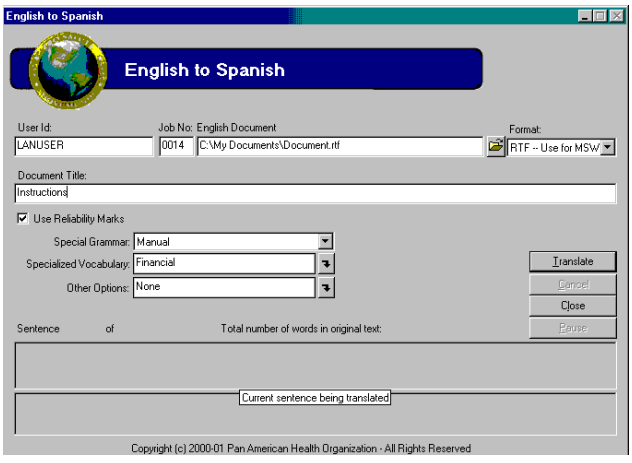

Figure 1: Translation dialog with runtime options

Translating test sentences

Test sentences can be translated from within the dictionary update module. All the runtime options for file translation are also available in this mode. The linguist or dictionary coder can view the parse of the sentence and ascertain which dictionary entries are being triggered.

Translating using the ActiveX component

Translation can also be done from within MS Word by highlighting short segments of the text and clicking a button on the PAHO translation toolbar. If no text is highlighted, a dialog box opens and the user is prompted to enter a short text to be translated. The user can select up to three specialized vocabularies in this mode.

Postediting macros

Many common postediting operations are automated and easily accessible to the translator on postediting toolbars. There are separate toolbars for postediting Spanish output and English output.

Dictionaries

SPANAM uses a Spanish source and an English target dictionary. ENGSPAN uses an English source and a Spanish target dictionary. Each dictionary has approximately 90,000 entries.

Types of entries

The source dictionaries contain the following types of entries:

- Uninflected single words (maximum 30 characters)
- Full forms of irregular or ambiguous words
- Multi-word expressions (2-5 words)
- Delayed verb+adjunct expressions
- Nested expressions (2-5 words or expressions)
- Analysis units (2-5 words or expressions)
- Transfer rules (trigger, context word(s) and conditions to be tested)
- Microglossary entries for single words

The target dictionaries contain the following types of entries:

- Single or multiple word translations (maximum of 102 characters)
- Microglossary translations

Microglossaries

If a specialized translation in a particular subject area conflicts with the main translation of a word, the specialized translation can be added as a microglossary target entry. If a source word has different syntactic or semantic features in a particular discipline, it can be added as a source microglossary entry. There are currently 13 microglossaries in use. They contain only those words that conflict with main entries. Up to 5 microglossaries can be specified at runtime in order of priority. The client can define six additional microglossaries as needed.

PAHO master dictionaries

The dictionaries distributed by PAHO contain all the entries and information used by PAHO's Translation Services for its own production translations. The dictionaries are updated daily based on feedback from the staff translators and contract posteditors. The record history fields indicate the coder who added the record, the date on which it was added, and the coder and date of the last change.

Client dictionaries

Clients are encouraged to add their own terminology and rules to the dictionaries. New client entries are assigned ID numbers in a reserved sequence. Whenever a client modifies a PAHO entry, that entry is flagged for inclusion in subsequent merges.

Translation Algorithm

Analysis

The system performs morphological analysis, single word lookup, and expression lookup, including analysis units. After this bottom-up strategy, a top-down, left-to-right parse is performed. The grammar is expressed as an augmented transition network (ATN). The ordering of the arcs depends on the text type selected by the user at runtime. The parser uses limited look-ahead, a well-formed phrase list, a hold list, and explicit backtracking. The parser returns a linked list of annotated phrase structure nodes for the first successful parse or the longest path. In the case of a partial parse, another set of bottom-up procedures is called to do a local analysis of the remaining words.

Transfer

The syntactic structure of the source sentence is examined to determine the functions and roles of its constituents. Then lexical transfer rules are applied to select the appropriate target glosses. The rules can test for the presence or absence of many syntactic and semantic relationships. Syntactic transfer rules produce the surface structure of the target sentence.

Synthesis

Codes contained in the target entry may trigger additional structural modifications. Phrase-level reordering and some sentence-level reordering is performed. Morphophonemic rules generate the surface forms.

Output Files

The program produces several different output files. In addition to the raw translation in the format of the input document, a side-by-side file with source and target texts aligned by paragraph and a list of not-found source words are produced automatically. The user may also request a listing of the output of the parser.

Dictionary Utilities

Browse

The user interface includes a browse function that allows all users to look up terms in the dictionaries. No password is required. Terms may be copied and pasted into another application.

Update

The dictionary update module allows the coder to add and modify all types of entries using the same interface. A tree control is used to display a source entry and all its possible target translations. Related source entries are displayed in a pop-up window. Powerful search functions produce lists of expressions and rules triggered by or

containing a particular word. For each part of speech, the applicable syntactic and semantic codes are modifiable using combo boxes and other customized controls. See Figure 2 for an example of a single-word entry.

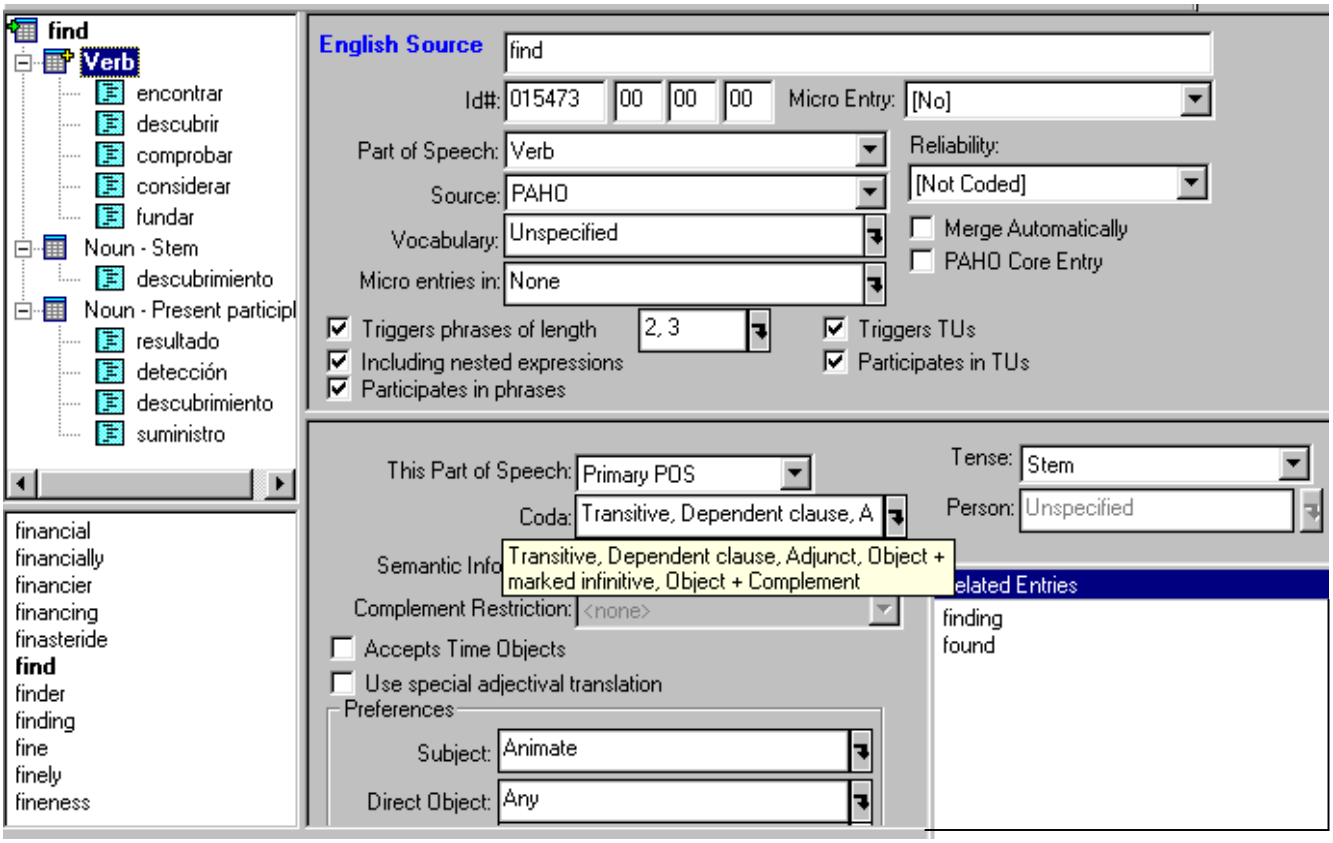

Figure 2: Source entry for the English word "find" with focus on Part of Speech "Verb"

Audit Log

All dictionary updates are logged by date and time, operator, and lexical key. The log can be reviewed to track the changes to a particular entry or to monitor the work being performed by different coders. The audit log is also used as input to the merge utility.

Import

The system includes a utility that permits the semiautomatic import of a list of terms from a text file. Terms and be imported into one or both sets of dictionaries from the same list.

Export

Clients may create lists of the dictionary entries that they have added or modified. They may also create a separate dictionary containing only the new terms added at their site.

Merge

Figure 3 on the next page shows the main menu for the dictionary merge utility. This utility allows clients to merge their working dictionary with a new PAHO master dictionary received as part of a system upgrade. The utility first creates lists of the entries that were created, modified, or deleted at the client site. The user then proceeds to view the lists, consult either set of dictionaries to check any doubtful entries, remove any undesired entries, and merge the rest. Lists of the successfully merged entries and the duplicate or discarded entries are stored in log files.

Administrative Functions

Job accounting

All file translations are registered in the job accounting file according to the user ID. The date, title, number of words, and other statistics are also recorded for each job. This file can be used to monitor the use of the system or to keep track of work done for different customers.

Password maintenance

Passwords are required to update the dictionaries and to perform administrative functions. Users do not need a password to run a translation.

File maintenance

Utilities are provided to compress and reindex the dictionaries, the dictionary audit log, and the job accounting file.

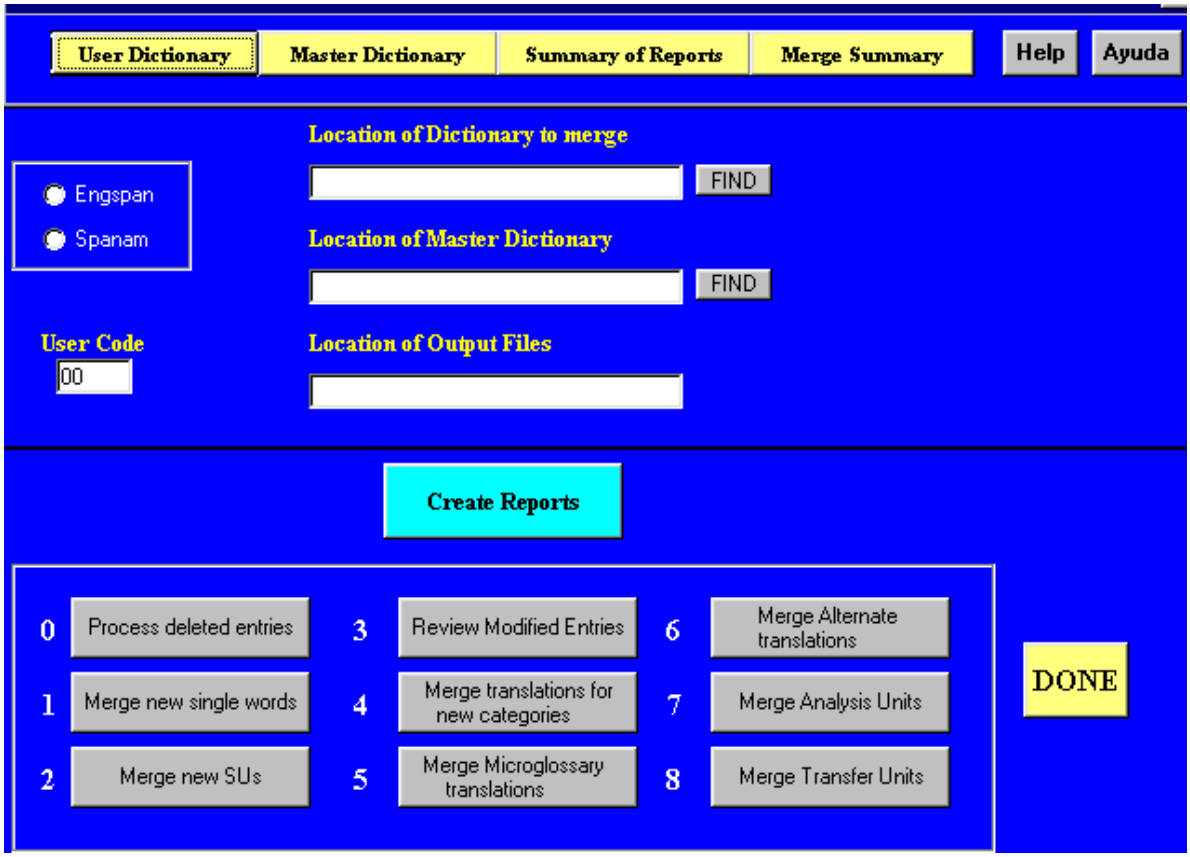

Figure 3: Main menu for the interactive dictionary merge utility

Licensing and Support

License agreement

The software is copyrighted by the Pan American Health Organization. Clients must sign a license agreement. License fees are used to defray the cost of maintenance, distribution, and support.

Security

Each licensee is provided with a hardware security device that must be present in order to run the software. On a local area network, the client software may be installed on any number of computers. The security device permits either 5 or 10 simultaneous users.

Training

PAHO offers a 5-day training course for new users. The course includes how to prepare and run translations, postedit the output, and update the dictionaries.

Support

Basic support for running translations is provided to all licensees. Support for dictionary updating is available with an annual support agreement.

Upgrades

Upgrades of the PAHO MT system are released every 3-6 months. They contain improvements to the translation engine and the user interface, expanded help files, and new dictionary entries. Upgrades are included in the annual support agreement.

Acknowledgements

The PAHO MT system is the result of 25 years of work by the small, but dedicated staff of the Translation Services unit. The development of the 32-bit Windows version began in 1999, under the direction of the author. Julia Aymerich was largely responsible for the design of the dictionary update interface and the dictionary merge utility.