De-Hyping Translation Memory: True Benefits, Real Differences, and an Educated Guess about its Future

By Jost Zetzsche

In a literal sense, translation memory programs are applications that extend the memory of the translation professional by allowing him or her to build up databases of translated material and leverage that against newly translatable content. However, this general definition does not take into account all of the capabilities of these programs.

These tools also allow the user to build up terminology databases that complement and extend the functionality of the translation memories. They enable translators to work in very complicated file formats that they may not understand or otherwise be able to support. Furthermore, many of these programs provide methods for analysis, quality assurance, and productivity that go way beyond the typical functions of a mere "translation memory" tool.

It may be surprising for some to learn that translation memory tools have been around for some time. The TRADOS translation company, founded in 1984, released its first commercial product, MultiTerm (TRADOS' terminology management component), in 1990, and Workbench (TRADOS' translation memory application for DOS) in 1992. In 1994, TRADOS released a Windows version with an MS Word interface. The translation agency Star—like TRADOS, founded in 1984—released a product that was originally designed for in-house use: Star Transit, with its terminology component TermStar. IBM released its Translation Manager (TM/2) product in 1992 (and buried it in 2002). And, as the first Windows-based commercial product, Atril's Déjà Vu was released in 1993.

While a large number of other tools have entered the market in the ensuing years, old tools are being discontinued at nearly the same pace (such as IBM's Translation Manager, Alpnet's TSS/Joust, SDL's Amptran, or Cypresoft's Trans Suite 2000).

Categories of Translation Memory Tools

The commercially available tools can be classified into two main categories:

- Tools that perform all or most of their work through macros in Microsoft Word that allow interaction between translation memory(s) and terminology database(s); and
- Tools that let the translator work in an independent, mostly tabular environment.

> "…While translation memory does make us all more efficient, ultimately any translation memory or terminology database is only as good as the translation professional who created it…"

Tools Using Microsoft Word as Their Main Translation Interface. The most well known application of the MS Word plug-in is TRADOS (see www.trados.com). Others include:

- MultiTrans (www.multicorpora.com);
- Wordfast (www.wordfast.net);
- MetaTexis (www.metatexis.com);
- WordFisher (www.wordfisher.com, now available as freeware);
- Fusion (www.orcadev.com); and
- Logoport (www.logoport.net).

This group can be categorized by its range of applications:

- TRADOS covers a very large range of applications and formats through its RTF and tagged text conversion utilities and, lately, with the increased use of its new interfaces, the TagEditor and the T-Windows collection.
- MultiTrans offers several other translation interfaces besides MS Word, including WordPerfect and PowerPoint.
- The other tools are primarily focused on native MS Word files or other formats that can be accessed through MS Word (either through a tagging mechanism or by "calling" into other applications).

MultiTrans does not completely fit in the above category. In fact, its developers would argue that it is not a "translation memory" tool at all, but a "bi-text" or "corpora" tool. Instead of matching on a sentence-by-sentence level, MultiTrans' corpora are full source- and target-language texts that contain the entire context of the original. These texts have an approximate matching capacity that allows alignment (the generation of translation memory databases out of previously translated text) to be done virtually on the fly. MultiTrans was designed to cater to the needs of the Canadian government, whose millions of pages of translated content from and to French and English proved too much to put through a manual alignment process.

Tools Using an Independent Translation Interface. The last group—the category that presents all files in an independent, uniform format—includes:

- Lionbridge's open source application ForeignDesk (www.foreigndesk.net);
- German newcomer across (www.across.net);
- Singaporean Heartsome (www.heartsome.net); and
- The better-known applications: SDLX (www.sdlx.com), Star Transit (www.star-transit.com), and Déjà Vu (see www.atril.com).

Lionbridge truly broke new ground when it released ForeignDesk as an open source software application in 2001, but it never really caught on, partly because of its interdependence with many of the TRADOS tagged formats.

Heartsome offers a whole new approach to translation memory because it uses the exchange formats TMX (for translation memories), TBX (for terminology databases), and XLIFF (for multilingual translation files) as its primary working format. Any supported file type is first converted to one of these data types before the actual translation occurs. Another first for Heartsome is that it equally supports Windows, Mac, and Linux platforms.

In the following analysis, I will concentrate primarily on three somewhat representative tools: the veterans TRADOS, Transit, and Déjà Vu, and the relative newcomer SDLX.

The Similarities

In some areas, the major tools have become increasingly similar (aside from their inherent similarity that they all have translation memories and display translatable text). For instance:

- All of them have a freely configurable terminology component that plays a fairly major role in the translation workflow.
- All of them process texts in Unicode, thus providing access to all languages that are supported by the Windows operating system.
- All of them have a feature that allows alignment.
- All of them provide support for TMX, the XML-based translation memory exchange format.
- All of them support a similarly broad range of file formats (including desktop publishing, word processing, and software development formats) and tagging standards (HTML and SGML).
- All of them allow concordance searches (searches for a single word or expression within larger segments) in their translation memories.
- All of them have some general word processing features, such as spell checking or a "search & replace" function.
- All of them provide fairly detailed analysis features.

This is more or less where the similarities end.

The Differences

Translation Memories. Transit is fundamentally different from TRADOS, SDLX, and Déjà Vu in the sense that rather than using an external translation memory, which stores previously translated information, it provides for a "virtual" translation memory—Transit refers to it as "reference material"—by associating already translated files. The benefit of this is that there is no need for an "additional" database that may tie up computer or server space. The drawbacks are the large amount of translated file pairs that have to be retained to provide for the necessary "reference material," plus the fact that the user has to have a very good overview of what is actually contained in the legacy material.

A major difference between the translation memory databases of TRADOS, Déjà Vu, and SDLX lies in their formats. In its non-enterprise versions, TRADOS uses a proprietary database format, while Déjà Vu and SDLX use a generic industry-standard format. TRADOS uses the same database engine for its databases as Microsoft Access. This differentiation is probably not so relevant for a beginning user, but it can make a difference for an advanced user. TRADOS' database concept limits the user to predefined ways of "communicating" with the databases, whereas the open database concept of Déjà Vu and SDLX allow for more extensive database access with readily provided tools and through standard SQL (Structured Query Language) commands.

Terminology Handling. What are terminology databases good for if all the relevant material is already located in the translation memories? For new users of translation memory programs, the use of the terminology databases often seems superfluous, if not downright confusing. There are several reasons for this:

- Obviously, the name "translation memory" program seems to suggest that the emphasis is on the translation memory. (Most of the major applications have recognized this and do not actually use

that terminology anymore.

- There is a more immediate gain through perfect and fuzzy matches on a sentence-by-sentence basis than there is with terminology databases.

- Translation memories can be built up relatively quickly by aligning existing translated file pairs automatically as you translate new texts.

- The construction of terminology databases is a comparatively tedious process. Terms have to be individually highlighted in the translation or even entered into the terminology management application, and additional information has to be entered.

If it is indeed so tedious to build up and use terminology databases, what makes them so important?

The terminology database is the place where translation professionals can invest effort into defining words and phrases grammatically, contextually, or even by contrast. If this is very helpful for a single translator, imagine how much more beneficial it is in a virtual translators' workgroup! Of course, none of this is news to anyone: any good dictionary offers the same concept. What makes these "dictionaries" much more exciting is that they are completely customizable. Furthermore, they are "living dictionaries" that present their findings for each segment being translated.

Why, then, is it helpful to have numerous translations for—let's say—"cat" ("feline animal," "computer-assisted translation," "Caterpillar," etc.) pop up during the translation of a text? Because of the close association of the terminology databases with a given translation project, and because of all the information that has been fed into the terminology database as the terms were entered, the application will actually recognize which of these terms is more relevant. Depending, for instance, on whether a text from the subject area "Flora and Fauna," "Translation Technology," or "Heavy Machinery" is being translated, the application will make the more likely choice for the correct term (while still allowing the translator to access the other ones).

For the most part, TRADOS and Transit have marketed and sold their terminology databases either as part of a combination with their translation memory solutions or as a stand-alone solution.

Transit probably concentrated most of its earliest efforts on developing a sophisticated terminology tool called TermStar. It allows you to enter a large variety of data: client, date, definition(s), homonyms, and, of course, translation. While this tool can be opened and used as an individual application, it becomes part of the workflow and interface when used in actual translation work. The terms that are displayed in the "Dictionary" windows in the main translation interface can be entered with the associated keyboard shortcut. New terms can be entered by highlighting source and target terms and entering them in the attached dictionary by choosing the appropriate menu commands.

TRADOS' new terminology program, MultiTerm iX, has a similar range of applications to TermStar's, and has a number of advantages over the older, relatively unpopular version of MultiTerm. For example:

- It is based on standard XML and the Microsoft Access engine rather than a proprietary database format.

- It exports into XML, HTML, and RTF.

- Term entry has become less cumbersome.

Déjà Vu and SDLX have gone a slightly different route with their handling of terminology. Their terminology components have always been closely integrated into the translation workflow—the terminology components were never sold and marketed apart from the main applications. In fact, terminology database files are accessed and maintained directly from the main interface. Just like TRADOS and Transit, the terminology databases are concept-based and completely configurable.

One unique aspect of the terminology treatment of Déjà Vu is the so-called "assemble" feature. This feature provides the possibility of piecing together segments that cannot be found in the translation memories or by fixing fuzzy matches from the translation memories and turning them into perfect matches.

The Work Environment for the Translator

There are three major areas where the work environment has an immediate impact on the work of the translator:

- The interface of the translation work;
- The file handling capabilities; and
- The code handling.

The Interface. Transit, SDLX, and Déjà Vu offer translations in an independent yet completely different interface from each other—Transit's interface is more that of a bilingual text editor, whereas SDLX's and Déjà Vu's is that of a bilingual table. Users of these tools describe the benefits of handling text in independent applications, such as in SDLX, Transit, and Déjà Vu, as follows: ➡

- Regardless of the source text and file format, the translation environment always stays the same.

- No "file conversion" is necessary to make texts display in an environment such as MS Word.

- The program is independent of any third party (i.e., an upgrade to MS Word does not have any effect on the computer-assisted translation tool).

On the other hand, TRADOS' traditional method of displaying translatable text in MS Word gives many users the comfortable feeling of operating in a familiar environment, and they find the full WYSIWYG ("what-you-see-is-what-you-get") interface helpful when translating DOC or RTF files. However, because this was originally true only for these formats, since every non-RTF or non-DOC file had to be converted to an RTF or DOC file before processing, TRADOS introduced an additional interface. The TagEditor was originally intended only for HTML- and SGML-type files, but it now allows for the processing of a variety of tagged file formats (including RTF, HTML, SGML, or various DTP formats), and even Excel, PowerPoint, and MS Word in TRADOS-specific TTX (TradosTag format) files. Other TRADOS interfaces for other file types, such as the so-called T-Window programs, include different environments for software resource files, executable files, or clipboard content.

Code Handling. With the exception of plain text files, every file type contains some kind of coding information. This coding information can serve a variety of purposes, including formatting, programming information, and hyperlinks. Translation memory tools have to deal with two different kinds of codes:

- Code that appears only outside (or between) segments; and
- Code that appears within segments.

All of the tools make only the text and the inner-segment codes accessible to the translator. However, TRADOS and Transit store the actual codes for the formatting, whereas Déjà Vu and SDLX only store placeholders for these codes. The benefit of the former method is that you can store a change of formatting (let's say, from bold to italics) between source and target languages. The benefit of the placeholders is that it is 100% compatible across file formats and formatting information.

File Handling. In the actual translation work, Déjà Vu and Transit think in terms of "projects." Everything is treated as if it were one large file, with all the translatable information being accessible at the same time. Though SDLX works with projects, the only source format where it allows true batch translation is HTML, through a process of "gluing" a number of tagged files into one large file. Otherwise, it performs translation on a file-to-file basis. TRADOS, on the other hand, always performs translation strictly on a file-to-file basis. The difference in this approach becomes particularly apparent when working on translation projects with an extremely large number of files, such as websites or software localization projects.

The Work Environment for the Project Manager

From the project manager's viewpoint, the work environments for these tools differ significantly from the translator's. File handling for the project manager (which consists of analyzing, pre-translating, and post-processing) is done through batch processes in all four tools. In the case of SDLX and TRADOS, translators may receive numerous individual files, but the project manager only processes them in batches.

While Transit, SDLX, and Déjà Vu offer the ability to prepare files so that translators can translate them using a free download of the appropriate program, TRADOS does not offer this functionality. On the other hand, TRADOS is clearly the most frequently used tool among freelance translators, so there is a high likelihood that translators already own TRADOS. And lastly, should the translator not own TRADOS, TRADOS files can be processed using many of the other translation memory tools. TRADOS offers Web-based translation memories, and TRADOS, SDLX, and Transit offer Web-based terminology databases. Déjà Vu does not offer any of these features.

Of course, all these tools do offer multilingual processing, but only Déjà Vu and SDLX allow the possibility of having all languages contained in one translation memory, thus reducing setup time for the project.

A Glance into the Future

The midterm developments for translation memory technology and use will be in the following areas:

- Stronger integration into content management systems;
- Stronger emphasis on integrated workflow/project management capabilities;
- Online access to databases as a common and well-used feature; and
- Translation memories as marketable assets.

Content Management Systems. Both TRADOS, SDLX, and Transit have been very active in this area; TRADOS and SDLX through a number of partnerships with content management providers, and Transit through the development of compatible content management systems. While this is a positive development, it presents somewhat of a challenge to the language provider. If there is a true integration of content management and translation memory, either the clients will be more heavily involved in translation memory management or else the vendors will have to broaden their service portfolio.

Workflow/Project Management Integration. Various systems, including TRADOS and SDLX, have integrated a strong workflow component. Other primarily workflow systems, such as Idiom, have an integrated translation memory component.

I believe that most multi-language vendors today would describe the need for an adequate workflow system as being equally if not more important than a translation memory system. Considering this, it is not surprising that computer-assisted translation vendors are trying to cater to that need. When the workflow systems are able to cover most or all project management needs, including accounting, these tools will have a tremendous impact on the industry.

Online Access. Virtually all of the translation memory systems that were released in the last couple of years—Fusion, Logoport, across (as well as the more well known tools, such as TRADOS, Transit, and SDLX)—have at least one online access component, thus clearly emphasizing the need for that feature. Virtual workgroups and routine high-speed access have long become a reality, and the only effective way to exchange data is through the use of common databases.

Translation Memory as a Marketable Asset. While to some this may seem like the most nebulous development, it is inevitable. So far the discussion surrounding this has not progressed beyond the copyright and quality concerns, but interest in a marketplace for translation memory data has long been present. (In fact, it has also long been a reality, as is evident by the popularity of the Microsoft "Glossaries.")

When end clients who hold most if not all the copyrights to their translation memories recognize that they can receive a more immediate return on the investment of their translation costs, it will only be a matter of time before they begin offering their translation memories for sale.

For a recent development on this, visit www.tmmarketplace.com.

Conclusion

The use of translation memory has become standard for translation professionals in the technical field, and this is increasingly true in the medical, legal, and financial fields as well. Has it made our profession more reliant on technology? Probably. But at the same time, has it replaced or removed the "human element" of translation in any shape or form? I would emphatically deny that. While translation memory does make us all more efficient, ultimately any translation memory or terminology database is only as good as the translation professional who created it.

ata

ATA Honors and Awards Committee Seeks Readers

The Honors and Awards Committee needs to expand its corps of readers for its two translation prizes: the Ungar German Translation Award for a distinguished literary translation from German into English, awarded in odd-numbered years, and the Lewis Galantière Translation Award for translations from any language, except German, into English, awarded in even-numbered years. The first reader for each book nominated must be fluent in the source language; the second reader need not be. Readers are furnished with a formal report form and have roughly two months in the summer to evaluate the book(s). There is no honorarium, but readers may keep the book(s) they evaluate. For more information on responsibilities, please e-mail Honors and Awards Chair Marilyn Gaddis Rose (**mgrose@binghamton.edu**). Anyone ready to volunteer now should e-mail Gaddis Rose, with a copy to Teresa Kelly (**teresak@atanet.org**) at ATA Headquarters.