Tech

Translation memory: state of the technology

Jost Zetzsche

It's time for the day of reckoning: What is the state of translation memory (TM) two years after SDL, developer and owner of the market's second-leading tool SDLX, bought the industry gorilla TRADOS in the summer of 2005?

Well, much to the surprise of some doomsayers, the market for TM has never been more alive!

As of this writing, SDL Trados has released version 8, STAR Transit has released Service Pack 20 for its current version, and Déjà Vu is at build 302 of its current version, to quote just a few examples that show the ongoing development efforts of existing tools. In addition, two full-fledged new tools (MemoQ and Lingotek) were released within the last 18 months, and at least two more tool suites will be released later this year. All together, about two dozen commercial and open-source tools on the market provide the main features of the typical translation environment tool (TM or bitext, terminology management, project management, and quality assurance [QA]), and a host of smaller tools offers other specific features in the realm of conversion, database management and QA.

So there must be a huge market out there, right?

Not so, says Common Sense Advisory's Donald A. DePalma: "The entire market for translation automation and localization tools, including translation management and machine translation, was somewhere in the neighborhood of US$100 million for 2006. When I followed the application development tool market back in the 1990s, you could find any number of single tools doing $10 [million], $25 [million] and even $100 million per quarter – that's for just one product, not for an entire industry."

And note that these numbers include many more tools – such as localization, machine translation (MT) and management tools – that are not part of the two dozen translation environment tools mentioned above. This raises the obvious question: Does this eagerness among developers to create new and better solutions stem from pure altruism, or do they see potentially lucrative market scenarios that still need to be tapped into?

Personally, I think the answer lies somewhere in between.

But before we delve into the current and upcoming trends in this field, let's make sure that we have our nomenclature straight. I've been on a campaign to supplant the terms *computer-assisted translation* (CAT) or *TM tool* with the more encompassing and therefore more accurate term *translation environment tool*. In an article I wrote in this magazine (*MultiLingual* #77 January/February 2006, www.multilingual.com/zetzsche77), I explained some of the details behind this personal terminology crusade. Since the same article also outlines the history of TM technology, I won't bore you with that here, either.

Written six months after the SDL-TRADOS deal, that article identified the following as the primary areas for "continuing development":

- terminology management
- content management
- workflow modules
- MT components
- the translation file exchange format XLIFF
- open source
- TM exchange

Let's look at these categories in terms of what has happened in each over the course of the last couple of years.

Terminology management

Terminology management could also be called the "thorn in the flesh of the translation industry." As has been lamented by many before, terminology work in general has been greatly

Jost Zetzsche is a co-founder of International Writers' Group and TM Marketplace and author of The Translator's Tool Box: A Computer Primer for Translators.

neglected (see the excellent *MultiLingual* #87 April/May 2007 about the current state of terminology work) – and this despite the fact that most translation environment tools provide relatively sophisticated terminology management facilities, which have now also increasingly been extended into QA tools.

Ironically, it is not so much the traditional terminology modules that have started to change perspectives on terminology work. Instead, there is a new emphasis on the subsegment level of the otherwise more traditional sentence-level translation unit. Tools such as Lingotek, Similis and MultiTrans have finally had some success in creating a general awareness that there is great value in small subsegments (Similis calls them "chunks") contained within larger structures. Other tool vendors are already working on solutions that will follow suit.

At the same time, terminology harvesting tools from SDL (PhraseFinder and MultiTerm Extract), LogiTerm, MultiTrans and Similis have made it possible to quickly build up terminology databases and are increasingly available for users outside the corporate environment.

Unfortunately, TBX, the TermBase eXchange format, has not had much of an impact. Aside from the standard's most ardent implementer, Heartsome, only a handful of other tools actively supports it. I recently spoke with a developer of a standalone terminology tool who mentioned that TBX was just too complex to support with his tool, so he decided instead to use the much simpler TMX format (Translation Memory eXchange) for exchanging his terminology databases. The TBX-Lite standard that is currently under development may end up helping this issue.

Content management

Content management was promising a year and a half ago and still is today. More tool developers have formed partnerships with content management system (CMS) providers. SDL, Idiom and across have been particularly active in that area by forming partnerships with the likes of the makers of Documentum and Interwoven or by creating APIs that easily adapt to CMS.

Still, the challenge inherent in this confluence of translation and content creation – particularly emphasized by tools such as SDL AuthorAssistant, Sajan's Authoring Coach, and across' crossAuthor – has yet to be embraced by the language industry. These tools allow technical writers to connect to a TM so that the content will be written with the greatest possible number of matches in the translation process. This would seem to be an opportunity ripe for the picking, allowing language service providers (LSPs) to broaden their service portfolios to include authoring services by using the TMs they helped their clients generate in the first place.

Workflow modules

Workflow modules within translation environment tools were still an exciting and fresh concept when I last wrote about it, but it's now become an expected and required component. At the multi-language vendor (MLV) level, TM and terminology requirements now have to fit into a larger framework of process management. Tools such as Idiom WorldServer, Lionbridge's Logoport, across, and SDL Trados Synergy are prime examples of tools that connect all three of the traditional layers in the translation process: translators, MLVs and translation buyers.

On the other hand, other traditional translation environment tools – including Déjà Vu, Similis, MemoQ, MultiTrans, SDLX and STAR Transit – are offering an interface between the LSP version and the freely downloadable editor for translators/editors. "Light" interfaces between some project management tools and the analysis module of the leading translation environment tools have also become commonplace.

A new development that will certainly continue in the future is actual partnerships between translation environment tools and project management solution providers, such as the newly announced partnership between MultiCorpora and Plunet.

A kind of partnership that represents a new and promising phenomenon is one that provides translation environment tool vendors with MT components.

MT components

Operating on the same keyword, another kind of partnership that represents a new and promising phenomenon is one that provides translation environment tool vendors with MT components. Both across and Idiom have announced formal partnerships with MT provider Language Weaver, a trend that will most certainly continue with increased cooperation of that nature and ever more seamless integration of MT into the process offered by translation environment tools. This quest for cooperation is mutually beneficial and will continue to be pursued by both sides. It is in the interest of MT tools to have a strong TM (and terminology management) component, and it is equally in the interest of translation environment tools to offer an easy and seamless interface and integration to MT. An immediate drawback of this symbiosis will be that the language agnosticism that most translation environment tools practice will be replaced by a "preferred" treatment – at least as far as MT goes – of the larger languages for which MT is available.

Other less formal integration will also continue, such as has been offered for some time now by tools such as Wordfast or MetaTexis with easy connections to MT packages with a Word interface.

Exchange formats: XLIFF

XLIFF, the exchange format that was poised to make translatable content completely exchangeable between supporting tools, has made some slow progress. While a number of tools – including most if not all of the localization tools as well as Lingotek, SDLX, TRADOS and others – are supporting XLIFF at various levels, it is probably fair to say that it has not achieved the prominence that it was supposed to have and should have. There are some hopeful signs, though. Following Heartsome's and Sun's Open Language Tools XLIFF Translation Editor's lead, both Lingotek and MultiTrans (in its XLIFF Editor) have adapted XLIFF as the internal default format for all (Lingotek) or HTML-based and XML-based formats (MultiTrans).

In the meantime, two other formats have become a *de facto* exchange format: SDL Trados' bilingual RTF-based format and,

What's next for TMS?

Benjamin B. Sargent

Since the publication of the Common Sense Advisory report "Translation Management Systems Scorecards" in February 2007, many language vendors and enterprise users have voiced their opinions about what is missing in these software products. Removing the obviously blue-sky suggestions, what follows is a rundown of realistic feature additions and enhancements that we believe TMS vendors should be working on this year and next.

1. *Plug-ins for content management system (CMS) environments.* Content contributors and content managers want to order/select translation from within their own environment. They don't want to have to load or learn a new application. The natural evolution of the "content connector" that shuttles content to and from the translation environment is a plug-in interface. Translations.com has pioneered this approach with graphical user interfaces consisting of one or more screens that open within an Interwoven TeamSite, Documentum Web Publisher or other CMS application.

2. *Tie-ins to authoring environments.* Documentation managers need their authors to select existing segments with equivalent meanings when translated matches already exist within the TM database. Ease of use is paramount here since tech writers would be asked to deviate from their current practice: finish writing paragraph, apply TM, select matches, perform final cleanup, begin writing next paragraph. SDL AuthorAssistant and Sajan Authoring Coach are good starts, but every vendor needs to be working on these capabilities.

3. *Business management capabilities.* Vendors and buyers alike write purchase orders for translation work, based on the results of running new content against existing TMs. Push-button purchase order generation is a winner for system users, but few vendors include this feature. One reason is that named resources with known rate cards are needed, thus making resource management functions a necessary precursor. Translation companies generate their invoices based on similarly generated data. "Buyers" servicing corporate translation requirements also generate the internal equivalent of invoices to charge back outsourced costs to other profit and loss centers within their organization. Software makers cavil and say that so many other charges pertain – such as DTP, localization engineering, QA and project management. Users say, "Yeah? So get busy!" The users are screaming for it. Which vendors will prove the nimblest?

4. *Cross-integrations with other enterprise systems.* This is a long term requirement, but inevitable. Enterprise class business process management, resource management, financial management (general ledger and procurement) systems all need to talk to a mature TMS, if and when someone builds one. LTC Organiser already talks to SAP, QuickBooks and Crystal Reports. Expect to see more connectors to general ledger applications, as well as sales force automation tools such as salesforce.com and SugarCRM.

The core functions of TMSs are well defined. Vendors with workflow, centralized memory management functions, and webtop tools for translators and reviewers need to focus next on bridging the budget and resource management gap between their software and other enterprise systems of record. It is all part of growing up and learning to do your household chores. **M**

Benjamin B. Sargent is a senior analyst at the research and consulting firm Common Sense Advisory.

because of the limitations of that format, the XML-based TTX format. You are hard-pressed to find tools that do not support one or both of these formats or are about to support them. While these formats do not offer all the manageability benefits of XLIFF, they are in fact more of an exchange format than XLIFF is.

Another originally distinguishing factor between translation environment tools is slowly dissipating as well: the support of the many desktop publishing or word processing formats. Since virtually all translation environment tools support XML and since many source formats are now in some form of XML or can be represented as XML, the costly development of new filters for new formats – which was typically only done by the larger tools – is increasingly becoming obsolete. Examples of this include the latest versions of InDesign, Microsoft Office and, of course, OpenOffice.org.

Open source

The openness that XML provides also now allows open-source tools to directly support formats such as Word, Excel and PowerPoint (rather than through a prior conversion via OpenOffice). OmegaT, the most actively developed open-source tool, has announced the inclusion of a filter for Office 2007 in its next release, and that should help it gain an even stronger footing in the freelance translator community. In May of this year, ProZ.com astonishingly reported that OmegaT was the fourth-most-used tool among its members.

TM exchange

TM exchange has arguably made the largest strides in the past couple of years, both in the form of tool development and tool-external initiatives.

Many tools now offer components to exchange project-based TMs interactively during the translation process (TRADOS, SDLX, across, Fusion, MemoQ, Idiom Workbench, Logoport, MetaTexis, MultiTrans and various others), and Lingotek in particular has made the sharing of TMs ("indexes" in Lingotek-speech) one of the cornerstones of its tool architecture.

Wordfast started the VLTM (Very Large Translation Memory) project, in which large public TMs for various language pairs are made available on a central server for all Wordfast users, and recently this feature was extended by creating space for private projects on the server as well.

Aside from the widely accepted TMX format and its as-of-yet-rarely-used extension SRX (the Segmentation Rules eXchange format, which makes sure that the same segmentation rules are used across user settings and tools), another interesting new technology has been introduced by MemoQ called "translation memory driven segmentation." With this technology, the underlying TM dictates the segmentation to make sure that potential matches are segmented according to the TM.

Methods of sharing data

Other tool-independent initiatives also show that there is an ever-growing awareness of the need to share data. The Translation Automation User Society (TAUS) organized a summit in March of this year with representatives of 26 multinational organizations "to explore how a co-operative platform for sharing language data can potentially increase levels of translation automation, through, for instance, advanced leveraging and training

of machine translation systems" (quote from the press release of TAUS, www.translationautomation.com/downloads/TAUSSummitNewsRelease.pdf). While the participants of the summit recognized that there are definite hurdles to overcome (such as legal questions, classification of data, and infrastructure), the next meeting to tackle some of the hands-on questions is already scheduled for this fall.

Another initiative is the licensing scheme that TM Marketplace has been offering for a couple of years. Rather than giving data assets away to competitors, this concept looks at the value of the TM data that has been assembled over many years, puts a price tag on it and sells licenses for its use. For instance, General Motors is offering TM Marketplace licenses for more than four million of its translation segments in six language combinations (English to German, European and Mexican Spanish, Canadian French, Italian, and Dutch) to other vendors in the automotive industry, LSPs who are active in the automotive industry, and MT developers.

And there is a third initiative, the "made-for-order model," done without any direct participation of the original data owner. Huge amounts of bilingual data in the form of PDFs, web pages or various other formats are available for download on the internet. While these data sources are not prepared to be used as TMs, the emergence of industrial-strength alignment tools and expertise makes it possible to turn these documents into bilingual TM data. These TMs can be "custom-ordered" and tailored for specific industries or products. You can find a white paper on the legal ramifications at www.tmmarketplace.com/whitepapers/align

What it all means

What does this all mean in a nutshell? Here's how I would summarize the developments in the translation environment tool market along the lines of the above categories:

- Terminology management is making a gradual shift because of a stronger use of subsegments.
- Workflow solutions have become a required feature in translation environment tools, either through partnerships or tool-internal features.
- There is a clear trend toward integration of MT and TM technologies, both through partnerships and through functional integration.
- Data exchange between tools, especially on the level of translation files and TMs, has become a reality.
- TM access has been increasingly opened up to allow workgroups simultaneous data access.
- TM sharing is supported by various other tool-independent initiatives.

So far, so good. Because of the proactive approach of many tool developers, tools are adapting to the market relatively quickly, or, as in the case of the handling of subsegments, they are encouraging the market to adapt to them. Increased competition – rather than decreased, as many had predicted in 2005 – and an increasingly level playing field through the openness of formats and the possibilities of data exchange are helping this market to be ever more flexible as it steps up the process of development. And the quest for partnerships and integration of tools on various levels is helping the "environment" of translation environment tools to become ever larger. Tools such as the "middleware" Clay Tablet (www.clay-tablet.com) could possibly further this process even more.

The Translation Summit (www.translationsummit.org) has initiated a task force that is supposed to help further reduce the gap between what translation users at all levels need and what tools provide. Though this task force is not limited to "translation environment tools" but looks at all the different tool-based processes of the translation workflow, translation environment tool developers should be able to benefit immensely from the results. This study is going to be launched with a comprehensive survey about which you're sure to hear more in the pages of this magazine and other forums.

To paraphrase the closing of the state of the union address delivered by Bill Clinton in January 2000: "As long as our dreams outweigh our memories, we will be forever young. That is our destiny. And this is our moment." **M**