# The Road toward Collaborative Translation Memories

*By Yves Champollion*

Looking back at the earliest translation memory (TM) tools from the late 1980s, it is evident that our expectations of TM software have evolved a great deal as these commercially available products have become more advanced. The first tentative efforts to create supporting software tools for translators saw memory systems that stored completely aligned source and target sentences in extensive databases, from which they could be recalled only when a complete (or perfect) match was discovered. The problem with this approach was that there was no guarantee that the new source-language sentence was from the same context as the original database sentence. Such tools naturally did little to automate some of the processes we now take for granted in our TM weaponry, as the translator still had to spend time reviewing all the matches for relevance and accuracy in the context of the translated document. Although cheaper than outright translation, this review still carried a cost. Therefore, the goal of reducing overhead (and hence translation costs) was not fully realized at that time. However, the benefits of increased consistency in phrasing and terminology were clearly evident, and the development of TM tools was regarded as a worthwhile pursuit.

First-generation TM tools were most useful in translation domains where the occurrence of perfect matches was common, such as technical documents. Unlike more creative texts, technical documents are often comprised of a series of exact component phrases and terms ("units") that are usually repeated throughout the text. Early TM tools had no trouble managing these significant text units.

As readers, we would be outraged if it were suggested that all texts were not "written," but merely constructed by ordering set units and pre-packaged blocks of meaning around to create an end product. This completely negates the idea of an original author whose word choices and painstakingly crafted sentences best convey his or her message. The advent of TMs essentially had the same effect on the work of the translator. Where once the translator's vocation was considered akin to that of a

The highly repetitive nature of some technical texts and manuals meant that early translation memory tools, although crude, were useful in certain translation domains.

wordsmith—carefully re-crafting the message of a text in his or her chosen language based on linguistic knowledge and academic prowess—the automated processes of TM tools seemed to devalue that role. The elegant art of the translator was somewhat reduced to the role of a typist, chained to a computer and paid by the mile for simply reassembling pre-translated text.

> By incorporating stylistic elements and offering increased flexibility, it is evident that translation memory tools are now much more intelligent than their predecessors.

This perception was due in part to a lack of sophistication in the tools themselves. Even TM tools that did start to support fuzzy, or nonexact, matches (initial examples include IBM's Translation Manager and early versions of Trados) were merely set up to offer purely statistical matches in unwieldy blocks, and there was little option but to reuse these blindly as they were presented. Thus, multiple-clause sentences proved too complex and awkward for TMs to deal with efficiently, resulting in inconsistencies in segmentation. Phrases were chopped and glued back together without any option of including a "feel" for the overall meaning of a text. The tools did not present the translator with various alternatives based on context.

## The Present: "Smarter" TMs

It was the lack of satisfaction with this statistical methodology that drove the development of the more advanced TM tools we see today. Powered by robust algorithms, programs became more intelligent and acquired the ability to distinguish inexact or fuzzy matches, as well as the ability to grade the level of suitability of the match for the translated text on a continuous scale from 0 to 100%. Tools now include a linguistic analysis engine, use chunk technology to break down segments into intelligent terminological groups, and automatically generate specific glossaries. While benefiting translators as a whole, the advantages of this development were perhaps most keenly felt in the translation of creative texts, where the likelihood of exact text repetition was diminished. This was an area in which TMs had previously only been of limited value. With these new match functions, translators were again empowered to make creative choices based on the more detailed data at their disposal. Previous translations could be reviewed along with an assessment of their suitability for the source text, and then be edited according to the new context. In this way, TMs began to approach the more flexible model that was required in order for them to fulfill the translation support role. Improved consistency was assured, but this improved flexibility meant that the tools were less dogmatic than previous versions. As a result, the role of the translator was revalued.

Today, translators are offered even more flexibility in their TM arsenal. Second-generation TM engines incorporate a host of features that go much further toward accommodating contextual influences. The most significant of these is that modern TMs now accept multiple translations of the same source. For reoccurring source text, current software offers up various options that relate to the original, with the preferred translation prioritized based on an automated assessment of the context. The program effectively "reads" each segment in context in the same way a translator would, thereby helping to solve previous issues related to segmentation, which, in turn, results in a more pleasing final product. The highly complex linguistic algorithms offer the best possible semantic match, rather than the simple statistical matches of yesteryear. This represents a key step in the quest for a linguistic tool that offers the level of flexibility required to handle a wide range of texts.

By incorporating stylistic elements and offering increased flexibility, it is evident that TMs are now much more intelligent than their predecessors. Yet, it is important for us to examine where this development is leading.

## The Future: Assets for All?

Now that TM tools incorporate a wide range of functions to support translators, it seems natural that future developments should focus on the shared use of these TM assets. Currently, even though an individual translator is likely to build up hundreds of thousands of words of TM each year, the actual memories themselves are typically seen as the property of the client for whom the project is being carried out. The buyers are the initiators of the source text, and so intellectual property rights are exercised over the TM that is generated. The buyer is purchasing a translation, but also a host of TM assets for use on other projects. The current attitude toward TM files is one of property and ownership, but the Internet's tendency toward knowledge sharing modes may be changing that approach.

Server-side TMs are already gaining popularity with large corporations that seek to improve consistency in enterprise-wide localization initiatives. Such platforms allow multiple translators or teams of translators to share their TM assets with key client-side stakeholders in real-time, thus producing a continuously updated set of terminology and phrasing that aligns with a company's brand voice. All translators can then ensure that the stock verbiage they use is drawn from a central TM repository that has been established and refined by other translators on the team and approved by the relevant client contacts. Here we hit on a key point for future TM models—collaboration.

Wiki-based knowledge sharing and peer-to-peer networking tools represent important and exciting developments that now form an integral part of much Internet-based research. The ability to network with others in an open environment and work together toward an end goal of improvement and refinement has shaped many Web-based activities. What could this tendency toward collaboration mean for the translation industry?

Few translators would deny the value and increasing importance of free-for-all repositories such as the Europa terminology portal. Unlimited access to many thousands of terms and phrases, which have been standardized and approved across 23 official languages, means that any team conducting a translation project for the European Union has a head start in its efforts to maintain consistency across geographically diverse locations. Such developments constitute a boon to both clients and translators seeking consistent texts in legal or technical domains, as well as a firm and consistent corporate voice.

The European Union stands out as a key example, but what if this practice were to become standard across a wide range of global and international bodies? If the United Nations, World Bank, World Trade Organization, and others were to engage in the same practice of releasing approved terms into the public domain, we would see the creation of a huge online terminology repository that is both open and collaborative. If continued over the coming years, this practice could help foster more streamlined communication between worldwide organizations, and we could even see this extend to the private sphere.

My own interest in this area led to the development of the Very Large Translation Memory (VLTM) project, which proposed an initial method for making blocks of TM assets available for free to online communities. Subsequent research leans toward the creation of a Web-based tool that can generate valuable TM assets automatically. As TM assets are heading toward an open model, the functionality of such a tool would need to reflect this. The next significant phase in the evolution of TM will see tools that can seek out multilingual websites in any domain chosen by the user. The tools will be able to harvest content from any of these sites, thus generating more TM options for the user. An online "TM harvesting" tool of this kind could potentially revolutionize the way translators approach translation projects.

Working within this model, TM assets would no longer be thought of as items to be hoarded, bought, sold, and traded, but more as public commodities to be shared and refined over time. The overall attitude would switch from exclusive to inclusive, and all participants would be working to free up TM content and contribute to a shared intellectual heritage. Of course, it is no coincidence that the majority of terminology released so far has stemmed from governmental institutions such as the European Union. It is worth noting that there is likely to be a certain amount of opposition to this free-for-all methodology in the private sphere. For instance, it is hard to imagine Microsoft suddenly deciding that all terminology relating to its software applications—which has been painstakingly researched, established, and refined over time in multiple languages—should be released into the public domain. Global corporate identities are largely drawn from approved multilingual content, and are not intended to be simply "harvested" at the touch of a button. That being said, certain forward-thinking corporations may see the advantage of providing terminology for all to use as part of a corporate citizenship initiative. Microsoft remains a pertinent case study here, as sharing terminology would be particularly applicable for firms that are innovators in highly technical or specialized industries.

As examined over the course of this article, it is evident that the role of translators is reflected to a certain extent by the tools they choose to use to complete projects. Though early tools were seen as reducing the value of human input, these proposed collaborative models perform precisely the opposite function. Collaborative efforts mean that all are contributing to a communal intellectual database, which does not comprise "property" as such. In this atmosphere, the value of individual translators can be seen as collective, as they are truly part of a united global community that is committed to furthering cross-cultural communications worldwide.