[*Translating and the Computer 21*. Proceedings… 10-11 November 1999 (London: Aslib)]

Translation Technology - The Next Generation

Sharon O'Brien, Language Technology Consultant
ALPNET Inc
sharonob@ie.alpnet.com

Summary

This paper addresses the topic of translation technology past, present and future. The historic time-span covered (1995-1999) might seem short, but one has to remember that one calendar year amounts to several years in the life-span of any technology.

The use of translation technology in the last four years, specifically translation memory and terminology management tools in the IT and localisation sectors, is considered. Original expectations and subsequent realities are explored.

We will then turn our thoughts to the (near) future to consider what translation technology will bring to the translation industry in the new millennium. The focus will specifically be on translation management technology, multi-lingual information management and technology currently being developed to support the entire Infocycle (the ALPNET-Sun Microsystems joint "TMCi" project).

Translation Technology 1995-1999

Prior to 1995

Prior to 1995, there was no widespread implementation of translation technology. Several translation service companies, like ALPNET for example, were using first generation translation memory and terminology management tools in-house but this was rather the exception than the rule and these tools were not widely distributed across the freelance and contractor base. Some industrial companies had full or pilot implementations of machine translation systems but, in the IT sector at least, machine translation was not considered a viable solution by most companies.

The use of translation memory and terminology management tools prior to 1995 was primarily for the benefit of the translation supplier. They were used as a means to gain competitive advantage through increased productivity. Apart from savings in time, benefits were not shared openly with Clients.

Some forward-thinking companies were examining the potential of translation technology at this time. They acquired beta versions of tools that were commercially available, evaluated their potential and carried out competitive analyses. The general mood towards translation technology at this time was one of curiosity. Optimists believed the tools had potential and could lead to a competitive advantage. Pessimists held that the primary end users, the translators, would never accept these tools.

1995-1997

The period between 1995 and 1997 saw a rapid subscription to, and implementation of, the translation memory and terminology management tools on offer, especially by commercial translation suppliers. This required significant investment in license fees, training and implementation, not to mention increased project administration.

While translation suppliers were busy selling the notion of translation memory to their customers, in order to differentiate themselves from the competitor, great battles ensued with the end users, the translators. It is a fact that most translators working in the IT and localisation sectors have a good understanding of computers and software. Nevertheless (or because of this?), they did not welcome translation tools with open arms. The reasons were manifold and not without foundation:

- Single user licences were expensive. Freelancers were being encouraged to "invest" in licences themselves, something they were not prepared to do.
- Many translators argued that translation memory tools would reduce the overall word count they translated in a year. Therefore, it seemed they were being forced to fund their own financial downfall.
- Translators considered the tools on offer to be unstable and felt that their use would slow down throughput instead of speeding it up.

The counter-arguments from translation suppliers were equally strong:

- It was only a matter of time before these tools became a *pre-requisite* as opposed to *desirable.* The development was frequently compared to that of moving from the dictaphone or type-writer to the word processor.
- The tools provided a mechanism for translators to avoid boring repetitive work and to concentrate instead on the more creative process of translating "new" text.
- Translation volumes were growing at such a rate that translators could look forward to processing more words annually using the tools than they had previously.

Meanwhile, the reaction from the purchasers of translation services was ambivalent. Greater problems of quality, time and cost pre-occupied their minds. Many did not want to consider translation tools as an answer to their woes. On the other hand, there were some companies who welcomed the development and supported it, especially when they began to realise that their costs could be reduced and their time-to-market increased by using these tools. Besides, most, if not all, of the investment was undertaken by the translation suppliers. The only price to be paid by the Customer was that of the risk involved in allowing your files to be processed by a relatively new, and as yet unproven, technology!

1997-1999

1997-1999 was a period of maturation for translation tools in the IT sector. Translators finally accepted the use of the tools. Customers began requesting their use and even purchased some licences of their own to use internally.

More translation memory and terminology management tools were launched onto the market to compete with the ones already there. The tools were demonstrated and discussed in workshops and at industry events. An increase in the stability and functionality of the tools was apparent. For example, support was provided for additional languages (especially Asian languages) and new file formats.

This period also saw a renewed interest in Machine Translation. This development was the result of many influences:

- Many MT systems were ported to the PC and were therefore more accessible to a greater number of users.
- MT gained prominence through the World Wide Web (e.g. the Babelfish site powered by Systran).
- Many translation memory tools supported MT (e.g. Trados Translator's Workbench with Systran and Logos and Star Transit with Logos).

The renewed interest led to pilot and evaluation projects focussing particularly on the combination of translation memory and machine translation.

With more and more tools on offer, compatibility became an obstacle. If one translation supplier used one TM tool and another used a different one, customers inherited translation memories which were incompatible. This crisis led to the setting up of a special interest group within "LISA" (the Localisation Standards Association) called "OSCAR". The OSCAR group developed a standard for the exchange of translation memories, which most of the commercial developers agreed to implement.

When it came to terminology management, the compatibility problem was not so prevalent because most terminology management tools used in this sector could export and import terminological entries to or from a tab- or comma-delimited file. (That's not to say that exchange of terminology is trivial, especially in the context of exchanging terms between different MT systems, or between MT systems and term management tools or indeed between different term banks.)

The future was starting to look very promising for translation tools at this point, until the purchasers of translation services started to analyse the ROI (Return on Investment) figures they had been promised in the early stages. The analysis was not always positive. It became apparent that the benefits promised earlier were not being fully achieved. For example, while the word rates were reduced for sentences that matched exactly or closely with previously translated sentences ("exact" and "fuzzy" matches), the overall cost of translation had not dropped significantly. Also, the time required for translation had not been reduced significantly. There were several, complicated reasons for this:

1. Time

When TM tools are used for the first time, there is no "memory". The memory has to be built either over time, with the translator adding sentences as s/he works, or automatically using an "alignment" tool to create a memory from legacy material.

The first option, where the translator builds the memory over time, does not provide any payback until there has been one or more updates to the original files.

The second option of alignment provides a faster route to building the memory. Unfortunately, texts can sometimes not be as "parallel" as might be believed. This, along with inadequacies in alignment technology, leads to less than 100% accurate memories and can sometimes lead to more work for the translator who has to clean up the memory as s/he works.

2. Updates

Updates of the source files are common in the IT industry. The documentation team works according to deadlines and rarely waits until a piece of software is "frozen" or complete before sending the documentation for translation.

Before translation tools, updates were handled in a cumbersome manner. The previous version was "compared", often automatically, with the new version. Differences were marked in the electronic version of the document and the translator had to wade his or her way through a multi-coloured maze of underline and cross-out in order to decipher the differences and update the translation.

Handling updates was supposed to be the strength of translation tools. It proved to be both a strength and a weakness. The method for handling updates was as follows: The newly revised file was automatically compared against the translation memory. The TM tool calculated how many sentences were new, the same or had changed somewhat. The TM tool could even automatically insert exact and fuzzy matches into the revised file. It sounds like an improvement on the former "cut n' paste" approach. However, there were several problems:

- Translators frequently worked remotely and before all files could be compared against the translation memory, all translation memories had to be collated and merged, which in turn frequently led to superior translations overwriting inferior translations for the same sentence.
- The cost of translation for the revised file was calculated on the basis of the number of new or changed words in the old method. In the new TM method, the entire file was put through the translation memory process again. And, because the contents of the translation memory were not always dependable, the translators demanded that they should be paid for reviewing *every* sentence again, not just the sentences that were new or had changed.

These inadequacies led to a situation where it was often more costly to do an update using Translation Memory tools than it was to do an update using the traditional cut n' paste method.

3. New processing tasks

Translation Memory and Machine Translation tools must support many different file formats. The most common formats used in the IT sector are Word (Doc and RTF format), HTML, FrameMaker, Interleaf, SGML, PageMaker and QuarkXPress. TM and MT tools support these very different file formats by converting them from the native format to a format which is "comprehensible" to the tool itself.

With the introduction of TM and MT tools, a new set of tasks was introduced. These included:

- File set-up (ensuring the files were set up optimally for the converter)
- File conversion (from the native format to the tool format)
- File conversion (from the tool format back to the native format)

- Integrity checking (making sure that the converter had not damaged or interfered with the functionality of the native file format)

Each of these tasks require human input, which in turn means additional time and cost when compared to the traditional method of translating files in their native file formats.

4. Stability

Early versions of some translation tools were not altogether stable. There were frequent "crashes", especially when the tools were used by groups of translators over a network. These crashes resulted in a delay in the translation work and sometimes even corrupted the translation memories and files being translated, which added time and cost to the project.

5. Compatibility

As mentioned previously, many translation departments and companies used different translation tools, which led to compatibility issues. Sometimes translation memories and glossaries had to be converted from one format to another before a project could even begin. This process was not always a foolproof one. Data was frequently lost and, again, time and cost were added to the project.

6. User framework

The optimum setting for the use of Translation Memory and Terminology management tools is when data can be shared by a group of users in a real-time manner. Many TM and terminology tools have this capability. However, the nature of the translation business does not fit well with the capabilities of the tools. Translators frequently work on a freelance basis from their home offices. In this environment, it is not possible to share a translation memory or glossary real-time with a group of other translators who are working on the same project.

This presents two disadvantages:

- Translators cannot benefit from each other's work on a day-to-day basis and consistency is hampered.
- At the end of any project, the project manager inherits multiple translation memories from each remote translator. These TMs frequently contain duplicate source sentences which have been translated differently by each translator. The TMs must be "merged" together, but it is very difficult to control which sentences should be overwritten and which should remain in the TM. Again, this effects quality and consistency and adds time and cost to the project.

7. Translation Process

Translated documentation, which is bound for publication, must look at least as good as the parallel source document. That's the reason why, at the end of the translation process, there is a DTP cycle ("Desk Top Publishing") for many file formats.

Frequently, a linguistic review also happens around this time in the process. Between the two events, linguistic changes to the final translation are inevitable.

The DTP process requires that the file format is converted from the translation tool format to the native format. So, any changes implemented in the file at this stage *are not reflected in the translation memory.* Ultimately, this leaves the user with two choices:

- Endure the additional cost of implementing the changes in two locations, i.e. in the translation itself (native format) and in the translation memory.
- Do not implement the changes in two locations and risk quality and consistency problems the next time that translation memory is used.

8. Infocycle

Put simply, the "infocycle" represents the entire process of content creation, translation and publication. The basis for comparison in any TM system is the source sentence. Additionally, the unit processed by machine translation and terminology management tools is the source sentence and term. However, those responsible for source creation and those responsible for translation rarely talk, never mind meet, to discuss how they could improve the entire process or reduce time and cost.

Those who hold the highest expectations for translation memory tools are those who have been given the responsibility by their companies for finding more efficient ways of producing multi-lingual information. Frequently, they are disappointed by the "leverage figures" (the statistics on re-use) obtained from translation memory tools. These disappointing figures are caused by many of the factors listed above. In addition, lack of consolidation of the source creation and translation processes also plays a significant role. Authors frequently make "minor" changes to new versions of documents. Expectations of a 90% leverage from translation memory often come in at a very disappointing 30%. Most of the loss of leveraging is a direct result of the "minor" changes in the source files!

Translation Technology 2000 and beyond

At the end of 1999, facing into 2000, the situation is one of grudging acceptance. Translation Tools, in particular Translation Memory and Terminology Management tools, are accepted as defaults in the day to day production of multi-lingual information but many are disappointed with the limitation of the tools and their seeming inability to meet the ROI promised in the days when these tools were first being sold.

So, what is the outlook like for 2000 and beyond?

Widening the focus - the "Infocycle"

As explained earlier, the "Infocycle" represents all the steps involved in the production and publication of multi-lingual information. In order to meet the expectations of higher quality information, faster and cheaper, the focus of tools must be expanded to encompass the entire Infocycle instead of just concentrating on translation memory or terminology management or machine translation.

Multi-lingual Information Management Technology

The term "Multi-lingual Information Management Technology" will become more widespread and will encompass the technologies we have spoken about already, i.e. Terminology Management, Machine Translation, Translation Memory. However, the term will not be limited to these technologies. It will also cover tools for Authoring, Text Indexing and Summarisation, Content Management, Controlled Language and Multi-Modal Publishing (i.e. publishing to the web, to CD and to paper).

TMCi - An example

Development of a multi-lingual information management suite of tools is already underway. The development project is code-named "TMCi" (Translation Management Centre Infrastructure). The TMCi project is a joint undertaking between ALPNET and Sun Microsystems. The starting point for the base technology was the "EPTAS" Client/Server based translation support system, which ALPNET acquired during 1999. (See Waldhör, K., "EPTAS - A Client/Server based Translation Support System", in Proceedings of Translating and the Computer 20, 1998).

TMCi is founded on the principles of open architecture and platform independence. Its objective is to produce a suite of tools for multi-lingual information management, which reduces the cost and time associated with producing multi-lingual information.

As mentioned, the core technology is a Client/Server translation management suite which allows the user to "leverage" translated text from *multiple* translation memories organised hierarchically. The "Client" is a browser-based tool, which provides access to the server applications from anywhere in the world. For example, a terminologist can submit a request to the server to perform terminology mining on a set of files. Or a project manager could request that a number of files be compared to a range of translation memories in order to assess the level of re-use that can be obtained from those memories.

Results from these kinds of activities will be processed by terminologists and translators using the Java-based Translation Editor and Terminology Management tool. The Translation Editor will be able to access translation databases on the server in "real-time", to update those databases and, indeed, to download the most current relevant translations. The Translation Editor will also support the project management cycle by submitting regular status reports to a project manager on the number of words translated or edited to date.

TMCi is already integrated into two well-known commercial MT systems. These MT systems will provide additional support for the translation cycle. In addition to these systems, part of the TMCi project involves research and further development of a prototype "Example-Based" MT (EBMT) system. The prototype MT system will focus on the German-English language pair. Tools which already exist today for the automatic creation of translation databases and for the extraction of core terminology from source files will be put to use to build the corpus and thesaurus for the EBMT system.

Apart from supporting translation and terminology activities, TMCi encompasses other aspects of the Infocycle.

Firstly, a tool will be developed to support the authoring process and ensure the highest quality source text. This tool will check the source for errors in spelling and general grammar and for adherence to customer-approved abbreviations and terminology. In addition, it will check the style of the source text and make comparisons with the customer-approved style guide. Reports will be generated and sent to the author along with recommendations for changes.

Secondly, TMCi will interface with an SGML-based multi-lingual content management tool. The content management tool first has to be selected. Following selection, the interfacing task will begin with the objective of reusing "information elements" from the content management tool before the source files are submitted to the translation databases and machine translation system for processing on the sentence level.

Finally, TMCi will support the information management cycle by feeding information and reports into a web-based information management system ("IMS"), which can then be queried by product managers, project managers and so on. For example, the leveraging results obtained by comparing a set of source files to a number of translation databases will be sent to the IMS automatically for storage with all other project information. As previously mentioned, the Translation Editor will have report-generating capabilities. These reports will also be automatically fed to the IMS.

Development for TMCi has been ongoing since August 1999. The core system is already available. By end of Q2, 2000, the Java-based Translation and Terminology Editor, the browser-based Client and the integration with the IMS will be complete. The end of the year 2000 will see the completion of the authoring support tool and the EBMT prototype as well as the integration with a commercial content management tool.

Conclusion

Although the period between 1995 and 1999 is a short one, a lot has happened in the domain of translation technology. Translation Memory tools, Terminology Management tools and Machine Translation have grown in popularity and use and, with the exception of MT, have become prerequisites in the day-to-day translation process.

The period between 1995 and 1999 has been long enough for most users to realise that, as they exist today, translation tools have reached their maximum potential and the ROI is somewhat disappointing when compared with original expectations.

To increase their potential and help realise a better ROI, integration with other technologies will be necessary and the focus will have to expand to encompass the entire Infocycle, including content creation and publishing.

References:

O'Brien, S., "Practical Experience of Computer-Aided Translation Tools in the Software Localisation Industry", in Bowker, L., Cronin, M., Kenny, D., Pearson, J. "Unity in Diversity: Current Trends in Translation Studies", St. Jerome Publishing, 1998.

O'Brien, S., "Translation Memory as a Linguistic Resource in the Localisation Industry", The ELRA Newsletter, Vol. 4, no. 2, April-June 1999.

Heyn, M., "Present and Future Needs in the CAT World", LISA Newsletter Volume V, No 3, pp. 15-33, Localisation Industry Standards Association (LISA), September, 1996

Waldhor, K., "EPTAS - A Client/Server based Translation Support System", in Proceedings of Translating and the Computer 20, Aslib, 1998

OSCAR, (1999): html://www.lisa.unige.ch/tmx/index.html