Evaluating Translation Memory Systems

Angelika Zerfass

Freelance Translation Tools Consultant
Holzemer Str. 38
53343 Wachtberg
Germany
azerfass@debitel.net

Abstract

Since the mid 1980s, translation tools have taken over more and more of the daily lives of translators and translation project managers. But a lot of time now has to be spent on evaluation, training and administrative tasks.
Translation tools were designed to make the translator's work easier, faster and more efficient. They range from conversion utilities to terminology management, translation memories, machine translation as well as workflow and project management systems.
They were developed with the aim to reduce repetitive translation work, but on the other hand they add different tasks to the workload, like administrating databases and the like.
This presentation will give an overview of one area of translation tools - the different translation memory systems on the market today and the technologies they use. It includes a comparison of common basic features like word count, analysis/statistics function and pre-translation, some tools' specialities as well as the description of data exchange possibilities between the systems by use of the TMX format.
As there is no "one best tool" for everything, the aim of this workshop is not, to recommend one tool, but to provide some guidelines for evaluating translation memory systems according to individual requirements.

1. Translation Memory Tools - Overview

Translation memory systems, as the name implicates, "memorise" the translations made by a human translator. Most translation memory systems (often also called "TM-systems"), consist of a database that stores the original text along with its translation - a database of segment pairs.

"Segment" here indicates that the units that is being translated and stored to the database can range from a single word (for example a heading or an item in a bulleted list) to phrases, complete sentences or even whole paragraphs. The tools recognise a segment by a set of internal rules that define, for example, that a segment ends with a full stop or a paragraph mark.

During translation itself, the tool will automatically look up every new source language segment to be translated in that bilingual translation memory. If the same segment is found in the database, the system will offer the translation that was saved with this segment as a suggestion to the translator for reuse. If it does not find the very same segment, it will start looking for similar segments. These are the so-called "fuzzy" matches, as the source language segments (in the document and in the database) only match to a certain percentage. When the translator gets such a fuzzy match from the database, they can decide if and how much of it can be reused for the current translation. Usually the translator can even set the level of "fuzziness", that is the percentage of similarity, so that the system will only offer translations that can be reused without having to make too many changes to the suggested translation.

Thus the use of a translation memory system can increase consistency and it cuts the time for writing a translation. This is especially true for the translation of repetitive documents like technical documentation, manuals, instructions and updates of already translated material.

Translation memory tools are usually the main component of a tools' suite. These suits also offer recycling tools, so called alignment systems. These are used to prepare translations made without translation memory systems for reuse in such a translation memory tool. They read in the source and the target language files, display them in parallel and propose connections of the source language segments to the corresponding target language segments. A translator will then review these connections. Then, the segment pairs can be imported into the translation memory. From now on they can be used just as if they had been translated interactively with the system itself. Another component of such a tools' suite is the terminology management system - another database that stores single terms (or phrases) together with their translation(s) into the target language(s). The translation memory database and the terminology database work together during translation. The translator will not only get suggestions for the translation of whole segments but also a list of all the terms within that segment that were found in the terminology database. Other components of such a tools' suite could be workflow or project management systems as well as filters and utilities for file format conversion.

Translation memory systems also start to be customisable for use with document or content management systems and some are even programmable via an API (application programming interface - programming commands that enable the user to call the translation memory system from other applications).

2. Translation Memory Tools Basic Principle

Basically all translation memory tools were developed with the same goal in mind: Something that has been translated before should not have to be translated again from scratch. It should come out of the database or reference material so that the translator only has to decide whether the previous translation can be reused or needs to be modified.

The technologies used to achieve this are different. Some tools use a model of referencing the files of a previous project, The referencing model uses those previously translated files (original source language files and translated files) as the source for suggestions of new translations. This model works especially well for projects with many updates containing a lot of small changes.

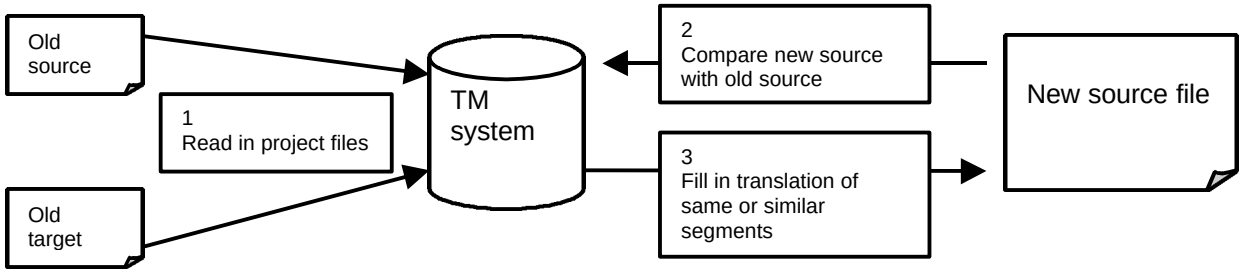

Figure 1. Reference Model

The database model on the other hand stores all translations ever made in one database, independent of context, which is useful if the same or similar segments appear in different projects and document types. Most of the commonly used translation memory systems are able to work with any language installed on the user's machine and they usually also allow the user to add project or user specific information to each translation.

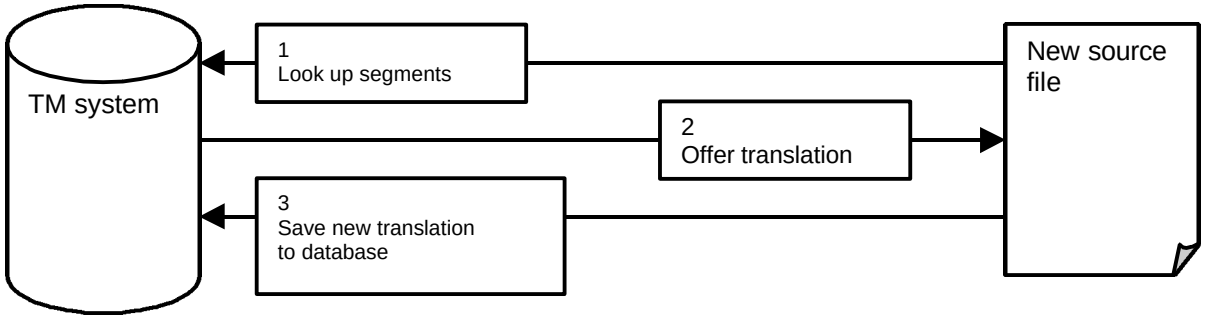

Figure 2. Database Model

3. Translating with Translation Memory Tools

The text to be translated consists of smaller units like headings, sentences, list items, index entries and so on. These text components are called "segments". Translation memory systems are equipped with a set of rules, which enables them to recognise, where a segment starts and where it ends. When translating with a translation memory system, the system goes through the text segment by segment, offering each of them to the translator together with any translation for this or a similar segment that has been stored in the database or can be found in the reference material. The translator decides whether to reuse the proposed translation, to adapt it or to create a new translation and then saves it to the system. Thus the translator builds up a store of segment pairs that can be referenced for future translation.

This store of segment pairs can also be used for analysing new files to determine the rate of recycling that can be achieved. Or it can be used to run a pre-translation, which creates files that contain segments with more or less matching translations already in them. This is very useful when working on a large batch of files or preparing files for other translators who are not working with a translation memory tool.

To be able to use translation memory tools on different file formats, from common Word files to DTP (desktop publishing) files, for example FrameMaker or Interleaf or files for the web in HTML, XML or SGML, some of these formats need to be converted to a format that the translation memory tool can work with. This happens either by use of separate conversion tools or filters integrated into the translation memory systems. Selecting a TM system therefore also depends on what file formats have to be worked on and how much time and effort needs to be spent on preparing and converting them to a usable format for translation and back to the original format afterwards.

Also, when it comes to software localisation for example, different tools have to be used for different parts of the project. The project might consist of text within the software from the user interface (GUI) to dialogs and messages as well as online-help files, documentation, packaging and marketing material and so on. And here different types of text require the use of different tools. GUI, software dialogs and messages are best translated with a software localisation tool, that is a translation memory tool that can read those special software file formats. They usually also contain testing features to check for consistent use of hot keys for example, or length related problems that might arise, if the translated text does not fit the button space it is supposed to appear on. But those systems are mostly specialised on the software itself.

For translation of the documentation, another translation memory tool is needed. And here the question arises how those tools for translating software and documentation interact, because what has been translated for one part might also be reusable in the other (this will be covered in the section about data exchange further down).

Online-Help files for example, could be translated with either a software localisation tool or with a translation memory system for documentation as both system types support this format.

4. Feature Comparison

All translation memory tools offer basic functionalities like word count or an analysis of recycling potential (how many of the segments in the file to be translated are present in the database or reference material as 100% matches or as similar, fuzzy matches). They also provide features for automatic pre-translation, search functionalities within the segment database, as well as access to terminology management components during translation. But each and every tool also has its specialities. These are the features that can influence the choice of tools.

Most translation memory systems read the files to be translated into the system itself, converting them into a table where one column contains the source language segments and another column that will be filled with the respective translation. Others connect to Microsoft Word so that any file that can be opened in Word does not have to be converted before translation and can be worked on in a WYSIWYG (what you see is what you get) mode. The translators can work in an environment that they are used to. Other file formats, for example DTP formats or so called tagged file formats like XML, HTML or SGML, are either converted or displayed in a separate editor. Colours are used to mark text to be translated as well as tags that make up the structure and formatting of the file.

More and more developers are enhancing the functionalities of the translation memory tools by adding new features like context sensitive pre-translation or machine translation-like components (for segments that have no match from the translation memory) as well as project management components.

5. Data Exchange between Translation Memory Systems

For some time, translators did not have the possibility to bring the data from one translation memory system into another system for reuse. A situation that was alleviated to some extent by the tools manufacturers by adding export functionalities for some proprietary formats of other manufacturers. But it was not feasible for each tool to support all export/import formats of all other tools - especially with new tools being developed and marketed all the time.

Now, the tool manufacturers have agreed to use one standard format for representing the data in their systems or at least to offer this format as one of the export formats. This allows an easier transfer of translation memory data from one system to another - even though the results are not always completely satisfactory. This standard is called TMX - Translation Memory Exchange format. It is an XML based representation of the data stored in a translation memory system.

5.1. Example of data representation in TMX format:

Segment pair:

This is a test. (English segment)
Dies ist ein Test. (German segment)

TMX representation:

```
<tu>
    <tuv lang="EN_US">
    <seg>This is a test.</seg>
    </tuv>

    <tuv lang="DE_DE">
```

```
        <seg>Dies ist ein Test.</seg>
        </tuv>
</tu>
```

Each segment pair is represented with a <tu> and </tu> tag that denote beginning and end of the segment pair. ("tu" stands for "translation unit", as those segments pairs are often called.) Then come the individual languages of the segments and the textual contents. This format could be produced and read by any translation memory system that works with TMX.

There are three levels of TMX compliance today. The first level only represents the text itself. The second level is able to represent the formatting information as well. And level three would be used to represent additional tool specific data like user IDs, project names and everything else the user has specified. Today, most tools comply at least to TMX level 1 or even to level 2.

6. Conclusion

Before investing in any translation tool, it is necessary to list the individual user requirements. This includes the file types that are to be translated. As most translation memory tools rely on structural and formatting information in the file, to segment and display the text, it should be tested if the way the files for translation are constructed, work well with this or that translation memory system. It could even mean to adapt the way of writing the documents in the first place, so that, at the translation stage, the tools that are used can handle the files more easily.

Another point is the networkability and the list of supported languages as well as the different supported file types.

Pricing for licenses, training and support should also be taken into consideration.

Then the tools should be tested for some time with real life examples to be able to evaluate, which tools answer the user's requirements best. Most tool manufacturers offer a trial period of about 30 days or a limited demo version of the software or, in case a longer evaluation period is needed, an extended trial with the full version of the software. This usually includes the need to buy a training session as well, to prepare the people who will be evaluating the software in the best possible way.

7. References

Some download sites for demo versions of translation memory tools:

- Trados - Translator's Workbench
  www.trados.com/products/download.htm

- Atril - Déjà Vù
  www.atril.com

- SDL - SDLX
  www.sdlintl.com

- Cypresoft - TransSuite2000
  www.cypresoft.com
  (supports only European languages)

- Star - Transit
  no download, contact Star for a demo CD at www.star-group.net

- Champollion - Wordfast
  Freeware
  www.geocities.com/wordfast/cat.htm

Some download sites for demo versions of software localisation tools:

- Pass Engineering - Passolo
  www.passolo.com

- Alchemy - Catalyst
  www.alchemysoftware.ie/demo4/

More information on TMX:
www.lisa.org/tmx