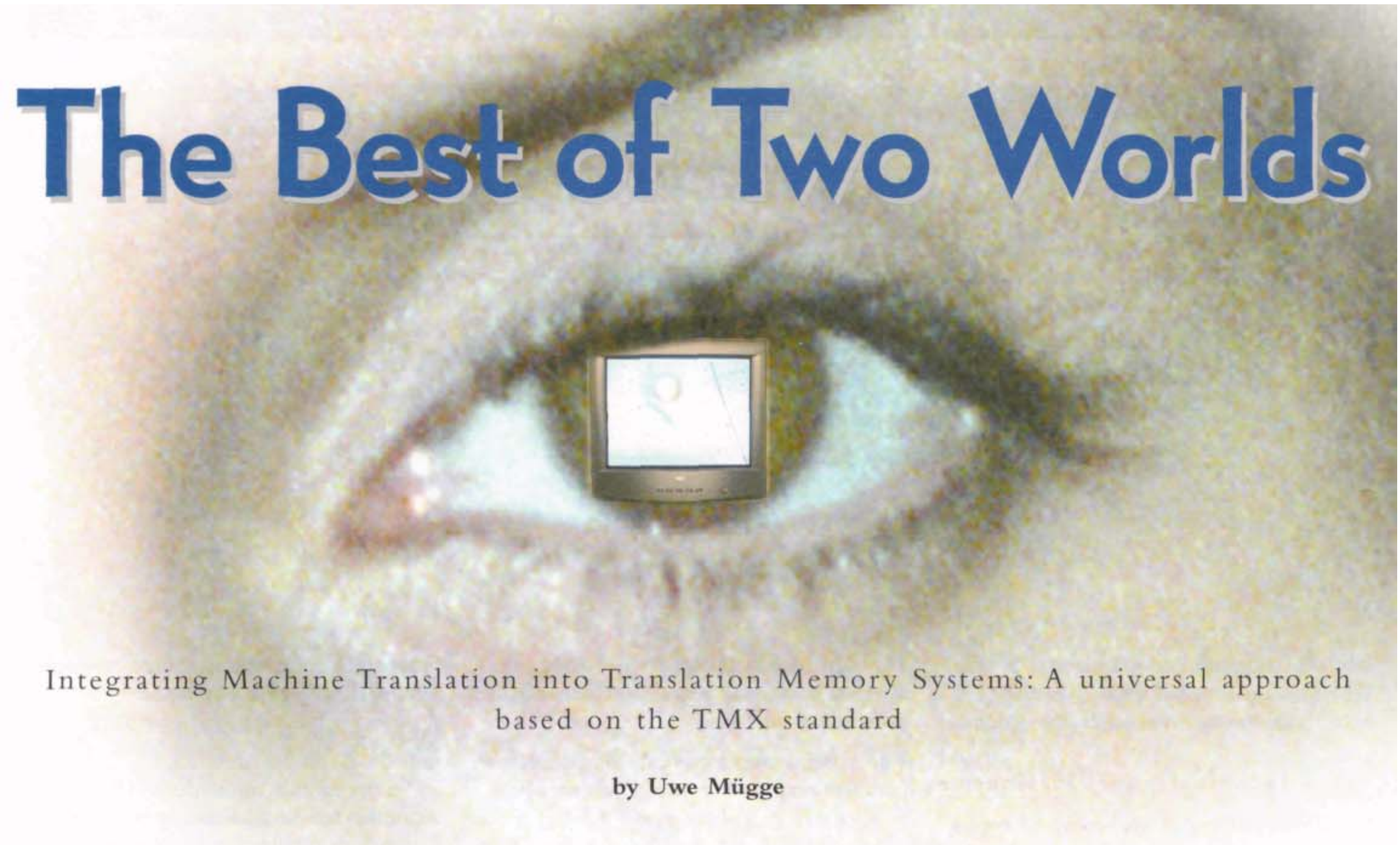


The Best of Two Worlds

Integrating Machine Translation into Translation Memory Systems: A universal approach based on the TMX standard

by Uwe Mügge

Let me begin by defining some terms. Machine Translation or MT refers to systems capable of translating documents from one language to another without user intervention. SYSTRAN and Power Translator are typical examples of commercial MT systems. Translation Memories or TM, on the other hand, refers to systems capable of recognizing previously translated material for reuse. In addition, many TM systems offer additional features such as powerful terminology management and conversion functionality. TRADOS Translator's Workbench and STAR Transit are among the best-selling commercial TM packages.

After more than fifty years of research and development, today there is a renewed interest in and enthusiasm for Machine Translation. This new interest is evident in the fact that many organizations and individuals actually use MT for various translation purposes. In addition, many institutions training language professionals take MT seriously enough to offer classes in MT theory and practice.

Low-Cost MT Systems

There are more low-cost, high-quality MT systems on the market covering more language combinations than ever before. For example, John Hutchins' Compendium of Translation Software identifies 230+ MT systems for English as source language alone. Target languages include most of the world's languages and many less commonly taught languages such as Afrikaans, Malay, and Zulu. Moreover, the 400+ products in the Compendium include many powerful low-cost MT systems even small translation agencies and freelancers can afford.

> **After more than fifty years of research and development, today there is a renewed interest in and enthusiasm for Machine Translation.**

The benefits of MT systems include:

- High translation speed

One of the undisputed advantages of MT systems is their impressively high translation speed. Many commercial MT systems are capable of translating hundreds, if not thousands, of pages per day (obviously without human revision).

- Terminological consistency

The majority of today's commercial MT systems belongs to the category of the so-called Transfer or Rule-based MT systems. These MT systems generally enable users to specify the terminology for a particular translation project. Rule-based MT systems generally follow users' terminology specifications, which ensures terminological consistency within and across documents or projects.

- Stylistical consistency

MT systems produce identical target-sentence structures for identical source-sentence structures. Using rule-based MT systems ensures stylistical consistency even in the largest translation projects.

- Completeness on the sentence level

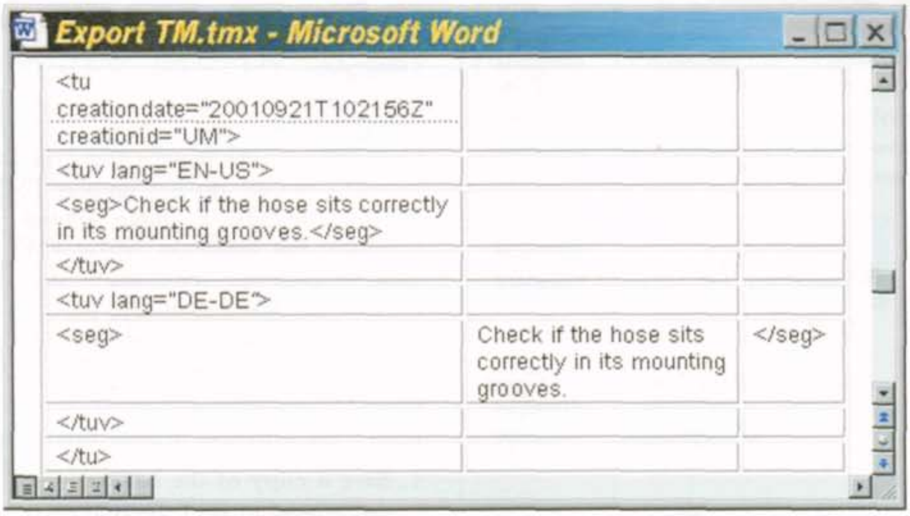
MT systems do not get tired, bored, or distracted—they reliably parse and translate every sentence in a source document. This is why using an MT system ensures completeness on the sentence level.

Organizations operating globally understand that documents authored for their domestic market require time-consuming and expensive localization for each foreign market. The internationalization of documents aims at authoring documents for easy localization. Internationalized documents avoid hard-to-translate information such as idiomatic expressions, jargon, humor, references to history, politics, religion, etc.

In addition, many organizations realize that the readability of their documents has an impact on the usability of their products. Highly readable documents reduce the number of operator errors, support-center calls and generally improve customer satisfaction. Authoring documents for improved readability means presenting information in an unambiguous and concise fashion. Documents authored for maximum readability are highly structured and consist of short sentences with a simple syntax.

John Hutchins' Compendium of Translation Software identifies 230+ MT systems for English as source language alone.

Internationalization and optimization for readability eliminate most of the obstacles for the successful use of MT (and also human translation). This is why MT produces excellent results with internationalized and optimized-for-readability documents, i.e. documents authored for translation.



<tu creationdate="20010921T102156Z" creationid="UM">		
<tuv lang="EN-US">		
<seg>Check if the hose sits correctly in its mounting grooves.</seg>		
</tuv>		
<tuv lang="DE-DE">		
<seg>	Check if the hose sits correctly in its mounting grooves.	</seg>
</tuv>		
</tu>		

Figure 1 Single entry in TMX format with unknown sentence isolated in middle column.

The Case for Integration

There are two approaches to correcting MT output:

1) Editing documents after translation (*post-editing*).

2) Editing the source text to prevent translation errors (*pre-editing*).

To avoid correcting the same problem repeatedly, using a Translation memory system is advisable for both types of editing work.

A number of MT vendors offer integrated MT/TM products, i.e. MT systems with a built-in TM component. However, the TM component of these MT systems is generally not nearly as powerful as a full-featured standard Translation Memory application. In addition, some standard TM products feature TM/MT interfaces for integrating MT. Unfortunately, these interfaces are product-specific, i.e. limited to a fraction of available MT products only.

The relevant functionality of TM products include:

- Pre- and post-editing

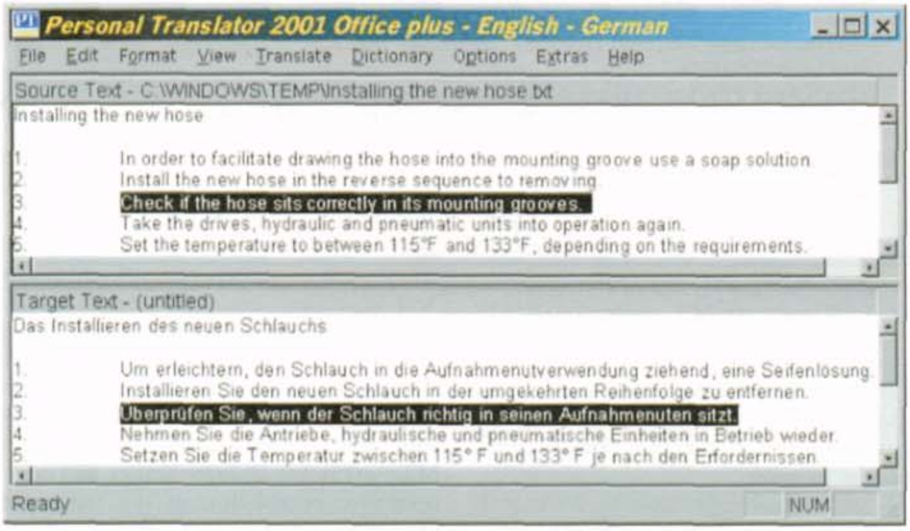
Integrated MT/TM systems generally only permit post-editing of MT output.

MT systems do not get tired, bored, or distracted–they reliably parse and translate every sentence in a source document. This is why using an MT system ensures completeness on the sentence level.

However, pre-editing is advisable when documents have not been properly internationalized and will be translated in multiple languages using MT. Many TM systems feature sophisticated spell-checking functionality, and some TM systems even offer grammar and style checking.

- Automatic recognition of terminology

One of the main benefits of MT systems is the consistent use of terminology. Setting up terminology is also one of the biggest challenges when using MT systems. There are normally limitations to the entries permitted in the dictionary of a given MT product, such as certain parts of speech, multi-word entries, entries with numbers, entries coded in the system dictionary, to name only a few. Even if the dictionary accepts an entry, the MT system may still use


Personal Translator 2001 Office plus - English - German
File Edit Format View Translate Dictionary Options Extras Help
Source Text - C:\WINDOWS\TEMP\Installing the new hose.txt
Target Text - (untitled)
Ready
NUM

Installing the new hose

1. In order to facilitate drawing the hose into the mounting groove use a soap solution.
2. Install the new hose in the reverse sequence to removing.
3. Check if the hose sits correctly in its mounting grooves.
4. Take the drives, hydraulic and pneumatic units into operation again.
5. Set the temperature to between 115°F and 133°F, depending on the requirements.

Das Installieren des neuen Schlauchs

1. Um erleichtern, den Schlauch in die Aufnahmenutverwendung ziehend, eine Seifenlösung.
2. Installieren Sie den neuen Schlauch in der umgekehrten Reihenfolge zu entfernen.
3. Überprüfen Sie, wenn der Schlauch richtig in seinen Aufnahmenuten sitzt.
4. Nehmen Sie die Antriebe, hydraulische und pneumatische Einheiten in Betrieb wieder.
5. Setzen Sie die Temperatur zwischen 115° F und 133° F je nach den Erfordernissen.

Figure 2 Single entry in TMX format with unknown sentence isolated in middle column.

an incorrect transfer. Most TM systems feature the automatic recognition of terminology stored in the dictionary component of the TM tool. This feature enables users to quickly spot and correct erroneous terminology transfers during post-editing.

Since many TM systems support TMX, this format is ideal for exchanging data between TM and MT systems.

- Automatic recognition and conversion of units of measurement

Converting numeric specifications in non-SI-units (Système International d'Unités) for SI-compliant markets, and vice versa, is a major localization challenge. For example, a length specification of '2.3 in' in a source text generally requires conversion to '5,84 cm' in the target text. Some MT systems recognize the abbreviation 'in' in the source text and produce the correct transfer in the target text. However, most MT systems will not automatically convert the figure '2.3' into '5,84'. A TM system like TRADOS Translator's Workbench, on the other hand, features the automatic recognition and conversion of units. Converting units of measurement inherently improves the speed and quality of the post-editing process.

- Automatic recognition and conversion of dates

In the U.S., the date '04/12/2001' refers to the twelfth day of April. In this example, the correct conversion for a German target text would be '12.04.2001'. Many MT systems simply do not handle date conversions. A TM system like TRADOS Translator's Workbench, on the other hand, reliably performs the automatic recognition and conversion of dates.

TMX: A New Integration Solution

In 1997, the Localisation Industry Standards Association (LISA) published the first version of the Translation Memory eXchange (TMX) format. The primary purpose of the TMX standard is to enable the exchange of translation memories between different TM systems. TMX is XML-compliant and supports the UNICODE standard. According to LISA, more than a dozen translation tools, both TM and MT systems, currently support the TMX format.

Since many TM systems support TMX, this format is ideal for exchanging data between TM and MT systems. Unfortunately, the popular low-cost MT systems generally do not support TMX. Below I describe a simple solution to overcome this limitation, where TMX is used to exchange information between the MT and TM system. This process enables you to post-edit translations created in any MT system using any TMX-compliant TM system.

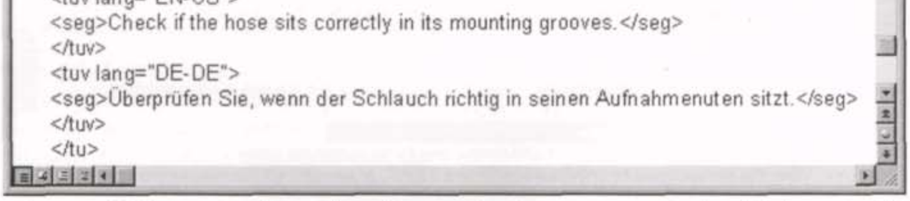
Here is the workflow in more detail:

1. Save a copy of the source document in RTF format

The step involves saving a copy of the source document first in a suitable text-only format. Open the text-only document and save this new source document in RTF format. The original source document retains all paragraph formatting, and the TM later applies paragraph formatting automatically to the translation. This step does not apply to source documents in HTML/XML/SGML format if the respective MT system supports these formats.

2. Analyze the source document and export unknown sentences as TMX

Avoiding the translation/post-editing of sentences more than once requires analyzing the source document against a TM database and exporting the unknown sentences only. Depending on the TM product, analysis may be automatic (e.g. as pre-translation) or require user intervention. After completing the analysis, create an export file with all unknown sentences for translation in the MT system. It is important to select TMX as the export format in order for the subsequent processes to work.



```
<tu creationdate="20010921T102156Z" creationid="MT!">
<tuv lang="EN-US">
<seg>Check if the hose sits correctly in its mounting grooves.</seg>
</tuv>
<tuv lang="DE-DE">
<seg>Überprüfen Sie, wenn der Schlauch richtig in seinen Aufnahmenuten sitzt.</seg>
</tuv>
</tu>
```

Figure 3 TMX entry complete with machine-translated target segment and modified creation ID.

3. Extract the unknown sentences from TMX

Since many MT systems do not support TMX yet, extracting the unknown sentences from the TMX export file is required. One way of extracting the unknown sentences involves a simple process in a word processor such as Microsoft Word.

a) First, open the TMX export file in your word processor.

b) Insert delimiters around the unknown sentences, and convert the body text into a table. Fig. 1 shows an entry in a TMX export file after separating unknown sentences in a table.

With this solution, linguists can create their own integrated MT/TM translation environment.

c) Copy the middle column with the unknown sentences, and paste the unknown sentences in a new document.

d) Save the new document with the unknown sentences in a suitable text-only format.

e) Save the TMX document in the TMX format.

4. Machine-translate the unknown sentences

Open the text-only source document in the MT system and set all required parameters (e.g. language combination, domain, dictionaries, etc.). Then follow the standard procedure for machine-translating the unknown sentences in the given MT system. Fig. 2 shows source and target text after completion of the translation process in an MT system. After completing the translation process, save the machine-translated document in a suitable text-only format.

5. Restore the TMX format

a) Return to the TMX document in your word processor.

b) Open the machine-translated document, select all, and copy the translation.

c) Paste the translation into the middle column of the TMX document.

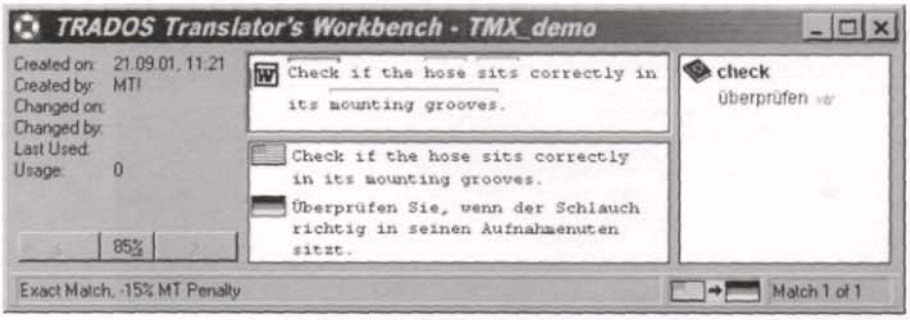
d) Convert the table back to text and remove the delimiters inserted at step 3.

e) Change the entry for "creationid" as needed.

f) After restoring the TMX format, save this document in TMX format. Fig. 3 shows an entry in the TMX document after the machine translation has been inserted.

6. Import the TMX document in the TM system

Importing the TMX document in the TM system means importing an external translation memory. Depending on the TM product, importing a TMX document may be a one-step process or may involve setting several parameters.


TRADOS Translator's Workbench - TMX_demo

Created on: 21.09.01, 11:21
Created by: MT!
Changed on:
Changed by:
Last Used:
Usage: 0

Check if the hose sits correctly in its mounting grooves.

Check if the hose sits correctly in its mounting grooves.

Überprüfen Sie, wenn der Schlauch richtig in seinen Aufnahmenuten sitzt.

check
überprüfen

85%

Exact Match, -15% MT Penalty

Match 1 of 1

Figure 4 TM window with source sentence and machine-translated proposal.

7. Import the original source document into the TM system

So far, we have translated a text-only copy of the source document. In order to produce a fully formatted target document, the original source document must be imported into the TM system. To import the original source document into your TM system, follow the standard procedures for importing a TMX translation memory.

8. Post-edit the machine-translated sentences

Now the TM system is ready for processing the original source document. The TM system will offer a translation for each sentence in the source document: either a previous translation (validated by a human linguist) or a new proposal (translated by the MT system). In this environment, the human linguist is post-editing the machine-translated sentences or validating relevant previous translations. Generally, TM systems indicate the source of a translation proposal (human linguist or MT system). At this stage, the TM system automatically applies all paragraph formatting in the source document to the target document.

The human linguist may now use features such as automatic recognition of terminology for maximum efficiency and quality. Fig. 4 shows a TM window with a source sentence and a translation proposal. Note that this window shows "MT" as source of the translation proposal. Also, note the horizontal bar over certain words, which indicates an entry in the terminology database of the TM system. Fig. 5 shows an overview of the complete translation process indicating the type of application used for each step.

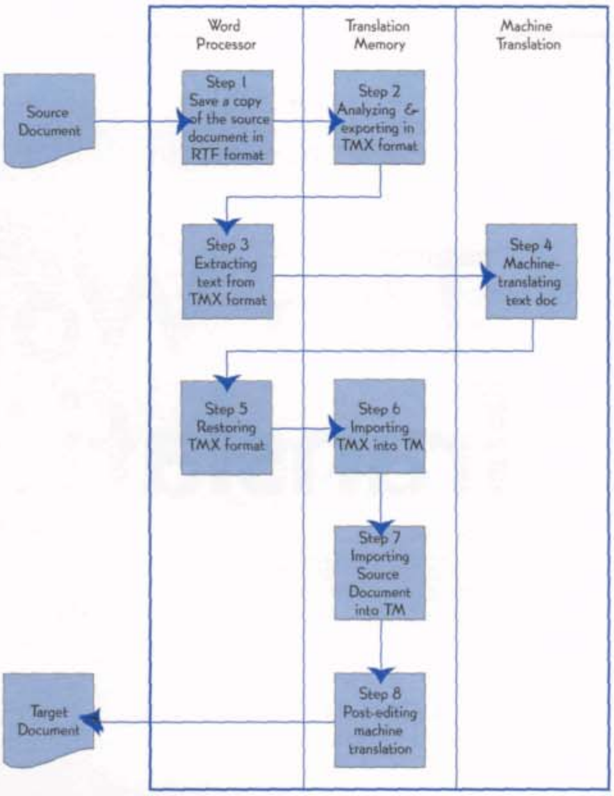

Word Processor
Translation Memory
Machine Translation
Source Document
Step 1 Save a copy of the source document in RTF format
Step 2 Analyzing & exporting in TMX format
Step 3 Extracting text from TMX format
Step 4 Machine-translating text doc
Step 5 Restoring TMX format
Step 6 Importing TMX into TM
Step 7 Importing Source Document into TM
Step 8 Post-editing machine translation
Target Document

Figure 5 Overview of the complete translation process.

Summary of benefits

Universal solution—The process described here is a truly universal solution for integrating any MT system into any TMX-compliant TM product. With this solution, linguists are no longer limited to the one or two MT systems supported by their TM system. Instead, the whole range of 400+ MT systems is now available for integration.

Improved productivity—Integrating MT into TM means that there is a translation proposal for each sentence in the source document. With properly internationalized and structured documents, linguists will primarily perform minor to modest post-editing instead of translating from scratch.

Immediate availability—The solution presented above is so simple that any technically knowledgeable linguist can implement it. With this solution, linguists can create their own integrated MT/TM translation environment.

Uwe Mügge

Uwe Mügge has more than ten years of experience in the localization industry serving in a wide variety of functions. He holds MA degrees from the University of Oregon (Telecommunication) and the Monterey Institute of International Studies (German Translation) and developed solutions for adding essential functionality such as automatic terminology extraction (patent pending) to standard translation environments. Uwe Mügge runs his own localization consulting business and teaches Translation Management & Terminology at IHL Lindau. He can be reached at info@muegge.cc.