Workflow in Wordfast and Invisible Machine Translator

MILAN CONDAK

INTRODUCTION

Several millions of customers throughout the world use MS Word, and about 12,000 of them use Wordfast. Everybody can easily get to know what Wordfast is at http://www.wordfast.net. One can download Wordfast/Plustools programs/manuals from that site. By December 2003, the site was presented in 18 languages: Albanian, Croatian, Czech, Dutch, English, Finnish, French, German, Greek, Hebrew, Hungarian, Italian, Japanese, Macedonian, Polish, Portuguese, Russian, and Spanish. The Manual is in 19 languages: Chinese simplified, Chinese traditional, Czech, Dutch, German, Greek, English, French, Hungarian, Italian, Japanese, Korean, Latvian, Polish, Portuguese, Romanian, Russian, Spanish, and Ukrainian.

Wordfast is available for free (without any license) for translation memories of up to 110 Kbytes and/or 500 translation units. In other words, Wordfast can be used for short to medium size jobs, for free, with all features activated.

Buying a licence removes this limitation and allows one to use Wordfast for jobs of any size. Buying a licence after training, a new user saves 33% of the licence cost.

The workflow is briefly described at:


Here is a practical example:

Supposing that Trados® is your typical translation environment. Then the following workflow can be used:

1. The source files are prepared/tagged in-house using your existing tools and filters.
2. Then the files are translated using Wordfast.
3. The target files are then polished/untagged in-house.
Wordfast is an open system: all the data can be maintained using such broadly known tools as MS Word, MS Excel, MS Access, and the like. This article describes Stage 2: Using Wordfast to translate the files.

The author of the program, Yves Champollion, France, made a presentation at the ASLIB Conferences and Exhibition in London, on 21 November 2003: “Convergence in CAT: Blending MT, TM, OCR, and Voice Recognition to Boost Productivity”:

http://www.aslib.com/conferences/programme.htm

That presentation, using live examples of various technologies, demonstrates what the integrated translation tool of the near future will look like.

The present paper discusses an actual combination of these technologies and the perspectives they open.

Translation Memory

Translation memory (TM) has become the tool of choice for most professional translators. But TM productivity quickly hits the ceiling, which can be significantly raised by linking TM to other technologies, such as OCR, MT, text-to-speech, voice recognition.

How can we translate files with Wordfast? We need MS Word and some operating system (MS Windows, Mac or Linux). The user can choose between several workflows, depending on the linguistic tools he/she uses.

Usually, we have MS Word + Wordfast + a file to translate. We can use any of the following 5 main methods and their combinations:

0th We have no other automatic tool, and we create a new TM.
1st We have a machine translator integrated with MS Word.
2nd We have a machine translator, which has no toolbar in MS Word.
3rd We have a glossary or glossaries.
4th We have a reliable TM.

The work can be performed in two modes:

---

- working with the file
- working with file segments.
1. Working with the file - Pre-translate

First, we select a document, after which we can use the translation tools in the following order:

   Analyze, Translate, Cleanup, Quality Check.

Let's now discuss Translate. This tool will pre-translate the selected document(s) using the current translation memory. The unknown (no-match) segments will be copied over the target segment if you specify CopySourceWhenNoMatch in the Pandora Box. However, if a link with a machine translation program is activated, the unknown segments will be machine translated. The Pandora Box is list of commands, the latter being case-insensitive, separated with paragraph marks.

If this function is started over an empty document with either Excel, Access or PowerPoint running in the background, the corresponding Excel/Access/PowerPoint document will be translated. If this function is started with the MT settings activated, machine translation will be performed on the unknown segments.

Once the pre-translation is done, the user starts a regular Wordfast session to translate his/her document(s) in the usual way. The work will be faster because segmentation and matching have already been performed.
2. Working with segments = a regular Wordfast session

Next validates the current segment and moves to the next one.

Copy Source copies the source segment over the target one, or translates the terms recognized in the glossary and shows the target-language terms.

Expand and Shrink influence the end of the segment. Expand expands the segment when the sentence actually extends beyond the final punctuation mark. Shrink reverses any use of the Expand segment icon.

Translate translates segments without opening them until a non-exact match is found.

Placeable is an untranslatable element that is simply copied from the source to the target. There are three icons for copying of placeables.
Dictionary looks up a word/expression in the currently active external dictionary.

Context scans the translation memory and displays all the text units containing a specific word.

Concordance scans the text files located in the folder specified in the Wordfast/Files/Set concordance folder. The following text file types are processed: TXT, HTM, HTML, CSV.

Glossary looks up a word/expression in the glossaries.

Memory displays the contents of the relevant text unit above a segment.

Quality check toggles real-time quality check during the translation session.

Quick cleanup cleans up a document without updating the translation memory. This command can be used if the user has revised the document by re-opening the segments, so that the changes are recorded in the TM.

F (fast) opens the set-up/working panels.

Examples of working with files

Let us have a Polish text:

<table>
<thead>
<tr>
<th>OŚWIADCZENIE I UPOWIĄZNIENIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zgodnie z obowiązującymi przepisami dotyczącymi podatku od towarów i usług oświadczamy, że:</td>
</tr>
</tbody>
</table>

0th method. Wordfast only makes segmentation, and the source segment is copied to the target one:

<table>
<thead>
<tr>
<th>OŚWIADCZENIE I UPOWIĄZNIENIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zgodnie z obowiązującymi przepisami dotyczącymi podatku od towarów i usług oświadczamy, że:</td>
</tr>
</tbody>
</table>

1st method: the target segment is translated by a machine translation system MT (in this case, it is PARS/P by Lingvistica '98 Inc.: www.linp98.com):

<table>
<thead>
<tr>
<th>DECLARATION AND AUTHORIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>According to obligatory rules referring tax from commodities and services we declare that:</td>
</tr>
</tbody>
</table>
The second method TM was created by aligning the original document and its machine translation "analog". First, we machine translate the source document. The following picture illustrates the Czech PC Translator program, which is another typical machine translator.

The text files (or html, csv) can be translated without opening the source documents:

Note: Soubor originálu stands for 'Original file', Soubor překladu stands for 'Translated file', Přeložit means 'Translate'.
Our files have respective names kde en.rtf and kde cz.rtf, and the source document was translated by opening it in MS Word.

A good way of explanation would be translating a document from an "unknown" language into the customer's native tongue. Let's analyze the text of the Czech national anthem as an example. The text itself is short enough: it has a title and some repetitions. The song below consists of two strophes, but only the first one represents the Czech anthem.

WHERE IS MY HOME
Where is my home?
Where is my home?
Water bubbles across the meadows,
Pinewoods rustle among crags,
The garden is glorious with spring blossom,
Paradise on earth it is to see.
And this is that beautiful land,
The Czech land, my home,
The Czech land, my home.

As the aligning tool, we can use the +Tools shareware program and its +Align feature. I use CS (CS-01) for the source Czech text, and CZ for the Czech text generated by the machine translator.

First of all, we align the two files in a joint table (you can only see the first strophe below).
<table>
<thead>
<tr>
<th>WHERE IS MY HOME</th>
<th>KDE JE MŮJ DOMOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is my home,</td>
<td>Kde je můj domov,</td>
</tr>
<tr>
<td>Where is my home?</td>
<td>Kde je můj domov?</td>
</tr>
<tr>
<td>Water bubbles across the</td>
<td>Voda bubliny přes</td>
</tr>
<tr>
<td>meadows,</td>
<td>louky,</td>
</tr>
<tr>
<td>Pine woods rustle among</td>
<td>Borové lesy suti</td>
</tr>
<tr>
<td>crags,</td>
<td>mezi útesy,</td>
</tr>
<tr>
<td>The garden is glorious</td>
<td>Zahraďa je skvělá</td>
</tr>
<tr>
<td>with,</td>
<td>s jarem květ,</td>
</tr>
<tr>
<td>spring blossom</td>
<td></td>
</tr>
<tr>
<td>Paradise on earth it is</td>
<td>Ráj na zemi to je</td>
</tr>
<tr>
<td>to see.</td>
<td>vidět.</td>
</tr>
<tr>
<td>And this is that</td>
<td>A toto je krásná</td>
</tr>
<tr>
<td>beautiful land,</td>
<td>země,</td>
</tr>
<tr>
<td>The Czech land, my home,</td>
<td>Česká země, můj</td>
</tr>
<tr>
<td></td>
<td>domov.</td>
</tr>
<tr>
<td>The Czech land, my home.</td>
<td>Česká země, můj</td>
</tr>
<tr>
<td></td>
<td>domov.</td>
</tr>
</tbody>
</table>

The correct Czech (CS-01) text of the Czech national anthem is presented in the example of 4th method.

If the source and target segments need adapting/aligning, the customer can use a small toolbar and the mouse.

The TM created can be used for translating.

The third method uses commands in the Pandora Box:
“CopySourceWhenNoMatch”, Propagate, Case=0
The Translate command generates the following result:

- WHERE IS MY HOME?
- KDE DOMOV MŮJ

Examples of working with segments = A regular Wordfast session

0th method - we have no automatic tool, and we create a new TM
The new translation memory has 0 translation units.

Nothing is offered by the TM for the first verse.

We get the first offer from the TM because the matching exceeds 75%:

The matching is 98%, the text is capitalized, and a comma is missing.

The text in the target segment is edited by the human translator.

The 1st method - we have a machine translator integrated with MS Word.
This is another "invisible MT system" mentioned in this article; it translates the opened segment.

Clicking Next, we get the machine translation, displayed in the shadowed field:

WHERE IS MY HOME

DOVE È LA MIA CASA

Clicking Next, we get the translation from the TM (see the green box below):
2nd method - we have an MT system that has no toolbar in MS Word

We have both an MT system and a glossary. You can see the words recognized in the opened segment and found in the glossary.

- We have raw translation memory
- File is pretranslated
- We have glossary

Offer from TM created by aligning source document
and raw translation from MT (PC TRANSLATOR) in PlusTools

We have TM with language pair EN-RU, but question marks are displayed in place of Cyrillic characters. In this case, we must convert TM to Unicode.

The word “HOME” has been found in the glossary. Standard matching from the aligned TM is 100%.
But we use the terminology from the glossary by clicking the Copy icon (the Alt+Insert command):

3rd method - we have a glossary or glossaries

3rd translation variant

You can see the words recognised in the opened segment and found in the glossary (the result is better than in the 0th method).
Matching is 98%, the text is capitalized, and a comma is missing (just like in the 0th method). The text in the target segment needs editing.

We have allowed to get offers from the TM if matching is below the actual threshold of 75%. Because the segments are short, we can get bad offers. Here, the Quality Control works on the fly.

For the 13th verse, we get a verification question:

For the 18th verse, we get a verification question again. If we open the glossary, we shall see:

The matching is 72%, and the offer is from the second verse of the anthem. The correct terminology is propagated from the glossary:
The next verse only differs in the full-stop and comma, the matching being 82%:

\[
\begin{align*}
&\text{Among Čechy, j mé domov.} \\
&\text{Mezi Čechy domov můj.}
\end{align*}
\]

We shall see another example of combining the 1\textsuperscript{st} method on the file and the 3\textsuperscript{rd} method on the segments as the last example of the workflow. We continue with the last main method.

4\textsuperscript{th} method - we have a reliable translation memory. We have:

a) the translation memory from our previous job, received from the client, or

b) the aligned translation memory of the same text in two different languages (from the previous job, website, files on a CD-ROM, etc).

Here is a TM from Esperanto - Czech documents, made in PlusTools.

<table>
<thead>
<tr>
<th>KIE HEJMO MIA</th>
<th>KDE DOMOV MŮJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kie hejmo mia, kie hejmo mia?</td>
<td>Kde domov můj, kde domov můj?</td>
</tr>
<tr>
<td>Akvo mugas sur herbejoj, ĉiuj sur laŭdoj, en jardeno je la floro, kaj la paradizo je unua vido.</td>
<td>Voda hučí po hučinách, bory šumí po skalinách, v šádě skvěle jara květ, zemský ráj to na pohled.</td>
</tr>
<tr>
<td>Kaj tie estas la bela lando, lando ĉeĥa, hejmo mia, lando ĉeĥa, hejmo mia.</td>
<td>A to je ta krasná země, země česká, domov můj, země česká, domov můj.</td>
</tr>
</tbody>
</table>
The right column of the table contains the correct Czech (CS-01) text of the Czech national anthem.

Matching is 100% in the green field. The next segment:

\[ \text{Kie hejmo mia} \]

Matching is 98% in the yellow field. The text in the target segment needs editing.

\[ \text{Kde domov můj} \]

The text in the target segment is edited by the human translator.

The last example is a combination of two modes and 3 languages. Supposing, we translate a manual:
We have a pre-translated file English-Polish (the 1st method) and an activated glossary EN-CS (the 3rd method), which was created in the previous jobs. The English-Polish translation is again performed by PARS/P by Lingvistica '98 Inc. (www.lmg98.com).

First, we have to set up Wordfast for performing machine translation, and wait till the translation is performed automatically.

Then we open the first segment. The source segment contains the source English text, while the target segment has the machine-made Polish translation, which we will only read as a “help” text. It can be useful if we understand Polish. The target segment contains no English words.

> Besides, the following keys are defined in the dictionary window:

> Pozna tym, następująca klawiatura się czynna w sposób okno.

>
You can see the words recognized in the source segment and found in the EN-CS glossary. Clicking the Copy icon (or pressing Alt+Insert), the recognized words will be translated into Czech:

\[9\]

\[\text{mínoto, the následující klávesy sou aktivní v slovníku okno.}\]

\[9\]

If we specify the CopyOnlyKnown command in the Pandora Box, the unrecognized words “the” and “in the” will not be copied to the target segment.

**Working with the files = Clean-up**

The translated bilingual text contains delimiter marks and source language fragments:

```
10-KIE HEIMO MIA KIE-DO MOVY-MMU<br />
10-Kie heimo mia Kde domov mňo Kie heimo mia? kde domov mňo<br />
10-Akvo mugas sur herbejoi Voda hučí po lučinách<br />
10-boskoj susuras sur rokaroi bory šumí po skalinách<br />
10-en ďardeno juvelas printemps floro v sádi skví se jara květ<br />
10-len surtera paradizo te unua video zemský ráj to na pohled<br />
10-Kai tio estas la bela lando A to je ta krásná země<br />
10-boskoj susuras sur rokaroi bory šumí po skalinách<br />
10-lando češa heimo mia země česká domov mňo<br />
10-lando češa heimo mia země česká domov mňo<br />
```

Thus we have three results: the cleaned translated document, the complete translation memory, and a report describing what has been cleaned up.

Only a few language pairs have machine translators integrated with MS Word. Czech customers can use desktop visible MT, while some language pairs have on-line Internet translators. All users can get the same text in two different languages. Please see [http://wordfast.net](http://wordfast.net)
and the Wordfast manuals for details. My advice is to use the 4th method in a combination with the 3rd one: a segmented document or a raw pre-translated document with a good glossary.

**Invisible machine translator**

**Standard visible MT**

Most of Czech customers use **PC Translator**. Such MT produces draft translations of foreign language documents. The customers can use it for creating:

a) a draft TM (for the 2nd and 4th methods)

b) a draft and an accurate glossary by extracting terms and general-usage words from the source document. For doing this, one can use the `Extract` feature in Plus Tools.
Invisible MT systems integrated with Microsoft Office

In such cases, we have a new item in the MS Word menu – 'Translate'. The submenu serves to select the language pair and translation direction as well as the work mode.

MT systems integrated in the file mode

If we use an MT system integrated in file mode, we do not align the documents. The opened document is segmented by Wordfast, and the resulting segments are pre-translated by the MT system.

<table>
<thead>
<tr>
<th>Month</th>
<th>January</th>
<th>February</th>
<th>March</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>40,982</td>
<td>65,852</td>
<td>63,936</td>
</tr>
<tr>
<td>Forecast</td>
<td>45,200</td>
<td>78,300</td>
<td>45,900</td>
</tr>
<tr>
<td>Cumulative Sales</td>
<td>40,982</td>
<td>106,814</td>
<td>172,743</td>
</tr>
<tr>
<td>Cumulative Forecast</td>
<td>45,200</td>
<td>123,500</td>
<td>199,400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Соседний</th>
<th>Гриль</th>
<th>Магнит</th>
</tr>
</thead>
<tbody>
<tr>
<td>Производство</td>
<td>40,982</td>
<td>85,832</td>
<td>65,936</td>
</tr>
<tr>
<td>Продажи</td>
<td>45,200</td>
<td>78,300</td>
<td>45,900</td>
</tr>
<tr>
<td>Суммарные</td>
<td>40,982</td>
<td>106,814</td>
<td>172,743</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Соседний</th>
<th>Гриль</th>
<th>Магнит</th>
</tr>
</thead>
<tbody>
<tr>
<td>Производство</td>
<td>40,982</td>
<td>85,832</td>
<td>65,936</td>
</tr>
<tr>
<td>Продажи</td>
<td>45,200</td>
<td>78,300</td>
<td>45,900</td>
</tr>
<tr>
<td>Суммарные</td>
<td>40,982</td>
<td>106,814</td>
<td>172,743</td>
</tr>
</tbody>
</table>
PARS by Lingvistica '98 Inc. is a typical MT system, or, to put it more exactly, a family of MT systems. In PARS, the pars97.dot template must be activated to provide system integration with MS Word.

The advantage of such a solution is high speed of text processing, while its disadvantage is the fact that the user cannot work immediately with segments.

**MT systems integrated in the segment mode**

An example of such a system is PeTra = Personal Translator for MS Word. To run it, the user activates the MS Word supplement, ptword.wll.

PeTra translates between English and Italian.

More language pairs are offered, for example, by L&H Power Translator Pro 7. That system produces draft translations of foreign language documents and is integrated with Microsoft Office by means of supplement PRWRD32.wll.
It is a fully-featured tool for translating the following language pairs:

- French to/from English
- German to/from English
- Italian to/from English
- Portuguese to/from English
- Spanish to/from English
- Japanese to/from English
- French to/from German
- Italian to/from French
- Italian to/from German
- Italian to/from Spanish.

Please see details at http://www.bmsoftware.com/powertranslator.htm.

The system also includes a text-to-speech feature. If the customer has installed the Dragon Naturally Speaking system, he/she can get the best offers in the target segment and edit them by dictating the corrections. No typing is required at all.