

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:

<http://www.geocities.com/wordfast/nworkflow.html>.

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.

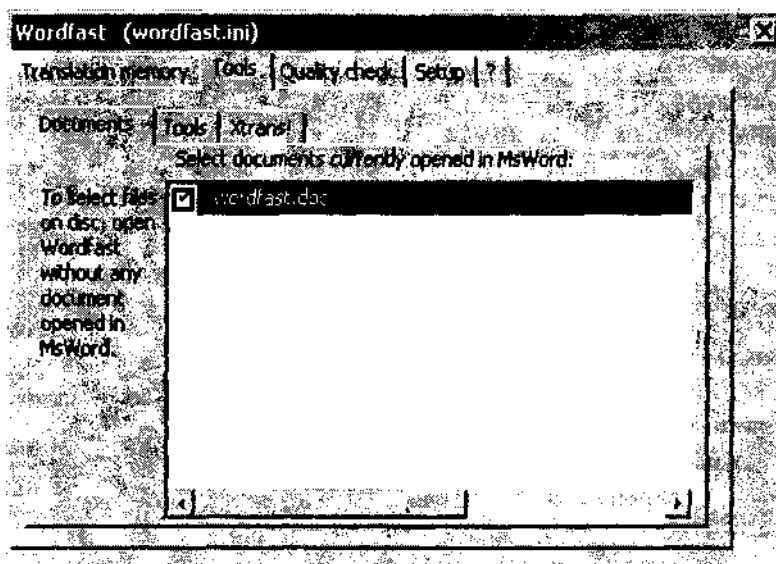
The work can be performed in two modes:

- 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:

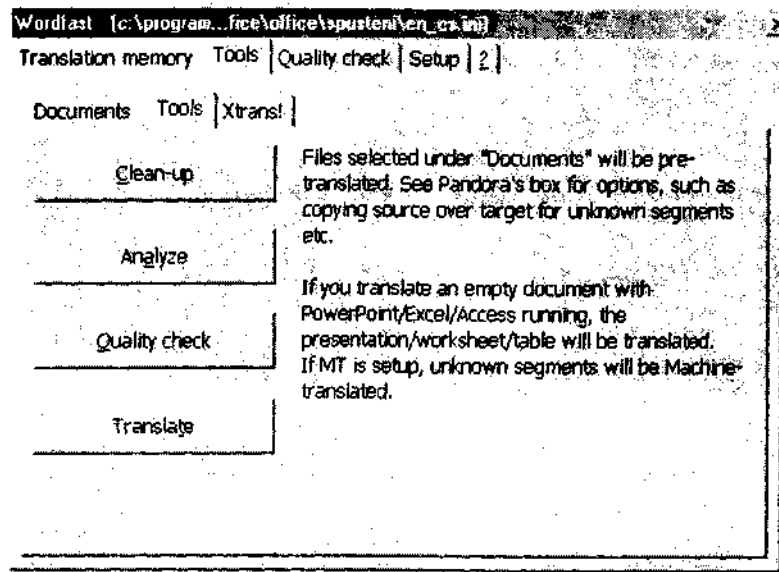
Analyze, Translate, Cleanup, Quality Check.



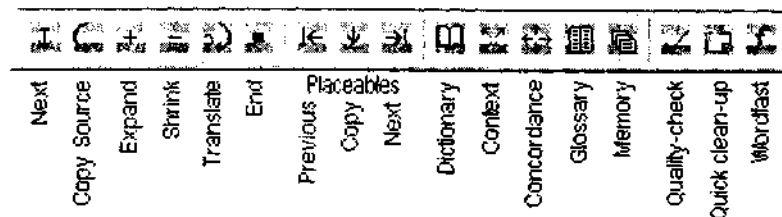
Let's now discuss **Translate**. This tool will pre-translate the selected document(s) using the current translation memory. The unknown (no-match) segments will be copied over the target segment if you specify **CopySourceWhenNoMatch** in the Pandora Box. However, if a link with a machine translation 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:

OŚWIADCZENIE I UPOWAŻNIENIE

Zgodnie z obowiązującymi przepisami dotyczącymi podatku od towarów i usług oświadczamy, że:

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

OŚWIADCZENIE I UPOWAŻNIENIE

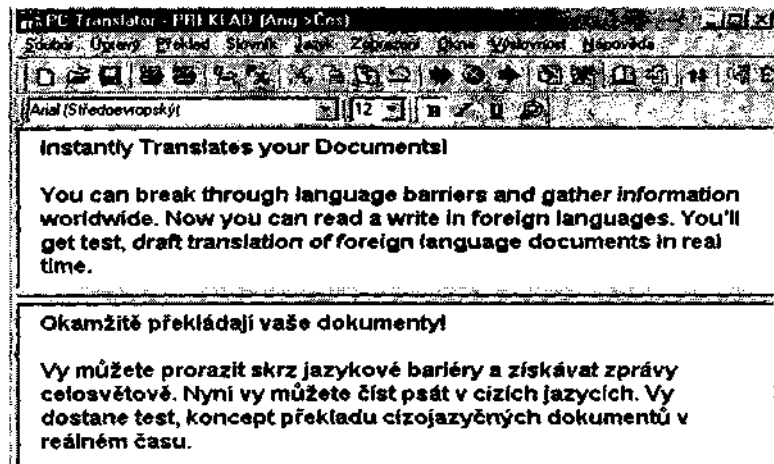
Zgodnie z obowiązującymi przepisami dotyczącymi podatku od towarów i usług oświadczamy, że:

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.ling98.com):

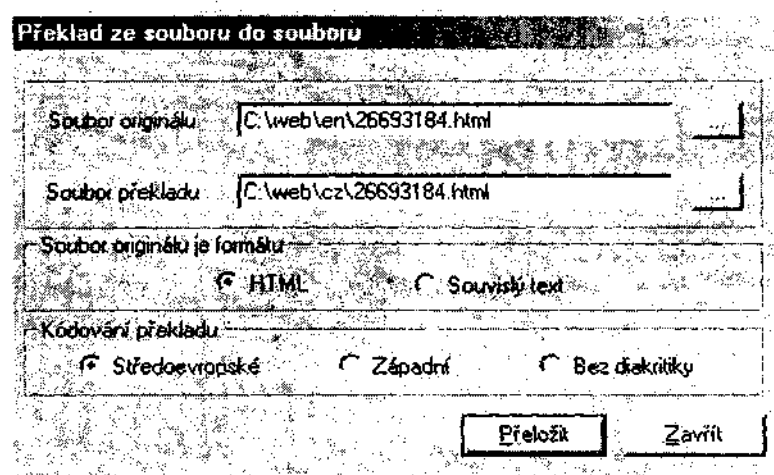
DECLARATION AND AUTHORIZE

According to obligatory rules referring tax from commodities and services we declare that:

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’.

WHERE IS MY HOME	KDE JE MŮJ DOMOV
Where is my home,	Kde je můj domov,
Where is my home?	Kde je můj domov?
Water bubbles across the meadows,	Voda bubliny přes louky,
Pine woods rustle among crags,	Borové lesy šustí mezi útesy,
The garden is glorious with, spring blossom	Zahrada je skvělá s jarem květ,
Paradise on earth it is to see.	Ráj na zemi to je vidět.
And this is that beautiful land,	A toto je krásná země,
The Czech land, my home,	Česká země, můj domov,
The Czech land, my home.	Česká země, můj domov.

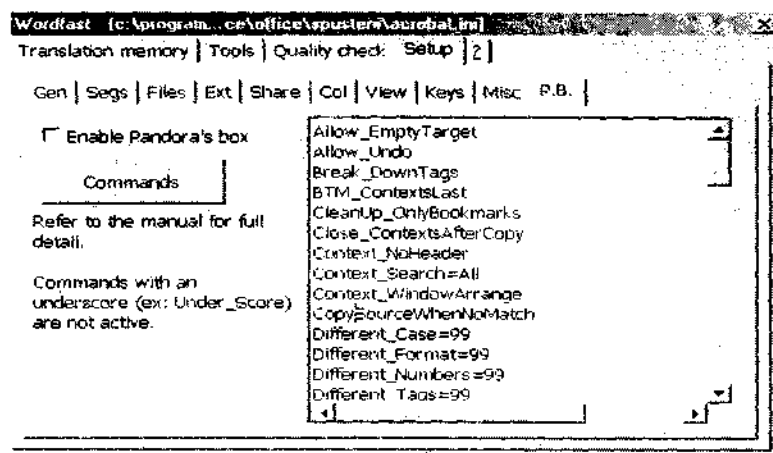
The correct Czech (CS-01) text of the Czech national anthem is presented in the example of 4th method.

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", Propagate1, Case=0

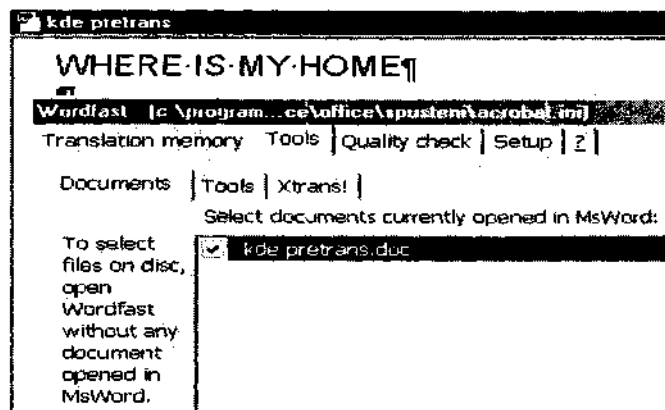


The Translate command generates the following result:

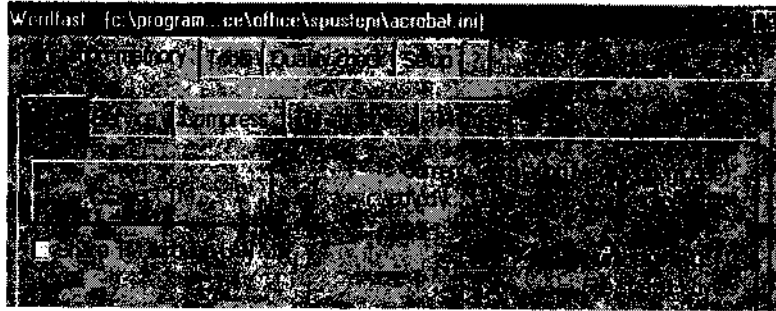
```
{0> WHERE·IS·MY·HOME.<|0|>KDE·DOMOV·MŮJ.<0|>¶
{0> Where·is·my·home.<|0|>KDE·DOMOV·MŮJ.<0|>¶
{0> where·is·my·home?<|0|>KDE·DOMOV·MŮJ?<0|>¶
{0> Water·bubbles·across·the·meadows.<|0|>Voda·hučí·po·lučinách.<0|>¶
{0> Pinewoods·rustle·among·craggs.<|0|>Bory·šumí·po·skalinách.<0|>¶
{0> The·garden·is·glorious.<|0|>The·garden·is·skví·se·spring·blossom.<0|>¶
    with·spring·blossom.
{0> Paradise·on·earth·it·is·to·see.<|0|>zemský·ráj·to·na·poled.<0|>¶
{0> And·this·is·that·beautiful·land.<|0|>a·to·je·ta·krásná·země.<0|>¶
{0> The·Czech·land·my·home.<|0|>země·Česká·DOMOV·MŮJ.<0|>¶
{0> The·Czech·land·my·home.<|0|>země·Česká·DOMOV·MŮJ.<0|>¶
{0> Where·is·my·home.<|0|>KDE·DOMOV·MŮJ.<0|>¶
{0> where·is·my·home?<|0|>KDE·DOMOV·MŮJ?<0|>¶
{0> If·in·a·heavenly·land·you.<|0|>If·v·bohumilém·land·you·setkal·se.<0|>¶
    ·have·met
{0> Tender·souls·in·agile·frames.<|0|>duše·milé·v·těle·čilém.<0|>¶
{0> Of·clear·mind·vigorous·and·prospering.<|0|>mysl·jasnou·v·znika·zdar.<0|>¶
{0> And·with·a·strength·that.<|0|>a·tu·sílu·vzdu·zmar!·defiance.<0|>¶
    ·frustrates·all·defiance.
{0> That·is·the·glorious·race·of·czechs.<|0|>to·je·te·Čechů·slavné·pléme.<0|>¶
{0> Among·Czechs·is·my·home.<|0|>mezi·Čechy·DOMOV·MŮJ.<0|>¶
{0> Among·Czech·is·my·home.<|0|>Among·Czech·DOMOV·MŮJ.<0|>¶
¶
```

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.

```
(0>¶
[WHERE IS MY HOME]
[< 0 ]¶
[WHERE IS MY HOME]
<0)¶
```

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

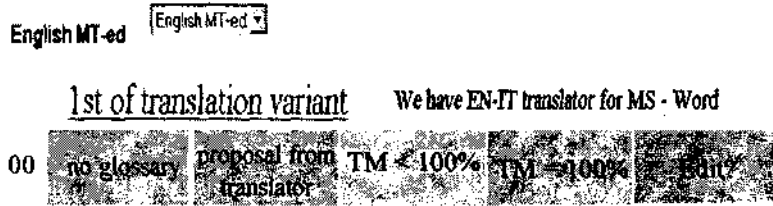
```
(0>¶
[Where is my home.]
[< 98 ]¶
[KDE DOMOV MŮJ]
<0)¶
```

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

```
(0>¶
[Where is my home.]
[< 98 ]¶
z [Kde domov můj,]¶
<0)¶
```

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:

<0>
 WHERE IS MY HOME
 <30<>
 DOVE È LA MIA CASA
 <0>

Clicking Next, we get the translation from the TM (see the green box below):

{0>
 where is my home?
 <100{>
 Dove ђ la mia casa?
 <0)

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.

English MT-ed English MT-ed

2nd and 3rd translation variant

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

00 edited by translator TM < 100% TM = 100%

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

{0>
 [WHERE IS MY HOME]
 <100{>
 [??? ??? ???]
 <0)

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.

{0>
 [WHERE IS MY HOME]
 <100{>
 [ГДЕ ДОМ МОЙ]
 <0)
]

The word "HOME" has been found in the glossary. Standard matching from the aligned TM is 100%:

(0>
 Water bubbles across the meadow.]
 <) 100 (>
 Вода жуужит вниз луг.]
 <0)

But we use the terminology from the glossary by clicking the Copy icon (the Alt+Insert command):

(0>
 Water bubbles across the meadow.]
 <) 100 (>
 Ручьи струятся по лугам.]
 <0)

3rd method - we have a glossary or glossaries

English MT-ed English MT-ed

3rd translation variant

We have glossary

found in glossary	edited by translator	TM < 100%	TM = 100%	
----------------------	-------------------------	-----------	-----------	--

(0>¶
 WHERE IS MY HOME?
 <) 0 (>¶
 KDE·DOMOV·MŮJ¶
 <0) ¶

You can see the words recognised in the opened segment and found in the glossary (the result is better than in the 0th method).


```
{0>¶
[Among Czechs is my home]¶
<|73+|>¶
[Mezi Čechy domov můj,]¶
<0>¶
```

The next verse only differs in the full-stop and comma, the matching being 82%:

```
{0>¶
[Among Czech, my home]¶
<|82|>¶
[Mezi Čechy domov můj,]¶
<0>¶
```

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

4th method - we have a reliable translation memory. We have:

- the translation memory from our **previous job**, received from the client, or
- the **aligned** translation memory of the same text in two different languages (from the previous job, web sites, files on a CD-ROM, etc).

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

KIE HEJMO MIA	KDE DOMOV MŮJ
Kie hejmo mia, kie hejmo mia? Akvo muĝas sur herbejoj, boskoj susuras sur rokaroj, en ĝardeno juvelas printempa floro, jen surtera paradizo je unua vido. Kaj tio estas la bela lando, lando ĉeĥa, hejmo mia, lando ĉeĥa, hejmo mia.	Kde domov můj, kde domov můj? Voda hučí po lučinách, bory šumí po skalínách, v řadi skví se jara květ, zemský ráj to na pohled. A to je ta krásná země, země česká, domov můj, země česká, domov můj.

The right column of the table contains the correct Czech (CS-01) text of the Czech national anthem.

English MT-ed English MT-ed ▾

{0>

[KIE HEJMO MIA] variant We have translation memory EO-CS.

<[100]> 75% TM<100% TM=100% Edit ?

[KDE DOMOV MŮJ]

<0)

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

{0>

[Kie hejmo mia]

<[99]>

[KDE DOMOV MŮJ]

<0)

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

{0>

[Kie hejmo mia]

<[98]>

[Kde domov můj]

<0)

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:

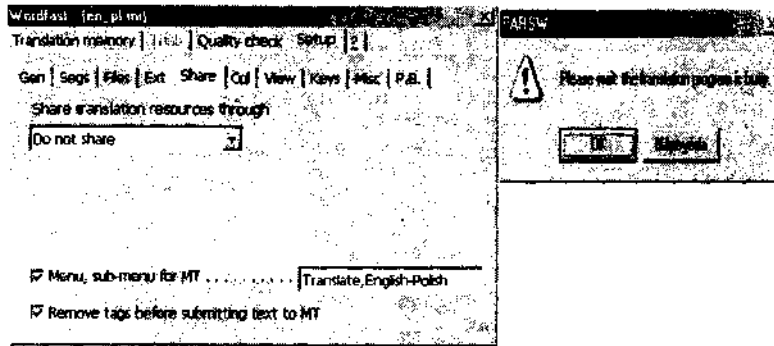
English MT-ed English MT-ed

- 1st + 3rd translation variant
- Translation EN-PL-CS
 - We have translator for MS-Word EN-PL
 - glossary EN-CS

00 Glossary EN-CS translator EN-PL TM < 100% TM = 100%

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.ling98.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.

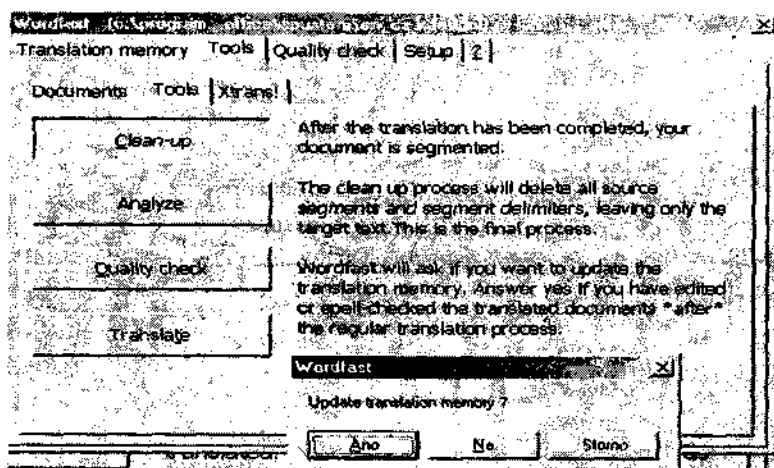
{0>

[Besides, the following keys are active in the dictionary window.]

<)<(>

[Poza tym, następująca klawiatura s¹ czynna w s³owniku okno:]

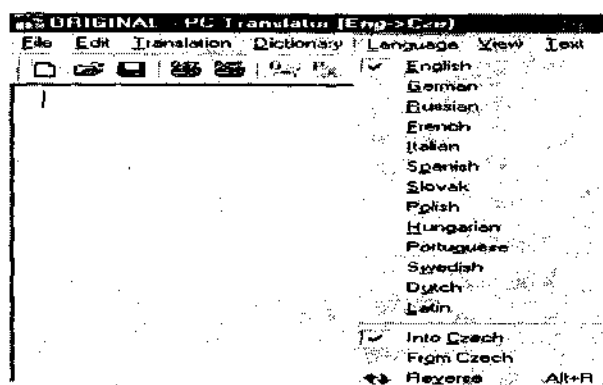
<0)



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:

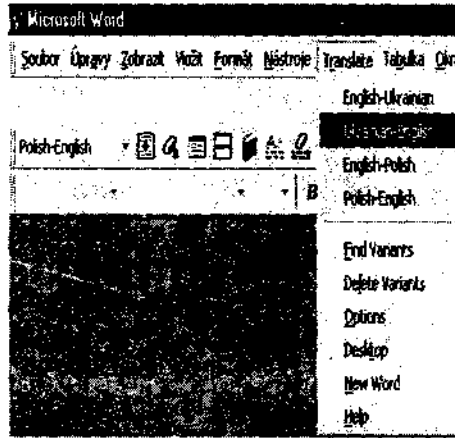
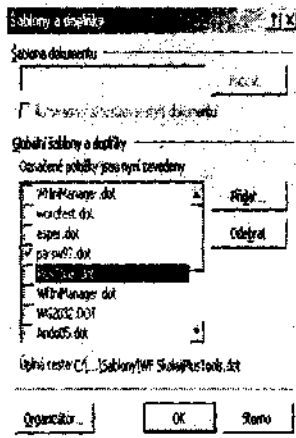
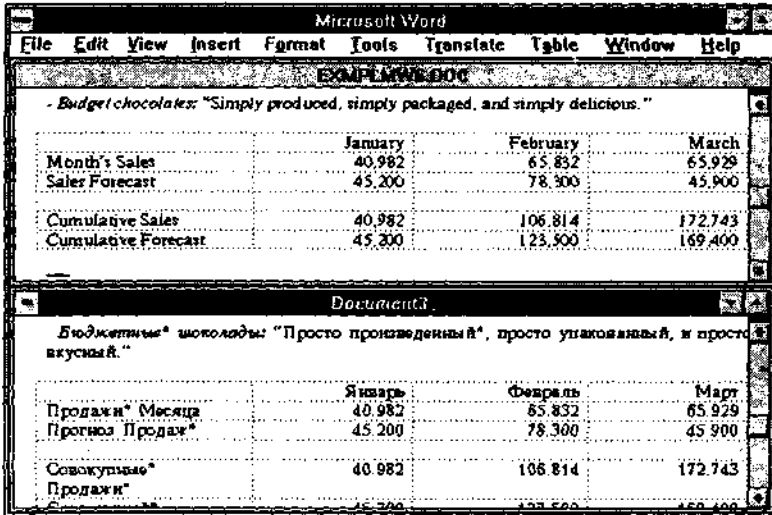
- a) a draft TM (for the 2nd and 4th methods)
- b) a draft and an accurate glossary by extracting terms and general-usage words from the source document. For doing this, one can use the +Extract feature in PlusTools.

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.



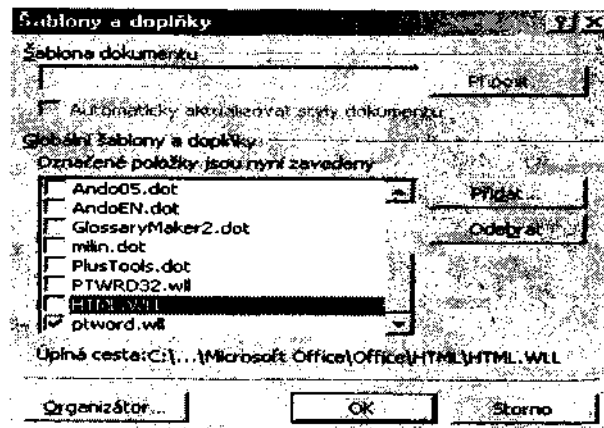
PARS by Lingvistica '98 Inc. is a typical MT system, or, to put it more exactly, a family of MT systems. In PARS, the *pars97.dot* template must be activated to provide system integration with MS Word.

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.

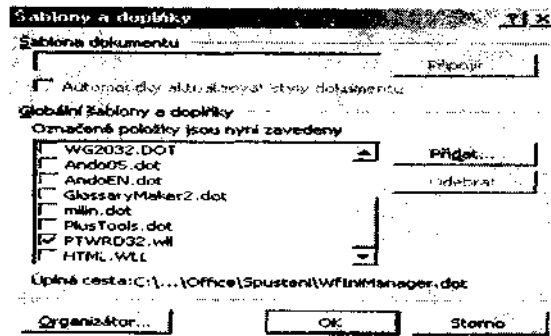
MT systems integrated in the segment mode

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.

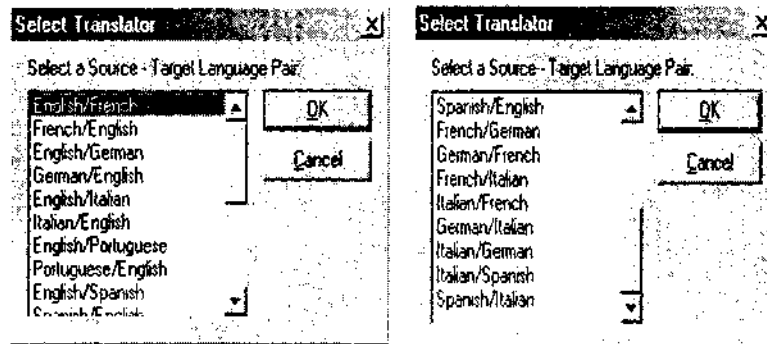


More language pairs are offered, for example, by L&H Power Translator Pro 7. That system produces draft translations of foreign language documents and is integrated with Microsoft Office by means of supplement *PRWRD32.wll*.



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.bmssoftware.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.

MILAN CONDAK
KRATKA 988, 735 41 PETRVALD
CZECH REPUBLIC
E-MAIL: mitan.condak@wo.cz
www.condak.cz