[*Translating and the Computer 26*, November 2004 [London: Aslib, 2004]

Problems with CAT tools related to translations into Central and Eastern European languages

Andrzej Nedoma, Lido-Lang Technical Translations

Andrzej_Nedoma@lidolang.com

Jurek Nedoma, Lido-Lang Technical Translations

Jurek_Nedoma@lidolang.com

Keywords: *CAT tools, CAT tools problems, Central and Eastern Europe, Central and Eastern European languages, Slavonic languages, Translation into Slavonic languages.*

This lecture has been prepared based upon Lido-Lang Technical Translations' experience in the use of CAT tools. Our office has been working with Trados since 2001. In 2002 we started using SDLX as a second CAT tool. Less frequently we deal with DejaVu, Transit and/or WordFast.

I. Analysis of the present knowledge and usage situation of CAT tools in Eastern European countries.

Unfortunately, no official statistics data is available with regard to the application of CAT tools in the translation industry. Therefore, our observations are rather subjective. However, we are convinced that our conclusions present the reality quite well.

Translators

As all Language Service Providers, we constantly receive the applications of freelancers from different countries, among them from Central and Eastern Europe - let's focus on these. One of the basic questions asked in our initial questionnaire refers to the experience in Computer-Aided Translation tools. Having analysed the responses obtained, we can formulate the following general conclusions:

1. Ca. 75% of responses are negative - candidates have very limited knowledge about CAT tools or they have heard about Trados (or CAT tools in general) but they do not use it yet.
2. Ca. 15% of candidates use Trados or Wordfast
3. Ca. 10% of candidates use Trados and SDLX and/or other tools

So the percentage of negative responses is relatively very high. There are two most frequently mentioned reasons for that:

a) Costs of CAT tools - average salaries in developed countries are much higher than in the 10 "new" EU member-countries, therefore prices of CAT tools are seen as very high in Central/Eastern Europe. Even if a translator has heard about CAT tools, he/she cannot afford to buy the licence.

b) Confusion of CAT tools with automatic translating machines. Translators do not believe automatic translation can be of any help and are proud to say that they do everything themselves (!).

Translation Agencies

Among the translation agencies (in the "new" EU member-countries) that we have visited personally or contacted in another way, the percentage of offices familiar with Trados, SDLX, DejaVu, etc. is less than 20%.

A completely different picture arises from the Internet survey - among the translation companies active in the Internet, the percentage of CAT tools users is ca. 60-70%.

However, more detailed discussion with managers of such agencies shows that in many cases the managing staff does not have any knowledge about CAT tools. These managers simply know that e.g. Trados exists and that certain of "their" freelancers can work with Trados. It means that the company itself cannot add any value to the product delivered by the translator. Some agencies simply act as a "contact box" transmitting messages and forwarding files between the translator and the customer.

Such a situation obliges us first to convince our freelancers of the merits of CAT tools and then to train them, to make them efficient in working with the new tools. On the one hand it costs us time and money. But on the other hand, such a policy creates a closer relationship between our company and "our" freelancers and gives us an opportunity to prepare our suppliers for more efficient work in state-of-the-art technologies.

II. Why CAT tools cause problems?

This chapter may - to some extent - seem very evident to a part of the audience. However, all the problems discussed below appeared while processing real projects for our Western European customers. Real questions asked by our clients and our explanations related to many aspects of processing projects convinced us that those problems are still not evident nor generally obvious. They still cause troubles to many of us - Language Service Providers.

We all are already well aware of the fact that CAT tools became a MUST in the translation industry. Not only do these tools raise the efficiency and quality of work, but they also enable considerable reductions in project costs.

Nevertheless, the application of CAT tools is linked with many problems that everyone should be aware of, before acceptance and during the entire process of performance of translation projects.

We have classified these problems into four groups:

- alphabetical problems
- grammatical problems
- linguistic problems
- technical problems

1) Alphabetical problems

"...please do not send Bulgarian in Cyrillic, I need it in Latin alphabet" - recently asked seriously by our respected customer.

It is very important to make our clients aware that the languages of Central and Eastern Europe have many special characters in the "common" Latin alphabet. Several languages use non-Latin alphabets. And also in the "common" Cyrillic alphabet, there exist special

characters used in some Cyrillic-written languages. Therefore, it must be stated at once whether a client's requirements regarding alphabet are logical or not. Knowledge of alphabets and their special characters is also very important during the post-translation processing of files, to check that the special characters were not corrupted. Table 1 presents a list of languages with indications about their used alphabet(s) and special characters.

Table 1. Alphabets and specials characters in Central / Eastern Europe

Language	Alphabet			Additional (special) characters
	Latin	Cyrillic	Other	
Belorussian		xxx		-
Bulgarian		xxx		-
Croatian	xxx			č ć đ š ž
Czech	xxx			á č ď é ě í ň ó ř š ť ú ů ý ž
Estonian	xxx			ä õ ō š ü ž
Greek			xxx	-
Hungarian	xxx			á é í ó ö ő ú ü ű
Latvian	xxx			ā č ē ģ ī ķ ļ ņ ō š ū ž
Lithuanian	xxx			ą č ę ė į š ų ū ž
Macedonian[1]		xxx		ѓ ѕ ј њ ќ џ
Maltese	xxx			ċ ġ ħ ż
Moldavian	xxx			ă â î ş ţ
Polish	xxx			ą ć ę ł ń ó ś ź ż
Romanian	xxx			ă â î ş ţ
Russian		xxx		-
Serbian[2]	xxx	xxx		č ć đ š ž ђ ј љ њ ћ џ
Slovak	xxx			á ä č ď é í ľ ň ó ř š ť ú ý ž
Slovenian	xxx			č š ž
Ukrainian		xxx		ґ є і ї

[1] It is worth noting that the standard MSWord package does not have the Macedonian alphabet in basic fonts (Times, Arial). This language is written in the Cyrillic alphabet and it requires installation of special Macedonian fonts. Without these fonts installed, the files are opened with the Latin alphabet.

[2] Serbian is written using the Latin or the Cyrillic alphabet. Both versions are correct, but recently Cyrillic has prevailed.

2) Grammar problems

2.1 "... *can we have full matches for free this time?"*

In order to understand properly the problem of full matches in Slavonic languages, one should have a quick look at the grammar rules in those languages, especially at the sentence structure, the type of suffixes playing a crucial role in declination, conjugation, etc. etc.

It's neither possible nor necessary to analyse the entire grammar here but we would like to give just a few examples showing how much more complicated Slavonic languages are in comparison with English. Consequently we will show why English 100% matches will not be that easy in Slavonic languages.

Example 1: Let's have a look at how different contexts entail changes in word form. The first sentence in our example enumerates several possible machining operations. Below there are titles of manual chapters in which these machining operations are discussed in detail. Of course in English the form of these words does not change, in Polish it does:

ENG		POL	
Details can be machined in the operations of: • turning. • threading. • milling.		Detale mogą być obrabiane w operacjach: • toczeni**a**, • gwintowani**a**, • frezowani**a**.	
Turning	(as a title)	Toczeni**e**	(as a title)
Threading	(as a title)	Gwintowani**e**	(as a title)
Milling	(as a title)	Frezowani**e**	(as a title)

The Trados analysis of the source file would report internal repetitions/full match in this example. But the words: **turning, threading** and **milling** in Polish have different grammar forms in these examples. The situation is analogical in all other Slavic languages: Czech, Slovak, Russian, Ukrainian, Bulgarian, Slovenian, Serbian, Croatian, Macedonian and Belorussian. It means that this 100% English match MUST be reviewed and changed in Polish (depending of the context).

Example 2: Let's analyse quickly plural forms in Slavonic languages.

ENG	POL	RUS
1 symbol	1 symbol	1 символ, 21 символ**л** 31 символ**л**
2 (and more) symbols	2 (3, 4) symbol**e** 22 (23, 24) symbol**e** 32 (33, 34, etc.) symbol**e**	2 (3, 4) символ**ла** 22 (23, 24) символ**ла** 32 (33, 34, etc.) символ**ла**
	5 (6 - 21) symbol**i** 25 (26 - 31) symbol**i** 35 (36 - 41, etc.) symbol**i**	5 (6 - 20) символ**лов** 25 (26 - 30) символ**лов** 35 (36 - 40, etc.) символ**лов**

The English language has only **one Singular** and **one Plural** form, whereas in Russian, Polish (and all other Slavonic languages) there is **one Singular** and **TWO Plural** forms.

Additional complication is visible in the column with Russian version. The numerals ended with "1" (e.g. 21, 31 or 1001) are treated in Russian as Singular (!). It means that (literally translating) in Russian one says "1001 symbol" - and not "1001 symbols".

It means that the client should not ask for general translation of segments like "XXX pieces included" and count on "replaceability" of "full" matches with any number instead of XXX.

It happened also that the client asked us to translate several words separately and then wanted to do an automatic translation of segments containing two or three of those words put together.

Another client tried to force us to translate the first half of the sentence and several versions of the endings of the sentence - in that way he wanted to have the liberty of creating a desired phrase.

A general comment for all these (and similar) examples: the Slavonic languages are much more complicated than English.

Please note that in the examples above we showed only the plural forms differences in the same declination case. When we add declination to it (for ex. in Polish there are 7 declination cases), it becomes evident that automation of translation is not possible.

Furthermore in these languages there are feminine, masculine and neutral genders, there are declinations and many other grammatical structures that must be appropriately applied to most of the words in the phrase. It is not possible to combine parts from different sentences because it is very probable that this would also require a change in the grammar structure. Such automatically created sentences would be definitely wrong.

And the following general conclusions can be drawn from these examples: very often in Slavonic languages 100% matches do require special attention. Certainly the TM entries cannot be uncritically used regardless of the context, otherwise critical mistakes may appear. Briefly speaking: the "full-matches" in English do not fully match in Slavonic languages.

Furthermore, in some CAT tools there is a possibility to prepare projects that do not contain internal repetitions - this option MUST NOT be used for the reasons mentioned above.

2.2 Why is proper segmentation so important?

Many times we have obtained for translation a file where sentences were divided into several segments. Normally a sentence structure in Slavonic languages is fairly different from the English one. In that situation source segmentation cannot be reflected in the target language.

Example 3: Let's analyse the translation of the title of our proceedings, written in 2 lines:

ASLIB Conference, segmented by Trados as:

{0>ASLIB<0>ASLIB<0}

{0>Conference<0>Conference<0}

The correct translation into Polish is "Konferencja ASLIB". If the translator will attempt to keep the segments, he/she must write

{0>ASLIB<0>**Konferencja**<0}

{0>Conference<0>**ASLIB**<0}

It is visible that the word "Conference" in Polish (Konferencja) must jump to the first segment. As a result, the TM will include incorrect pairs:

ENG: "ASLIB" = (?) POL: "Konferencja", and

ENG: "Conference" = (?) POL: "ASLIB"

Conclusion - Automatic segmentation of titles (as well as full sentences) into 2 (or more) segments leads to the introduction of incorrect pairs into the TM.

And reciprocally, there is another risk: short sentences or notions segmented into even shorter segments may be incorrectly translated from the TM.

Example 4: "Signature date" (i.e. the date when a document has been signed) should be translated into Polish as "Data podpisu". However in the project we received it was segmented as:

{0>Signature<0} {0>date<0}.

The TM that was used for pre-translation already contained correct translations of segments:

{0>Signature<100>Podpis<0} and {0>date<100>data<0}.

As the final effect "Signature date" was automatically translated as "Podpis data" which is completely wrong. To make things worse - pre-translated segments were protected against changes.

2.3 Can we expect special reduction of volume because there is a lot of numbers in tables?

Seemingly YES, but in reality NO. This seemingly simple part of a translation project also requires special attention. This is due to different decimal separators between languages - in certain languages there are decimal points, in other languages - decimal commas.

In traditional translations in text formats it is enough to check whether the translator remembered to transcript numbers appropriately. However in CAT tools it may be more complicated. In fact some CAT tools omit numbers as "non-translatables", which causes mistakes in the final documentation. Therefore, it is necessary to be aware of possible mistranslations, to remember about it and to check the correctness of decimal separators at the final stage of DTP review.

The list of proper decimal separators in chosen languages is as follows:

Decimal separator	Language
comma	Belorussian, Czech, Estonian, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Russian, Slovak, Slovene, Spanish, Turkish
point	English

Certain customers ignore this problem and try to dispute, saying that this difference is not important. As a rule the following question asked by ourselves is convincing: Can you imagine the surprise of a manager who wanted to buy 1.000 m^3 of timber and found delivered 1,000 cubic meter of timber (because the translator was informed that he will not be paid for translation of numbers and consequently he "forgot" the difference in decimal separators).

3) Linguistic problems

Why reference materials are so important

Projects prepared in CAT tools often do not show the real structure of the text, its final layout, drawings etc. In fact translators work on pure text presented in a way that does not help to solve linguistic problems that appear during translation.

Therefore providing a reference documentation (e.g. in the form of original source files) constitutes customer's interest as a measure aimed at assuring the highest quality of translation.

Frequently we receive spot projects in which we should translate just 3 or 5 words. The text is already translated into English, Italian and French (for example) and we should add translation into Central and Eastern European languages. What appears from a received file is that all translations already inserted might have been done literally (without knowledge of the context) thanks to the similarity of many Western European languages. However in Slavonic languages we encounter dilemmas about the real meaning of the source text. Let's take an example of an abbreviation.

Example 5: The English abbreviation "int." could be translated into Italian, French or Spanish as "int." regardless of the actual meaning - whether it means (1) international, (2) internal or (3) interval. In Polish, however, depending on actual meaning we should write:

(1) "m-nar." (międzynarodowy), (2) "wewn." (wewnętrzny), (3) "int." (interwał).

The situation is similar in other Slavonic languages. Translators need help to decide, which translation is correct and this help absolutely must be provided by the customer in a form of short explanations, reference materials, etc.

4) Technical problems

There exist also several technical problems which appear in the post-translation processing of files. Most common are hyphenation problems, the already mentioned font problems and text-length related problems

None of these problems arises during translation itself in CAT tools. The translator works normally in a simple text environment, separated from final layout aspects (hyphenation, text flow on the page, etc.). Assuming that each translation is done by a native speaker of the target language, working in his/her localised version of the system (as always in our Agency activity), specific fonts do not cause any trouble at all.

All troubles blossom at the stage of importing translation to the final format: the software applies automatic rules of hyphenation (word breaking), special characters do not pass through correctly and (as it usually happens in Slavonic languages) translation overflows its text box area.

Why does it happen?

4.1 Hyphenation

Problems with hyphenation are normally caused by automation of processes. As a rule, the final user of the translation (a printing company) does not realise that the English hyphenation module of their software is not adequate for the foreign version of the text. Therefore, in the final document there appear lots of wrong word breakings, which is very evident and irritating for the final readers of the document.

In text received for final quick review (in a PDF format, as a rule) we also find many words that are separated with hyphen/dash, even though they are in the middle of a printed row. This results from leaving hyphenation tags during preparation of files for translation. The translator (asked to take care on all tags in his translation) leaves these hyphens in his translation - as hidden in certain tags - and the final effect is very bad.

To avoid such problems it is necessary to switch off the hyphenation during export/import of files as well as to ensure that final files (exported to PDF) are quickly checked by a native reviewer for any hyphenation mistakes at the very end of the process.

4.2 Fonts

Font problems may appear during processing target files on a system working in another language than the target translation. Problems with fonts were already mentioned above in the section describing alphabet-related issues.

A very important aspect is the occurrence of special characters in these alphabets - all letters with accents, hooks, etc. The problem may appear during post-translation processing of files that such characters become corrupted. Therefore it is necessary to know the list of characters in each target language to be able to check whether all characters are displayed correctly. A full list of all characters in the discussed languages is attached at the end of this paper.

4.3 Text length

Problems of the length of translated text in comparison with source text also require a short explanation.

It is important to remember that the texts translated into such languages as e.g. Polish, Czech, Slovak, Slovenian, Hungarian, Russian, Greek are in general some 10-20% longer than respective English source text. This fact often forces the DTP department to change font sizes to match the original page layout or to change the number of pages. In many cases it is impossible to keep both the layout and font size and to fit the text in source text boxes.

The problem of text length in different languages is also a very important aspect in the localisation of display screens, fault messages etc. Clients often force us to shorten the translated text in order to keep a given number of characters. It is crucial to understand that Central and Eastern European languages are longer and less concise than English, so respecting string length is problematic. Abbreviating translation too much entails the risk of reducing overall comprehension.

Ideally the end user should have in mind these objective circumstances and take them into consideration when identifying the spaces for individual inscriptions. If the author does not make necessary space reservation, a compromise between general understanding and keeping predefined length is not feasible in many cases.

III. CONCLUSIONS

In the present paper we have analysed only a few examples of several types of problems. It is not possible to mention all of them here; however, we believe that this paper provides a general idea of possible difficulties one may encounter during execution of translation projects into Central and Eastern European languages. All the Language Service Providers should be aware of these problems and should be able to explain them to their clients.

Unfortunately, projects often pass through many intermediaries, who do not feel themselves obliged to participate in solving these problems. This may complicate project management and may cause incorrect preparation of the project, as well as many accompanying problems caused by a damaged, over extended communication path.

The only solution to avoid that unproductive situation, is to avoid mid-agencies, acting as contact-boxes only. The best results (optimal quality, shortened time of handling projects) can be achieved if the customer co-operates with an aware agency, which can solve in-house the majority of problems, without endlessly forwarding e-mails between the customer and the translator - i.e. an agency able to add real value to the translator's / proof reader's / editor's / reviewer's efforts.