10 Golden Rules for Creating Translation-Oriented Source Documents

Here are some tips for preparing source documents optimally for translation even during their initial creation. A well-prepared and cleanly-formatted source document can save a lot of time and money during translation with a translation memory system (TM system) since the recognition capabilities of the TM system only make sense if the segments to be translated are actually identical or similar.
The following rules are derived from practical experience with Across. At first they may appear insignificant to the author of a text, but for the translator, some of these things present large problems.

1 PDF Files vs. Original File Formats

Whenever possible, avoid using PDF files as the source document format for translation. Always try to provide the original file format that served as the basis for the creation of the PDF files since PDF files cannot currently be edited in Across; instead, they have to be transformed into another format (usually Word) before translation. The transformed documents must generally be edited again before translation since the converted text usually contains too many formatting errors to be able to translate it sensibly with a TM system. This editing is always associated with additional time and costs and delays the start of the translation.

2 Hard Line Breaks

Avoid hard line breaks (paragraph marks) within sentences; otherwise, no sensible segments can be offered for translation. Line breaks should only be used if a new paragraph should actually be started. TM systems decide using segment end limiters where a translation unit (normally a sentence) ends. These characters are generally **.**, **!**, **?**, and **¶**.
A line break is always detected as a segment end and manual editing is required if the line break is within a sentence and subdivides it into two segments as a result. Manual adaptation by the translator requires additional time, and the initial analysis will find fewer hits in the TM (matches), which will make the translation unnecessarily more expensive.
Frequently, hard line breaks in PowerPoint or in DTP programs are put in the wrong place because translators do not have sufficient knowledge about how to work with these programs.

Here is an example of the display in crossDesk of a sentence that contains an incorrect line break:

In·the·middle·of·this·first·sentence·there·is·an·incorrect·line·break·
because the text was copied from a PDF file. The sentence is therefore subdivided into two illogical segments as a result. Manual adaptation by the translator is required.

3 Soft Line Breaks

Soft line breaks (**Ctrl+Enter**) should also be avoided. TM systems do not interpret them as segment ends, which is why such units are not detected correctly and they have to be re-worked manually by the translator.
The following text sample contains a soft line break at the end of each bullet point; this causes the whole text to be offered as a single unit for translation in crossDesk.

Please · check · the · following:↵
-·spell · check↵
-·formatting↵
-·dates · and · numbers↵
-·terminology

Soft line breaks are frequently inserted unintentionally by copying texts from various applications into source documents. This happens quite frequently if the text to be translated is copied from an e-mail into a Word document, for example.

4 Manual Page Break

Very often manual page breaks are inserted for formatting purposes, e.g. because a headline falls at the bottom of a page. To improve the layout and the readability of the text, the author inserts a manual page break at a specific location. However, during translation texts usually grow or shrink in length depending on the language combination, so it is very unlikely that the manual page breaks from the source text should be placed in the same location as in the target text.
Usually the manual page breaks are not “translated” but are skipped during translation and they are inserted into the final version of the text after the translation is finished and the text is converted back into its original document format.

5 Blank Spaces and Tabs for Formatting

Try to use tabs or indents to indent texts and do not use a series of blank spaces to do this. After reading the document into a TM system, these characters are all displayed and the display in the translation editor is confusing on the one hand; on the other hand, such manual adaptation is oriented towards the source language. In 99% of cases, the translation with the same blank spaces will look different than it does in the source document. The text then has to be re-worked after the translation in nearly every case.

Sample text in Word:

Bulleted·list·with·a·series·of·blank·spaces·instead·of·tabs·or·indents¶
----This·sentence·is·to·demonstrate·how·a·text·looks·like·in·crossDesk↵
----if·a·series·of·white·spaces·is·inserted·instead·of·tabs·or·indents¶

The same sample text in crossDesk:

-····This · sentence · is · to · demonstrate · how · a · text · looks · like · in · crossDesk↵
·····if · a · series · of · white · spaces · is · inserted · instead · of · tabs · or · indents

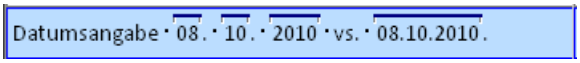
While working, a translator can only assess with difficulty whether the blank spaces actually have a particular function or whether they serve only formatting purposes. If we assume, for example, that the translator does not delete the blank spaces from the translation and tries to put them in same place in the target text as in the source document, then after the export, the translation would look like this:

Aufzählung mit Leerzeichen statt Tabulatoren oder Einzügen
- Dieser Satz soll zeigen, wie eine Text in crossDesk aussieht, wenn man mehrere Leerzeichen statt Tabulatoren oder Einzügen verwendet.

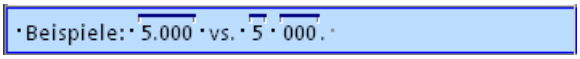
6 Date, Time, and Number Formats

For the detection of date, time, and number formats, TM systems orient themselves according to specified rules. Thus in the Across system settings, it is specified whether date formats in German have the format DD.MM.YYYY or DD.MM.YY. If there is a blank space between the numbers, the date is no longer recognized as a coherent number group, and it cannot be checked for correct usage in the translation. This happens frequently with dates and numbers in the thousands.

Example: 08.10.2010 vs. 08. 10. 2010



Example: 5.000 vs. 5 000

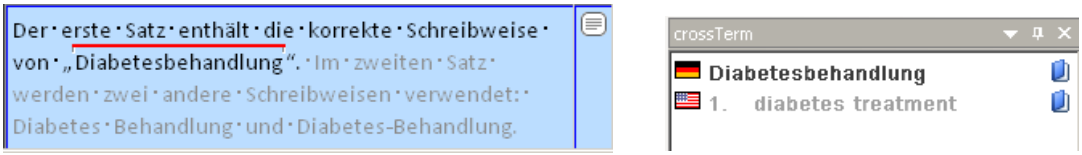


The blue lines indicate to the translator that there is a number and which number areas have been detected as coherent units. If the blue line is interrupted, several units were detected. In this case, no sensible checking can be done to ensure the correct takeover of the number formats.
It is therefore recommended that for texts that contain a lot of date, time, and number formats, you specify a uniform format and use it consistently.

7 Uniform Spelling of Specialized Terminology

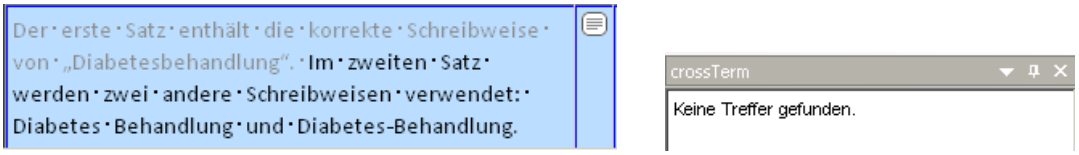
The uniform spelling of specialized terminology is essential for correct terminology detection. In German, for example, writing a term as one word (Diabetesbehandlung), as two words (Diabetes Behandlung) or connecting two words with a hyphen (Diabetes-Behandlung) is a frequent cause of inconsistent translation since the automatic terminology detection is not activated in these cases.

Example of correct spelling:



The red marking indicates that the term is present in the crossTerm terminology database. The stored translation is suggested to the translator. The translator can then take over this suggestion directly into the translation.

Example of alternative/incorrect spelling: vs.



The system does not recognize the specialized term and the terminology window remains empty. Thus the translator does not know that there is existing terminology information and he may, in case of doubt, translate the term inconsistently or incorrectly.

8 Usage of Correct Abbreviations

Always use the correct spelling of abbreviations. If you use different spellings for one and the same word, the segmentation of the sentences will be incorrect. For example if you look up the word “approximately,” you will find app., approx., apx. as abbreviations. In Across app. and approx. are defined as abbreviations for approximately, but if you use apx. the sentence will be subdivided in two segments and manual adaptation by the translator is required.

9 Superfluous Formatting

If you work frequently with colored marking in the text in order to visually emphasize text passages, you should make sure that this has been removed completely before translation and that there are no line breaks and blank spaces remaining to which this formatting is assigned. Otherwise, this "invisible" formatting is offered to the translator as possible formatting and may result in the translator writing directly in the wrong font/color.

Example from MS Word:

At the end of the paragraph is a superfluous formatting mark ¶

10 Hyphenation

If you want to use hyphenation in your text, make sure that you either use the “automatic hyphenation” function or insert an optional hyphen manually. A lot of people just insert a normal hyphen instead of using the automatic hyphenation or the manually inserted optional hyphen. In this case the translator and the translation memory system face the following problems:
Standard hyphens are recognized as normal characters and add an additional character to the word in which they are placed. On the one hand this means that the translation unit stored in the translation memory will not be a 100% match even if exactly the same sentence appears again without hyphenation. On the other hand, terminology recognition for this specific term will not work since the term is separated by an incorrect character.

First sample sentence with correct hyphenation in Word:

The first sentence contains an optional hyphen¬ation that was inserted manually. ¶

First sample sentence with correct hyphenation in Across. Term was recognized by the system.

The · first · sentence · contains · an · optional · hyphen¬ation · that · was · inserted · manually. ·

Second sample sentence with incorrect hyphenation in Word:

The second sentence just contains a hyphen in the middle of the word hyphen-ation. ¶

Second sample sentence with incorrect hyphenation in Across. Term was not recognized by the system.

The · second · sentence · just · contains · a · hyphen · in · the · middle · of · the · word · hyphen-ation.