MS WORD AND ITS HIDDEN SECRETS

HOW TO MAKE AN MS WORD FILE MORE TRANSLATION FRIENDLY

INTRODUCTION:

In general, MS Word is not the best file format for localization. While you can be quite creative and give your content a nice look and feel, by introducing too many edits the underlying code becomes very messy. This has a negative impact on localization. This paper explains why.

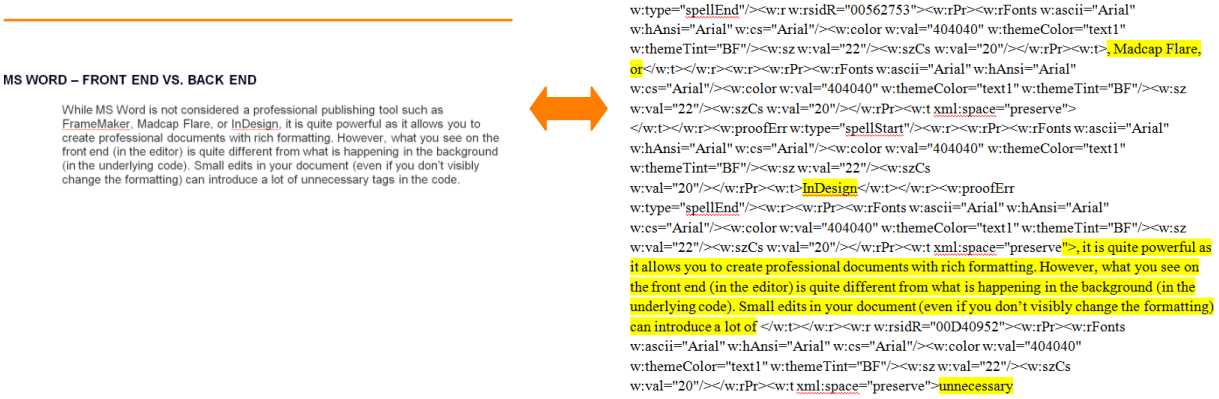
If you enter MS Word through the back door, you can optimize the code for translation. EzGlobe has developed a solution to clean up the code and thus minimize the impact of messiness on translation and translation reuse.

KEYWORDS:

MS Word, Microsoft Word, XML, translation file format, translation memory

MS WORD – FRONT END VS. BACK END

While MS Word is not considered a professional publishing tool such as FrameMaker, Madcap Flare, or InDesign, it is quite powerful as it allows you to create professional documents with rich formatting. However, what you see on the front end (in the editor) is quite different from what is happening in the background (in the underlying code). Small edits in your document (even if you don't visibly change the formatting) can introduce a lot of unnecessary tags in the code.

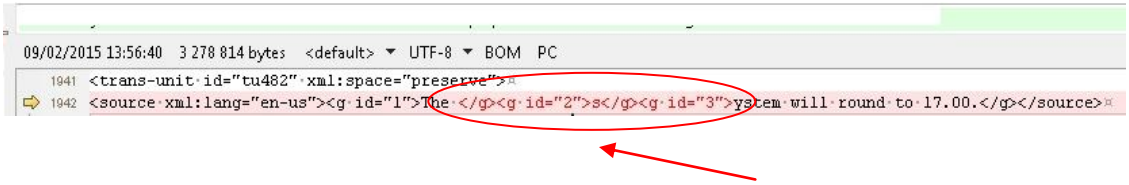


WORKING WITH TRANSLATION MEMORY

If we want to translate an MS Word file using a computer-aided translation tool (also known as "translation memory") we cannot use the file "as is". The translation memory works in conjunction with text-based files in which we can separate the content (text to translate) from the format (formatting information in form of tags). This is why we always convert rich files to text-best files. The best format to use is the XML version of MS Word.

In general, XML is a great format for translation because it is very structured and can be easily parsed. However, XML files generated from MS Word are often very messy. Each time you edit your MS Word document, MS Word adds formatting information to the underlying code even if this information is not always useful and pollutes the XML. Most of the time, you end up with unnecessary tags within individual sentences or even words.

The following graphic shows how edits change the underlying code of this simple sentence: "The system will round to 17.00." The writer edited the word "system" and Word added tags within the word itself.



NOTE ON MS WORD EDITING

In MS Word you can apply style or format not only to parts of sentences but also parts of individual words. Even if you apply **the same** style or format to the same word but do it in two steps, then MS Word will add additional tags.

These tags are found in the middle of the word "system". They break the sentence down to multiple fragments:

"The" + "s" + "ystem will round to 17.00."

These tags do not add any value and can be safely removed.

WHAT DOES THIS REALLY MEAN?

When sentences are broken down to fragments you will face some localization problems:

1. Text with excessive tags is not very translation friendly as it is difficult for the translators to see a sentence as a whole.

2. The tags break down the sentences to fragments that are then stored, as such, in the translation memory. These fragments have very little use when it comes to reusing the translation memory especially on different documents and/or in different formats.

3. There is a high risk of errors. A slight change to a tag can break the file and make the output almost impossible to generate.

NOTE ON TRANSLATION MEMORY

Very simply said, translation memory is a database of translated segments (sentences) that can be reused.

So, if your new text contains the same sentence that has been translated previously, translation memory will replace that sentence with the stored translation. However, if your translation memory contains segments that are not logical (such as the example above) it will be of a very little use. The tool will not find this sentence:
"The system will round to 17.00.".

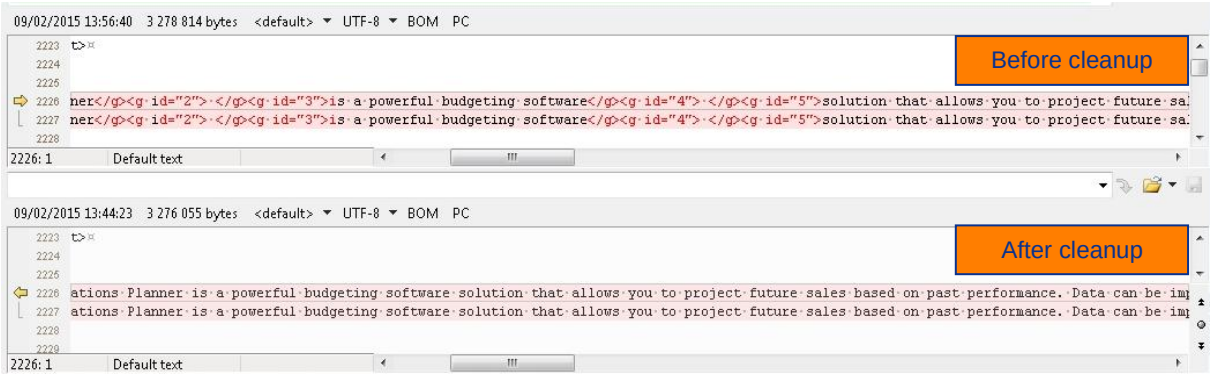
EZGLOBE'S SOLUTION AND BENEFITS

At EzGlobe, we translate thousands of MS Word files. Indeed, for some of our clients MS Word is a practical solution if they don't have the extra resources or manpower to invest into professional publishing tools.

Motivated by the demand we have invested into finding solutions that better fit the translation environment of documents authored in MS Word.

Since it was the XML file generated from the MS Word document that was problematic (messy), **we developed a programmatic solution that cleans up the XML file and verifies its integrity** (i.e. makes sure none of the formatting has been lost or altered). Simply said, it removes those tags that are not necessary and gives us a cleaner XML that can be used more efficiently in conjunction with the translation memory.

The following image shows an excerpt from an MS Word-generated XML file before and after the cleanup.



On average, EzGlobe's solution decreases the number of tags by 90% and yields much cleaner files to translate. The benefits are clear:

1. Easier translation

The translator can read the text easily and thus produce better translation quicker. The overall quality of the translated document increases.

2. Higher reuse

When complete sentences (as opposed to sentence fragments) as stored in the translation memory, we can reuse them more efficiently. This decreases your translation cost.

3. Smaller and lighter files

When complete sentences (as opposed to sentence fragments) as stored in the translation memory, we can reuse them more efficiently. This decreases your translation cost.

4. Decreased risk of file corruption

Less tag manipulation means less risk that tags will be inadvertently deleted or altered.

So how much can you save?

Request a FREE evaluation of your doc.

Contact EzGlobe

www.ezglobe.com