Avoiding Segmentation Issues

June 29, 2021

www.idemtranslations.com

It's hard to get excited about segmentation. So, instead, let's talk about saving money!

You want your translations to be consistent over time, right? But you shouldn't have to pay for your translation provider to translate the exact same content over and over again. This is why professional translators use computer-assisted translation (CAT) tools.

CAT tools are a working interface where translators can work on translating your content into their language while connecting to a repository called a "Translation Memory." A Translation Memory is a database that is built, over time, by adding each bit of translated content as it is completed. Translation providers typically create one Translation Memory for each language pair (e.g., English into German) for each client. What this means is that all your content lives together, subdivided into one miniature database for each language combination that you ask your provider to translate for you.

But what does that database look like? There must be some way to break your documents down into the right-sized pieces to store in this Translation Memory, right? This is where segmentation comes in. CAT tools parse your source documents into individual segments for translators to work on individually, so that each segment can be stored in the Translation Memory.

This executive brief will introduce you to how CAT tools decide what constitutes a segment, the risks associated with bad segmentation, and what you can do to save your organization money by eliminating or reducing those segmentation risks.

What is a segment?

It's best to think of a segment as a minimal unit of logical meaning. So that sentence you just read? It's a segment. So is the header above it.

CAT tools define segments based on delimiters, which are specific to each kind of document that gets fed into the tool. The delimiters are different for software code, website content, and formatted documents. This brief will focus on the delimiters used in formatted documents as these tend to cause greater risks to the translation process than standardized coding languages used for software and websites. Be aware, however, that CAT tools also parse the strings for your website or software GUI in a similar manner, so the main themes of this executive brief are still relevant for those projects, albeit with some minor nuances.

How do CAT tools segment a document?

CAT tools use surprisingly simple rules to define a segment. They look for any of the following and treat what comes before it and what comes after it as two separate segments:

- hard return ¶
- period/full stop .
- colon :
- semicolon ;

These rules translate into each of the following being considered a segment:

- a heading
- a sentence
- a standalone clause within a sentence, separated by either a colon or a semicolon
- each cell in a table

Here is an example: the table below shows what this section of this executive brief would look like when parsed into a CAT tool. The left-hand column is for the parsed source content; the right-hand side is blank for a hypothetical translator to input the associated translation.

Source	Translation
CAT tools use surprisingly simple rules to define a segment.	
They look for any of the following and treat what comes before it and what comes after it as two separate segments:	
hard return ¶	
period/full stop .	
colon :	
semicolon ;	
These rules translate into each of the following being considered a segment:	
a heading	
a sentence	
a standalone clause within a sentence, separated by either a colon or a semicolon	
each cell in a table	
Here is an example:	
the table below shows what this section of this executive brief would look like when parsed into a CAT tool.	
The left-hand column is for the parsed source content;	
the righthand side is blank for a hypothetical translator to input the associated translation.	

What are the risks to bad segmentation?

Different languages put words in different orders, but they share a common baseline for what constitutes a minimal unit of logical meaning. Bad segmentation is risky when it cuts apart (or mushes together) content in a way that is grammatically illogical, forcing translators to pick between a rock and a hard place. Do they try to match each half of a split segment with something semi-equivalent to the source at the expense of proper grammar in their native language? Or do they translate with proper grammar knowing that the two segments that will ultimately be fed into the Translation Memory don't actually match the funky halves of the split source? Let's look at a few examples.

Example 1

Average¶
Whole-Body SAR

Why it's problematic: Many languages put modifiers, like "average," after the noun they modify. If the translator handles each of these two segments individually, the translation will be grammatically incorrect when it is output from the CAT tool. If the translator ensures that the output displays in a grammatically correct fashion, then the two segments being added to the Translation Memory will be incorrect (because the translation for "Whole-Body SAR" will be in the space for "Average" and vice versa).

How to fix it: Never put a hard return in a table heading, document heading, or text box. Adjust display of the English with spacing or, in a pinch, a soft return.

Example 2

Carefully read these--SECTION BREAK--
instructions for use.

Why it's problematic: That section break operates like a hard return, so the sentence is cut in two, offering the translator two separate segments to work with and forcing them into a similar dilemma to the one described to the left: choosing between content that will output correctly in the target document versus accurate segments to feed the Translation Memory.

How to fix it: Never separate content with a break (page, section, column, or otherwise); only insert such a break between sentences.

Another segmentation risk appears when ***too much*** information gets into the same segment. The culprit in this case is nearly always the use of tabs to align English content. CAT tools do not treat tabs as a delimiter and, when they output the translated dcoument, they retain all the tabs used in the source. However, since the words are different lengths in different languages, this means that additional formatting is required to ensure a professional, presentable document.

Example 3

Principal Investigator:--tab--tab--Dr. Smith
Address:-tab--tab--tab--tab--tab- 123 Acme Road

Why it's problematic: Dr. Smith's name and address align nicely in English because the sum of "Principal Investigator" plus two tabs is equal in length to the sum of "Address" plus five tabs. But the translations for "Principal Investigator" and "Address" will have a different number of letters in translation, skewing the formatting.

How to fix it: Favor tables over tabs. It's easier to get a gorgeous alignment with tables and you can simply hide the lines so they are not visible to your readers. Tables also work beautifully for signature sections as partial borders allow you to perfectly align the descriptor with the input line (e.g. signature, date, etc.).

What can you do to save money?

Now that you understand the risks, what can you do to avoid these problems and reduce the long-term costs associated with managing your translations?

DON'T do this...	...DO this instead!
Use tabs to align content.	Use tables to align content; show/hide the borders as necessary. (This works in headers/footers, too!)
Place content into textboxes that might split over two pages.	If you need to format a block of content so it pops (e.g., with a different color background), use a single-cell table that is allowed to break across pages.
Insert hard returns to force text-wrapping.	Adjust margins or, in a pinch, use a soft return.
Insert page breaks to force text onto the next page.	Select the text that should move to the next page and use "Keep with next" type commands so that it hugs the text that follows it.

The rigor that you put into creating clear, well-organized, and consistent source documentation pays dividends when it comes time to translate. This is not only true of the actual content (since clear content is more easily translated), but equally true of how you choose to format a given document. The tips above will ensure that CAT tools process and parse your documents properly, thereby providing your translators with clear, unambiguous segments to translate. In the long-term, these clear segments and your own consistency in using best practices in formatting will increase the value of your Translation Memory, enabling your translation provider to re-use past translations and charge you less.

If you found this executive brief beneficial, you may also enjoy our briefs on:

- building glossaries and style guides,
- the anatomy of a multilingual IFU,
- best practices for Translation Memories, and
- when to transition to a full desktop publishing application.

About Idem Translations

Founded in 1983, Idem Translations, Inc. is a full-service provider of translation and localization services. Idem specializes in certified translations for medical device, biomedical, and pharmaceutical companies, as well as other organizations and entities working in the life sciences sector, such as contract research organizations, healthcare research centers, and institutional review boards. The company is a WBENC-certified woman-owned business and holds certifications to ISO 9001:2015, ISO 13485:2016, and ISO 17100:2015.

Get Help

For more information about how we can take the risk out of translations for you and your team, please visit us online:

WEBSITE
www.idemtranslations.com

TWITTER
twitter.com/IdemTransInc

LINKEDIN
www.linkedin.com/company/143474