

On time...on target...on budget

The Importance and Pitfalls of Accurate Word Counting

On time...on target...on budget

The Importance and Pitfalls of Accurate Word Counting

The purpose of this document is to provide our customers with a better understanding of the difficulty associated with establishing accurate document word counts.

From a word counting perspective, there are two types of documents:

- Documents that can be readily counted electronically
- Documents that cannot be counted electronically or are difficult to count electronically

Whenever possible, Apex bases pricing on the electronically established word count of the source document. The obvious advantage of counting the words of the source text lies in the fact that this approach provides a firm basis for cost assessment upfront, i.e., before a financial commitment is made. It is essential that the word count be accurate and verifiable.

If the word count of a source document cannot be reliably established, Apex bases quotations on an estimate of the word count in the target document, i.e., the translation.

A. Documents that can be readily counted electronically:

1.) Microsoft Word (DOC, RTF, WPS, etc.)

DOC, RTF, and WPS file types are the most common file-formats available. MS Word comes with a built-in word counting module. Unfortunately, this module is severely limited in its ability to accurately determine the actual number of translatable words in a document. For one, the only two delimiters (characters that indicate the separation between two words) that MS Word recognizes are tabs and spaces. There exists a number of other characters that could well be considered delimiters, such as the following: .,;:\/(){}[]*+&!'"~<>=

The word counting delimiters cannot be reconfigured in MS Word.

In addition, MS Word counts the text repeatedly in headers and footers, even if this text is 100 percent identical on each page. There is no way to configure MS Word to count repetitive headers and footers only once per document or section.

MS Word also does not exclude numerals from its word counts. Apex's policy is to exclude numerals from the billable word count for all projects, because they do not require translation - although in some cases numerals may require conversion into the appropriate notation.

Most importantly, MS Word does not count any words in text boxes!

On time...on target...on budget

Because of the significant shortcomings of MS Word's word counting module, Apex currently uses PractiCount (<http://practiline.com>) as our primary word counting software tool. PractiCount provides a panoply of options for configuring delimiters and including/excluding headers, footers, comments, annotations, textboxes, numerals, and hidden text. These options allow for a much more accurate word count that benefits all involved in the translation process.

2.) Other Common Editable File Formats (XLS, CSV, PPT, TXT, etc.)

Apex currently uses PractiCount (<http://practiline.com>) to count the words in the above-mentioned file types.

3.) PDF Documents

Not all PDF documents are electronically countable. To establish whether a PDF is countable, select the text in the document using the "Select All" feature in Adobe Acrobat. If the text can be highlighted, the document is very likely countable. Countable PDFs can be counted with third-party word counting tools such as PractiCount (<http://www.practiline.com>), AnyCount (<http://www.anycount.com>), or using a plug-in counter for Adobe Acrobat available at <http://www.intellipdf.com/stat.htm>.

Please note that word counts obtained from PDFs should always be regarded with some suspicion and, if possible, should be corroborated with a manual count, or a word count established with Optical Character Recognition software (OCR). This is because some PDFs may contain special characters that can be erroneously counted as words.

4.) HTML Pages and Similar File Types (ASP, JSP, PHP, SHTML, etc.)

Apex determines the number of words contained in HTML pages and similar file types using WebBudget XT (<http://webbudget.com>). As with conventional documents, Apex adheres to the policy of excluding numerals by appropriately configuring the delimiters of this software. All translatable text in HTML files is counted, including text found in non-body elements of pages, such as meta, alt, on-mouse, and script text, etc.

Because HTML and similar file types generally contain a significant amount of 100 percent redundant text (text that reoccurs in the same context, commonly found in non-body page elements), Apex attempts to determine the exact number of redundant words. For this purpose, we use the application, DeJaVu X Workgroup Edition (<http://www.atril.com>). Apex charges a lower rate for redundant words.

5.) File Formats Requiring Desktop Publishing/Typesetting (FrameMaker, PageMaker, Illustrator, Photoshop, InDesign, CorelDraw, QuarkXPress, Freehand, etc.)

Typically, for the purpose of generating a cost and turnaround time proposal for the translation of these file types, it suffices to analyze the contents of the documents without actually obtaining the original, editable layout files from customers.

On time...on target...on budget

Layout files are usually quite bulky in terms of the disk space they occupy, and we therefore prefer to analyze a PDF copy (or other image file type) of the document in question. PDFs are preferable because they allow us to conduct electronic word counts in most cases and are far easier to transmit electronically (particularly by email). This means that while translations of the above-mentioned file types are actually produced from an editable layout file, the word count for quotation is most easily obtained from a non-editable, although still countable, PDF copy. Countable PDFs are counted according to item 3.) above.

A second option is to conduct word counts directly in the desktop publishing software. Some DTP applications have built-in word counting modules; others do not. This option of course requires that the particular DTP software that was used to create the document is available, which can be quite expensive and can consume significant disk space. Word counts from PDFs are generally sufficiently reliable and much easier to obtain.

B. Documents that cannot be counted electronically or are difficult to count electronically:

Document types that cannot be counted electronically are: hardcopy documents, faxed documents, and electronic document formats such as PDF documents with scanned content, TIF, GIF, JPG, BMP, etc.

The first course of action is to convert these documents into an editable format using Character Recognition Software (OCR), such as ABBYY FineReader (<http://www.abbyy.com>).

OCR is a process by which the software attempts to recognize the characters in a non-editable image file and match them to known characters, and reconstruct the text in an editable format, which is usually Microsoft Word. Once the text has been saved in MS Word, it can be counted electronically. We recommend using a third-party tool such as PractiCount (<http://practiline.com>), for example, to count MS Word documents generated with OCR applications, since they generally consist almost entirely of text boxes, which MS Word's built-in word counting module does not count.

It should be noted that, depending on the quality of the source document, word counts established by the OCR method are not always accurate and it is recommended that a visual check be performed before accepting the word count.

If the word count of a source document cannot be reliably established using OCR software, Apex bases quotations on an estimate of the word count in the target document, i.e., the translation. This word count is calculated based on the estimated word count of the source document, to which a multiplier is applied to account for text expansion or shrinkage.

Text expansion/shrinkage invariably occurs when translating text from one language into another, although the level of fluctuation varies between language combinations. When translating from English into Turkish, for example, you will likely encounter around a 15 percent increase in the number of words.

When translating from English into Spanish, for example, the number of words increases by roughly 7 percent. If this were the case for a project involving documents that are not electronically countable, the estimated number of words in the non-editable source document would be multiplied by 1.07 to estimate the number of words in the translation. Apex's cost estimate is then based on this derived word count.

Whenever Apex bases a proposal on an estimate of the number of target words, our invoice is adjusted to reflect the actual electronic word count of the translation, which can be established electronically in almost

On time...on target...on budget

White Paper

Lastly, if all of the above-mentioned methods fail, word counts can also be established by manually counting the words contained in a document. This does not always mean that each word needs to be counted off by hand. Rather, one approach for manual counting involves establishing an average number of words per line, as well as an average number of lines per page, using a few representative sample pages. By multiplying the product of these two numbers by the total number of pages, a rough word count estimate can be produced.

Additional Information:

1.) Documents with Edits, Images and Graphics, and Multilingual Documents

For documents that contain edits or comment, Apex's electronic word count will include all words in edits and comments, unless otherwise specified. For this reason, it is important to have finalized documents prepared before beginning the translation process whenever possible to eliminate unnecessary cost and quotation/production time.

If a document contains text in more than one language, it may be impossible or difficult to electronically count only the words in one language, and not in the others. Unless the text can be easily separated by language, a (usually very tedious and time-consuming) manual count may be the only option. Disentangling multi-language texts may also require considerable input from the customer, so it is generally in everyone's best interest not to combine different languages in an inextricable fashion.

If a document contains text that is embedded in graphics, or that is part of non-editable imported objects, and can therefore not be directly accessed and replaced, Apex usually transcribes the text into a fresh MS Word document. In this manner, the additional text in these elements of the document can also be counted.

2.) Languages with Non-Phonetic Scripts

In many languages that are written using non-phonetic scripts, such as Chinese, Japanese, and Korean, it is impossible to define word delimiters. When Apex translates from English into such a language, we use the English source word count as the basis for quotation/billing - provided the source document is electronically countable. If the source document is not electronically countable, the number of English source words is estimated.

When translating from a language with non-phonetic script into English, Apex uses the estimated target word count for quotation purposes (character counts are never the basis for quotation/billing - only the corresponding word counts, which are derived by applying a multiplier). Estimates are based on counted or estimated source characters, multiplied by the applicable multiplier. The translation is usually a document that permits electronic word counting, and its word count is therefore the basis for billing.

If you have any word counting problems or questions, please contact us at your convenience at 1-800-634-4880.