ApexTranslations®
Language Solutions

The Importance and Pitfalls of Accurate Word Counting

On time. On target. On budget.

Apex Translations, Inc.
June, 2010

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.

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.

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

DOC, RTF, and WPS file types are the most common file-formats available. These file formats are counted using MS Word. It comes with a built-in word counting module that is used to establish an accurate word count of your documents.

Apex's policy is to include numerals in the billable word count for all projects, because the translation of the text and the numerals need to follow certain syntax rules of the target language and the translator must make sure the numerals are placed within the text appropriately. We will exclude numerals from the overall word count in cases where the numerals are isolated within the document, and no conversions need to be performed to meet target language conventions.

Documents that are to numerous to be counted one at a time with MS Word, or documents that possess a certain amount of complexity in terms of what should be included and/or excluded from the overall word count, we use the widely accepted word-counting software PractiCount (http://practiline.com). PractiCount provides a panoply of options for configuring delimiters and including/excluding headers, footers, comments, annotations, textboxes, numerals, and hidden text, just to name a few.

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

Apex 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, the text in the document can be highlighted using the "Select All" feature in Adobe Acrobat. If the text can be highlighted, the document is very likely electronically countable and can be counted with PractiCount (http://www.practiline.com), for example, or using a plug-in word counter for Adobe Acrobat available at http://www.intellipdf.com/stat.htm.

APEX Translations, Inc., 201 West Water Street, Suites 1 & 2, Plymouth, NC 27962, USA
Telephone: (US) 1-800-634-4880, (INTL.) 001-252-791-9902, (Fax) 1-815-461-9916
E-mail: info@apex-translations.com

PDF documents that contain scanned content cannot be counted electronically. Apex employs OCR (Optical Character Recognition) software such as ABBYY FineReader (http://www.abbyy.com) to convert the scanned content into editable text in MS Word, which can then be counted electronically (ref. item 1.).

If the quality of the PDF document is poor and a reliable word count cannot be established using OCR software, see item 7b.)

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

Apex determines the number of translatable words contained in HTML files and similar file types using DJVX software (http://www.atril.com). This software is able to separate translatable text from non-translatable text such as html code so that only translatable text will be counted.
Because HTML and similar file types typically contain a significant amount of duplicate text, we also quantify text that is repetitive. Apex charges a 50% of our regular rate for any text that occurs in duplicate.

5.) File Formats Created by Desktop Publishing (DTP) / Typesetting Software

For the purpose of generating a cost and turnaround time proposal for the translation of documents that were created with DTP software, such as (FrameMaker, PageMaker, Illustrator, Photoshop, InDesign, CorelDraw, QuarkXPress, Freehand, etc.), it suffices to establish the word count of the documents in PDF format instead of performing this analysis directly in the DTP files, which tend to be very bulky and difficult to transfer, especially by e-mail.
Countable PDFs are counted according to item 3.) above.

A second option is to conduct word counts directly in the desktop publishing software that was used to create the document. However, some DTP software packages do not offer word counting capabilities.

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

For documents that contain edits or comments, Apex's electronic word count will typically include words contained in edits and comments, unless otherwise instructed.

If a document contains text in more than one language, if may be very difficult, in some cases even impossible to electronically count only the words in one language. 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 counted, Apex typically counts this text manually which would be added to the overall word count.

7.) Documents that cannot be counted electronically

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

7a.) 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), for example. 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 (ref. item 1.).

7b.) If the quality of the source document is poor and a reliable word count cannot be established using OCR software, Apex provides you with a cost estimate that is based 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 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 based on this derived word count. However, Our invoice will be adjusted to reflect the actual electronic word count of the translation, which can be established electronically in almost all cases.

8.) Languages with Non-Phonetic Scripts

In many languages that are written using non-phonetic scripts, such as Chinese, Japanese, and Korean, we determine, to the best of our abilities, the cost of the translation based on the English target word count which mostly depends on number of characters of the source text and the subject matter to be translated.
The estimated target word count will be used for quotation and invoicing purposes.

In some cases, the above method may result in too much uncertainty. Should this be the case, we determine, to the best of our abilities, the cost of the translation as described above.
Apex's cost estimate is based on this derived word count. However, our invoice will be adjusted to reflect the actual electronic word count of the translation, which can be established electronically in almost all cases.

If you have any questions regarding our word counting methods, please contact us at your convenience at 1-800-634-4880 or send us an e-mail at info@apex-translations.com.