

AnyCount 7.0 White Paper

Table of Contents

Foreword	0
Part I Introduction: Measuring Your Work	1
Part II Text Counter Requirements	3
Part III Resume	7

1 Introduction: Measuring Your Work

In any business, customers and contractors have to deal with volumes of what they buy and sell. Word processing industry is no different: translation, proofreading, desktop publishing and other services are sold and bought in certain volumes. What makes word processing industry special, however, is that measuring its volumes requires a very special approach.

Text Volume Units Consistency

There is a number of problems to be solved by those working in text processing industry. The first of them is measurement units and their consistency. For example, text volume units alone are different in different countries.

Table 1. Text count standards in different countries

Country	Text Count Standards
UK	Per 1000 words of source text
Germany	Per line of 55 keystrokes
Belgium	Per line of 60 characters
Italy	Per 25 lines of 60 characters (1,500 characters per standard page)
USA	Per word of source text
France	Per word of target text But: In Information Technology (localization) - per word of source text
Japan	Per byte of source text
Mexico	Per page

Thus, a German localization company, working with UK market will naturally have to account one same text in both words, and keystrokes (or characters with spaces). But two texts with same number of words, can have totally different volumes of characters, simply because words can be as short as "I", and as long as "Pneumonoultramicroscopicsilicovolcanoconiosis" (longest word ever to appear in English language dictionary, according to <http://en.wikipedia.org/wiki/>

Pneumonoultramicroscopicsilicovolcanoconiosis).

Such inconsistency creates major problem for word processing companies (translation and localization agencies in particular) - the objective need to re-calculate whole text in both client's volume units and own volume units.

Text Count Tools Consistency

Volume unit consistency is not the only problem with text count. What if volumes of same text, measured by client and by vendor in the same units are still different? It is obvious, that because text volumes are quite large, nobody would ever think of counting these manually. Actually, industry has a great variety of software, elaborate or simple, to rely on when counting text. It becomes customary to include text volumes count features into word processors, for instance. However, it is this variety of software, used and sold on the market, that makes determining exact text volumes even more complex task.

Obviously, vendors cannot have clients use exactly the same software to count texts. And it is as obvious that different clients can use totally different software for text processing. If even two different versions of same product (for example, Microsoft Office Word 2003 and 2007) return different results in terms of text volumes for one same file, then what there is to say about tools developed by different manufacturers? Since text volume is the kind of volume, that is put into invoice, determining precise and mutually agreed volumes of texts becomes a vital negotiation topic.

Format Consistency

Since most documents nowadays are created and stored on electronic carriers, format topic cannot also be ignored. Not all files are plain text, and some of the files (for instance, scanned documents) are not even text files. Software, tasked with determining the volumes of such files, will have to locate text itself, and differentiate it from something which should not be counted (like code strings, for example).

The problem becomes more apparent when same text, saved in different formats, has different volumes for each, not to say that users will probably have to purchase different text counter versions designed to count different file formats.

2 Text Counter Requirements

Depending on the industry needs we can define general requirements to text count software. Let us have a glance at how various text count cases are handled by specialized text count software, like AnyCount.

Characters and Words

If you look closely at the measurement standards, mentioned in table 1, you will see that all of them can be broken down into either characters, or words. Obviously, words cannot be broken into characters, since one word can have different number of characters in it.

Ability to count text in characters with spaces, characters without spaces, and words is a must for any text count software in the industry, because very often one units are used by client, and other are used internally, and you cannot accurately convert words into characters by simple multiplication.

Is Technical Character Count Enough?

Suppose the text contains a list like this:

- one
- two
- three

Some software will return total count of characters without spaces as 11, and some will return 14. The problem is that for one program marked list bullets are characters, and for the other - they are not. Answer to question, whether such characters should be counted, depends on the other question: whether they *need* to be counted? In translation industry, for instance, these items are not translated or processed in any way: a bullet will remain a bullet in English, Arabic, Russian, Chinese, or any other document. So, must the number of bullets increase overall text volume?

The philosophy behind AnyCount engine has always been oriented at the needs of industry, rather than technical approach (i.e. technically, bullet is a character). For someone working in word processing industry, bullet is not a character, and bullet is not a word, because each character and each word has certain price and certain cost.

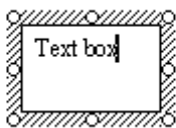
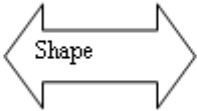

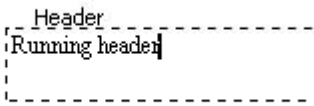
Complete Text Count In a File


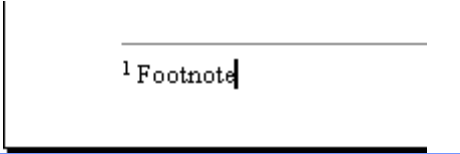
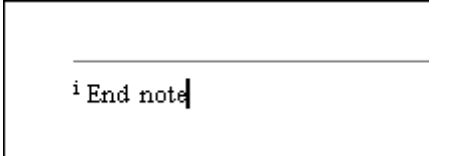
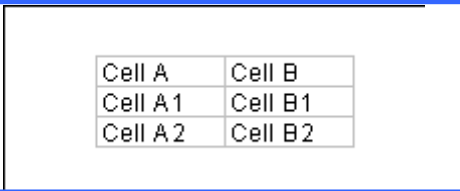
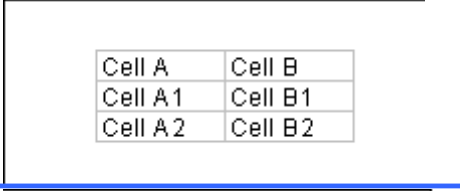
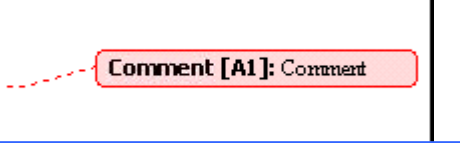
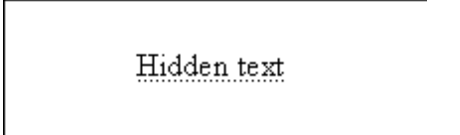
Another issue is that usually, you do not simply count text - you need to count text in a file. How is this different? Even common text files have elements which are not the part of standard text. Headers and footers can be a good example: they need to be processed, and therefore must be accounted for. However, footers and headers usually repeat - if not throughout all the text, then through certain sections in text.

This is another example of how AnyCount's approach is different: technically, you must count all the headers and footers, and on practice, these are ignored by most text file statistics. In this case determining whether identical headers and footers should or should not be counted is up to the user, while the fact that they have to be counted is not in dispute.

Headers and footers are by far not the only elements which are not usually included into word counts. There are many more, like text boxes, shapes (like wordArt), embedded objects (like tables), to name a few. Good text counter must allow user to optionally include or exclude text elements like these. Here is a brief list of what is and is not counted by AnyCount and MS Office Word 2003:

Table 2: Text count comparison for a DOC file between AnyCount 6.0 and MS Word 2003.

Object	Example	AnyCount counts	MS Office counts
Text	-	Yes	Yes
Textboxes		Yes	No
Shapes		Yes	No
WordArt		Yes	No
Running headers		Yes	No

Running footers		Yes	No
Footnotes		Yes	Yes
End notes		Yes	Yes
Embedded documents (i.e. XLS table embedded in DOC file)		Yes	No
Linked documents (i.e. XLS table linked with DOC file)		Yes	No
Comments		Yes	No
Hidden text		Yes	No

Consistent Counting of Different File Formats

There is a simple test for consistency, which any text count software absolutely must pass:

- Create a DOC or RTF document with certain amount of text in it.
- Count the volume of text in this document, then export this document to PDF.
- Count the volume of text in PDF document.
- Select all text in PDF document and paste it into a new DOC or RTF document, save file and count

text once again.

When doing so with AnyCount, be sure to exclude all format-specific elements, like comments, from the count results.

What does it mean if results of all three counts are consistent? First of all, it means that you are using same approach to counting text in different file formats. If approach is different, then different files with the same text would have different volumes, and that working with some formats would be less profitable than with the other. Clients will most certainly have different formats, and even internally a company may be limited to certain format. For instance, translation agency can be limited to working with DOC format because their translation memory software can work only with DOC.

A good example of the need to process different file formats may be a user submitting a scanned document for translation. This document is a JPG image and technically does not contain any text. Recognizing the need of processing scanned images, we equipped AnyCount with built-in text recognition module, because clients do sometimes submit scanned texts.

Another objective requirement of consistent counting different file formats is that one text processing project can contain not one, but a number of formats: for instance, software documentation can contain CHM help, HTML web-based help and PDF or MS Publisher files for printed documentation. Text Count software must be able to return a unified, precise and consistent volume for the whole such project. Using one tool that counts them all is the most logical solution.

3 Resume

As we can see, text count by software, which was not designed specifically for the industry (i.e. statistical or technical count) may prove inaccurate, or not suitable to be used for determining text volumes as far as word processing goes. We have also established, that even specialized text count software must be thoroughly tested for consistency and accuracy, at least where different file formats are used.

We have covered only a small fraction of functions and requirements of text count software. More detailed descriptions and comparisons can be found on <http://www.anycount.com>.