COMPARISON OF WORD COUNTING PROGRAMS

How Word Counting Software Reads and Interprets Text Files

(test conducted 8/11/13 with 12 popular programs)

What are word counting programs?

There are a number of programs designed to count words in documents. Some word processors have a simple word counting function built in. Other programs are designed specifically to count characters, words, or phrases. Many of the word-counting programs are designed to assist translators and proofreaders that need to bill their work by the number of words handled. Another common usage is to allow writers and students to determine the length of their document in 'words' to meet requirements of their work or school. However, the results vary with almost every program. We ran tests using some of the most popular word counting programs to determine which programs were the most accurate. To determine their accuracy, we created a test file containing some common words, punctuation, and numbers. These items were selected because they represent the most common sources of counting differences.

Word counting strategies

The programs tested appeared to use similar counting strategies for most of the text analyzed. However, there were a few key areas where they differed. Depending on the content of the text being analyzed, these differences could be substantial. Some of the differences, such as in number counting or handling email and URL addresses, would be much greater in technical documents. There are also different opinions as to what counts as a 'word', even among experts. Should contractions such as *you'll* be treated as one or two words? Should hyphenated words always be treated as two words? In the color-coded analysis below, we attempted to use what we believed was the most commonly accepted standard as the benchmark.

Some individual uses of a word-counting program may have requirements that differ from this standard so that requirement should be taken into consideration when evaluating programs for your use. This will most likely be most critical in evaluating how any given program handles contractions, possessives, and numbers. More importantly, it may be more useful to always use the same program when comparing documents to each other or to some standard, as the accuracy of the word counts should remain consistent relative to other documents.

We did not have access to the actual algorithms used by each program to count words. Thus, we had to infer their counting strategies by having each program analyze snippets of text containing the items in question until we could determine how each item was counted. The general strategies seem to fall into one of several categories:

- **Space-separated words**: Many of the products, including MS Word, use a simple algorithm for most cases that looks for spaces (or an equivalent like line ending) to determine the word boundaries. Thus, they treat 'one/two', 'one-two', and even 'one!two!three' all as single words. These will be accurate in most cases and will tend to produce lower word counts.
- **Words separated by spaces and other characters**: Other programs also treat other characters, in addition to spaces, as 'word separators'. They might treat an embedded colon as a word separator, or always honor sentence ending characters (.!?) as dividing words, even if there are no spaces on either side. These programs tend to report slightly higher word counts.
- **Specialized handling of numbers and apostrophes:** The last grouping included programs handling numbers and apostrophes differently. They may count commas or periods in numbers as word separators, as well as apostrophes in contractions and possessives. Thus, *123,343.22* may be counted as three words, and *you'll* will count as two words. These will tend to produce the highest word counts but may be required in situations requiring these capabilities.

Test file

wordtest.txt download file

The test file we created is a plain text file that has been saved in UTF-8 format to accommodate upper ASCII characters. It was created by entering text in a simple text editor and pasting in some special characters from a Word document. It is not very long, but is meant to test simple word counting and also counting of words that contain special characters that may be represented internally by more than one character. This analysis is for the UTF-8 format. Files in other formats will have different internal representations but we have found UTF-8 to be the most reliable way to display and edit text files. The items added to the file are described below and match the results table found below.

Item to test	String in file
Embedded colon in word with no spaces	**in:word**
Words separated by slash	**in/slash**
Email addresses	**bob@abc.com**
Numbers	**123.43** **$543.12** **98%**
URLs	**www.abc.com**
Embedded underscore	**under_score**
Embedded periods in common words	**a.m.**
Contractions	**isn't**
Possessive with regular apostrophe	**Bob's**
Embedded hyphen	**with-hyphen**
Embedded em or en dash (from Word)	**en-dash**
Possessive with curly apostrophe (from Word)	**smart's**

Word counting program comparison

We tested a number of word counting programs by opening our test file (or pasting it into the program in some cases) and running the count analysis and recording the results. Several of the programs had fairly consistent results but they were all influenced by the factors we discussed above. Several programs had options to turn on or off certain word counting options so we tried to include results from both options. The column marked 'Possible number of words' indicates the number of separate words contained in the string that are separated by some kind of separator character. This column is not meant to necessarily indicate the 'correct' count as several of the strings have different but valid interpretations. The numbers in each of the other columns indicates the word count for that item reported by each program.

We have indicated with green backgrounds those results that agree with what, we believe, is the accepted standard for that item. Items with yellow background vary from that standard, but may be perfectly acceptable in a given situation. Those marked in red indicate a result that would almost always be seen as incorrect or misleading (such as counting 1,234.44 as three words) but may be required in some situations. The programs that had an obvious option or preference setting for a specific situation have both values reported, along with *option. Many of the programs can report the total number of words with and without counting numbers. The two rows with no background colors, email and URL addresses, were left this way because there was no clear right or wrong setting, and the desired result would vary by usage and document contents. This test focused mainly on the word counting capabilities of each program. Many of the programs had other features, such as being able to count multiple files at one time, or producing invoices based on the number of words found.

Results of testing wordtest.txt download file

String in document	Possible number of words	MS Word	myWordCount	FineCount	Word List Expert	AKS Word Count	Word Count Manager	TotalAssistant	TexTally	Any Count	Sobolsoft	PractiCount	Smart Edit
colon:word	2	1	2	2	1	2	1	1	1	1	1	1	1
slash/word	**2**	1	2/1 *option	2	1	2	1	1	2	1	1	1	1
bob@abc.com	**3**	1	3	2	1	2	1	1	2	2	1	1	1
1234	**1**	1	1	1	1	1	1	1	1	1	1	1	1
1,234.44	**1**	1	1	3	1	3	1	1	2	1	1	1	1
$123.00	**1**	1	1	2	1	2	1	1	1	1	1	1	1
$1,222.32	**1**	1	1	3	1	3	1	1	2	1	1	1	1
www.abc.com	**3**	1	3	3	1	3	1	1	1	3	1	1	1
under_score	**1**	1	1	1	1	1	1	1	2	1	1	1	1
a.m.	**2**	1	2/1 *option	2	1	2	1	1	1	2	1	1	1
isn't	**1**	1	1	2	1	1	1	1	1	1	1	1	1
Bob's	**1**	1	1	2	1	1	1	1	1	1	1	1	1
with-hyphen	**2**	1	2/1 *option	2	1	2	1	1	2	1	1	1	1
en--dash	**2**	2	2	2	1	1	1	1	1	1	1	2	1
smart's (curly)	**1**	1	1	2	1	1	1	1	1	1	1	1	1
Number of words in our test file as reported by each program													
Without numbers		**n/a**	**322/297**	**343**	**n/a**	**358**	**n/a**	**n/a**	**n/a**	**313**	**n/a**	**244**	**n/a**
Including numbers		**327**	**392/367**	**431**	**314**	**392**	**327**	**314**	**340**	**355**	**314**	**314**	**314**

*option: the numbers shown for the myWordCount column are for counting without the options on, and then with the options on.

Indicates a program's word count agrees with the accepted number

Indicates a program's word count varies from the accepted number but may be acceptable for your purposes

Indicates a program's word count may be misleading or incorrect, but may still have specific uses

Analysis

The top bar of the table shows how each program treats specific word strings by indicating how many words will be counted in each case. We have marked with green backgrounds those results that agree with the accepted value for that item. Some programs, like myWordCount, have options that handle how hyphens, slashes, and abbreviations are handled. Programs that have options to include or not include numbers in the count have entries for both rows at the bottom of the table. The two rows at the bottom indicate how each program reported the word count for our sample wordtest.txt file. These results reflect the information contained in the previous section of the table. The relative differences between the program totals will change

depending on the content of the document, particularly in the number of URL and email addresses, and the use of numbers. Some of the programs had options to give the user some control over selecting the word-ending characters, but these are difficult for many users and their usage would result in varying results for the same document Some specific comments on the results include:

- **Numbers**: Most programs treated numbers, no matter how long, as one word. Some, however, treat commas and periods inside numbers as word separators, resulting in much higher counts.
- **Embedded underscores.** All but one program report words containing embedded underscores as one word. this seems to be the accepted standard now, particularly since Google has recently embraced this usage.
- **Abbreviations**: Most programs that simply count space-separated words will treat all abbreviations like a.m. or U.S.A as single words. Some always treat the periods as word separators, resulting in high word counts for these abbreviations. Some programs, like myWordCount, have options allowing you to specify which abbreviations should be treated as single words.
- **Hyphenated words**. The common practice is to treat hyphenated words as single words. Most of the programs follow this convention, while a few do not and some, like myWordCount, give you the option to count these as single or multiple words.
- **En and em dashes**: Another common practice is to treat en and em dashes as word separators which differentiate them from hyphens. However, less than half of the programs tested treated our en and em dashes this way. Part of this discrepancy may be due to the fact that the dashes were pasted in from a Word document and were multi-byte characters.

Results of testing large file

`PrincessOfMars.txt` **download file**

We created another test file (`PrincessOfMars.txt`) by downloading the Edgar Rice Burroughs story from Project Gutenberg, *The Princess of Mars*, and converting it to a text file. It is a fairly long file with a few special characters. We include this test here to test the various programs on the speed of loading, and the accuracy of their number counting routines. We list the total word count from each program in the table below. We just included the total including all numbers. Besides providing a total word count, several of the programs we tested also produced tables of word counts for each word in the document. We included the extra time each program took to load and analyze the count by word. We also randomly selected three words to test the accuracy of each program's count. The results are shown below. Some of the programs combined upper and lower case letters, others like myWordCount give you the option to combine upper and lower case words or not. Several of the other programs reported the found words in several different places with different capitalization or punctuation.

Word in text	Actual count in text	MS Word 2003	myWordCount	FineCount	WordList Expert	AKS Word Count	WordCount Manager	Total Assistant	TexTally	Any Count	Sobolsoft	PractiCount	Smart Edit
Total Word Count including numbers		70,494	70,607	70,986	70,493	71,154	70,494	70,494	71,107	70,811	70,472	70,494	70,493
time to load word table		n/a	:07	n/a	5:15	:02	n/a	n/a	n/a	n/a	1:50	:03	:03
the	**4816**		**4816**		**4490**	**4813 ***					**4812**	**4797***	**4623**
another	**57**		**57**		**55**	**57***					**57**	**57***	**55**
Barsoom	**76**		**76**		**65***	**76***					**76***	**76***	**71**
Comments and subtotals where word totals were broken down by case or punctuation			upper and lower case can be combined by turning on an option		Barsoom: 64, Barsoom!: 1	THE:15 The:283 the:4496 "The:13 "the:6 "Another:1 Another:3 another:52 another?:1 Barsoom:74 Barsoom's: 1 Barsoom?: 1					Barsoom:75 barsooms:1	(the:1 The:283 the:4494 the,:1 "the:5 "The:13 "Another:1 Another:3 another,.?:53 The Barsoom total was 76 but was listed 9 times with different chars at the end	upper and lower case combined

Indicates a program's word count agrees with the actual number in the document verified independently

Indicates a program's word count varies from the actual number in the document

Analysis

The total counts reported vary from 70,472 to 71,154 words. These differences can mostly be explained by the information in the previous table indicating how each program counts different types of word combinations. There is no 'correct' count for these as there are too many variables. The analysis of the programs that reported word counts by individual words showed some important differences, including:

- **Time to load**: Most of the programs performed the extra analysis fairly quickly, between two and seven seconds. The exceptions were *Word List Expert* which took over 5 minutes, and the Sobolsoft program which took almost 2 minutes.
- **Grouping words**: Some of the programs reported every variation of the target words as separate line items and counts. This included differences in capitalization and ending punctuation. This made determining the true count of any specific word difficult as each instance was found in a different part of the table, and some of the programs took a long time to resort the list.

Details of the word counting programs tested

Program	myWordCount	FineCount	WordList Expert	AKS Word Count	WordCount Manager	Total Assistant	TexTally	Any Count	Sobolsoft	PractiCount	Smart Edit
Version tested	*3.08*	*2.6.1934*	3.2.1	1.3.0.60	*2.5.2*	*2.6.0.4*	1.10	*8.0.7*	n/a	3.2	*3.001*
Count listed for each word	**yes**	**no**	**yes**	**yes**	**no**	**no**	**no**	**no**	**yes**	**yes**	**yes**
Can program highlight each word in original document?	**yes**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**yes**
Can it also count characters?	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**no**	**yes**	**yes**
Can it also count phrases?	**yes**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**no**	**yes**
Price	**$14.95**	**$49.95**	**$14.95**	**$24.00**	**$29.95**	**$29.95**	**$19.95**	**$49.95**	**$29.95**	**$59.95**	**$59.95**
Website	website	website	website	website	website	website	website	website	website	website	website

Programs Tested

myWordCount
AKS Word Count
TexTally
PractiCount
FineCount
Word Count Manager
AnyCount
Smart Edit
Word List Expert
Total Assistant
Sobolsoft Word Count

Tests performed: 8/10/13