Character Counter 4.4

Stephen Palmstrom 8 July 2015

# Purpose

As part of checking the conversion of legacy fonts to Unicode, it is useful to obtain statistics for the numbers of different glyphs and their respective fonts. This program will analyse Microsoft Word documents including Rich Text Format and produce an Excel spreadsheet of the various glyphs and their fonts. You can then compare the legacy and Unicode files to check that the conversion has worked satisfactorily. It will also analyse text documents (anything that does not have the .doc, .docx or .rtf extensions). It is possible to analyse several documents and obtain an aggregate count of the various glyphs and their fonts. Note that this is an interim release – in a future version it will be possible to do a bulk analysis and obtain a list of glyphs by filename and font.

# Licensing and Copyright

The program was developed for MissionAssist, so it is copyright to that organisation, but distributed under the GNU General Public License (<http://www.gnu.org/licenses/gpl.txt>).

# System Requirements

The program uses both Microsoft Word and Excel in its analysis, so needs both programs, and runs under Microsoft Windows (7 or above).

# Architecture and Troubleshooting

The program opens the Word document you specify and goes through it character by character counting each one, but distinguishing their respective fonts. It gives you the option of counting the decomposed characters that make up individual glyphs or graphemes as units. It also checks, if the font is Unicode, if they could be composed and makes suggestions for such composition.

The program invokes hidden instances of Word and Excel unless they are already running, when it connects with and hides them. In the event of it crashing or being terminated abnormally, you will have to use Task Manager (Right-click on the taskbar and select Task Manager to run it) or its equivalent to terminate those instances. If you terminate one of them using Task Manager or because it becomes visible, the program may crash, though there is error trapping so if, for example, Word has been closed, the program will attempt to reopen it. Similarly for Excel. If the Excel workbook you want to write to is open, the program cannot delete it before writing, but it will let you close the workbook and try again.

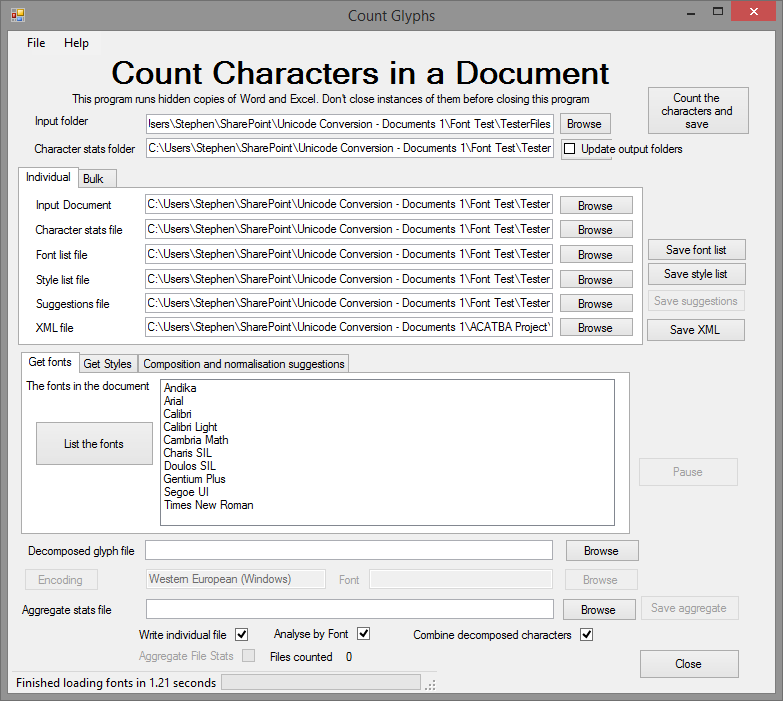
You can also analyse text document, i.e. documents without the .doc, .docx or .rtf extensions, but the program cannot determine the font with which to display the glypns. You can choose the encoding of the text file.

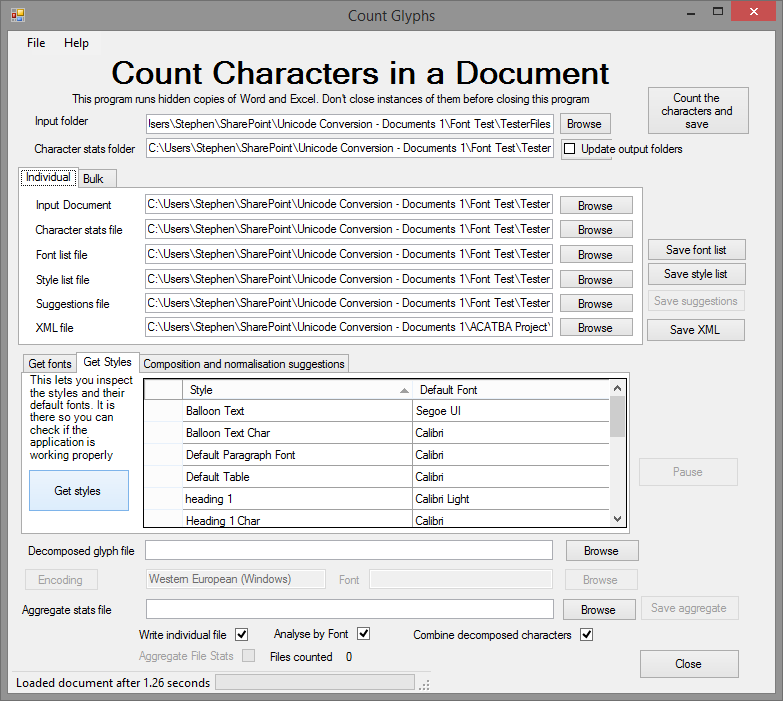
If you give it an Excel file name in the Aggregate stats file box, it will default to accumulate aggregate statistics on a number of files that you might be processing one at a time. The Bulk tab lets you select a number of files to process in one go from a given folder and get individual statistics files.

# Installation and Execution

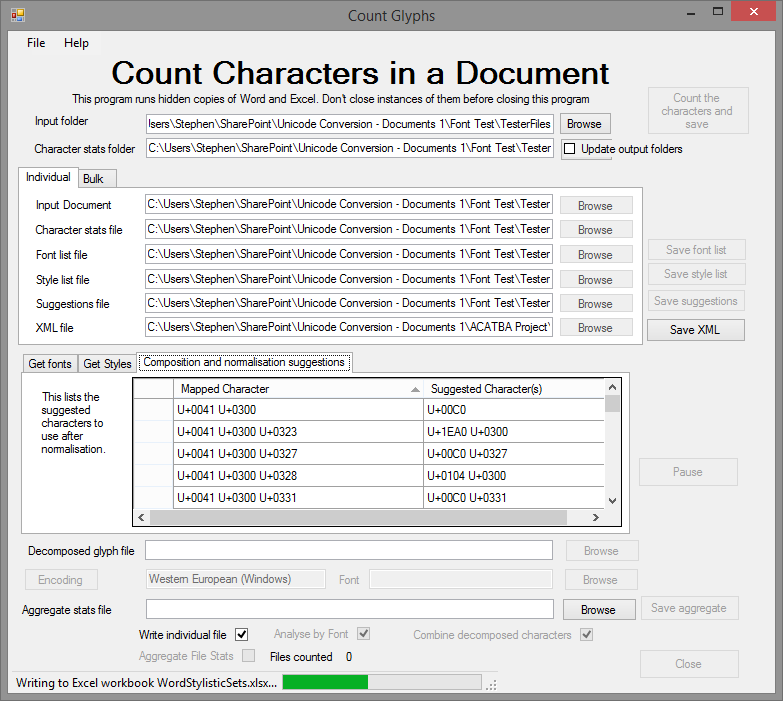
Install the program by clicking Setup.exe in the CharacterCounter folder. A link to it should then appear in the Start Menu.

Browse to select the Word document you wish to analyse. The Excel output file will default to a file of the same name in the same folder. The program remembers the last folder you used, so once you have done one analysis, subsequent analyses in the program run will go to the folders you used.

The *List the fonts* button will be enabled once the program has an input and output file. This will list the fonts the Word thinks it has in the document.

The *Count Characters* button will be also enabled once the program has input and output files. Once you have clicked it, the *Pause* button will be enabled and the Close button disabled. Clicking *Pause* will give you the option to close whilst doing the analysis. The process os sp fast that you are unlikely to need to pause it. Progress is shown at the bottom of the window together with an estimate of the time to completion calculated by multiplying the average clock ticks per character by the remaining characters. This means that the estimated time to completion may even increase if the program encounters many fonts in the middle of the document. The calculation does not take into account the writing of the Excel file at the end of the analysis. If you uncheck the Analyse by Font checkbox the program will run faster, but it will not attempt to analyse the glyphs by font. The *Count Characters* button will be disabled once the program has counted the characters in a file.

Clicking on the *Get Styles* tab will expose the Get Styles button that lists styles in use and their default fonts. It will also list some notional styles used by the program to remember the default styles in the Word document. The function is provided in case you need to troubleshoot the analysis.



Clicking on *Count Characters* will do the analysis using the options you have chosen. If the program is counting decomposed characters as one, it will also flag up any normalisation anomalies it finds. You have the option of saving them to an Excel workbook so you can modify the mapping if you so wish.

# Output

This is an example of the first few rows of the Excel file output when Analyse by Font is checked:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Font** | **Dec** | **MS Hex** | **USV** | **Glyph** | **Count** |
| Calibri | 2 | 2 | U+0002 |  | 2 |
| Arial | 9 | 9 | U+0009 |  | 2 |
| Napa Heavy SF | 12 | C | U+000C |  | 1 |
| Arial | 13 | D | U+000D |  | 3 |
| Calibri | 13 | D | U+000D |  | 9 |
| Elementary Heavy SF | 13 | D | U+000D |  | 1 |
| Footlight MT Light | 13 | D | U+000D |  | 1 |
| Formal436 BT | 13 | D | U+000D |  | 2 |
| Freehand575 BT | 13 | D | U+000D |  | 1 |
| High Tower Text | 13 | D | U+000D |  | 1 |
| Horizon BT | 13 | D | U+000D |  | 1 |
| Napa Heavy SF | 13 | D | U+000D |  | 3 |
| Symbol | 13 | D | U+000D |  | 2 |
| Tennessee Heavy SF | 13 | D | U+000D |  | 4 |
| Times New Roman | 13 | D | U+000D |  | 2 |
| Troutkings BTN | 13 | D | U+000D |  | 1 |
| Napa Heavy SF | 14 | E | U+000E |  | 1 |
| Napa Heavy SF | 30 | 1E | U+001E | ‑ | 2 |
| Napa Heavy SF | 31 | 1F | U+001F | ­ | 1 |
| Arial | 32 | 20 | U+0020 |  | 19 |
| Calibri | 32 | 20 | U+0020 |  | 48 |
| Elementary Heavy SF | 32 | 20 | U+0020 |  | 9 |
| Footlight MT Light | 32 | 20 | U+0020 |  | 4 |
| Formal436 BT | 32 | 20 | U+0020 |  | 4 |
| Freehand575 BT | 32 | 20 | U+0020 |  | 11 |
| High Tower Text | 32 | 20 | U+0020 |  | 3 |
| Horizon BT | 32 | 20 | U+0020 |  | 7 |
| Napa Heavy SF | 32 | 20 | U+0020 |  | 19 |
| Tennessee Heavy SF | 32 | 20 | U+0020 |  | 11 |
| Times New Roman | 32 | 20 | U+0020 |  | 20 |
| Troutkings BTN | 32 | 20 | U+0020 |  | 11 |
| Calibri | 33 | 21 | U+0021 | ! | 1 |
| Calibri | 36 | 24 | U+0024 | $ | 1 |
| Calibri | 37 | 25 | U+0025 | % | 1 |
| Calibri | 38 | 26 | U+0026 | & | 1 |
| Calibri | 40 | 28 | U+0028 | ( | 1 |
| Calibri | 41 | 29 | U+0029 | ) | 1 |
| Calibri | 42 | 2A | U+002A | \* | 1 |
| Calibri | 43 | 2B | U+002B | + | 1 |
| Napa Heavy SF | 44 | 2C | U+002C | , | 2 |
| Napa Heavy SF | 45 | 2D | U+002D | - | 2 |
| Arial | 46 | 2E | U+002E | . | 2 |
| Calibri | 46 | 2E | U+002E | . | 6 |
| Freehand575 BT | 46 | 2E | U+002E | . | 1 |
| Napa Heavy SF | 46 | 2E | U+002E | . | 1 |
| Times New Roman | 46 | 2E | U+002E | . | 1 |
| Troutkings BTN | 46 | 2E | U+002E | . | 1 |
| Footlight MT Light | 48 | 30 | U+0030 | 0 | 1 |
| Footlight MT Light | 49 | 31 | U+0031 | 1 | 1 |
| Horizon BT | 49 | 31 | U+0031 | 1 | 1 |

The first column is missing if Analyse by Font is not checked and there is only one value for each code point. Unchecking Analyse by Font speeds the process a bit. The default paragraph font is used to display the characters when you don’t check Analyse by Font.

If you choose to count decomposed characters as one, all the codes will appear in the one line, e.g.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Times New Roman | 100 821 | 64 335 | U+0064 U+0335 | d̵ | 1 |

# Composed and decomposed characters and normalisation

|  |
| --- |
| Glyphs |
| r̃ |
| ṣ̌ |
| š̯ |
| ʊ̀ |
| į̈́ |
| į̈ |
| ɨ́ |
| ɨ̨ |
| ɨ̨́ |

There are times when a given glyph or grapheme needs more than one Unicode or legacy character to represent it, for example to add an accent, circumflex etc. These would be defined on a font by font basis for legacy fonts, whilst Unicode has a whole list of combining characters to perform that function. Because they are standard in Unicode, standard routines can be made available to handle them, whilst they are non-standard in the legacy fonts. We handle those by allowing you to provide an Excel spreadsheet containing the combining characters and their fonts in cells A2 downwards. The program creates a regular expression that matches any character preceding any of those glyphs, or a range of characters. Thus, it will find any characters associated with the suspected combining character. This may help you in mapping legacy character combinations to Unicode combined characters. This feature only makes sense if you are analysing by font. Unicode fonts are handled using standard Windows routines. We also check to see if decomposed characters can be made up of composed characters and whether the character string is normalised ([http://unicode.org/reports/tr15/#Description\_Norm](http://unicode.org/reports/tr15/%23Description_Norm)), again using standard routines and list those which could be composed and/or normalised. You can save the list as an Excel workbook for future reference – see the Help menu. A sample workbook for legacy characters is distributed with this program.

The program will count the combinations it recognises as single characters, for example.

|  |  |
| --- | --- |
| **Mapped Character** | **Possible Character** |
| U+0061 U+0303 | U+00E3 |
| U+0065 U+0301 | U+00E9 |
| U+0069 U+0328 | U+012F |
| U+00E1 U+0328 | U+0105 U+0301 |

# A note on regular expressions as used to analyse combining characters.

A regular expression is a very powerful, if complex, way of analysing and manipulating text. In this case, we create an expression of this form:

(.[<character combination1><character combination 2>...<Character combination n>])

This means: Match any character followed by any of the characters in the list supplied. Prefix any character that is significant in a regular expression with the \ escape character.

# ‘Under the bonnet’

Whilst working on getting statistics on characters used in Word, we discovered that it was not possible readily to discover characters inserted using *Insert Symbol*. Instead, they were being returned as a left bracket ‘(‘ or U+0028. However, in newer versions of Word (2003 onwards) it is possible to get an XML (eXtensible Mark-up Language) representation of the document, which exposes the symbols to programs. So, we analyse the XML representation of the document to get the statistics on the text and symbols therein. It allows us to get the fonts and styles used very easily.

This does mean, however, that we do not handle characters below U+0020, i.e. control characters in the same way. This is because they are represented in a different way in XML. This program attempts to count them as if they had the value that appears in Word, so, for example, a tab is counted as U+0009 and a paragraph marker as U+000D.

We do the counting in the body of the document, which also holds headers, footers, footnotes and end notes. There are some Stories in Word that it doesn’t cover, like footnote continuations that don’t seem to be used in ‘real’ documents, though the program could be updated to do so if there were a felt need.

|  |  |  |
| --- | --- | --- |
| **Object** | **Character representation** | **Comment** |
| Footnote mark | U+0002 | ^f in Word when searching |
| Endnote mark | U+0002 | ^e in Word when searching |
| Separator | U+0003 | Seems only to be in special Stories |
| Continuation separator | U+0004 | Seems only to be in special Stories |
| Tab | U+0009 | ^t in Word when searching |
| Column break | U+000£ | ^n in Word when searching |
| No break hyphen | U+001E | ^~ in Word when searching |
| Soft or optional hyphen | U+001F | ^- in Word when searching |
|  |  |  |

**Warning** – the reliability of the counting and font specification for characters below U+0020 is less than for characters above it as we have had to deduce how Word handles these by examining the XML representations of Word documents.