

---

## NAME

countWords.pl

## SYNOPSIS

```
% perl countWords.pl x y z
```

```
x = <1=keep_tagged_words | 0=discard_tagged_words>
```

```
y = <0=no_consolidations | 10=consolidate_and | 20 =consolidate_thorn | 30 =consolidate_both>
```

```
z = <w=sortByWords | c=sortByCounts>
```

```
e.g. % perl countWords.pl 1 0 c
```

means keep\_tagged\_words, no\_consolidations, and sortByCounts

## DESCRIPTION

### SUMMARY

Assumes the script "sortIntoDirectories.pl" has already been run (see README for that script to learn of assumed directory structure). Script to generate frequency counts of each word in each text. Goes through each file and counts the number of times each word appears, the number of unique words in the file, the number of total words in the file, and the number of words that appear only once (hapax legomena). The information is then outputted to a .csv file, and the same information is calculated for each text in each manuscript in each genre.

## INPUT

### Old English .txt files

Input is from the sorted\_texts folder created by the script "sortIntoDirectories.pl" that this script assumes has already been run. We assume the texts to be input are stored up one directory level, that is, in: "../1\_sort\_Into\_Directories/sorted\_texts/". See the variable \$SORTED\_TEXTS in code below. If the texts are in a different directory, you should change this variable.

### Command line arguments

Whether to keep tagged words or not, whether to consolidate or not. This information is input on the command line. See example above.

## OUTPUT

A new word\_counts folder will be produced, containing folders for each genre, which contain sub folders for each manuscript, which in turn have .csv files for each text.

Example: The output for Genesis AB text will be found in A01.001\_GenA\_B\_T00010.csv. This file can be found by going to:  
A\_Poetry --> A01 --> A01.001\_GenA\_B\_T00010\_keepTagged\_none\_c.csv  
(genre) (manuscript) (file)

The output filename (A01.001\_GenA\_B\_T00010\_keepTagged\_none\_c.csv) contains the command line arguments. In this example, the tagged words have been kept, no consolidations have been made, and the words are sorted by count.

The .csv file contains the genre, manuscript, filename, number of unique words, number of total words, number of times a word appears only once

in the text (hapax legomena), the code for each word as it would appear in Anglo-Saxon, as the word would appear in the .sgml document, its counts in the text, and its relative frequency (number of times the word appears in the text proportionate to the number of words in the text). The counts are sorted in order from most to least frequent.

Example (Partial) Output for Genesis AB:

```
A_Poetry A01 A01.001_GenA_B_T00010.txt 4576 2773 17094
RANK HTML WORD RAW WORD COUNT RELATIVE FREQUENCY
1 and <code>and</code> 555 0.0324675
2 on <code>on</code> 452 0.026442
3 &thorn;a <code>&amp;t;i;a</code> 358 0.020943
```

## AUTHORS

Mark D. LeBlanc  
Christina L. Nelson

## MODIFICATION HISTORY

### June 12, 2007 (mdl) --

just getting started

### June 14, 2007 (mdl) --

removing punctuation (period, double-quote) from words

### June 15, 2007 (mdl) --

adding command-line switch to keep or discard tagged words (<foreign>, <corr>) or // (double-slash)  
From: C<OldEnglishCorpus/oecorpushtml/corpus.htm>  
"Words which are fragmentary in manuscript or emended by the editor of the text are enclosed by '< >'. This may also indicate that there is a problem with the manuscript in the space adjacent to the word. Editorial punctuation has usually been adopted; for most texts it follows modern norms. Text that is originally in runic script is enclosed in double slashes '//".

### June 15, 2007 (mdl) --

assuming the begin and end tags for runes,  
<hi...> ... </hi>, respectively are on the same line

### June 25, 2007 (mdl) --

consolidate some words and types:  
(i) all words forced to lowercase (lc), which also handles these forced equalities  
(ii) &AE; == &ae;    &D; == &d; == &T; == &t;    &E; == &e;  
     &Omega; == &omega;

### Sept 14, 2007 (mdl) --

&amp; == and == ond  
&d; == &t; (eth/ev == thorn)

**May 12, 2008 (cIn) --**

comments

**May 21, 2008 (mdl, cIn) --**

remove summary statistics

**May 22, 2008 (cIn) --**

update to be able to print out what the word looks like  
in Anglo-Saxon, change from .xls output to .dat, write pod

**May 23, 2008 (cIn) --**

finish removing summary statistics, finish pod

**May 28, 2008 (cIn) --**

add relative frequency to output, change to comma delimited

**May 30, 2008 (cIn) --**

change files from .dat to .csv. Add a header row  
to the output that is meaningful to the data. Finish pod

**June 2, 2008 (cIn) --**

finish comments

**June 4, 2008 (cIn) --**

subsort alphabetically before printing sorted by counts

**June 11, 2008 (cIn) --**

add rank to output

**June 16, 2008 (cIn) --**

change filenames to include command line information,  
i.e. A01.001\_GenA\_B\_T00010\_keepTagged\_none\_c.csv

**May 11, 2009 (mdl, cIn) --**

update documentation

**COPYRIGHT INFORMATION**

=====  
Copyright (C) 2009 Wheaton Lexomics Research Group, Norton MA

This program is free software: you can redistribute it and/or modify  
it under the terms of the GNU General Public License as published by  
the Free Software Foundation, either version 3 of the License, or  
(at your option) any later version.

This program is distributed in the hope that it will be useful,  
but WITHOUT ANY WARRANTY; without even the implied warranty of  
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the  
GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.