

---

**NAME**

as3\_countWords.pl

**SYNOPSIS**

```
% perl as3_countWords.pl w x y z
```

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

```
x = <1=consolidate | 0=leave_as_is>
```

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

```
z = <1=disemvowel | 0=leave as is>
```

```
i.e. % perl as3_countWords.pl 1 0 c 0
```

means keep\_tagged\_words, no\_consolidations, sortByCounts, and do not disemvowel

Note: When disemvowel is turned on, all vowels from each word are removed

**DESCRIPTION****SUMMARY**

Script to generate frequency counts of each word in each text of those text files found in the local directory "split\_texts".

Starting in the "split\_texts" directory, goes through each subfolder and each file within those subfolders 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). For each input file, the information is then outputted to an associated .csv file.

**INPUT**

The "split\_texts" folder, with subfolders, either created by the "cutter.pl" script that has already been run or generated elsewhere by the user.

**OUTPUT**

A new "analysis\_word\_counts\_w\_x\_y\_z" folder (where w, x, y, z are the command line values used) containing subfolders like those found in the input directory, each of those holding .csv files for each input text. Each .csv file contains the input 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). If sorting is done by counts (command line argument y=c), the counts are sorted in order from most to least frequent.

For example:

```
analysis_word_counts_1_0_c_0
```

```
is the output folder name for a test run keeping tagged words, not  
consolidating, sorted by counts, and not disemvoweled.
```

**AUTHORS**

Mark D. LeBlanc  
Christina L. Nelson  
Amos C. Jones

**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 (cln) --**

comments

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

remove summary statistics

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

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 (cln) --**

finish removing summary statistics, finish pod

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

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 16, 2008 (acj) --**

remove clutter and update pod  
fixed some output file problems

**June 17, 2008 (acj) --**

fixed first two lines of header to read as genre and manuscript

**May 12, 2009 (mdl, cIn) --**

updated 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/>>.