Feliks Dushatsky
A05248825
Math 187

Letter Frequency Computation

The text I used for my analysis of letter frequencies was one consisting of all of Shakespeare's works. I obtained this text file from the website of Project Gutenberg (http://www.gutenberg.org/etext/100). The file itself was around 5 MB in size.

I computed initial letter frequencies without altering the punctuation within the file. As expected, "E" and "T" were the most frequent; more detailed results for this are attached. Before I removed any grammar, I removed all punctuation from the text. The following is the punctuation that I removed: commas (82414), periods (76897), colons (1812), semicolons (17194), question marks (10475), exclamation marks (8827), tildes (0), dashes (7835), double quotes (450), open parenthesis (165), and closed parenthesis (164). I replaced these punctuation marks with spaces.

There remained the single quotes punctuation to remove, but before I could remove that I had to take care of the grammar word (phrase) *'s*. So, I replaced *'s* with *space 's space*, then searched for *'s* in the file to see how many *'s* entries were present. The fact that *'s* contained some punctuation within it distinguished it as a special case for the grammar words. Once I found how many *'s* there were in the file, I replaced that grammar word with a space, and replaced the remainder of the single quotes (22169) with spaces.

At this point all of the punctuation should have been deleted from the file. I proceeded to scan and remove grammar words from the file; when I scanned for these words, I ignored case. There were a total of 27 grammar words that I searched for in the file, they accounted for approximately 25% of the total words. Once I removed all of the grammar words from the text, I computed the letter frequencies again. Some of the letter frequencies changed from before, in fact the frequency of the most common letter ("E") went up, more detailed results are attached.

Feliks Dushatsky
A05248825
Math 187

The theory was that the letter frequencies would “flatten” out and become more “even-handed” once the grammar words were removed. The standard deviation for the letter frequencies without any grammar words was lower compared the standard deviation with the grammar words. The decrease in standard deviation, along with a visual comparison of the two graphs, is evidence that removing grammar words does in fact yield more “even-handed“ letter frequencies.

However, even though the results satisfy the theory, the evidence isn’t very strong. The standard deviation decreased by a small amount, and the relative change in individual letter frequencies was not that significant. Looking at the graph of letter frequencies with the grammar words removed, we see that the frequencies flattened out somewhat, yet the graph maintained its overall figure, with peaks in the same areas.

An explanation as to why certain letters have higher frequency may relate to our reliance on vowels. The vowels in English are *a, e, i, o, u,* and *y* where y is the least common vowel. We see that the most common vowels are *e, a,* and *o*, and they move up in frequency rank once grammar words are removed. Furthermore, the frequency of *a, e, i, u,* and *y* increased once the grammar words were removed. English sentence structure relays on the use of grammar words, yet the removal of these grammar words doesn’t hinder the occurrence of vowels. This shows that there exists a certain reliance on vowels, this may stem from the fact that every word is composed of at least one vowel.

If we consider these six vowels in twenty six total letters, we see that vowels make up about 23.1% of the alphabet. Taking some random words, “cryptography” or “mathematics” for instance, we see that vowels make more than their fair share of the letters in each word. “Cryptography” consists of 33.3% vowels, and “Mathematics” consists of 36.3% vowels. Vowels in general seem to be frequent within individual words, this may be an explanation as to why their overall frequencies are so significant. Moreover it is evident that vowels are a necessity, they “glue” together consonants to compose words.

Letter Frequencies with grammar words

Unsorted

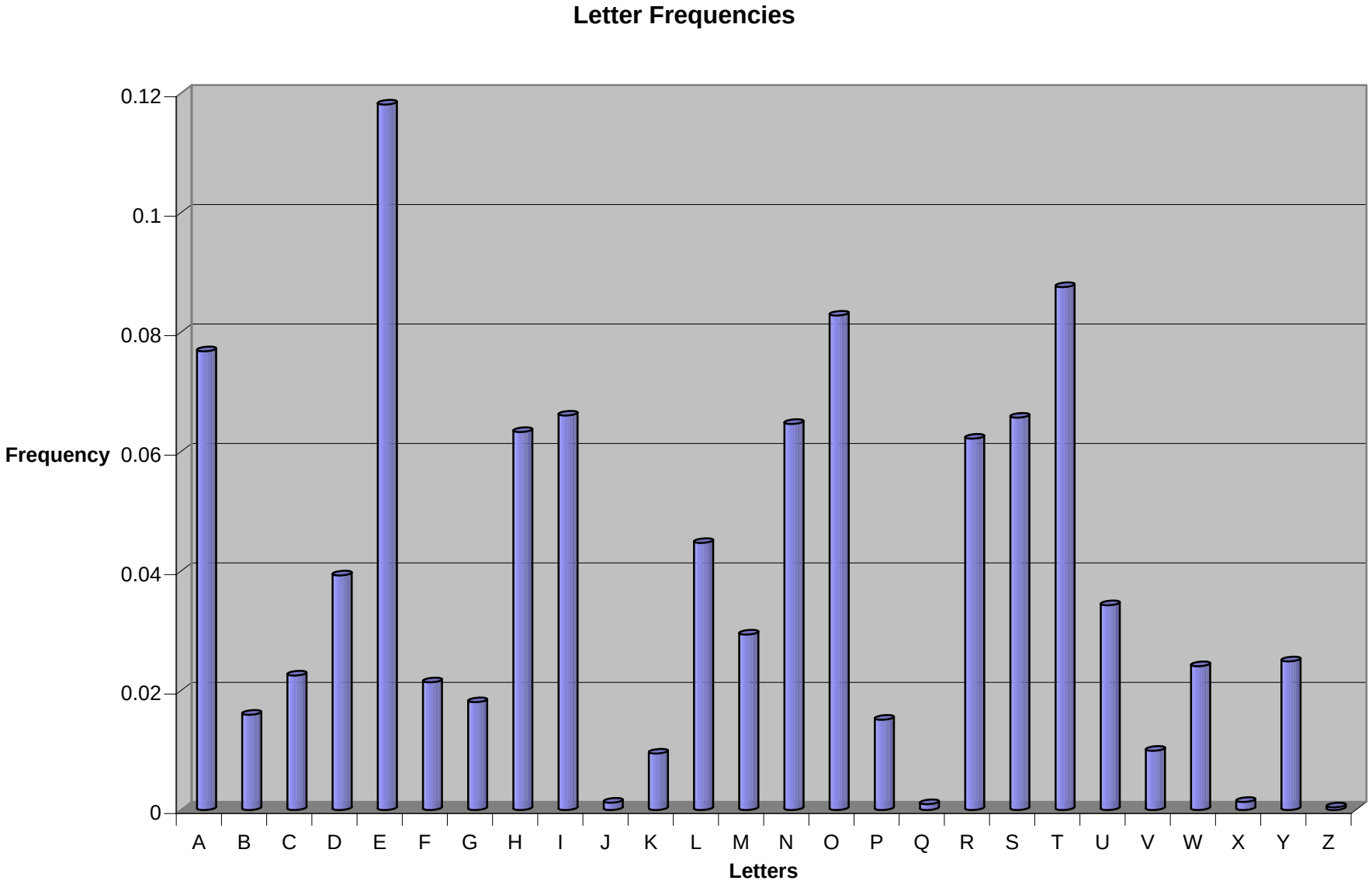
A	0.0768058
B	0.01594794
C	0.02251054
D	0.039275795
E	0.118117586
F	0.021374356
G	0.018053435
H	0.06328114
I	0.066043034
J	0.001226515
K	0.009446465
L	0.044710867
M	0.029375508
N	0.06463396
O	0.082780436
P	0.015108177
Q	0.000967
R	0.06217011
S	0.06568711
T	0.08755424
U	0.034278594
V	0.00991354
W	0.02406755
X	0.001347409
Y	0.024882704
Z	0.00044

Sorted

E	0.118117586
T	0.08755424
O	0.082780436
A	0.0768058
I	0.066043034
S	0.06568711
N	0.06463396
H	0.06328114
R	0.06217011
L	0.044710867
D	0.039275795
U	0.034278594
M	0.029375508
Y	0.024882704
W	0.02406755
C	0.02251054
F	0.021374356
G	0.018053435
B	0.01594794
P	0.015108177
V	0.00991354
K	0.009446465
X	0.001347409
J	0.001226515
Q	0.000967
Z	0.00044

Standard Deviation: **0.031858**

Letter Frequencies with grammar words



Letter Frequencies without grammar words

Unsorted

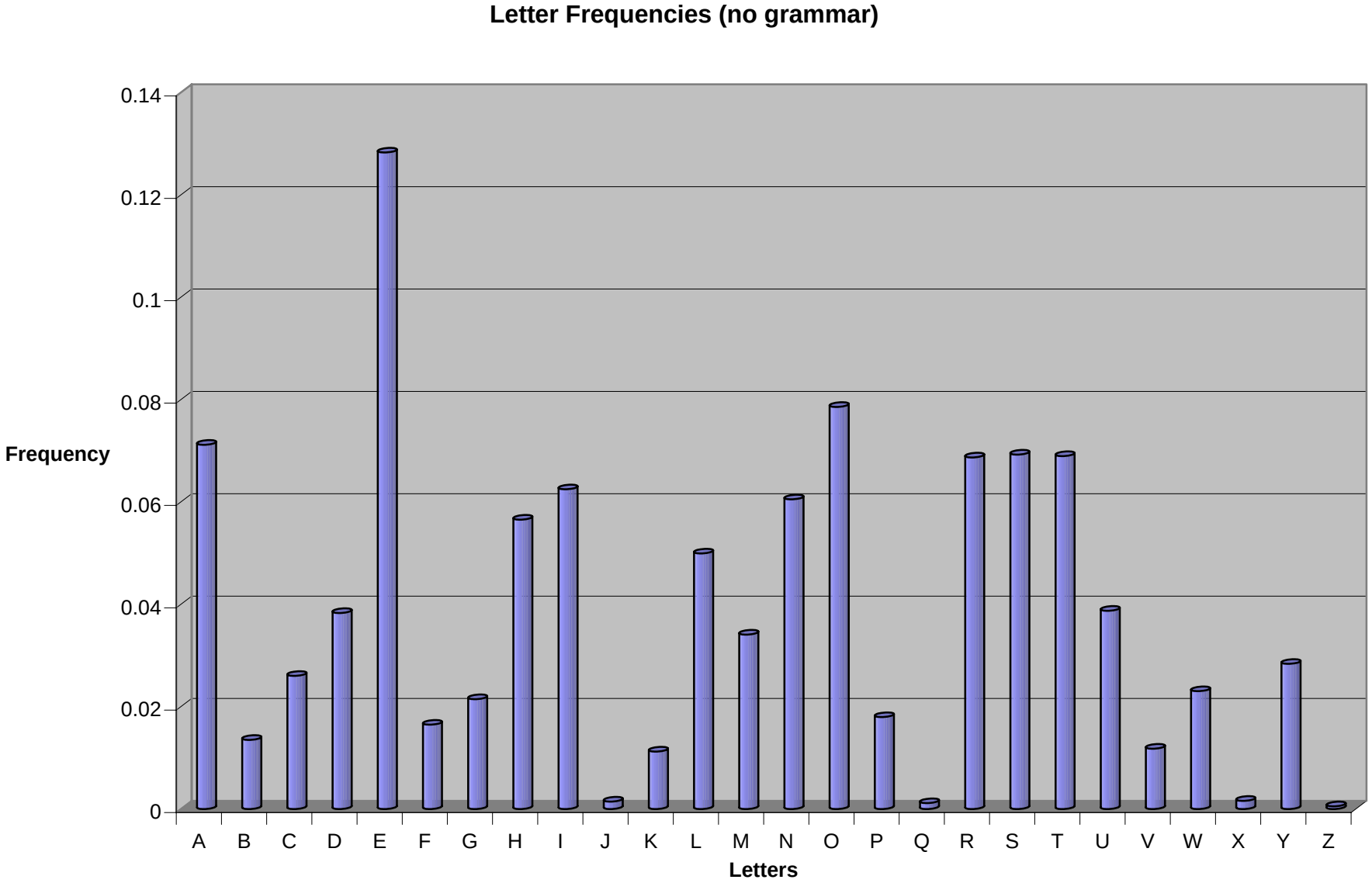
A	0.071168765
B	0.013502114
C	0.026011799
D	0.038300943
E	0.12827796
F	0.016475543
G	0.02145987
H	0.05653598
I	0.06246773
J	0.001457942
K	0.011228883
L	0.049927182
M	0.034067288
N	0.06050216
O	0.07856938
P	0.017958881
Q	0.001149957
R	0.06870007
S	0.06925399
T	0.06896369
U	0.03872852
V	0.01178409
W	0.023024866
X	0.001601646
Y	0.028358007
Z	0.000523

Sorted

E	0.12827796
O	0.07856938
A	0.071168765
S	0.06925399
T	0.06896369
R	0.06870007
I	0.06246773
N	0.06050216
H	0.05653598
L	0.049927182
U	0.03872852
D	0.038300943
M	0.034067288
Y	0.028358007
C	0.026011799
W	0.023024866
G	0.02145987
P	0.017958881
F	0.016475543
B	0.013502114
V	0.01178409
K	0.011228883
X	0.001601646
J	0.001457942
Q	0.001149957
Z	0.000523

Standard Deviation: **0.031242**

Letter Frequencies without grammar words



Grammar word frequencies

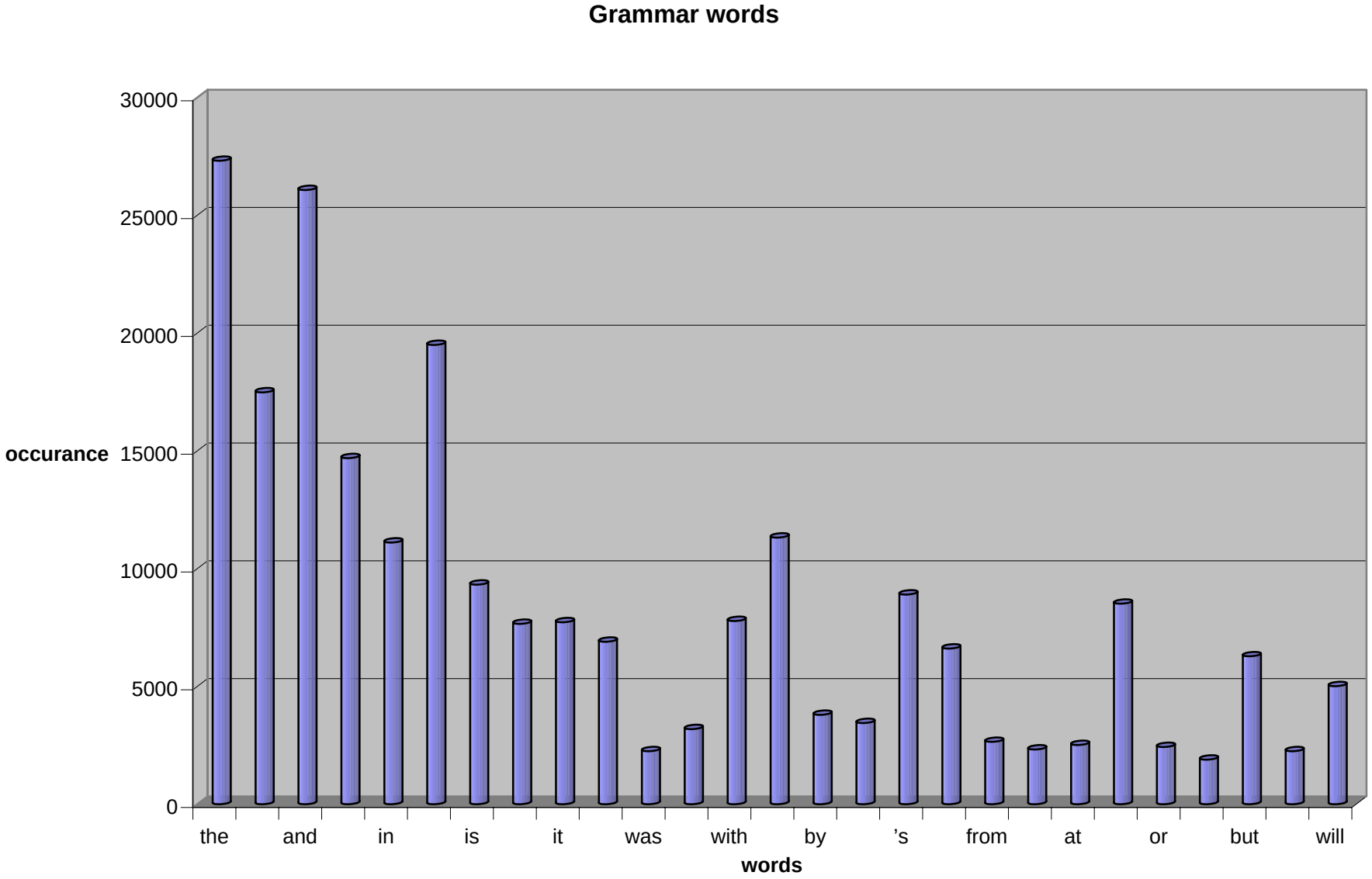
Unsorted

the	27312
of	17481
and	26069
a	14691
in	11105
to	19497
is	9318
for	7655
it	7723
be	6897
was	2249
on	3188
with	7777
that	11318
by	3794
are	3446
’s	8898
this	6608
from	2646
which	2321
at	2514
not	8511
or	2429
an	1890
but	6277
has	2249
will	5008

Sorted

the	27312
and	26069
to	19497
of	17481
a	14691
that	11318
in	11105
is	9318
’s	8898
not	8511
with	7777
it	7723
for	7655
be	6897
this	6608
but	6277
will	5008
by	3794
are	3446
on	3188
from	2646
at	2514
or	2429
which	2321
was	2249
has	2249
an	1890

Total grammar words:	**228871**
Total words:	**903614**
Percent grammar words	**25.33%**


Grammar words
occurance
30000
25000
20000
15000
10000
5000
0
the
and
in
is
it
was
with
by
’s
from
at
or
but
will
words