# Data Visualization on Pompe Diseases

Aman Berhe

May 13, 2016

The Dimention of the Positive and negative words:

Table 1: Pompe Diseases with all the words and total number of sentences

| Pompe data | positive | negative | vector |
|---|---|---|---|
| Words | 2663 | 11029 | 3 |
| sentence | 69 | 416 | 1 |

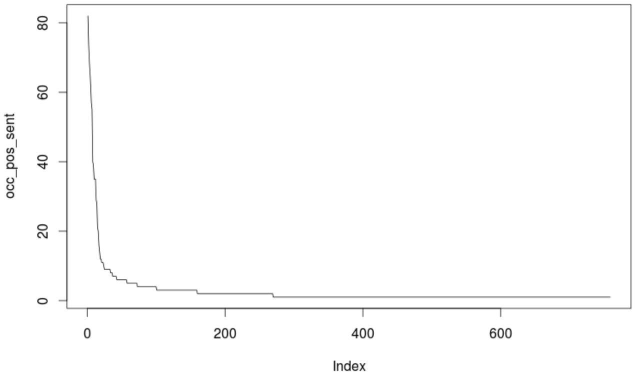
Table 2: Pompe Diseases with punctuation removed words in a sentence and total number of sentences

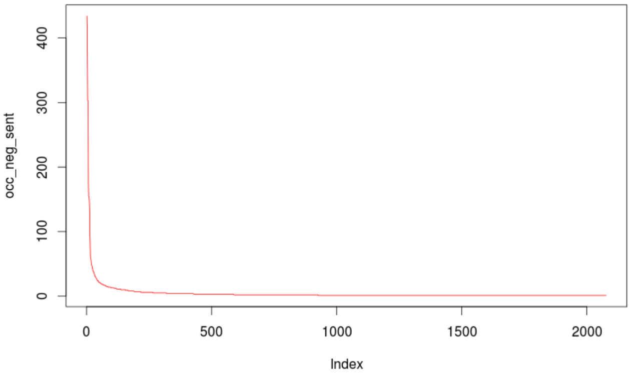
| Pompe data | positive | negative | vector |
|---|---|---|---|
| Words | 2155 | 9684 | 3 |
| sentence | 69 | 416 | 1 |

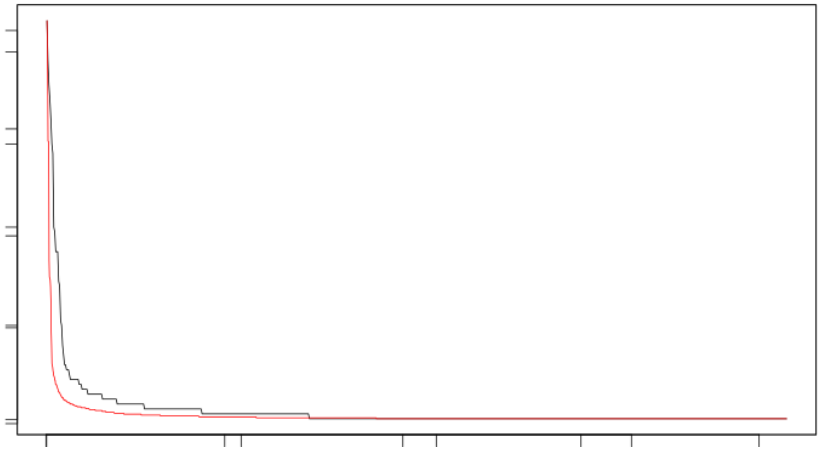
Table 3: Pompe Diseases with punctuation removed words in a sentence and total number of sentences

| Pompe data | positive | negative | vector |
|---|---|---|---|
| Words | 1789 | 8138 | 3 |
| sentence | 69 | 416 | 1 |

Here are some pictures which shows the occurrence of words. the words satisfy the Zipf's Law

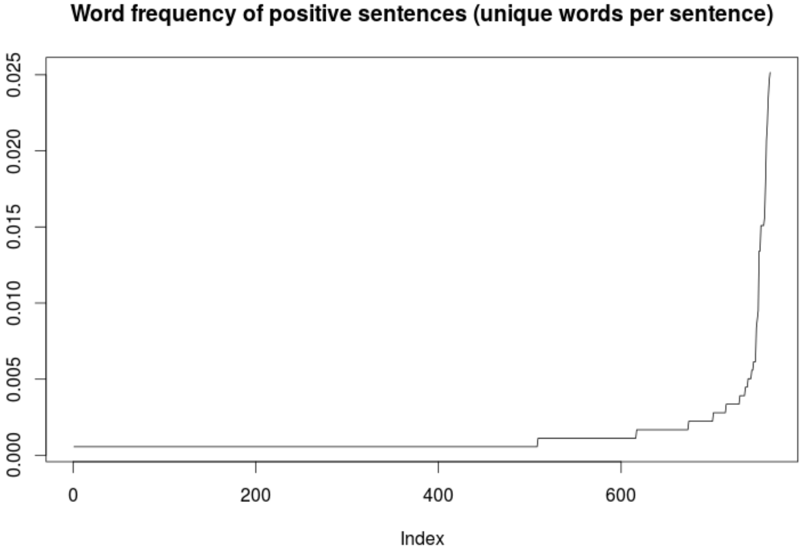


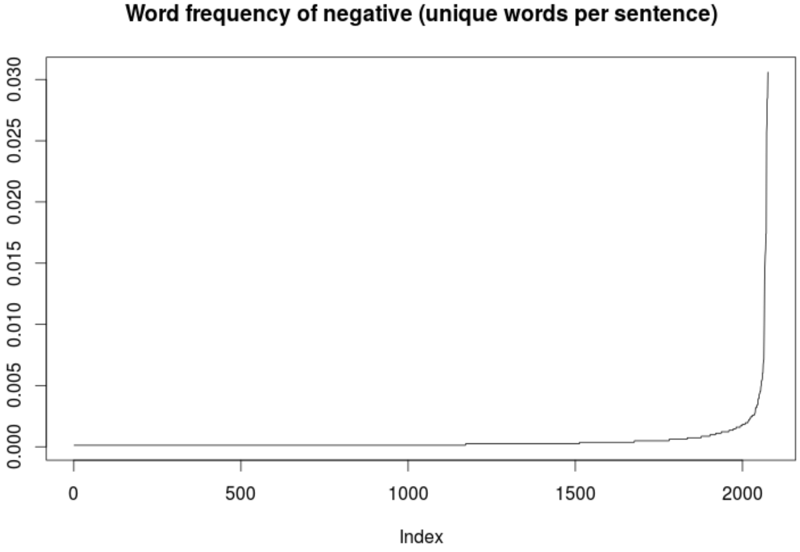




The More the data the curve is very smooth. The curves of the unique words in a sentence is very similar to the above pictures

Plots of Word frequency: here I take only one word per sentence

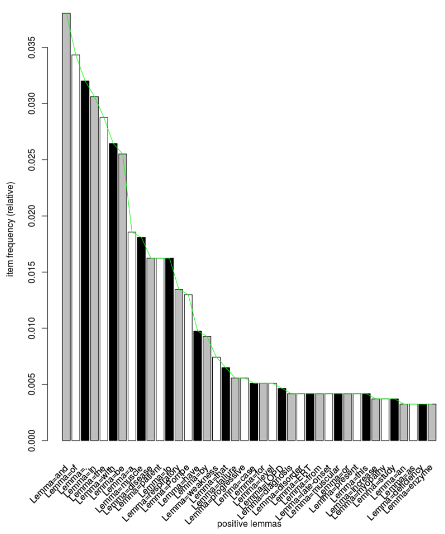


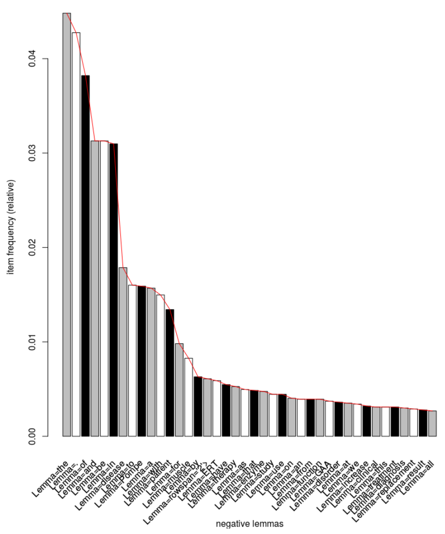


The plots of word frequency and unique word per sentence occurrence look almost alike since the occurrence is used to compute the frequency

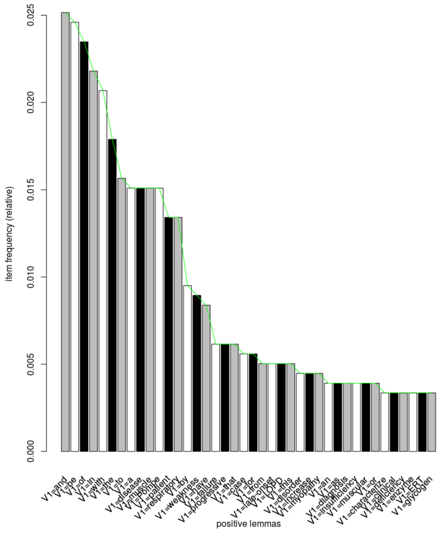
Here are the top 20 words

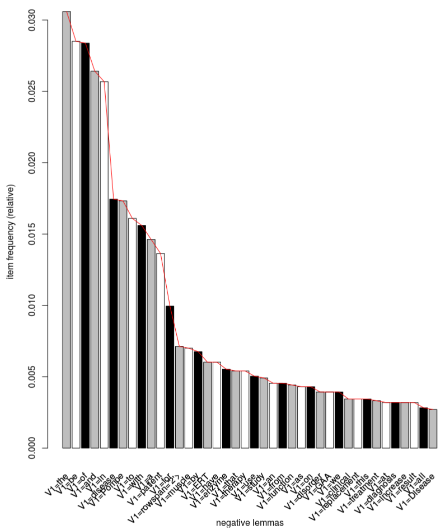
for all words in a snetence


positive Amytrophic

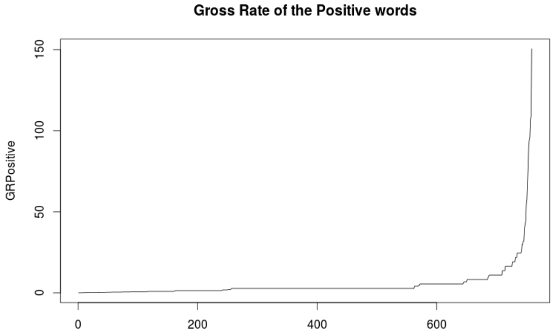

negative Amytrophic

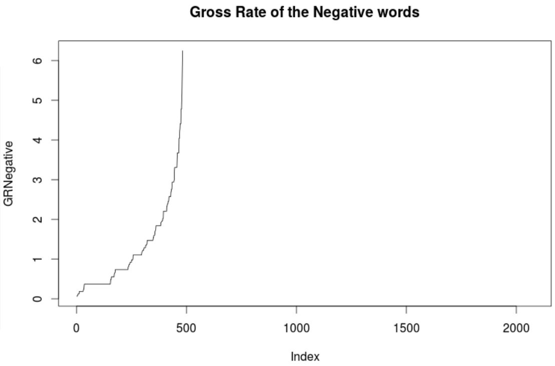
for the unque words per sentence


positive Amytrophic

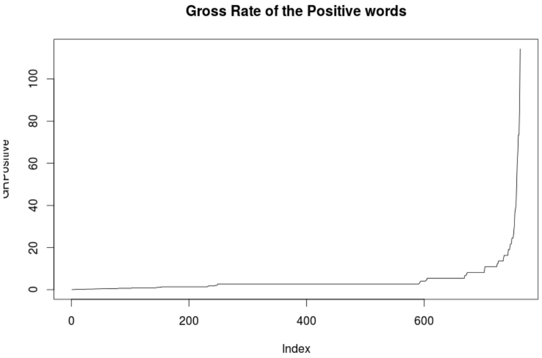

negative Amytrophic

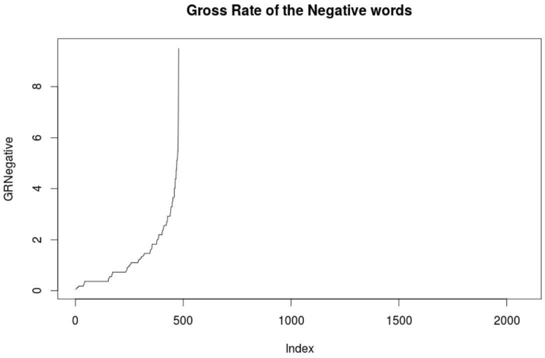
The following shows the gross rate of the words on total number of words



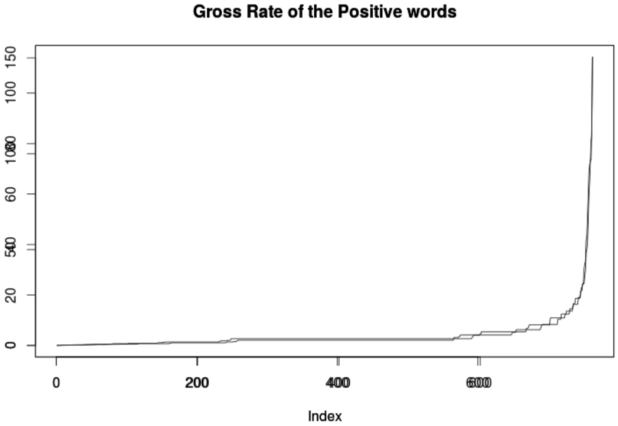


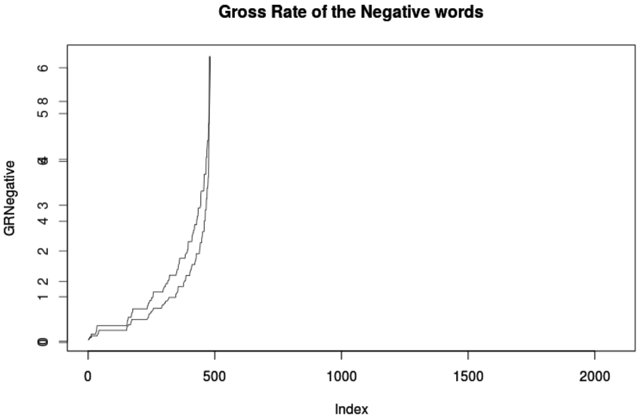
The following shows the gross rate of the unique words in a sentence on total number of words



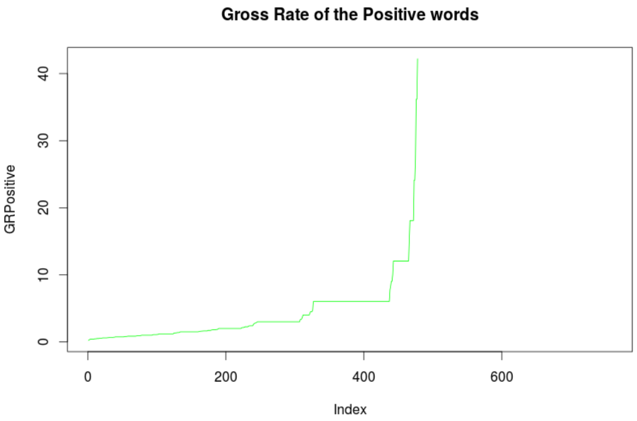


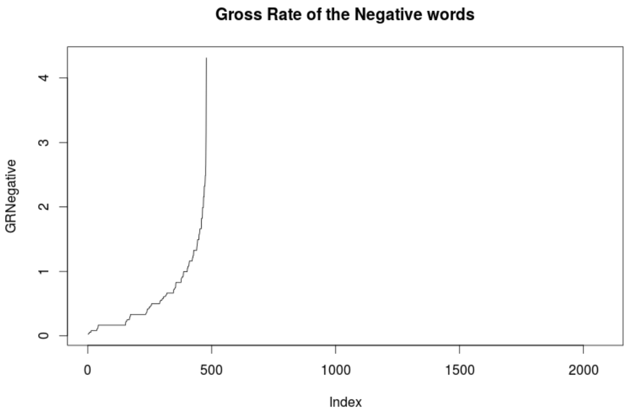
comparison between the unique words per sentence and total words in a



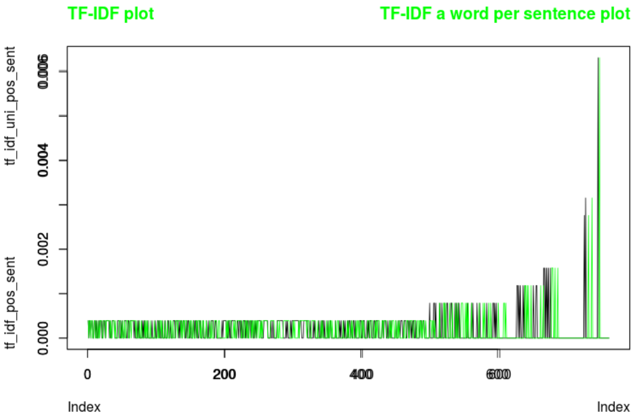


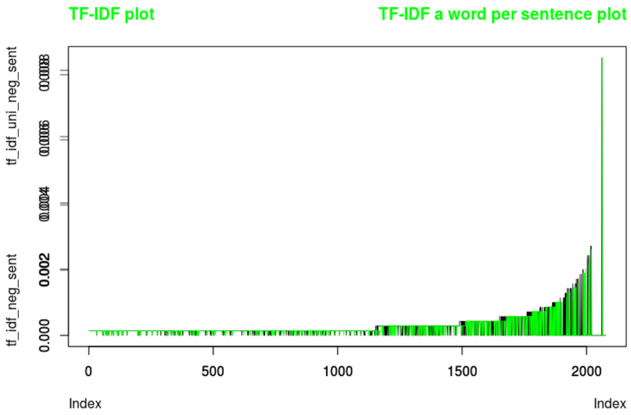
The following shows the gross rate of the words over total number sentences





TF-IDF representation of the words look like this



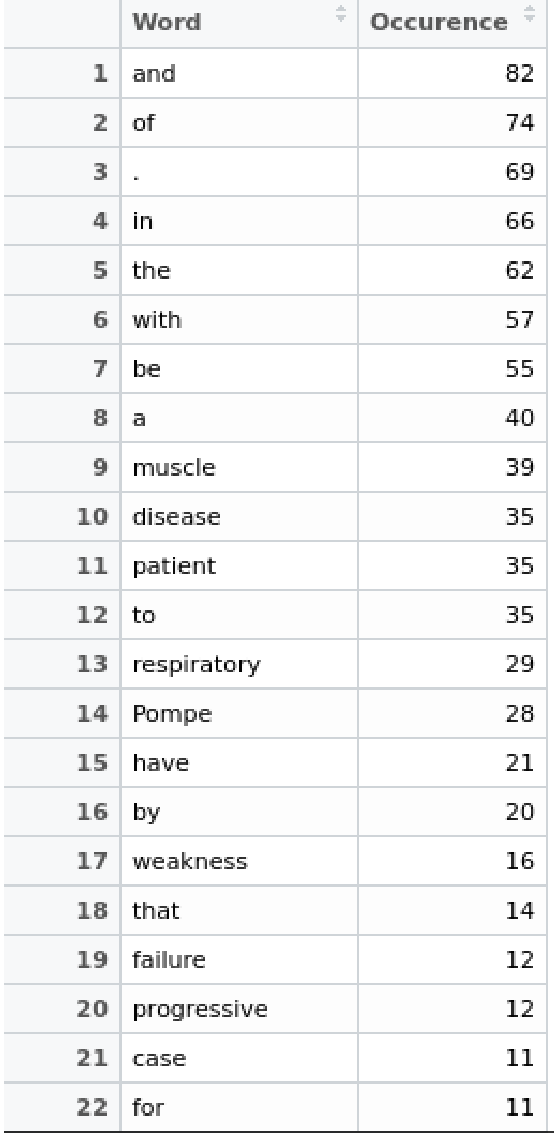


The Tf-IDF is computed by: if the word exists in both documents it will be 0 if not its frequency will be multiplied by 0.3 (log(2))

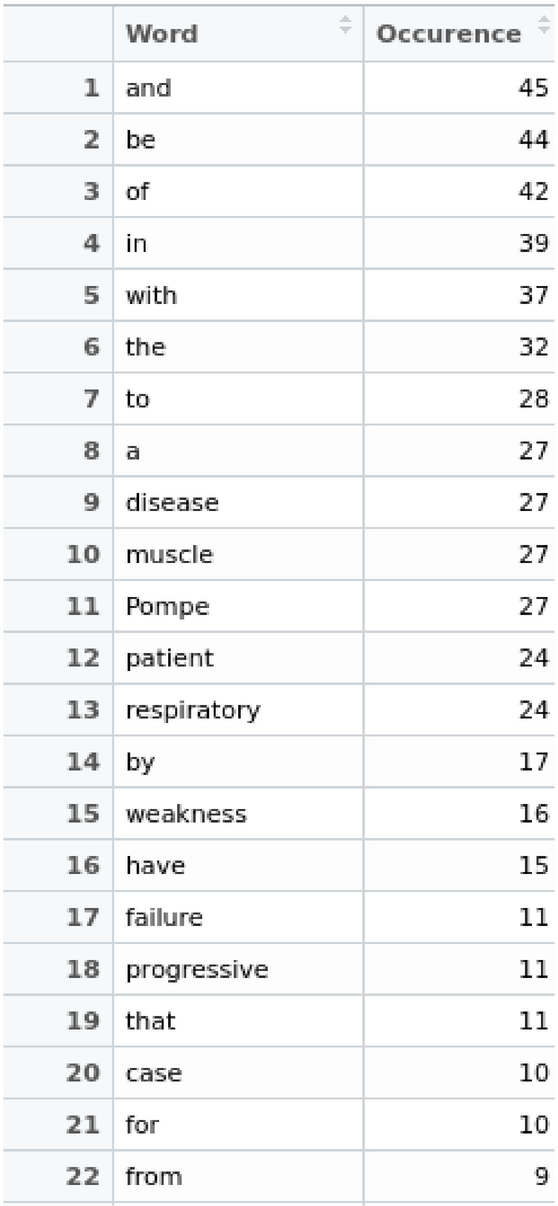
# Words and their Occurrence

The words and their ocuurence has benn put side by side like this

positive data

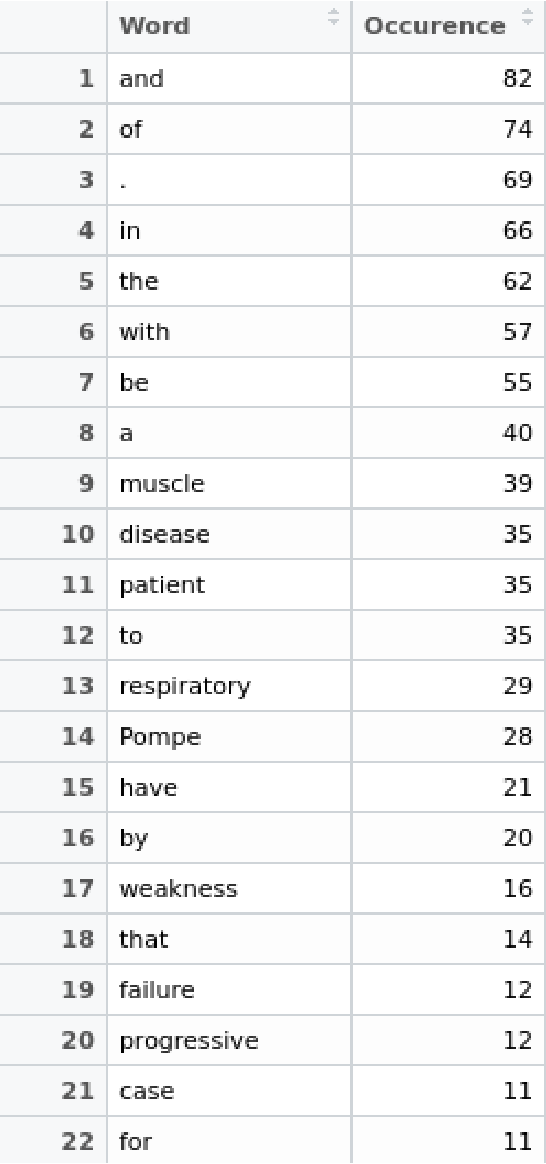


|  | Word | Occurence |
|---|---|---|
| 1 | and | 82 |
| 2 | of | 74 |
| 3 | . | 69 |
| 4 | in | 66 |
| 5 | the | 62 |
| 6 | with | 57 |
| 7 | be | 55 |
| 8 | a | 40 |
| 9 | muscle | 39 |
| 10 | disease | 35 |
| 11 | patient | 35 |
| 12 | to | 35 |
| 13 | respiratory | 29 |
| 14 | Pompe | 28 |
| 15 | have | 21 |
| 16 | by | 20 |
| 17 | weakness | 16 |
| 18 | that | 14 |
| 19 | failure | 12 |
| 20 | progressive | 12 |
| 21 | case | 11 |
| 22 | for | 11 |

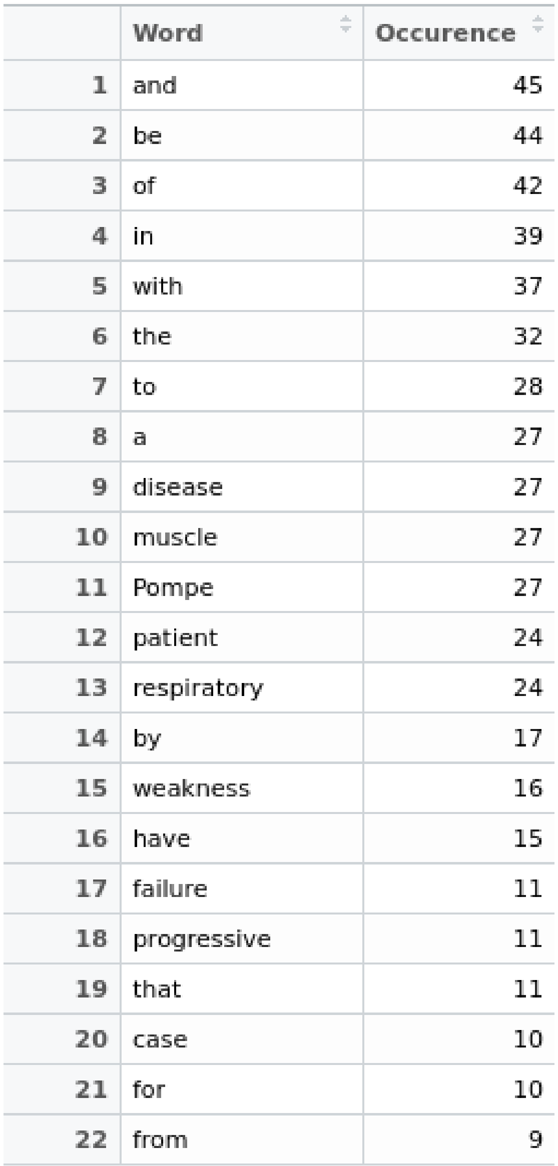


|  | Word | Occurence |
|---|---|---|
| 1 | and | 45 |
| 2 | be | 44 |
| 3 | of | 42 |
| 4 | in | 39 |
| 5 | with | 37 |
| 6 | the | 32 |
| 7 | to | 28 |
| 8 | a | 27 |
| 9 | disease | 27 |
| 10 | muscle | 27 |
| 11 | Pompe | 27 |
| 12 | patient | 24 |
| 13 | respiratory | 24 |
| 14 | by | 17 |
| 15 | weakness | 16 |
| 16 | have | 15 |
| 17 | failure | 11 |
| 18 | progressive | 11 |
| 19 | that | 11 |
| 20 | case | 10 |
| 21 | for | 10 |
| 22 | from | 9 |

negative data



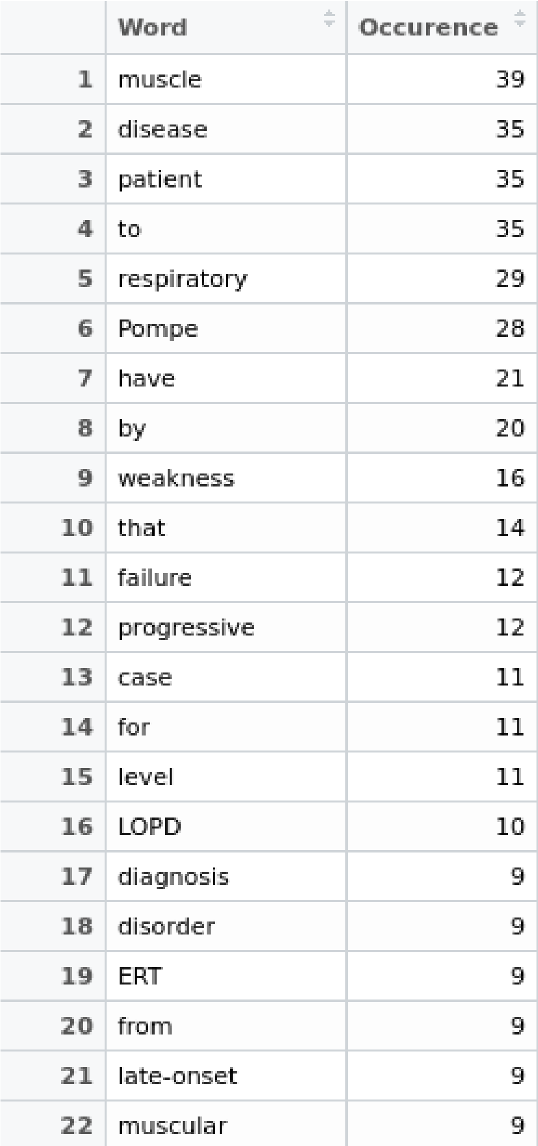
| | Word | Occurence |
|---|---|---|
| 1 | and | 82 |
| 2 | of | 74 |
| 3 | . | 69 |
| 4 | in | 66 |
| 5 | the | 62 |
| 6 | with | 57 |
| 7 | be | 55 |
| 8 | a | 40 |
| 9 | muscle | 39 |
| 10 | disease | 35 |
| 11 | patient | 35 |
| 12 | to | 35 |
| 13 | respiratory | 29 |
| 14 | Pompe | 28 |
| 15 | have | 21 |
| 16 | by | 20 |
| 17 | weakness | 16 |
| 18 | that | 14 |
| 19 | failure | 12 |
| 20 | progressive | 12 |
| 21 | case | 11 |
| 22 | for | 11 |



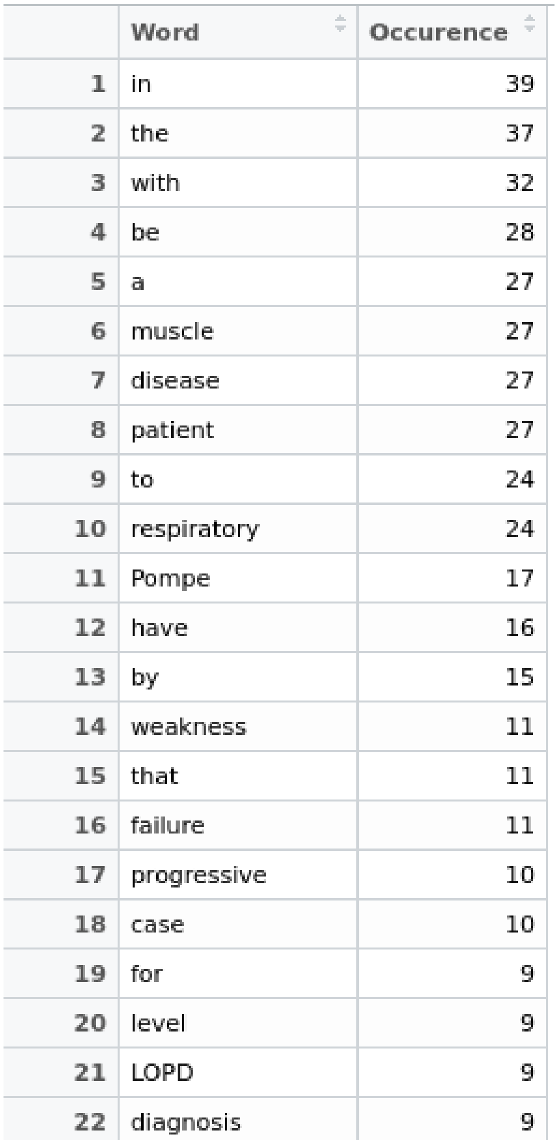
| | Word | Occurence |
|---|---|---|
| 1 | and | 45 |
| 2 | be | 44 |
| 3 | of | 42 |
| 4 | in | 39 |
| 5 | with | 37 |
| 6 | the | 32 |
| 7 | to | 28 |
| 8 | a | 27 |
| 9 | disease | 27 |
| 10 | muscle | 27 |
| 11 | Pompe | 27 |
| 12 | patient | 24 |
| 13 | respiratory | 24 |
| 14 | by | 17 |
| 15 | weakness | 16 |
| 16 | have | 15 |
| 17 | failure | 11 |
| 18 | progressive | 11 |
| 19 | that | 11 |
| 20 | case | 10 |
| 21 | for | 10 |
| 22 | from | 9 |

# Removed Stop Words

words with greater than 40 occurrence has been removed and the result is as follows positive data

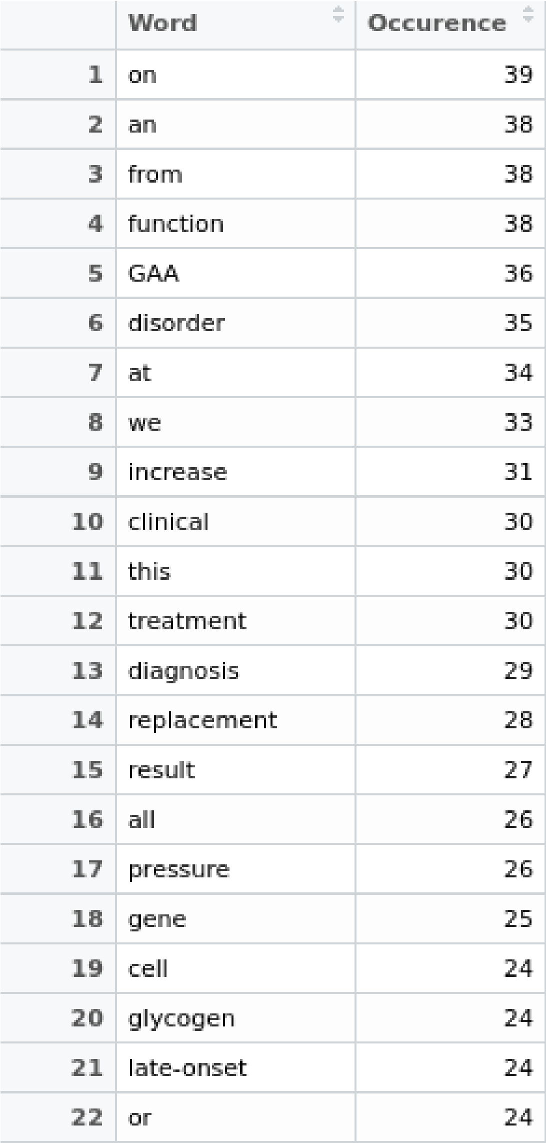


| | Word | Occurence |
|---|---|---|
| 1 | muscle | 39 |
| 2 | disease | 35 |
| 3 | patient | 35 |
| 4 | to | 35 |
| 5 | respiratory | 29 |
| 6 | Pompe | 28 |
| 7 | have | 21 |
| 8 | by | 20 |
| 9 | weakness | 16 |
| 10 | that | 14 |
| 11 | failure | 12 |
| 12 | progressive | 12 |
| 13 | case | 11 |
| 14 | for | 11 |
| 15 | level | 11 |
| 16 | LOPD | 10 |
| 17 | diagnosis | 9 |
| 18 | disorder | 9 |
| 19 | ERT | 9 |
| 20 | from | 9 |
| 21 | late-onset | 9 |
| 22 | muscular | 9 |

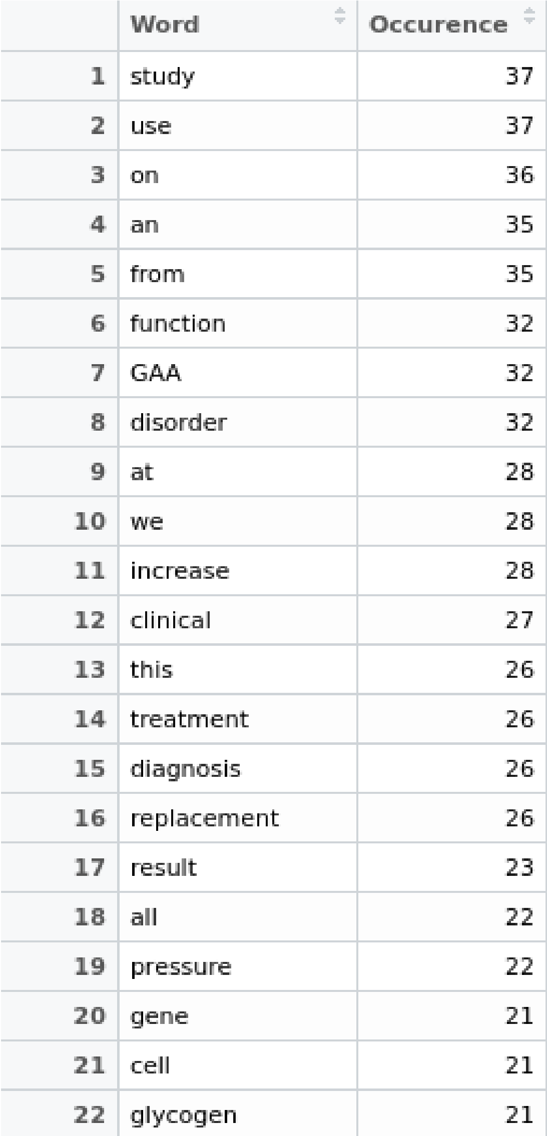


| | Word | Occurence |
|---|---|---|
| 1 | in | 39 |
| 2 | the | 37 |
| 3 | with | 32 |
| 4 | be | 28 |
| 5 | a | 27 |
| 6 | muscle | 27 |
| 7 | disease | 27 |
| 8 | patient | 27 |
| 9 | to | 24 |
| 10 | respiratory | 24 |
| 11 | Pompe | 17 |
| 12 | have | 16 |
| 13 | by | 15 |
| 14 | weakness | 11 |
| 15 | that | 11 |
| 16 | failure | 11 |
| 17 | progressive | 10 |
| 18 | case | 10 |
| 19 | for | 9 |
| 20 | level | 9 |
| 21 | LOPD | 9 |
| 22 | diagnosis | 9 |

positive data



| | Word | Occurence |
|---|---|---|
| 1 | on | 39 |
| 2 | an | 38 |
| 3 | from | 38 |
| 4 | function | 38 |
| 5 | GAA | 36 |
| 6 | disorder | 35 |
| 7 | at | 34 |
| 8 | we | 33 |
| 9 | increase | 31 |
| 10 | clinical | 30 |
| 11 | this | 30 |
| 12 | treatment | 30 |
| 13 | diagnosis | 29 |
| 14 | replacement | 28 |
| 15 | result | 27 |
| 16 | all | 26 |
| 17 | pressure | 26 |
| 18 | gene | 25 |
| 19 | cell | 24 |
| 20 | glycogen | 24 |
| 21 | late-onset | 24 |
| 22 | or | 24 |



| | Word | Occurence |
|---|---|---|
| 1 | study | 37 |
| 2 | use | 37 |
| 3 | on | 36 |
| 4 | an | 35 |
| 5 | from | 35 |
| 6 | function | 32 |
| 7 | GAA | 32 |
| 8 | disorder | 32 |
| 9 | at | 28 |
| 10 | we | 28 |
| 11 | increase | 28 |
| 12 | clinical | 27 |
| 13 | this | 26 |
| 14 | treatment | 26 |
| 15 | diagnosis | 26 |
| 16 | replacement | 26 |
| 17 | result | 23 |
| 18 | all | 22 |
| 19 | pressure | 22 |
| 20 | gene | 21 |
| 21 | cell | 21 |
| 22 | glycogen | 21 |