Vocabulary Analysis of Project Gutenberg

Zachary Booth Simpson
May 2000

(c)2002 ZBS. http://www.mine-control.com/zack
Please sign my guestbook if you find this work useful.

Introduction

While reading Moby Dick in April 2000, I was astounded by Melville's enormous vocabulary. I wondered what was Moby Dick's total vocabulary and how it compared to other works. Thanks to the Project Gutenberg, an online resource for literature, (and copious spare-time) I was able to download a considerable sample of works and perform a word analysis. The following are the results from this informal study including relative vocabulary densities and anomalous word usage.

Please sign my guestbook if you find this interesting or helpful. Thanks, Zack.

Sample Database

The works represented in this study come exclusively from the Project Gutenberg (PG). While most PG works are included, the sample is not complete; some works have been eliminated for obvious reasons (e.g. Pi to 10,000 digits) while others works were eliminated because they were malformed or unavailable. Some books in the Project Gutenberg are split into several separate volumes or alternatively several works are combined into one; this may effect the sample slightly, especially the Anomalous Word Charts. In some cases, I have manually combined multiple volumes into one for logical consistency.

The sample index was derived from the Thallason Index of the Project Gutenberg because the master indices from the PG itself were inconsistent. I extend my thanks to their efforts as well as to all contributors to the Project Gutenberg.

View the sample database by TITLE
View the sample database by AUTHOR

NOTE: Due to a change of server, I no longer have sufficient room to store the entire sample database online. My apologies.

Total Vocabulary

'Total Vocabulary' is the measure of unique words in a book. A word is defined as a set of case-insensitive alpha characters and apostrophes (to include contractions such as can't) thus excludes numbers and punctuation. Each work is scanned in its entirety including titles, indices, and page numbers after eliminating the Gutenberg Preamble which prefixes each work.

Largest Vocabularies (Regardless Of Book Size)

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Vocabulary Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decline and Fall of the Roman Empire, vol 1-6</td>
<td>Gibbon, Edward</td>
<td>43113</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>Anonymous / Various</td>
<td>39023</td>
</tr>
<tr>
<td>Gargantua and Pantagruel</td>
<td>Rabelais, Francis</td>
<td>25985</td>
</tr>
<tr>
<td>1998 CIA World Factbook, The</td>
<td>US CIA</td>
<td>24220</td>
</tr>
<tr>
<td>Les Miserables</td>
<td>Hugo, Victor</td>
<td>23334</td>
</tr>
<tr>
<td>Anomalies and Curiosities of Medicine</td>
<td>Gould/Pyle</td>
<td>22930</td>
</tr>
<tr>
<td>Brann The Iconoclast, vol 1,10,12</td>
<td>Brann, William Cowper</td>
<td>22507</td>
</tr>
<tr>
<td>Plutarch's Lives, trans by A. H. Clough</td>
<td>Plutarch</td>
<td>20237</td>
</tr>
<tr>
<td>History Of The Conquest Of Peru (2nd ver), The</td>
<td>Prescott, William H.</td>
<td>19235</td>
</tr>
<tr>
<td>Warfare of Science/Theology</td>
<td>White, Andrew Dickson</td>
<td>19187</td>
</tr>
</tbody>
</table>
Vocabulary Density

'Vocabulary Density' is a measurement of vocabulary usage in comparison to the length of the book. This ratio is expressed as the 'Inverse Absolute Vocabulary Density' and is computed dividing the Total Words by the Unique Words (W/V). This statistic may be thought of as: 'how many words will be read on average before a new word is encountered.' For example, Moby Dick has a (W/V) score of approximately 12 -- a new word is introduced on approximately every line of the book! That is quite an accomplishment for a work that is almost a quarter of a million words long.

Ideally, the (W/V) statistic allows comparison of one book's style to another. However, this simplistic metric is complicated by the simple fact that a short work will inevitably be denser than a larger work due to the fact that practically every word in a short work is unique. To understand, consider the case of writing a multi-million word essay. Given that there are only a limited number of words in the English language (~400,000 in this sample), one would eventually run out of words and thus the vocabulary density of such a titanic treatise would drop accordingly. This effect can be seen in the flattening trend of the scatter plots seen below.

Scatter-plots of inverse vocabulary density (y-axis) vs. total words (x-axis). Samples below the pink trend line have denser vocabularies than average, those above, sparser. Note that trend line fits less well for smaller works.

In order that the vocabulary densities of large and small works may be compared, a 'normalizing' curve is fit to the sample creating a 'normalized density score' useful for comparison. Unfortunately, the one-size-fits-all trend curve (found empirically by minimizing least mean square error of a square-root scale coefficient) fails to fit the smaller works well as can be seen in Figure 3. Thus, comparison of large works (> 30,000 words) to smaller ones (< 30,000) is ill-advised. Therefore, the following tables isolate these two sample groups.

NOTE: Due to a change of server, I no longer have sufficient room to store the entire sample database on-line. My apologies.

Most Dense Vocabularies, Normalized For Book Size. Books Over 30,000 Words

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Vocabulary Words</th>
<th>Normal Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decline and Fall of the Roman Empire, vol 1-6</td>
<td>Gibbon, Edward</td>
<td>43113</td>
<td>1543676</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>Anonymous / Various</td>
<td>39023</td>
<td>203886</td>
</tr>
<tr>
<td>Gargantua and Pantagruel</td>
<td>Rabelais, Francis</td>
<td>25985</td>
<td>323013</td>
</tr>
<tr>
<td>Brann The Iconoclast, vol 1,10,12</td>
<td>Brann, William Cowper</td>
<td>22507</td>
<td>300783</td>
</tr>
<tr>
<td>1998 CIA World Factbook, The</td>
<td>US CIA</td>
<td>24220</td>
<td>422744</td>
</tr>
</tbody>
</table>

http://mine-control.com/zack/guttenberg/index.html
### Anomalies and Curiosities of Medicine
- **Author:** Gould/Pyle
- **Words:** 22930
- **Normal Density:** -7.43

### Hackers' Dictionary of Computer Jargon, The
- **Author:** Anonymous / Various
- **Words:** 16757
- **Normal Density:** -6.03

### History Of The Conquest Of Peru (2nd ver), The
- **Author:** Prescott, William H.
- **Words:** 19235
- **Normal Density:** -5.87

### Moby Dick
- **Author:** Melville, Herman
- **Words:** 17227
- **Normal Density:** -5.75

### Warfare of Science/Theology
- **Author:** White, Andrew Dickson
- **Words:** 19187
- **Normal Density:** -5.46

### Poems And Songs Of Robert Burns
- **Author:** Burns, Robert
- **Words:** 14968
- **Normal Density:** -5.46

### Les Miserables
- **Author:** Hugo, Victor
- **Words:** 23334
- **Normal Density:** -5.17

### Travels through France & Italy
- **Author:** Smollett, Tobias
- **Words:** 14625
- **Normal Density:** -5.05

### Waverley
- **Author:** Scott, Walter
- **Words:** 15325
- **Normal Density:** -4.79

### Poems And Songs Of Robert Burns
- **Author:** Burns, Robert
- **Words:** 14968
- **Normal Density:** -5.46

### Least Dense Vocabularies, Normalized For Book Size. Books Over 30,000 Words

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Vocabulary Words</th>
<th>Normal Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book of Mormon, The</td>
<td>Anonymous / Various</td>
<td>5612</td>
<td>275887</td>
</tr>
<tr>
<td>Bible, Both Testaments, King James Version, The</td>
<td>Anonymous / Various</td>
<td>12867</td>
<td>790126</td>
</tr>
<tr>
<td>Le Morte D'Arthur, vol 2</td>
<td>Malory, Thomas</td>
<td>5717</td>
<td>194249</td>
</tr>
<tr>
<td>Bible, Douay-Rheims Version, Challoner Revision, The</td>
<td>Anonymous / Various</td>
<td>18559</td>
<td>1029084</td>
</tr>
<tr>
<td>Le Morte D'Arthur, vol 1</td>
<td>Malory, Thomas</td>
<td>5826</td>
<td>169703</td>
</tr>
<tr>
<td>Treaty of the European Union [Maastricht], The</td>
<td>Anonymous / Various</td>
<td>2826</td>
<td>59469</td>
</tr>
<tr>
<td>Nada the Lily</td>
<td>Haggard, H. Rider</td>
<td>5040</td>
<td>117857</td>
</tr>
<tr>
<td>White Knight: Tirant Lo Blanc (tr R.S. Rudder), The</td>
<td>Martorell, Joanot</td>
<td>6343</td>
<td>161871</td>
</tr>
<tr>
<td>Story of Burnt Njal (Njal's Saga) Icelandic, The</td>
<td>Anonymous / Various</td>
<td>5468</td>
<td>129135</td>
</tr>
<tr>
<td>Moll Flanders</td>
<td>Defoe, Daniel</td>
<td>6139</td>
<td>139300</td>
</tr>
<tr>
<td>Heimskringla [Norwegian Kings]</td>
<td>Sturlson, Snorri</td>
<td>10405</td>
<td>306474</td>
</tr>
<tr>
<td>Twilight Land</td>
<td>Pyle, Ernie Howard</td>
<td>4113</td>
<td>74003</td>
</tr>
<tr>
<td>First Book of Adam and Eve</td>
<td>Platt, Rutherford</td>
<td>2287</td>
<td>32820</td>
</tr>
<tr>
<td>On the Origin of Species</td>
<td>Darwin, Charles</td>
<td>6993</td>
<td>155549</td>
</tr>
<tr>
<td>Personal Memoirs of U.S. Grant, vol 2</td>
<td>Grant, Ulysses S.</td>
<td>6965</td>
<td>154177</td>
</tr>
<tr>
<td>Two Years in the Forbidden City</td>
<td>Der Ling, Princess</td>
<td>4962</td>
<td>92456</td>
</tr>
<tr>
<td>Princess of Cleves, The</td>
<td>Lafayette, Madame de</td>
<td>3779</td>
<td>61809</td>
</tr>
<tr>
<td>Emma</td>
<td>Austen, Jane</td>
<td>7228</td>
<td>161099</td>
</tr>
<tr>
<td>Flower Fables</td>
<td>Alcott, Louisa May</td>
<td>2501</td>
<td>34525</td>
</tr>
<tr>
<td>Parmenides</td>
<td>Plato</td>
<td>2616</td>
<td>36337</td>
</tr>
</tbody>
</table>

### Most Dense Vocabularies, Normalized For Book Size. Books Under 30,000 Words

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Vocabulary Words</th>
<th>Normal Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biog Study of A. W. Kinglake</td>
<td>Tikwell, Rev. W.</td>
<td>6794</td>
<td>29001</td>
</tr>
<tr>
<td>Waifs and Strays, etc</td>
<td>O'Henry</td>
<td>5826</td>
<td>29482</td>
</tr>
<tr>
<td>50 Bab Ballads (vol 1)</td>
<td>Gilbert, W.S.</td>
<td>5689</td>
<td>28588</td>
</tr>
<tr>
<td>Style</td>
<td>Raleigh, Walter</td>
<td>5385</td>
<td>24331</td>
</tr>
<tr>
<td>Cicero's Orations [selected orations in Latin] Cicero</td>
<td>Anonymous / Various</td>
<td>4525</td>
<td>13219</td>
</tr>
<tr>
<td>New Poems</td>
<td>Thompson, Francis</td>
<td>5392</td>
<td>25181</td>
</tr>
<tr>
<td>Chita: A Memory of Last Island</td>
<td>Hearn, Lafcadio</td>
<td>5495</td>
<td>26874</td>
</tr>
</tbody>
</table>

http://mine-control.com/zack/guttenberg/index.html
Word Anomalies

It would be interesting to know for a given book what words are used uncommonly often or, likewise, uncommonly infrequently. To compute this, the relative frequency of each word is sampled from the database at large and then compared to the frequency in each book.

Not surprisingly, these 'Anomalous Word Summaries' paint an incredibly accurate picture of the work. For example, among Moby Dick's most anomalous words are: whale, sperm, and harpooneer. Of course, proper names tend to dominate these lists; for example, ahab, stubb, and queeqqueg top out Moby Dick. Just as interesting is what the book is NOT about. Among Moby Dick's most infrequently used words (i.e. words which are common in other books, but not in this one) are: miss, government, happiness, smiled, and machine.

The Infrequently Used Summaries list only words which are actually used in the work. While it might be logical to list words that are frequently used in other books but that never show up in this book, it would be useless because such a list
would be dominated by anachronistic words such as 'thou' and 'thy' that are common in the database but unused in most works.

Misspellings significantly skew both the Infrequent and Unique Word Lists and are fairly common due to the use of Optical Character Recognition (OCR) software which is extremely prone such mistakes.

The following table is a sample of Word Anomalies picked by hand from the database to illustrate the technique. To view Anomaly Summaries for any work, click on the book name in either the author index or title index.

NOTE: Due to a change of server, I no longer have sufficient room to store the entire sample database on-line. My apologies.

View the index by TITLE
View the index by AUTHOR
(Click on any title to view the Anomaly Summary)

### Sample of Word Anomalies

**The Bible (King James Edition); Anonymous / Various**
- **Frequent:** unto, lord, isreal, shall, god, moses, jesus, david, offering, tabernacle
- **Infrequent:** girl, boy, school, success, condition, listen, princess

**Wonderful Wizard of Oz; Baum, Frank**
- **Frequent:** woodman, scarecrow, witch, tin, emerald, monkeys, kansas, brains, winged
- **Infrequent:** mother, money, soul, natural

**White Fang; London, Jack**
- **Frequent:** musher, beaver, sled, dogs, cherokee, snarl
- **Infrequent:** letter, person, window, green, sweet, loved, party, paper

**The Republic; Plato**
- **Frequent:** guardians, unjust, true, injustice, state, gymnastic, rulers, democractical
- **Infrequent:** miss, girl, boy, prince

**Alice's Adventures In Wonderland; Carroll (C.L. Dodgson), Lewis**
- **Frequent:** gryphon, turtle, caterpiller, mock, dodo, mouse, rabbit, hedgehog
- **Infrequent:** death, country, happy, fair, common

**Origin of the Species; Darwin, Charles**
- **Frequent:** species, varieties, subaerial, selection, sterility, plants, modification, forms, variability
- **Infrequent:** person, government, love, thinking, god, evil, fire

**Communist Manifesto; Marx, Karl/Engels, Friedrich**
- **Frequent:** bourgeois, proletariat, communists, antagonisms, revolutionising, socialism, production, class, feudal, reactionary, exploitation, conditions, crises
- **Infrequent:** said, love, why, heart, mother, poor, felt

**Paradise Lost; Milton, John**
- **Frequent:** wonderful, heaven, satan, dominations
- **Infrequent:** country, church, horses, sister

**Apology; Plato**
- **Frequent:** corrupter, accusers, demigods, socrates, oracle, indictment
- **Infrequent:** she, work, morning, replied, body

**Gargantua and Pantagruel; Rabelais, Francis**
- **Frequent:** codpiece, catchpole, ballocks, dingdong, fart, chatterlings, gymnast, arse
- **Infrequent:** smile, existence, feelings, british, professor, suffering

**1st Inaugural Speech; Roosevelt, Franklin Delano**
- **Frequent:** foreclosure, interdependence, uneconomical, leadership, outgo, unsolvable, values, redistribution, national, emergency
- **Infrequent:** you, her, his

**The Jungle; Sinclair, Upton**
- **Frequent:** packingtown, packers, stockyards, fertilizer, slaughterhouses, streetcar, lituanian
- **Infrequent:** influence, village, pray, gods, example

**20,000 Leagues Under The Sea; Verne, Jules**
Frequent: manometer, canadian, captain, frigate, harpoon, cuttlefish, submarine
Infrequent: garden, justice, ladies, laughed, wife

Time Machine; Wells, H. G.
Frequent: psychologist, sphinx, traveller, machine, i, lever, dimension
Infrequent: mother, dear, money, friends, horse, peace

War of the Worlds; Wells, H. G.
Frequent: martians, leatherhead, artilleryman, londonward, cylinder, pit, scullery
Infrequent: love, king, truth, gentleman, joy, youth

Moby Dick; Melville, Herman
Frequent: whale, sperm, harpooner, pequod, leviathan, fishery
Infrequent: miss, fortune, happiness, smiled, angry, enemies
The Latin element in the English vocabulary consists of a large number of words of Latin origin, adopted directly into English at various periods. The principal periods, during which Latin words were brought directly into English are:

1. At the introduction of Christianity into England by the Latin Catholic missionaries, A.D. 596.

The use of Project Gutenberg (PG) as a text corpus has been extremely popular in statistical analysis of language for more than 25 years. However, in contrast to other major linguistic datasets of similar importance, no consensual full version of PG exists to date. In fact, most PG studies so far either consider only a small number of manually selected books, leading to potential biased subsets, or employ vastly different pre-processing strategies (often specified in insufficient details), raising concerns regarding the reproducibility of published results. Project Gutenberg began in 1971 by Michael Hart as a community project to make plain text versions of books available freely to all.
Quantitative narrative analysis (QNA) is used to explore a cleaned subcorpus, the Gutenberg English Poetry Corpus (GEPC), which comprises over 100 poetic texts with around two million words from about 50 authors (e.g., Keats, Joyce, Wordsworth). The GLEC provides a collection of over 3,000 English texts from the Gutenberg project, spanning a wide range of genres, both fiction and non-fiction (novels, biographies, dramas, essays, short stories, novellas, tales, speeches and letters, science books, poetry; e.g., Austen, Bronte, Byron, Coleridge, Darwin, Dickens, Einstein, Eliot, Poe, Twain, Woolf, Wilde, Yeats) with about 12 million sentences and 250 million. Previous (Prohibition on alcohol (United States)). Next (Projectile). Project Gutenberg, abbreviated as PG, is a volunteer effort to digitize, archive, and distribute cultural works. Founded in 1971, by Michael S. Hart, it is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of December 2007, Project Gutenberg claimed Project Gutenberg is a volunteer effort to digitize and archive cultural works, to "encourage the creation and distribution of eBooks." It was founded in 1971 by American writer Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of public domain books. A majority of the titles were originally published before 1950, as these titles do not fall under copyright protections. I engineered 16 style metrics indicative of vocabulary richness, lexical complexity, sentence and word length, part-of-speech ratios, and readability scores:

Standardized Type-Token Ratio: Type-token ratio (TTR) is a measure of vocabulary richness. It is calculated by dividing the number of unique tokens in a document (types) by the total number of tokens in the document. The use of Project Gutenberg (PG) as a text corpus has been extremely popular in statistical analysis of language for more than 25 years. However, in contrast to other major linguistic datasets of similar importance, no consensual full version of PG exists to date. @article{Gerlach2020ASP, title={A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics}, author={M. Gerlach and Francesc Font-Clos}, journal={Entropy}, year={2020}, volume={22} }. M. Gerlach, Francesc Font-Clos. Published 2020. Computer Science, Physics, Medicine. Entropy. The use of Project Gutenberg (PG) as a text corpus has been extremely popular in statistical analysis of language for more than 25 years.