# Vocabulary Analysis of Project Gutenberg

**Zachary Booth Simpson**
**May 2000**



## 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 guestbo0k 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 seperate 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 on-line. 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)

| Title (click on work to view word anomalies) | Author | Vocabulary Words | |
|---|---|---|---|
| Decline and Fall of the Roman Empire, vol 1-6 | Gibbon, Edward | 43113 | 1543676 |
| Roget's Thesaurus | Anonymous / Various | 39023 | 203886 |
| Gargantua and Pantagruel | Rabelais, Francis | 25985 | 323013 |
| 1998 CIA World Factbook, The | US CIA | 24220 | 422744 |
| Les Miserables | Hugo, Victor | 23334 | 570508 |
| Anomalies and Curiosities of Medicine | Gould/Pyle | 22930 | 393856 |
| Brann The Iconoclast, vol 1,10,12 | Brann, William Cowper | 22507 | 300783 |
| Plutarch's Lives, trans by A. H. Clough | Plutarch | 20237 | 742013 |
| History Of The Conquest Of Peru (2nd ver), The | Prescott, William H. | 19235 | 300976 |
| Warfare of Science/Theology | White, Andrew Dickson | 19187 | 322799 |

| | | | |
|---|---|---|---|
| Bible, Douay-Rheims Version, Challoner Revision, The | Anonymous / Various | 18559 | 1029084 |
| Moby Dick | Melville, Herman | 17227 | 211763 |
| Cloister and the Hearth, The | Reade, Charles | 16911 | 282120 |
| Hackers' Dictionary of Computer Jargon, The | Anonymous / Various | 16757 | 169716 |
| Sketches by Boz | Dickens, Charles | 16413 | 262440 |
| Vanity Fair | Thackeray, William Makepeace | 16349 | 360049 |
| Our Mutual Friend | Dickens, Charles | 16337 | 338266 |
| Dombey and Son | Dickens, Charles | 16332 | 366517 |
| Pickwick Papers, The | Dickens, Charles | 16253 | 313143 |
| Don Quixote (tr John Ormsby) | Cervantes | 16160 | 425814 |
| Count of Monte Cristo, The | Dumas, père, Alexandre | 16110 | 464256 |
| Terminal Compromise/NetNovel | Schartau, Win | 15898 | 213672 |

# 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.

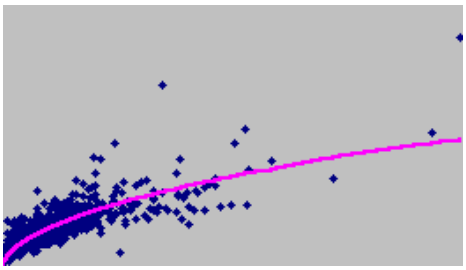

**Figure 1. 800,000 word domain**

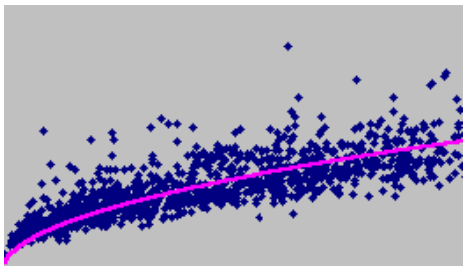

**Figure 2. 100,000 word domain**

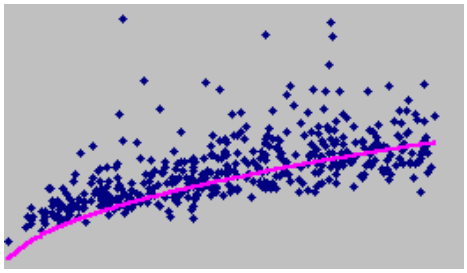

**Figure 3. 30,000 word domain**

**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

| Title (click on work to view word anomalies) | Author | Vocabulary | Words | Normal Density |
|---|---|---|---|---|
| Decline and Fall of the Roman Empire, vol 1-6 | Gibbon, Edward | 43113 | 1543676 | -12.92 |
| Roget's Thesaurus | Anonymous / Various | 39023 | 203886 | -12.48 |
| Gargantua and Pantagruel | Rabelais, Francis | 25985 | 323013 | -9.86 |
| Brann The Iconoclast, vol 1,10,12 | Brann, William Cowper | 22507 | 300783 | -8.14 |
| 1998 CIA World Factbook, The | US CIA | 24220 | 422744 | -8.04 |

| | | | | |
|---|---|---|---|---|
| Anomalies and Curiosities of Medicine | Gould/Pyle | 22930 | 393856 | -7.43 |
| Hackers' Dictionary of Computer Jargon, The | Anonymous / Various | 16757 | 169716 | -6.03 |
| History Of The Conquest Of Peru (2nd ver), The | Prescott, William H. | 19235 | 300976 | -5.87 |
| Moby Dick | Melville, Herman | 17227 | 211763 | -5.75 |
| Warfare of Science/Theology | White, Andrew Dickson | 19187 | 322799 | -5.46 |
| Poems And Songs Of Robert Burns | Burns, Robert | 14968 | 129551 | -5.46 |
| Les Miserables | Hugo, Victor | 23334 | 570508 | -5.17 |
| Travels through France & Italy | Smollett, Tobias | 14625 | 142922 | -5.05 |
| Waverley | Scott, Walter | 15325 | 185273 | -4.79 |
| Tracks of a Rolling Stone | Coke, Henry J. | 13259 | 106525 | -4.77 |
| Terminal Compromise/NetNovel | Schartau, Win | 15898 | 213672 | -4.69 |
| Main Street | Lewis, Sinclair | 14580 | 169912 | -4.51 |
| Sketch-Book of Geoffrey Crayon, The | Irving, Washington | 13241 | 129907 | -4.32 |
| Devil's Dictionary, The | Bierce, Ambrose | 11172 | 60906 | -4.23 |
| Roads of Destiny | O'Henry | 12196 | 98952 | -4.22 |
| Leaves of Grass | Whitman, Walt | 12924 | 124036 | -4.21 |
| Lucasta Poems, The | Lovelace, Richard | 11153 | 62900 | -4.20 |

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

| Title (click on work to view word anomalies) | Author | Vocabulary | Words | Normal Density |
|---|---|---|---|---|
| Book of Mormon, The | Anonymous / Various | 5612 | 275887 | 28.56 |
| Bible, Both Testaments, King James Version, The | Anonymous / Various | 12867 | 790126 | 26.55 |
| Le Morte D'Arthur, vol 2 | Malory, Thomas | 5717 | 194249 | 16.69 |
| Bible, Douay-Rheims Version, Challoner Revision, The | Anonymous / Various | 18559 | 1029084 | 15.67 |
| High History of the Holy Graal, The | Anonymous / Various | 5327 | 158488 | 14.14 |
| Le Morte D'Arthur, vol 1 | Malory, Thomas | 5826 | 169703 | 12.97 |
| Treaty of the European Union [Maastricht], The | Anonymous / Various | 2826 | 59469 | 11.48 |
| Nada the Lily | Haggard, H. Rider | 5040 | 117857 | 9.92 |
| White Knight: Tirant Lo Blanc (tr R.S. Rudder), The | Martorell, Joanot | 6343 | 161871 | 9.74 |
| Story of Burnt Njal (Njal's Saga) Icelandic, The | Anonymous / Various | 5468 | 129135 | 9.52 |
| Moll Flanders | Defoe, Daniel | 6139 | 139300 | 8.05 |
| Heimskringla [Norwegian Kings] | Sturlson, Snorri | 10405 | 306474 | 7.74 |
| Twilight Land | Pyle, Ernie Howard | 4113 | 74003 | 7.32 |
| First Book of Adam and Eve | Platt, Rutherford | 2287 | 32820 | 7.25 |
| On the Origin of Species | Darwin, Charles | 6993 | 155549 | 6.78 |
| Personal Memoirs of U.S. Grant, vol 2 | Grant, Ulysses S. | 6965 | 154177 | 6.74 |
| Two Years in the Forbidden City | Der Ling, Princess | 4962 | 92456 | 6.71 |
| United States Copyright Act of 1976, The | Anonymous / Various | 2271 | 30635 | 6.63 |
| Princess of Cleves, The | Lafayette, Madame de | 3779 | 61809 | 6.61 |
| Emma | Austen, Jane | 7228 | 161099 | 6.55 |
| Flower Fables | Alcott, Louisa May | 2501 | 34525 | 6.52 |
| Parmenides | Plato | 2616 | 36337 | 6.41 |

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

| Title (click on work to view word anomalies) | Author | Vocabulary | Words | Normal Density |
|---|---|---|---|---|
| Biog Study of A. W. Kinglake | Tcikwell, Rev. W. | 6794 | 29001 | -2.41 |
| Waifs and Strays, etc | O'Henry | 5826 | 29482 | -1.67 |
| 50 Bab Ballads (vol 1) | Gilbert, W.S. | 5689 | 28588 | -1.61 |
| Style | Raleigh, Walter | 5385 | 24331 | -1.60 |
| Cicero's Orations [selected orations in Latin] | Cicero | 4525 | 13219 | -1.59 |
| New Poems | Thompson, Francis | 5392 | 25151 | -1.55 |
| Chita: A Memory of Last Island | Hearn, Lafcadio | 5495 | 26874 | -1.54 |

| | | | | |
|---|---|---|---|---|
| Poems | Henley, William E. | 5301 | 24303 | -1.53 |
| Georgics [English], The | Virgil | 5089 | 21668 | -1.51 |
| Letters on Literature | Lang, Andrew | 5550 | 29479 | -1.42 |
| Shelley | Waterlow, Sydney | 4907 | 21390 | -1.38 |
| Sword Blades and Poppy Seed | Lowell, Amy | 5255 | 26996 | -1.31 |
| Bab Ballads, vol 2, The | Gilbert, W.S. | 4769 | 20582 | -1.31 |
| Who Was Who: 5000 B. C. to Date | Gordon, Irwin L. | 4802 | 21807 | -1.25 |
| Ginx's Baby, A Satire | Jenkins, Edward | 5370 | 29763 | -1.22 |
| Bab Ballads, vol 3, The | Gilbert, W.S. | 4853 | 23153 | -1.20 |
| Foolish Dictionary, The | Wurdz, Gideon | 3826 | 11615 | -1.19 |
| Lays of Ancient Rome | Macaulay, Thomas Babbington | 4987 | 25043 | -1.18 |
| Essay on Comedy, Comic Spirit | Meredith, George | 4344 | 17204 | -1.18 |
| Reginald in Russia and Other Sketches | Saki (H.H. Munro) | 4715 | 22184 | -1.14 |
| Philobiblon of Richard de Bury, The | Bury, Richard de | 4921 | 24906 | -1.13 |
| Reading of Life, and Other Poems, A | Meredith, George | 3878 | 12990 | -1.12 |

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

| Title (click on work to view word anomalies) | Author | Vocabulary | Words | Normal Density |
|---|---|---|---|---|
| New McGuffey First Reader, The | McGuffey (compiler), W.H. | 630 | 8276 | 9.57 |
| Ethics, part 2 (tr Elwes) | Spinoza, Benedict de | 1485 | 18314 | 7.03 |
| Ethics, part 3 (tr Elwes) | Spinoza, Benedict de | 1866 | 22877 | 6.33 |
| Somebody's Little Girl | Young, Martha | 983 | 9795 | 6.08 |
| Ethics, part 1 (tr Elwes) | Spinoza, Benedict de | 1422 | 14046 | 5.23 |
| Berne Universal Copyright Convention [1988], The | Anonymous / Various | 967 | 8023 | 4.78 |
| Adventures of Reddy Fox | Burgess, Thornton W. | 1572 | 14948 | 4.71 |
| Ethics, part 5 (tr Elwes) | Spinoza, Benedict de | 1269 | 10805 | 4.44 |
| Well of the Saints, The | Synge, J.M. | 1695 | 15540 | 4.28 |
| Lady Windermere's Fan | Wilde, Oscar | 2056 | 19942 | 4.16 |
| Organic Syntheses | Conant (Editor), James Bryant | 2202 | 21695 | 4.08 |
| Alice's Adventures in Wonderland | Carroll (C.L. Dodgson), Lewis | 2649 | 27785 | 3.95 |
| White People, The | Burnett, Frances Hodgson | 2262 | 21593 | 3.78 |
| Dreams | Schreiner, Olive | 2137 | 19817 | 3.75 |
| True Story of Christopher Columbus, The | Brooks, Elbridge S. | 2805 | 29141 | 3.69 |
| Woman of No Importance, A | Wilde, Oscar | 2374 | 22496 | 3.59 |
| Meno, second part | Plato | 933 | 6200 | 3.56 |
| Tom Sawyer Detective | Twain (Samuel Clemens), Mark | 2475 | 23467 | 3.47 |
| Rosmersholm | Ibsen, Henrik | 2815 | 28053 | 3.40 |
| Ballad of Reading Gaol | Wilde, Oscar | 1196 | 8287 | 3.36 |
| Deirdre of the Sorrows | Synge, J.M. | 1968 | 16415 | 3.32 |
| Story of Doctor Dolittle, The | Lofting, Hugh | 2759 | 26835 | 3.30 |

## 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 words 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 queequeg 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:** wonderous, 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, chitterlings, 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

Project Gutenbergâ€™s The Mathematical Analysis of Logic, by George Boole. This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org. Title: The Mathematical Analysis of Logic Being an Essay Towards a Calculus of Deductive Reasoning. Author: George Boole. Release Date: July 28, 2011 [EBook #36884]. Language: English. Character set encoding: ISO-8859-1. *** start of this project gutenberg ebook the mathematica Start of this project gutenberg ebook new word-analysis ***. Produced by Keith Edkins and the Online Distributed Proofreading Team at http://www.pgdp.net. NEW WORD-ANALYSISÂ  6. 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:â€” 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.Â  Book from Project Gutenberg: The Hacker Crackdown, law and disorder on the electronic frontier Library of Congress Classification: HV Topics: Computer crimes -- United States, Computer programming -- United States -- Corrupt practices, Source: http://www.gutenberg.org/ebooks/101. Project Gutenberg. Project Gutenberg , in full Project Gutenberg Literary Archive Foundation , a nonprofit organization (since 2000) that maintains an electronic library of public domain works that have been digitized, or converted into e-books , by volunteers and archived for download from the organizationâ€™s Web site : www.gutenberg.org. The project got its start on July 4, 1971, when Michael Hart, a student at the University of Illinois , began typing the U.S. Declaration of Independence into the schoolâ€™s computer system for distribution free of charge. He soon followed with the works of William Shakespeare an...Â  Get a Britannica Premium subscription and gain access to exclusive content. Subscribe Now.

Project Gutenberg is zealously noncommercial, digitizes books in the public domain alone, and publishes an accurate rendition of the full electronic text (but not the formatting). The Million Book Project, Internet Archives, and Open Content Alliance are noncommercial, target public-domain books, but also aspire to process copyrighted â€œorphanâ€ books, and show users page images backed up by full-text searching.Â  Genesis. In the beginning was Project Gutenberg, and that began with the Declaration of Independence. The date was 4 July 1971, Michael Hart said that he'd been to a fireworks show that evening and didn't feel like going home â€“ opting instead for the computer lab at the University of Illinois, where there was good air conditioning. Project Gutenberg is a volunteer effort to digitize and archive cultural works, to "encourage the creation and distribution of eBooks.â€ 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. For more information on Project Gutenberg, see gutenberg.org/. ETL Pipeline. In order to obtain the raw text of all books available on Project Gutenberg, I used wget to scrape a Project Gutenberg mirror for all English-language .txt files. I obtained about 80,000 files from the mirror. I next needed to do two things. We show that this technique provides excellent results when applied to over 380 texts from the Project Gutenberg archives, as well as to two previously published data sets. Introduction. Modern methods of authorship attribution are reviewed for Russian techniques by Milov (1994) and for Western methods by Holmes (1998).Â  He shows that it gives substantially better results than the analysis of individual letters. In the present paper we apply the method to English texts including two published data sets. Application. We will consider three data sets to illustrate the technique described above: Authors of texts in English, obtained from the Project Gutenberg archives,Project Gutenberg web site: http://promo.net/pg/. Data from the Baayen et al. 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. Continues at: http://www.mine-control.com/zack/guttenberg/. Project Gutenberg's The Measurement of Intelligence, by Lewis Madison. Terman This eBook is for the use of anyone anywhere at no cost and with. almost no restrictions whatsoever. You may copy it, give it away or re-use. it under the terms of the Project Gutenberg License included with this. eBook or online at www.gutenberg.net Title: The Measurement of. Intelligence. An Explanation of and a Complete Guide for the Use of.