# Exact Methods in the Study of Language and Text

**Dedicated to Professor Gabriel Altmann**

**On the Occasion of His 75th Birthday**

**Edited by**

**Peter Grzybek & Reinhard Köhler**

**Mouton de Gruyter**
**Berlin – New York**

# Contents

# Text ranking by the weight of highly frequent words

*Ioan-Iovitz Popescu*

"I am ill at these numbers..."
*Hamlet Act 2, Scene 2*

Almost every scientist, by ordering their own published articles or those of others from the most to the least cited paper, will conclude that only the head of the list is truly significant and existent for the scientific community. I also did this with my papers when posting them in descending order on my website a few years ago (Popescu 2001). The question was if there exists a simple and objective "head cutoff" for this purpose. A proposal in this connection has only recently been set forth for the quantification of scientific output of individuals by a single and easily computable scientometric parameter (Hirsch 2005). This is the "$h$-index", defined as the number $h$ of papers with citation counts higher or equal to $h$. For instance, a scientist cumulating a $h$-index of, say, $h = 20$, will have published 20 papers that have received at least 20 citations each. Obviously, the corresponding Hirsch's point $H(h,h)$ on the (rank, frequency) citation curve appears as a "turning point", the closest to the (rank, frequency) origin, as illustrated in Figure 1.

Generally, by construction proper, the (rank, frequency) citation distribution starts with the rank number one, corresponding to the most highly cited paper ("there's one in every crowd") and ends with the rank equal to the total number of papers having at least one citation. Consequently, the total number of citations is given by the area under the (rank, frequency) citation curve. Hirsch also found that this area is proportional to $h^2$, i.e. *Total Citation Count* $= ah^2$, with the constant $a$ ranging between the values 3 and 5 for the papers in the field of physics. For university teachers in Physics, as suggested by Hirsch, a value of $h \approx 12$ would be a minimal threshold for an associate professor, while a value of at least $h \approx 18$ is needed for advancement to full professor. At the very top of this scale there are scientists cumulating up to about $h \approx 100$ for physical sciences and almost $h \approx 200$ for biological and

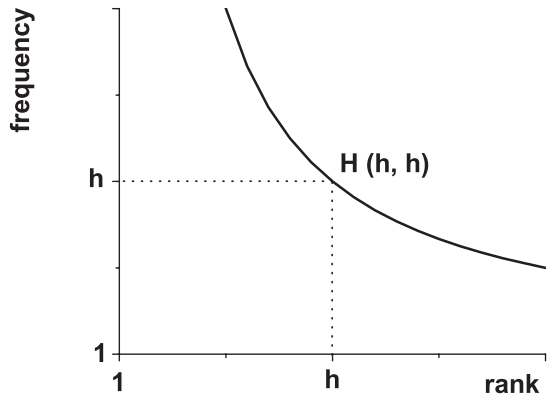


*Figure 1:* At Hirsch's point $H(h,h)$ the frequency and rank (always positive integers) have the same (or the closest possible) value, frequency = rank = $h$ being the $h$-index of the evaluated (rank, frequency) distribution

biomedical sciences. I will quote only the following relevant assertions from Hirsch's paper regarding the scientific output evaluation by the $h$-index: "I argue that two individuals with similar $h$ are comparable in terms of their overall scientific impact, even if their total number of papers or their total number of citations is very different. Conversely, that between two individuals (of the same scientific age) with similar number of total papers or of total citation count and very different $h$-values, the one with the higher $h$ is likely to be the 'better' scientist". To this, I will add, however, that perhaps a fairer assessment criterion would be the cumulated citation percentage of the most cited first $h$ papers out of the overall number of citations.

The present work is aimed to bring empirical arguments for the transfer of the $h$-index concept from scientometrics to linguistics, in other words, to switch the problem from paper citation ranking to word frequency ranking. Three main classes of web text sources were used for this purpose, namely The Bible (Table 1, p. 562), classical works (Table 2, p. 563), and Nobel lectures (Table 3, p. 565). More specifically, the (rank, frequency) word distributions of these widely known literary or scientific texts (see references) have been produced with the help of web available word frequency counters (see references) and processed and cleaned up of "non-words" with a Microsoft Excel program. Three important quantities describing the (rank, frequency)

word distribution were worked out in this way and introduced in the Tables 1 to 3 (p. 562ff.), as follows:

1. text length or total word count (equivalent to the total citations count), representing the area under the (rank, frequency) word curve from the first rank (rank one) up to the last rank (as given by the total number of unique words or the vocabulary), denoted in the corresponding table column headings as Total;
2. *h*-index for words, by analogy to that introduced by Hirsch (2005) for paper citations, indicates the "word distribution width" and is defined as the number *h* of unique words with counts higher or equal to *h*;
3. weight or percentage of the first *h* highly frequent words (*hfw*) out of the total word count (equivalent to the scientometric percentage of the first h highly cited papers).

Two other quantities, fixing the distribution scales, but not loading the tables, are (4) the vocabulary, giving the maximum value of the rank scale by the number of unique, different words, and (5) the value of the highest frequency of the word distribution, that is the number of words populating rank one, thus fixing the frequency scale.

A large variety of texts of various fields and of different size have been compared by sorting the data by these indicators. Thus, pasting the data of all mentioned three tables together, summing up a total of 151 texts, and sorting them by the first quantity (1) we can see that the investigated text lengths cover an interval between 53841 total word count of Goethe's *Faust 2* in Kline's English translation and 295 total word count of *The Third Epistle* of John. Likewise, sorting the data by the second quantity defined above (2), the *h*-index, we find out that its value ranges between the Books of Ezekiel or Jeremiah, both having $h = 83$, and again 3 John with $h = 6$. Generally, as expected, the rankings by text length and by *h*-index are closely similar, inasmuch as the square of the *h*-index represents a fairly accurate estimate of the total number of words according to the relationship *Total Word Count* $= ah^2$ (the proportionality constant, $a$, corresponding to the 151 tabulated texts, ranges between 4.5 and 9.5),

Last but not least, sorting the data by the weight of the first *h* highly frequent words (3), that is by the normalized word inventory "hard core", the joint listing reveals the top position of Bible texts (with 15 *Holy Books* having a *hfw* weight from 65 to 60 per cent), followed by classical texts (*hfw* weight from Newton's 63 and Einstein's 55 down to Dante's 40 per cent)

and, finally, by Nobel lecture texts (*hfw* weight from 47 to 27 per cent) and, almost within the same *hfw* bandwidth, current scientific papers, newspapers and random texts. In other words, the *hfw* criterion appears as a consistent estimator of the ineffable grace under which the text has been created. The present text, for instance, excepting tables and figures, has an *h*-index of 13 and a *hfw* percentage of 33.

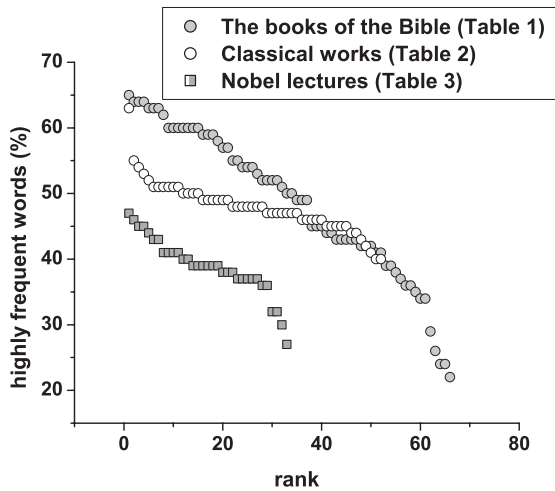


*Figure 2:* Three grace rankings as revealed by the weight of highly frequent words

Figure 2 illustrates graphically the separate *hfw* rankings of the above tabulated texts. Here again the ranking within and between the three considered text levels is evidenced. Clearly, more comparative research on similarities and differences of the word distributions is needed for a better understanding of the meaning of the *hfw* criterion and of the divine art of using the threads of highly frequent words in the text tapestery. Particularly striking appears the *hfw* synergism of various text parts as illustrated in Table 4 for Dostoevsky's *Crime and Punishment*. Thus, though the six novel parts have almost the same *hfw* percentage when taken separately, this value increases significantly when counted together. This and other related text features will be detailed elsewhere.

In summary, a simple and objective measure is proposed for the text evaluation by a single criterion, namely the percent of the cumulated number of

the first *h* decreasingly ranked words out of the total word count. Any (electronic) text can be evaluated in this way in a matter of seconds. The highest *hfw* synergies found sofar are *The Bible* (77 percent), *The Old Testament* (76 percent), *The Pentateuch* (72 per cent), *The New Testament* (70 percent), *The Four Gospels* (67 per cent), Dickens' *David Copperfield* (68 per cent) and *Great Expectations* (65 per cent), Tolstoy's *War and Peace* (65 per cent), Dostoevsky's *Crime and Punishment* (64 percent), Homer's *Iliad* (64 percent) and *Odyssey* (64 percent), and so on.

**Acknowledgments.** Thanks are due to linguist Professor Gabriel Altmann and to biophysicist Professor Daniel Lavalette for stimulating my interest in ranking matters, to physicists Professors Nicholas Ionescu-Pallas and Rudolf Emil Nistor for helpful discussions, and to chemist Professor Alexandru Balaban for pointing out Hirsch's recent scientometric paper.

*Table 1:* The books of the Bible sorted by decreasing weight of highly frequent words

| ID | Text | Total | *h* | *hfw* | ID | Text | Total | *h* | *hfw* |
|---|---|---|---|---|---|---|---|---|---|
| 3 | Leviticus | 24567 | 65 | 65 | 47 | 2 Corinthians | 6061 | 33 | 50 |
| 2 | Exodus | 32692 | 75 | 64 | 21 | Ecclesiastes | 5586 | 32 | 49 |
| 24 | Jeremiah | 42671 | 83 | 64 | 28 | Hosea | 5181 | 30 | 49 |
| 26 | Ezekiel | 39428 | 83 | 64 | 30 | Amos | 4217 | 28 | 49 |
| 4 | Numbers | 32918 | 76 | 63 | 33 | Micah | 3156 | 25 | 45 |
| 5 | Deuteronomy | 28377 | 66 | 63 | 49 | Ephesians | 3003 | 23 | 45 |
| 9 | 1 Samuel | 25051 | 71 | 63 | 58 | Hebrews | 6902 | 33 | 45 |
| 1 | Genesis | 38315 | 81 | 62 | 25 | Lamentations | 3409 | 25 | 44 |
| 6 | Joshua | 18858 | 58 | 60 | 48 | Galatians | 3083 | 25 | 44 |
| 11 | 1 Kings | 24507 | 64 | 60 | 8 | Ruth | 2567 | 22 | 43 |
| 12 | 2 Kings | 23521 | 61 | 60 | 36 | Zephaniah | 1617 | 18 | 43 |
| 14 | 2 Chronicles | 26085 | 65 | 60 | 39 | Malachi | 2567 | 22 | 43 |
| 23 | Isaiah | 37037 | 71 | 60 | 50 | Philippians | 2155 | 21 | 43 |
| 42 | Luke | 25942 | 65 | 60 | 52 | 1 Tessalonians | 1834 | 19 | 43 |
| 43 | John | 19116 | 62 | 60 | 29 | Joel | 2032 | 18 | 42 |
| 7 | Judges | 18953 | 59 | 59 | 37 | Haggai | 1124 | 13 | 42 |
| 10 | 2 Samuel | 20598 | 60 | 59 | 51 | Colossians | 1976 | 19 | 42 |
| 19 | Psalms | 41551 | 74 | 59 | 22 | Song of Solomon | 2656 | 21 | 41 |
| 40 | Matthew | 23696 | 63 | 58 | 60 | 1 Peter | 2471 | 21 | 41 |
| 13 | 1 Chronicles | 20350 | 50 | 57 | 53 | 2 Tessalonians | 1018 | 15 | 39 |
| 44 | Acts | 24262 | 66 | 57 | 59 | James | 2304 | 20 | 39 |
| 20 | Proverbs | 15056 | 49 | 55 | 31 | Obadiah | 665 | 11 | 38 |
| 41 | Mark | 15157 | 50 | 55 | 34 | Nahum | 1278 | 13 | 37 |
| 46 | 1 Corinthians | 9450 | 42 | 54 | 32 | Jonah | 1321 | 15 | 36 |
| 62 | 1 John | 2506 | 23 | 54 | 54 | 1 Timothy | 2244 | 19 | 36 |
| 66 | Revelation | 12001 | 45 | 54 | 35 | Habakkuk | 1463 | 15 | 35 |
| 18 | Job | 18107 | 51 | 53 | 55 | 2 Timothy | 1661 | 17 | 34 |
| 15 | Ezra | 7445 | 33 | 52 | 61 | 2 Peter | 1554 | 16 | 34 |
| 27 | Daniel | 11588 | 42 | 52 | 56 | Titus | 886 | 11 | 29 |
| 38 | Zechariah | 6449 | 33 | 52 | 63 | 2 John | 295 | 7 | 26 |
| 45 | Romans | 9417 | 42 | 52 | 57 | Philemon | 423 | 8 | 24 |
| 17 | Esther | 5633 | 32 | 51 | 65 | Jude | 608 | 8 | 24 |
| 16 | Nehemiah | 10487 | 38 | 50 | 64 | 3 John | 295 | 6 | 22 |

*Table 2:* Classical works sorted by decreasing weight of highly frequent words

| ID | Author [trans.] | Text | Total | *h* | *hfw* |
|---|---|---|---|---|---|
| N1 | Newton | Principia (an excerpt) | 35982 | 73 | 63 |
| E1 | Einstein [R.W. Lawson] | Relativity | 29368 | 63 | 55 |
| N11 | Newton | Principia, Book III | 7066 | 38 | 55 |
| SC11 | Shakespeare | The merry wives of Windsor | 23779 | 67 | 53 |
| SC15 | Shakespeare | Twelfth night | 21483 | 64 | 52 |
| ST07 | Shakespeare | Othello | 27939 | 69 | 51 |
| SC07 | Shakespeare | Much ado about nothing | 22579 | 62 | 51 |
| D1I | Dante | Divina Commedia 1 Inferno | 22934 | 58 | 51 |
| D1E1 | Dante [H.W. Longfellow] | Divine Comedy 1 Hell | 37031 | 69 | 51 |
| SC06 | Shakespeare | Measure for measure | 23137 | 63 | 51 |
| ST03 | Shakespeare | Hamlet | 32223 | 73 | 51 |
| SC03 | Shakespeare | As you like it | 22832 | 61 | 50 |
| SC12 | Shakespeare | The taming of the shrew | 22155 | 64 | 50 |
| SC02 | Shakespeare | All's well that ends well | 24368 | 64 | 50 |
| ST02 | Shakespeare | Coriolanus | 29278 | 72 | 50 |
| SH10 | Shakespeare | Richard III | 31426 | 71 | 49 |
| SC16 | Shakespeare | Two gentlemen of Verona | 18244 | 56 | 49 |
| ST04 | Shakespeare | Julius Caesar | 20843 | 60 | 49 |
| SC14 | Shakespeare | Troilus and Cressida | 27614 | 73 | 49 |
| SC17 | Shakespeare | Winter's tale | 25996 | 68 | 49 |
| SC09 | Shakespeare | The comedy of errors | 16181 | 50 | 49 |
| SH01 | Shakespeare | Henry IV part 1 | 26152 | 65 | 48 |
| SC04 | Shakespeare | Cymbeline | 28985 | 72 | 48 |
| SC10 | Shakespeare | The merchand of Venice | 22210 | 61 | 48 |
| SH02 | Shakespeare | Henry IV part 2 | 27980 | 68 | 48 |
| ST05 | Shakespeare | King Lear | 27803 | 70 | 48 |
| ST01 | Shakespeare | Antony and Cleopatra | 26963 | 68 | 48 |
| SH06 | Shakespeare | Henry VI part 3 | 25896 | 65 | 48 |
| ST08 | Shakespeare | Romeo and Juliet | 25917 | 67 | 47 |
| ST10 | Shakespeare | Titus Andronicus | 21723 | 64 | 47 |
| D3E1 | Dante [H.F. Cary] | Divine Comedy 3 Paradise | 35345 | 70 | 47 |
| SH07 | Shakespeare | Henry VIII | 25973 | 65 | 47 |
| SH05 | Shakespeare | Henry VI part 2 | 26806 | 66 | 47 |

*(continued on next page)*

*Table 2 (continued from previous page)*

| ID | Author [trans.] | Text | Total | *h* | *hfw* |
|---|---|---|---|---|---|
| SC05 | Shakespeare | Love's labour's lost | 23048 | 63 | 47 |
| G1E2 | Goethe [G.M. Priest] | Faust 1 | 32455 | 68 | 47 |
| D1E2 | Dante [H.F. Cary] | Divine Comedy 1 Hell | 36476 | 69 | 46 |
| SH03 | Shakespeare | Henry V | 27557 | 64 | 46 |
| D2E1 | Dante [H.F. Cary] | Divine Comedy 2 Purgatory | 36560 | 70 | 46 |
| SH09 | Shakespeare | Richard II | 23894 | 60 | 46 |
| G1E1 | Goethe [A.S. Kline] | Faust 1 | 32874 | 68 | 46 |
| SC01 | Shakespeare | A midsummer night's dream | 17167 | 57 | 45 |
| G2E1 | Goethe [A.S. Kline] | Faust 2 | 53841 | 78 | 45 |
| SH08 | Shakespeare | King John | 21775 | 57 | 45 |
| SH04 | Shakespeare | Henry VI part 1 | 22846 | 62 | 45 |
| SC08 | Shakespeare | Pericles, prince of Tyre | 19560 | 59 | 45 |
| SC13 | Shakespeare | The tempest | 17453 | 57 | 44 |
| ST09 | Shakespeare | Timon of Athens | 19623 | 55 | 44 |
| G1G | Goethe | Faust 1 | 30625 | 64 | 43 |
| ST06 | Shakespeare | Macbeth | 18213 | 53 | 42 |
| G2G | Goethe | Faust 2 | 44452 | 74 | 41 |
| D2I | Dante | Divina Commedia 2 Purgatorio | 15400 | 42 | 40 |
| D3I | Dante | Divina Commedia 3 Paradiso | 9577 | 36 | 40 |

*Table 3:* Nobel lectures sorted by decreasing weight of highly frequent words

| Year and Field | Author | Total | *h* | *hfw* |
|---|---|---|---|---|
| 1965 Phys | Richard P. Feynman | 11265 | 41 | 47 |
| 1908 Chem | Ernest Rutherford | 5082 | 26 | 46 |
| 1938 Lit | Pearl Buck | 9090 | 39 | 45 |
| 2004 Lit | Elfriede Jelinek | 5746 | 33 | 45 |
| 1979 Peace | Mother Teresa | 3822 | 26 | 44 |
| 1902 Phys | Hendrik A. Lorentz | 7301 | 31 | 44 |
| 1911 Chem | Marie Curie | 4319 | 25 | 43 |
| 1925 Med | Frederick G. Banting | 8193 | 32 | 41 |
| 1925 Med | John Macleod | 4862 | 24 | 41 |
| 1963 Peace | Linus Pauling | 6246 | 28 | 41 |
| 1984 Lit | Jaroslav Seifert | 5243 | 26 | 41 |
| 1920 Phys | Max Planck | 5203 | 24 | 40 |
| 1970 Lit | Alexandr Solzhenitsyn | 6516 | 32 | 40 |
| 1902 Phys | Pieter Zeeman | 3480 | 21 | 39 |
| 1950 Lit | Bertrand Russell | 5703 | 29 | 39 |
| 1973 Lit | Heinrich Böll | 6094 | 28 | 39 |
| 1983 Peace | Lech Walesa | 2587 | 19 | 39 |
| 1989 Peace | Dalai Lama | 3601 | 23 | 39 |
| 1991 Peace | Mikhail Gorbachev | 5693 | 26 | 39 |
| 1905 Med | Robert Koch | 4283 | 24 | 38 |
| 1975 Econ | Leonid V. Kantorovich | 3924 | 22 | 38 |
| 1989 Econ | Trygve Haavelmo | 3186 | 21 | 38 |
| 1930 Lit | Sinclair Lewis | 5007 | 25 | 37 |
| 1953 Peace | George C. Marshall | 3249 | 19 | 37 |
| 1959 Lit | Salvatore Quasimodo | 3698 | 21 | 37 |
| 1976 Lit | Saul Bellow | 4775 | 26 | 37 |
| 1986 Econ | James M. Buchanan Jr. | 4623 | 23 | 37 |
| 1975 Med | Renato Dulbecco | 3675 | 22 | 36 |
| 1993 Lit | Toni Morrison | 2972 | 22 | 36 |
| 1935 Chem | Irène Joliot-Curie | 1105 | 12 | 32 |
| 1986 Peace | Elie Wiesel | 2693 | 19 | 32 |
| 2002 Peace | Jimmy Carter | 2330 | 16 | 30 |
| 1996 Lit | Wislawa Szymborska | 1983 | 16 | 27 |

*Table 4:* Illustrating *hfw* synergism of various parts of Dostoevsky's *Crime and Punishment*

| Text | Total | *h* | *hfw* |
|---|---|---|---|
| Part 1 | 35365 | 70 | 52 |
| Part 2 | 38653 | 76 | 52 |
| Part 3 | 29924 | 71 | 51 |
| Part 4 | 28342 | 67 | 51 |
| Part 5 | 28226 | 66 | 51 |
| Part 6 | 35900 | 74 | 53 |
| Epilogue | 6336 | 30 | 44 |
| Parts 1+2 | 74066 | 103 | 57 |
| Parts 1+2+3 | 104028 | 123 | 59 |
| Parts 1+2+3+4 | 132370 | 137 | 61 |
| Parts 1+2+3+4+5 | 160617 | 154 | 62 |
| Parts 1+2+3+4+5+6 | 196519 | 171 | 64 |
| All Parts + Epilogue | 202853 | 174 | 64 |

## References

Hirsch, Jorge E.
2005 "An index to quantify an individual's scientific research output". In: *arXiv:physics/0508025 v4 23 Aug 2005*. `http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf`

Popescu, Ioan-Iovitz
2001 "Cited Papers Ranked by Descending Citation Frequency". `http://www.geocities.com/iipopescu/CITSH.htm`

***Main electronic text sources and tools used in this paper***

*The Bible* (English King James Version).
`http://www.fourmilab.ch/etexts/www/Bible/`

Shakespeare, *The Complete Works*.
`http://www-tech.mit.edu/Shakespeare/`

Dante, *Divina Commedia* and Goethe's *Faust*.
`http://jollyroger.com/library/`

Newton, *The Principia*.
`http://members.tripod.com/~gravitee/`

Einstein, *Relativity: The Special and General Theory*.
`http://www.bartleby.com/173/`

*The Nobel Lectures*.
http://nobelprize.org/nobel/

Dostoevsky, *Crime and Punishment*. http://www.bartleby.com/318/

***Word Frequency Counters:***

http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html

http://www.writewords.org.uk/word_count.asp