# Victor Davis

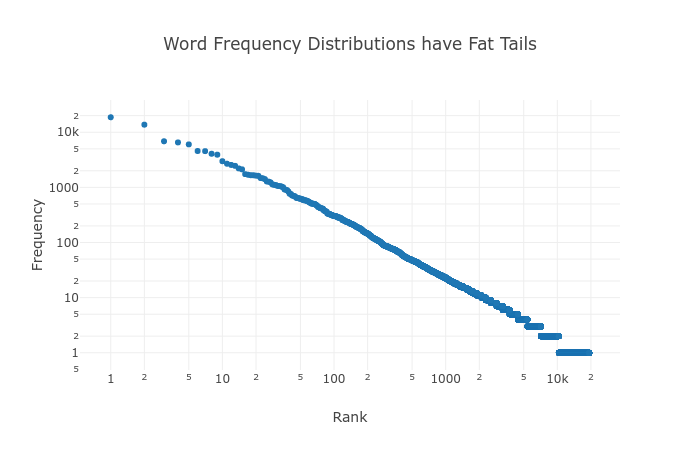
## Types, Tokens, and Hapaxes: A New Heap's Law

### Abstract

Heap's Law[[1]](#footnote-2) states that in a large enough text corpus, the number of types as a function of tokens grows as for some free parameters . Much has been written[[2]](#footnote-3)[[3]](#footnote-4)[[4]](#footnote-5)[[5]](#footnote-6)[[6]](#footnote-7) about how this result and various generalizations can be derived from Zipf's Law[[7]](#footnote-8). Here we derive from first principles a completely novel expression of the type-token curve and prove its superior accuracy on real text. This expression naturally generalizes to equally accurate estimates for counting hapaxes and higher -legomena.

### Introduction

Zipf's Law is usually formulated[[8]](#footnote-9) as Freq 1/Rank, generalized by raising Rank to an exponent . This is benchmarked against real text and shown to be curiously accurate for corpora of varying sizes but exhibiting a fat tail. Therefore it performs most poorly on rare words, which contribute the most to the type-token curve. The common interpretation is that any inaccuracy in modeling a type-token growth curve can be explained away by this fat tail in which the unpredictability of the frequency of rare words introduces noise into an otherwise correct signal. This paradigm is completely false. We show that Zipf's Law can be reformulated to model the frequencies of rare words as accurately as common ones and that this reformulation leads to a simple, intuitive derivation of a logarithmic (*not* exponential) type-token growth curve.



### Keywords

**Tokens**: Instances of words in a text, elements of the corpus in which words are considered an ordered list. (Using "words" and "tokens" interchangeably.)

**Types**: Distinct words of a text, elements of the lexicon/vocabulary or the size of the *set* of tokens. (Using "distinct words" and "types" interchangeably.)

**Hapaxes**: Types that occur exactly once in a corpus.

**Dis, Tris, Tetrakis, Pentakis, -Legomena**: Types that occur exactly times in a corpus.

**-corpus**: A text corpus consisting of tokens and types.

**Optimum Sample**: A text corpus exhibiting a perfect Zipf distribution.

**Perfect Zipf Distribution**: A word frequency distribution following Zipf's law literally and perfectly, reformulated to directly estimate the number of rare words. (Defined in more detail later.)

### Hypothesis

For a text corpus consisting of tokens and types, randomly sampling tokens of that corpus yields types, and the expected value of is given by (1), parametrized by , the size of an optimum sample.

(1)

The procedure for obtaining the parameters is described in the "Results" section later, equations (19) and (20), respectively.

### Methods

For some small , construct a -corpus by collecting every word that occurs *exactly* times. This mini-corpus, a sub-selection of the whole, has a perfectly uniform word frequency distribution, a text consisting of total words (tokens): distinct words (types) each repeating exactly times. What is , the expected number of types in a random selection of words out of this mini-corpus? If sampling sequentially without replacement, the expected value is the partial sum of the probability at each step of drawing a new type.

(2)

Suppose tokens are drawn at random without replacement, resulting in types. There are then tokens left to draw, of which are of a type not yet drawn. Thus,

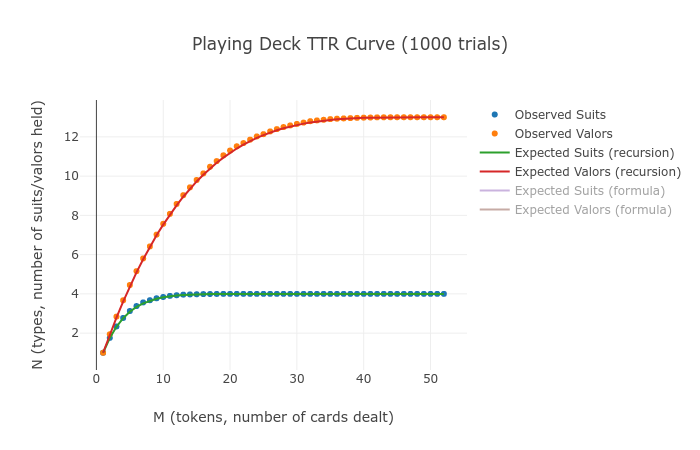
(3)

(3a)

Substituting the *actual* number of types drawn, , for the *expected* number of types drawn, , we have a recursive expression for the growth of the curve starting at .

(4)

This recursion performs quite well with real data. Consider a shuffled deck of cards, analogous to either a - or -corpus, depending whether suits or valors are considered "types."



We can use calculus to derive a non-recursive version of this function. Since the derivative is approximately equal to the vertical distance between successive elements in the series, we can rewrite (4) and integrate.

(5)

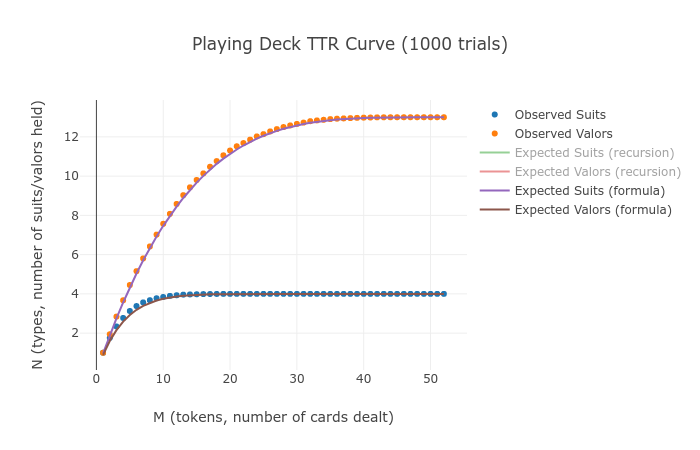
Using as a boundary condition, we find , giving:

(7)

(7a)

(7b)

Again, we find a near-perfect match between the discrete values calculated by recursing and the continuous values given by (7). *Near* perfect because the discrete case (actual reality) effectively stopped the integral limit approaching zero once . Thus, (7) is an analytical estimate of (4) smoothing out the discontinuities found in a real deck of cards.



### Text Corpora

A -corpus can be partitioned into a number of -mini-corpuses, and the expected type-token curve of the whole can be gotten by summing over the individual parts. Start by partitioning the corpus into decks: all hapaxes into the first deck, all dis legomena into the second deck, tris into the third deck, etc. The th deck shall be a -mini-corpus, the whole corpus encoded by some vector . (Call the "zeroth deck" the null set for now, for number of types occurring exactly zero times in the whole corpus, which will come in handy shortly.)

The whole corpus consists of all these decks randomly shuffled together. Sampling tokens should yield on average tokens from the th deck. The number of types yielded from each deck is therefore . Substitute this into (7) and sum over all decks.

As for the bounds on the summation, we can consider a finite vector in infinite-dimensional space with zeroes imputed into missing indexes (numbers for which no words occur *exactly* times). This expression (8), as we'll see later in the "Results" section, performs nearly perfectly on real data. We can also rewrite it as an infinite series over the interval by substituting the proportion of the corpus sampled rather than the raw token count.

(8)

(8b)

What does (8b) really *mean*? What happens when we sample tokens at random, or some proportion of the corpus? Choosing a sample size of gives a permutation of the original corpus, and it's easy to see from (8b) that as expected, i.e. all types are drawn. Sampling some proportion of the corpus, we expect types to be drawn, and types *not* to be drawn. What is the expected value of the latter? For the hapaxes, the probability of *not* drawing each token is , so the expected number of hapaxes *not* drawn is . The probability of *not* drawing *both* instances of each dis legomenon is so the expected number of dis legomena *not* drawn is . Continuing this argument, we define , the expected number of types *not* drawn when sampling proportion of the corpus:

(9)

It can be seen that (9) is just a way of rewriting (8b) given this new intuition. But let's go further. How many *hapaxes* can be expected when sampling proportion of the corpus? The expected number of existing hapaxes drawn is . But a hapax will be *created* in the sample if *one* instance of a dis legomenon is drawn but not the other. We should expect to be created this way. Likewise, we expect to create a hapax in our sample by drawing only *one* of the three instances of each tris legomenon, resulting in created this way. In general, using the shorthand ,

(10)

(11)

There is a beautiful way of visualizing this as a matrix transformation acting on involving Pascal's Triangle. Since each of the columns sum to , it's easy to prove that both the input vector and the resultant vector sum to . The equation (11) above expresses the row sums.

(12)

(12a)

(12b)

(If the idea of infinite matrices doesn't sit well, remember is always finite in practice.) Notice (11) can be expressed in terms of the th derivatives of for :

(14)

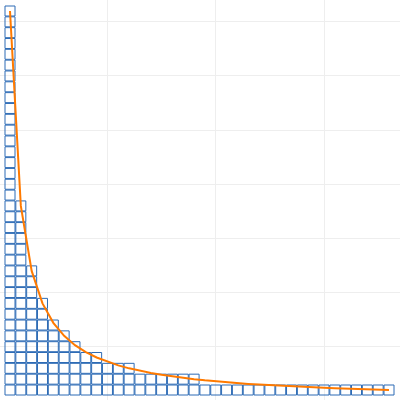
By now we've achieved three results, having not yet made a single empirical assumption about the word frequency distribution of a corpus. Any type-token system describable in terms of the evolution of some follows these rules.

1. Equation (12) gives an algorithmic approach to predicting -legomena counts when sampling a corpus. Using the observed counts from the *whole* corpus as input, the word frequency distribution of any sample can be simulated by constructing and transforming .
2. Any function such as Heap's Law which approximates the type-token growth curve can be generalized via (14) to make testable predictions for -legomena counts as well.
3. If can be approximated by a series and converges on some analytic function, then (14) gives analytic functions for -legomena as well.

Can be approximated by a series? In real text, do the tabulations of -legomena actually follow some regular, enumerative pattern?

### Zipf's Law

Suppose that for a given -corpus there exists a corresponding optimum sample -corpus following a *perfect Zipf distribution*. If the original corpus is "too big" then some number of randomly selected tokens will produce a sub-selection of text exhibiting Zipf's Law. If the original corpus is "too small" then it can be thought of as a sub-selection of some larger corpus consisting of tokens exhibiting Zipf's Law. (Of course, some corpora may be "just right" and .)



Define a *perfect Zipf distribution* as follows: For some -corpus with a ranked word frequency distribution , the expected number of types repeating or more times is roughly equal to . Alternatively, there exists about types repeating or more times. "Roughly" and "about" to allow for normal statistical noise in the still "perfect" sample.

**Corollaries:**

* the expected number of types repeating between and times is
* The expected number of types repeating fewer than times is
* No types occur exactly zero times:
* the expected number of types repeating *exactly* times, or the expected number of -legomena (for ) equals
* The expected *proportions* of -legomena are
* This distribution is already normalized:
* The expected frequency of the most common word is
* The expected frequency of the th most common word is
* Freq 1/Rank

This reformulation of Zipf's Law is practically equivalent to the original, with the caveats that (a) for high ranks (rare words) we round *down* to the nearest integer to obtain the expected frequency, and (b) the total number of types is equal to the frequency of the commonest word. (So, curiously, there is one "the" for each vocabulary word in the lexicon.) This leaves us with a specific prediction for the counts of rare words (-legomena for low ). The expected number of hapaxes, dis, tris, and higher -legomena, for is given by:

(15)

Substituting (15) into (8b) above, and taking ,

It can be shown that (16) converges on (17). Taking successive derivatives, per (14) we now have expressions not just for the number of types with respect to tokens, but also hapaxes and higher -legomena:

(17)

(17.0)

(17.1)

(17.2)

(17.3)

(17.4)

(17.5)

### Results

The King James Bible is a -corpus. The most common word "the" appears times (not ), and % of the types are hapaxes (not %). It is nowhere near an optimum sample, so how do we find a suitable ? The proportion of hapaxes is a decreasing function, starting at , falling to , and continuing to fall as goes to infinity. We can model this by expressing (17.1) as a proportion of (17) to obtain:

(18)

In the limit, as expected, even though the function is undefined at those points. Using a binary search algorithm, we find that . Could this mean the Bible is ten times "too large" to exhibit Zipf's Law? Taking and scaling such that , we fit (1) to our sample corpus using optimum sample parameters . Calculating , we have:

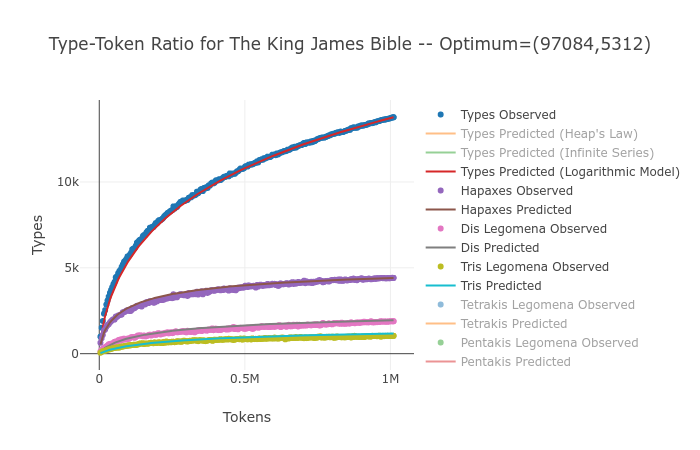
(19)

(20)

(1.1)

#### Observed vs Predicted: King James Bible

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| tokens | types | hapax | types pred | hapax pred |
| 4,042 | 993 | 632 | 734 | 535 |
| 105,092 | 5,766 | 2,482 | 5,525 | 2,726 |
| 206,142 | 7,781 | 3,124 | 7,561 | 3,310 |
| 307,192 | 9,075 | 3,531 | 8,946 | 3,633 |
| 408,242 | 10,087 | 3,750 | 10,010 | 3,846 |
| 509,292 | 10,898 | 3,881 | 10,878 | 4,001 |
| 610,342 | 11,664 | 4,050 | 11,613 | 4,120 |
| 711,392 | 12,303 | 4,213 | 12,252 | 4,215 |
| 812,442 | 12,855 | 4,283 | 12,817 | 4,294 |
| 913,492 | 13,314 | 4,353 | 13,324 | 4,359 |
| 1,010,654 | 13,769 | 4,414 | 13,767 | 4,413 |

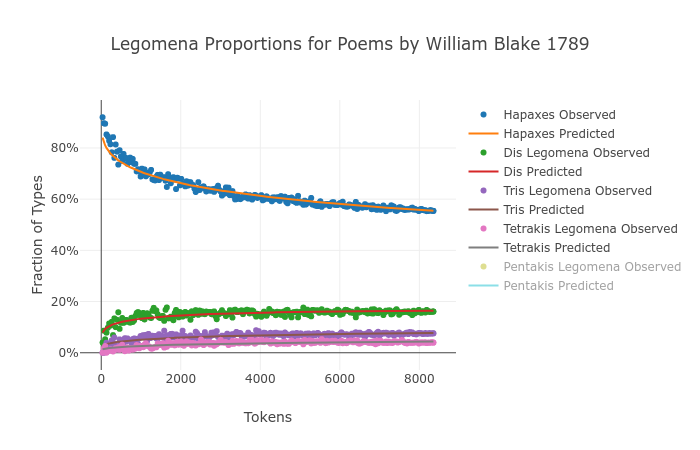


William Blake's *Poems* is a -corpus. The most common word "the" appears times (not ) and % of its types are hapaxes. Using a binary search, we find that , making *Poems* "too small" by half. Using the same procedure as above, we find .

(1.2)

#### Observed vs Predicted: William Blake's Poems

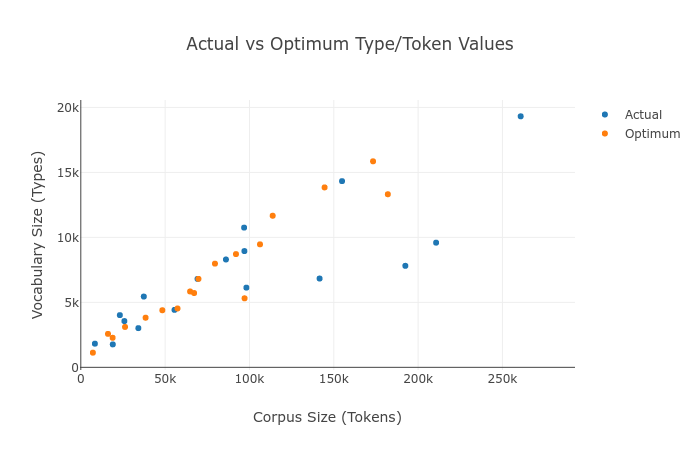
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| tokens | types | hapax | types pred | hapax pred |
| 33 | 25 | 23 | 33 | 27 |
| 858 | 439 | 324 | 424 | 304 |
| 1,683 | 704 | 491 | 678 | 457 |
| 2,508 | 909 | 587 | 882 | 571 |
| 3,333 | 1,066 | 653 | 1,057 | 662 |
| 4,158 | 1,211 | 740 | 1,212 | 739 |
| 4,983 | 1,366 | 809 | 1,352 | 805 |
| 5,808 | 1,495 | 878 | 1,480 | 864 |
| 6,633 | 1,604 | 931 | 1,598 | 916 |
| 7,458 | 1,707 | 953 | 1,708 | 963 |
| 8,283 | 1,812 | 1,010 | 1,811 | 1,005 |

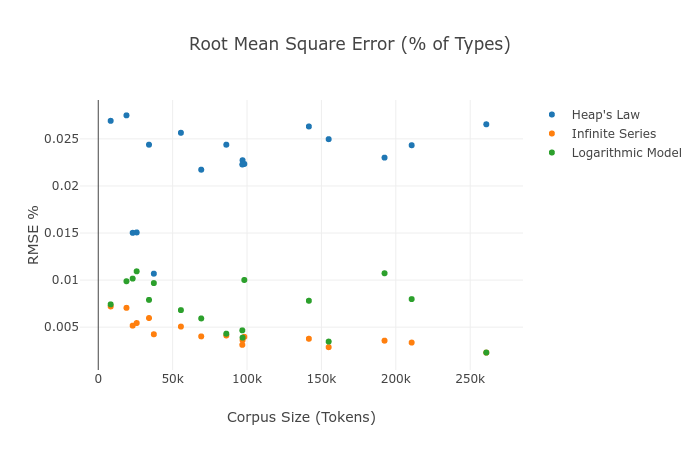


Repeating the procedure, we obtain values for that estimate types, hapaxes, and higher -legomena for a number of different books. We'll evaluate the model fit using root mean square error as a percent of observed types. Below are the fitting errors for real data. The fit for (8) using actual -legomena counts as the coefficients gives a good baseline for measuring the model, since it doesn't assume Zipf's Law. Making the leap from (8) to (17) requires the assumption of the existence of an optimum sample following a perfect Zipf distribution, an assumption that appears to be more accurate in some cases, like Melville's *Moby Dick*, and less accurate in others, like Shakespeare's *Julius Caesar*. Still, Heap's Law is consistently accurate only to within %, while the logarithmic function derived above is nearly always accurate to within %.

#### RMSE % Comparison for Various Books

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| title |  |  | Heaps | Series | Model |
| Moby Dick by Herman Melville 1851 | 173,272 | 15,854 | 2.66% | 0.23% | 0.23% |
| Leaves of Grass by Walt Whitman 1855 | 144,542 | 13,845 | 2.50% | 0.29% | 0.35% |
| The Ball and The Cross by G . K . Chesterton 1909 | 91,990 | 8,714 | 2.27% | 0.36% | 0.39% |
| The Wisdom of Father Brown by G . K . Chesterton 1914 | 79,540 | 7,980 | 2.44% | 0.41% | 0.43% |
| Paradise Lost by John Milton 1667 | 113,779 | 11,666 | 2.23% | 0.31% | 0.47% |
| The Man Who Was Thursday by G . K . Chesterton 1908 | 69,764 | 6,807 | 2.17% | 0.40% | 0.59% |
| Stories to Tell to Children by Sara Cone Bryant 1918 | 26,223 | 3,108 | 2.56% | 0.51% | 0.68% |
| Poems by William Blake 1789 | 16,121 | 2,574 | 2.69% | 0.72% | 0.74% |
| Sense and Sensibility by Jane Austen 1811 | 38,414 | 3,817 | 2.63% | 0.38% | 0.78% |
| Alice ' s Adventures in Wonderland by Lewis Carroll 1865 | 18,864 | 2,275 | 2.44% | 0.60% | 0.79% |
| The Parent ' s Assistant , by Maria Edgeworth | 67,166 | 5,716 | 2.43% | 0.34% | 0.80% |
| The Tragedie of Hamlet by William Shakespeare 1599 | 182,050 | 13,320 | 1.07% | 0.42% | 0.97% |
| The Adventures of Buster Bear by Thornton W . Burgess 1920 | 7,158 | 1,127 | 2.75% | 0.71% | 0.99% |
| Persuasion by Jane Austen 1818 | 48,342 | 4,393 | 2.23% | 0.40% | 1.00% |
| The Tragedie of Macbeth by William Shakespeare 1603 | 106,245 | 9,465 | 1.50% | 0.52% | 1.02% |
| Emma by Jane Austen 1816 | 57,337 | 4,528 | 2.30% | 0.36% | 1.07% |
| The Tragedie of Julius Caesar by William Shakespeare 1599 | 64,786 | 5,838 | 1.51% | 0.54% | 1.09% |
| The King James Bible | 97,084 | 5,312 | 2.25% | 0.28% | 1.13% |





### Discussion

It's worth reinforcing that Zipf's Law is only accurate when applied to a certain optimum sample size of a given corpus. Smaller samples *won't* follow Zipf's Law, but as text accumulates, the sample will reach a phase at which it exhibits Zipf's Law peculiarly well. As text continues to accumulate, the pattern deteriorates in the opposite direction. The logarithmic functions (17.x) accurately model this behavior even if they can't explain the motivation for it. But it has to be stressed that Zipf's Law is *not* a pattern accurately describing all texts; it is better thought of as describing a phase text passes through as it accumulates.

This phase is the point at which (a) there are approximately as many instances of the most common word (usually "the") as there are distinct words in the vocabulary, and (b) the proportions of hapaxes, dis, tris, etc legomena with respect to the number of types is roughly . It should not be expected that (a) and (b) *always* hold, and that some texts follow this more accurately than others. Our hypothesis is that it seems to be the case that *all* texts follow (a) and (b) peculiarly well at some optimum sample size, and that by discovering this optimum, accurate predictions can be made for how these proportions evolve as the sample size is varied.

### Conclusion

Most investigations of Zipf's and Heap's respective Laws and their inter-relationship tolerate a conspicuous degree of error when applied to real text. This may reflect an implicit bias that these "laws" are mere curiosities of nature, that no model *should* be able to fit cogent, rational human language to a tidy physics equation. Shockingly, the discovery of the formulas above and their accurate fit to real data imply that, far from being mere curiosities, Zipf's Law and its corresponding derivations for estimating types, hapaxes, and higher -legomena must be more fundamental to the nature of language than we give it credit. It's still an open question *why* Zipf's Law seems to hold. Taking it in a very literal formulation as an axiom, accurate predictions can be made about real text, bolstering the idea that *something* fundamental must be going on, and Zipf's original observation must be the inevitable cumulative effect of some deeper, underlying information-theoretic property of human language.

### Correspondence Address

Victor Davis [victor.davis.us@member.mensa.org](mailto:victor.davis.us@member.mensa.org)

### References

| [[9]](#footnote-10)|[[10]](#footnote-11)|[[11]](#footnote-12)|[[12]](#footnote-13)|[[13]](#footnote-14)|[[14]](#footnote-15)|[[15]](#footnote-16)|[[16]](#footnote-17)|[[17]](#footnote-18)|[[18]](#footnote-19)|[[19]](#footnote-20)|[[20]](#footnote-21)|[[21]](#footnote-22)|[[22]](#footnote-23)|

1. https://dl.acm.org/citation.cfm?id=539986 Heaps, H S 1978 *Information Retrieval: Computational and Theoretical Aspects* (Academic Press) [↑](#footnote-ref-2)
2. http://iopscience.iop.org/article/10.1088/1367-2630/15/9/093033 Font-Clos, Francesc 2013 *A scaling law beyond Zipf's law and its relation to Heaps' law* (New Journal of Physics 15 093033) [↑](#footnote-ref-3)
3. http://iopscience.iop.org/article/10.1088/1367-2630/11/12/123015 Bernhardsson S, da Rocha L E C and Minnhagen P 2009 *The meta book and size-dependent properties of written language* (New Journal of Physics 11 123015) [↑](#footnote-ref-4)
4. http://iopscience.iop.org/article/10.1088/1742-5468/2011/07/P07013 Bernhardsson S, Ki Baek and Minnhagen 2011 *A paradoxical property of the monkey book* (Journal of Statistical Mechanics: Theory and Experiment, Volume 2011) [↑](#footnote-ref-5)
5. http://milicka.cz/kestazeni/type-token\_relation.pdf Milička, Jiřı́ 2009 *Type-token & Hapax-token Relation: A Combinatorial Model* (Glottotheory. International Journal of Theoretical Linguistics 2 (1), 99–110) [↑](#footnote-ref-6)
6. https://www.nature.com/articles/srep00943 Petersen, Alexander 2012 *Languages cool as they expand: Allometric scaling and the decreasing need for new words* (Scientific Reports volume 2, Article number: 943) [↑](#footnote-ref-7)
7. http://dx.doi.org/10.1037/h0052442 Zipf, George 1949 *Human behavior and the principle of least effort* (Reading: Addison-Wesley) [↑](#footnote-ref-8)
8. http://dx.doi.org/10.1037/h0052442 Zipf, George 1949 *Human behavior and the principle of least effort* (Reading: Addison-Wesley) [↑](#footnote-ref-9)
9. https://dl.acm.org/citation.cfm?id=539986 Heaps, H S 1978 *Information Retrieval: Computational and Theoretical Aspects* (Academic Press) [↑](#footnote-ref-10)
10. http://iopscience.iop.org/article/10.1088/1367-2630/15/9/093033 Font-Clos, Francesc 2013 *A scaling law beyond Zipf's law and its relation to Heaps' law* (New Journal of Physics 15 093033) [↑](#footnote-ref-11)
11. http://iopscience.iop.org/article/10.1088/1367-2630/11/12/123015 Bernhardsson S, da Rocha L E C and Minnhagen P 2009 *The meta book and size-dependent properties of written language* (New Journal of Physics 11 123015) [↑](#footnote-ref-12)
12. http://iopscience.iop.org/article/10.1088/1742-5468/2011/07/P07013 Bernhardsson S, Ki Baek and Minnhagen 2011 *A paradoxical property of the monkey book* (Journal of Statistical Mechanics: Theory and Experiment, Volume 2011) [↑](#footnote-ref-13)
13. http://milicka.cz/kestazeni/type-token\_relation.pdf Milička, Jiřı́ 2009 *Type-token & Hapax-token Relation: A Combinatorial Model* (Glottotheory. International Journal of Theoretical Linguistics 2 (1), 99–110) [↑](#footnote-ref-14)
14. https://www.nature.com/articles/srep00943 Petersen, Alexander 2012 *Languages cool as they expand: Allometric scaling and the decreasing need for new words* (Scientific Reports volume 2, Article number: 943) [↑](#footnote-ref-15)
15. http://dx.doi.org/10.1037/h0052442 Zipf, George 1949 *Human behavior and the principle of least effort* (Reading: Addison-Wesley) [↑](#footnote-ref-16)
16. www.hup.harvard.edu/catalog.php?isbn=9780674434929 Zipf, George 1932 *Selective studies and the principle of relative frequency in language* (Cambridge: Harvard University Press) [↑](#footnote-ref-17)
17. https://catalog.hathitrust.org/Record/000359461 Zipf, George 1935 *The psycho-biology of language: An introduction to dynamic philology* (Boston: Mifflin Company) [↑](#footnote-ref-18)
18. https://www.journals.uchicago.edu/doi/abs/10.1086/364570 Herdan, Gustav 1960 *Type-token mathematics* (The Hague: Mouton) [↑](#footnote-ref-19)
19. https://dl.acm.org/citation.cfm?id=325476 Baeza-Yates, Ricardo 1997 *Block Addressing Indices for Approximate Text Retrieval* (Journal of the American Society for Information Science, v.51 n.1, p.69-82, Jan. 2000) [↑](#footnote-ref-20)
20. http://dx.doi.org/10.1155/2012/480196 Chen, Yanguang 2012 Zipf’s law, *Hierarchical Structure, and Shuffling-Cards Model for Urban Development* (Discrete Dynamics in Nature and Society, Volume 2012) [↑](#footnote-ref-21)
21. http://iopscience.iop.org/article/10.1088/1367-2630/13/4/043004 Seung, Ki Baek 2011 *Zipf's law unzipped* (New Journal of Physics 13 043004) [↑](#footnote-ref-22)
22. https://doi.org/10.1016/j.physa.2011.05.003 Eliazar, Iddo 2011 *The growth statistics of Zipfian ensembles: Beyond Heaps’ law* (Physica A, Volume 390, Issue 20) [↑](#footnote-ref-23)