

Zipf's Law Revisited

Omer Tripp and Dror Feitelson
The School of Engineering and Computer Science
Hebrew University
Jerusalem, Israel

Abstract

Zipf's law states that the frequency of occurrence of some event as a function of its rank is a power-law function. Using empirical examples from different domains, we demonstrate that at least in some cases, increasingly significant divergences from Zipf's law are registered as the number of events observed increases. Importantly, one of these cases is word frequency in a corpus of natural language, which is—undoubtedly—the most prominent example of Zipf's law. We analyze our findings mathematically and attempt a natural explanation of the regularities underlying them.

1 Introduction

Zipf's law, named after the Harvard linguistic professor George Kingsley Zipf, states that the frequency of occurrence of some event E , $P(E)$, as a function of its rank, r , is a power-law function:

$$P(E_r) \approx \frac{1}{r^\alpha}$$

The first application of Zipf's law was in the domain of natural language. Since, however, the law has been applied to many other domains, including income distribution and the size distribution of cities.

Interestingly, this law is often reported to be at work with respect to limited collections of events. For example, Kundsen affirms the validity of Zipf's law for cities based on empirical data on the 61 largest Danish cities [2]. Similarly, Hatzigeorgiu et al. investigate the application of Zipf's law to the 1,000 most

common words and lemmas in Greek [5], and Nemeth and Zainko confirm that the coverage of corpora by the most frequent words follows a parallel logarithmic rule for Hungarian, English and German in the 40-85% coverage range [4].¹ In all cases, a larger and more complete collection of events could have been considered.

Fortunately, attempts to cover wider collections of events have been made. Ferrer, Cancho and Sole [1] and Montemurro [3] systematically explored exceptions to Zipf's law in general corpora of English text. They found that the total vocabulary can be divided into two or more power-law registers, differing in their mathematical distribution as well as their content. Their crucial finding was that the registers represent two kinds of vocabulary items: Very frequent items belonging to a basic vocabulary (which comprises 4,000-6,000 words) and less frequent items which are specific words. The decay exponent of less frequent items in the power-law distribution of rank by frequency is more steep than the decay exponent of very frequent items.

In what follows, we present results that echo these findings. Our results are spread over several domains. These are: Downloads from the Sourceforge website, accesses to pages in the 1998 World Cup website and word frequency in corpora of natural language. In all cases, we observe the same phenomenon: The results we obtain increasingly diverge from Zipf's law as the number of events we consider grows larger. Furthermore, in all cases the frequencies associated with events that rank high (numerically) decay more steeply than expected.

2 Empirical Results

The results we present below are presented as Zipf graphs with log axes. Ranks appear on the x-axis. The distance between subsequent ranks is equal the number of members occupying the (numerically) higher rank.

Some of the figures include dotted lines at their bottom. These should allow the reader to better appreciate the degree to which the data diverges from the straight line predicted by Zipf's law. They were added for convenience, however, and do not have any numerical counterpart in our results and analysis.

¹Actually, Nemeth and Zainko's investigation is more complete; we refer to other findings they present, which coincide with our results, later when we analyze our findings.

2.1 Sourceforge Downloads

The complete data set consists of approximately 41,000 active projects and a total of 700,000,000 downloads. Four samples of the SourceForge projects population were taken. Their respective sizes are: 700,000, 3,500,000, 7,000,000 and 70,000,000. Sampling proceeded according to the relative weights of the projects.

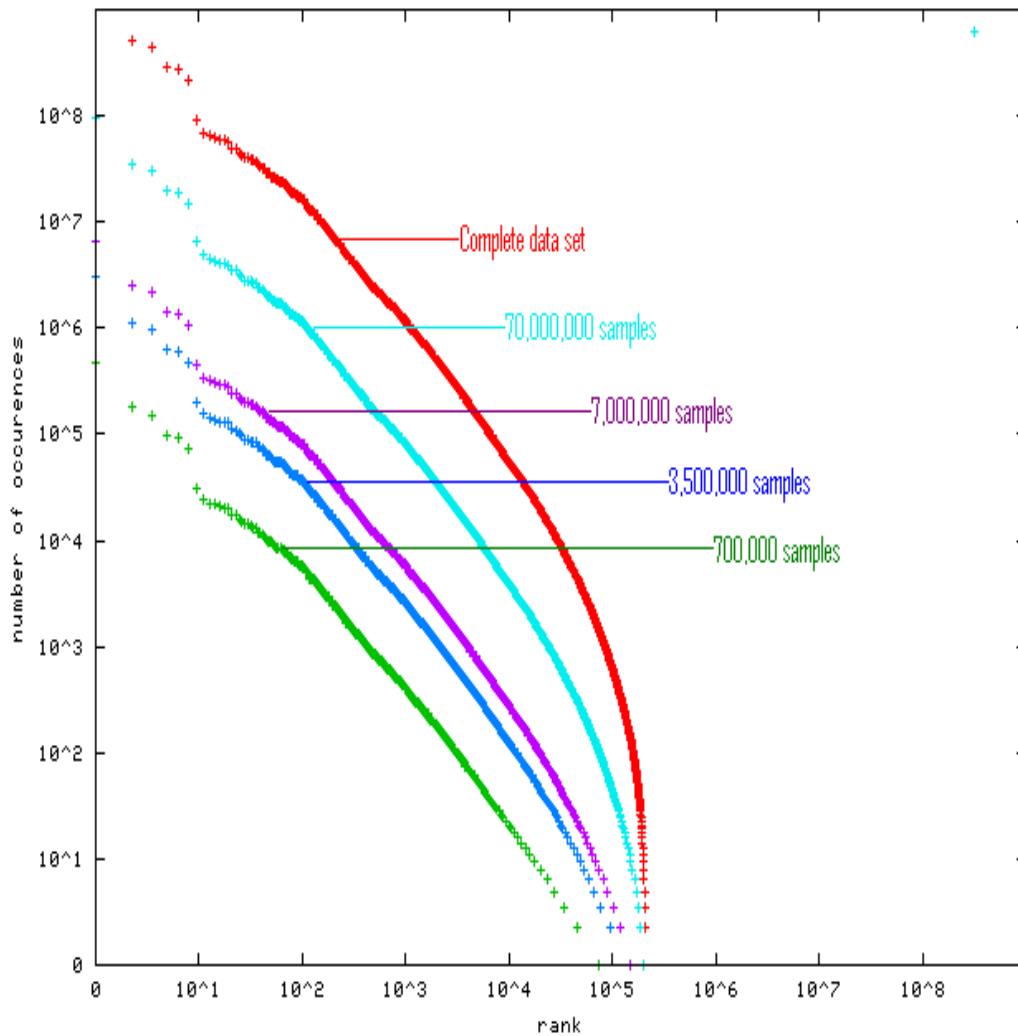


Figure 1: Sourceforge downloads.

As **Figure 1** makes clear, the full data set diverges from Zipf's law. It drops much more steeply than predicted by the law. This trend becomes increasingly

less pronounced as the population under consideration decreases in size. The other data sources presented below display varying instances of the same pattern.

2.2 1998 World Cup Website

The complete data set consists of approximately 21,000 web pages and a total of 1,350,000,000 downloads. The collection of pages in the World Cup website was sampled 6,500,000 times (according to the relative weights of the web pages). The resultant data set consists of approximately 14,500 points. Another source of information is downloads at a specific time interval. On the 30th of June, 1998, about 6,500,000 downloads were recorded between 00:00 and 06:45.

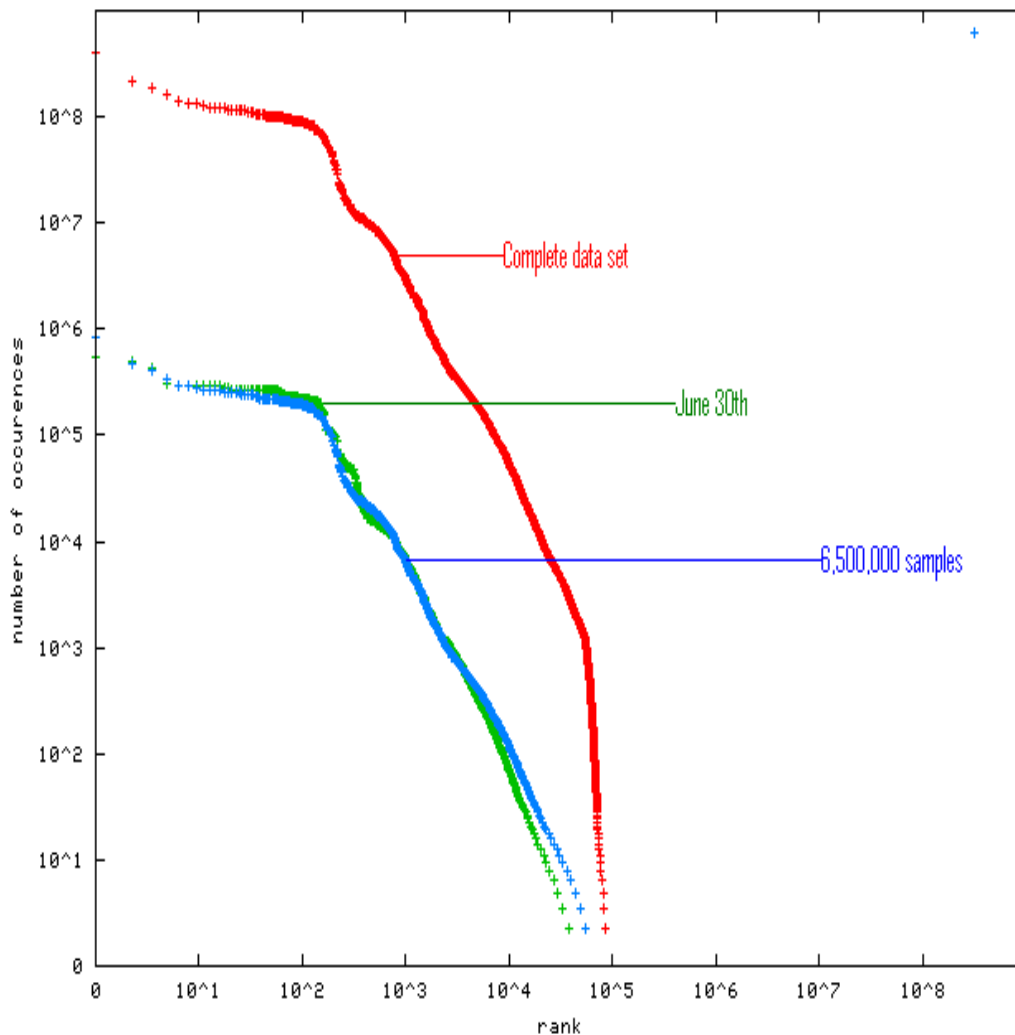


Figure 2: Visits to the World Cup website.

2.3 Word Frequency in Corpora of Natural Language

The results listed below represent the original data minus punctuation marks; numbers were not omitted from the word frequency calculation.

1. Scientific books.

The book *Optical Computing: A Survey for Computer Scientists*, which was written by Prof. Feitelson, contains 6,088 different words, and a total

of 98,989 words. We generated four sets of samples from the original book. Their respective sizes are: 7,500, 15,000, 30,000 and 60,000. Sampling proceeded according to the relative frequencies of the words.

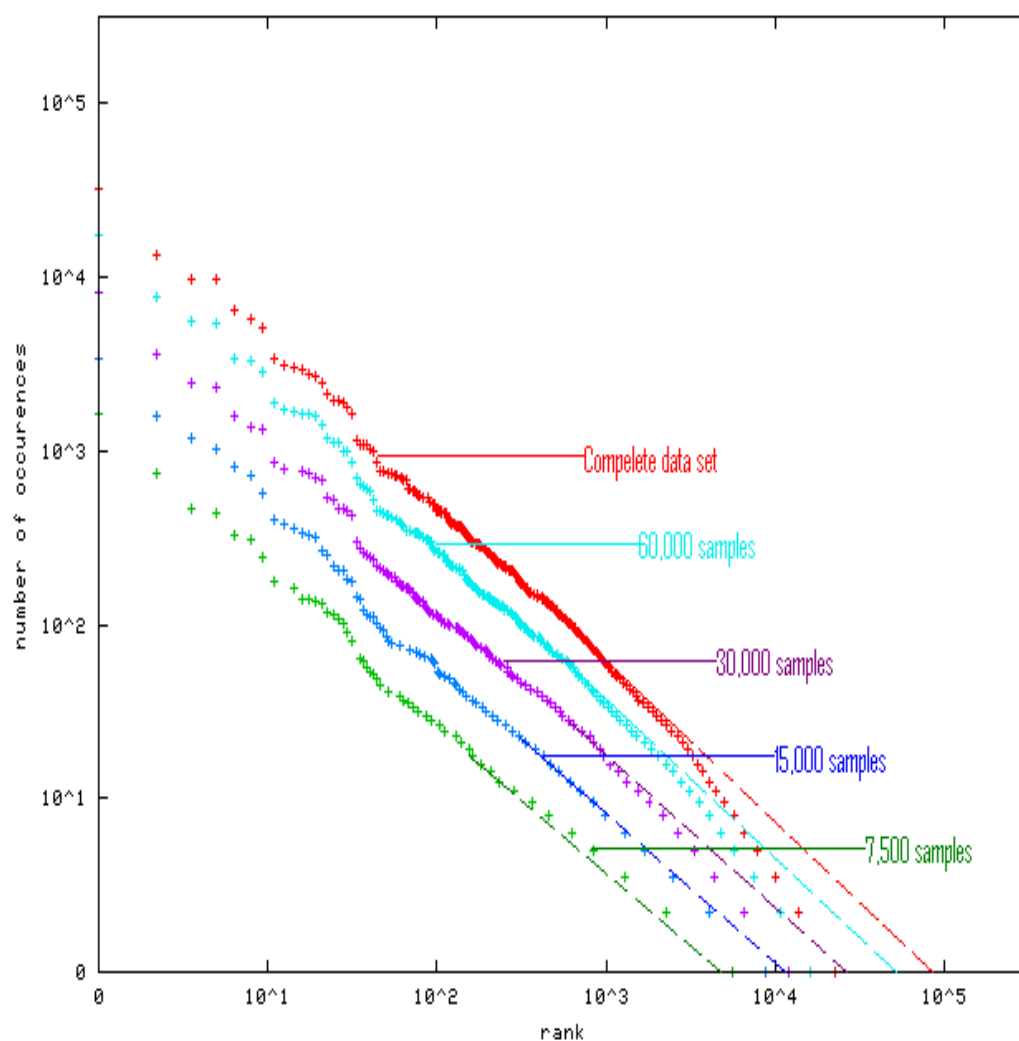


Figure 3: Word frequency in the book *Optical Computing*.

2. The Bible.

The Bible—including both the Old and the New Testament—contains 17,859 different words, and a total of 796,888 words. We generated four sets of samples from the original book. Their respective sizes are: 40,000, 80,000,

160,000 and 320,000. As in the previous case, sampling proceeded according to the relative frequencies of the words.

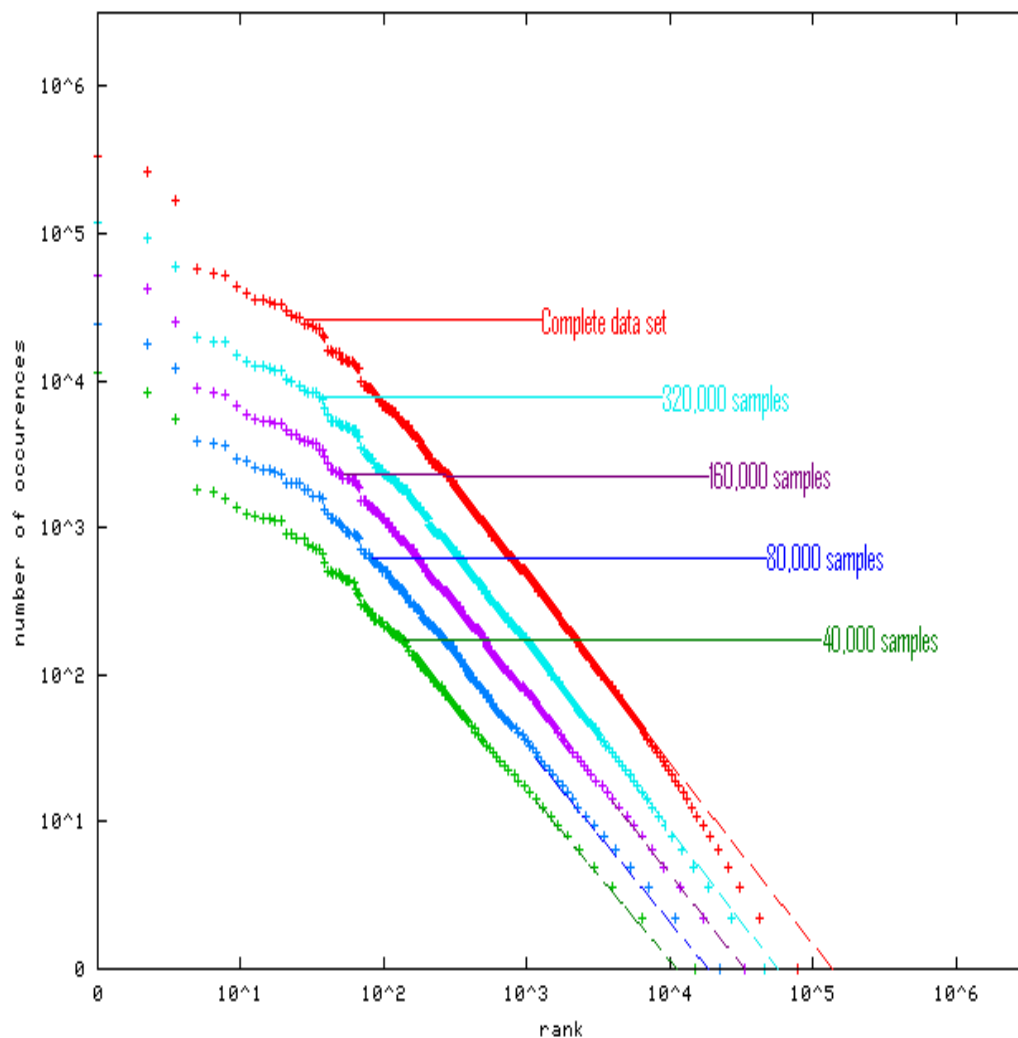


Figure 4: Word frequency in the Bible (including both the Old and the New Testament).

3. Novels.

To obtain a vivid comparison between the validity of Zipf's law in the context of small versus large collections of events, we gathered books from various authors and put them into the same data set. We compared the resultant collection with smaller data sets. The smaller data sets are divided

into two groups: The first contains five data sets, each based on a single piece, whereas the second contains four sub-samples of the large data set.

The books we grouped together are:

- **Mark Twain.** *A Tramp Abroad, Following the Equator, Life on the Mississippi, Roughing It, The Gilded Age, The Prince and the Pauper, Tom Sawyer* and *What is Man?*.
- **Fyodor Dostoyevsky.** *Crime and Punishment, Poor Folk, The Gambler* and *The Idiot*.
- **Leo Tolstoy.** *Anna Karenina, The Cossacks, Childhood* and *War and Peace*.
- **William Shakespeare.** The complete works of William Shakespeare.

The five data sets we examined separately are based on the books *Tom Sawyer, Ulysses, The Idiot, Anna Karenina* and *Hamlet*. **Table 1** lists the number of words as well as the number of different words in each of the data sets.

Data set name	Total number of words	Number of different words
All novels grouped together	3,516,165	155,530
Anna Karenina	349,665	27,206
Hamlet	31,656	6,780
The Idiot	243,830	23,073
Tom Sawyer	70,045	12,339
Ulysses	265,154	43,836

Table 1: Total number of words and number of different words in each of the novel-based data sets.

The four samples taken from the large data set are of sizes: 350,000, 175,000, 70,000 and 35,000.

The results we obtained are presented below:

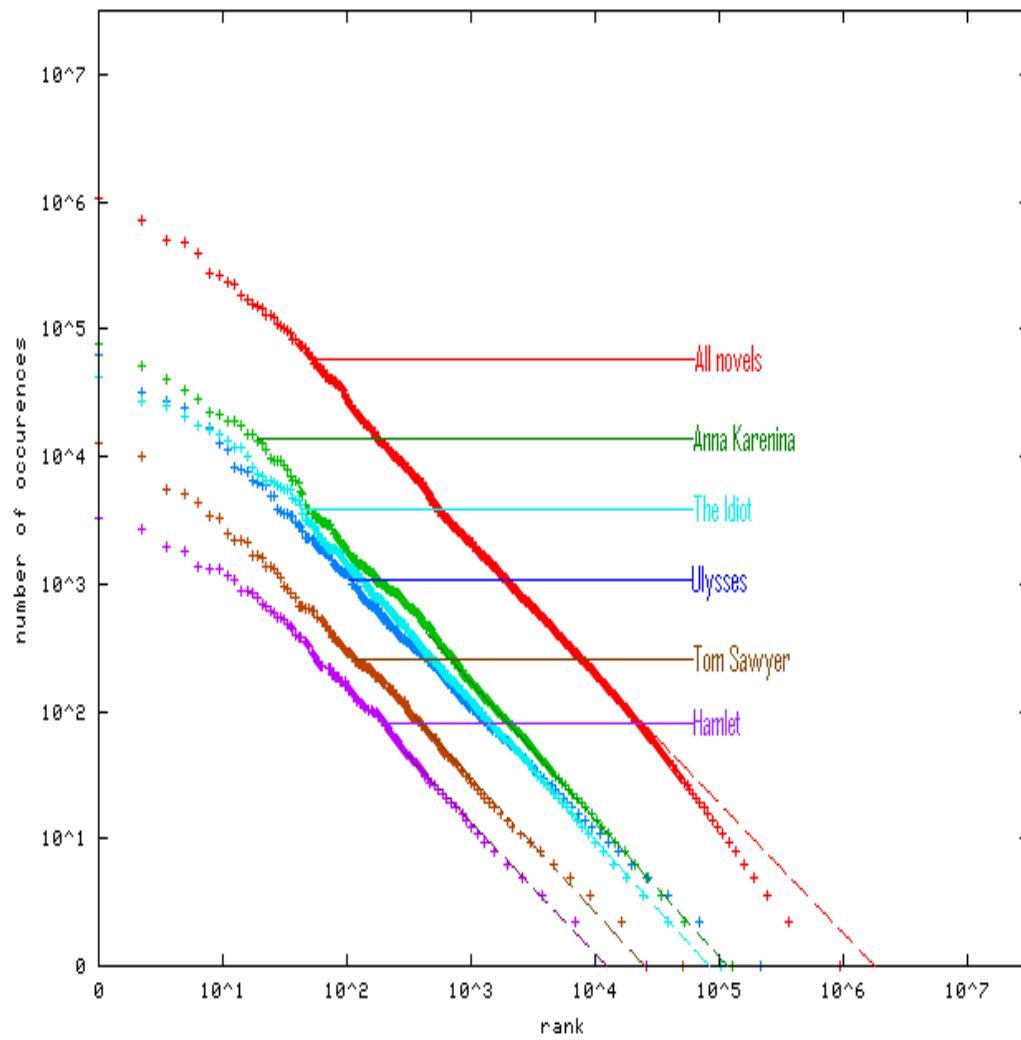


Figure 5: Word frequency in all the novels listed above when grouped together, as well as in five pieces when considered separately.

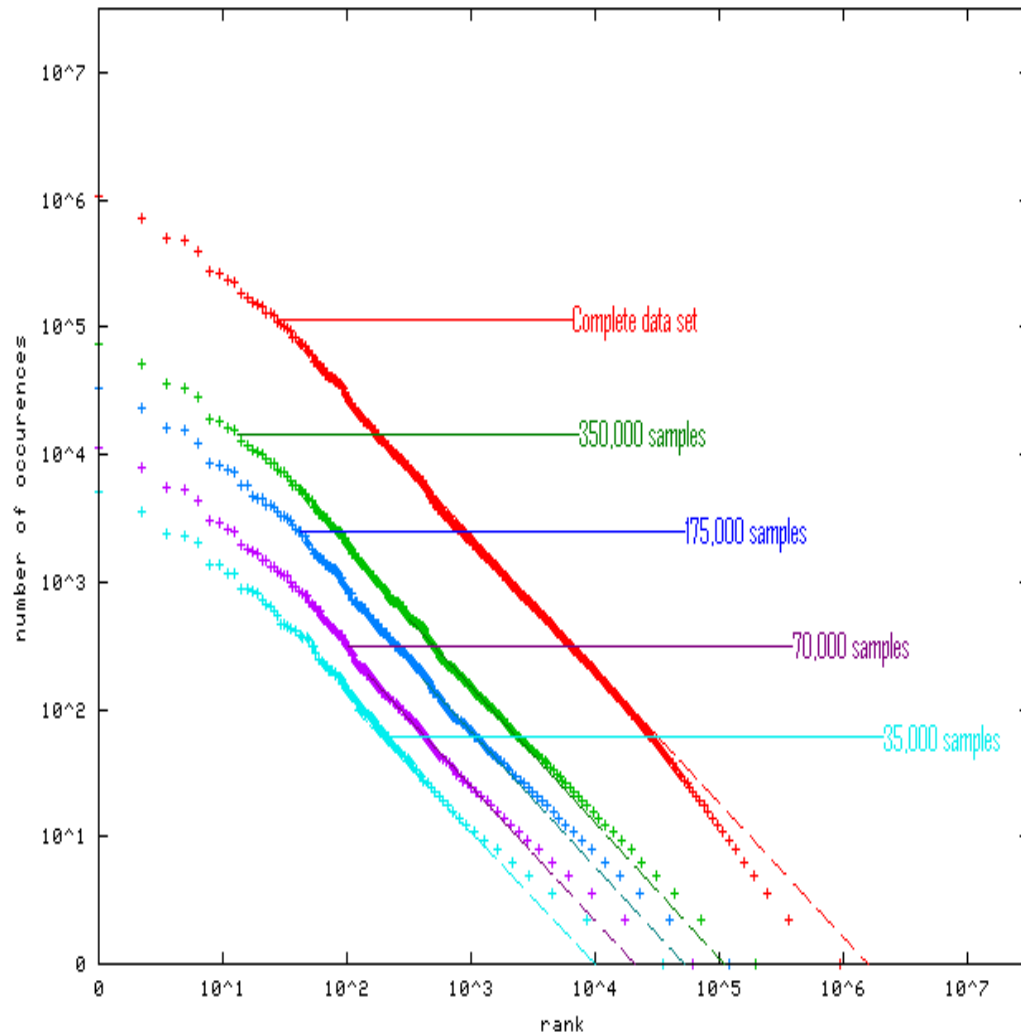


Figure 6: Word frequency in all the novels listed above when grouped together, as well as in four sub-samples of decreasing size.

An interesting trend in **Figure 6** (which can also be observed in **Figure 4** and in **Figure 5**, albeit to a lesser degree) is that small enough sub-samples diverge from Zipf's law in being above the straight line at the numerically (high) ranks (i.e., at the bottom-right part of the figure). This trend is the opposite of the trend displayed by large populations that we have considered, which tend to drop below the straight line at the high ranks.

To gain a more accurate notion of the degree to which each of the data sets diverges from the straight line predicted by Zipf’s law, we used simple linear regression.² The results we obtained are presented in **Table 2** below:

Data set name	Prediction error
Optical Computing—7,500 samples	0.4964
Optical Computing—15,000 samples	0.6721
Ulysses	1.2380
Optical Computing—30,000 samples	1.2720
Tom Sawyer	1.3395
Optical Computing—60,000 samples	1.8595
All works grouped together—35,000 samples	2.5730
All works grouped together—70,000 samples	2.9909
Bible—40,000 samples	3.1892
Optical Computing	3.2081
All works grouped together—350,000 samples	3.2497
Hamlet	3.2549
All works grouped together—175,000 samples	3.3020
The Idiot	3.5737
Anna Karenina	3.6353
Bible—80,000 samples	5.0632
All works grouped together	5.3616
Bible—160,000 samples	6.8803
Bible—320,000 samples	9.3697
Bible	14.4449

Table 2: Prediction error of simple linear regression when applied to the data sets based on corpora of natural language.

3 Discussion

All the results listed above point to the same trend: Increasingly significant divergences from Zipf’s law are registered as the number of events considered grows larger. More specifically, our results seem to imply that as the number of events considered rises, “the weak grow weaker”; events that rank high (numerically) tend to appear less frequently than they are expected to appear according to Zipf’s law.

²Simple linear regression is also known as least squares regression. The line it considers to be the best fit is the one for which the sum of squares of the residuals is least. Calculation of the line’s equation proceeds according to the following equations:

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

and

$$b = \bar{y} - a\bar{x}$$

The predictor line is thus given by $y = ax + b$, and prediction error is the sum of squares of the residuals.

In the Zipf graphs presented above, the straight line observed when a large number of events is considered is replaced by a line whose slope becomes more and more steep as it approaches the x-axis, when the number of events taken into account is relatively small. Numerical results obtained in the word frequency domain affirm this trend. Linear predictors produce higher prediction error when the number of words taken into account grows larger. This phenomenon is consistent across all the texts we analyzed. Interestingly, results of the same nature are reported by Nemeth and Zainko [4], who investigate word frequency in large corpora of natural language. Their conclusion is that Lavalette’s law is better suited to low frequency items than Zipf’s law.³

We believe that all the cases we reviewed—as well as many other cases—share in common the following properties:

- The collection of events that are available for consideration is finite.
- The available events follow a strong version of the Pareto principle: The probability of occurrence of a very small number of events is very high, and the opposite is true of the rest of the events in the collection. For example, a small number of words in English tends to appear very frequently in English texts, whereas all the rest of the words appear very rarely on average.
- The high frequency events follow (an approximation of) the Pareto distribution.

Assuming that a certain collection of events, C , satisfies these three properties, if we draw n independent samples from C according to the empirical distribution underlying the events in C , we would expect the following:

- For small values of n (relative to the size of C), high frequency events are likely to obscure the existence of low frequency events almost completely, and thus the result is expected to approximate a Pareto distribution (and hence fit nicely with Zipf’s law).
- For high values of n , the presence of low frequency events should be felt. These events may or may not follow the Pareto distribution, but—regardless

³Lavalette’s law is given by the equation

$$P(E_r) = C \times \left(\frac{r \times r_{max}}{r_{max} - r + 1} \right)^{-a}$$

where E_r is an event of rank r and r_{max} is the highest attainable rank.

of the distribution they follow—the decay exponent corresponding to high frequency events cannot be made to fit these events. The outcome of n samples—expressed in terms of a Zipf graph—is thus a linear fit in the (numerically) low ranks that breaks at the high ranks into a line whose slope becomes increasingly steep.

At this point, we leave open the question whether it is worthwhile to view the low frequency events as following (an approximation of) a Pareto distribution with a greater decay exponent.

Since Zipf’s law is, in fact, a claim to the effect that certain populations of events are distributed according to the Pareto distribution (hence the straight line), the results we obtained can be explained analytically. The law—in its classic formulation—tells us that on a log-log scale, rank and frequency are related by a straight line whose slope is -1 . This implies that if the size of the population is n , then the number of samples should be

$$\sum_{i=1}^n \frac{n}{i} = n \sum_{i=1}^n \frac{1}{i} \approx n \log(n)$$

Figure 7 illustrates this state of affairs.

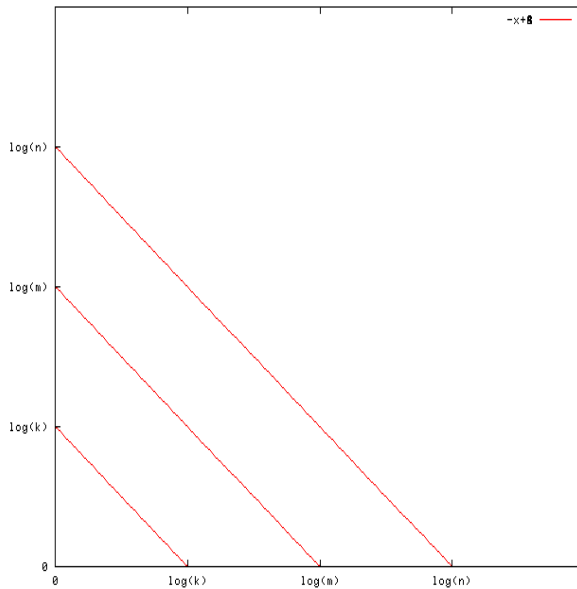


Figure 7: The relationship between frequency and rank as predicted by Zipf’s law. If we sample $k \log(k)$ times, then we expect to see k different members of the population. This ratio should still hold if we increase the number of samples.

However, if the population under consideration is finite, then at some point we cease to see new members. Suppose that the size of the population is t , and let $s > t$. We expect to see s different members of the population if we sample $s \log(s)$ times, but this is an impossibility given that the size of the population is t . This means that our series of straight lines should break at some point. This state of affairs is illustrated by **Figure 8**.

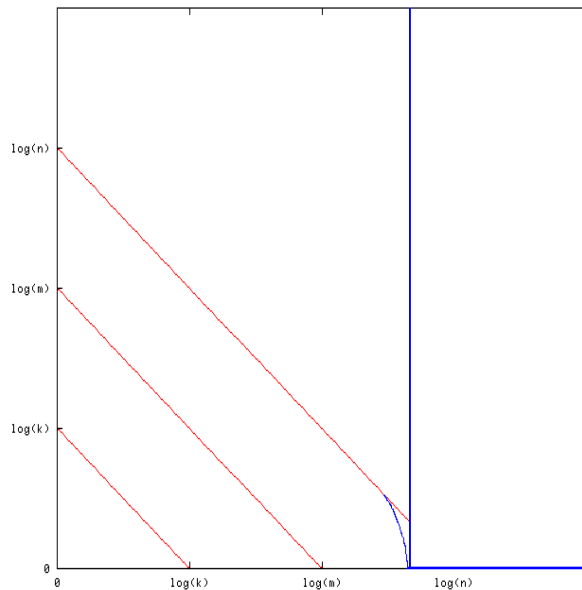


Figure 8: When the population under consideration is finite, Zipf's law cannot be satisfied. If we sample enough times, then the ratio between the number of samples and the number of different members in the samples breaks. The straight line that runs all the way to the x-axis becomes an impossibility, and thus—as we reach the higher ranks—the blue line takes its place.

Intuitively, our claim—as applied to the examples we surveyed above—is that there is a very limited number of projects in Sourceforge which people tend to download. The rest of the projects are rarely ever visited. Similarly, there are few web pages in the 1998 World Cup website that draw people's interest. The remaining pages are hardly ever accessed. When considering only the 10,000 most popular projects in Sourceforge or the 1,000 most visited pages in the World Cup website, we should be able to (allegedly) witness Zipf's law in action. However, as we have shown, this ceases to be true when these collections of events are sampled more intensively.

4 Future Work

Our work has shown that Zipf's law should be taken with a grain of salt, if not more carefully. We have shown that it is worthwhile—at least in some cases—to distinguish between two types of events whose probabilities of occurrence are dramatically different. High frequency events (of which there are few) seem to follow the Pareto distribution. As for low frequency events, the question which distribution—if at all—best fits them is left open.

A more fundamental question is if there is one law that can effectively govern both types of events. Lavalette's law is reported to be more successful in the low frequency range, albeit with some reservations [4]. It remains to be seen whether Lavalette's law fares better with the examples we present above, as well as with other examples. Mathematical analysis of Lavalette's law's ability to cope with collections of events that satisfy the properties we list above may also be useful.

We also leave open the question of how general the results we present here really are. More research in other domains is likely to answer this question.

References

- [1] Ramon Ferrer-i-Cancho and Ricard V. Sole. Zipf's law and random texts. *Advances in Complex Systems*, 5(1):1–6, 2002.
- [2] Thorbjørn Knudsen. Zipf's law for cities and beyond: The case of Denmark. *American Journal of Economics and Sociology*, 60(1):123–146, 01 2001. available at <http://ideas.repec.org/a/bla/ajecsc/v60y2001i1p123-146.html>.
- [3] M. A. Montemurro. Beyond the zipf-mandelbrot law in quantitative linguistics. *Physica A Statistical Mechanics and its Applications*, 300:567–578, November 2001.
- [4] Geza Nemeth and Csaba Zainko. Multilingual statistical text analysis, zipf's law and hungarian speech generation. *Acta Linguistica Hungarica*, 49(3–4):385–401, 2002.
- [5] George Mikros Nick Hatzigeorgiou and George Carayannis. Word length, word frequencies and zipf's law in the greek language. *Journal of Quantitative Linguistics*, 8(3):175–185, 2001.