2021. *Proc Ling Soc Amer* 6(1). 394–402. https://doi.org/10.3765/plsa.v6i1.4975.

Speech tested for Zipfian fit using rigorous statistical techniques

Paul De Palma, Leon Antonio Garcia-Camargo, Jeb Kilfoyle, Mark VanDam & Joseph Stover*

Abstract. Zipf's law describes the relationship between the frequencies of words in a corpus and their rank. Its most basic form is a simple series, indicating that the frequency of a word is inversely proportional to its rank:

$$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots$$

The past two decades have seen the emergence of usage-based and cognitive approaches to language study. A key observation of these approaches, along with the importance of frequency, is that speech differs in substantial and structural ways from writing (Bybee, 2010; Miller & Weinert, 1999; Tomasello, 2003). Yet, except for a few older analyses performed on very small corpora, all studies of Zipf's law have been done on written corpora. Further, a judgement of Zipfianness in much of this work is based on loose and informal criteria (Ha, *et al.*, 2002). In fact, sophisticated statistical techniques have been developed for curve fitting in recent years in the mathematics and physics literature (Newman, 2005; Clauset *et al.*, 2009). These include the use of the Kolmogorov-Smirnov statistic, along with maximum likelihood estimation to generate p-values and the use of the complementary error function for normal distributions. The latter helps determine if a corpus, failing a Zipfian fit, might be better described by another distribution. In this paper, we will:

- Show that three corpora of recorded speech follow a power law distribution using rigorous statistical techniques: Buckeye, Santa Barbara, MiCase.
- Describe preliminary results showing that the techniques outlined in this paper may be useful in the diagnoses of those conditions that can include disordered speech (MacWhinney *et al.*, 2011).
- Explain how to do the analyses described in this paper.
- Explain how to download and use the R/Python code we have written and packaged as the *Zipf Tool Kit*.

Keywords. computational linguistics; Zipf; ASD

1. Introduction. Zipf's law describes the relationship between the frequencies of words in a corpus and their rank. Its most basic form indicates that the frequency of a word is proportional to the inverse of its rank, expressed in this series:

* Acknowledgements: The authors would like to thank Robert and Claire McDonald and the McDonald Work-Award Program for their generous and continuous support for undergraduate research assistants. They would also like to thank former students, Bethany Bogensberger and Allison Hayes, who contributed to earlier versions of this work. Authors: Paul De Palma, Gonzaga University (depalma@gonzaga.edu), Leon Antonio Garcia-Camargo, Gonzaga University (lgarcia-camargo@zagmail.gonzaga.edu) , Jeb Kilfoyle, University of New Mexico (jkilfoyle@.unm.edu), Mark Vandam, Washington State University (mark.vandam@wsu.edu) & Joseph Stover, Gonzaga University(stover@gonzaga.edu).

$$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots$$

Work on what has come to be called Zipf's law has been continuous, engaging researchers who have made significant contributions elsewhere. Benoit Mandelbrot, of the Mandelbrot set, and Nobel Prize winner and early AI researcher Herbert Simon are two examples (Fedorowicz, 1982). Much of the work on word frequency distribution is forbiddingly mathematical (Fedorowicz, 1982; Baayen, 2001), some quite informal in its judgement of Zipfianness (Ha, *et al.*, 2002), some packaged with psychological and linguistic assumptions which may or may not be accurate (Simon, 1955, cited in Fedorowicz, 1982), and some are concerned primarily with the underlying cognitive/linguistic mechanisms which produce the Zipfian distribution of words (Piantadosi, 2014). A few have looked at speech, but with tiny samples (Ridley & Gonzales, 1994). Those examining speech with somewhat larger samples focus not on speech, but on deriving yet more accurate closed form formulae for Zipf's law (Baixeroes, *et al.*, 2011).

No one appears to have examined large corpora of transcribed speech for Zipfianness. This seems curious. The past two decades have seen the emergence of usage-based and cognitive approaches to language study. A key observation of this work is that speech differs in substantial and structural ways from writing (Bybee, 2010; Tomasello, 2003; Miller & Weinert, 1999). Yet many Zipfian studies, perhaps most, and including those of Zipf himself, appear not to distinguish between speech and writing. Baixeroes, *et al.*2011) observe that for most of the seventy odd years that Zipf's law has had a name, large recordings of speech, along with the ability to process it, have simply not existed. We offer another hypothesis. Most of the work on Zipf's law has occurred outside of linguistics. Mandelbrot was a mathematician, Simon was an economist and cognitive psychologist, Le Quan Ha is a computer scientist, Newman is a physicist, Piantadosi is a cognitive scientist, to offer just a few of many possible examples. We assume that these scientists, though eminently competent in their own fields, were simply unaware of emerging work in linguistics.

Our approach to Zipf's law is different from what we find in the literature, in another way, though motivated by Zipf's own investigation of the speech of schizophrenics (Zipf, 1949). We are not interested in a general formulation of Zipf's law that would cover all of language in all of its richness. Rather, we begin with a general power law distribution and ask how closely an actual distribution, say adult conversational speech, or the speech of children clinically diagnosed with autistic spectrum disorder, approximates it. After a quick look at our materials in section 2, we turn to just such a discussion section 3. In Section 4, we present the Zipfian Tool Kit, a collection of programs that researchers can use in their own investigations of Zipf's Law.

1.1. A FEW DETAILS. After Zipf's initial work, much analysis appears to have been efforts to modify Zipf's original formulation to account for all of language. Here is Zipf's 1949 formulation:

$$r \times f = C \tag{1}$$

where f is the frequency of of occurrence of a word in a text, r the rank of that word by frequency and C a proportionality constant. A plot of rank order of words against their frequency

in *Ulysses* is shown in Figure 1. It is the characteristic plot found in all studies of Zipf's Law.

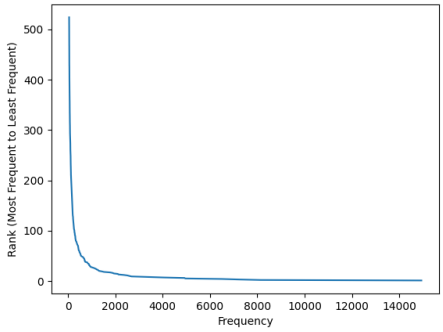

Figure 1. Word rank against word frequency

One of the more surprising aspects of the law is that it describes phenomena far afield from word distributions. Magnitudes of earthquakes, intensities of solar flares, citations of scientific papers, web hits, wealth, family names, and city populations all appear to follow a Zipfian distribution (Newman, 2005). Figure 2 is a plot of the rank order of the world's richest people against their wealth (*Bloomberg Billionaire's Index*, 2020). What these phenomenon have in common is intuitively obvious. There are very few of each with high numbers (the richest of the rich, large earthquakes, heavily cited papers, frequently used words, etc.), and a great many with low numbers (the merely rich, small earthquakes, rarely cited papers, rarely used words, etc.).

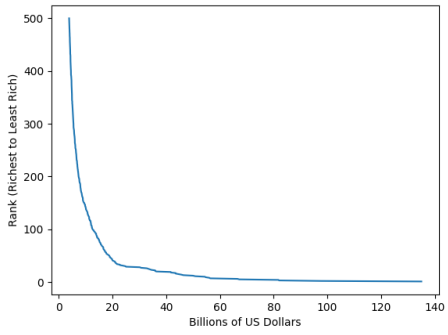

Figure 2. Rank of the rich against wealth

A little algebra (see Newman, 2005) lets us derive a closed-form formula, Eq. 2, to describe Zipfian phenomena. In this more general form, it is usually referred to as a *power law*, in fact, exactly the form which describes the magnitude of earthquakes, the intensity of solar flares, and so on. If we take α to be 1, as Zipf did, then we have Eq.(1), which Zipf used to

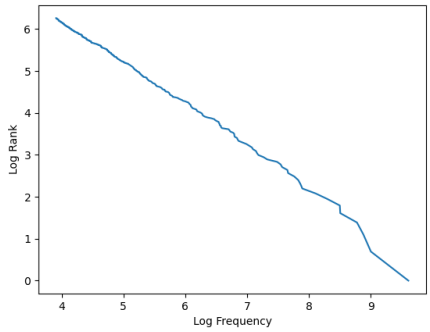

Figure 3. Log rank against log word frequency

express the relationship he discovered empirically between word rank and word frequency in *Ulysses* (Zipf, 1949).

$$R(f) = Cf^{-\alpha} \tag{2}$$

The question we ask is not how to develop ever more general formulations of Zipf's Law, but rather, to ask how closely data from a variety of corpora fit a power law. In the next two sections we will present preliminary results 1) indicating that conversational speech forms a Zipfian distribution; and 2) indicating that the speech of children clinically diagnosed with Autistic Spectrum Disorder (ASD) does not. In our usage, conversational speech is Zipfian, and the speech of ASD children is not.

2. Materials and methods.

2.1. MATERIALS. We used the following materials in this work:

- The Python programming language (python.org) and matplotlib (matplotib.org) to tokenize the corpora and do rank/frequency plots.
- The R Project for Statistical Computing software (r-project.org) to encode the techniques described in the next sub-section.
- The corpora listed in Tables 1 and 2, along with James Joyce's *Ulysses* from Project Gutenberg (Project Gutenberg, 2020).

Corpus	Type	No. Words
Buckeye	SS	277,795
Santa Barbara	SS	234,259
MiCase	AS	1,652,747

Table 1. Conversational speech corpora

Buckeye (Pitt *et al.*, 2007) and Santa Barbara (Du Bois *et al.* 2000-2005) are collections of spontaneous speech. MiCase (Simpson, *et al.*, 2002) is also spontaneous speech, but in the academic register. Since Miller and Weinert (1998) suggest that such speech has elements of written language, we will call this corpus *academic speech*, AS.

Corpus	Development	Age	No. Words
Assymetries	ASD	7-12	10,436

Table 2. Traditional developing and autistic spectrum disorder speech corpora

The Asymmetries transcripts, containing transcribed speech of children diagnosed with autistic spectrum disorder (ASD) were taken from TalkBank/ASDBank (MacWhinney, 2000). The number of words is small compared to the conversational speech corpora. Nevertheless it is sufficient for the statistical techniques we use, and which are described in the next sub-section.

2.2. METHODS. In recent years, researchers int the mathematics, physics, and econometrics communities have developed a collection of statistical techniques that can be used to judge whether a given distribution of empirical data fits a power law or is better described by another distribution (Vuong, 1989; Newman, 2005; Clauset, 2009). Though a detailed explanation of these techniques is beyond the scope of this paper, the basic idea is clear enough. A straight line with slope of -1 begins to break down at a certain point in the distribution (see Figure 3). Following Clauset, we compute the Kolmogorov-Smirnov statistic to compare the maximum distance between the area under the curve describing a given corpus with data of a theoretical model, as in Figure 4, (though, in the figure, for an arbitrarily chosen data set) (Rojas-Lima, *etal*., 2019). This lets us generate p-values and compare the empirically generated distribution with a theoretical power-law distribution. Our method allows us to compute p-values with an uncertainty less than ± 0.01. We use a p-value of less than 0.05 as the cutoff for power-law behavior. P-values greater than 0.1 are considered strong evidence of a power-law.

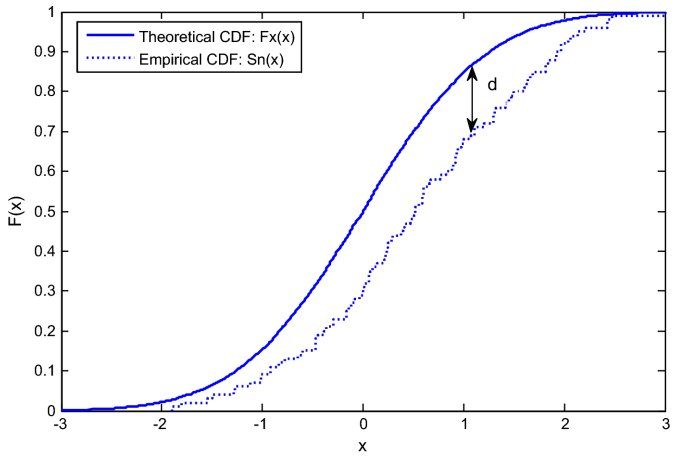

Figure 4. KS Distance

It is also possible that another distribution fits empirical data more closely than does a power law. To compare power-law distributions to other distributions, we use a log likelihood ratio test proposed by Vuong (1989), and cited in Clauset, *et al.*, (2009). A sufficiently large negative *R* value favors the alternative distribution, whereas a small negative or positive value favors the power law (see Table 4). The alternative distributions we consider are log-normal, Poisson, and exponential.

3. Is speech Zipfian?

3.1. CONVERSATIONAL SPEECH. In this section, we present preliminary results indicating that conversational speech is Zipfian, whereas the speech of children diagnosed with autistic spectrum disorder is not. We first consider conversational speech as represented by Buckeye, MiCase, and Santa Barbara corpora. Table 2 presents p-values for the three corpora under investigation.

Corpus	P-Value
Buckeye	0.06
MiCase	0.15
Santa Barbara	0.58

Table 3. P-values by conversational corpus

Using a conservative p-value of $p \geq 0.10$ both MiCase and Santa Barbara are strongly Zipfian. If we use the more liberal cutoff of $p \leq 0.05$, all three corpora may be described using a power law. Why Buckeye is the least Zipfian of the three corpora will be the subject of further investigation, though we should note that p-value differences are not necessarily linear.

As noted in the previous section, relatively low p-values, as in the case of Buckeye, could indicate that the a different distribution might describe the data better than the power law distribution. Table 3 shows the comparison of the power law distribution to other distributions. A p-value ≥ 0.10 indicates a significant comparison. A sufficiently large negative *R* value favors the alternative distribution, whereas a small negative or positive value favors the power law. The take-away is that Buckeye appears to be described by a log-normal distribution.

Finally, none of the corpora are either Poisson or exponential. In fact, when they deviate from the power law distribution, they deviate towards log-normal. This is not surprising given that with respect to tail decay–large x-values becoming rare–power law and log-normal distributions are similar, as are Poisson and exponential distributions.

P-Value, R	Log-normal		Poisson		Exponential	
Buckeye	0.06	-1.82	0.00	6.85	0.00	7.80
MiCase	0.50	0.67	0.00	6.48	0.00	7.11
Santa Barbara	0.43	-0.79	0.00	6.91	0.00	13.67

Table 4. Comparison test by corpus

3.2. SPEECH OF CHILDREN DIAGNOSED WITH AUTISTIC SPECTRUM DISORDER. Children with a clinical diagnosis of autism spectrum disorder (ASD) have long been known to exhibit various kinds of language performance deficits, particularly in pragmatics, and more generally in discourse. They also exhibit developmental difficulties with the lexical-semantic and grammatical aspects of language (Condouris, Meyer, Tager-Flusberg, 2003; Kjelgaard, Tager-Flusberg, 2001). Though there have been at least two studies with very small samples that attempt to determine if the speech of adults with aphasia and Alzheimer's follow power law distributions (Van Egmond, *et al.*, 2015; Hernandez-Fernandez, Diaquez-Vide, 2013), we know of no similar studies of children diagnosed with autism.

As indicated in Table 5, the p-value for the sample is far below even the liberal 0.05 cut-off for Zipfianness.

Corpus	Age	P-Value
Asymmetries	Age	0.0151

Table 5. P-value for ASD child speech

4. The Zipfian Tool Kit. In this section, we describe the software we developed to analyze frequency of words and test for Zipfianess. It is based on the techniques described in Clauset (Clauset, *et al*, 2009). and is freely downloadable from GitHub. The package includes software to tokenize data, generate frequency tables, generate rank vs frequency log-log plots, generate p-values to determine if distribution follows a power-law, and to compare a power law fit with poisson, exponential, and log-normal distributions.

4.1. DETAILS. The code for this package has was developed using the R Project for Statistical Computing (r-project.org) to encode the techniques, the Python programming language (python.org), matplotlib (matplotib.org), and the Natural Language Toolkit (NLTK) to tokenize written corpora and create rank/frequency plots.

URL https://github.com/lgarcia-camargo/Zipf-Toolkit

Depends R ($\geq$ 3.4.0), python ($\geq$ 2.7)

Imports python: NLTK, pandas, matplotlib. R: poweRlaw

Encoding UTF-8

4.2. DESCRIPTION. When analyzing the distribution of words in a corpora it is necessary to tokenize and count frequencies, and visualize the data. Earlier studies compared the distribution of a corpus on a log-log plot to that of a line with the slope -1 in order to determine its Zipfianness.

The following command will tokenize a file, using the NLTK tokenizer (NLTK.tokenize, 2020), count frequencies of words, and output the results to a .csv file:

```
$ make csv
```

Once the .csv file is created, the distribution can be analyzed with a visualization of a log-log frequency vs. rank plot. Note that in the exposition above, for reasons of clarity, we plotted the rank against the frequency, though Zipf himself plotted frequency against rank. The following command will create this plot, given the .csv file.

$ make create

The output will be similar to Figure 3.

Yet simply looking at a log-log plot of the distribution is not enough to determine its fit. With the .csv file, the distribution of frequencies can be analyzed using techniques described in Clauset et. al. (Clauset, 2009). To do this we provide an .R file that will generate a file with information about the corpus, including word count, xmin, xmax, p-value, r and p comparison values. To generate statistical analysis of distribution do the following: 1) Invoke the R environment by entering 'R' into terminal. 2) Enter the following command and follow prompts:

> source('getInfo.R')

The program will run for a set amount of time depending on the size of the corpora. After completion, the data will be output to a text file. When analyzing the output of the statistical test the important steps to determine Zipfianess are the following:

1) Using the word count, xmin, and xmax (not shown in this paper) determine if enough of your data is being sampled for the methods. If the xmin or xmax is removing the majority of the data, the results of the test are not viable.

2) Using the conservative cutoff value of .05 or liberal cutoff of .1, determine if the p-value indicates a Zipfian distribution.

3) If the p-value indicates a positive result, using the p, R values for each comparative distribution, determine which distribution is favored, how strongly it is favored, and whether the difference is significant.

5. Discussion and Future Work. Since Zipf first formulated his observations over a twenty year period beginning in 1929, there has been almost continuous interest, mostly in attempting ever-better refinements of what has come to be called Zipf's Law. The development of inexpensive computing hardware and accompanying software in the last two decades have allowed us to look at Zipf's Law using sophisticated, but compute-intensive, statistical techniques. The parallel development of inexpensive recording and storage devices has opened speech to computational investigation. In future work, we plan 1) a full explanation of the statistical techniques used in this research, but one readable by non-mathematicians; 2) a further investigation as to whether speech is Zipfian by using more and more diverse speech corpora; 3) a more complete investigation into the speech of children in general, and ASD children in particular, to determine Zipfianness. We hypothesize that the speech of ASD children is non-Zipfian. Including the speech of typically developing children as a (presumably Zipfian) benchmark will offer evidence for that hypothesis. If the hypothesis is confirmed, the techniques described here may be a useful complement to current techniques for diagnosing autism in children.

References

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Bloomberg Billionaire Index. 2020. https://www.bloomberg.com/billionaires.

Clauset, Aaron, Cosma Rohilla Shalizi & M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4). 661–703. https://doi.org/10.1137/070710111.

Condouris, Karen, Echo Meyer & Helen Tager-Flusberg. 2003. The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology* 12. 348–358. https://doi.org/10.1044/1058-0360(2003/080).

Fedorowicz, Jane. 1982. The theoretical foundation of Zipf's law and its application to the bibliographic database environment. *Journal of the American Society for Information Science* 33(5). 285-293. https://doi.org/10.1002/asi.4630330507.

Francis, W. Nelson & Henry Kucera. 1979. *Brown corpus manual. Manual of information to accomapany a standard corpus of present-day edited American English for use with digital computers*. Providence, RI: Brown University.

Ha, Le Quan, E. I. Sicilia-Garcia, Ji Ming & F. J. Smith. 2002. Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*. https://www.aclweb.org/anthology/C02-1117.

Hernández-Fernández, Antoni & Faustino Diéguez-Vide. 2013. La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer. Anuario de Psicología 43(1). 67–82.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk* (3rd Edition). Mahwah, NJ: Lawrence Erlbaum.

Newman, Mark E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46(5) 323–351. https://doi.org/10.1080/00107510500052444.

NLTK.tokenize Package. 2020. https://www.nltk.org/api/nltk.tokenize.html.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin Review* 21(5). 1112–1130. https://doi.org/10.3758/s13423-014-0585-6.

Ridley, Dennis R. 1982. Zipf's Law in transcribed speech. *Psychological Research* 44. 97–103. https://doi.org/10.1007/BF00308559.

Simon, Herbert A. 1955. On a class of skew distribution functions. *Biometrika* 43(3,4). 425–440.

Rojas-Lima, J., F. A. Domínguez-Pacheco, C. Hernández-Aguilar, L. M. Hernández-Simón & A. Cruz-Orea. 2019. Kolmogorov–Smirnov test for statistical characterization of photopyroelectric signals obtained from maize seeds. *International Journal for Thermophysics* 40(4). https://doi.org/10.1007/s10765-018-2462-4.

van Egmond, Marjoelein, Lizet van Ewijk & Sergey Avrutin. 2015. Zipf's Law in non-fluent aphasia. *Journal of Quantitative Linguistics* 22(3). 233–249. https://doi.org/10.1080/09296174.2015.1037158.

Vuong, Quang H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2). 307–333. https://doi.org/10.2307/1912557.

Zipf, George K. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley.

Zipf, George K. 1932. *The psycho-biology of language*. Cambridge, MA: Harvard University Press.

Zipf, George K. 1929. Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology* 40. 1–95.