Word Frequency Distributions across Languages

Trudie Strauss[1] Michael J. Von Maltitz[1] Damián E. Blasi[2]

[1]Department of Mathematical Statistics and Actuarial Science
University of the Free State
South Africa

[2]University of Zurich
Switzerland

TBI Winterseminar, Bled, 2018


Take
Note

Outline

1. Introduction
   - Importance of Word Frequency
   - History
2. Approach
   - Overview
   - Data and Methods
   - Expectations
3. Preliminary Results
   - Results
   - Explanation

Importance of Word Frequency

- Word frequency distributions are a central object of study in the language sciences
- frequency of words determines many important phenomena in language
- *e.g.* age of acquisition, rate of change through time...

History

- Simple and explicit parametric models: power-law distributions
- Zipf's Law:

Zipf's Law

$$f(r) \propto \frac{1}{r^{\alpha}}$$

for $\alpha \approx 1$

- Adapted, "improved" models, higher complexity
- Why do word frequencies follow the distribution they do?

Our Approach

- data-oriented
- available data
- computing power

Calculate 32 word frequency distribution measures of lexical diversity as multidimensional space describing the distribution, *e.g.*:

- mean frequency of words
- skewness
- kurtosis
- entropy
- number of hapax/dis legomena

Token / Type

Tokens - total number words in a text
Types - number of unique words in a text

  - best Zipf parametric fit for each text; compare every measure with simulations from simulated theoretical distribution

Data

For each language in the Leipzig Corpus [1]

- download largest, most recent Wikpedia text file (sentences)
  - 90000 to 20000000 words
- create word list with R-package, `tidytext`

[1]D. Goldhahn, T. Eckart & U. Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the 8th International Language Ressources and Evaluation (LREC'12), 2012

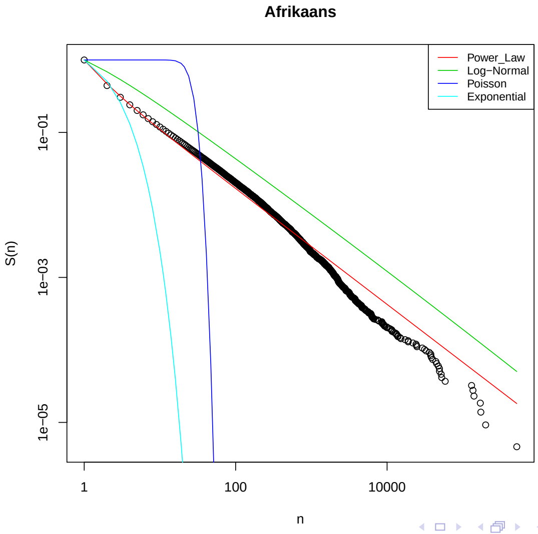
Method

We fit the following parametric models to the data:

- power law
- log-normal
- exponential
- Poisson

With parameters estimated from the empirical data, using package `poweRlaw`

Example language: Afrikaans


Afrikaans
Power_Law
Log−Normal
Poisson
Exponential
S(n)
1e−01
1e−03
1e−05
1
100
10000
n

Method

Empirical Data:

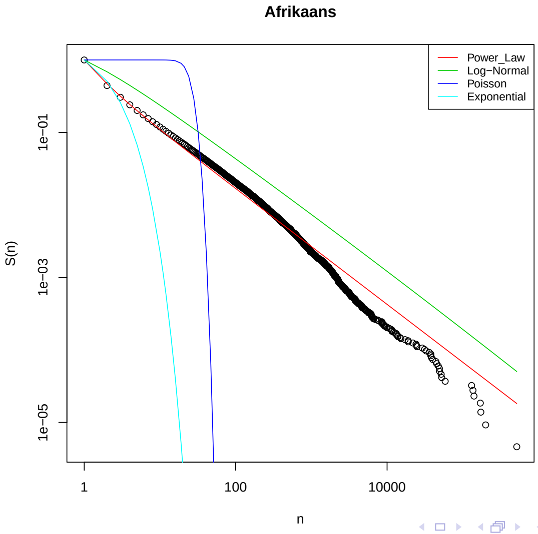
- Define $n_i = 1000 : N$ (for $i = 1 : 100$)
- sample n_i words from the initial word list
- determine value for each of the measures
- Each language: $100_{\text{values of } n_i} \times 35_{\text{measures}}$

Simulated Data

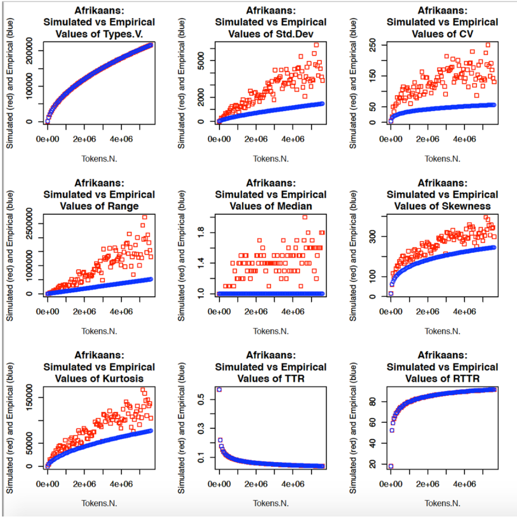
- Fit power law distribution to each sample, using the α value calculated on entire data set
- Simulate from theoretical distribution for $n_i = 1000 : N$ (for $i = 1 : 100$)
- value of each measure, mean over all simulations for n_i
- $100_{\text{values of } n_i} \times 35_{\text{mean of simulated measures}}$

Expectations

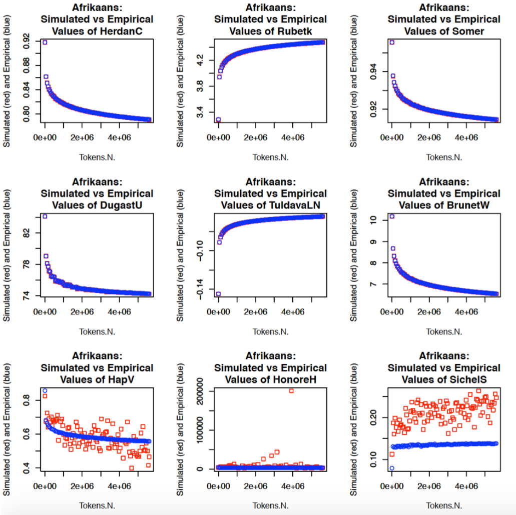
- Go beyond simple parametric models, and determine
  - individual measures differ across languages
  - to what extent
  - influence of N
- From the power law fit we see that the power law distribution is reasonable for some N
- We therefore expect that
  - empirical measures should correspond to simulated measures for certain values of N
  - we can identify "optimal" N for which measures correspond to theoretical distribution.


Afrikaans
Power_Law
Log−Normal
Poisson
Exponential
S(n)
1e−01
1e−03
1e−05
1
100
10000
n

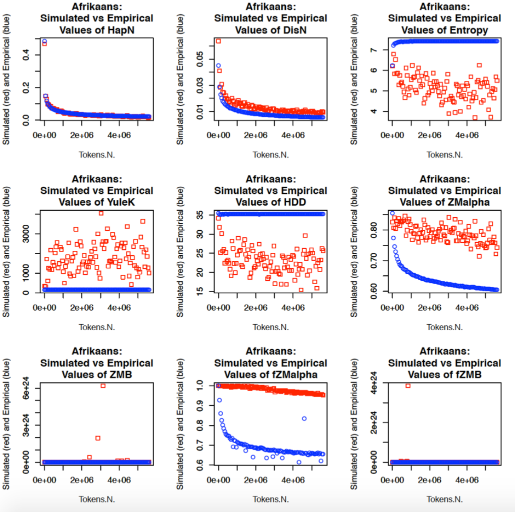
Individual Measures



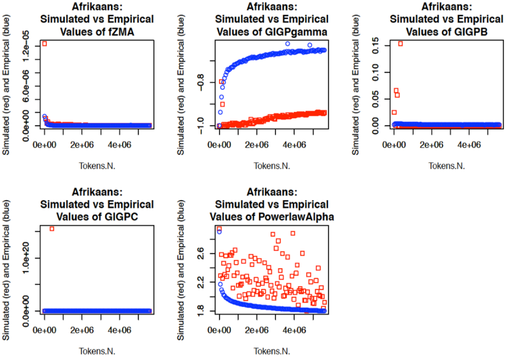
Individual Measures



Individual Measures



Individual Measures



But why?

assuming analysis was done correctly, this discrepancy between the power law fit and the individual measures, could make sense:

- Expected:
  - while Zipf seems to be a good approximation of the distribution as a whole
  - when you zoom in, it fails to deliver in many respects
- data/reasoning/algorithmic errors

Thank you

- Damián E. Blasi

University of the Free State

- Michael J. von Maltitz
- Sean van der Merwe

Leipzig University

- Peter Stadler
- Nancy Retzlaff
- Sarah Berkemer

Funding

- South African National Research Foundation
- Knowledge, Interchange and Collaboration Grant