Estimating Lexicon Size Based upon Zipf's Law: A Novel Mathematical Approach

William Kariampuzha, Aetizaz Sameer, Olivia Chen, Hannah Kang
Aaron Flores, Dr. Aaron Braver

TEXAS TECH UNIVERSITY College of Arts & Sciences
TEXAS TECH UNIVERSITY Honors College

Abstract

Estimating the number of words in a language or lexicon size is a complicated endeavor due to the the ever-evolving nature of language. Previous attempts to estimate the lexicon size have been capped by the size of databases available for written texts. Here, we demonstrate a novel mathematical method for estimating the size of the lexicon using Zipf's Law. Zipf's Law approximates how often a particular word in a language is used. It does this based on the statistical rank of the word and the frequency of the most used word in a language at any point in time. Using data from the Brown Corpus and the Corpus of Historical American English, we obtained preliminary estimates for the size of the lexicon.

Quantifying the size of the lexicon can allow us to better understand and model the rate of change of any language. Natural language processing systems (e.g. Siri, Amazon Alexa, Google Assistant) will need to evolve with ever-changing words and word meanings. These derived formulas may allow these voice-based systems to adjust their databases in order to function accurately in our rapidly changing world.

$$f_i = f_1\left(\frac{1}{i}\right)$$

Zipf's Law

Introduction

Estimating lexicon size is complicated due to the ever-evolving nature of language. Existing procedures to estimate the size of a lexicon are limited by the sources of the corpora that are used to create existing estimates.

Michel et al. estimated the size of the lexicon using the truncated Google Ngrams corpus with random sampling to obtain 597,000 existing words in 1950 and 1,022,000 words in 2000.

This novel approach uses Zipf's Law to find the number of words in a language (lexicon size). Zipf's Law is a statistical phenomenon following a harmonic progression that was first described in the relative word frequencies in corpora.

A corpus (pl. corpora) is a very large database of words and other lexical items that allow quantification of macro-linguistic phenomena such as word use (frequency) over time.

The harmonic series is the sum of all of the terms in the harmonic sequence:

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}$$

The *p*-series or hyperharmonic series is the generalization of the harmonic series:

$$\frac{1}{1^p}+\frac{1}{2^p}+\frac{1}{3^p}+\ldots+\frac{1}{n^p}$$

where p is obtained from a logarithmic regression on the frequency data.

Hypothesis

If Zipf's Law approximates the frequency of every word in a language, then the sum of the frequencies will yield the size of the lexicon.

Derivations & Methods

To obtain an accurate fit to the frequency data obtained from corpora, we used the **Zipf-Mandelbrot (ZM)** Law, where β and a are obtained from regression analysis.

$$f_i = \frac{f_1}{(i+\beta)^a} \longrightarrow \sum_{i=\beta}^{n} \frac{f_1}{(i+\beta)^a}$$

ZM Law — ZM Series

Integral Test: For the ZM Series to be used, it must reach 100% of the lexicon size. The ZM Series is in the form of a *p*-series and does not diverge in all cases like the harmonic series. Thus, we tested the viability of all cases of the *p*-series with the Integral Test.

$$\int_1^\infty \frac{f_1}{i^a}\,di$$

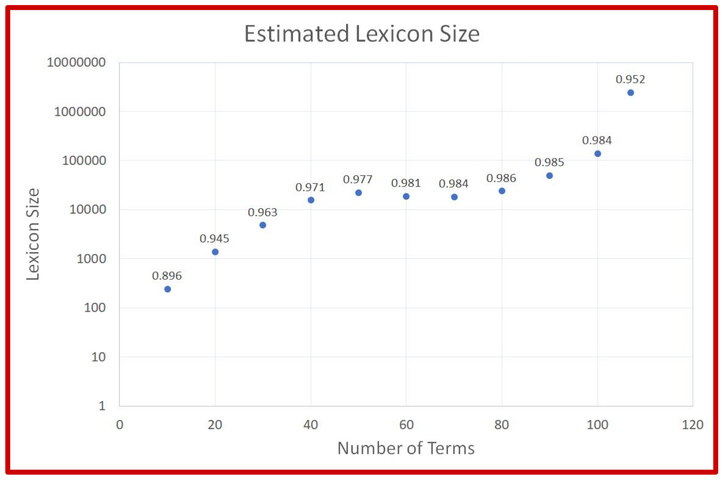
Power Law Regression:

Piecewise Analysis: With a and β ascertained from regression analysis, a piecewise series can be created for any point in time with data from different corpora for any language.

Results

We estimated the size of the lexicon in 1960 by plotting the Zipf-Mandelbrot distribution (a = 1.03, β = 0, R^2 = 0.952) for the Corpus of Historical American English (COHA). We did so by first summing the frequencies of the 107 highest statistically ranked words in COHA, then using the resulting regression to sum terms after rank 107 in Python. In addition, using the purported Zipfian distribution of the Brown corpus, we estimated the size of the lexicon in 1961.

English Lexicon Size Estimates using Zipf's Law	
Brown Corpus	1,108,990
Corpus of Historical American English	2,141,375



Conclusions

- Estimating lexicon size with a mathematical equation is theoretically proven
- Quantity and quality of data can heavily determine results
- New cultural insights can be identified with lexicon size estimates.
- This novel approach has the ability to account for regional colloquialisms that never reach a book or a dictionary.

Future Research

To strengthen the validity of this groundbreaking method, we aim to:

- Utilize the Oxford English Corpus and other corpora for better controlled data
- Verify the validity of this process by applying it to non-English corpora and lexicons

Future aims include:

- Quantify the rate of language change as a differential equation with respect to population size, migration patterns, and cultural and technological change.
- Apply differential equation to the evolution of natural language processing systems.

Sources

1. Davies, Mark. The Corpus of Historical American English (COHA): 400 million words, 1810-2009. (2010-)
2. Fagan, Stephen; Gençay, Ramazan. An introduction to textual econometrics", in Ullah, Aman; Giles, David E. A. (eds.), Handbook of Empirical Economics and Finance, CRC Press, pp. 133–153, ISBN 9781420070361. (2010
3. Michel et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)
4. Francis, W.N. and Kučera, H. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Brown University. Providence, Rhode Island. Compiled in 1964, 1971, 1979.
5. Mandelbrot, Benoît. Information Theory and Psycholinguistics. Reprinted in R.C. Oldfield and J.C. Marchall (ed.).Penguin Books (1968)[1965].
6. Wolfram Research, Inc., Mathematica, Version 12.0, Champaign, IL (2019).
7. Lin et al.. Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Volume 2: Demo Papers (ACL '12) (2012)

Acknowledgements

The project presented could not have been developed without the following individuals:

- Orin Gotchey, Texas Tech University Department of Mathematics
- Hunter Hageman, Texas Tech University Whitacre College of Engineering
- James McCracken, Oxford English Dictionaries
- Dr. Cecily Zacharias, Indian Institute of Technology Madras Department of Mathematics

We would like to thank the TTU Honors College **Undergraduate Research Scholarship (URS)** Program supported by the CH and Helen Jones Foundations as well as the **Center for Transformative Undergraduate Experiences (TrUE).**