CALIFORNIA STATE SCIENCE FAIR
2010 PROJECT SUMMARY

Name(s)

Ishan S. Puri

Project Number

S0320

Project Title

Linguistic Creativity and the Zipfian Distribution: An Entropic, Stylometric, and Computational Analysis

Abstract

Objectives/Goals

For almost five decades, linguists have actively pursued the questions: What is "creativity", how do we think, and where does meaning lie in our communication? Combining fundamental linguistic theory with information and source theory, a novel and critical mathematical relation is derived that sheds light on what creativity is and how it can be measured mathematically. Language representation is important practically in the world of communication (e.g. cell phone industry), and academically (e.g. bioinformatics, computer science,astrophysics). The paper reaches to the core of science: how do we think and how are thoughts translated into writing?

Methods/Materials

After several months of investigation into papers on Zipf's Law, studies of entropy, Huffman encoding, and mathematical linguistics, a relationship between Zipf's Law and entropy was considered. First, using 5 corpora from last year's study (32,574 words each, reviewed eight times independently and finally for statistical significance)I wrote several Python programs to measure Huffman encoding size and entropy to see empirically if such a relationship was possible. After seeing a positive correlation, using Shannon's and Sayood's fundamental theorems I mathematically showed a relationship between entropy and the Zipfian coefficient k. A bigram model was later considered to verify results empirically with a linear regression and a MATLAB model was employed.

Results

Initially, an empirical evaluation was developed via programs in Python, Cygwin, and packages that allowed calculations of the Levenberg-Marquardt linear regression of the compressed sizes of the corpora. A direct relationship was found between compression size through Huffman encoding and the Zipfian coefficient. Then a novel formula was derived linking total entropy of a natural text to the Zifpian coefficient. The results were verified mathematically and graphically with MATLAB and a bigram Python model.

Conclusions/Discussion

Mathematically we have come closer to understanding fundamentally where meaning lies in a test, how we interpret the idea of "creativity", and how a text is "surprising" ((i(A)=surprisal)). These ideas have great implications in all of science and most notably in linguistics, psychology, and sociology. This novel relationship can be used in mainstream linguistic analysis as we have shown with the much-improved bigram model.

Summary Statement

I dervied a novel mathematical formula that brings us closer to understanding linguistic creativity , empirically verified with Python programs and MATLAB, and showed wide implications in practical and theoretical fields.

Help Received

Dr. Stabler, professor of linguistics at UCLA, gave me introductory books and papers to read.