Do Texts in Different Languages Distribute According to Zipf's Law?

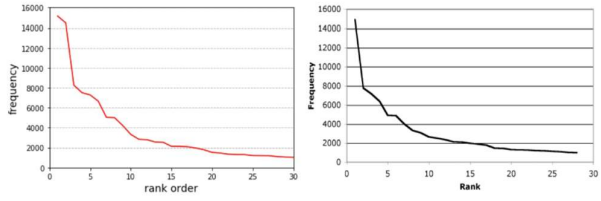
By: Leia Gill and Hadas Yosef-Zada Supervisor: Avi Ben Simon

Introduction:

Zipf's law describes many natural & artificial phenomena that are distributed according to a power law distribution such as- earthquake magnitudes, city sizes and incomes. The phenomena described are phenomena in which 'large events' are very rare and 'small' ones are quite common. Here we show how word frequencies in natural languages act according the Zipf's law (the 'large events' are the most frequently used words in the language - such as 'the' and 'and' in English texts - while the 'small events' are words that are used seldomly, that usually appear only once or a couple of times in a text) and continue on to check whether the law holds in languages other than English that haven't been previously studied.

Research Process:

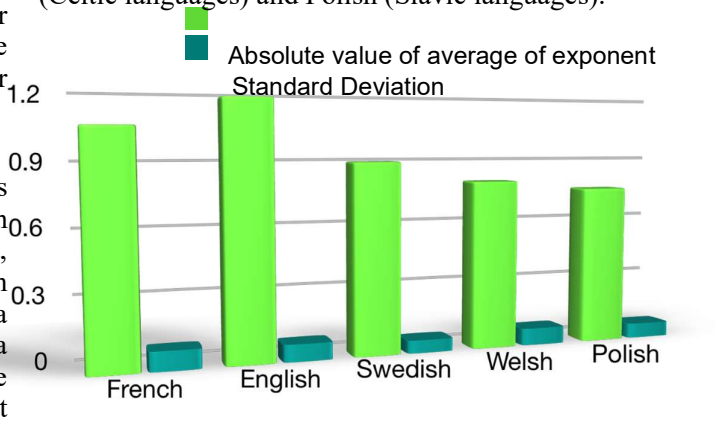
In our research we wanted to examine how Zipf's distribution is expressed in different languages. In order to do that, we first had to write a program that, when given a text, counts the number of times each word appears throughout the text and returns it as a graph in which the vertical axis is the frequency a word (the number of times it appeared throughout the text) and the horizontal axis is the rank (the word that appears the most times is given the first rank, and so on). In Zipf's original research from 1936 "The Psychobiology of Language", he researched the book 'Ulysses' by James Joyce and showed that the rank-frequency graph acts as an exponential decay; meaning, every graph of a text has an exponent, which eventually determines Zipf's distribution. Zipf's distribution occurs when that exponent in absolute value is bigger than 1. In order to make sure our program really puts out a correct graph we ran it on the book "Ulysses" and compared our results to Zipf's original research:



Left: the graph our program created, Right: the graph Zipf created in his research of the book. Since we know every graph has a unique exponent we decided to use our program to find the unique exponents of fifty books in five different languages (each language from a different language family). Then to average the exponents of the books from each language to check if they act according to Zipf's law (the absolute valu-
e of the exponent needs be >1).

Results:

This chart shows the results of our research, the average of the exponents from ten books in the English language (from the West-Germanic family), as well as French (from the Romanic languages), Swedish (from the Scandinavian languages), Welsh (Celtic languages) and Polish (Slavic languages).



The average exponent of the languages ranges from 0.75 to 1.2 with small standard deviations, the smallest being 0.06 in the Swedish language and the highest being 0.09 in the French language.

Conclusions:

Using the program we wrote we were able to find an average exponent for each language we studied. Since for Zipf's distribution to work the exponent has to be bigger than 1, we discovered that some languages don't act according to Zipf's distribution (Swedish, Welsh, Polish). Later we tried to understand, why do some languages do not act according to zips law distribution? We found that Germanic and Romanic such as English and French use more conjunctions than the Scandinavian and Slavic languages. The conjunctions are the most frequently used words in the language hence the exponent is *higher* (high enough to act according to Zipf's law).

Future research:

Future research in this subject could be checking if it can drive a method for identifying languages. If, given a text in an unknown language and finding the

text’s exponent, we can uniquely identify the language in which it was written.