



**CALIFORNIA STATE SCIENCE FAIR
2016 PROJECT SUMMARY**

Name(s) Kathryn C. Forrest	Project Number 36112
Project Title Using Zipf's Law to Analyze Word Complexity through the Past Three Centuries	
Objectives/Goals The objective is to determine whether Zipf's law applies to state of the union (SOU) speeches given throughout U.S. history. Methods/Materials Get the full text of 57 SOU speeches (the 1st speech from every 4-year term) from the UC Santa Barbara archives. Write computer code in R to organize the words into a data frame that includes each word's frequency (the total number of times each word appears) and rank (its position in the list of words ordered by frequency, with the most frequent word first). For example, the word "the" might appear 1,000 times (The frequency is 1,000.) and might be the most common word in a speech. (The rank is 1.) Once this data frame is created, make graphs (using a log10 scale on each axis) to illustrate whether Zipf's law applies to the speech. Results The equation that describes Zipf's law is $y=k/x^a$ (where the exponent "a" is the slope (on the log-log scale), and "k" is the total number of words in the speech). Zipf's law says that a should equal 1, so that the frequency of any given word is inversely proportional to its rank. The law was approximately true for most of the addresses. None of the graphs in this project had a slope of exactly 1, but many came close. The graphs of SOUs with slopes closest to 1 were given between the years 1845 and 1934. On average, these slopes were about 0.9. The steeper the slope is in the graph of a speech, the less complex the speech is. Conclusions/Discussion Zipf's law approximately applied to most of the speeches, although it was not applicable to speeches given before the 1820s. The law reveals how speech complexity has changed over time. It was concluded that human speech has grown less complex since the 1800s.	
Summary Statement Using R code, I proved that Zipf's law (which states that the frequency of a word in a text sample is inversely proportional to its rank) applied to state of the union speeches given since 1790.	
Help Received Father helped in learning R code. I did the programming, analyzing, and assembly of the project myself.	