**Language of our Presidents:**

**Analyzing the First State of the Union Addresses by US Presidents over Time**

Michael Beauchamp

Data 620: Data Visualization

Professor Carrie Beam

August 9, 2021

**Executive Summary**

The presidents of the United States are required by law to provide congress and the American people a report on the current state of affairs for the country, better known as a State of the Union address. These addresses, when turned to text show both similarities and differences in speech between our presidents over time.

In this report, I analyze the first State of the Union addresses by 14 different presidents. While some words such as government, people and state are present throughout all the speeches, we see words such as Russia, terrorism and others appear due to the political climate of the time. While the average number of words for the State of the Unions addresses is 1,440, Teddy Roosevelt far exceeded the average with 3,274 words while George Washington was far below with just 388 words.

Readability and grade level are just some of the analytic tools used by governments and organizations around the world to analyze text. Washington, who had the lowest total word count, had the hardest document to read with a score of 25.80 on the Flesch Reading Ease and a Flesch-Kincaid Grade Level of 18.40, or above graduate grade level. The highest score for reading ease belongs to Biden at 64.40, or very easy to read with a grade level of 8.4, or eighth grade. The trend over time is showing that the reading ease scores are increasing, meaning the speeches are becoming easier to read. Reading levels are also going down indicating that the difficulty of understanding the speech is decreasing.

Highlighting single words makes it difficult to ascertain the vast change in language since 1790, however just reading Washington’s State of the Union you notice the difference in diction and prose. After the analysis, I believe it would be more beneficial to group the words. This analysis would focus on either bi-gram or tri-grams to see if the change of language is more apparent when the speeches are sectioned into small phrases.

**Introduction**

Language can be described as the single most important aspect of human evolution. Our ability to express our thoughts through speech and written word draws a distinct line between humans and other animals. Over time the way we communicate can best be portrayed as fluid, ever changing with the invention of new mediums through which we record our history. From cave drawings to rudimentary paper, from the first printing press to computers and the internet, language changes with time. Interestingly, from a period of 1790 to 2021, which is a mere fragment of our timeline you can clearly see a difference in speech and the written word. Our presidents hold the highest office in our country; however, they too are not exempt from this change. In this paper, I explore the words of our presidents through their first State of the Union addresses and intend to highlight both the similarities and differences in language through an analysis of words and readability.

**Data Gathering, Cleansing and Preparation**

**Data Gathering**

Presidential speeches are easily obtained with a simple internet search. I chose the following State of the Union addresses to include in the word analysis:

|  |  |
| --- | --- |
| 1. George Washington - 1790 | 8. John F. Kennedy - 1961 |
| 2. James Madison - 1814 | 9. Richard Nixon - 1970 |
| 3. Andrew Jackson - 1829 | 10. Ronald Reagan - 1981 |
| 4. Abraham Lincoln - 1861 | 11. William Clinton - 1993 |
| 5. Theordore Roosevelt - 1901 | 12. Barack Obama - 2009 |
| 6. Franklin D. Roosevelt - 1933 | 13. Donald Trump - 2017 |
| 7. Dwight D. Eisenhower - 1953 | 14. Joe Biden - 2021 |

Once I was able to locate the speeches online, I copied the text and pasted it into a new notepad file and saved them as a text file (.txt). After gathering all 14 speeches, it was time to start on the Python code that would split each files’ words into a counted dictionary, remove stop words and eventually convert the remaining words to a single Excel file for Tableau visualization.

**Cleansing**

The Python code I wrote for the analysis first opens the text file for the speech. Once the file is open, the code separates each word and puts those distinct words into a dictionary with the number of times that word occurred in the speech. During this process, we rely on two tools called Lemmatizer and Stemmer to remove common words from the speeches. The lemmatizer works off a downloaded file of common words, called Stopwords which contains words such as “a, the, an” and more which are not needed and can skew the data. The lemmatizer and the stemmer also strip some endings off words to get to their root. For example, the words testing, tests, tester, tested would all become their root word test.

Once the dictionary has been setup and the words counted, I added the president’s name and the year that the State of the Union address occurred to each line. This step is essential to analyzing the data in Tableau. You must have some sort of date field to conduct any kind of analysis over time and using the president’s name created a field similar to a category. Then I called a function which sorted and reversed the dictionary, putting the most common words and their counts at the top. Finally, my code creates a csv file (.csv) for the speech and adds the columns named “Count, Word, President, Year”. The csv file that is created has the word “Final” in its title, which is important for the next step of the code This process must be followed for each of the 14 presidential speeches, changing the name and year in the code to match the file.

**Preparation**

Once I had all 14 files converted to a list with their word counts saved as csv files, I wanted to combine them into an Excel file. To do this, I used pandas and glob to concatenate the files that start with the word “Final” into a single csv file called “CombineSpeeches”. The importance of the world final here is to make sure that if there are other csv files in the directory, those are not concatenated to the file we are creating. The last step is to call a function that uses pandas to write the csv file into an Excel file for Tableau. Once the code is complete, we end up with an Excel file that contains all the words from all 14 speeches with their counts in columns A and B with the president’s name and year of the address in columns C and D.

**Analysis**

**Top Words**

Text

Description automatically generated

Language changes over time are apparent in the word analysis that I conducted using the State of the Unions. The word cloud shown in figure 1 above highlights the top 100 words that occur through all 14 of the speeches analyzed. The largest of the words in the chart are the most common including people, government, congress, states and must. A word cloud is a useful visualization tool as it plots many words in a small but still easy to understand area where something like a bar chart would take up too much space to convey the same data. Also, the word cloud provides a generalized idea as to the major topics of the text in question, and in this case does a fantastic job of illustrating a political theme.

Chart, bar chart

Description automatically generated

Figure 2 above, demonstrates the top 10 words that were included in all the speeches along with the data point indicating the number of occurrences. The word “must” is of particular interest here, as it is not a word that was prevented by the Stopword tool we used nor did I want to exclude it from the analysis by adding it. The word “must” is very powerful, signaling a command or perhaps an obligation to something. Presidents using the word “must” show that they believe what they are speaking of to be important, where outcomes have consequences for those it is directed to. Instead of saying something like “We have to” or “We should”, presidents use “We must” to emphasize the topic should be taken seriously and instill a sense of urgency in the audience.

If you were to guess which words would have been at the top of the list for the State of the Union addresses, it wouldn’t be a surprise to see the words people, government, states and congress. The word people, after all is the focal point of the preamble of the US Constitution, “We the People”, one of the most popular segments of writing in US history.

**Words Over Time**

Seven words that come to mind when considering the State of the Union are America, congress, democracy, government, must, people and states. Most of the words make up the most popular shown above in the discussion of the top words. From the chart below in figure 3, we can assume that if a president is going to use the word government, they may also use the word states in conjunction. This would make sense, especially in the earlier years of the United States when we were consistently adding new states to the Union. The presidents also seem to use the word “must” with the word people. This indicates that presidents may be using the word “must” to implore the people to listen and take action on the topics of their speech.

Some interesting words here are democracy and America which show up sparring through all the speeches. The word America didn’t show up in a speech until Teddy Roosevelt in 1901 but has recently seen more usage by presidents Obama, Trump and Biden. The word people is also on the rise during this period while the words government and states remains low. This suggests that the most current presidents are trying to reach the citizens of the United States with their speeches instead of pandering to congress or other government figures.

Chart, line chart

Description automatically generated

Not only do communication mediums affect our language, world events also dictate the language that is used in the State of the Union addresses. Consider some of the words by Jackson, Lincoln and Teddy Roosevelt who at the time of their speeches were dealing with westward expansion. The word Indian only appeared in those three speeches. Obviously at this time, dealings with Native Americans were ongoing as people moved west. The word work shows up frequently in 1901 and 1933, for both Roosevelts, however Franklin Roosevelt in 1933 used work in more relation with people and the economy. Roosevelt in 1933 was giving his first State of the Union address in the beginning time period of the great depression explaining his use of the term work.

If you were asked which presidents were more likely to use the word Russia, you would guess Eisenhower in 1953 and Trump in 2017. The cold war and the allegations of Russian collusion in the elections both lead to these presidents being the only ones to use the words Russia in the speeches. Similarly, if asked about the word terrorism, one would expect to find that word after speeches given in 2001 when the attack on the World Trade Center occurred. Looking at the data, we find that is the case with terrorism only being mentioned by Obama in 2009, Trump in 2017 and Biden again in 2021. These world events changed the language used by the presidents and help to drive words specifically tied to them.

**Words by President – Readability and Grade Level Score**

Shown in figure 3 below, the average number of words used by the presidents’ averages around 1450 words in their speeches. Teddy Roosevelt used by far the most words of the presidents, coming in with a total over 3,200 with the words “great” being the most frequent with 60 instances. Contrary to Roosevelt, George Washington’s speech contained only 388 words. Washington also had a very low count of words that were repeated, with public being the top repeated word with only 5 occurrences.

Chart, bar chart

Description automatically generated

One way in which we can analyze text is to look at how easy the text is to read and at what grade level the text is written. According to readable.com, the Flesch Reading Ease score gives a score between 1 and 100 where the higher the score the easier a text is to read. The Flesch-Kincaid grade level test was developed by the Navy who was using the Flesch Reading Ease test. Both tests use a mathematical formula that takes into account the number of words, words per sentence and the total syllables used (Flash reading ease, 2021).

I used an online readability tester at readabilityformulas.com to get the Flesch Reading Ease and Flesch-Kincaid grade level for each of the 14 speeches. I copied the results down into a Excel file and added that to Tableau to create the following figure:

Chart, bar chart

Description automatically generated

Over time, there is a drastic change in the readability and the grade level of the speeches by each of the presidents. The test shows that the score for the ease of reading to be rising, with rapid jumps between Eisenhower and Nixon and then again from Reagan to Obama. Rising score in this case indicates that the speeches were easier to read and understand based on the criteria designated by the test. As time goes by the grade level required for understanding the speech has also declined, demonstrating that the readability and grade level required move in unison with eachother. According to the test, speeches given by our first presidents require a much high reading grade than our current presidents. Washington for example, required what would be considered a graduate level on the testing scale where Biden would require a little better than an 8th grade level.

**Conclusion**

The first State of the Union addresses given by these 14 presidents show a dramatic change of language through history. One interesting conclusion to the analysis is that some of the presidents with the lowest word counts had the lowest score for readability and the highest required reading grade. The analysis shows that over time, there are certain key words that presidents use to draw people into what they are trying to convey, and there are some words that are just to interwoven into the politics for them not to be included. After this analysis, I think that it would be beneficial to run the same code but to capture 3-word groupings as to better see how the words have changed over time. These groupings would provide a clearer picture of change than single words.

**References**

Automatic readability checker, a free readability formula consensus calculator. (n.d.). https://readabilityformulas.com/free-readability-formula-tests.php.

*Flesch reading ease and the Flesch Kincaid grade level*. Readable. (2021, July 9). https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/.