In overall the word **people** has highest frequency, as well as with the lowest central tendency, since most of the points are more scattered and further from the regression lane, resulting high standard errors. In the other hand the word **nation** seems to be the word with the less frequency.

More analytically, we can assume that the reason the word people was used that much, is due to the nature of public speeches, since the speaker was addressing to the people. The slope of the regression line is negative and close to zero.

Furthermore Donald Trump, appears that he spent a lot of his speech time dealing with his political opponent Joe Biden and China. The word **Biden** is the next more frequent word. We can notice a slow fall in the usage of the word after its peak on the 8th of September, due to the slight negative slope of the curve. On this date ex-president Trump spent a lot of time attacking his opponent (<https://edition.cnn.com/2020/09/08/politics/donald-trump-north-carolina-rally-fact-check/index.html>). Also, the word **China** was used 206 times in total. According to the relevant graph, we can notice a moderate decrease in its overall usage, with a moderate variance, since the distance of the points are not that away (standard errors). China was accused by Trump, in the majority of his speeches, about Chinese tech companies funnel American citizens personal data (huawei, Tik-Tok), as well as that China was incompetent to prevent the spread of Covid-19 to the rest of the world (<https://www.politico.com/news/2020/09/11/trumps-tiktok-china-412053>, <https://slate.com/news-and-politics/2020/09/trump-woodward-book-panic-coronavirus-china.html> ......)

We can notice also that Trump preferred the usage of words **country** and **America**, compared to **nation**. We can assume that country had the most appearances (country:255, America:202) because it might refer to any country in general, while America and nation are referring to America. The word country appears also to have the lowest standard errors, since it had a stable percentage of usage in all speeches (slope close to zero), as well as nation. On the other hand the word America seems to be overused Henderson speech on 13/09, and its points are more scattered.

Finally, is worth to mention that in the words we have chosen, the wall is the only with positive regression line slope. Also, the ex-President had not used this word until the 8th of September, however it gets more popular especially till the end of our search range which reached its peak at around 0.5%. Wall refers to the wall at the borders between the US and Mexico, which was the signature promise of President Donald Trump's 2016 election campaign.

**Tf-idf**

Tf-idf gives the weight of the word in an overall corpus. Word with high tf-idf, have a high term frequency but low document frequency (the don’t appear a lot in an entire set of corpus). In our case, it can show if a word has meaning or relationship to a specific speech. As we can notice that names such as: **Jason, Bob, Dana, Colby, Mick** appear to have the highest tf-idf scores ranging from 0.007-0.010. This is not unusual since those names have a particular connection to their respective speech location.

In the Bemidji Minnesota, Jason refers to Jason Lewis, former U.S. Representative for Minnesota's 2nd congressional district from 2017 to 2019 and member of the Republican Party.

The name Bob in the Ohio speech is refers to Bob Paduchik who was the Senior Advisor of Trump’s re-election campaign, who was quoted by Trump that they were going to win Ohio. Also, to Bob Latta U.S representative of Ohio and member of Republican party.

In Henderson, the word Colby, refers to the UFC fighter Colby Covington, who is an outspoken supporter of the Republican Party and President Donald Trump, who joined him at the speech, while Dana, to Data White, President of the Ultimate Fighting Championship (UFC), another outspoken supporter.

Furthermore, after checking the original speech text itself for all those names, it appears that there is a pattern in which Trump repeats himself when calling someone’s name:

***We go to again win Ohio. I understand from Bob Paduchik, you know Bob? That we're going to win it by more than we did.***

***Representative Bob Latta. Thank you, Bob. Great job. They like you, Bob. That's very good. Good job, Bob. Thank you very much.***

***they don't fight like Colby. You know who Colby is. You're going to say hello to Colby. They don't fight like Colby……***

***Let's say you'd had a fight and you happen to meet Colby Covington. You say, "What's your name?" And he said, "My name's Colby Covington." And the first time I saw Colby Covington,……..***

***But I'd like to introduce Colby Covington. Great fighter. Great, great fighter. Incredible. He is a great fighter.***

According to (economist) When he needs to plan his next sentence—as everyone must—he often buys time by repeating himself. This reinforces the impression that he is supremely confident and that what he’s saying is self-evident.

Also the word elite has a high score in Freeland, where trump was referring to the crowed as elite

Finally, we can notice the words with not that high score, such as tiktok, timber, cows, milk, hold a moderate score (0.02-0.03). The word tiktok appears to have a particular weight in Fayetteville, probably since one day before the speech (September 18), TikTok filed a lawsuit to Trump (TikTok v. Trump). The words timber, cows and milk show a relative weight in Mosinee, Wisconsin. This can be explained by the fact the Wisconsin is leading dairy producers and is known as "America's Dairyland"; it is particularly famous for its cheese.[10][11] Manufacturing (especially paper products), information technology, cranberries, ginseng,[12] and tourism are also major contributors to the state's economy.

**Zipfs Law**

The plot above shows, the application of Zipf’s law in 10 Trump speeches. The scales are in logarithms with base 10. Furthermore, the black line is the regression line of our plot, while the red line is the theoretical regression line of Zipf’s law. According to Zipf’s law, the frequency of any word is inversely proportional to its rank in the frequency table.

From the graph we can notice that all the speeches follow almost the same trend, since all are concentrated together, especially for the ranks from 10 to 500. This can indicate that almost the same vocabulary, with the same word frequencies was used in all the speeches. However, for the words with lower rank (500-1000), the lines can be distinguished. This means, that different words with greater meaning (word with lower frequencies tend to be words with higher meaning like nouns/verbs), were used in different locations. This is confirmed by the tf-idf graphs above (different words with high tf-idf appear on different locations).

We calculated the slope of our regression line in order to figure out the degree of deviation from Zipf’s theoretical line (red line, slope = -1). In theory a Zipf’s law curve will have a -1 slope, starting from the top left corner to the bottom right. In our case we notice that all the curves have some degree of deviation from this theory, especially for words with high rank. For words with rank 1 to 10, the curve is flatter, with a slope between -1 and zero. A flatter curve can indicate broader vocabulary for those ranks. However, the words in high ranks tend not to be that meaningful, mostly consisting of conjunctions.

In the other hand we can notice that from rank 10 the speeches are almost sticked to their regression line. The regression line of the speeches (black line) has a slope of -1.11 indicating a sharper degree of change. That means the lower the rank of a word (greater x-value), the less term frequency it has, compared to the Zipf’s law theory. So, a sharper slope (less than -1) can indicate a poorer vocabulary, for the words from ranks 10 and after (more meaningful words)