A Corpus Analysis of Presidential Speeches & Remarks

Cassady Shoaff

Montclair State University

A focused corpus analysis on the speeches and remarks of U.S. presidents Biden and Trump. Data collection examined a three-month period within the coronavirus pandemic for each president, March-May 2020 for Trump and January-March 2021 for Biden. Previously compiled presidential speech corpora do not yet include the two most recent presidents, so this project aimed to expand the corpora. The data was collected automatically via Python web scraping and analyzed using the Natural Language Toolkit to investigate keywords and ngrams, which were compared to former presidents Obama and G.W. Bush. The findings indicate Trump uses a speech style different from career politicians and did not use keywords related to the coronavirus pandemic during the time period examined, whereas Biden's word use does reflect a focus on the healthcare crisis.

Keywords: president, speech, Biden, Trump, NLTK

1. Introduction

This project concentrated on the creation of a small, specialized corpus of presidential speeches and remarks from current President Joseph Biden and preceding President Donald Trump. The Corpus of Presidential Speeches by David Brown contains texts for U.S. Presidents from George Washington to Barack Obama, but does not yet contain any transcripts from Trump or Biden, so this project aimed to contribute an updated corpus of recent presidential speech. Additionally, conversational dialogue, e.g. the presidents' responses to reporters' questions, was included in this project, unlike the existing corpus which focuses only on oration. Data collection and analysis were performed automatically via Python scripting using the Natural Language Toolkit.

2. Literature Review

Previous research of U.S. presidential speeches in recent years has centered on 45th President Donald Trump. Several studies utilized the Clinton-Trump Campaign Speech Corpus from the Grammar Lab by linguist David Brown. One such study observed that Trump used metaphors to marginalize disadvantaged groups in the U.S. via creation of an "other" with negative metaphors. "When Trump uses the words *pour* and *flow,* it usually correlates to people, refugees, immigrants, and, according to Trump, the subsequent crime and drugs they bring with them" (Stifton, 2020). Another corpus-based study comparing Clinton and Trump's linguistic style using the same campaign speech texts examined the keywords favored by each politician.

> Clinton stuck to her core messages in her speeches and appeared to address voters' concerns by constantly using such keywords as my, young, women, help, can, kids, everyone, etc. With these words, she was trying to find commonalities with ordinary people by showing a sense of empathy. In contrast, Trump was concentrated on launching attacks on his opponents (e.g., Hillary, Clinton, bad, she), or criticizing the policies of the current administration (e.g., Obama, Mexico, border, etc.) (Chen et. al. 2019).

Again using the same Clinton-Trump campaign corpus, an investigation with machine analysis (word2vec) examined the theme and sentiment in the Clinton-Trump campaign speeches, revealing that Clinton's strategy utilized conventional appeals to inclusiveness and reason, whereas Trump used repetition and appealed to negative sentiments (Liu 2018).

Donald Trump's 2015-2016 campaign speeches have also been evaluated for readability or complexity, finding that Trump's language requires a 4th/5th-grade reading level (age 9-11) whereas all other political candidates examined had a 9th-grade reading level (age 14-15), which "suggests that Trump uses low readability and simplicity of language as a rhetorical strategy to gain popularity, in accordance with the trend of anti-intellectualism" (Kayam 2017).

In an additional contrastive analysis, six speeches from analogous contexts by Barack Obama and Donald Trump were compared for number of words and length of sentences, morphological composition, and use of pronouns. The analysis found Trump used more words but spoke shorter sentences; additionally, Trump's discourse seems to be less formal than Obama from the use of deictic words, and he used more first-person singular pronouns than Obama who used "we", "our", and "us" more frequently. The study concludes "the discourse of President Trump is more narcissist and populist" whereas "the discourse of President Barack Obama seems to be delivered with a humble tone and a higher degree of formality" (Casaño-Pitarch, 2018).

However according to 2015 psychological research, usage of first-person singular pronouns or “I-talk” was not significantly correlated to narcissistic traits.

> There is a widely assumed association between I-talk and narcissism among both laypersons and scientists despite the fact that the empirical support for this relation is surprisingly sparse and generally inconsistent. (Carey 2015)

An examination of linguistic markers of narcissism applied the Linguistic Inquiry and Word Count, finding again that first person singular pronouns was unrelated to narcissism (Holtzman 2019). Based on the evidence of these two studies, researchers should be cautious to infer that certain language use indicates narcissistic personality traits because the connection remains unclear. In the literature reviewed above, it can be observed that Trump’s language use in campaign speeches exhibits negative metaphors to create “otherness,” criticism of political opponents, and appeals to negative sentiments. This overall negativity contrasts with the language use of other more typical politicians.

3. Methodology

Following the format of the Corpus of Presidential Speeches (CoPS) by David Brown of the Grammar Lab, which contains the speeches of 44 U.S. Presidents, from George Washington to Barack Obama, I collected speeches and remarks for the 45th and 46th U.S. Presidents. For Joe Biden, the transcripts from the White House Briefing Room were dated from the January 2021 inauguration speech until the end of March 2021. The Donald Trump texts were taken from the Trump White House Archive from the three-month period of March to May 2020. This date range was chosen because the WHO declared a global pandemic and U.S. coronavirus lockdowns began during this period, so it was this author’s assumption that the pandemic would be a topic during those three months, whereas Biden’s texts were from the beginning of the vaccination process, following the height of the hospitalizations. The three-month timeframe was chosen to approximate the same sample size as the data from the Biden presidency (all that was available at the time of this writing). These texts included both standalone oratory speeches as well as remarks made as part of interviews or discussion panels (as opposed to CoPS which used only oratory speeches).

I used a webscraper to automatically collect the data, accomplished via a Python script I wrote for this purpose. Using the BeautifulSoup HTML parser, the script traversed the White House Briefing Room or Trump White House Archive websites to accumulate a list of URLs to each page that contained presidential speeches and remarks, while excluding speeches by the Vice President or other officials by filtering the title section. Then the second stage of the script visited each URL and extracted the raw text from the page. (I chose to divide the script into multiple parts due to Internet connectivity issues caused by working remotely, but in ideal circumstances they could be combined into one file.) Small modifications to the script were made due to differences in the HTML structure of Biden's White House Briefing Room webpages versus the Trump White House Archive, resulting in two versions of the program. Because websites vary widely in design and internal organization, anyone undertaking similar data collection would likely also need to make customizations to this script to adapt to the particular website of interest.

Each page visited by the script was saved into individual text files, so each transcript can be viewed singly. It also inserted date and title header tags after the fashion of CoPS. Because the Trump transcripts cover only a partial period of Trump's presidency, and Biden's presidency is ongoing, this data is not intended to be as comprehensive as CoPS; if this dataset were to be contributed to CoPS in the future, further removal of non-oratory transcripts and addition of the totality of speeches from Trump's presidency would need to be added in order to complete the corpus. However, the non-oratory presidential remarks represent an additional source for analysis, which is especially needed due to small timeframe of Biden's presidency, so they have been included for this project.

The third stage of the script cleaned the raw text data to remove irrelevant parts of the transcripts such as parenthetical comments e.g. (*The document is signed.*) or (*Laughs.*), dialogue tags/name labels, header and footer information, and non-presidential persons' speech, such as audience members, reporters, foreign dignitaries, and White House officials. Again, due to differences in the two archives, such as inconsistency in the formatting used by transcribers and non-Unicode special characters ("fancy" quotation marks) in the Trump Archive website, a separate version of the cleaning script was needed for Biden and Trump's texts. These represent two strategies for the preprocessing stage. In Biden's version, the cleaning script matches non-presidential speech and removes it, utilizing Python's Regular Expressions (regex). It detects the interlocutor's dialogue tags and deletes everything from the tag until Biden's next line. The same script was also used to preprocess the Obama and G.W. Bush text from the Corpus of Presidential Speeches, only to remove the title and date header tags.

In the Trump version, the transcripts' inconsistent formatting necessitated another method. An opposite approach was used: this version finds matches of the president's dialogue tag and captures everything between that tag and the next ALLCAPS dialogue tag. It also implements the cleaning phase on each individual file due to formatting differences, whereas the original loaded all raw text files into a single string first before running the pattern match; this is likely more inefficient since it runs the same regex multiple times, but required due to the unfortunately inconsistent formatting choices of the transcribers.

The cleaned text was then saved as one long string into a single text file for each president. The original transcripts remain available separately. As mentioned previously, due to remote work connectivity issues, the data collection and preprocessing were divided into smaller stages, but with better conditions it would be possible to combine all of these steps into one program. Additionally, some of the cleaning steps may be better suited to incorporate at the same time as data collection; for example, the text files of each speech may retain unwanted title header information. However, I was wary of overprocessing in case it proved necessary to go back and examine the inter-dialogue contexts, so this information was left intentionally. I have included many (perhaps too many) details about this text collection and cleaning process simply because the two datasets differed so drastically that two approaches were required. It is easy to underestimate the data collection and preprocessing stages. While doing this project I gained enormous respect for everyone working to collect new corpora.

4. Analysis

The Natural Language Toolkit (NLTK) Python package was used to analyze the data. First the cleaned text is tokenized and stop words removed automatically using NLTK's word tokenization tool. The default stop words list from the NLTK English corpus was used with small additions such as personal titles (e.g. "Mr." and "Mrs.") and contractions. Then the frequency of the tokens could be analyzed using NLTK's FreqDist() class which measures the amount of times each token appeared in the text. The script also outputs the readability score, for complexity of the vocabulary, as a grade level according to the Coleman-Liau formula (the code snippet which I had written previous to this project).

Coleman-Liau	0.0588 * average letters per 100 words - 0.296 * average sentences per 100 words - 15.8

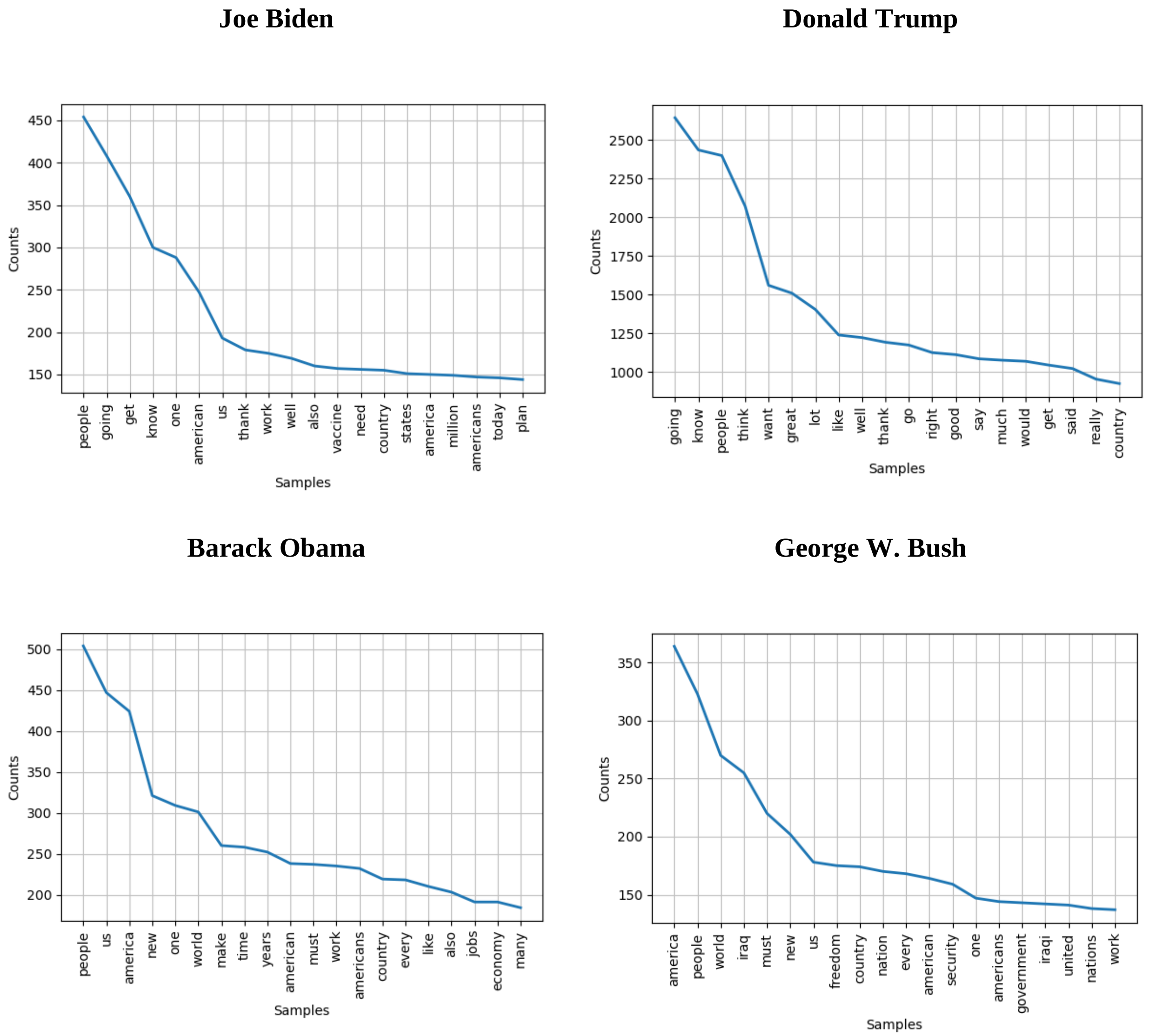
Visualizations of the data were created using Python packages Matplotlib and Seaborn. Unlike the preprocessing, this script did not require different versions for each dataset; one can simply change the "president" variable at the top to match the last name of the corpus to be examined, and the script will open the appropriate cleaned text file. (Small exception: George W. Bush is "gwbush" in order to differentiate from his father.) Tokens and types are shown in Figure 1. It is worth noting that Trump's text contained significantly more tokens (but a comparable number of types) than the other presidents. In contrast to the three career politicians, Trump talks more.

President	**Biden**	**Trump**	**Obama**	**G.W. Bush**
Tokens	35,205	157,412	57,349	37,578
Types	5,432	8,796	8,028	6,226
TTR	15.4%	5.8%	13.9%	16.6%
Readability	Grade 5	Grade 5	Grade 9	Grade 10

Figure 1. Type-Token Ratio and Readability scores

The readability levels of Obama and G.W. Bush (from CoPS) were comparable to previous research (Kayam 2017) that showed typical politicians' speech was around the 9th grade reading level. Trump's readability level at Grade 5 was also in agreement with previous findings.

However, Biden’s readability at Grade 5 was surprising, since he is a career politician (as opposed to Trump’s business and entertainment sector origins), it would have been expected that the readability level would be more similar to other career politicians such as Obama or Bush. Perhaps “the trend of anti-intellectualism" mentioned by Kayam influenced Biden to choose a simpler style as part of his strategy to appeal to a wider audience. Alternatively, since the new corpus collected for this research was not exclusively standalone speeches, but included dialogue between the presidents and reporters or other policy makers, this readability score may simply reflect the less-formal vocabulary used in such conversation as opposed to oration.



Joe Biden

Donald Trump

Barack Obama

George W. Bush

Figure 2. Top 20 Keywords

Top 20 keywords (shown in Figure 2) used in the speeches and remarks were extracted for Donald Trump and Joe Biden using the cleaned text data as well as keywords for Barack Obama and George W. Bush using the Corpus of Presidential Speeches. All four presidents had “people” in their top 3 most commonly used words. “American” was common to Biden, Obama, and Bush, but not Trump, and “country” also appeared in the keywords of all four politicians. Some keywords were relevant to the specific time period, e.g. “vaccine” was Biden’s 12th most used word, and “Iraq” was Bush’s 4th keyword. Although the Trump sample texts were also from the time of the pandemic, none of his keywords related to the pandemic. He had on average shorter keywords than the other three presidents and favored more monosyllabic words. He also used many superlatives, “great,” “[a] lot,” “much,” and “really” appearing in the top 20 keywords. This encourages the notion that his speech style is more informal than the career politicians.

Top 20 bigrams and trigrams for Biden and Trump were generated using Seaborn visualization package, shown in Figure 4 and 5. Many of Trump’s most frequently used ngrams were actually filler words or polite phrases such as “please_go_ahead” and “thank_much” shown in Figure 3. Observing the large amount of filler words helps to illustrate Trump’s speech style. Nine out of Trump’s 20 top trigrams were these polite filler phrases. However, this makes for a difficult comparison to Biden’s top keywords, so a small addition to the stop words list was made for Trump’s ngram section, removing additional tokens "ahead", "well", "thank", "much", "yeah", "please" in order to generate a more insightful set of keywords. The other three presidents did not require any extra preprocessing to remove filler from their top 20 keywords. This extra stop word removal revealed certain well-known Trump catchphrases such as “never_seen_anything” and “done_incredible_job” as well as priorities such as “small_business.”

Although many of the transcripts were collected from the beginning of the coronavirus during March-May 2020, directly pandemic-related vocabulary did not appear in Trump’s top 20 bigrams or trigrams even after additional preprocessing. The closest keywords would be “new_york_jersey” which may refer to the high number of cases in that region. It is possible that choosing a different three-month period during the 2020 pandemic would have resulted in more pandemic-specific vocabulary, but at least during the months of March through May, Trump’s word use does not clearly reflect a focus on the coronavirus crisis.

Biden's top 20 bigrams included vocabulary specifically pertinent to the pandemic. "Get_vaccinated" and "million_shots" were the 14th and 12th most common bigrams to appear in the text. Many health-related terms appear in the trigrams, such as "community_health_centers," "enough_vaccine_supply," "health_human_services," "retired_doctors_nurses," and "least_one_shot." His trigrams also reflect his Catholic religious beliefs with phrases such as "may_god_bless" and "may_god_protect." Overwhelmingly, the most common key trigram was "American_rescue_plan" which shows Biden concentrated on this particular piece of policy. Overall, with the majority of his key phrases were healthcare related. Biden had a very clear focus on the health crises that was occurring during the first three months of his presidency.

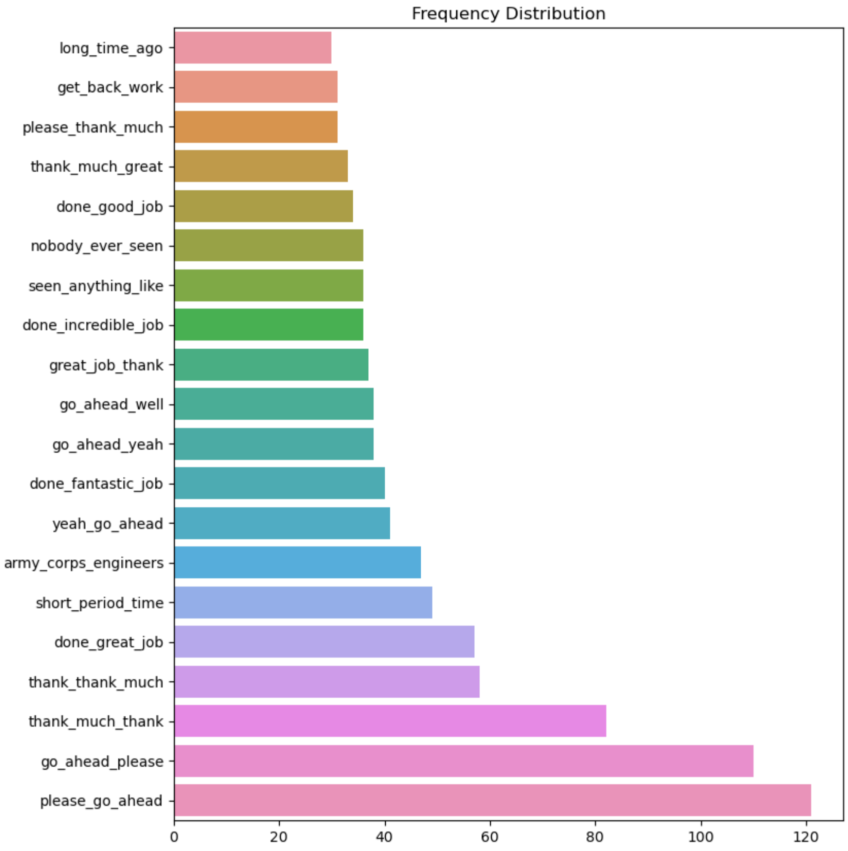


Figure 3. Trump trigrams before removing additional stop words

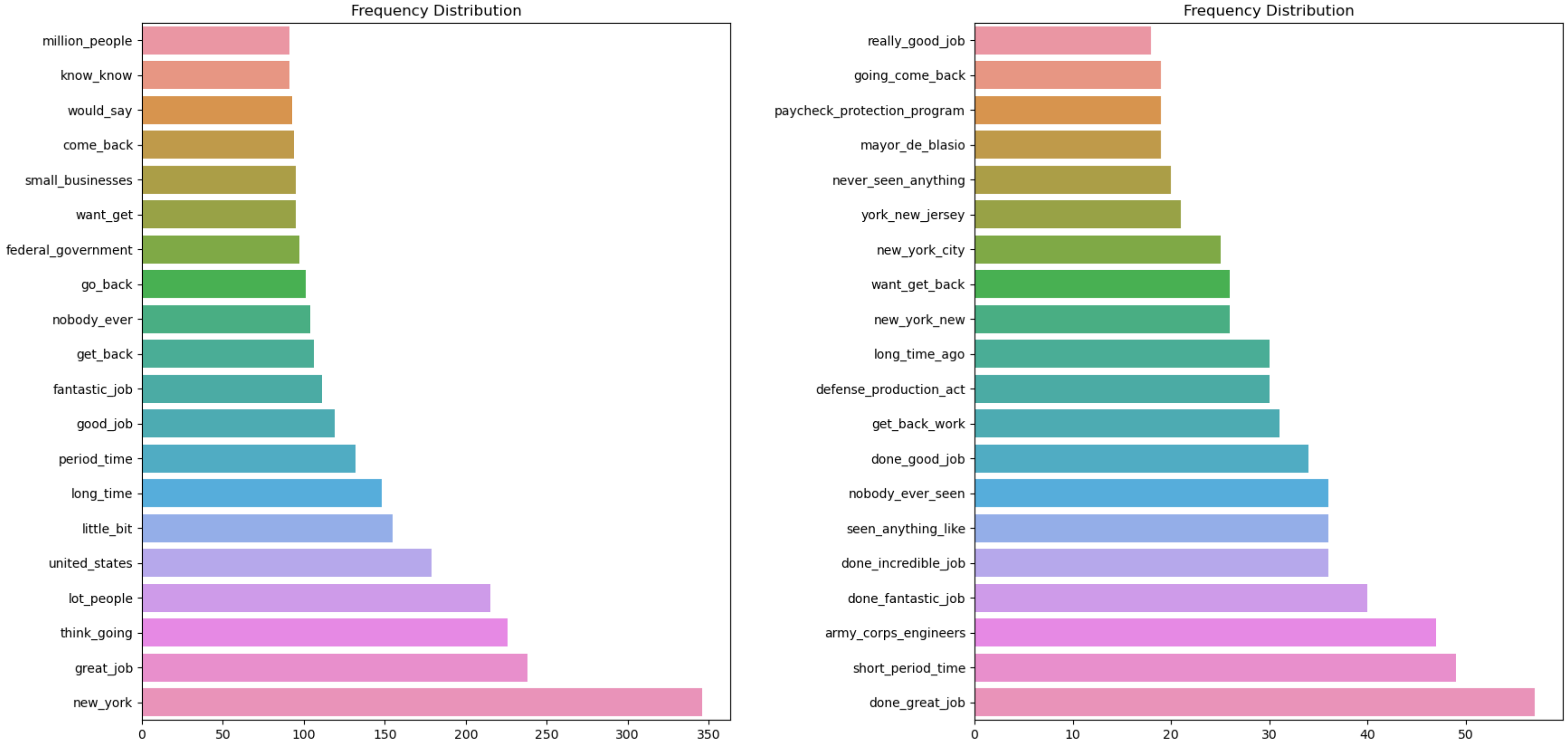


Figure 4. Trump ngrams

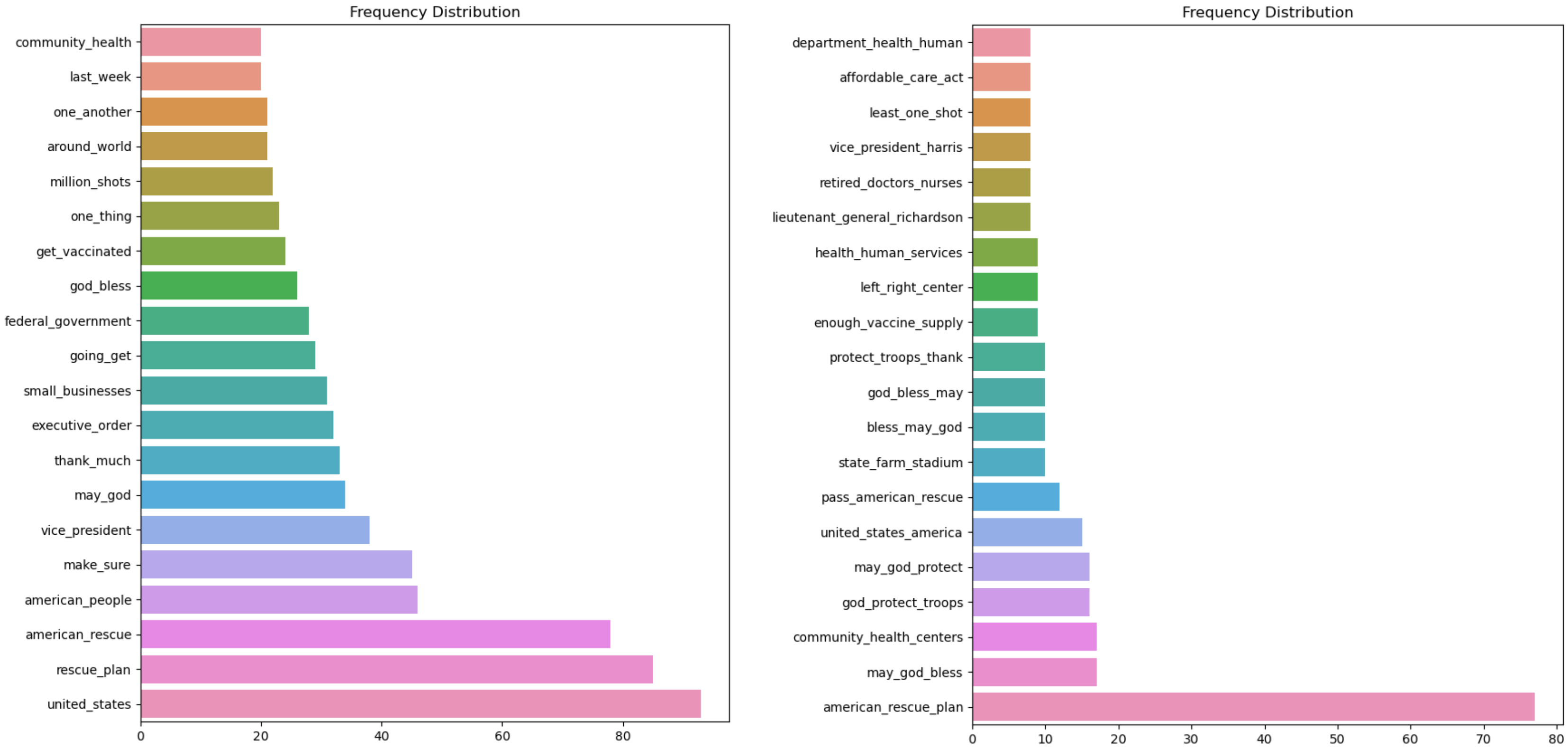


Figure 5. Biden ngrams

5. Conclusion

Both Biden and Trump favored simpler vocabulary, which could reflect intentional political strategy to appeal to a wide voter base, or could be a result of including conversations in the corpus, since the Corpus of Presidential Speeches includes only oration whereas the new corpus data collected for this project included dialogue. By number of tokens used, Biden was similar to fellow career politicians Obama and G.W. Bush, whereas Trump's high token count, despite both Biden and Trump's data being collected from only a three-month time period, indicates he talks much more than other presidents. Trump favored short monosyllabic words and superlatives, and his most common bigrams and trigrams before additional preprocessing were mainly filler phrases, giving the impression of a preference for an everyday style of speech. Trump's keywords do not reflect a focus on the coronavirus, at least during the March-May 2020 time period that these texts were collected from. In comparison, Biden's top keywords from his first three months of presidency did include "vaccine" and his bigrams and trigrams were overwhelmingly healthcare focused. This creates a stark contrast between Donald Trump and Joe Biden.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* https://www.nltk.org/

Casaño-Pitarch, R. (2018). Mr. President, discourse matters: a contrastive analysis of Donald Trump and Barack Obama's discourse. *RUDN Journal of Language Studies Semiotics and Semantics*, 9(1), 173-185. http://dx.doi.org/10.22363/2313-2299-2018-9-1-173-185

Carey, A. L., Brucks, M. S., Küfner, A. C. P., Holtzman, N. S., große Deters, F., Back, M. D., Donnellan, M. B., Pennebaker, J. W., & Mehl, M. R. (2015). Narcissism and the use of personal pronouns revisited. Journal of Personality and Social Psychology, 109(3), e1–e15. https://doi.org/10.1037/pspp0000029

Chen, X., Yuanle Y., & Hu J. (2019). A Corpus-Based Study of Hillary Clinton's and Donald Trump's Linguistic Styles. *International Journal of English Linguistics*, 9(3), 13-22. http://dx.doi.org/10.5539/ijel.v9n3p13

Brown, D. W. (2016). *Corpus of Presidential Speeches.* http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic Markers of Grandiose Narcissism: A LIWC Analysis of 15 Samples. Journal of Language and Social Psychology, 38(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

Kayam, O. (2018). The Readability and Simplicity of Donald Trump's Language. Political Studies Review, 16(1), 73–88. https://doi.org/10.1177/1478929917706844

Liu, D., Lei, L. (2018). The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media*, 25, 143-152. https://doi.org/10.1016/j.dcm.2018.05.001

Stifton, S. (2020). A Corpus-Based Analysis of the Pervasive and Effective Metaphor Use in Donald Trump’s 2016 Presidential Campaign. *Satura*, 2, 2-12. https://doi.org/10.17879/satura-2019-3061