



6-24-2019

Analyzing a Corpus of President Trump's Public Statements to Reveal Possible Media Bias

Kevin Zalewski
Brigham Young University

William Eggington
Brigham Young University

Follow this and additional works at: <https://scholarsarchive.byu.edu/jur>



Part of the [Political Science Commons](#)

Recommended Citation

Zalewski, Kevin and Eggington, William (2019) "Analyzing a Corpus of President Trump's Public Statements to Reveal Possible Media Bias," *Journal of Undergraduate Research*: Vol. 2019: Iss. 2019, Article 28.

Available at: <https://scholarsarchive.byu.edu/jur/vol2019/iss2019/28>

This ORCA is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in *Journal of Undergraduate Research* by an authorized editor of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

Analyzing a Corpus of President Trump's Public Statements to Reveal Possible Media Bias

JUNE 24, 2019 BY [ADMIN](#)

Kevin Zalewski, William Eggington

Introduction

In recent years, America has become extremely politically divided. As political polarization has increased, so has distrust of the media, especially during President Trump's current term of office. The Media Insights Project reports that "just 17 percent of Americans give the news media high marks for being 'very accurate.'" In this partisan political landscape, it can be difficult to know where to turn for unbiased, unspun news coverage. Where can a person learn what the president has said on a given topic without some type of filter coloring the information? This possible bias is a problem that corpus linguistics can address. We demonstrate that a corpus of the president's public statements can reveal degrees of media bias, or non-bias, thus allowing the truth-seeking political commentator and the general public to test media claims using a bias free, data-driven research tool.

Methodology

The intention in assembling the corpus was to gather all of President Trump's public discourse since his inauguration, including speeches, press conference transcripts, and tweets and to prepare the texts so that a concordancer could analyze them accurately. Speeches and press conference transcripts were pulled exclusively from www.whitehouse.gov/briefings-statements/. Tweets were collected from the Trump Twitter Archive. The presidential corpus as discussed in this paper covers from January 2016 to April 2018.

All texts were scrubbed of extralinguistic data (for example, crowd noises: cheering, applauding, booing etc.) as well as any speech from sources other than President Trump. These other sources include reporters, foreign dignitaries, etc. as well as retweets on Twitter which represent the President merely reposting someone else's digital speech). The body of text composing the corpus was tagged such that the concordancer would ignore anything that wasn't President Trump's direct speech. This tagging was done largely through a custom-built automated text tagger that used the unique formatting of whitehouse.gov transcripts to recognize when speakers other than the President began talking. The transcripts used full caps and a colon to introduce new speakers (for example, "PRESIDENT TRUMP:" or "PRESIDENT MOON:". The tagger would then, using angle brackets, ignore this introductory name and, if the name were anything other "PRESIDENT TRUMP" or "THE PRESIDENT," all text after the name until the next name.

The process of scrubbing President Trump's tweets was much simpler because the Trump Twitter archive allows the tweets to be downloaded en mass with all the requisite information automatically removed.

The completed corpus contains 984,351 word tokens and 14,136 unique word types. Once the corpus was created, generalized media claims about the President's speech from the timeframe were gathered and tested against the corpus.

Results

Using the presidential corpus we had created, we were able to test several media claims, including “Trump vows to make Americans say ‘Merry Christmas’ again, over and over” (The Washington Post) and “The president suddenly has lost the urge to discuss the stock market” (Marketwatch). By applying the corpus and corpus linguistic methodology to these claims we reveal shades of truth and untruth in each case.

By inserting “Merry Christmas” into the corpus, we find many examples which initially appear to corroborate the substance of The Washington Post’s claim. For example, President Trump has said, “We’re saying Merry Christmas again” and “We can say Merry Christmas again.” There are at least seven occurrences of President Trump promising or celebrating the return of the holiday greeting, “Merry Christmas.” However, not one of these occurrences uses the word make as an active verb, and make does not appear as a collocate of these results. This creates a misleading statement. Make creates a potential ambiguity meaning both “to cause” and “to force.” President Trump has on many occasions promised to cause Americans to say “Merry Christmas” and we can easily verify that with the presidential corpus. He has never promised to force Americans to say it and the corpus shows no evidence of such. The claim by The Washington Post that “Trump vows to make Americans say ‘Merry Christmas’ again, over and over” is accurate in substance but misleading in form.

On March 24, 2018, Steve Goldstein writing for marketwatch.com claimed that “The president suddenly has lost the urge to discuss the stock market.” The last recorded tweet by President Trump before March 24, 2018 using the term stock market was on February 7, 2018. The term comes up 45 times before this point in President Trump’s tweets alone, so it’s inarguably fair to say that the drop of the term is striking. The article touches on an event the day before in which a reporter asked President Trump about the stock market. However, Goldstein made sure to point out that the comments were prompted by a reporter, suggesting that the President is unwilling to discuss the subject of his own accord. The corpus results reveal that while Goldstein discloses the interview in which President Trump was prompted to discuss the stock market, he neglected to mention a March 20th speech in which the president—unprompted said, “The economy is maybe the best it’s ever been ever, I think ever. How has it ever been better? The stock market is at an all-time high, jobs are the number one, 154 million jobs.” It would seem from this, that the president may not have been as unwilling to discuss the stock market as Goldstein suggests.

Discussion

These examples demonstrate the apolitical nature of corpus linguistics; the corpus does not provide an opinion, only examples of the president’s speech. It is up to the user to interpret the results. It’s possible, even likely, that neither article intended to deceive its readers. It is likely just as difficult for news people to keep tabs on the bulk of the president’s speech as it is for the general public. A publicly available and regularly updated corpus could help reporters and other media representatives to self-regulate, ensuring that their journalistic integrity is above reproach without consuming untold hours of presidential discourse. Simultaneously, such a resource would allow news consumers to hold media representatives accountable for the claims they make about generalized presidential discourse.

Conclusion

In two cases, we have shown how various techniques can be applied to a corpus of presidential discourse in order to test the veracity of media claims. Such a corpus, publicly available and rigorously maintained, could provide an additional layer of accountability for our news media and provide a source of information that is entirely devoid of spin or bias. Such a corpus could provide an extra layer of defense against fraudulent fake news sources and also give honest reporters an additional tool to ensure that their claims are above reproach.

FILED UNDER: [COLLEGE OF FAMILY, HOME, AND SOCIAL SCIENCES](#), [ORCA-2018](#), [POLITICAL SCIENCE](#)