UC Berkeley

How Do Presidents Speak?

A text analysis of State of the Union Addresses since 1961

Bjørn August Skeel-Gjørling
12-11-2017

Table of Content

1. Introduction 2

2. Data 3

3. Data Analyses 4

3.1. Association between growth rate and speech sentiment 4

3.2. Decreasing language difficulty over time 5

3.3. No egocentric trend 6

4. Conclusion 6

5. Appendix 7

5.1. Links to code and dataset 7

5.2. Figure 1: Sentiment and growth rate 7

5.3. Figure 2: Language complexity (LIX) 8

5.4. Figure 3: Language complexity (Proportion of unique words) 8

5.5. Figure 4: Frequency of egocentric words 9

6. Bibliography 10

1. Introduction

Political speeches are extremely important in a democracy as they are one of the main ways politicians communicate with the population. This makes speeches important both during political campaigns and when politicians are explaining policies. One very popular topic for politicians to talk about is the economy. Are we creating jobs, how is our international trade balance doing and are we experiencing enough economic growth? Thus, economics has been a central focus within the discipline of political science. The literature has shown a clear relationship between economics and voting behavior (Downs, 1957; Duch, 2007). "Economic Voting" is used as a term to describe that if the economy is on an upward going trend voters will have a larger tendency to vote for the incumbent. This is true both on a macroeconomic level and on an individual economic level, also called pocket-book voting (Lewis-Beck & Stegmaier, 2013). Unfortunately, no scholars have, to my knowledge, tried to understand how the economic situation affects political speeches. In other words, can we talk about "Economic Speaking"? This is exactly what this paper is trying to do by posing the following research question:

To what extent are presidential speeches affected by the macroeconomic situation of the United States?

To answer this question, I have lined up three hypotheses (see below). The first hypothesis directly addresses the research question, whereas the other two test potential time trends.

1. American political speeches are more positive when the US economy is on an upward going trend.
2. Language used in American political speeches has become simpler over time.
3. American political speeches have become more egocentric over time.

The paper will be structured in three sections. First, I explain what data I'm using and how the data was collected. Second, I will describe the methodology, results and potential biases of each data analysis used to test the three hypotheses. Lastly, I will conclude and describe the implications of my results.

2. Data

As data, I chose all State of the Union Addresses by American Presidents since 1961. All written addresses (speeches that wasn't delivered orally) was excluded. The reason is that written addresses doesn't fit the definition of a political speech, but should rather be defined as a rapport, and because I found significant language differences between oral and written addresses. The year 1961 was selected as the cut-off point because the addresses from before 1961 was primarily written and because the world bank only has historical GDP rates from 1961 and onward. The presence of GDP-data is important in the data analysis of the first hypothesis.

The speeches were collected from the American Presidency Project which is a research project by Gerhard Peters and John T. Woolley from UCSB (Peters & Woolley, 2017). The research project makes all State of the Union Addresses available on their website. In practice, the speeches were collected using two scraper-functions. The first function collected links to all webpages with State of the Union Addresses. The second function collected the following information from each speech: 1) Speaker, 2) Year, 3) Date and 4) Text content. All web scraping was conducted using the python package "BeautifulSoup". The final dataset included 55 speeches from 1961 to 2016. 2017 is not included as we don't know the GDP-rate for the current year yet.

3. Data Analyses

This section describes the methodology and results of each of the three data analyses corresponding to the three hypotheses.

3.1. Association between growth rate and speech sentiment

To test the first hypothesis, I need both a measure of the speech sentiment and a measure of the macroeconomic trend in each year. As dependent variable, I use the average sentiment value of all sentences in the speech. The sentiment values are computed using a python package called Textblob. Each speech could receive a sentiment value ranging from +1 to -1, where +1 is equal to a perfectly positive speech and -1 is equal to a perfectly negative speech. As independent variable, I use the U.S. GDP growth rate in that given year. Linear regression is used as the statistical method to analyze the association.

The statistical result corresponds to the hypothesis. I find a positive association between growth rate and sentiment values, so high growth rates are associated with positive speeches. The association is statistically significant with a p-value of 0.046[1]. A good example of this association is Barack Obama's speech in 2009. The speech was carried out during one of the worst economic crisis in decades with a growth rate of -2.78 %. This is clearly seen in the speech as well with a sentiment value of 0.07, which is more than one standard deviation below the average of 0.1. See figure 1 in appendix for a scatterplot of the association.

A critique of this analysis is that Textblob can be a rather imprecise method for sentiment analysis. Two sentences from the Obama 2009 speech exemplify this as both "*It's the worry you wake up with and the source of sleepless nights*" and "*The answers to our problems don't lie beyond our reach*" receive a neutral sentiment score of 0.0 even though the first is clearly negative and the second clearly positive. Fortunately, measurement error on the dependent

[1] The association has a correlation coefficient of 0.3, which indicates a weak positive association.

variable only leads to prediction inefficiency but not bias in the estimates of the linear regression model (King, Keohane & Verba, 1994). Therefore, we should still trust the association between growth rates and speech sentiment.

3.2. Decreasing language difficulty over time

To measure how simple the language is in a speech, I use the so called LIX-score. The LIX is used to measure how difficult a text is to understand and read (Björnson, 1971). LIX uses the following equation:

$$Lix = \frac{Number\ of\ words}{Number\ of\ sentences} + \frac{Number\ of\ words\ over\ 6\ letters * 100}{Number\ of\ words}$$

Once again, the results correspond to the posed hypothesis. I find that since 1961 the language used in State of the Union Addresses has become much simpler. This is clear as John F. Kennedy's speech in 1961 has a LIX of 50, whereas Obama's speech in 2016 has a LIX-value of only 36. The decreasing time trend is statistical significant with a p-value of 0.00001[2]. See figure 2 in appendix for a scatterplot of the association. To test the robustness of the result, I run the same test with an alternative measure of language complexity. The alternative measure is the number of unique words as a proportion of the total number of words. Again, I find a negative and statistically significant trend with a p-value of 0.01. In other words, more recent speeches use fewer unique words relative to older speeches.

A critique of the LIX measure is that it doesn't take word combinations or context into consideration, but only focuses on word and sentence length when measuring language complexity. An implication of this is, that LIX won't be able to detect increased complexity due to political topics that are more difficult to understand. This is a weakness of the study.

[2] The association has a correlation coefficient of -0.57, which indicates a clear negative association.

3.3. No egocentric trend

To investigate the presence of egocentrism in the State of the Union Addresses, I use the word frequency of egocentric words. More precisely, I calculate the normalized proportion of the words “I”, “me”, “my”, “myself” in each speech.

The statistical analysis finds no association over time. In other words, the frequency of egocentric words has not changes noticeable since 1961. The linear regression has a p-value of 0.95 which is not even close to statistically significant. Thus, I falsify my third hypothesis that American political speeches have become more egocentric over time.

One reason behind the lacking trend could be the selected sample of political speeches. The main purpose of State of the Union Addresses is to describe the countries situation. This means that the president himself often play a minor role in these addresses.

4. Conclusion

In conclusion, the macroeconomic situation of the United States affects presidential speeches. I find that higher growth rates are associated with a more positive State of the Union Address in that particularly year. Thus, we can talk about presence of “Economic Speaking”. This study is a most-likely study because the purpose of the State of Union Addresses is to describe the economic and political situation of the country. Future studies should therefore investigate if “Economic Speaking” is also present when we look at a broader variety of political speeches. Further, I find a clear time trend showing that the language in presidential speeches has become simpler since 1961. This has positive democratic implications as simpler language increases the voters’ ability to both understand and evaluate the content of political speeches. This should to some extent increase the voters’ ability to hold the President accountable for the state of the union.

5. Appendix

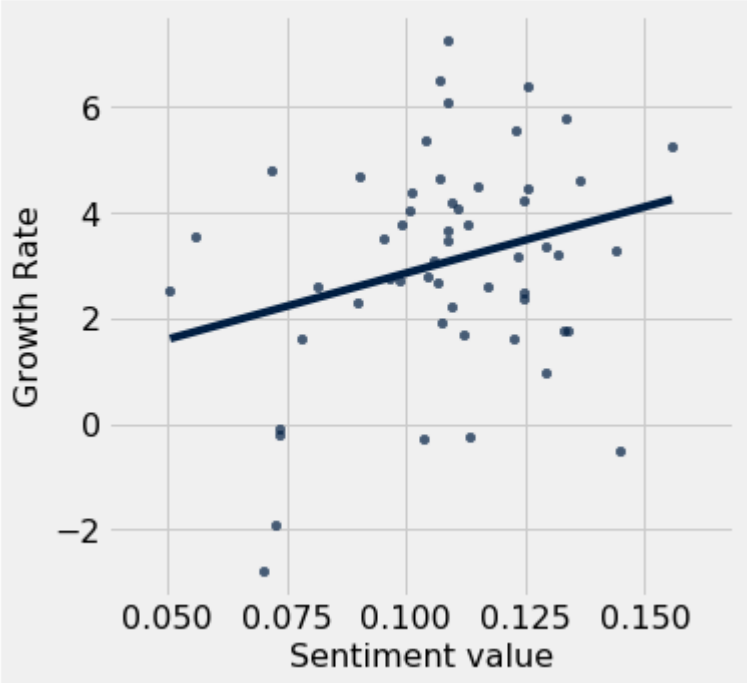
5.1. Links to code and dataset

The dataset is available here: https://raw.githubusercontent.com/basgpol/Text-as-data/master/final_data.csv

The Jupyter notebook is available here: https://github.com/basgpol/Text-as-data/blob/master/webscraping%20US_speeches.ipynb

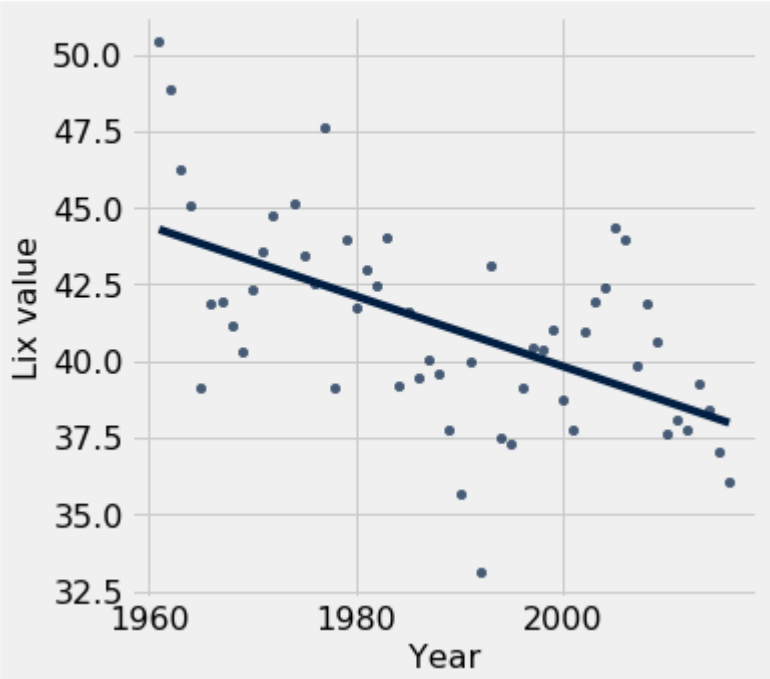
5.2. Figure 1: Sentiment and growth rate

```
The correlation coefficient is 0.2705931954096512
The p-value is 0.045702901253453526
```



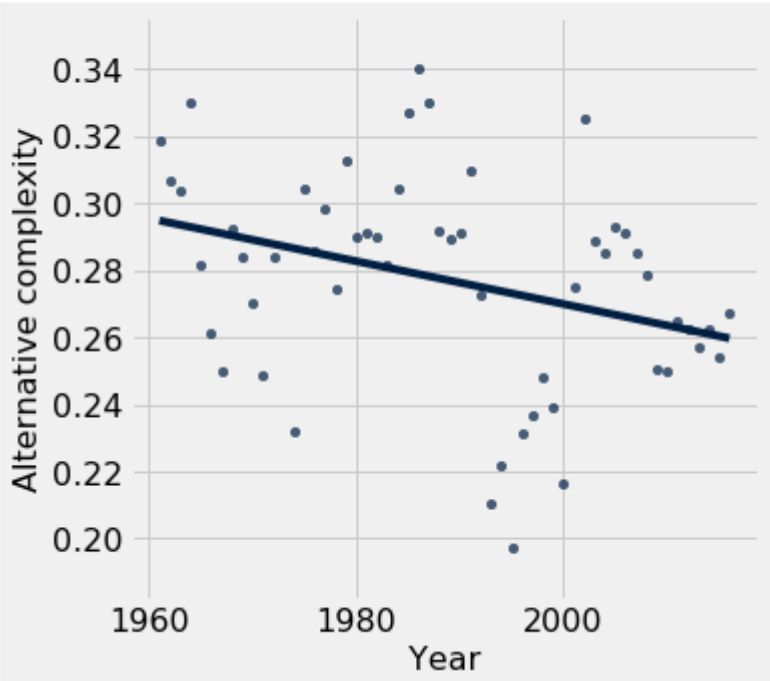
5.3. Figure 2: Language complexity (LIX)

```
The correlation coefficient is -0.5674696261887595
The p-value is 6.247580152329038e-06
```



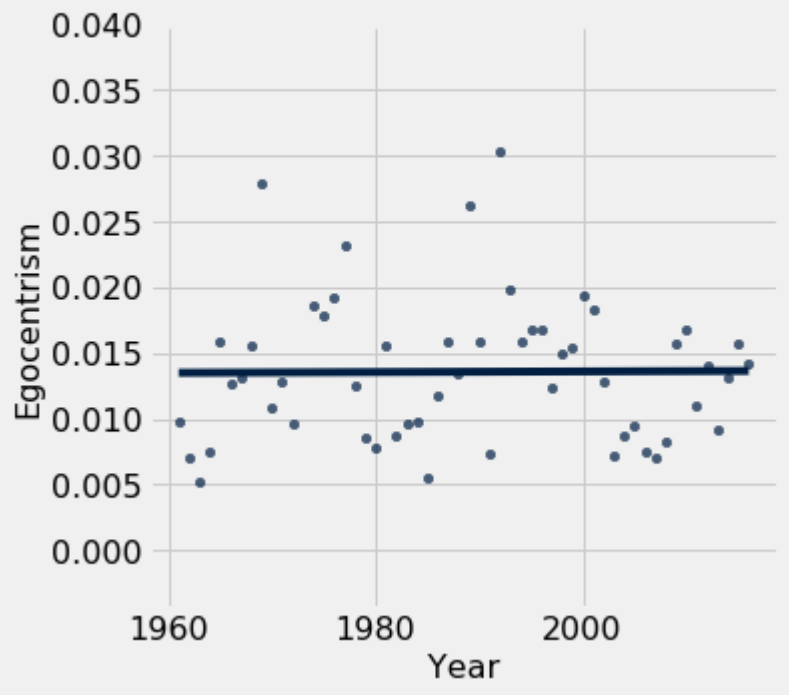
5.4. Figure 3: Language complexity (Proportion of unique words)

```
The correlation coefficient is -0.3271359892102729
The p-value is 0.014774042566723902
```



5.5. Figure 4: Frequency of egocentric words

```
The correlation coefficient is 0.007908164779331324
The p-value is 0.9543043930238604
```



6. Bibliography

Björnsson, Carl-Hugo. (1971). "Læsbarhed". *Gad Forlag*.

Downs, Anthony. (1957). "An Economic Theory of Political Action in a Democracy." Journal of Political Economy 65, no. 2: pp. 135-50.

Duch, Raymond M. (2007) Comparative studies of the economy and the vote. In Carles Boix & Susan C. Stokes (red.): *The Oxford Handbook of Comparative Politics*. Oxford: Oxford University Press, pp. 805-845

King, Gary, Robert O. Keohane & Sidney Verba. (1994). "Designing Social Inquiry: Scientific Inference in Qualitative Research". *Princeton University Press*

Lewis-Beck, Michael S. & Mary Stegmaier (2013). "The VP-function revisited: A survey of the literature on vote and popularity functions after 40 years". *Public Choice*, 157(3-4), pp. 367-385

Peters, Gerhard & John T. Woolley. (2017). "The American Presidency Project". *University of California, Santa Barbara.* Located at: http://www.presidency.ucsb.edu/sou.php