Replication Exercise #1: Report

Priscila Stisman | Samuel Cohen

Table of contents

Introduction

This report will detail our efforts to replicate specific code and outcomes from Arthur Spirling's 2016 paper "Democratization and Linguistic Complexity." Spirling's paper explores how the readability of parliamentary speeches increases overtime, especially as a result of the Second Reform Act of 1867, which took away property requirements for voting and enfranchised a significant portion of Britain's population (Spirling 2016). He hypothesizes that cabinet members' speech interpretability will become increasingly understandable overtime due to their prominent roles in government and newfound need to appeal to a new electorate— one that was less wealthy, less literate, and less educated as a whole (Ibid). Contrarily, backbenchers would not need to change their speech, as they are generally considered the "rank and file" and are not given the same level of public attention (Ibid). Spirling uses temporal trends of readability metrics (FRE scores, for instance) and multivariate regressions to assist in his

findings, which were relatively similar to his hypothesis. Our project attempts to replicate significant portions of this study, specifically temporal trends of readability, and the primary multivariate regression Spirling runs. We also add original research to this domain using the same data: readability of speeches by party, and TF-IDF and cosine similarity score analysis.

Differences & Similarities

FRE scores

The FRE Statistics results are very similar, though not identical. Our values for the minimum, first quartile, median, mean, and third quartile closely match those in the paper, with only a slight difference in the third quartile. However, the maximum differs significantly: while the paper reports a maximum FRE of 205.80, our result is 121.22. The bulk of the distribution is between 0 and 100, as in the paper.

The average readability score, in both the paper and our replication exercise, indicates that around the year 1860, the average cabinet speech becomes more comprehensible than the average non-cabinet speech, whereas before that, their mean comprehension scores were quite similar.

FRE Trends

Spirling plots the mean FRE scores over time by cabinet position (Ibid, 128). The general trend is stable and relatively unchanging before 1867, after which the scores increase dramatically for cabinet members, and only slightly for backbenchers (Ibid). Likewise, in our replication, we get a similar result. One slight difference is that the point of convergence between cabinet members and backbenchers in terms of readability happens much earlier than it does in Spirling's plot. However, this might be due to us aggregating based on year and not by quarter.

We also recreate the average syllable count per word score over time plot (Ibid, 129). We find almost the exact same trends (a decrease for cabinet MPs, stability for backbenchers), albeit at a smoother rate due to us aggregating by year and not quarter.

Overall, the general trends seem to match, with cabinet FRE scores overtaking those of backbenchers several years before 1867.

OLS Regression

The OLS regression results for comprehension scores by cabinet position, using the same set of controls as the authors, yield very similar coefficients, though not identical. However, the sign of the coefficients is consistent in all cases. Some of our coefficients have lower p-values

than those in the paper. For example, the Reform Act dummy is significant at the 1% level in our results, while in the paper, it is significant at the 5% level.

Autopsy

Data Retrieval

We use two primary data sets, which are essentially different iterations of each other. The first data set is the "bigframe" data set used for quantitative analysis in Spirling's research. This data set contains metadata on the year a speech took place, the word and syllable count, FRE score, political party of the MP in question, whether or not they were a cabinet member, and the competitiveness of their seat. We also use the raw speech data in both our attempt to recreate much Spirling's analysis, as well as in our original additions. Like Spirling, we subset for speeches only between the years 1832 and 1915. We also derive a random sample of 10,000 from the raw speeches due to the sheer size of the data.

Retrieving this data was not very intuitive. While Spirling has a Harvard Dataverse page for his replication materials, it only included the code and bigframe numerical data—that is, there was no raw data in this repository. To find the raw speech data, we navigated to the Arthur Spirling and Andy Eggers Database, which included the CSVs for the raw data.

Replication Challenges

One of the primary challenges we faced was the fact that we used only a sample of the raw speech data. While we believed 10,000 observations out of well over 600,000 would be representative, there were still some discrepancies in our results. For instance, while the mean and median FRE scores for our sampled speeches were relatively similar to those in the paper, the standard deviation was a lot larger, and the minimum and maximum values we quite different as well.

Another challenge we encountered was interpreting very large and very small FRE scores. While FRE scores typically range between 0 and 100, both the paper and our replication indicated scores much higher and lower than these thresholds. After some research, we discovered that it is indeed possible to have FRE scores below 0 (if language is particularly complex).

Replication Successes

Perhaps our greatest replication success was our results for Spirling's multivariate regression. Coefficients differed only slightly, if at all. This success was likely due to us using the bigframe data rather than our samples and cleaned raw data. While a LaTex illustration of the regression was not provided in the paper, we were still able to recreate it with good results.

Extension

The following are our additions to the research presented in this paper, as well as suggested improvements.

Party Affiliation and Speech Readability

For our original additions, we decided to look at how readability evolves along party lines. The majority of Spirling's paper focuses on how cabinet status affects speech readability, rather than party affiliation. We decided to explore speech readability over time by political party.

We graphed the mean FRE scores by year over time, and then disaggregated the trend lines by political party: Conservative and Liberal. We hypothesized that conservative MPs would likely not change their speech as much as Liberals, due to our belief that Liberals would garner more working class support.

Our findings indicate that speech becomes easier over time after 1867 for all MPs, regardless of party affiliation. However, interestingly, after 1867, as one party increases their speech readability over a short time, the other often decreases, and vice versa.

TF-IDF Weighting

For our original analysis, we also used the raw speech data for text analysis. Spirling's analysis mostly focuses on readability, so we decided to explore a bit more in the weeds. We first tokenized, preprocessed, and created a data frequency matrix (DFM). We then took the weighted TF-IDF scores for each token to find which words carried the most importance throughout the corpus.

We find that simple procedural words have the highest scores, as speeches can often be very formulaic. For instance, "yes," "sir," "amend," "order," and "bill" had the highest weighted scores.

Cosine Similarity

Finally, we used cosine similarity to find the speeches that are most similar with each other. We utilize our TF-IDF matrix to find the closest cosine similarity for each document. We then created a cosine similarity matrix, and pivoted this table to long format for ease of analysis. After ordering by cosine similarity score, we analyzed the document diads with the highest scores.

Similar to the highest TF-IDF scores, the speeches with the highest similarity scores exhibited very formulaic language that appears to be common in parliamentary procedure. For instance, the speech duo with the highest cosine similarity score between each other was an MP in 1876

saying "he will repeat the question on Monday," and an MP in 1890 saying "I will repeat the question on Monday." The next highest duo was an MP in 1872 saying "He would withdraw the amendment," and another in 1899 saying "I will withdraw the amendment."

Suggested Improvements

- Easier to access raw data: One of our primary challenges was finding the raw speech data Spirling used in his analysis. Adding this to the Harvard Dataverse repository would be a welcome update, and could help future researchers replicate and innovate at a much easier rate.
- Party affiliation analysis: While we provide a preliminary analysis on readability and party affiliation, additional research could help determine how political parties develop their speech over time. The British Political Manifestos corpus could be of significant use here.
- More analysis of raw text: TF-IDF, cosine similarity, KWIC, etc.: Future research should focus more on analysis of the text in speeches, not just readability, as interesting and insightful details may get lost in the weeds. In addition, sentiment analysis would also be a welcome addition– it would be interesting to see how MPs in different cabinet positions or political parties change not only their speech, but the tone of their message over time in response to various extraneous factors.
- Comparing different readability scores: One final action future researchers can take is to compare FRE scores to other metrics of readability, such as Dale Chall or SMOG. This would add robustness and would make it easy to spot any discrepancies.

References

Eggers, A. & Spirling A. (Accessed 2025). Eggers and Spirling Database. https://andy.eggers.rs/eggers_spirling_database.html

Prokopets, M. (Accessed 2025). The Beginner's Guide to Flesch Reading Ease Scores. Nira. https://nira.com/flesch-reading-ease/

Replication Materials for: 'Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915' - The Journal of Politics. (Accessed 2025). Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DDQ

Spirling, A. (2016). Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discource, 1832-1915. Journal of Politics. 78(1).