Methods of Corpus Linguistics

Final report

A comparative analysis of recent U.S. Republican party platforms and presidential campaigns

Mariann Lengyel
(r0977681)

Contents

A comparative analysis of recent U.S. Republican party platforms and presidential campaigns

1. State-of-the-art and research questions

Every four years, before the presidential elections, U.S political parties convene to approve a party platform upon which their chosen candidate will run." A party platform is a set of principles, goals and strategies designed to address pressing political issues." (Constitutional Rights Foundation) Party platforms - approved by the national conventions of the parties - are supposed to provide the basis of and guidance for the candidate's presidential campaign. As such, these documents provide great insight into the official position of the party (a.k.a. the voice of the party establishment) and a great way to measure the "individualism" of each actual presidential campaign. In my analysis, I am looking to find out how presidential campaigns digress from the tenets of the platform (not necessarily by contradicting it, but also by shifting the focus and altering the characteristic features, topic representation of the document).

As Fairclough (1989) puts it, politics is not merely conducted through language, but politics *is* language. It is thus understandable, that the corpus-based analysis of political texts has been an increasingly popular method. This genre of analysis is further facilitated by the growing availability and the meticulous recording of political texts by public bodies, as Mayaffre & Poduat (2013) notes, who used the method to assess the European identity of French presidents through their speeches. In the United States, the quantitative analysis of presidential speeches has also garnered significant attention. Notable studies in this area include Chen, Yan & Hu (2019), who conducted a corpus-based study on the linguistic styles of Hillary Clinton and Donald Trump and Hamed (2020), who focused on keywords and collocations in U.S. presidential discourse since 1993, offering a longitudinal perspective. These studies collectively underscore the growing importance of corpus-based analysis in understanding the nuances of political communication.

Party platforms, as the primary expression of the views of US political parties have also been subject to many analyses. Comparative studies between Democrat and Republican platforms (Smith, 1992, Kidd, 2008) used these documents to reveal the (lack of) differences between the two parties' positions. Party platforms were also subject to longitudinal analyses, where researchers were looking at the evolution of a specific concept in the parties' rhetoric over time. Jackson & Heath (2023) used corpus linguistics methods to analyse the Republican Party's ideological view on higher education and how it has changed between 1948 and 2020. Motyl (2012) was testing the theory of political realignment through an analysis of party

platforms between 1856-2008. Meanwhile Conger (2010) contrasted state-level Republican party platforms to detect the power of different coalitions within the Republican party.

Finally, a few words on the notion of party discipline. As Kam (2014, p. 399) puts it: "party cohesion is the degree to which members of a party are observed to work together in pursuance of a party's goals. Party discipline refers to party cohesion that is generated and sustained by the party's leaders." Party discipline and cohesion in the US politics have been primarily studied within the context of the Congress (Mcarty, Poole & Rosenthal, 2001), and particularly in the House of Representatives (Pearson, 2015) and Krehbiel (2000) who used roll-call votes to measure party cohesion. Examining a US party's unity through the rhetoric of its different leading figures however has been a so far undiscovered area of research.

2. Description of chosen corpus and type of analysis

In my analysis I was looking to examine the party cohesion through comparing party platforms and presidential campaigns in the last 4 elections. As I found out, in 2020 however, the Republican Party did not publish a platform, as they were unable to safely hold a convention amid the pandemic. Thus, I have decided to only look at the three presidential campaigns featured in the table below. As for the temporal dimension, I have decided to exclude the texts of the candidates before the election year, to give their campaigns a uniform timeframe, regardless of the different campaign start dates.

Election year	*Candidate*
2008	John McCain
2012	Mitt Romney
2016	Donald J. Trump

To assemble the datasets, I have used the website of The American Presidency Project, a curated archive of all presidential speeches and campaign documents, collected and maintained by the University of California. From this site, I first manually extracted the texts of the three party platforms. Next, I built a web scraper program incorporating the BeautifulSoup Python library to extract all the campaign documents of the three presidents in plain text format from the website. The program was built to be customizable by the name of the target president and the required timeframe. With another Python script, I then unified all the campaign documents of each president into single .txt files. Next, I used the text mining

package of R to preprocess and clean the texts. The resulting six corpora can be observed below:

Name	*Number of tokens*
Rpp_2008	13936
Rpp_2012	18281
Rpp_2016	21129
McCain_2008	474614
Romney_2012	664621
Trump_2016	152468

For my research, I have decided to do a two-step analysis with two different methods. First, I was looking at comparative word frequencies to identify both the main rhetoric and content differences for each president and the corresponding party platform of that year. Then I also conducted a collocation analysis by looking at the most common bigrams of words that had the same proportional frequency in the party platforms and the presidential campaign documents. The reason for this additional step was to understand whether, beyond the visualization of the comparative frequency analysis, there is a difference between the framing and usage of the shared frequency words.

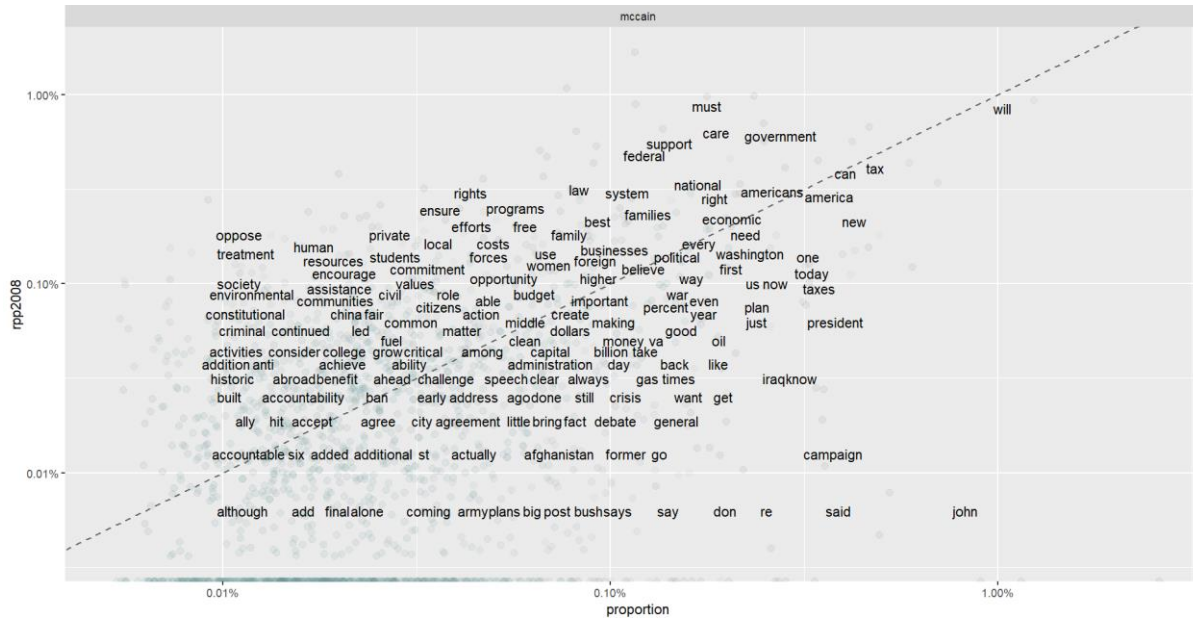
Finally, a few words on the resources I used outside of the class materials for my project. Apart from the documentation of the libraries I used, for both analyses, I have followed the methods of Silge & Robinson (2017), particularly Chapter 1.3 and Chapter 4.1 in their book Text Mining with R: a Tidy Approach. Additionally, I have used Stackoverflow and ChatGPT for debugging and correction purposes (most prominently at the token counter function) and for explanation of code snippets in the material I could not understand. Also, I would like to mention that my campaign scraper program is based on another script, called I have written for another project, that scrapes all documents from a chosen president from the same website. I have rewritten the code to scrape the campaign documents instead, and added customizability by date and candidate name.

3. Discussion of results

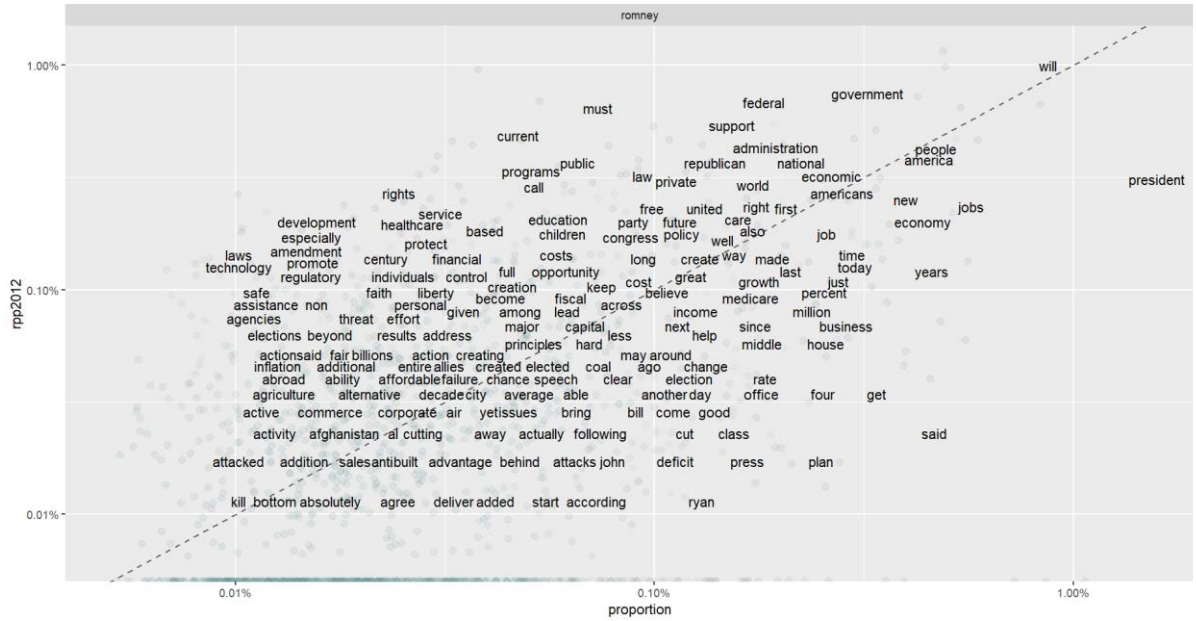
I. Comparative frequency analysis

The results of the comparative frequency analysis can be observed on the plots below. Words that are closer to the X-axis have a higher proportional frequency in the presidents'

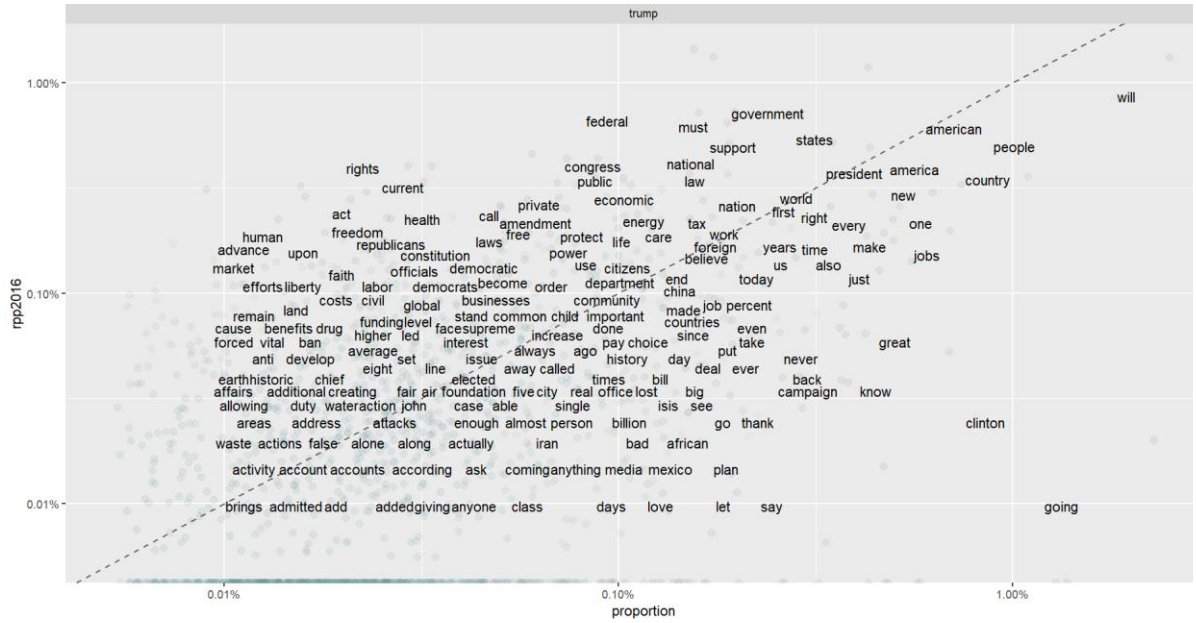
corpora, while words that are closer to the Y-axis have a higher proportional frequency in the party platform corpora. Words that are close to the diagonal line have similar proportional frequencies in both the campaign documents and the party platform.



Comparative frequency of the McCain and the RPP_2008 corpora



Comparative frequency of the Romney and the RPP_2012 corpora



Comparative frequency of the Trump and the RPP_2016 corpora

For better visualization, in the table below I have collected some of the interesting results, highlighting the reoccurring words and concepts. On the plots, I ignored the word *will,* which came on top for every plot, also, for obvious reasons, the presidents' own names as well as their opponents' names.

	Party Platform	Both	Presidential Candidate
2008	Environmental, students, **rights,** programs, support, **care**, family, constitutional, **China**, women	**Economic**, **tax,** budget, businesses, **Americans,** can, higher, leadership, believe, political	Iraq, **Afghanistan,** army, war, oil, crisis, gas, always, even, just, want, hope, **good**, **big, actually**
2012	Education, healthcare, development, technology, **faith**, **children**, inflation, **liberty**, agriculture, **rights,** religious	**Afghanistan**, **economic**, corporate, kill, attacked, **care**, right, **Americans,** people, can, create(d), capital, principles, great	**Jobs**, business, deficits, Medicare, **class,** growth, **plan,** cut, **attacks**, **good**, change, **actually**
2016	**Rights**, **liberty**, **faith**, health, education, support, constitution, energy, care	**Tax, attacks**, community, **China**, work, **children,** child, foreign, fair, believe, American, president, attacks	African, Mexico, **job(s),** ISIS, Iran, **class**, bad, **plan,** never, just, **big**, **ever, actually**

As I was expecting, the party platforms tended to center policy words (e.g. *health, education, agriculture*) whereas the campaign texts were less characterized by them. Interestingly, in all three of them, the word *rights* had a visibly higher frequency than in the corresponding presidential campaigns, suggesting that while it is an important topic for the party, the candidates did not rely on it for their campaigns. Religion-themes words such as *faith* or *religious* were also typically used by the party platforms in a higher frequency. On the other hand, the word *jobs* was a term preferred by the presidential candidates. Their campaign texts were also heavier on the adjectives and other non-topic words, such as *always, ever, good/bad* etc.

In 2008, both the McCain campaign and the party platform had *economic* (and related words such as *tax, budget,* and *businesses*) as a central theme. However, while the party platform frequently used words related to domestic policy (*environmental, students, family, care, support*), the McCain campaign was characterized by an emphasis on foreign policy *(Iraq, Afghanistan, army, war, oil*). This is not surprising given the candidate's military career.

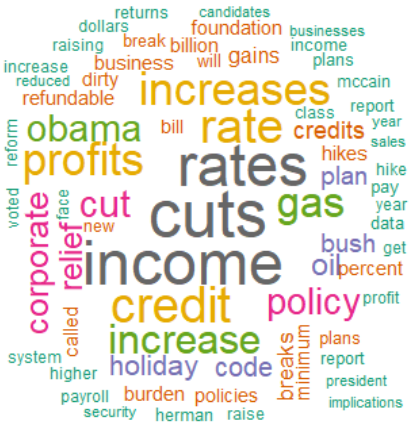
In 2012 on the other hand, foreign policy-related words, such as *Afghanistan, kill* and *attacked* move closer to the center i.e. being used by both the candidate and the party platform equally frequently, while economy also remained a shared theme. The candidate, Mitt Romney did not stray far from the party platform as his campaign documents were also characterized by words connected to domestic issues (*deficit, growth, jobs, Medicare*).

In 2016 however, there was an even more characteristic digression in the direction of foreign policy by the presidential candidate. Whereas the party platform continued to focus on domestic topics, such as *health, education, energy*, the Trump campaign texts digressed towards words, such as *Iran*, *Africa, Mexico,* and *ISIS*.

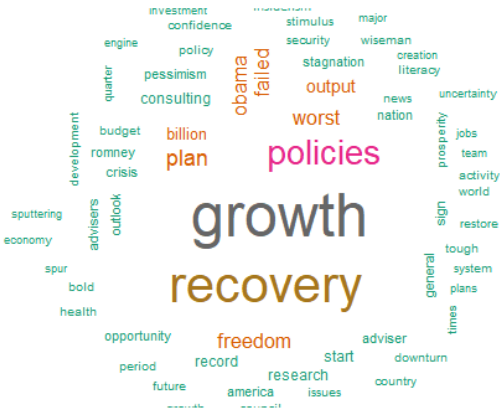
II. Bigram analysis

For the second part of my analysis, I was looking at the middle column of the table above. I was looking to find out whether there is a difference between the context and usage of words that in the first analysis appeared to be common reoccurring themes in both corpora. For all three pairs I was looking for the policy-related word that was highest on the diagonal line, i.e. had similarly high frequency in both the party platform and the presidential campaign texts.

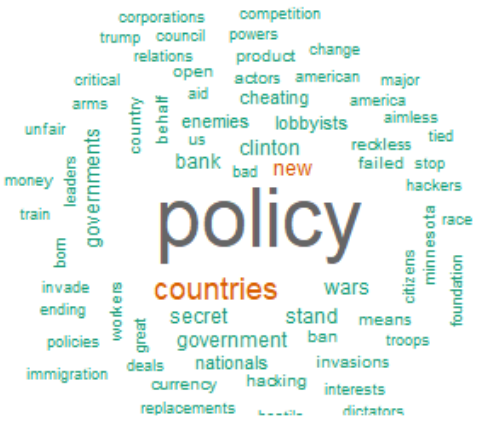
Thus, I have identified 3 target words: *tax* for the year 2008, *economic* for 2012, and *foreign* for 2016. With another R script I created frequency lists out of all the bigrams where the target word was either in the first or second position. Then I stripped the target word from the bigram and visualized the resulting single-word frequency lists on pairs of word clouds. The results can be observed below.



Collocates of the word "tax" in the RPP_2008 and the McCain_2008 corpora



Collocates of the word "economic" in the RPP_2012 and the Romney_2012 corpora



Collocates of the word "foreign" in the RPP_2016 and the Trump corpora

In the 2008 corpus-pair, the word *rates* was dominant in both bigram-frequency lists as the collocate of tax, while the words *code* and *credit* were prominent in the bigrams of the party platforms, and the words *cuts* and *income* characterized the bigrams of the McCain texts. Some of the words that had the most different proportional frequency in the texts alone also showed up as the most frequent collocates of the word tax, namely *support* and *family* for the party platform, and *gas* and *oil* in the McCain texts.

From the 2012 word cloud it is visible, that the bigram *economic growth* dominated both corpora,, suggesting a presence of a common party line. In the bigrams of the party platform, the word *economic* tended to be paired up with the words *greater, development, prosperity, challenges, freedom, nation, downturn, liberty, stability.* In the Romney corpus, the other prominent collocates of *economic* were *plan, policies, recovery, plan, billion, Obama, failed, worst, output, freedom.*

The largest difference was between the bigram frequencies of the target word *foreign* of the 2016 corpora. Whereas the top bigram-pair for Trump was the words *foreign policy* by far, in the party platform text, this did not appear prominently. There, the word *foreign* was most strongly associated with *assistance, nationals, governments, economy, investments, workers, relations* and *markets.* It was also noticeable that the texts of the Trump campaign were heavily associated with negative collocates, such as *bad, enemies, hacking, unfair, reckless, failed, hackers, ban, invasions, invade, dictators, aimless* and *wars.* The word cloud of the party platform text however, only contained three negative words namely *entanglement, exploit* and *encroachment,* hinting towards a possible difference in the degree of isolationism in the rhetoric of the candidate and that of the party.

4. Conclusions and ideas for follow-up research

In conclusion, in the last three party platforms and corresponding presidential campaigns, there is definitely a difference in content. Party platforms tended to be more focused both on ideas and ideals (*rights, faith, liberty*) and domestic policies and their execution (*health, education, support, programs*), whereas presidential campaigns, as they are more rhetoric in nature, were much more characterized by verbs, adjectives and adverbs (*actually, ever, always, good/bad*).

When it came to the words connected to policy, there were more digression in topics in the McCain and Trump campaigns. The proportional comparative frequency of the words used in their campaign documents suggests that they were considerably more focused on foreign policy matters than the party platforms of the corresponding years. This cannot be said about the Romney-campaign corpora, where even the words that were much higher in frequency than in the text of the party platform tended to revolve around domestic policy. Apart from terms that, given the context, were expected (*American, political*), the most prominent of the shared policy-related words were *economic* and *tax* between the party platforms and the presidential campaign texts.

In the course of the bigram analysis of the most frequent shared policy-terms for each pair of corpora, it was revealed that the McCain-, and Romney-documents were closer to their party platforms, whereas the Trump-corpora demonstrated less shared frequency.

Finally, as I will discuss it in the next section, these results should be interpreted keeping in mind the limitations of the analysis. Looking at the campaign documents provides insight into only one genre produced within a presidential campaign. Even if it is representative of the general direction of the campaign, one genre does not provide the full picture. A potential way forward would be to look at one specific presidential campaign and gather data from a range of genres (e.g. interviews, social media, live debates etc.) while also trying to expand the dataset that captures the voice of the party establishment from the same year, beyond the party platform. Using these two comprehensive datasets, one conduct more refined analyses to understand the differences and similarities between them. (This is in fact a research topic I am planning to explore in my thesis.)

5. Reflection on, and evaluation of chosen approach

My chosen method has proven to be useful for understanding some of the trends and patterns in the presidential campaign rhetoric and how those relate to the party platforms. There were two considerable problems however that possibly distorted the results and/or not showed the full picture.

First, as social media is gaining momentum, traditional media and press formats are on the decline even in presidential campaigns. This is visible from the difference between the sizes of the McCain-Romney corpora and that of the Trump corpus, the latter being less than a third of size than the other two. For the Trump-campaign, I should have included his social media activity, as that was arguably more impactful than the official campaign documents. But then, I should have also included the social media activities of the other two presidents, which were considerably less sizeable. Also, since discourse used on social media is entirely different in tone and formalities than language used in traditional press, that would have introduced an additional variable into my research.

The other aspect in which my analysis may fall short is the lack of lemmatization in the pre-processing of the corpora. This was a problem I could not overcome, since all the lemmatization packages and functions resulted in a lot of distorted or badly lemmatized words for some reason, which were rather distracting during the analysis. (e.g. hillari instead of hillary) I ended up deleting this step from my corpus cleanup function, but then some words were represented twice (e.g. job and jobs) which took away from the accuracy of the analysis.

6. Description of “difficulties encountered during project” (plus chosen solutions)

One difficulty I have encountered was the non-interoperability of corpus objects resulting from using different libraries. After a lot of research, I have chosen the text mining package of R as it seemed to be the easiest, fastest way to process the corpora. However, as I learnt, the analyses that followed required the data as tidy objects, whereas tm’s output were vcorpus objects. I had to convert all six of them.

Another difficulty of the text mining package was that I first tried to use the folder of individual scraped documents to create the corpora. For, the vcorpus format is a complex dataframe that stores documents as a list where each element is a complex object representing a document. This introduced unnecessary complexity into my data, since I was not interested in the differences between the individual campaign documents, while preventing the application of my chosen analysis methods, which would have required the documents to be treated as one single line. As I encountered several difficulties trying to use the unnest function, I decided to go back to Python and write a script that unifies the individual text files of the folder into one text file.

Furthermore, it is perhaps worthy to mention that I had quite some difficulties incorporating the frequency comparison method of Silge & Robinson (2017), especially the pivot wider and pivot longer parts, since they were working with three tables instead of two. To overcome this problem, I have opted for a rather awkward solution, that resulted in a table with a column entitled “other” that just contained the singular value of the president’s name.

Lastly, I have not found the exported plots of the first analysis sufficiently detailed, so instead of exporting, I have decided to screenshot the zoomed versions, to be able to feature a larger variety of words.

7. Bibliographical references

Campaign documents. (n.d.). The American Presidency Project. Retrieved November 1, 2023, from https://www.presidency.ucsb.edu/documents/app-categories/elections-and-transitions/campaign-documents?items_per_page=60&page=0

Chen, X., Yan, Y., & Hu, J. (2019). A Corpus-Based Study of Hillary Clinton's and Donald Trump's Linguistic Styles. International Journal of English Linguistics. https://doi.org/10.5539/ijel.v9n3p13

Conger, K. H. (2010). Party Platforms and Party Coalitions: The Christian Right and State-Level Republicans. Party Politics, 16(5), 651-668. https://doi.org/10.1177/1354068809346003

CRAN - Package tidytext. (n.d.). The Comprehensive R Archive Network. Retrieved November 7, 2023, from https://cran.r-project.org/web/packages/tidytext

CRAN - Package tm. (n.d.). The Comprehensive R Archive Network. Retrieved November 6, 2023, from https://cran.r-project.org/web/packages/tm/index.html

Fairclough, N. (1989). *Language and power*. Pearson Education.

Hamed, D.M. (2020). Keywords and collocations in US presidential discourse since 1993: a corpus-assisted analysis. Journal of Humanities and Applied Social Sciences.

Jackson, R. A., & Heath, B. L. (2023). An ideological view of college: A textual analysis of Republican Party platforms from 1948 to 2020. *Open Journal of Political Science*, *13*(04), 347-368. https://doi.org/10.4236/ojps.2023.134022Martin, S., Saalfeld, T., & Strøm, K. (2014). *The Oxford handbook of legislative studies*. Oxford Handbooks.

Kidd, Q. (2008). The real (Lack of) difference between Republicans and Democrats: A computer word score analysis of party platforms, 1996–2004. *PS: Political Science & Politics*, *41*(3), 519-525. https://doi.org/10.1017/s1049096508080694

Krehbiel, K. (2000). Party Discipline and Measures of Partisanship. *American Journal of Political Science*, *44*(2), 212–227. https://doi.org/10.2307/2669306

Mayaffre, D., & Poudat, C. (2013). Quantitative Approaches to Political Discourse: Corpus Linguistics and Text Statistics. In *Speaking of Europe: Approaches to complexity in European political discourse* (pp. 65-85). John Benjamins Publishing.

McCarty N, Poole KT, Rosenthal H. The Hunt for Party Discipline in Congress. *American Political Science Review*. 2001;95(3):673-687. doi:10.1017/S0003055401003069

Motyl, M. (2012). Party evolutions in moral intuitions: A text-analysis of US political party platforms from 1856-2008. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2158893

Partington, A., & Johnson, J. H. (2020). Corpus Analysis of Political Language. In *The concise encyclopedia of applied linguistics* (pp. 327-339). John Wiley & Sons.

Pearson, K. (2015). *Party discipline in the U.S. House of Representatives*. University of Michigan Press.

Political parties, platforms, and planks. (n.d.). Constitutional Rights Foundation -. Retrieved November 6, 2023, from https://www.crf-usa.org/election-central/political-parties-platforms.html

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media. Retrieved November 1, 2023, from https://www.tidytextmining.com/