A data science approach to 138 years of congressional speeches

Ethan. C. Tucker, Colton. J. Capps, Lior Shamir*

Department of Computer Science, Kansas State University, Manhattan KS 66502, USA
Email: lshamir@mtu.edu

Abstract

The availability of automatic data analysis tools and large databases have enabled new ways of studying language and communication that were not possible in the pre-information era. Here we apply a quantitative analysis to a large dataset of USA congressional speeches made over a period of 138 years. The analysis reveals that the readability index of congressional speeches increased consistently until the 96th congress, and then started to decline. Congressional speeches have also become more positive over time, and in general express more sentiments compared to speeches made in the 19th century or early 20th century. The analysis also shows statistically significant differences between Democratic and Republican congressional speeches.

Keywords: Speeches; congress; data science

1. Introduction

Speeches have been a primary methods of communication in politics and public administration, and their pivotal role in the government and Democratic process has been noted since the ancient Greek and Roman government systems (Champion, 2000; Pepe, 2013; Triadafilopoulos, 1999; Fantham, 2003). As these speeches reflect the agenda of the speaker, analysis of the speeches can provide important insights about the way the speaker use language to communicate their views (Dowding et al., 2010; Eshbaugh-Soha, 2010; Schaffner, 1996; Boromisza-Habashi, 2010; Remer, 2008). The information era enables the digitization of archives, making very large databases accessible to large-scale manual and machine analysis of the data. That introduces a new approach to the use of language analysis that can reveal insights about political communication that are difficult to identify manually (Cardie and Wilkerson, 2008; Grimmer and Stewart, 2013; Wilkerson and Casas, 2017).

Converting transcripts of political speeches into numbers can help identify differences and trends in the language used in the speeches that are impractical to detect in

large databases by manual inspection of the text (Grimmer and Stewart, 2013). It is clear that such quantitative analysis cannot capture all information, as machine analysis of text has not yet elevated to the analysis power of the human brain. Namely, computer analysis cannot yet fully "understand" the content and context of the speech or define its political meaning. However, machine analysis can reduce the speeches to text elements that can be measured, and therefore allows quantitative analysis and statistical inference of the speeches.

Substantial work has been done in the application of discourse analysis to political communication (Mahdiyan et al., 2013; Adeyanju, 2016; Bonikowski and Gidron, 2015; Reyes, 2015). Natural language processing techniques have also been used to determine ideology proportion in political speeches (Sim et al., 2013), and multi-modal methods that combine text analysis with automatic eye tracking provided additional information to the analysis of the text alone (Scherer et al., 2012). It has also been shown that a political position can be identified automatically from the speech transcript (Laver et al., 2003). Frequency of certain terms and words has also been used to show similarities and differences between politicians, and can be measured directly through their speeches (Savoy, 2010), or indirectly through social media and other content related to the politicians (Chung and Park, 2010). Analysis of political speeches was also applied to identify gender-related language differences in parliamentary speeches (Sensales et al., 2018).

The US Congressional Record is one of the longest spanning and most significant collections of political documents available. Examining this set of speeches has the potential to yield useful information about historical trends in legislative priorities, speech patterns, and other features of debate and political communication in congress. As discussed above, making such discoveries manually is difficult due to the large size of the data.

Previous approaches to analyzing trends in the US Congressional Record using automation have focused on features designed specifically for legislative speeches. Quinn et al. (2006) examined the probability of speeches to be related to a given legislative topic between 1997 and 2005. Their analysis provided clear descriptions of the length of debate on various issues (Quinn et al., 2010). Another method of analysis sought to measure changes in partisanship over time based on the association bigrams with a political party (Gentzkow et al., 2019). The study determined that partisanship remained stable from 1873 until 1994, after which an increase in partisanship of congressional speeches was identified (Gentzkow et al., 2019). Yu (2013) used computational methods to show gender differences between congressional speeches of male and female legislators between the years of 1989 and 2008. Analysis of the frequency of congressional speeches in the 103rd (1993-1994) and 109th (2005-2006) congresses showed that female legislators speak at higher rate than male legislators (Pearson and Dancey, 2011). Yu et al. (2008) used automatic document classification to identify the party of the speech automatically, and identified changes

of the speeches across different years, reflected through changes in the classification accuracy of speeches based on the time difference between the training and test data. Diermeier et al. (2012) used a support vector machine (SVM) classifier to identify terms that distinguish between liberal and conservative speeches in the 101st to 108th congresses. Thomas et al. (2006) used automatic text classification to identify automatically whether a speech supports or opposes its relevant bill.

Here we applied quantitative text analysis to examine changes in congressional speeches over 138 years. The approach is based on multiple text measurements computed from each speech, and averaged in each year to obtain statistical signal reflecting the trends of these measurements.

2. Data

The initial dataset used in this study is a corpus of nearly $1.9 \cdot 10^6$ congressional speeches made between 1873 and 2010, retrieved from the Congressional Record[1]. Clearly, many of these speeches were made prior to the information era, and when no digital storage devices were available. The speeches were therefore digitized by applying Optical Character Recognition (OCR) to the records provided by HeinOnline[2] (Gentzkow et al., 2018). The data used in this study is the subset parsed from the bound editions of the Record, which currently spans from the 43rd (1873) through the 111th (2010) congress. The dataset also contained many transcripts of short comments that were not political speeches. An example of such comment is "Mr. Chairman, how much time do we have remaining?". Another example is "Mr. Speaker , I demand a recorded vote". Clearly, these are comments that do not express a political view, and therefore cannot be analyzed as speeches. To exclude such comments, text files that contained less than 1000 characters were not included in the dataset. After the exclusion of short comments, the dataset contained 959,237 text files such that each text file is a single congressional speech.

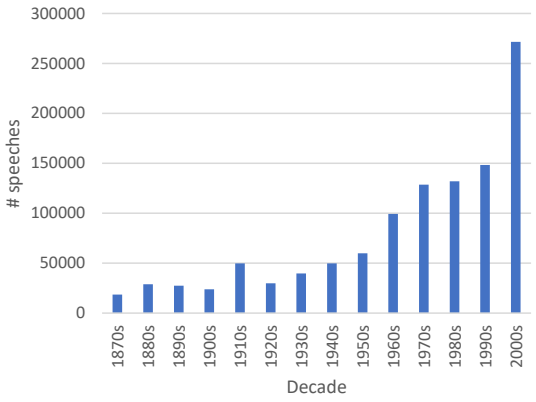


Figure 1: The number of congressional speeches in different decades.

The number of speeches in each year can be different, and the speeches are not equally distributed. Figure 1 shows the number of speeches in different decades. As the figure shows, the number of speeches generally increases in time.

[1] https://www.congress.gov/congressional-record/

[2] https://home.heinonline.org/

3. Text analysis

Analyzing data at the scales described in Section 2 cannot be done manually, and require automation. For that purpose, the open source UDAT text analysis software (Shamir, 2020) was used. UDAT computes multiple different aspects of the text, providing a comprehensive numerical analysis of large text datasets. Unlike some document classifiers, UDAT does not rely solely on the frequency of certain keywords appearing in the text, but also on elements that reflect the structure and writing style (Shamir et al., 2015; Alluqmani and Shamir, 2018). These text elements are quantified to show differences between the different classes of the text data.

To analyze and compare the speeches, several quantifiable text descriptors were extracted from each speech:

1. **Coleman–Liau readability index**: The purpose of the Coleman–Liau readability index (Coleman and Liau, 1975) is to estimate the reading level of the text, and associate the text with a grade. For instance, a Coleman–Liau readability index of 3 means that the text is estimated to be at a reading level suitable for a third grade student. The index is computed by $0.0588 \cdot \frac{100 \cdot w}{c} + 0.296 \cdot \frac{100 \cdot s}{w} - 15.8$, where w is the number of words in the text, c is the number of characters in the text, and s is the number of sentences.
2. **Word diversity**: Word diversity is determined by $\frac{W}{w}$, where w is the total number of words in the speech, and W is the size of the vocabulary of the speech (total number of unique words). If the same word appears in the text more than once, every appearance of the word after its first appearance in the text will increment w but will not affect W. If no word appears in the text more than once, the number of unique words is equal to the total number of words, and therefore the word diversity of the text is 1, which is the maximum possible value. The words are stemmed using CoreNLP (Manning et al., 2014) to correct for different forms of the same word.
3. **Word homogeneity**: The word homogeneity measures the change in the frequency of words throughout the speech. The text file of the speech is separated into 10 equal-sized segments, and the homogeneity h_i of word i is determined by $h_i = max(F_i) - min(F_i)$, where F_i is a set of the frequencies of the word i in the each of the text segments. Words that have frequency of less than 0.001 are ignored to avoid the impact of rarely used words. The homogeneity is then determined by the mean h_i. The word homogeneity is measured with an inverse scale. If the same words are used consistently throughout the text the word homogeneity is expected to be relatively low, while if a different set of words is used in different parts of the speech the word homogeneity is expected to be high.
4. **Total number of words**: The total number of words is a measurement of the length of the speech.
5. **Sentiments**: The sentiment expressed in each sentence in each speech was estimated using CoreNLP (Manning et al., 2014), such that each sentence is assigned with a sentiment value between 0 through 4. Sentiment of 0 means very negative, 1 is negative, 2 is neutral, 3 is positive, and 4 is very positive. The sentiment of each sentence is computed

by using the 215,154 labeled phrases and a parse tree of 11,855 sentence combinations, and the text is analyzed with a deep recurrent neural tensor network (Socher et al., 2013). The sentiment tree bank can be found in the CoreNLP website[3].

6. **Topic words**: The frequency of words related to certain topics. The frequency of topic words is measured by the number of words associated with the topic divided by the total number of words of the text file. The topics are sports, mathematics, science, states, shapes, school, positive words, negative words, languages, elections, food, money, driving, military, law, countries, dance, emotions, boats, energy, family, music, land, art, astronomy, colors, animals, dog breeds, cat breeds, fish, birds, reptiles, cars, beach, weather, fall, spring, summer, winter, vacation, farm, medicine, trees, transportation. times, clothing, shoes, hats, buildings, birthday, monsters, office, tools, camping, castles, fruits, circus, cooking, geography, kitchen, jobs, leaders, house, restaurants, roads, rocks, weapons, containers, acquaintances, yard, flowers, self, female, male. The topic words are taken from the Enhanced Learning thesaurus[4]. CoreNLP is used to identify the words by their stems, so that different forms of the same word are not counted as different words.

The analysis was done such that the feature values were averaged for each year, and the standard deviation and standard error were determined. In addition to the mean and standard deviation of all speeches made in each year, the features were also computed for the Democratic and Republican speeches separately, to identify possible patterns of differences between parties. More information and access to the software and source code used in this paper can be found in (Shamir, 2020).

4. Results

The results show substantial differences between speeches in different years. Some of the differences were natural to the wide span of years examined in this study. For instance, the use of words related to computers and the Internet were extremely low in speeches made in the 19th century, as these terms did not exist at the time, or had different meaning (e.g., the word ''computer") that was less of a political concern during that time. Other differences can be directly linked to the political situation of the time. For instance, Figure 2 shows the frequency of words and terms related to energy (e.g., "oil", "gas", "electric", etc) in congressional speeches.

As the graph shows, the frequency of energy-related words in congressional speeches has been increasing gradually since 1873. A spike in the frequency of energy terms can be identified around the year of 1973, and can be linked to the Organization of Arab Petroleum Exporting Countries (OAPEC) embargo, naturally attracting the attention of the legislators during that time. Another increase in the frequency of energy terms is noticed in 2008, when gas prices

[3]https://nlp.stanford.edu/sentiment/treebank.html
[4]http://www.enchantedlearning.com

Figure 2: Mean frequency of energy words in congressional speeches in different years. The error bars show the standard error of the mean.

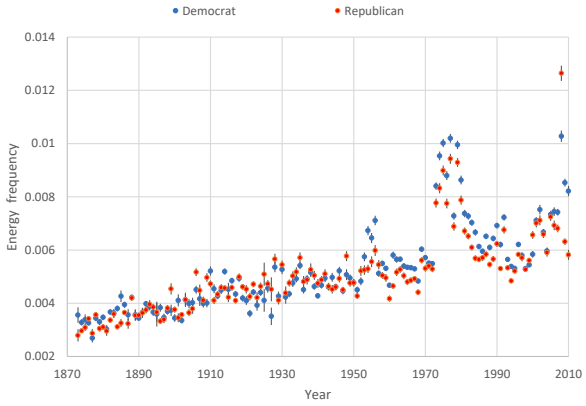


Table 1: Frequency of words that identify women in Democratic and Republican speeches during the 2000s and the 1870s.

Decade	Democrats	Republicans	t-test P
2000s	$0.002954 \pm 3 \cdot 10^{-5}$	$0.002525 \pm 3 \cdot 10^{-5}$	$< 10^{-5}$
1870s	$0.000537 \pm 2 \cdot 10^{-5}$	$0.000514 \pm 2 \cdot 10^{-5}$	0.41
t-test P	$< 10^{-5}$	$< 10^{-5}$	

soared. Interestingly, during the OAPEC embargo Democrats used energy terms more frequently than Republicans, while in 2008 Republicans mentioned energy in their speeches more frequently than their Democrat colleagues.

Figure 3 shows the change in frequency in words that identify women such as "she", "her", "hers", etc, as well as equivalent words that identify men. The figure clearly shows a sharp increase in the use of words that identify women starting the 1980s. The frequency of words that identify men has been decreasing, until reaching almost the same level as the frequency of words that identify women.

The data also show that in the years of 2000-2010 Democratic speeches used more words that identify women compared to Republican speeches. Table 1 shows the mean frequency of women identity words in the 2000s and in the 1870s. The table shows that in the 1870s the frequency of women term was very low, and nearly equal between the different parties. In the 2000s, words related to women identity are more than five times more frequent than in the 1870s, and there are also differences between the frequency of such words in Democratic and Republican speeches.

Figure 4 shows the change in the Coleman-Liau readability index. The graph shows a stable readability index of 7.5 to 8 until the 1930s, showing that in the late 19th century and the early 20th century congressional speeches were indexed at around the high middle school reading level. Since the 1930s, Starting the 1930s congressional speeches showed a constant increase in the readability index, peaking at around 10 in 1976, which is a $\sim$23% increase since 1939. In the late 1970s the trend reversed, and the readability index started to decrease gradually until an average index of below 9 in the beginning of the 21st century. One possible explanation to the simpler language of the speeches starting the 1970s can be related to the growing presence of the media that started during that time (Graber and Dunaway, 2017). With the media coverage of the congress activities, politicians could speak to an audience of legislators, but at the same

Figure 3: Mean frequency of words that identify women (left) and words that identify men (right) in congressional speeches in different years.

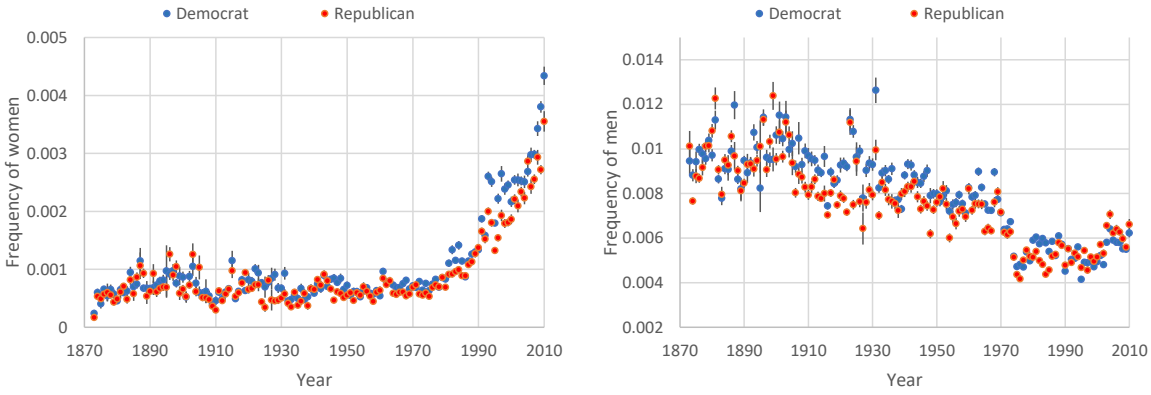


Figure 4: The mean Coleman-Liau readability index of congressional speeches in different years.

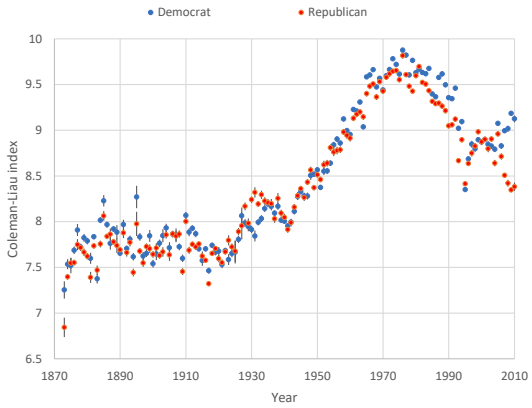


Table 2: Coleman-Liau readability index of Democratic and Republican speeches in different decades.

Decade	Democrats	Republicans	t-test P
2000s	8.95±0.008	8.67±0.009	$< 10^{-5}$
1980s	9.58±0.008	9.41±0.008	$< 10^{-5}$
1930s	8.04±0.011	8.18±0.013	$< 10^{-5}$
1870s	7.66±0.014	7.52±0.013	$< 10^{-5}$

time also communicate the speech to the general public through the media.

The graph also shows that Democratic speeches have a higher readability index compared to Republican speeches, and the difference has been becoming wider in the more recent years. Table 2 shows the mean readability index of Republican and Democratic speeches in several different decades. While Democratic speeches normally have a higher readability index than Republican speeches, in the 1930s that difference was reversed, and Republican speeches had higher readability index during that time.

Although word diversity is not mathematically related to the Coleman-Liau readability index, the diversity of words in a speech can provide another measurement of the complexity of the speech. Figure 5 shows that word diversity decreased gradually in the 19th century, and then increased during the 20th century until the 1970s. Starting the 1970s word diversity in speeches declined,

Figure 5: The mean word diversity in congressional speeches in different years.

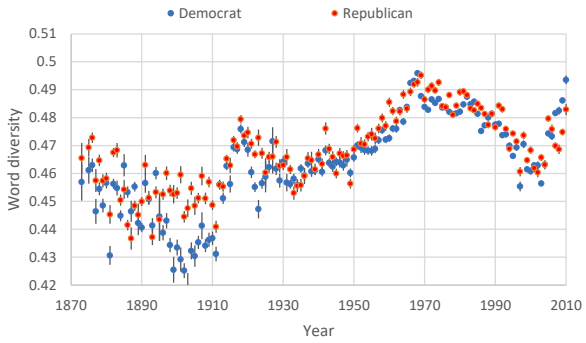


Table 3: Diversity of words in Republican and Democratic congressional speeches.

Year	Democrats	Republicans	t-test P
2000s	0.472±0.0003	0.47±0.0004	< 0.0001
1990s	0.47±0.0003	0.474±0.0003	< 0.0001
1980s	0.482±0.0003	0.484±0.0003	< 0.0001
1970s	0.484±0.0003	0.488±0.0004	< 0.0001
1960s	0.482±0.0003	0.486±0.0004	< 0.0001
1950s	0.469±0.0004	0.472±0.0005	< 0.0001
1940s	0.463±0.0005	0.465±0.0007	0.02
1930s	0.459±0.0005	0.461±0.0007	0.02
1920s	0.461±0.0007	0.468±0.0006	< 0.0001
1910s	0.456±0.0006	0.462±0.0005	< 0.0001
1900s	0.431±0.0009	0.453±0.0007	< 0.0001
1890s	0.443±0.0008	0.451±0.0007	< 0.0001
1880s	0.450±0.0007	0.452±0.0007	< 0.04
1870s	0.457±0.0009	0.466±0.0007	< 0.0001

and then increased again in the 21st century. The profile of change in word diversity is largely in agreement with the change in the readability index, although the two measurements are mathematically independent from each other.

As the figure shows, word diversity increased starting around the 1930s, peaked in 1969, and then gradually decreased until the beginning of the 21st century. That profile is very similar to the profile of change in the readability index, although the two measurements are independent. The graph also shows sudden increases of words diversity, mostly in Republican speeches in 1917, 1942, 1951, as well as a certain increase also in 1991. Interestingly, these are all years in which a war started. The year of 1874 was not used in the analysis due to a higher number of typos in the transcripts in these years, making 1874 different from other years in which the transcripts were accurate.

Table 3 shows the differences between word diversity in Republican speeches and Democratic speeches in different decades. The table shows that Republican speeches had a higher diversity of words until the 21st century, in which the word diversity in Democratic speeches became higher. The difference has also increased during the first decade of the 21st century, and in 2007 through 2010 the difference was 0.01 or higher, which is the highest difference since the beginning of the 20th century.

Figure 6 shows the change in word homogeneity. The graph shows that word homogeneity had been declining, which means that more recent congressional speeches tend to use the same set of words throughout the speech. Democratic speeches are more homogeneous than Republican speeches. The mean homogeneity of a Democratic speech is $0.0604 \pm 4 \cdot 10^{-5}$, while Republican speeches have an average measured homogeneity of $0.0617 \pm 4 \cdot 10^{-5}$. The two-tailed t-test statistical significance of the difference is ($P < 10^{-5}$).

Congressional speeches have also changed

Figure 6: The mean word homogeneity in congressional speeches in different years.

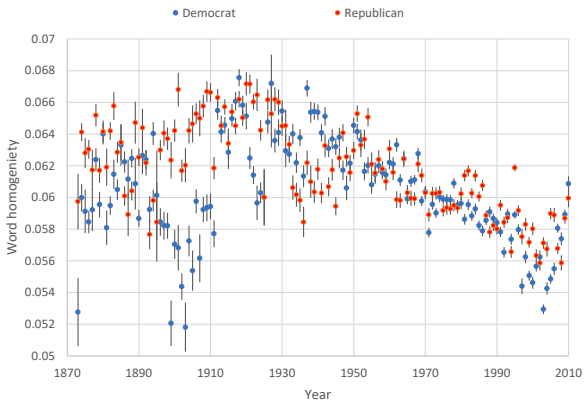
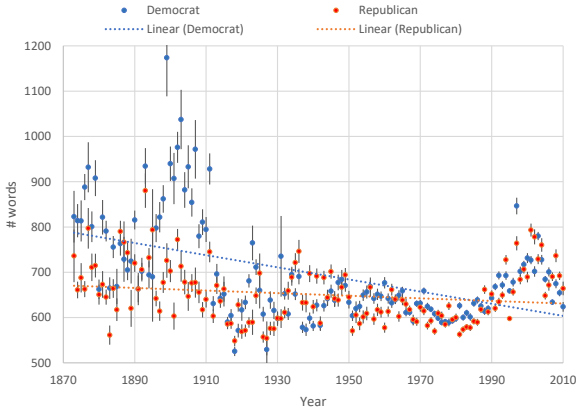


Figure 7: The average number of words in speeches of Republican and Democrat legislators in different years.



in length. Figure 7 shows the change in the mean number of words in a congressional speech in each year. The graph shows that congressional speeches were generally longer in the end of the 19th century, and became much shorter during the 20th century. Starting the 1980s, congressional speeches gradually became longer until the beginning of the 21st century, when the trend reversed and congressional speeches started to become shorter. For instance, in the 43rd through the 46th congress (1873-1876) an average congressional speech was $\sim$758$\pm$19 word long, while in 57th through the the 61st congress (1901 through 1909) the average congressional speech was reduced to $\sim$652$\pm$5 words.

The graph also shows substantial differences between the length of Democratic and Republican speeches. For instance, in the decade of 1900-1909 the average length of a Democratic speech was 933$\pm$12.5 words, while the average Republican speech during the same time was 679$\pm$7.68 words long. In 2000 through 2009 the differences became much smaller, with slightly longer Republican speeches. During that time, the average Democratic speech was 703$\pm$2.7 words, while the average Republican speech was 718$\pm$3.5 words.

Another element that changed in congressional speeches over time is the sentiment expressed in the speeches. Figure 9 shows the change in negative, very negative, positive, and very positive sentiments expressed in congressional speeches. The graph shows that both very positive and very negative sentiments became generally more common in the more recent years. The expression of stronger sentiments in congressional speeches forms a trend that changes in different years. Staring the 1980s, congressional speeches became less negative, and expressed more positive sentiments. The early 1960s were the

years in which the positive sentiments expressed in speeches increased, peaking in 1964. The year of 1964 also shows substantial difference between the sentiments expressed in Democratic speeches and the sentiments expressed in Republican speeches. The peak in sentiments during that year and the differences between Democratic and Republican speeches could be related to the Civil Rights Act that was signed during that time.

The difference in sentiments expressed in speeches also changed between Republican and Democratic speeches. In the 2000s the average frequency of positive sentences in Democratic speeches was 0.1129±0.0003, while it was 0.1147±0.0004 in Republican speeches ($P < 0.0001$). The frequency of negative sentences in the same decade showed higher average frequency of 0.284±0.0006 in Democratic speeches, compared to average frequency of 0.278±0.0007 in Republican speeches ($P < 0.0001$).

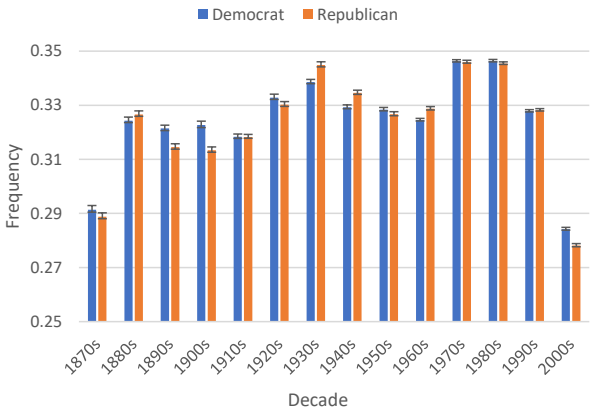
The difference in sentiment could be related to the political affiliation of the president at the time. Table 4 and Figure 8 show the frequency of negative sentiments from 2000 through 2010. Interestingly, the table shows that Democratic speeches expressed more negative sentiments in the years of 2000-2006, when the president was Republican, while in 2008-2010, when the president was Democrat, Republican speeches became more negative. In 2007 and 2008 no statistically significant difference in the negative sentences was identified.

However, the difference in negative sentiments does not change consistently with the political party of the president. For

Table 4: Frequency of sentences expressing negative sentiments in Republican and Democratic speeches.

Year	Democrats	Republicans	t-test P
2000	0.2422±0.0027	0.2366±0.0027	0.14
2001	0.302±0.002	0.2886±0.0022	<0.0001
2002	0.2887±0.0027	0.2812±0.0030	0.06
2003	0.3152±0.0019	0.3008±0.0021	<0.0001
2004	0.2448±0.0029	0.2229±0.0031	0.0002
2005	0.2983±0.0021	0.2775±0.0022	<0.0001
2006	0.2864±0.0026	0.2717±0.0027	<0.0001
2007	0.3047±0.0017	0.3048±0.0019	0.97
2008	0.3183±0.0020	0.3204±0.0022	0.48
2009	0.2519±0.0020	0.2715±0.0022	<0.0001
2010	0.2734±0.0028	0.2832±0.0034	0.02

Figure 8: Frequency of sentences expressing negative sentiments.



instance, in 1995 through 1999 Democratic speeches expressed more negative sentiments compared to Republican speeches despite a Democrat president during that time. Overall, no statistically significant difference in negative sentiments between Democratic and Republican speeches was identified between 1993 through 2000. During the Reagan administration, sentences that express negative sentiments were slightly more frequent in Democratic speeches (0.3466 ± 0.0007) compared to Republican speeches (0.3448 ± 0.0007), but with no statistical significance ($P \simeq 0.07$). While Democratic speeches tend to include more negative sentences than Republican speeches between 1980 through 2010, Republican speeches were more negative during the 1930s and 1940s. In the 1930s, the frequency of negative sentences in Democratic speeches was 0.339 ± 0.001, compared to 0.345 ± 0.001 in speeches of Republican legislators.

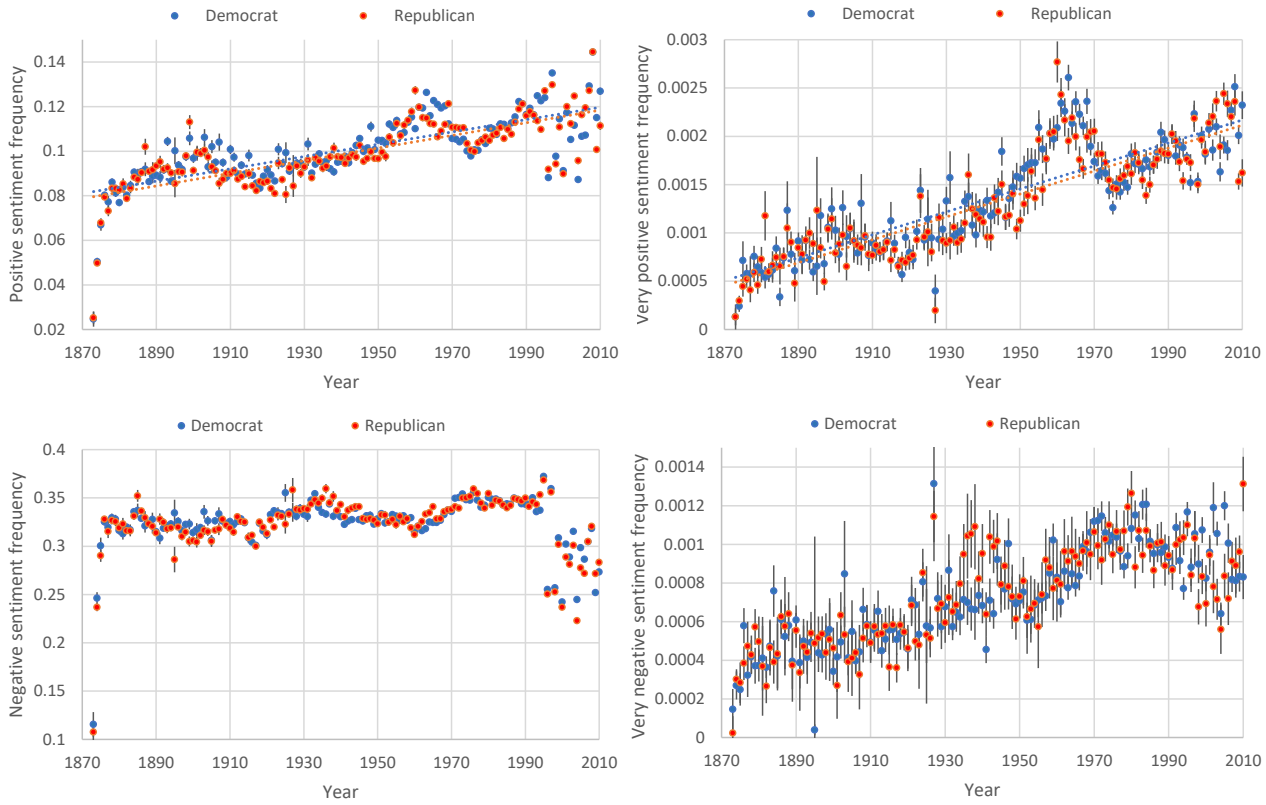
5. Conclusions

The accessibility of digital archives and availability of computational tools enables data-driven analysis of text, providing a new approach to studying language and communication (Cardie and Wilkerson, 2008; Grimmer and Stewart, 2013; Wilkerson and Casas, 2017). Here we used a large corpus of $\sim 10^6$ congressional speeches to analyze changes and trends in congressional speeches over time, as well as differences between speeches made by Republican and Democrat legislators. The large dataset of speeches covering a wide range of more than 100 years enables the analysis of the trends of changes in congressional speeches. Given the comprehensive analysis of a large number of text descriptors, this work can be used as a resource for further analysis of the long-term links between political speeches and other events or processes with political or societal nature.

The analysis shows a sharp increase in words related to women identity starting the 1980s. That change can be related to the change in the number of women in the congress, which has been increasing consistently since the 97th congress (1981). It can also be related to the higher number of bills related to topics relevant to women, which is naturally also a function of the number of women representatives.

An interesting trend was revealed by the Coleman-Liau readability index. The analysis shows a gradual increase in the readability index from a middle school level to high school level in the late 1970s, followed by a gradual and consistent decrease. The analysis also shows that speeches made by Democrat legislators have higher readability index compared to speeches of Republican legislators, and the difference has been becoming larger since the beginning of the 21st century. A very similar observation was made with the diversity of words, which is mathematically unrelated to the readability index but shows a very similar profile. The partisan split in the Coleman-Liau index correlates with the rise in partisanship beginning around 1995, as examined by Gentzkow et al. (2018), who identified textual framing of the Republican platform in 1994 with the increasing linguistic differences between parties.

Figure 9: The frequency of positive sentiments (top left), very positive (top right), negative (bottom left) and very negative (bottom right) sentiments expressed in congressional speeches.



The divergence in Coleman-Liau index suggests that such a modern partisan split may affect the style and form of their speeches as well as their content. The consistent decline in the readability index and diversity of words used in speeches can also be related to political speeches aiming at communicating with the general public through the media. The role of the media in increasing the power of the president was noticed by legislators at the time. For instance, in 1970 senator William Fulbright told congress that "Television has done as much to expand the powers of the president as would a constitutional amendment formally abolishing the co-equality of the three branches of the government" (Graber and Dunaway, 2017). The congress resisted radio and television broadcasting of most sessions until the 1970s (Graber and Dunaway, 2017). It is therefore possible that the increase in the broadcasting of speeches through the media and the presence of journalists in the congress gradually changed the purpose of speeches, as legislators started to address the media and the general public through their speeches.

Sentiment analysis shows that more recent speeches express stronger sentiments compared to speeches made in the 19th century, but negative sentiments expressed in speeches have been declining since the 1980s. Differences between parties show more negative sentiments in Republican speeches during the 71st through the 79th congress (1930s and 1940s). That changed in the following years, when Democratic speeches became somewhat more negative than Republican speeches. The analysis shows that since 2000, speeches of legislators from the opposite party of the president at the time the speech was made were more negative than speeches from legislators from the same political party as the president.

Due to the large size of the data, it is clear that the analysis done in this study is not possible without automation. The availability of computational tools for automatic text analysis enables new type of research of political communication, providing insights that are difficult to identify and quantify with traditional manual analysis. The method used in this study can be used for quantitative analysis of other large datasets of text data, enabling the detection of subtle trends and differences that are difficult to identify by manual analysis of the text.

Acknowledgements

We would like to thank the two knowledgeable anonymous reviewers for the comments that helped to improve the manuscript.

References

Adeyanju, D., 2016. Pragmatic features of political speeches in english by some prominent nigerian leaders. Journal of Political Discourse Analysis 2, 49–64.

Alluqmani, A., Shamir, L., 2018. Writing styles in different scientific disciplines: a data science approach. Scientometrics 115, 1071–1085.

Bonikowski, B., Gidron, N., 2015. The populist style in american politics: Presidential campaign discourse, 1952–1996. Social Forces 94, 1593–1621.

Boromisza-Habashi, D., 2010. How are political concepts 'essentially'contested? Language & communication 30, 276–284.

Cardie, C., Wilkerson, J., 2008. Text annotation for political science research.

Champion, C., 2000. Romans as bapbapoi: Three polybian speeches and the politics of cultural indeterminacy. Classical Philology 95, 425–444.

Chung, C.J., Park, H.W., 2010. Textual analysis of a political message: The inaugural addresses of two korean presidents. Social Science Information 49, 215–239.

Coleman, M., Liau, T.L., 1975. A computer readability formula designed for machine scoring. Journal of Applied Psychology 60, 283.

Diermeier, D., Godbout, J.F., Yu, B., Kaufmann, S., 2012. Language and ideology in congress. British Journal of Political Science 42, 31–55.

Dowding, K., Hindmoor, A., Iles, R., John, P., 2010. Policy agendas in australian politics: The governor-general's speeches, 1945–2008. Australian Journal of Political Science 45, 533–557.

Eshbaugh-Soha, M., 2010. The politics of presidential speeches, in: Congress & the Presidency, Taylor & Francis. pp. 1–21.

Fantham, E., 2003. The contexts and occasions of roman public rhetoric, in: Roman Eloquence. Routledge, pp. 102–116.

Gentzkow, M., Shapiro, J., Taddy, M., 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts, in: URL: https://data.stanford.edu/congress_text.

Gentzkow, M., Shapiro, J.M., Taddy, M., 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. Econometrica 87, 1307–1340.

Graber, D.A., Dunaway, J., 2017. Mass media and American politics. CQ Press.

Grimmer, J., Stewart, B.M., 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21, 267–297.

Laver, M., Benoit, K., Garry, J., 2003. Extracting policy positions from political texts using words as data. American Political Science Review 97, 311–331.

Mahdiyan, M., Rahbar, M., Hosseini-Maasoum, S.M., 2013. Applying critical discourse analysis in translation of political speeches and interviews. Academic Journal of Interdisciplinary Studies 2, 35.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit, in: Proceedings

of 52nd Annual Meeting of the Association for Computational Linguistics, pp. 55–60.

Pearson, K., Dancey, L., 2011. Elevating women's voices in congress: Speech participation in the house of representatives. Political Research Quarterly 64, 910–923.

Pepe, C., 2013. The genres of rhetorical speeches in Greek and Roman antiquity. Brill.

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R., 2006. An automated method of topic-coding legislative speech over time with application to the 105th-108th us senate, in: Midwest Political Science Association Meeting, pp. 1–61.

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R., 2010. How to analyze political attention with minimal assumptions and costs. American Journal of Political Science 54, 209–228.

Remer, G., 2008. Genres of political speech: Oratory and conversation, today and in antiquity. Language & Communication 28, 182–196.

Reyes, A., 2015. Building intimacy through linguistic choices, text structure and voices in political discourse. Language & Communication 43, 58–71.

Savoy, J., 2010. Lexical analysis of us political speeches. Journal of Quantitative Linguistics 17, 123–141.

Schaffner, C., 1996. Political speeches and discourse analysis. Current Issues in Language & Society 3, 201–204.

Scherer, S., Layher, G., Kane, J., Neumann, H., Campbell, N., 2012. An audiovisual political speech analysis incorporating eye-tracking and perception data., in: Conference on Language Resource and Evaluation, pp. 1114–1120.

Sensales, G., Areni, A., Giuliano, L., 2018. Pronouns and verbs as gender markers in italian parliamentary speeches: Intersecting gender, communication, and politics. Rassegna di Psicologia 34, 51–66.

Shamir, L., 2020. UDAT: Compound quantitative analysis of text using machine learning. Digital Scholarship in the Humanities , fqaa007.

Shamir, L., Diamond, D., Wallin, J., 2015. Leveraging pattern recognition consistency estimation for crowdsourcing data analysis. IEEE Transactions on Human-Machine Systems 46, 474–480.

Sim, Y., Acree, B.D., Gross, J.H., Smith, N.A., 2013. Measuring ideological proportions in political speeches, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 91–101.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank, in:

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642.

Thomas, M., Pang, B., Lee, L., 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 327–335.

Triadafilopoulos, T., 1999. Politics, speech, and the art of persuasion: Toward an aristotelian conception of the public sphere. The Journal of Politics 61, 741–757.

Wilkerson, J., Casas, A., 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. Annual Review of Political Science 20, 529–544.

Yu, B., 2013. Language and gender in congressional speech. Literary and Linguistic Computing 29, 118–132.

Yu, B., Kaufmann, S., Diermeier, D., 2008. Classifying party affiliation from political speech. Journal of Information Technology & Politics 5, 33–48.