US Congressional Speech Analysis

Charlotte Zhao

March 1, 2025

1 Stylistic Variation

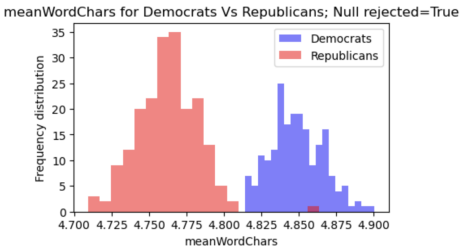
I examine and compare the complexity of congressional speeches within four groups of subjects: Republicans vs. Democrats, congressmen vs. congresswomen, Congress 107 vs. Congress 111, and US speeches vs. UK speeches. The null hypothesis is that the complexity of the speeches remains similar between the two subgroups within each category. I use mean word characters, a measure of average word length in a speech, and type-token ratio, a measure of word uniqueness, to evaluate the word complexity of speeches. Here are the equations for the two metrics.

$$\text{meanWordChars} = \frac{\text{number of characters}}{\text{number of words}}$$

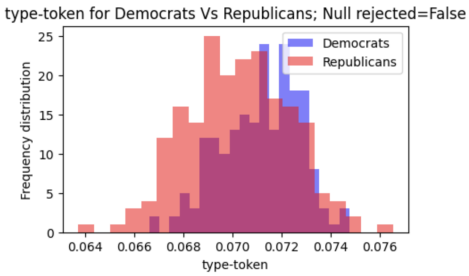
$$\text{type-token ratio} = \frac{\text{number of unique words}}{\text{corpus length}}$$

Applying bootstrapping, I calculate the 95%confidence interval of each metric for the four groups of subjects. For each group, I reject the null hypothesis if the confidence intervals of the two subgroups do not overlap. My findings are as follows.

- **Democrats vs Republicans**: Speeches from republican congresspersons are less textually complex than speeches from democrat congresspersons. While the confidence intervals for the type-token ratio from the two subgroups are not significantly different, I reject the hypothesis that the meanWordChars for speeches from democrats are the same as those from republicans.
- **Congressmen vs Congresswomen**: Speeches from congressmen are less textually complex than speeches from congresswomen.
- **Congress107 vs Congress111**: Textual complexity of speeches is decreasing over time.
- **US Congress vs Parliamentarians**: Textual complexity of U.S. congresspersons' speeches is lower than that of U.K. parliamentarians.

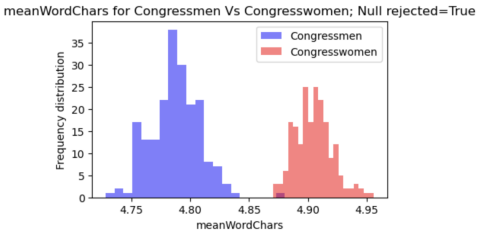


(a) meanWordChars

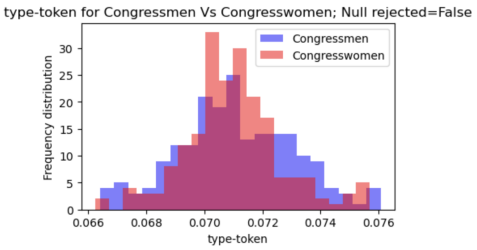


(b) Type-token Ratio

Figure 1: Bootstrap Distribution for Republicans vs Democrats

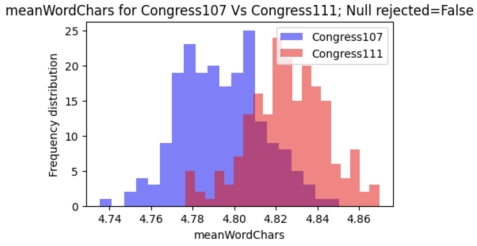


(a) meanWordChars

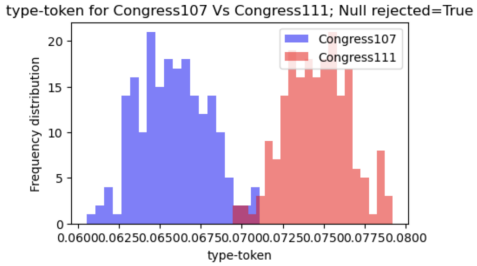


(b) Type-token Ratio

Figure 2: Bootstrap Distribution for Congressmen vs Congresswomen

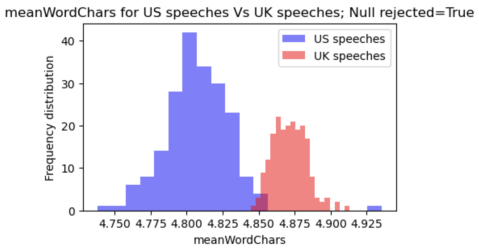


(a) meanWordChars

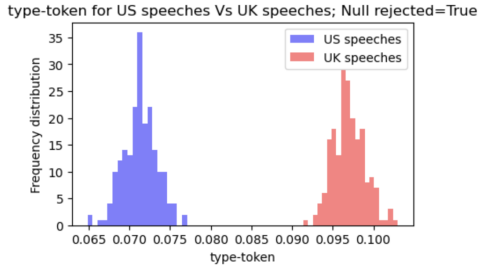


(b) Type-token Ratio

Figure 3: Bootstrap Distribution for Congress 107 vs Congress 111



(a) meanWordChars



(b) Type-token Ratio

Figure 4: Bootstrap Distribution for US Congress vs UK Parliament

2 Lexical Variation

I identified the exclusive words and examined overrepresented or underrepresented words within each comparing group. To quantify, I calculated the over/underrepresentation of words by taking the difference of PMI between 2 subgroups.

$$PMI(X, subgroupA) - PMI(X, subgroupB) = \log_2(Pr(X|A)) - \log_2(Pr(X|B))$$

Based on my bootstrap samples, the 95% confidence interval for the PMI difference is calculated for each group. A word is classified as overrepresented if its confidence interval is above 0. Conversely, a word is classified as underrepresented if its confidence interval is below 0. My results are below. The exclusive words are in the format (word, count).

- **Top Exclusive Words for Republicans**: ('obamacare.', 220), ('entity:', 181), ('hubbert', 121), ('compean.', 111), ('pantano', 110), ('hubberts', 109), ('bow:', 93),('eximbank', 59)

- **Top Exclusive Words for Democrats**: [('sonoma', 141), ('drinan', 123), ('apa', 98), ('accutane', 74), ('lgbt', 72), ('30somethings', 69), ('kalaupapa', 67)]

Table 1: Over/Under represented words based on bootstrap samples

Category	Over	Under
Republicans	'illegals, ' 'ida.'	none
Democrats	none	'ida.'

Table 2: Examples of Words with top PMI difference

Republicans	Democrats
smuggler	antichoice
steamroller	lgbt
abortionists	cardholder

- **Top Exclusive Words for Congressmen**: [('emp', 198), ('302(a)', 179), ('302(f)', 149), ('aggregates', 145), ('iowans', 125)]
- **Top Exclusive Words for Congresswomen**: [('ecologist.', 25), ('hipc', 19), ('brain-trust', 19), ('idalia', 18), ('ncb', 18)]
- Bootstrap samples return none are over-represented by Congressmen, and none are under-represented by Congresswomen.
- From **exclusive words from Congress 107**, some words have been abandoned in congressional speeches are: [('ins.', 297), ('tpa', 162), ('worldcom', 125), ('enrons', 101), ('traficant', 92), ('ergonomic', 88)]
- From **exclusive words from Congress 111**, some words have been introduced are [('obama', 4519), ('tarp', 1493), ('capandtrade', 817), ('congre', 484), ('barack', 467), ('acorn', 465)],('capandtax', 383)]

Table 3: Over/Under represented words based on bootstrap samples for US vs UK

Category	Over	Under
US Congress	program, theyre...	none
UK Parliament	however, governments...	none

Table 4: Examples of Words with top PMI difference for men vs women in congress

Congressmen	Congresswomen
socialist	ecologist
esoteric	weea
rescissions	wheatley

- **Top Exclusive Words for US Congress**: [('h.r.', 59842), ('dont', 40676), ('thats', 28244), ('program.', 25463), ('didnt', 10764)]
- **Top Exclusive Words for UK Parliament**: [('however,', 117660), ('that,', 100156), ("government's", 57824), ('and,', 39078), ('said,', 35900)]

3 Semantic Variation

Within each group, I trained a new Word2Vec model from each subgroup. Calculating the semantic variation of a word by its nearest neighbors from the resulted models, I identified the words with the greatest variation in meanings in each comparison subgroup.

- **Republicans vs Democrats**: Some words that differ most in meaning are "NSF", "liability", "taxation", "income", "hurricanes", and "tuition". The variation in meanings of these words may indicate the two parties' different opinions on healthcare, tax cuts, research, and education funding.

- **Congressmen vs Congresswomen**: Some words that differ most in meaning are "girl", "women", "impeachment", "African American," and "immigrants". We may find more words are gender-oriented, suggesting possible gender inequalities; congressmen may use the word "women" in a different literary context than congresswomen.
- **Congress 107 vs Congress 111**: Some words that differ most in meaning are "whistleblowers", "Iraqis", "defensive", "expensing", and "overregulation". This may reflect the shift in social issues, while "Iraqis" and "defensive" can be related to foreign policies, "expensing" and "overregulation" may be connected to economic expansions and depressions.
- **US Congress vs UK Parliament**: Some words that differ most in meaning are "infections", "mainland", "surcharge", "revenge".

4 Semantic Difference for Key Words

Based on my finetuned Word2Vec models, I calculated semantic variation score for key words representing critical social issues such as immigration, healthcare, gun reform, LGBTQ, and abortion. In the figures below, a semantic variaion score with a smaller magnitude indicates larger semantic difference between two comparison groups.

	Issue	Word	Semantic Variation Score
0	immigration	immigration	-57
1	immigration	border	-71
2	healthcare	health	-42
3	healthcare	insurance	-63
4	gun_reform	gun	-50
5	gun_reform	safety	-38
6	gun_reform	rights	-31
7	lgbtq	gay	-65
8	lgbtq	rights	-31
9	lgbtq	marriage	-58
10	abortion	abortion	-78
11	abortion	choice	-55

(a) Republicans vs Democrats

	Issue	Word	Semantic Variation Score
0	immigration	immigration	-45
1	immigration	asylum	-79
2	immigration	border	-65
3	healthcare	health	-52
4	healthcare	insurance	-58
5	gun_reform	gun	-46
6	gun_reform	safety	-35
7	gun_reform	rights	-25
8	lgbtq	rights	-25
9	lgbtq	marriage	-40
10	abortion	abortion	-56
11	abortion	choice	-53
12	abortion	parenthood	-59

(b) Congressmen vs Congresswomen

Figure 5: specific issues within each group

- From Republicans vs Democrats, gun reform has the highest semantic variation. Parenthood (-59) and Abortion (-56) also show a noticeable difference in the two parties' understanding of reproductive right.
- From Congressmen and Congresswomen, the interpretation of human right seems to exhibit the largest semantic variation. Social issues like immigration (-45) and abortion choice (-59) also show larger gender-based differences in language.