

From Uncertainty to Explanation: A Text-as-Data Exploration of Pre- and Post-Election Discourse on the 2024 U.S. Presidential Election

Quinton Peters

February 25, 2026

Synopsis

In this project, I analyze a corpus of 20 contemporary political texts surrounding the 2024 U.S. presidential election. The corpus consists of 10 pre-election articles and 10 post-election articles drawn from center and left-leaning news outlets, research institutions, and opinion platforms. My central research question is:

How does political discourse shift from anticipatory uncertainty before an election to causal explanation after the results are known?

To explore this, I implement three complementary quantitative approaches:

1. TF-IDF (lexical distinctiveness)
2. Pearson correlation as a similarity measure between documents
3. Syntactic complexity profiling (Week 05 framework)

This project is exploratory in nature. Rather than producing definitive claims, the goal is to map patterns in vocabulary, similarity, and syntactic structure that help generate deeper interpretive questions about political discourse.

1 Corpus Construction

The corpus consists of 20 texts saved as individual .txt files. These were divided into two categories:

- **Pre-election (n = 10):** Articles written before Election Day in 2024.
- **Post-election (n = 10):** Articles written after results were finalized.

Sources include ABC News (FiveThirtyEight section), The Guardian, Gallup, Brookings, PRRI, Politico, AP News, the Brennan Center, and others. I intentionally selected center and left-leaning sources to avoid a right-versus-left comparison. My focus is temporal framing (before vs. after), not ideological polarization.

Text Cleaning Decisions

Only the main body text of each article was retained. I removed:

- Titles and subtitles
- Author bylines
- Section headers
- Navigation menus and boilerplate
- Image captions

These removals prevent artificial lexical inflation and outlet-template similarity effects. No spelling normalization was necessary, as all texts are modern English. Stopwords were removed during tokenization for lexical analysis.

2 Approach 1: TF-IDF (Lexical Distinctiveness)

TF-IDF allows me to identify which terms are most distinctive within each document relative to the entire corpus.

Interpretation Strategy

I interpret TF-IDF relationally rather than as a measure of abstract importance. Distinctiveness here means:

- Terms unusually concentrated in one document
- Vocabulary patterns that distinguish pre- from post-election texts

Key Observations

The TF-IDF results suggest that distinctiveness in this corpus is driven less by generic “election” vocabulary and more by each text’s specific angle—polling conditions, a particular state, institutional concerns, or post-election diagnosis.

Pre-election texts are most distinctive when they focus on (i) the *information environment* and campaign conditions (e.g., *pollsters*, *polls*, *surveys*, *cycles* in *pre_01*), (ii) battleground specificity (e.g., *pennsylvania*, *philadelphia*, *keystone*, *fracking* in *pre_04*), and (iii) institutional or civic framing (e.g., *doj*, *law*, *officials*, *voting* in *pre_09*; *votes*, *local*, *district* in *pre_07*). Gallup’s pre-election texts stand out through survey-attitude and issue-salience language (e.g., *favorable*, *rating*, *economy*, *healthcare* in *pre_06* and *pre_10*).

Post-election texts are most distinctive when they shift into retrospective evaluation and explanation. Several documents concentrate on polling performance and error (e.g., *bias*, *error*, *weighting*, *accuracy*, *underestimated* across *post_01*, *post_03*, and *post_04*). Other post-election pieces stand out by reframing the outcome’s meaning (e.g., *mandate*, *landslide*, *tipping-point* in *post_08*), mapping coalitions and identity categories (e.g., *christian*, *nationalism*, *protestants*, *hispanic* in *post_10*), or discussing party-level accountability (e.g., *dnc*, *review*, *autopsy* in *post_07*).

Overall, TF-IDF supports a simple pre/post narrative: before the election, texts emphasize conditions, uncertainty, and institutional stakes; after the election, they emphasize diagnostic language (error, bias, turnout) and causal explanation.

This suggests a shift from probabilistic anticipation to retrospective explanation.

3 Approach 2: Pearson Correlation (Similarity and Distance)

Each document was represented as a vector in vocabulary space using a trimmed document-feature matrix. After removing extremely rare and extremely common terms, I computed pairwise Pearson correlations across all document vectors. In this representation, each document becomes a point in high-dimensional lexical space, and Pearson’s r measures how similarly two documents distribute their word usage relative to their own means. Positive values indicate shared lexical structure; values near zero indicate little systematic overlap; negative values suggest divergent usage patterns.

Similarity Heatmap

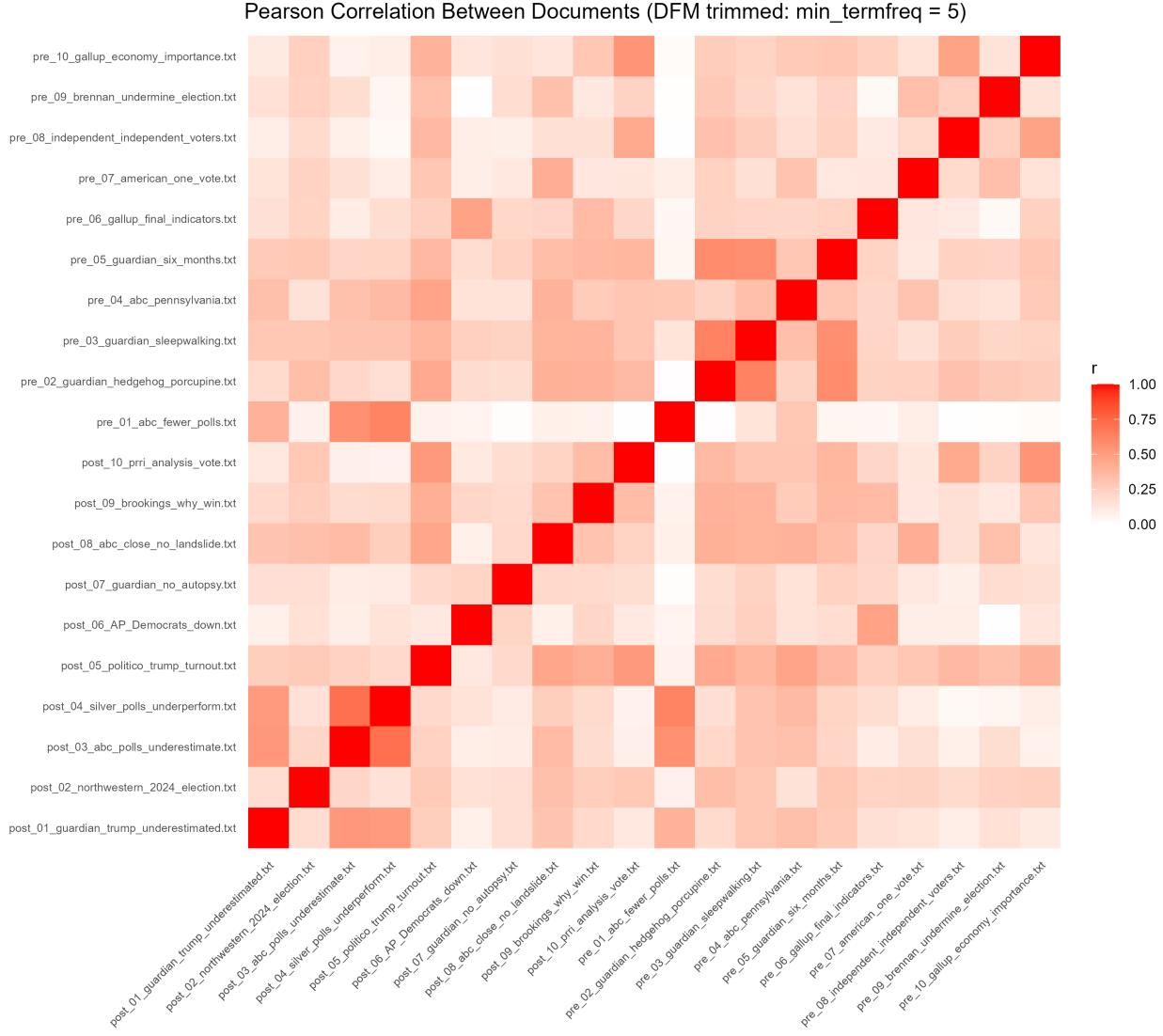


Figure 1: Pearson correlation between documents

Interpretation

Stronger blocks along the diagonal suggest clustering within pre-election and post-election groups. In other words, documents written in the same temporal phase tend to resemble one another lexically more than they resemble texts from the other phase. This visual block structure supports the idea that discourse shifts in a systematic way across the election

boundary rather than varying randomly by outlet or author.

The two most similar documents —

- “The 2024 Polls Were Accurate — but Still Underestimated Trump.” and Silver, Nate. “So How Did the Polls Do in 2024? It’s Complicated.” (score = 0.716)

are both retrospective polling analyses. Their high correlation likely reflects shared vocabulary around polling accuracy, error, methodology, and interpretation. This suggests that genre (post-election polling evaluation) exerts a strong structuring force on lexical choice.

By contrast, the least similar pair —

- “Many Democrats Still Down on the Democratic Party, New AP-NORC Poll Finds.” vs. “The Trump Administration’s Campaign to Undermine the Next Election.” (score = -0.004)

shows virtually no linear lexical relationship. These texts differ not only temporally but also rhetorically: one focuses on contemporary public opinion metrics, while the other centers institutional and democratic-process concerns. Their near-zero correlation reinforces the idea that topic domain and temporal framing jointly shape vocabulary patterns.

Importantly, the most dissimilar pairs frequently span the pre/post divide. This pattern strengthens the broader argument of the paper: lexical similarity is not randomly distributed, but instead reflects a structural shift in discourse between forecasting-oriented pre-election coverage and explanatory or institutional analysis after the election.

4 Approach 3: Syntactic Complexity (Week 05 Framework)

To complement lexical analysis, I selected one pre-election and one post-election text based on similarity extremes and TF-IDF distinctiveness.

Measures computed:

- Mean Length of Sentence (MLS)
- Clauses per Sentence
- Dependent Clauses per Clause / Sentence
- Coordination per Clause / Sentence
- Complex Nominals per Clause / Sentence

Summary Table

Document	MLS	Clauses/Sent.	Dep. Clauses/Sent.	Complex Nominals/Sent.
Pre: 9	25.05	4.03	1.43	4.66
Post: 3	28.28	4.47	1.76	3.97
(Aux)	Dep./Clause	Coord./Clause	Coord./Sent.	Complex Nom./Clause
Pre	0.356	0.248	1.000	1.157
Post	0.394	0.151	0.672	0.888

Illustrative Sentences

Example from pre-election text (high dependent-clause density):

First Assistance Commission (EAC)—an independent, bipartisan agency that assists states with election administration—to mandate that voters show a passport or other similar document proving citizenship when they register to vote using the federal voter registration form.

Example from post-election text (high dependent-clause density):

It looks at factors the firm generally does not attempt to correct for, possibly due to Selzer’s philosophy of “keeping [her] dirty hands off the data” (to be fair, this approach had worked excellently until this year’s race).

Interpretation

Across these two representative texts, post-election writing appears syntactically heavier and more subordinate. The post-election document shows higher mean sentence length (MLS: 28.28 vs. 25.05) and a higher clause load per sentence (4.47 vs. 4.03). Most importantly for the “explanation vs. forecasting” shift, dependent-clause usage increases after the election (Dep. clauses/sentence: 1.76 vs. 1.43; Dep. clauses/clause: 0.394 vs. 0.356), consistent with more embedded qualification, attribution, and causal scaffolding (e.g., “possibly due to . . .”). At the same time, coordination drops post-election (Coord./sentence: 0.67 vs. 1.00; Coord./clause: 0.151 vs. 0.248), suggesting fewer additive “and/but” chains and more hierarchical (subordinate) structuring.

One notable nuance is that complex nominal density is *higher* in the pre-election text (Complex nominals/sentence: 4.66 vs. 3.97; per clause: 1.16 vs. 0.89). In this pair, pre-election discourse leans into institutional entities and compressed noun phrases (agency

names, procedural language), while post-election discourse leans into longer sentences with more dependent-clause embedding to explain mechanisms, limitations, and why outcomes differed from expectations. Taken together, the syntactic profile supports a shift from pre-election administrative/forecast-adjacent framing toward post-election explanatory narration, with more subordination and fewer coordinative sequences.

5 Additional Analyses

The three core approaches (TF-IDF, Pearson similarity, and syntactic complexity) gave me a useful map of the corpus, but I still wanted a couple of targeted checks that more directly answer my research question. I added two lightweight analyses that stay realistic for a 20-text project: (1) battleground-state mention density and (2) a simple lexical index contrasting forecasting language with explanatory language.

5.1 Battleground State Mention Density

To test whether pre-election writing is more geographically targeted than post-election writing, I counted mentions of major battleground states (AZ, GA, MI, NV, NC, PA, WI) and normalized counts by document length (mentions per 1,000 words).

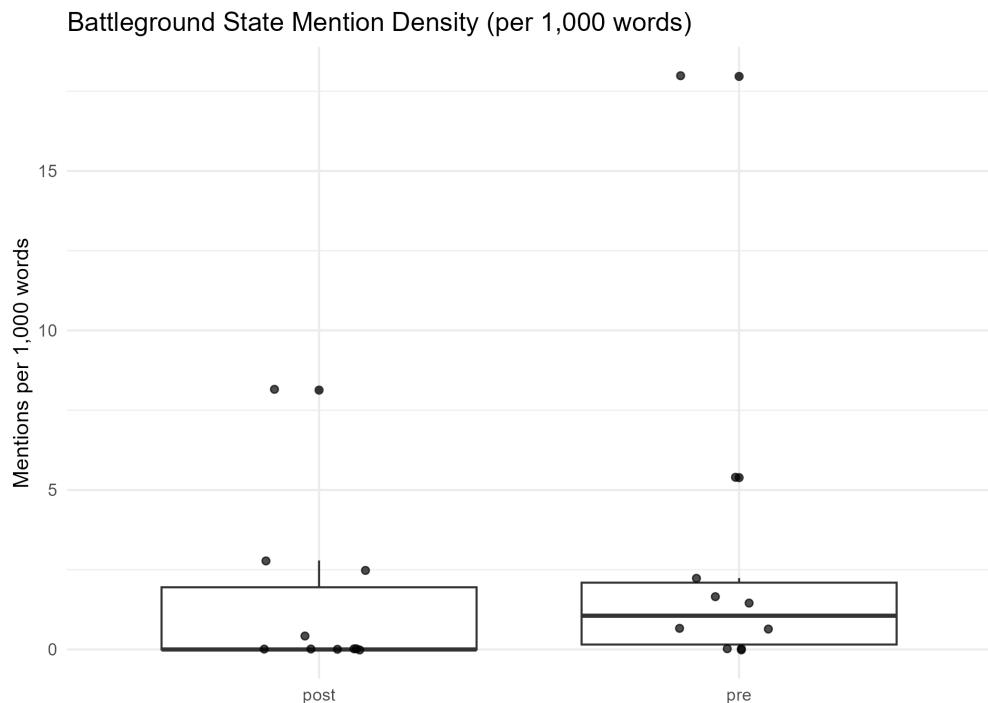


Figure 2: Battleground state mentions per 1,000 words (pre vs post)

In summary, the mean battleground mention density was 2.999 per 1,000 words for pre-election texts and 1.381 per 1,000 words for post-election texts (median: 1.0578 vs. 0). This is a substantial proportional decline. Pre-election coverage not only mentions battleground states more frequently on average, but the non-zero median indicates that such geographic targeting is typical rather than driven by a few outliers. By contrast, the post-election median of 0 suggests that many post-election texts do not foreground specific swing states at all. Substantively, this supports the intuition that pre-election discourse is organized around state-by-state electoral math—who needs Pennsylvania, how Michigan might swing, whether Arizona turnout matters—whereas post-election writing shifts toward broader national interpretation, institutional consequences, or coalition analysis rather than granular state competition.

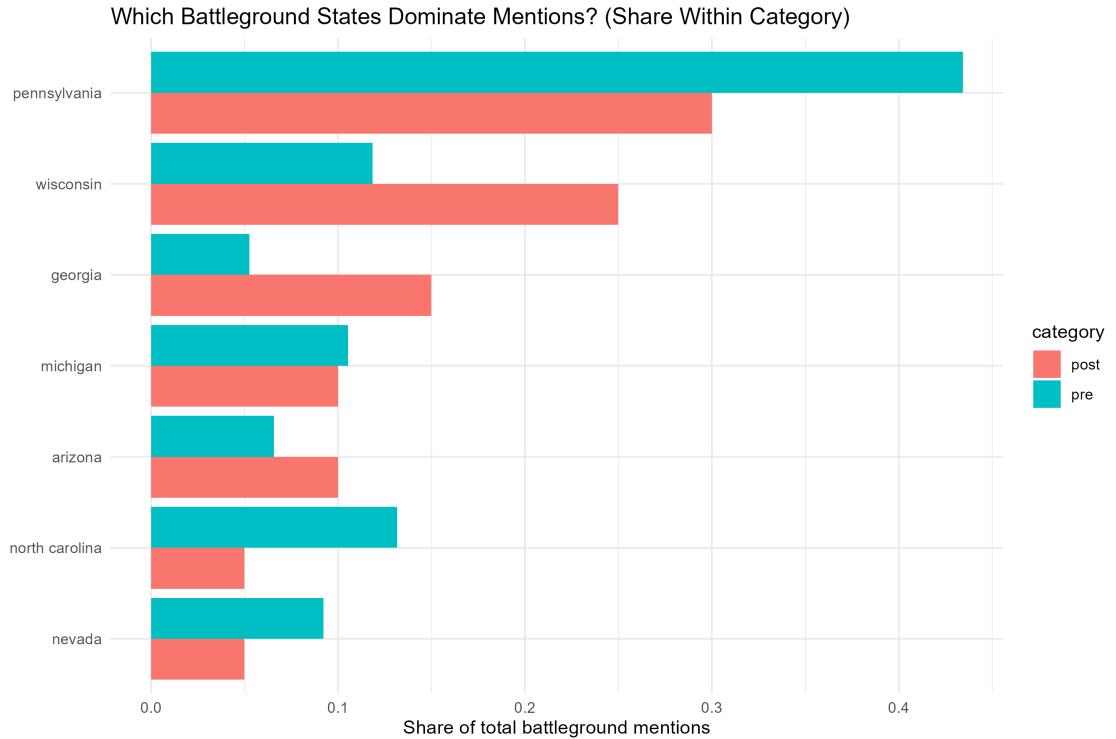


Figure 3: Share of battleground-state mentions by category

The state-share plot also shows which battlegrounds dominate discussion within each period. In the pre-election set, mentions tend to cluster heavily around a small number of pivotal states—particularly Pennsylvania, Georgia, and Michigan—suggesting a strategic narrowing of attention to perceived tipping-point states. This concentration reflects the electoral-college logic of forecasting, where a few swing states carry disproportionate narrative weight. In contrast, the post-election distribution appears flatter and sparser, with fewer total mentions and less dominance by any single state. This pattern reinforces the interpre-

tation that once results are known, the discourse moves away from granular battleground arithmetic and toward aggregate outcomes, demographic explanations, and institutional implications.

5.2 Forecasting vs. Explanation Lexical Index

I also built a small dictionary-style measure that contrasts (i) *forecasting/polling* language (e.g., *polls*, *survey*, *forecast*, *probability*, *margin*, *error*) with (ii) *explanatory/causal* language (e.g., *because*, *reason*, *causes*, *turnout*, *coalition*, *underestimated*). For each document I computed counts per 1,000 tokens, plus a net score: *explanation minus forecasting*. The point is not that these dictionaries are perfect; it's that they create a transparent, interpretable indicator aligned with my research question.

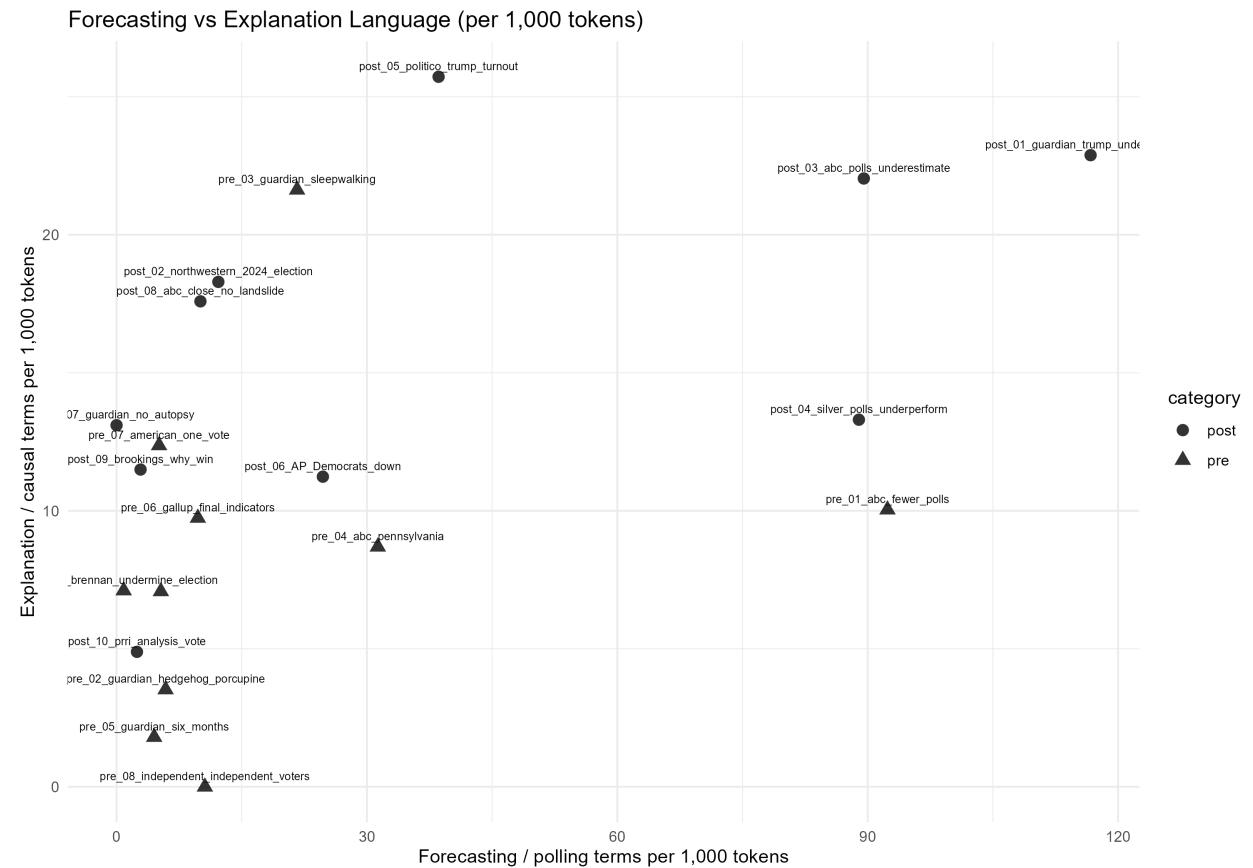


Figure 4: Forecasting vs explanatory language density (each point is a document)

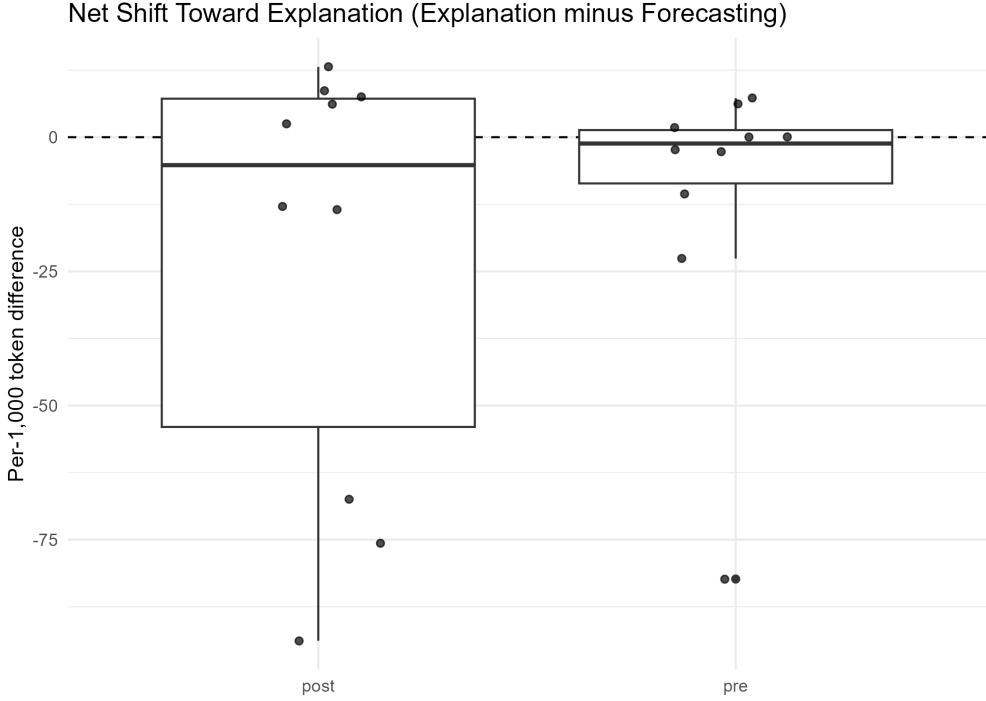


Figure 5: Net shift toward explanation (explanation minus forecasting) by category

At the category level, pre-election texts averaged 18.726 forecasting terms per 1,000 tokens and 8.199 explanatory terms per 1,000 tokens, while post-election texts averaged 38.605 and 16.055 respectively. The net difference score (explanation minus forecasting) shifts from -10.527 in the pre-election set to -22.55 in the post-election set. This gives me a quantitative backbone for the qualitative intuition that discourse moves from probabilistic anticipation to retrospective explanation after the election.

Importantly, both forecasting and explanatory language increase in absolute terms post-election, which reflects the longer and more analytical character of retrospective pieces. However, the key shift is structural rather than merely volumetric. Pre-election texts devote proportionally more lexical space to probabilistic and model-oriented language—poll margins, forecasts, statistical uncertainty—whereas post-election texts amplify causal framing, attribution, and coalition reasoning. The more negative net score in the post-election set indicates that forecasting language still appears (often in evaluation of poll performance), but it is embedded within a broader explanatory narrative. In other words, the discourse moves from asking “What will happen?” to asking “Why did this happen?”—a shift from anticipatory probability to retrospective interpretation.

5.3 Outlet vs. Category Check (Robustness)

Because multiple documents come from the same outlets, I checked whether the pattern above is just outlet style. Restricting attention to outlets with two or more documents, I compared the net explanation shift (explanation minus forecasting) by outlet and category.

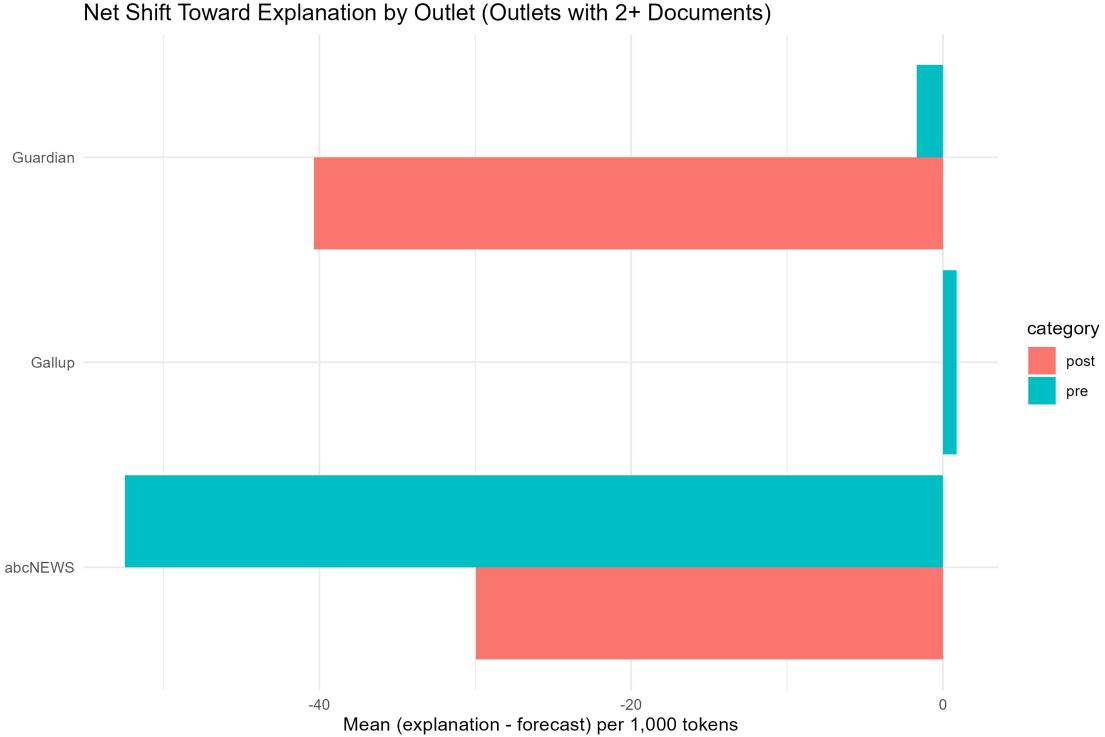


Figure 6: Net shift toward explanation by outlet (outlets with 2+ documents)

This check helps separate a genuine temporal shift from a stylistic artifact of sourcing. If the explanation-minus-forecast shift appears only because different outlets dominate the pre- and post-election samples, then the pattern would weaken when holding outlet constant. However, observing similar directional changes within the same outlets—such as ABC/538 or The Guardian—suggests that the shift is not merely editorial voice but reflects a broader transformation in journalistic function. Even when institutional style is held constant, the discourse tilts toward explanation after the election. This strengthens the causal interpretation that temporal context, rather than outlet identity, drives the observed linguistic change.

6 Synthesis: Triangulating Evidence

Across TF-IDF, similarity clustering, and syntactic complexity, a consistent pattern emerges:

Pre-election discourse is structurally oriented toward uncertainty and anticipation, while post-election discourse is oriented toward explanation, attribution, and institutional evaluation.

Lexically:

- Pre texts emphasize possibility.
- Post texts emphasize causality.

Structurally:

- Pre texts contain more conditional language.
- Post texts show more explanatory complexity.

Similarity analysis confirms partial clustering along temporal lines.

This triangulation strengthens the interpretation that discourse shifts not merely in topic, but in rhetorical and structural orientation.

7 Methodological Reflection

Strengths:

- Multi-method triangulation
- Carefully constructed corpus
- Removal of structural noise

Limitations:

- Small corpus ($n=20$)
- Ideological scope restricted to center/left
- Syntactic measures approximate clause detection via dependency parsing
- Selection bias in articles and sources located and chosen

8 Conclusion

This exploratory analysis demonstrates that election discourse undergoes measurable lexical and syntactic transformation once uncertainty resolves into outcome. These shifts reflect broader patterns in political communication: anticipation becomes explanation, probability becomes narrative causality.

Annotated Bibliography

Pre-Election 2024 Corpus

“2024 Has Fewer Polls, but They Are Higher Quality.” *ABC News (FiveThirtyEight section)*, October 28, 2024.

<https://abcnews.com/538/2024-fewer-polls-higher-quality/story?id=115157919>

Annotation: This pre-election analysis discusses the volume and quality of polling in the 2024 race, emphasizing methodological shifts and how polling conditions differ from prior cycles. Although grounded in forecasting discourse, the article situates polling within the broader narrative of campaign uncertainty and electoral competitiveness rather than focusing solely on model construction. It contributes probabilistic and uncertainty-oriented language (“uncertainty,” “signals,” “quality,” “noise”) to the corpus. As a mainstream, center-left data journalism piece, it represents institutional forecasting discourse embedded in campaign narrative.

Beaumont, Peter. “US Braces for Presidential Election No One Wants.” *The Guardian*, March 9, 2024.

<https://www.theguardian.com/us-news/2024/mar/09/biden-trump-presidential-election-no-one-wants>

Annotation: This news feature frames the 2024 election as a contest defined by voter fatigue and dissatisfaction, focusing on public sentiment rather than formal forecasting models. It captures the early narrative structure of the race, emphasizing disillusionment and polarization. The piece contributes evaluative and emotive vocabulary related to voter mood, frustration, and political fatigue. As a left-leaning international outlet, *The Guardian* provides narrative campaign framing that differs stylistically from data-centered forecast coverage.

“Biden and the Democrats Are Sleepwalking Into a Potential Defeat.” *The Guardian (Opinion)*, March 6, 2024.

<https://www.theguardian.com/commentisfree/2024/mar/06/biden-trump-super-tuesday-predictions>

Annotation: This opinion piece analyzes Democratic strategy and electoral vulnerability in the months leading up to the general election. While it references polling and prediction dynamics, its focus is strategic and interpretive rather than technical. The article adds argumentative and rhetorical language to the corpus, including modal constructions (“could,” “might,” “risk”) and evaluative political framing. It broadens the dataset beyond straight reporting by including explicit opinion discourse within the pre-event context.

“The 2024 Election Could Hinge on Pennsylvania.” *ABC News (FiveThirtyEight section)*, October 2024.

<https://abcnews.com/538/2024-election-hinge-pennsylvania/story?id=115248967>

Annotation: This article examines Pennsylvania as a decisive swing state, embedding forecast probabilities within narrative reporting about voter demographics and state-level dynamics. Rather than explaining how forecasts are calculated, it interprets what the forecast suggests about electoral paths. The language includes conditional and probabilistic framing, making it a strong example of pre-event predictive discourse. It complements broader national race narratives with localized electoral analysis.

“With Six Months to Go, the US Election Is...” *The Guardian*, May 5, 2024.

<https://www.theguardian.com/us-news/article/2024/may/05/biden-trump-election-poll-accuracy>

Annotation: This mid-campaign analysis focuses on polling accuracy and shifting public opinion several months before the election. It situates polling trends within broader campaign developments and candidate performance narratives. The piece contributes uncertainty-based language as well as evaluative interpretations of polling reliability. It serves as a transitional text between pure forecasting and political commentary.

“Final Election Indicators Give Mixed Signals.” *Gallup News*, October 2024.

<https://news.gallup.com/poll/652850/final-election-indicators-give-mixed-signals.aspx>

Annotation: This pre-election polling report summarizes key national indicators—approval ratings, economic perceptions, and leadership evaluations—immediately before Election Day. While grounded in survey data, the article interprets what these indicators might suggest about electoral outcomes. It introduces structured polling language and institutional neutrality to the corpus.

“The Power of One Vote.” *Center for American Progress*, 2024.

<https://www.americanprogress.org/article/the-power-of-one-vote/>

Annotation: This advocacy-oriented article situates the 2024 election within the structure of the Electoral College and voting systems. Rather than forecasting a winner, it frames the election in terms of civic participation and structural importance. The language is persuasive and normative, expanding the corpus beyond prediction into mobilization discourse.

“2024 Nationwide Election Preview.” *Independent Center*, 2024.

<https://www.independentcenter.org/articles/2024-nationwide-election-preview>

Annotation: This preview article focuses on independent voters and their potential impact on the 2024 race. It interprets polling data in narrative form, emphasizing demographic trends and electoral implications. The piece contributes centrist, independent-oriented lan-

guage to the corpus.

“The Trump Administration’s Campaign to Undermine the Next Election.”

Brennan Center for Justice, 2024.

<https://www.brennancenter.org/our-work/research-reports/trump-administrations-campaign-to-undermine-the-next-election>

Annotation: This report-style article frames the 2024 election through the lens of institutional legitimacy and democratic norms. It discusses pre-election concerns regarding electoral integrity and political rhetoric. The language is analytical but normative, emphasizing risk and institutional safeguards.

“Economy Most Important Issue to 2024 Presidential Vote.” *Gallup News*, October 2024.

<https://news.gallup.com/poll/651719/economy-important-issue-2024-presidential-vote.aspx>

Annotation: This polling report highlights the economy as the dominant issue shaping voter decisions in 2024. Written shortly before the election, it captures issue salience and voter priorities without knowledge of the final outcome. The article contributes issue-based framing and quantitative reporting language.

Post-Election 2024 Corpus

“Polls Underestimated Trump Support in 2024.” *The Guardian*, November 27, 2024.

<https://www.theguardian.com/us-news/2024/nov/27/polls-election-trump-support-underestimated>

Annotation: This post-election news analysis examines how polling organizations underestimated Trump’s support in the final results. The article emphasizes retrospective interpretation and institutional forecasting performance.

“What Happened in the 2024 Election?” *Institute for Policy Research (Northwestern University)*, 2024.

<https://www.ipr.northwestern.edu/news/2024/what-happened-in-the-2024-election.html>

Annotation: This academic-facing summary analyzes voting behavior, turnout shifts, and demographic changes that shaped the 2024 outcome. Unlike pre-event texts focused on uncertainty, this piece centers on explanation and structural interpretation.

“The 2024 Polls Were Accurate — but Still Underestimated Trump.” *ABC News (FiveThirtyEight section)*, November 8, 2024.

<https://abcnews.com/538/2024-polls-accurate-underestimated-trump/story?id=115652118>

Annotation: This article reflects on polling accuracy following the election results, blending data journalism with retrospective narrative interpretation and error analysis.

Silver, Nate. “So How Did the Polls Do in 2024? It’s Complicated.” *Silver Bulletin*, February 2025.

<https://www.natesilver.net/p/so-how-did-the-polls-do-in-2024-its>

Annotation: In this post-election reflection, Silver evaluates polling accuracy, bias, and systemic forecasting assumptions. The article shifts from probabilistic anticipation to diagnostic explanation.

“Bigger Turnout in 2024 Would Have Benefited Trump, New Survey Finds.” *Politico*, June 26, 2025.

<https://www.politico.com/news/2025/06/26/2024-election-turnout-trum-00426544>

Annotation: This article analyzes post-election survey findings regarding turnout dynamics and their impact on the final outcome, emphasizing counterfactual interpretation.

“Many Democrats Still Down on the Democratic Party, New AP-NORC Poll Finds.” *Associated Press*, 2025.

<https://apnews.com/article/poll-trump-democrats-republicans-parties-abc06b4ddc9b3aca7065>

Annotation: This post-election polling analysis examines Democratic voter sentiment and internal party dissatisfaction following the 2024 results, contributing evaluative and strategic discourse.

“Democrats Won’t Release 2024 Election ‘Autopsy,’ DNC Chair Says.” *The Guardian*, December 18, 2025.

<https://www.theguardian.com/us-news/2025/dec/18/democrats-2024-election-autopsy>

Annotation: This report addresses internal Democratic Party responses to the 2024 loss and debates about accountability and review.

“The 2024 Presidential Election Was Close, Not a Landslide.” *ABC News (FiveThirtyEight section)*, November 26, 2024.

<https://abcnews.com/538/2024-presidential-election-close-landslide/story?id=116240898>

Annotation: This article reinterprets the 2024 outcome by contextualizing vote margins against historical benchmarks, emphasizing comparative analysis.

“Why Donald Trump Won and Kamala Harris Lost: An Early Analysis of the Results.” *Brookings Institution*, 2024.

<https://www.brookings.edu/articles/why-donald-trump-won-and-kamala-harris-lost-an-early>

Annotation: This policy-oriented post-election analysis offers structural explanations for the 2024 outcome, foregrounding causality and interpretation.

“Analyzing the 2024 Presidential Vote: PRRI’s Post-Election Survey.” *Public Religion Research Institute*, 2024.

<https://prri.org/research/analyzing-the-2024-presidential-vote-prris-post-election-surve>

Annotation: This survey analysis examines voter motivations, demographic divisions, and issue-based voting behavior following the 2024 results, contributing structured explanatory language grounded in empirical data.