CourtShadow: Quantifying Courtroom Linguistic Environments with Interpretable Logistic Regression

Diya Karan
Project Website: https://courtshadow.org

1 Introduction

Racial disparities in the United States (U.S.) criminal justice system manifest not only in sentencing outcomes, but also in the linguistic patterns underlying courtroom interactions (Conley & O'Barr, 1990; Gibbons, 2003). Interruptions, stance markers, framing, and topic emphasis can influence how a defendant or case is perceived. Despite these implications, courts have historically treated such linguistic variation as legally irrelevant. In the landmark case *McCleskey v. Kemp* (1987), even large-scale statistical evidence of racial disparities in sentencing was deemed insufficient.

Recently enacted Racial Justice Acts (RJAs) in states such as California and North Carolina now explicitly authorize the use of systemic statistical evidence. This allows for the language of courtroom proceedings to be collectively interpreted as an institutional environment. Consequently, these linguistic dynamics may be analyzed for application to real world legislation such as RJAs. Yet, few large-scale studies analyzing these patterns have been performed.

CourtShadow addresses this gap by building an interpretable, statistically-grounded framework that examines the language dynamics of trial transcripts. In particular, it detects whether the linguistic environment of a transcript resembles that of Group A–coded or Group B–coded cases (two anonymized labels representing different defendant groups). The project does *not* predict race or assess individual prejudice. Instead, it examines whether the recorded structure, framing, pronoun usage, and criminal-legal topics of a transcript reflect systematic discursive differences between defendant groups.

This project questions whether the linguistic environments of criminal trial transcripts differ systematically across defendant groups. Additionally, it investigates the ways in which these differences (if any) can be quantified through interpretable logistic regression.

2 Methods

Data and segmentation

I collected 42 publicly-available criminal trial transcripts (comprising 1,915 total text segments). Due to widespread variation in formatting and speaker structure across transcripts, each document cleaned and organized via a text pre-processing pipeline. Based on speaker turns and approximate length, these transcripts were then divided into short, linguistically-meaningful segments. Each segment inherits the group label of its parent case.

Feature families

Every segment is represented using **38 transparent features** organized into four families:

- **Structure** (length, sentence count, question density);
- **Discursive Framing** (politeness, harshness, mitigation, certainty);
- **Pronouns & Voice** (first- and second-person usage);
- **Criminal-Legal Topics** (violence, policing, drugs, financial crime, harm).

Feature definitions, heuristics, and full keyword lists appear in Appendix A.

Logistic regression model

Each segment is a pair (x_i, y_i), where $x_i \in \mathbb{R}^{38}$ is the feature vector and $y_i \in \{0, 1\}$ is the inherited case label. The model treats labels as Bernoulli random variables:

$$y_i \sim \text{Bernoulli}(p_i), \qquad p_i = \sigma(\theta^\top x_i) = \frac{1}{1 + e^{-\theta^\top x_i}}.$$

The negative log-likelihood (cross-entropy loss) is

$$J(\theta) = -\sum_{i=1}^{n} \big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \big].$$

To avoid overfitting and stabilize explanations, I use an L2-regularized objective:

$$J_{\text{L2}}(\theta) = J(\theta) + \lambda \sum_j \theta_j^2.$$

A moderate penalty ($\lambda \approx 10^{-3}$) produces the best held-out performance.

Case-level aggregation

CourtShadow's goal is to characterize *case-level* linguistic environments. Segment-level probabilities are averaged into a **Linguistic Environment Score** (LES):

$$\bar{p}_{\text{case}} = \frac{1}{m} \sum_{j=1}^{m} p_j,$$

where m is the number of segments in a case. Due to the linear nature of logistic regression, feature contributions can be decomposed cleanly as

$$\theta^\top x = \sum_{k \in \text{family}} \theta_k x_k,$$

This supports deep family-level interpretability, distinguishing the CourtShadow tool from black-box classification algorithms (Appendix B).

3 Results

Model performance

On an 8-case held-out test set, CourtShadow achieves:

- 87.5% chunk-level accuracy;
- Receiver Operating Characteristic (ROC) chunk-level Area Under the Curve (AUC) = 0.95;
- Stable, well-behaved calibration curves.

The model substantially outperforms majority and random baselines. This indicates that the linguistic environment contains non-trivial signal even with simple, interpretable features.

Feature-Family Behavior

Aggregating weights by their feature family reveals consistent patterns:

- **Structure**: Long, dense, and highly interrogative segments behave differently across groups;
- **Framing**: Harshness and strong certainty markers raise LES, while mitigating or polite framing pushes it downwards;
- **Pronouns**: Second-person emphasis ("you") versus first-person narrative ("I") signals different institutional stances;
- **Topics**: Weapons, drugs, policing, and harm are discussed more frequently in high-LES cases.

These patterns emerge only in aggregate; no single keyword drives the model.

Case Comparisons

Two test cases can be compared and examined to illustrate how LES captures institutional patterns. The first transcript is from the Wilmington Ten case, in which ten civil rights activists were wrongfully convicted (U.S. Department of Justice, n.d.) This transcript receives an LES score in the 0.90–0.98 range, with contributions concentrated in harsh framing, high-stakes topic indicators, and dense adversarial questioning. By contrast, a white-defendant case such as *United States v. Cohen* receives a substantially lower score of approximately 0.32. This score was driven by neutral framing, different structural patterns, and a distinct mix of criminal-legal topics. These contrasts demonstrate that CourtShadow captures the *texture* of the linguistic environment as opposed to any speaker's intent or identity.

4 Discussion

CourtShadow provides evidence that courtroom linguistic environments contain structured, statistically-detectable differences across defendant groups. These findings align with previous sociolinguistic research on courtroom dynamics. For instance, prior works have revealed the ways in which this language is correlated with power, credibility, and perceived threat in regard to race (Conley & O'Barr, 1990; Johnson, 2020). The project extends these insights by offering an interpretable, transcript-level method for quantifying such patterns.

The work also speaks directly to emerging RJA frameworks, which allow defendants to present system-level linguistic or behavioral evidence of discrimination. Although CourtShadow is not a legal diagnostic, it evinces the ways in which statistical tools can reveal patterns in institutional language.

Limitations of this tool include its limited dataset size (42 transcripts), imperfect or incomplete documents, and the confounding role of case type, witness composition, or attorney strategy. As in any observational analysis, results are correlational rather than causal. Still, these findings suggest that courtroom language, which was previously unmeasured, constitutes a meaningful dimension of criminal proceedings. The observed patterns indicate that these linguistic environments merit greater attention in research and policy.

5 Conclusion

CourtShadow reveals that courtroom linguistic environments are systematically different across Group A– and Group B–coded cases. In particular, it demonstrates that such differences can indeed be quantified via a transparent Bernoulli–logistic model with interpretable feature families. While the model does not infer intent or recommend outcomes, it offers a structured way to examine how institutional language may reflect broader inequities.

Future work may incorporate larger corpora, richer linguistic features, and more detailed speaker-role analysis. Appendices provide full feature definitions, pre-processing details, derivations, and case-level interpretability figures.

Appendices Note

Complete appendices, additional plots, and full methodological derivations are hosted at: https://courtshadow.org.

References

Conley, J. M., & O'Barr, W. M. (1990). *Rules versus Relationships*. University of Chicago Press.

Cotterill, J. (2003). *Language and Power in Court*. Palgrave.

Eberhardt, J. L. (2019). *Biased*. Viking.

Gibbons, J. (2003). *Forensic Linguistics*. Blackwell.

Johnson, E. (2020). Language and the administration of justice. In *Oxford Handbook of Language and Law*.

Maynard, D. W. (1984). *Inside Plea Bargaining*. University of Chicago Press.

U.S. Department of Justice, Civil Rights Division. (n.d.). *Transcripts: State of North Carolina v. Benjamin Franklin Chavis, Marvin Patrick, Connie Tyndall, et al.* Retrieved from https://www.justice.gov/crt/transcripts-casestate-north-carolina-v-benjamin-franklin-chavis-marvin-patrick-connie-tyndall-et

California Racial Justice Act, Cal. Penal Code §§ 745–747 (2020).

North Carolina Racial Justice Act, N.C. Gen. Stat. §§ 15A-2010–2012 (2009/2020).