Stylometric Profiling and Speaker Classification in Formula 1 Press-Conference Language

Jiashen Wang

2025-12-10

Abstract

This study examines how Formula 1 drivers use language in official FIA press-conference transcripts from 2022–2025. First, I analyze token counts for the top seven drivers to track changes in media presence across seasons. Second, I compare linguistic style using hierarchical clustering and PCA based on part-of-speech distributions, revealing groupings such as similarities among world champions and among rookie drivers. Finally, I apply a LASSO-regularized logistic regression model to predict whether a transcript was spoken by Max Verstappen, achieving strong validation and test accuracy. Together, these findings show that linguistic behavior in press conferences reflects competitive dynamics in Formula 1 and enables reliable speaker identification using linguistic features.

Introduction

Formula 1 is a sport defined not only by engineering and race performance, but also by the narratives drivers construct through language. Press-conference transcripts offer a unique opportunity to observe how drivers present themselves, communicate pressure, and reflect competitive dynamics both within teams and across seasons. Understanding how language varies across drivers is therefore valuable for characterizing public personas and identifying what makes certain speakers, like championship contenders, distinctive.

This project investigates three research questions:

RQ1: How do token volumes change across seasons for drivers between 2022–2025?
We expect media presence to vary by performance, suggesting that token counts can act as a proxy for driver prominence.

RQ2: How do linguistic styles differ between drivers?
We hypothesize that some drivers will demonstrate more descriptive, emotionally expressive

Table 1: Summary of Corpus

	Year	Total Docs	Total Tokens
	2025	282	266,716
	2024	228	247,277
	2023	285	240,316
	2022	253	193,165
Total	—	1,048	947,474

language, while others favor direct, action-oriented phrasing. To evaluate this, we apply hierarchical clustering and PCA using part-of-speech (POS) distributions extracted via UDPipe.

RQ3: Can linguistic features predict whether an interview was spoken by Max Verstappen? As a multi-world-champion known for concise and decisive communication, we expect Verstappen's speech to exhibit a distinct linguistic signature. We test this using a predictive model trained on lexical metrics, keyword indicators, pronoun usage, and POS proportions.

The corpus consists of official Formula 1 press-conference transcripts from 2022 through 2025. They are are well suited to the research questions as they provide consistent language samples across drivers and seasons. Together, these analyses contribute to understanding how language reflects performance, pressure, and identity in one of the world's most competitive sporting environments.

Data

The corpus used in this study consists of official FIA press conference transcripts from Formula 1 drivers between 2022 and 2025, scraped from formula1.com. Each Grand Prix weekend includes a mix of sessions: Thursday, post-qualifying, post-sprint, and post-race press conferences. Each session typically features 3 to 6 drivers per conference. Each press conference is accessed via a unique webpage, and all remarks made by a single driver in a session are aggregated into one document.

Only transcripts containing 200+ tokens of driver speech were retained to ensure meaningful linguistic analysis. After preprocessing, the final corpus included 1,048 documents totaling 947,474 tokens, with each document labeled by driver, team, Grand Prix, season, and conference type. This structure enables both season-level analysis and driver-level style comparison. There is a balanced representation across the four seasons, with approximately 250 documents and over 230,000 tokens per year (see Table 1).

The stacked bar chart in Figure 1 illustrates total token volume per season, with the blue section indicating the proportion attributable to Max Verstappen. Across all four years, Verstappen contributes a substantial and stable share of total corpus tokens, ensuring that the classification task later in this report has sufficient sample size for training and predictive reliability. Token volume grows slightly each season, suggesting that the corpus is not concentrated

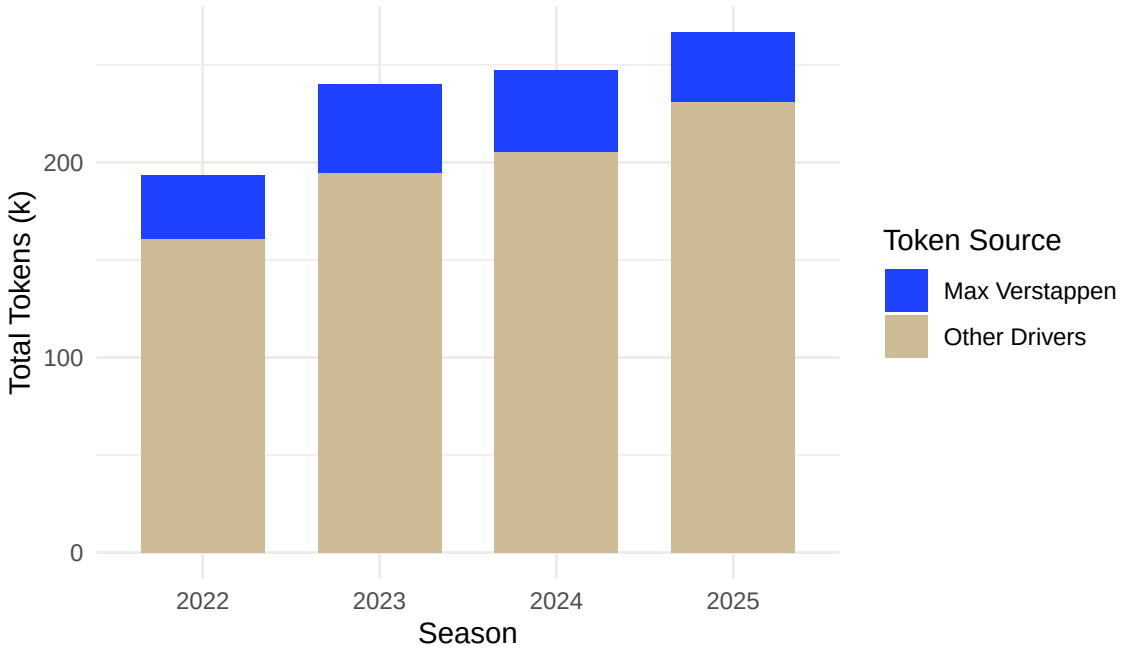

Figure 1: Token Count by Season with Verstappen Share

in early or late years.

A primary limitation of the corpus is the uneven press attendance and document volume across drivers, particularly those who achieve podium finishes more frequently. Drivers such as Verstappen produce far more words than those consistently outside podium positions. This may bias predictive results toward well-sampled drivers.

Methods

To investigate how communication volume changes over time, I computed the total number of tokens spoken by each driver in every season (2022–2025). Because some drivers appear infrequently across years, I restricted analysis to the top seven drivers by total token count over the full corpus. This ensures comparability and avoids misleading trends caused by drivers who have minimal press coverage. Annual totals were visualized using line charts, allowing us to observe whether high-performing drivers maintain consistent media presence or fluctuate due to competitive performance, team strategy, or season context. This method directly answers RQ1 by quantifying how much drivers speak and how consistently they appear in press conferences.

To evaluate whether drivers share stylistic tendencies, I applied hierarchical clustering on part-of-speech distributions. All transcripts for each driver were concatenated into a combined document, then parsed using UDPipe to produce proportional counts of Universal POS categories (NOUN, VERB, ADJ, PRON, etc.). Features were normalized so stylistic differ-

Table 2: Test, Train, and Validation set split

Dataset	Non-MV	MV	Total Docs	% of Entire Corpus
Train	502	94	596	58
Validation	131	18	149	15
Test	242	40	282	27

Table 3: Features Used for LASSO Classification

Feature Type	Extracted Variables
Token-level features	Word count, average word length, average sentence length
Keyword presence	they, struggle, happy, difficult, tough, win, fast, fight
Pronoun usage	Per-thousand-token frequency of "I" and "we"
POS distributions	Percentage of tokens in POS classes (NOUN, VERB, ADJ, etc.)

ences reflect grammatical preference rather than speech volume. Clusters were computed using Ward's D2 linkage. A dendrogram were used to assess the clusters, and PCA was performed to visualize how drivers separate along major linguistic dimensions. This method shows whether speech patterns cluster by personality, first language, team environment, or performance.

The classification task is focused on whether a transcript was spoken by Max Verstappen (1) or any other driver (0). To evaluate generalization realistically, the dataset was split chronologically: training data included transcripts from 2022 to 2024, while 2025 transcripts formed the test set. Within the training data, an 80/20 split created a validation set. The split is shown in Table 2. A chronological held-out test set ensures that results reflect predictive ability. In all sets, class balance is roughly maintained, with Verstappen representing about 15-20% of documents. This avoids biasing the model toward the majority class.

Because the feature space is moderately high-dimensional and correlated, I applied LASSO logistic regression using cv.glmnet() with alpha = 1. LASSO performs variable selection by shrinking some coefficients to zero, reducing overfitting and improving interpretability. Cross-validation determined the optimal penalty λ, and the final model was fit using that value. Only non-zero coefficients were interpreted as meaningful markers of speech style.

The features used for the classification task are shown in Table 3. Regularization was selected because the linguistic features might not be independent and many may contribute very weak signal. Penalization improves stability, reduces noise, and highlights the strongest predictors that are best for classification and prediction.

Results

Figure 2 presents total token counts for the seven most frequently interviewed Formula 1 drivers across seasons 2022–2025. Token count serves as a practical reflection for media presence.

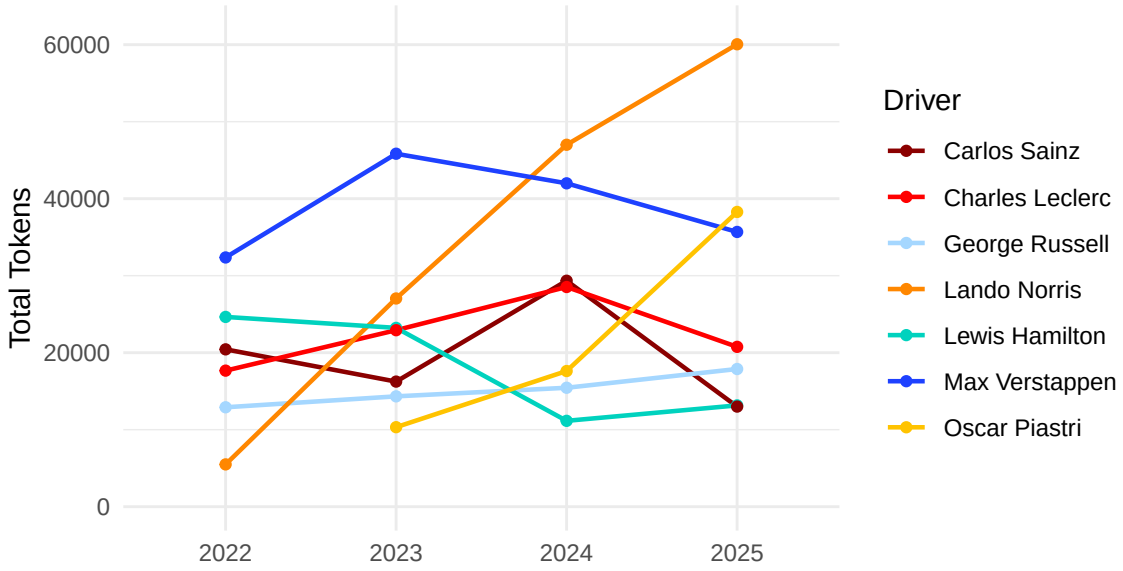

Figure 2: Token Count by Driver per Year

Across all four seasons, Max Verstappen consistently appears among the top speakers, peaking in 2023 and remaining strongly represented through 2025. This pattern is consistent with his role as a dominant on-track competitor, aligning with the expectation that drivers who consistently contend for race wins and championships receive more media attention.

Lando Norris displays the most striking upward trajectory, beginning with comparatively limited coverage in 2022 and rising sharply to become the most dominant speaker by 2025. This coincides with his emergence as the new world champion in 2025, illustrating how competitive success translates into increased media exposure. In contrast, Lewis Hamilton shows a stable but modest decline in total word output over the same period, while Ferrari drivers Charles Leclerc and Carlos Sainz fluctuate season-to-season with similar trends.

Overall, this analysis confirms that token volume reflects competitive relevance and media prominence in Formula 1. This finding provides a strong foundation for the other two research questions. The variation in exposure across seasons supports the importance of exploring stylistic differences and testing whether language features can classify Verstappen against the grid.

Figure 3 visualizes hierarchical clustering of POS-tag distributions across drivers, cut into five clusters. One small cluster at the bottom contains two rookies, Zhou Guanyu and Logan Sargeant, reflecting similar linguistic behavior and relatively lack of experience in press conferences. Another pair contains the two Spanish drivers, Carlos Sainz and Fernando Alonso. This is consistent with their shared speech styles due to the influence of their first language and culture. Another four-driver cluster includes three world champions—Lewis Hamilton, Lando Norris, and Max Verstappen. This grouping suggests that media experience and championship status may correspond with stylistic habits. Remaining drivers form two larger clusters with greater internal variability, consistent with a wider range of lexical choices and press experience

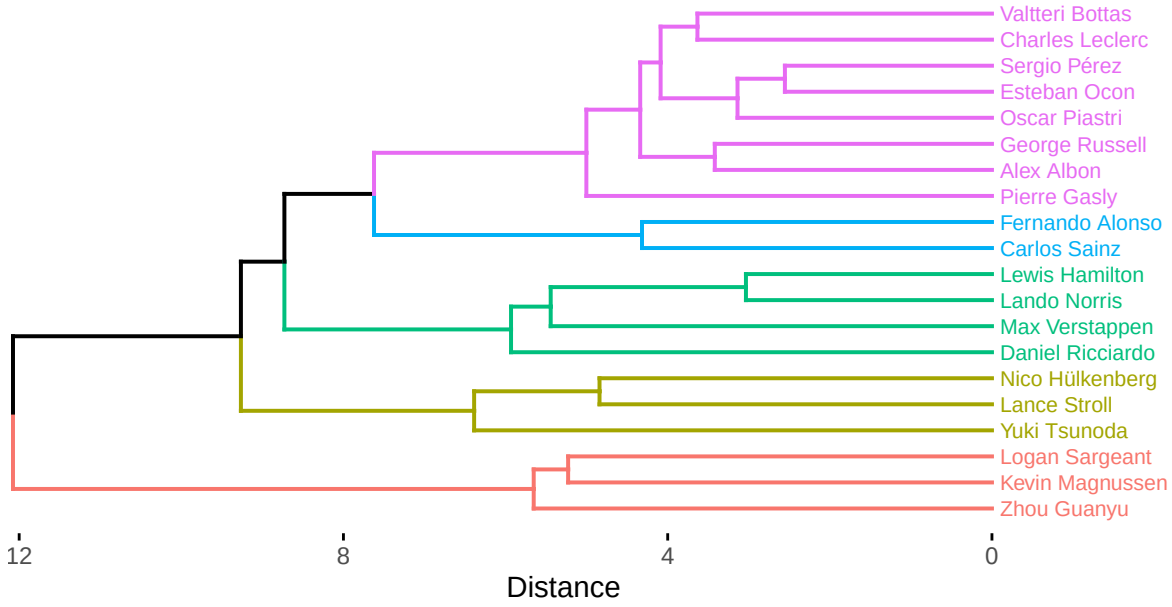

Figure 3: Hierarchical Clustering Dendrogram of Drivers (5 Clusters)

levels.

Figure 4 provides a PCA visualization illustrating how POS features contribute to stylistic differences. The first two components explain approximately 48% of total variation. Zhou and Sargeant lie at the extreme left of PC1, consistent with high use of adpositions and verbs and supporting their distinct placement in the dendrogram. Sainz and Alonso are positioned high on PC2, driven by elevated frequency of numerals and proper nouns. Many other drivers occupy a dense central region, indicating moderate and balanced POS use without strong directional loading. Variation in POS-based style provides the foundation for predictive modelling, demonstrating that classification of speakers is both statistically reasonable.

To address the third research question, a LASSO logistic regression model was trained to predict whether a press conference transcript was spoken by Max Verstappen or by another driver. After splitting the corpus into training, validation, and a held-out test set, cross-validation selected a minimal lambda of 0.0092. Out of the 30 candidate linguistic predictors, the model retained 20 non-zero coefficients. The classifier reached 87.9% accuracy on the validation set and 85.8% on the 2025 test set.

Figure 5 displays the selected coefficients, where green ones are positive and red ones are negative. Predictors with positive coefficients include frequencies of “they”, “difficult”, “win”, and “happy”. They indicate higher likelihood of a Verstappen transcript. These patterns suggest that Verstappen’s speech regularly includes references to external agents like other teams (“they”), performance evaluations (“difficult,” “win”), and positive reactions (“happy”).

Conversely, the strongest negative predictor is average word length, indicating that longer or more complex words are associated with other drivers rather than Verstappen. Other negative coefficients include the frequencies of “fight”, “struggle”, and the proportion of particles

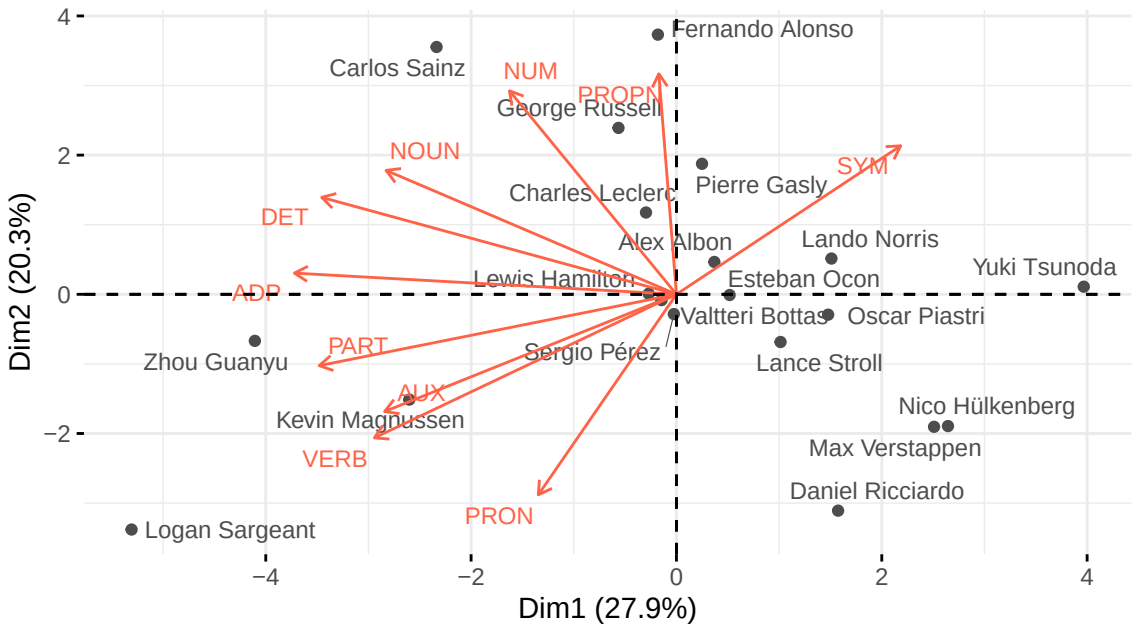

Figure 4: PCA Biplot of Drivers by POS Features

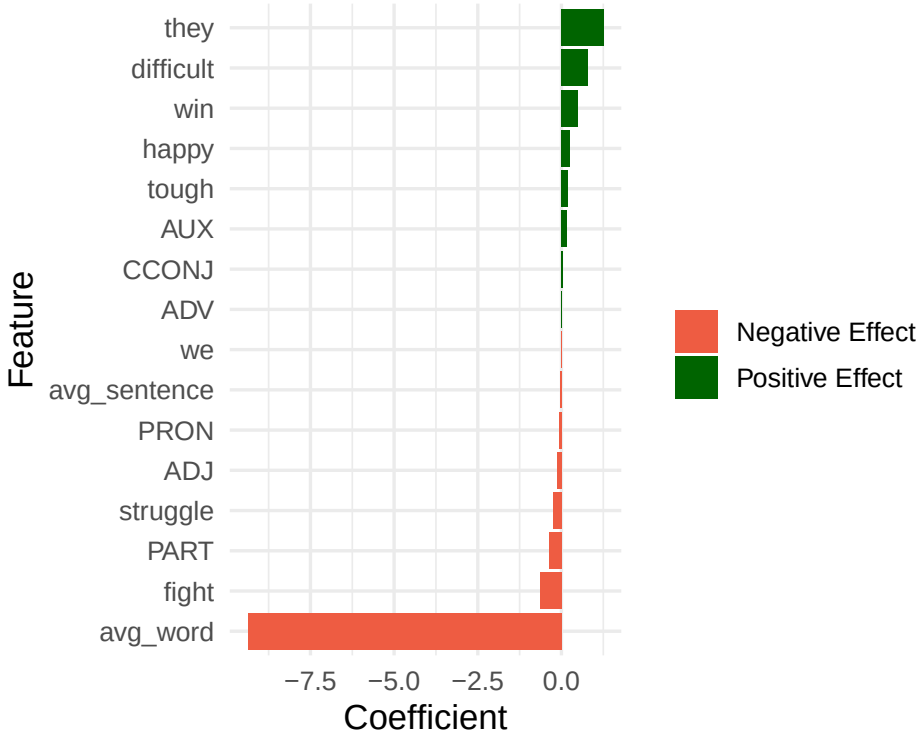

Figure 5: LASSO Coefficients for Predicting Max Verstappen Transcripts

(PART) and adjectives (ADJ). When these features appear more frequently, a transcript is less likely to belong to Verstappen.

Overall, the results demonstrate that Verstappen's linguistic signature is identifiable and stable across seasons. The stylistic features collectively highlight Verstappen's characteristic style of have more reference to others, some particular evaluations of race performances, shorter words, and less self-referential constructions.

Discussion

The analyses in this study show that Formula 1 drivers display unique linguistic patterns in their press-conference speech, with connections to competitiveness, experience, and personal communication style. The token-count analysis demonstrated that media presence reliably reflects on-track performance. Verstappen consistently produced high word totals across all four seasons, while Norris' sharp rise in 2025 paralleled his emergence as world champion.

The POS-based clustering analysis showed that drivers do not share a uniform linguistic profile. The hierarchical clustering and PCA revealed groupings consistent with competitive status, linguistic background, and media experience. Rookie drivers Zhou and Sargeant clustered together due to simpler syntactic patterns and limited press exposure, while Alonso and Sainz formed a pair consistent with shared Spanish-influenced English usage. Notably, the three recent world champions, Hamilton, Verstappen, and Norris, clustered closely, suggesting that extensive media experience may shape a similar professional language style.

The classification results further demonstrated that linguistic features provide a distinctive signature of speaker identity. The LASSO model achieved strong validation (87.9%) and test accuracy (85.8%) in identifying Verstappen's transcripts, selecting features consistent with his direct and performance-focused style. Positive predictors such as frequencies of "they", "difficult", and "win" highlight his evaluative framing of events, while shorter average word length strongly distinguished him from other drivers.

Limitations

First, the corpus reflects imbalance in press-conference participation. Drivers with more podiums naturally appear more often, introducing bias that may amplify stylistic features tied to media exposure rather than linguistic tendencies. Second, the analysis draws exclusively from FIA press conferences over a four-year window, omitting earlier seasons and excluding other major forms of driver communication such as interviews. Third, while the feature set captures important lexical and POS-based characteristics, it remains limited. Future work could incorporate richer indicators such as lexical diversity and measures of syntactic complexit to provide a more complete account of individual speaking style.

Acknowledgments

Generative AI is used in the process of scraping online data. It was helpful in providing a starting point for the data collection process. Nevertheless, I still spent a large amount of time modifying and debugging to fit my specific needs for the corpus.

Works Cited

Frederick Mosteller & David L. Wallace (1963) Inference in an Authorship Problem, Journal of the American Statistical Association, 58:302, 275-309
Xiao, Richard. (2009) Multidimensional analysis and the study of world Englishes, World Englishes, 28(4), 421–450
Granger, Sylviane (2017) Academic Phraseology: A Key Ingredient in Successful L2 Academic Literacy
Le, Xuan, Ellen Riloff & Mark Tanner. (2018) Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, Literary and Linguistic Computing, 33(1), 1–22
Lab 1: Mosteller & Wallace, 36668 Text Analysis, Fall 2025, CMU
Lab 7: Part-of-Speech Tagging and Dependency Parsing, 36668 Text Analysis, Fall 2025, CMU
Lab 10: Cluster Analysis, 36668 Text Analysis, Fall 2025, CMU