Linguistic Fingerprints on the Ice: Differentiating Coach and Player Discourse in NHL Interviews

Axel Strid
Faculty of Science and Engineering
Linköping University, Sweden
axest556@student.liu.se

Abstract

This study investigates the linguistic and thematic distinctions between National Hockey League (NHL) Coaches and Players during Stanley Cup Final interviews (1997–2019). While sports interviews are often dismissed as cliché-filled, we hypothesized that the distinct pressures of the roles; performance execution (Players) versus strategic management (Coaches); would create detectable linguistic signatures. Using a dataset of 2,096 interview transcripts, we trained supervised classifiers reaching an exceptional 92% accuracy. Moving beyond prediction, we interpret our findings using LIME, Log-Odds Ratio, and validate the results with unsupervised clustering. We find that players utilize a "subjective" dialect characterized by first-person pronouns and sensory vocabulary, while coaches employ a "managerial" dialect focused on collective nouns ("the group") and abstract processes. Furthermore, we demonstrate that unsupervised clustering and topic modeling only successfully separates these roles when the dataset is balanced, mitigating the noise of the dominant player class.

1 Introduction

In the high-pressure environment of professional sports, post-game interviews are a ritualized form of communication. For the casual observer, these interactions may seem like uniform and identical, filled with repetitive "hockey clichés" about "working hard" and "getting pucks deep". However, the speakers fill fundamentally different roles. Players are task-focused, describing immediate physical experiences and personal emotions. Coaches are question-focused and strategic, often acting as spokespeople for the whole organization.

The primary problem addressed in this project is the following research question: **Can Natural Language Processing (NLP) models accurately distinguish between the discourse of NHL coaches and players, and if so, what linguistic features drive this distinction?**

Solving this problem is interesting from both a linguistic and a machine learning perspective. Linguistically, it offers insight into how professional roles shape idiolects (individual speech patterns) within a shared domain. Technologically, it challenges us to interpret *why* models perform well. If a simple classifier achieves almost perfect accuracy, it implies a distinct separation in the feature space that justifies explanation. By solving this, we learn to look past high accuracy scores and understand the semantic fingerprints of human authority and compliance in sports.

2 Theory

While standard text classification methods formed the baseline of this project, the primary contribution lies in the interpretability techniques not covered in the standard curriculum.

2.1 Supervised Baseline: Multinomial Logistic Regression (MLR)

Logistic Regression serves as our baseline. It models the probability of a class y given a text vector x using the logistic function. Unlike "black box" models, it offers inherent interpretability: the learned coefficients (β) directly indicate the direction and magnitude of a word's influence on a specific class (UCLA: Statistical Methods and Data Analytics, 2024).

2.2 Interpretability Framework

To move beyond accuracy, we employed an interpretability framework:

Global Feature Importance (Average Feature Effect) To understand general model behavior, we analyze global feature importance. For linear models, this is derived from the learned coefficients.

However, raw coefficients can be misleading if features have significantly different scales. We utilize the Average Feature Effect, which weights the coefficient by the frequency of the feature, highlighting words that are both impactful and frequent (scikit-learn developers, 2025).

Local Interpretability (LIME) To understand model behavior, LIME (Local Interpretable Model-agnostic Explanations) was employed. Unlike global feature importance, LIME approximates the model locally around a specific prediction to explain *individual* instances. It alters the input by removing words to see how the prediction probability changes, allowing for a visualization of which specific words drive a specific text to be classified as "Player" or "Coach" (Ribeiro et al., 2016).

Log-Odds Ratio with Dirichlet Prior To understand stylistic differences, we need to measure *distinctiveness*. Following Monroe et al. (2008), a Log-Odds Ratio with an informative Dirichlet prior was utilized. This metric calculates the probability of a word occurring in one class relative to another class, penalizing common words such as stop words. High positive log-odds indicate words unique to Coaches, while high negative log-odds indicate words unique to Players.

2.3 Unsupervised Grouping

To validate these patterns without labels, we utilize **K-Means Clustering** (optimizing centroid distance) and **Latent Dirichlet Allocation (LDA)**. LDA is a generative probabilistic model that assumes documents are mixtures of latent topics, and topics are mixtures of words (Mahmoud, 2023).

3 Data

The dataset consists of 2,096 interview transcripts from the NHL Stanley Cup Finals, spanning a 21-year period (1997-2019). It was sourced from Kaggle and originally scraped from ASAP Sports.

3.1 Data Exploration

The 2,096 interviews are divided between 1,498 Players, 515 Coaches, and 83 Other (Experts, General Managers, etc.). The scope contains 470 unique interviewees across 24 different teams. The dataset includes metadata such as the speaker's name, the teams involved in the specific game, the date, and the transcript text.

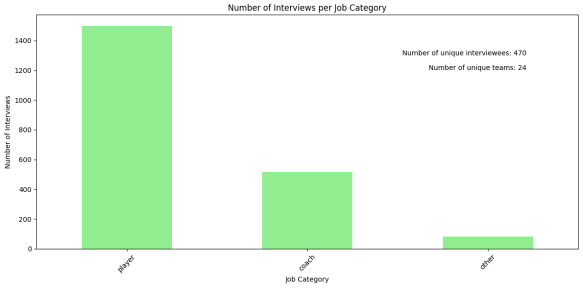

Figure 1: Number of interviews per job category.

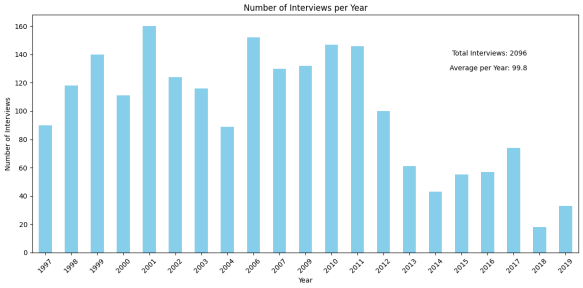

Figure 2: Number of interviews per year.

3.2 Preprocessing and Balancing

Standard preprocessing included tokenization and lowercasing. Removal of English stop words was tested. We experimented with two main vectorization strategies: Count Vectorizer and TF-IDF. The data was also manipulated for experiments of models trained on the full 3-class (Coach, Player, Other) data, on a full 2-class (Coach, Player), as well as on a balanced 2-class dataset, since the original data is heavily imbalanced (3:1 Player-to-Coach ratio).

Balancing the dataset turned out to be a critical step for achieving promising results of the unsupervised phase of the project.

4 Method

The approach followed a pipeline from supervised prediction to deep interpretation, concluding with unsupervised validation.

4.1 Supervised Classification

Multiple scikit-learn classifiers were trained: Multinomial Naïve Bayes, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, K-Neighbors, and a Zero-Shot Gemini-Flash-2.5 LLM model, to establish a performance baseline. Models were evaluated using Accuracy, Precision, Recall, and F1-scores, both class-wise and macro, and weighted average to account for class imbalance. Various configurations were tested as described in the Data section of the report, including

the removal of English stop words versus keeping them, to determine if function words (like "I" vs. "we") were predictive.

4.2 Name Stripping (Entity Removal)

A potential source of data leakage was the presence of person names. If a model learns that "Quenneville" is a coach and "Crosby" is a player, it classifies based on identity and not discourse style.

Therefore, a strict Named Entity Recognition (NER) pipeline was implemented using spaCy to strip all person names from the text. Models were re-trained on this anonymized data to ensure the learned patterns were linguistic rather than metadata-based.

4.3 Interpretability and Unsupervised Analysis

The core of the methodology focused on *why* the models achieved such high predictive accuracy. Following the supervised experiments, the best-performing model (MLR) was selected for deep analysis and interpretation:

Average Feature Effect: Highlights the global importance of individual tokens.

LIME & Log-Odds Ratio: Used to extract the specific distinctive vocabulary defining each role.

K-Means Clustering: Applied to the anonymized text (with $k = 2$) to see if the roles would naturally separate without labels.

LDA: Used to identify latent topics and measure their distribution across roles to verify if coaches and players speak about different subjects.

5 Results

5.1 Classification Performance

Simple algorithmic classifiers performed exceptionally well. From the extensive experimentation across seven model architectures, Logistic Regression consistently yielded the highest performance and interpretability. Consequently, Logistic Regression is the primary focus of our subsequent linguistic analysis.

Notably, there was no single "best practice" for text vectorization; optimal configurations (Count vs. TF-IDF, stop-word removal) varied significantly between model architectures.

Comparing the full (Table 1) and balanced (Table 2) datasets, we observe that while global accuracy remained high, the convergence of Weighted and Macro F1-Scores in the balanced setting confirms model robustness. This demonstrates that the distinct "dialects" remain identifiable even without the statistical advantage of class imbalance.

Model	Acc.	F1-W	F1-M	Config.
Multinomial NB	0.90	0.90	**0.81**	CountV
Logistic Regression	**0.92**	**0.91**	0.76	CountV
Decision Trees	0.84	0.84	0.66	CountV
Random Forests	0.88	0.87	0.70	TF-IDF
Gradient Boosting	0.91	0.90	0.76	TF-IDF
K-Neighbors	0.86	0.86	0.86	CountV
Gemini-Flash-2.5	0.90	0.90	0.60	Zero-Shot

Table 1: Classification performance on the **full, imbalanced 3-class dataset** (Coach, Player, Other).

Model	Acc.	F1-W	F1-M	Config.
Multinomial NB	0.57	0.60	0.54	TF-IDF
Logistic Regression	**0.91**	**0.92**	**0.89**	TF-IDF
Decision Trees	0.77	0.79	0.73	TF-IDF
Random Forests	0.87	0.87	0.73	TF-IDF
Gradient Boosting	**0.91**	0.91	0.88	TF-IDF
K-Neighbors	0.46	0.46	0.46	TF-IDF
Gemini-Flash-2.5	-	-	-	Zero-Shot

Table 2: Classification performance on the **balanced 2-class dataset** (Coach vs. Player).

5.2 Robustness to Preprocessing

To verify that models were learning discourse style rather than identity, we evaluated performance under aggressive preprocessing: removing stop words and performing NER to strip proper names (for example "Crosby" and "Quenneville").

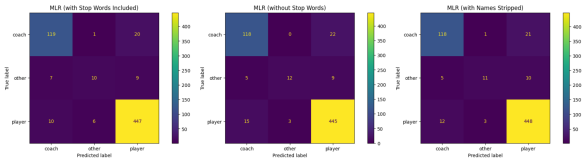

Figure 3: Confusion matrices for MLR: (Left) Standard, (Center) No Stop Words, (Right) Name Stripped. Performance remains stable, confirming linguistic substance over metadata.

As shown in Figure 3, these steps resulted in negligible performance loss. This confirms that the distinct "dialects" of Coaches and Players are systemic and identifiable even without specific names or functional stop words. While performance stayed flat, this step ensured the analysis was driven by linguistic substance rather than metadata.

5.3 Interpretability: The "Me" vs. "The Group"

We utilized Log-Odds Ratio analysis to identify words that are statistically over-represented in one class relative to the other.

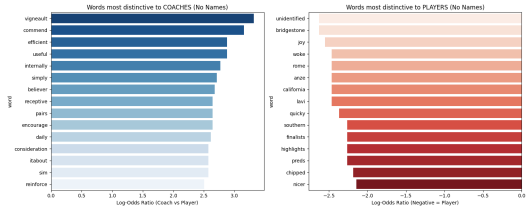

Figure 4: Top distinctive words (Log-Odds Ratio) for Coaches (Blue) and Players (Red) from the name-stripped dataset.

Figure 4 reveals a sharp contrast in vocabulary. **Coach Distinctive Words** include abstract, managerial terms such as *efficient, useful, internally, consideration, encourage, receptive*, and *reinforce*. **Player Distinctive Words** are grounded in sensory experience and emotion, including *joy, woke, quicky, chipped*, and *nicer*.

Complementing this, the Average Feature Effect on the balanced dataset highlights the use of "filler" words. The word *"just"* is the rank-1 keyword for Players, often accompanied by *"obviously"* and *"yeah"*. In contrast, Coaches favor collective nouns, utilizing *"we"* and *"team"* significantly more frequently than the first-person *"I"*.

5.4 Local Validation with LIME

To verify these global patterns drive individual predictions, we examined instances via LIME.

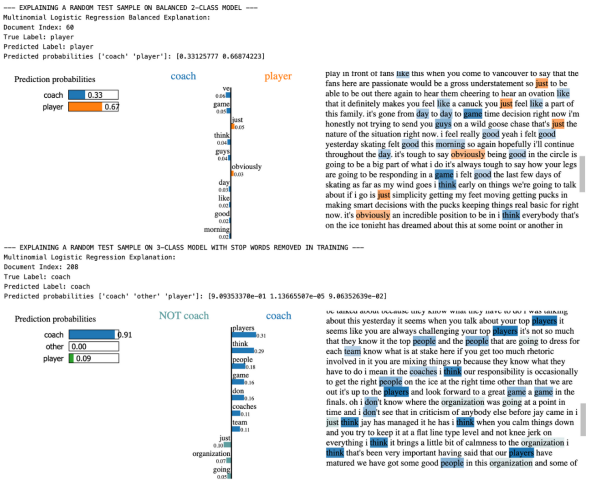

Figure 5: LIME explanations for a predicted Player (top) and Coach (bottom).

As illustrated in Figure 5, the classifier's decision-making aligns with our linguistic hypothesis. In the Player example, the prediction is heavily driven by the linguistic fillers *"just"* and *"obviously"*, which outweigh the content words. Conversely, the Coach prediction is represented by formal and collective markers (*"team", "people", "players"*), confirming that the model has successfully learned the distinct "managerial" register.

5.5 Unsupervised Analysis: The Impact of Balancing

To test if these dialects could be discovered without labels, we applied K-Means clustering.

The Subset Phenomenon On the full imbalanced dataset, K-Means with $k = 2$ failed to separate the roles. Cluster 0 was 97% Players, but Cluster 1 was a mix of 50% Player and 50% Coach. The model essentially found a "Casual Player" cluster and a "General Hockey" cluster, treating Coaches as a subset of the latter.

Emergence of Distinct Dialects However, upon balancing the dataset, the unsupervised separation improved dramatically (Figure 6).

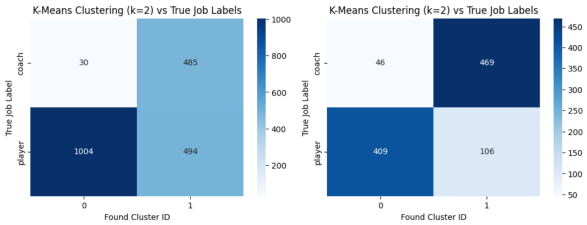

Figure 6: Impact of balancing on K-Means clustering. The balanced dataset (Right) successfully separates Player discourse (Cluster 0) from Coach discourse (Cluster 1).

Cluster 0 (Player-Dominant, 89.9%) is defined by filler words: *just, obviously, going, know, really, lot*. **Cluster 1 (Coach-Dominant, 81.6%)** is defined by structural terms: *team, play, good, guys*.

This finding was reinforced by LDA Topic Modeling on the balanced data. Topic 1 captured the **Subjective Experience** (68% Player), dominated by first-person pronouns (*my, am, feel, mean*). Topic 2 captured the **Managerial Perspective** (82% Coach), dominated by collective strategy terms (*group, opportunity, trying*).

6 Discussion

The results demonstrate that NHL Coaches and Players speak distinct "dialects" of the same lan-

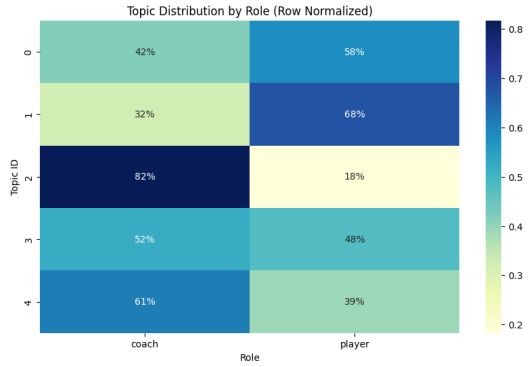

Figure 7: LDA Topic Distribution (Balanced Data). Topic 2 shows a strong Coach signal (Managerial), while Topic 1 shows a strong Player signal (Subjective).

guage. The high classification accuracy (92%) is not an artifact of metadata leakage (names), but a result of differences in register and perspective.

6.1 Register: Casual vs. Formal

The dominance of fillers like "*just*" and "*obviously*" indicates a spontaneous, lower-register speech pattern in Players. Players are often interviewed immediately after physical exertion, leading to simpler sentence structures. Coaches, conversely, use "managerial" vocabulary (*efficient, internally, reinforce*) reflecting a higher register and a strategic, detached perspective.

6.2 Perspective: Subjective vs. Collective

The most profound difference is the use of pronouns and collective nouns. Players dominate the first-person topic identified in LDA (*I, me, my, feel*). They narrate the game through their own sensory experience. Coaches almost exclusively use the third-person or collective perspective. They refer to the team not just as "the team" but as "the group". This reflects their role as managers of a unit rather than participants in the action.

6.3 Relation to Prior Work

Our findings align with recent research demonstrating that linguistic signals in sports interviews contain latent information about role and mental state. Oved et al. (2020) found that NBA players' interviews contain "risk" and "strategy" signals that effectively predict in-game deviations, a conclusion supported by our finding that NHL players and coaches operate in distinct semantic spaces.

Additionally, Weissbock and Inkpen (2014) established that textual data from pre-game reports enhances predictive models for NHL game outcomes when combined with statistical metrics. While Weissbock and Inkpen (2014) treated all text as a single data source, our findings suggest that future models should separate 'Coach' and 'Player' interviews. Since these groups speak differently, analyzing them separately could uncover unique predictive signals that are lost with mixed data.

Furthermore, while Velichkov and Koychev (2019) demonstrated that pre-game interview text alone can predict match outcomes in individual sports, this presents a promising direction for future research in the NHL context. Our study supports the feasibility of such forecasting by confirming the core hypothesis: that sports interviews contain robust, distinguishable linguistic signatures rather than empty clichés, providing the necessary feature set for future predictive modeling.

6.4 Limitations

The primary limitation was the class imbalance. As shown in the unsupervised analysis, the sheer volume of player interviews can drown out the "Coach signal" if not addressed via undersampling. Additionally, the analysis is limited to the Stanley Cup Finals; regular-season interviews might display different emotional valences (less pressure).

7 Conclusion

This project successfully solved the problem of differentiating NHL discourse, achieving over 90% accuracy in classifying roles. Beyond classification, we isolated the specific linguistic markers that define these roles. We conclude that Players speak a dialect of "Action and Emotion", characterized by first-person pronouns, fillers like "*just*" and "*obviously*", and sensory words. On the other hand, Coaches speak a dialect of "System and Process", characterized by collective nouns like "group", abstract descriptors, and third-person analysis.

The takeaways extend beyond hockey: in hierarchical organizations, authority figures and subordinates may share a vocabulary ("the game"), but they occupy distinct semantic spaces. The "manager" speaks of the system; the "worker" speaks of the task.

References

Mohamed Bakrey Mahmoud. 2023. All about latent dirichlet allocation (lda) in nlp. Medium. Accessed: 2026-01-12.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Nadav Oved, Amir Feder, and Roi Reichart. 2020. Predicting in-game actions from interviews of nba players. *Computational Linguistics*, 46(3):667–712.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, San Francisco, CA, USA. ACM.

scikit-learn developers. 2025. *Classification of text documents using sparse features*. scikit-learn. Accessed: 2026-01-10.

UCLA: Statistical Methods and Data Analytics. 2024. Multinomial logistic regression | r data analysis examples. Accessed: 2026-01-10.

Boris Velichkov and Ivan Koychev. 2019. Deep learning contextual models for prediction of sport events outcome from sportsmen interviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1240–1246.

Josh Weissbock and Diana Inkpen. 2014. Combining textual pre-game reports and statistical data for predicting success in the national hockey league. In *Canadian Conference on Artificial Intelligence*, pages 251–262. Springer.

A Appendix

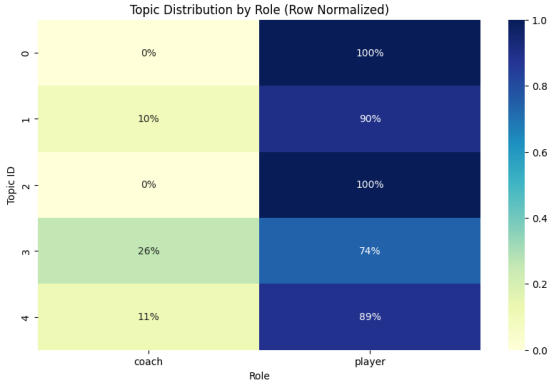

Figure 8: LDA Topic Distribution on the **imbalanced dataset**. Note that every topic (rows) is dominated by the Player class (Dark Blue), illustrating how class imbalance drowns out the distinct Coach signal.

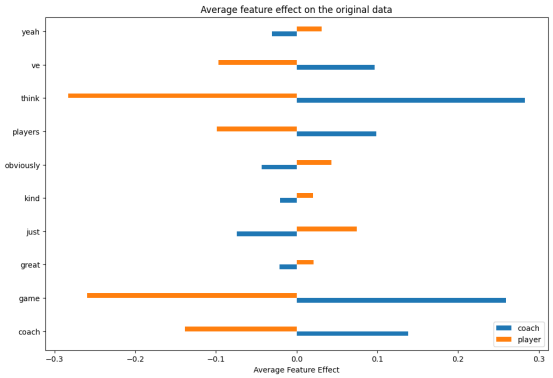

Figure 9: Average Feature Effect (Global Importance) on the balanced dataset (no stop words). The word *"just"* is the single most predictive feature for the Player class, confirming the "casual register" hypothesis.

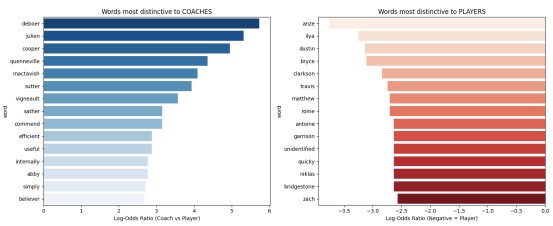

Figure 10: Log-Odds Ratio analysis **before name stripping**. The most distinctive features are proper names (e.g., *"Quenneville", "Sutter", "Ilya"*), confirming the necessity of the NER pipeline to force the model to learn linguistic style rather than identity.