ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

LEXICAL DIVERSITY AND EMOTIONAL LANGUAGE IN POP, ROCK, AND RAP LYRICS

Esha Tul Razia
Bs English, Punjab College Mian Channu
eshatulrazia42@gmail.com
Ayesha Rauf
BS English, Punjab College Mian Channu
ayesharauf2003@gmail.com
Burera Marium
MPhil Scholar, University of Education
barirakhalid890@gmail.com
Iqra Saifullah
MPhil Scholar, University of Education
iqrasaif3890@gmail.com

Abstract

Song lyrics constitute a rich and underexplored domain of linguistic inquiry, offering a window into the intersection of popular culture, emotional expression, and vocabulary use. The present study employs a corpus-based computational approach to examine lexical diversity, sentiment polarity, and thematic structures in contemporary English song lyrics drawn from three dominant musical genres: Pop, Rock, and Rap. A balanced corpus of 3,000 songs—1,000 per genre—was compiled from a publicly available Kaggle dataset. Lexical diversity was operationalized through the Type-Token Ratio (TTR), while sentiment polarity was measured using the TextBlob framework. Keyword frequency analysis and Latent Dirichlet Allocation (LDA) topic modeling were applied to uncover dominant vocabularies and latent semantic themes. One-way Analysis of Variance (ANOVA) and Tukey's Honest Significant Difference (HSD) post-hoc tests were used to determine the statistical significance of inter-genre differences. Results indicate that Rap exhibits significantly greater lexical diversity (M = 0.5070) than both Pop (M = 0.4707) and Rock (M = 0.4724), while Pop demonstrates significantly more positive sentiment polarity (M = 0.0668) than the other genres. Topic modeling identified four major thematic clusters—Narrative Action, Romance, Reflective Existentialism, and Socio-Cultural Realism—which vary in prominence across genres. These findings demonstrate how computational linguistic methods can systematically uncover meaningful differences in language use across contemporary musical genres.

Keywords: corpus linguistics, lexical diversity, sentiment analysis, topic modeling, song lyrics, computational linguistics, Type-Token Ratio, LDA

1. Introduction

Language plays a central role in music, serving not only as a medium of artistic expression but also as a mirror of cultural values, emotional experience, and social reality. Song lyrics constitute a uniquely accessible form of discourse, situated at the intersection of popular culture and linguistic production. As corpus linguistics has expanded beyond canonical written and spoken texts to embrace digital and popular-cultural data, song lyrics have attracted increasing scholarly attention as a site for systematic linguistic investigation (Brackett, 2016; Tsur & Gafni, 2019).

Different musical genres are widely understood to exhibit distinct linguistic profiles. Pop music is commonly associated with repetitive, positively framed language designed to achieve broad commercial appeal. Rock frequently incorporates introspective, narrative, and emotionally complex content that explores personal experience. Rap, by contrast, is widely recognized for its lexical density, intricate internal rhyme structures, and socially engaged storytelling (Bradley, 2017; Herd, 2016; Orejuela, 2015). Despite these widely held characterizations, systematic quantitative comparisons of linguistic features across musical genres remain relatively scarce in the corpus linguistics literature.More recently, scholars have extended lyrical analysis to political songs, examining how language is used to legitimize

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

ideological positions and construct political identities through discourse strategies such as emotion, rationality, and appeals to expertise (Muhammad et al., 2025).

Recent advances in computational linguistics have dramatically expanded the analytical toolkit available to researchers. Techniques such as Type-Token Ratio (TTR) calculation, automated sentiment analysis, keyword frequency extraction, and Latent Dirichlet Allocation (LDA) topic modeling now enable the large-scale investigation of lexical and thematic properties within extensive text corpora. These methods have been productively applied to social media discourse, literary texts, news corpora, and political speech, yet their application to musical lyrics remains underdeveloped (Blei, 2012; Maier et al., 2018). The present study addresses this gap by applying a multi-method computational approach to a balanced corpus of contemporary English-language song lyrics.

Furthermore, recent work in discourse and corpus linguistics has underscored the importance of examining emotional and ideological dimensions of language across different text types. Muhammad et al. (2025), in their analysis of political songs from Pakistan's PTI and PPP parties, demonstrate that song lyrics function not merely as entertainment but as vehicles for ideological legitimization, emotional engagement, and identity construction. While their study employs qualitative Critical Discourse Analysis, the present investigation complements such work by providing a quantitative computational framework for examining how emotional language varies systematically across musical genres.

1.1 Research Objectives

The study pursues the following objectives:

Objective 1: To measure, compare, and contrast lexical diversity and sentiment polarity across Pop, Rock, and Rap lyric corpora.

Objective 2: To identify, extract, and map dominant semantic themes and characteristic vocabularies unique to each genre.

1.2 Research Questions

The study is guided by the following research questions:

RQ1: How do Pop, Rock, and Rap lyrics differ in terms of lexical diversity and sentiment polarity?

RQ2: What semantic themes and dominant lexical patterns characterize each musical genre?

1.3 Significance of the Study

This study contributes to the growing field of computational corpus linguistics by demonstrating the applicability of quantitative methods to contemporary popular music. By integrating lexical diversity measurement, sentiment analysis, statistical inference, and topic modeling within a unified analytical framework, the research offers a comprehensive account of linguistic variation across major musical genres. The findings extend understanding of how vocabulary richness, emotional orientation, and thematic structures operate within musical discourse. The study also provides a replicable methodological framework adaptable to future investigations of popular cultural texts.

2. Literature Review

2.1 Corpus Linguistics and Song Lyrics

Corpus linguistics provides systematic frameworks for investigating language through large collections of naturally occurring texts. While the discipline has traditionally focused on canonical written and spoken genres, it has progressively expanded to include digital, popular cultural, and multimodal texts (McEnery & Hardie, 2012). Song lyrics constitute a particularly valuable corpus-linguistic resource because they combine features of spoken and written language, including colloquial vocabulary, rhythmic structure, emotional content, and culturally embedded reference systems.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Researchers have examined lyrical content to investigate vocabulary usage, emotional expression, cultural identity, and discourse patterns (Kreyer & Mukherjee, 2007; Orejuela, 2015). Tsur and Gafni (2019) demonstrate that prosodic and phonological properties of lyrics interact significantly with semantic content, contributing to the emotional impact of songs. Brackett (2016) situates lyrical analysis within broader musicological and cultural frameworks, arguing that genre distinctions are as much linguistic as they are musical.

2.2 Lexical Diversity in Language Analysis

Lexical diversity, broadly understood as the variety of vocabulary used within a text, is a core concept in corpus linguistics and language assessment. The Type-Token Ratio (TTR) remains one of the most widely employed indices of lexical diversity, computed as the ratio of unique word forms (types) to total word occurrences (tokens) within a text. A higher TTR indicates a broader and more varied vocabulary (Johansson, 2009). Prior research has demonstrated that lexical diversity serves as an indicator of cognitive complexity, educational attainment, and stylistic sophistication across discourse types (Kyle & Crossley, 2015; Zenker & Kyle, 2021).

Within musical research, lexical diversity has been employed to distinguish between lyrical styles and compositional practices. Rap music has consistently been identified as exhibiting greater lexical density than other popular genres, a characteristic linked to the genre's tradition of complex wordplay, narrative elaboration, and sociolinguistic distinctiveness (Bradley, 2017; Herd, 2016). Pop and Rock, by contrast, tend toward higher degrees of lexical repetition through the structural use of choruses, hooks, and refrains designed to maximize memorability and commercial appeal.

2.3 Sentiment Analysis in Linguistic Research

Sentiment analysis is a computational technique for identifying the emotional orientation of texts along a positive-negative polarity continuum. Automated sentiment analysis tools such as TextBlob, VADER, and transformer-based language models have been widely applied to social media data, product reviews, news articles, and literary corpora (Hutto & Gilbert, 2014; Liu, 2015). In linguistic research, sentiment analysis provides a quantitative means of examining how emotional expression varies across genres, registers, and discourse contexts.

Applied to musical lyrics, sentiment analysis has revealed that genre-level emotional profiles differ in systematic ways. Pop music tends to exhibit higher positive sentiment scores, consistent with its commercial orientation toward emotional uplift and romantic optimism (Malheiro et al., 2016). Rock and Rap, by contrast, demonstrate greater emotional variability, incorporating negative, ambivalent, and socially critical sentiments alongside positive emotional expression. Kolchyna et al. (2015) argue that the sentiment patterns of song lyrics are culturally and generically conditioned, reflecting the social and communicative functions of different musical traditions.

2.4 Topic Modeling and Thematic Analysis

Topic modeling is an unsupervised machine learning approach to discovering latent thematic structures within large text corpora. Latent Dirichlet Allocation (LDA), originally proposed by Blei et al. (2003), remains the most widely used topic-modeling algorithm and has been productively applied to political discourse, news archives, literary texts, and social media data (Blei, 2012; Maier et al., 2018). By analyzing patterns of word co-occurrence, LDA enables researchers to identify recurring semantic themes without requiring prior knowledge of the corpus content.

Applied to song lyrics, topic modeling has revealed genre-specific thematic tendencies that align with broader cultural and sociological understandings of musical genres. Pop lyrics tend to cluster around themes of romantic relationships and interpersonal emotion. Rock

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JOURNAL OF APPLIED LINGUISTICS AND TESOL (JALT)
Vol.9.No.2 2026

exhibits both narrative-driven and introspective thematic orientations. Rap lyrics consistently generate topic clusters centered on socio-cultural commentary, urban experience, identity, and personal struggle (Fell & Sporleder, 2014). These findings underscore the utility of computational approaches in providing systematic, reproducible evidence for genre-based linguistic distinctions.

2.5 Language and Ideology in Musical Discourse

Beyond purely formal linguistic analysis, scholars have examined how song lyrics function as sites of ideological construction, emotional legitimization, and cultural identity formation. Fairclough (2015) emphasizes that linguistic choices in any discourse type reflect and reproduce underlying ideological frameworks. Within the specific domain of political music, Muhammad et al. (2025) demonstrate through Critical Discourse Analysis how political parties in Pakistan deploy emotional appeals, hypothetical futures, and altruistic framing in their song lyrics to legitimize their ideological positions. Their application of Halliday's (2014) transitivity model reveals how material, mental, relational, and existential processes serve distinct persuasive functions within lyrical texts.

While the present study focuses on popular rather than political music, it engages with the same fundamental question of how language encodes emotional and thematic meaning across different lyrical traditions. The computational methods employed here complement qualitative discourse-analytic approaches by providing statistically rigorous evidence for genre-based differences in emotional orientation and thematic content.

3. Methodology

3.1 Research Design

This study adopts a quantitative corpus-linguistic research design to investigate lexical diversity, emotional sentiment, and thematic patterns in contemporary English song lyrics. Computational linguistic techniques were employed to analyze a large corpus of lyrics from three popular musical genres: Pop, Rock, and Rap. The study combines descriptive statistics, inferential statistical testing, and machine-learning-based topic modeling to identify linguistic differences across genres. This multi-method approach reflects best practices in corpus linguistics for ensuring the reliability and replicability of findings (McEnery & Hardie, 2012; Biber & Egbert, 2018).

3.2 Corpus and Data Collection

The dataset was obtained from a publicly available song lyrics corpus hosted on Kaggle, containing English-language lyrics categorized by musical genre. To ensure balanced representation across genres, a sample of 3,000 songs was selected, consisting of 1,000 Pop songs, 1,000 Rock songs, and 1,000 Rap songs. Only songs with complete lyric transcripts and valid genre labels were retained for analysis. The resulting corpus represents a diverse collection of contemporary English-language music from across the late 20th and early 21st centuries.

Table 1
Corpus Composition

Genre	Number of Songs
Pop	1,000
Rock	1,000
Rap	1,000
Total	3,000

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.3 Data Preprocessing

Prior to analysis, the corpus underwent a series of standardized preprocessing procedures implemented in Python. The preprocessing pipeline included: (a) removal of missing or incomplete lyric entries; (b) conversion of all text to lowercase; (c) removal of punctuation marks and special characters; (d) tokenization of lyrics into individual word tokens; (e) removal of common English stopwords for keyword frequency and topic modeling analyses; and (f) construction of clean textual representations suitable for subsequent computational analysis. These procedures were performed to minimize corpus noise and ensure analytical consistency across all three genre sub-corpora.

3.4 Computational Tools

All analyses were conducted using Python (version 3.10) in Google Colab. The following libraries were employed:

Table 2
Python Libraries Used

Library	Purpose
Pandas	Data manipulation and corpus management
NumPy	Numerical computation
NLTK	Tokenization and text preprocessing
TextBlob	Sentiment analysis
Gensim	Topic modeling (LDA)
Scikit-learn	Machine learning support
Matplotlib	Data visualization
Seaborn	Statistical visualization
SciPy	Inferential statistical analysis
Statsmodels	ANOVA and Tukey HSD testing

3.5 Lexical Diversity Analysis

To measure vocabulary richness, the study employed the Type-Token Ratio (TTR), one of the most widely used indicators of lexical diversity in corpus linguistics (Kyle & Crossley, 2015; Zenker & Kyle, 2021). The TTR was calculated by dividing the number of unique words (types) by the total number of words (tokens) within each song:

$$TTR = Types / Tokens$$

Higher TTR values indicate greater lexical diversity and a broader range of vocabulary usage. Mean TTR scores for each genre were subsequently calculated and compared.

3.6 Sentiment Analysis

Emotional orientation was examined using the TextBlob sentiment analysis framework (Loria, 2018), which assigns each text a polarity score ranging from -1.0 (highly negative) to 0.0 (neutral) to +1.0 (highly positive). Each song lyric received an individual polarity score, and genre-level averages were calculated. TextBlob's lexicon-based approach to sentiment scoring has been validated in multiple prior studies of informal and creative texts (Kolchyna et al., 2015; Malheiro et al., 2016).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.7 Keyword Frequency Analysis

Following stopword removal, word frequencies were calculated using Python's NLTK frequency distribution tools. The five most common content words per genre were extracted and compared to identify dominant lexical patterns and thematic orientations. Frequency analysis provides a foundational quantitative indicator of topical focus and vocabulary prioritization within each genre sub-corpus (McEnery & Hardie, 2012).

3.8 Topic Modeling

To uncover latent semantic structures within the corpus, Latent Dirichlet Allocation (LDA) was employed using the Gensim library (Blei, 2012; Řehůřek & Sojka, 2010). LDA is an unsupervised probabilistic model that identifies hidden thematic patterns by examining word co-occurrence relationships within and across documents. The preprocessed lyric corpus was transformed into a document-term matrix. Following iterative parameter tuning, four topics were extracted and interpreted based on their highest-probability keywords. These topics were subsequently analyzed to identify recurring themes and genre-specific semantic tendencies.

3.9 Statistical Analysis

To determine whether observed inter-genre differences were statistically significant, one-way Analysis of Variance (ANOVA) was conducted separately for lexical diversity (TTR) and sentiment polarity. Where significant effects were detected ($p < .05$), Tukey's Honest Significant Difference (HSD) post-hoc test was applied to identify which specific genre pairs differed significantly from one another. A significance threshold of $p < .05$ was adopted throughout the study.

3.10 Ethical Considerations

The study utilized publicly available song lyrics obtained from an existing online dataset. No human participants were involved in the research, and no personal or sensitive information was collected or processed. Ethical risks were consequently minimal. The research focused exclusively on aggregated linguistic patterns and thematic structures within the corpus.

4. Quantitative Analysis and Findings

The quantitative analysis was conducted on the balanced corpus of 3,000 English song lyrics. To address the two research objectives, lexical diversity measurement, sentiment polarity scoring, keyword frequency analysis, and topic modeling were performed using Python-based computational tools. Inferential statistical tests were subsequently applied to evaluate the significance of inter-genre differences.

4.1 Lexical Diversity and Sentiment Analysis

The consolidated metrics for lexical diversity (TTR) and sentiment polarity across the three genres are presented in Table 3.

Table 3

Lexical Diversity and Sentiment Metrics by Genre

Genre	Mean TTR	Mean Sentiment Polarity	Emotional Orientation
Pop	0.4707	0.0668	Positive
Rock	0.4724	0.0376	Moderately Positive
Rap	0.5070	0.0374	Moderately Positive

As shown in Table 3, Rap recorded the highest lexical diversity score (M = 0.5070), indicating that for every 100 words in a Rap lyric, approximately 51 are unique. Pop (M = 0.4707) and Rock (M = 0.4724) exhibited substantially lower and nearly identical TTR values.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

From a corpus-linguistic standpoint, these metrics confirm that Rap relies on a broader vocabulary to sustain complex rhyme schemes, rapid narrative delivery, and dense thematic content (Bradley, 2017; Herd, 2016). Pop and Rock, by contrast, favor highly repetitive lexical structures—choruses, hooks, and refrains—to maximize audience retention and commercial appeal.

Regarding sentiment polarity, Pop demonstrated the highest positive score (M = 0.0668), while Rock (M = 0.0376) and Rap (M = 0.0374) shared nearly identical and substantially lower values. All three genres scored marginally above the neutral threshold of 0.00, indicating a slight overall positive bias across the corpus. Pop's relatively elevated sentiment polarity reflects its commercial orientation toward emotionally optimistic and romantically framed content (Malheiro et al., 2016). Rock and Rap, by contrast, incorporate darker, more socially critical, and emotionally ambivalent themes that depress the average positive polarity score.

4.2 Statistical Analysis of Lexical Diversity

Table 4

One-Way ANOVA Results for Lexical Diversity

Source	F-value	p-value	Significance
Genre	21.01	< .001	Significant

The one-way ANOVA revealed a statistically significant effect of genre on lexical diversity, $F(2, 2997) = 21.01$, $p < .001$, confirming that vocabulary richness differs significantly across the three musical genres.

Table 5

Tukey HSD Post-Hoc Comparisons for Lexical Diversity

Comparison	Mean Difference	p-value	Significant
Pop vs. Rap	0.0363	< .001	Yes
Pop vs. Rock	0.0017	.963	No
Rap vs. Rock	0.0347	< .001	Yes

Tukey's HSD post-hoc analysis reveals that Rap differs significantly from both Pop and Rock in lexical diversity, while no significant difference exists between Pop and Rock. The overall ANOVA effect is therefore attributable primarily to Rap's substantially elevated vocabulary richness relative to the other two genres.

4.3 Statistical Analysis of Sentiment Polarity

Table 6

One-Way ANOVA Results for Sentiment Polarity

Source	F-value	p-value	Significance
Genre	10.17	< .001	Significant

A one-way ANOVA confirmed a statistically significant effect of genre on sentiment polarity, $F(2, 2997) = 10.17$, $p < .001$, indicating that emotional orientation varies systematically across musical genres.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Table 7
Tukey HSD Post-Hoc Comparisons for Sentiment Polarity

Comparison	Mean Difference	p-value	Significant
Pop vs. Rap	-0.0294	< .001	Yes
Pop vs. Rock	-0.0292	< .001	Yes
Rap vs. Rock	0.0002	.9996	No

Post-hoc analysis demonstrates that Pop differs significantly from both Rap and Rock in emotional orientation, while Rap and Rock do not differ significantly from each other. These findings confirm that Pop music consistently deploys more positively valenced emotional language than either of the other two genres.

4.4 Keyword Frequency Analysis

Table 8 presents the five most frequently occurring content words within each genre sub-corpus following stopword removal.

Table 8
Top Five Content Words by Genre

Rank	Pop	Freq.	Rock	Freq.	Rap	Freq.
1	love	1,189	know	1,136	know	2,052
2	know	1,054	love	805	aint	1,364
3	oh	1,051	one	705	na	1,242
4	na	798	never	693	go	1,238
5	go	772	oh	688	cause	1,187

The dominance of "love" as the highest-frequency content word in Pop directly corresponds with the genre's elevated positive sentiment score and its commercial focus on romantic and interpersonal themes (Malheiro et al., 2016). Rock demonstrates a greater emphasis on reflective and negatively inflected vocabulary, with words such as "one" and "never" signaling introspective and existential orientations. The Rap corpus is notably characterized by substantially higher absolute frequency counts across all top words—"know" appears 2,052 times in Rap versus 1,136 in Rock—mathematically reinforcing the immense linguistic density characteristic of the genre (Bradley, 2017). The presence of informal markers such as "aint" and "cause" in Rap's top words further reflects the genre's strong association with vernacular and colloquial language use.

4.5 Topic Modeling Analysis

The LDA topic modeling procedure identified four major latent thematic structures within the lyric corpus, as detailed in Table 9.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 9
LDA-Derived Semantic Themes

Topic	Representative Keywords	Interpretation
Topic 1	go, take, time, night, us, let	Narrative Action and Movement
Topic 2	love, yeah, come, ooh, little, baby	Romance and Interpersonal Relationships
Topic 3	know, love, never, time, one, see	Reflection and Existential Thought
Topic 4	shit, man, cause, life, know, aint	Socio-Cultural Realism and Personal Struggle

Topic 1 (Narrative Action) reflects physical, situational, and chronological storytelling patterns particularly prevalent in Rock and narrative-driven Pop tracks. Topic 2 (Romance and Interpersonal Relationships) represents the thematic core of commercial Pop music, encompassing affection, call-and-response vocalizations, and standard relationship narratives. Topic 3 (Reflective Existentialism) captures introspective and philosophical content characteristic of deeper Rock compositions. Topic 4 (Socio-Cultural Realism) is defined by strong vernacular markers and existential nouns, mapping directly to the gritty, real-world narrative frameworks characteristic of Rap (Fell & Sporleder, 2014; Herd, 2016).

The distribution of topics across genres confirms genre-specific thematic priorities: romantic and interpersonal themes predominate in Pop, while socio-cultural realism is most strongly associated with Rap. Rock occupies an intermediate position characterized by narrative and reflective content. These thematic distinctions align closely with findings from prior qualitative and quantitative analyses of genre-based lyrical content and are consistent with broader sociolinguistic understandings of how different musical communities use language to construct identities, relationships, and social realities (Brackett, 2016; Muhammad et al., 2025).

5. Discussion

The findings of this study provide robust quantitative evidence for systematic linguistic differences among Pop, Rock, and Rap song lyrics across the dimensions of lexical diversity, emotional sentiment, vocabulary use, and thematic content. Several key patterns merit extended discussion.

First, the significantly elevated lexical diversity of Rap relative to Pop and Rock confirms prior qualitative and theoretical accounts of the genre's linguistic distinctiveness. Bradley (2017) characterizes Rap as fundamentally a literary art form in which the breadth and precision of vocabulary serve as primary aesthetic resources. The TTR data generated in the present study supply quantitative empirical support for this characterization. Herd (2016) similarly argues that the linguistic complexity of Rap lyrics reflects the genre's origins in African American oral traditions that place high cultural value on verbal ingenuity, wordplay, and narrative craft. The present data suggest that this vocabulary richness is not merely stylistic but constitutes a defining computational signature of the genre.

Second, Pop music's significantly higher positive sentiment polarity is consistent with both commercial music industry logic and prior corpus-linguistic findings. Malheiro et al. (2016) demonstrate that the acoustic and lyrical properties of Pop music are systematically optimized to elicit positive affective responses in listeners. The present sentiment data confirm that this optimization extends to the lexical and propositional content of lyrics, with Pop consistently employing more positively valenced language than Rock or Rap. Kolchyna et al. (2015) argue that genre-level emotional profiles reflect the social and communicative functions served by different musical traditions: Pop's function as commercially oriented entertainment

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

drives its emotional optimism, while Rock's and Rap's social-commentary functions permit and often require emotional negativity and complexity.

Third, the topic modeling results underscore the social and cultural dimensions of genre-based linguistic variation. The identification of a dedicated Socio-Cultural Realism topic (Topic 4) as a defining feature of Rap's thematic profile is consistent with the genre's widely recognized role as a vehicle for social critique, cultural commentary, and the documentation of marginalized experience. This finding resonates with Muhammad et al.'s (2025) demonstration, in a different but related musical context, that song lyrics serve as powerful instruments of ideological expression and emotional legitimization. While their analysis focuses on Pakistani political songs, the underlying linguistic mechanisms—emotional language, identity construction, and thematic framing—operate similarly across musical traditions.

Fourth, the absence of a statistically significant difference in lexical diversity between Pop and Rock, and between Rap and Rock in sentiment polarity, reveals important nuances in the genre landscape. Rock occupies a genuinely intermediate position—more lexically varied than Pop but less so than Rap, and emotionally more complex than Pop but without Rap's extreme socio-cultural grounding. This intermediate profile suggests that Rock functions as a bridge genre, combining elements of commercial emotional appeal with greater linguistic and thematic depth than Pop alone offers. Biber and Egbert (2018) note that register variation is rarely binary but more typically distributed along continuous gradients, and the present data confirm this for musical genres as well.

6. Conclusion

This study has demonstrated that Pop, Rock, and Rap song lyrics differ systematically and significantly in their lexical, emotional, and thematic properties. Through a multi-method computational approach combining TTR-based lexical diversity measurement, TextBlob sentiment analysis, keyword frequency extraction, LDA topic modeling, and ANOVA-based inferential statistics, the study has produced a comprehensive quantitative profile of linguistic variation across three major musical genres.

Rap exhibits the highest lexical diversity, driven by its tradition of complex wordplay and narrative elaboration. Pop demonstrates the most consistently positive emotional language, reflecting its commercial orientation toward audience uplift and romantic themes. Rock occupies an intermediate position across both lexical and emotional dimensions. Topic modeling reveals four major thematic clusters—Narrative Action, Romance, Reflective Existentialism, and Socio-Cultural Realism—whose distribution across genres reflects and reinforces the social and cultural functions served by each musical tradition.

The study makes several contributions to corpus linguistics and the computational analysis of popular cultural texts. It provides statistically robust, quantitatively grounded evidence for genre-level linguistic distinctions that have often been characterized impressionistically. It demonstrates the methodological utility of combining lexical, sentiment, frequency, and topic-modeling analyses within a unified framework. And it positions the linguistic analysis of song lyrics within broader conversations about the relationships between language, emotion, culture, and ideology—conversations to which qualitative discourse analysts such as Muhammad et al. (2025) and Fairclough (2015) have made foundational contributions.

Several limitations should be noted. TTR is sensitive to text length, and genre differences in average song length may partially confound the lexical diversity comparisons. TextBlob's lexicon-based sentiment scoring may not capture the full complexity of figurative, ironic, or genre-specific emotional expression in lyrics. The corpus, while balanced at 1,000 songs per genre, does not control for temporal variation within genres or for sub-generic diversity within Pop, Rock, and Rap.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Future research should address these limitations by applying length-normalized lexical diversity indices (e.g., the Moving Average Type-Token Ratio), employing transformer-based sentiment models trained on musical data, and investigating sub-generic variation within broad genre categories. Cross-linguistic extensions of the present framework to non-English language lyrical corpora would also contribute significantly to understanding the universality or cultural specificity of genre-based linguistic patterns.

References

Biber, D., & Egbert, J. (2018). Register variation online. Cambridge University Press. https://doi.org/10.1017/9781316388228

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Brackett, D. (2016). Categorizing sound: Genre and twentieth-century popular music. University of California Press. https://doi.org/10.1525/9780520966796

Bradley, A. (2017). Book of rhymes: The poetics of hip hop (2nd ed.). Basic Books.

Fell, M., & Sporleder, C. (2014). Lyrics-based analysis and classification of music. In Proceedings of COLING 2014 (pp. 620–631). Association for Computational Linguistics. https://aclanthology.org/C14-1059

Fairclough, N. (2015). Language and power (3rd ed.). Routledge. https://doi.org/10.4324/9781315838250

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). Halliday's introduction to functional grammar (4th ed.). Routledge. https://doi.org/10.4324/9780203783771

Herd, J. (2016). Change gonna come: Music, race, and the soul of America (Updated ed.). University of Michigan Press.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (pp. 216–225). AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. Working Papers in Linguistics, 53, 61–79.

Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. In Proceedings of the 24th International Conference on Artificial Intelligence (pp. 1271–1277). AAAI Press. https://arxiv.org/abs/1507.00955

Kreyer, R., & Mukherjee, J. (2007). The style of pop song lyrics: A corpus-linguistic pilot study. Anglia, 125(1), 31–58. https://doi.org/10.1515/ANGL.2007.31

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly, 49(4), 757–786. https://doi.org/10.1002/tesq.194

Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Loria, S. (2018). TextBlob documentation (Version 0.15.3). https://textblob.readthedocs.io

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Wurzer, U., Potthast, M., Voigt, M., & Gronemeyer, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures, 12(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Emotionally relevant features for classification and regression of music lyrics. IEEE Transactions on Affective Computing, 9(2), 240–254. https://doi.org/10.1109/TAFFC.2016.2598569

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Muhammad, S., Shakoor, Z., & Ikram, H. (2025). Language of legitimization in analysis of political songs (PPP and PTI parties). International Journal of Systems Science Bulletin, 3(11), 213–225. https://doi.org/10.5281/zenodo.17587803

Orejuela, F. (2015). Rap and hip hop culture. Oxford University Press.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). ELRA. https://is.muni.cz/publication/884893/en

Tsur, R., & Gafni, D. (2019). Phonetics and the poem. Cambridge Scholars Publishing.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. Assessing Writing, 47, 100505. https://doi.org/10.1016/j.asw.2020.100505