Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10(2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

STYLOMETRY OF LEXICAL PATTERNS AND IDENTITY FORMATION IN EVERETT'S ERASURE AND I AM NOT SIDNEY POITIER: A CORPUS STYLISTICS STUDY

Dania Hameed
Department of Linguistics and Communications (DLC), University of Management and Technology, Sialkot, Pakistan

Samira Amjad
Department of Linguistics and Communications (DLC), University of Management and Technology, Sialkot, Pakistan

Muhammad Zahid
Associate Professor of English, Higher Education Department, Punjab, Pakistan

Muhammad Nadeem Anwar (Corresponding Author)
Higher Education Department, Punjab, Pakistan
mnadeemanwar@gcslahore.edu.pk

ABSTRACT

This paper evaluates the stylistic features and lexical patterns in Percival Everett's novels Erasure and I Am Not Sidney Poitier to find out the formation of identity through a corpus stylometric lens. Using a digital text mining software, Voyant Tools, this research applies computational methods to extract word frequency, vocabulary density, collocation patterns, and syntactic features that contribute to identity formation in Everett's fiction. The evaluation reflects significant variations in lexical patterns, repetitions, and collocations across the two novels. Erasure reveals a linguistically dense and emotionally charged style, interrogating the exploitation of Black identity through metafictional narrative. In contrast, I Am Not Sidney Poitier is more dialogic and has an ironic tone, illustrating a protagonist's journey through racial and social misrecognition. Through this integration of digital humanities and literary theory, the study highlights how Everett's language resists fixed notions of Blackness and authorship, offering a dynamic model for identity construction in contemporary African American literature.

Keywords: Stylometry, Lexical Patterns, Identity Formation, Voyant Tools, Afro-American Literature

Introduction

Stylometry, a branch of applied linguistics, investigates the subtle use of language in texts, with particular emphasis on the interconnection between linguistic form and literary function. It studies how elements such as diction, syntax, figurative

Online ISSN 3007-3197
Print ISSN 3007-3189

language, and narrative structure participate in the origination of meaning within a text. Stylometry is a vital instrument in literary studies, used to investigate the ways in which authors manipulate language to elicit emotions, create characters, and establish themes. By examining sentence structure and syntactic components, stylometric analysis reveals that these formal features significantly contribute to an author's unique style and the literary purpose of their works (Hollingsworth, 2012). According to Burrows (2002), lexical repetition allows scholars to identify subtle differences in stylistic fingerprinting, which are crucial for understanding how identity is constructed within a text. Lexical patterns, such as word frequency, collocations, frequent words and vocabulary density, are crucial for stylometric and corpus-based analyses. They function as tangible stylistic markers that reveal not only the writer's distinctive narrative style but also themes and ideologies running through a text. In stylometric research, repeated lexical items serve as quantitative indicators of style, often unveiling patterns not visible through traditional close reading (Burrows, 2002; Craig & Kinney, 2009). McEnery and Hardie (2012) stress that corpus tools allow researchers to process large-scale linguistic data, uncovering lexical trends that shape the emotional and ideological underpinnings of literary texts.

Within this field, stylistics emerges as a computational sub-discipline that applies the use of computers and different software tools to the analysis of style. By evaluating quantitative linguistic features, such as the most frequent words, the sentence length, lexical diversity, and collocation patterns, stylistics enables an objective and repeatable analysis of authorial style. (Boukhaled et al., 2015). This scientific method permits researchers to correlate text and determine the stylistic endorsement based on measured markers rather than exclusively on individual perception.

Identity is not a collection of intrinsic characteristics but rather a social and linguistically constructed process. According to Hall (1996), identity emerges from active cultural discourse and is always situated within a historical and ideological framework. It is never isolated or constructed but rather arises from representation, language, and difference. Identity is understood as a dynamic and continuous process of "becoming" rather than a permanent essence. It is historically contextual and is always changing as a result of cultural and power dynamics, as well as discourse and representations. As Hall (1996) explains, identities are "not a priori; they do not transcend culture, history, time, and place," but rather are created within specific social and ideological circumstances (Mirza, 2020).

The current study focuses on contemporary African American author Percival Everett's two metafictional novels: Erasure (2001) and I Am Not Sidney Poitier (2009). Both novels are exemplary of Everett's interest in metafictional techniques and postmodern satire, especially in their interrogation of race, authorship, and the politics of literary representation. These works rebuild cultural and literary hypotheses by self-consciously playing with narrative form, genre expectations, and reader assumptions (Everett, 2001; Everett, 2009).

The study incorporates a corpus stylistics approach, utilizing computerized digital tools such as Voyant tools to conduct a data-driven analysis of Everett's lexical and stylistic choices. Corpus stylistics combines traditional close reading with computational methods to uncover patterns that might not be immediately apparent

Online ISSN 3007-3197
Print ISSN 3007-3189

through manual analysis. By analyzing word frequency, repetitions, collocations and vocabulary density across both novels, the research aims to reveal how linguistic features contribute to the construction of racial identity, authorial voice, and metafictional commentary in Everett's fiction.

Everett is very famous for his writings; he is known worldwide for his themes and writing style. He got the Nobel Prize for his novel Erasure, in which he discusses the theme of identity formation and confusion. His novel, I Am Not Sidney Poitier, is also about identity crises. This research evaluates his lexical choices and how he builds his characters and their struggles for identity formation.

Problem Statement

While stylometry and corpus stylistics have proven valuable in authorship attribution and literary pattern recognition, there remains a significant gap in their application to postmodern African American literature, particularly in analyzing identity formation. The Erasure and I Am Not Sidney Poitier by Everett have been extensively discussed in terms of race and metafiction, but few studies utilize digital tools to systematically investigate how his lexical choices reflect and construct identity. This research addresses the lack of empirical, data-driven exploration of Everett's stylistic practices, aiming to bridge literary theory and stylistic computational analysis to generate new insights into his narrative strategies.

Research Objectives

To apply Voyant Tools to identify and compare significant lexical patterns and stylistic features in Erasure and I Am Not Sidney Poitier.

To explore the way variations in vocabulary, word choices, and stylistic techniques contribute to different representations of identity in the two novels.

Research Questions

How can Voyant Tools assist in identifying and comparing significant lexical patterns and stylistic features in Erasure and I Am Not Sidney Poitier?

How do variations in vocabulary and word frequency between the two novels illustrate different representations of identity?

Significance of the Study

This research highlights the innovative integration of stylometric tools with literary analysis to study identity construction. It elaborates on Everett's style by offering a new method of textual investigation, enhances the application of digital humanities in contemporary text, and provides deeper insight into how linguistic patterns contribute to the discourse of race and identity formation in present fiction.

Literature Review

Stylometry, defined as the quantitative study of literary style, has evolved significantly since its introduction by Mosteller and Wallace (1964), who used data-based methods to resolve attribution in the Federalist Papers. Holmes (1998) expanded the field of stylometry, highlighting the role of computational methods in examine stylistic features such as sentence length, lexical density, and word frequency. With the rise of digital humanities, instruments like Voyant Tools (Sinclair

Online ISSN 3007-3197
Print ISSN 3007-3189

& Rockwell, 2016) have assisted the large-scale study of literary texts. These tools offer visualizations of word frequencies, keyword in context (KWIC), and collocation patterns, allowing researchers to move beyond traditional close reading. Moretti (2013) advocated for "distant reading," a method that favors large scale analysis to discover more inclusive patterns in literature especially useful when analyzing huge texts or corpora.

Percival Everett's Erasure (2001) is widely recognized for its satire on the commercialization of African American identity in the literary platform. The novel centers on Thelonious "Monk" Ellison, a Black academic who adopts the stage name "Stagg R. Leigh" to write a satire of conventional Black fiction, which paradoxically becomes a commercial success. Scholars such as King (2011) and Young (2010) have examined how Erasure critiques racial fundamentalism and literary filtering. Recent studies have resumed attention in Erasure through the lens of despondent satire and parody. For example, Larrue (2018) discusses how Everett critiques dominant racial narratives by confronting the reader with ironic distance and literary absurdity. The 2023 film adaptation American Fiction has also reignited scholarly and public discussions about the novel's themes of racial identity and authorial authenticity (Christman, 2023).

I Am Not Sidney Poitier (2009) furthers Everett's critique of race and representation, using intertextuality and absurdist narrative to challenge perceptions of Black identity. The protagonist, named "Not Sidney Poitier," is continually mistaken for the actor Sidney Poitier and finds himself in surreal situations that parody narratives and American racial politics. Raynaud (2013) explores how Everett employs nonsense and repetition to expose the constructed nature of identity. Similarly, Schmidt (2021) interprets the novel as a multi-layered parody of post-Blackness—a cultural movement that rejects monolithic portrayals of Black experience.

In African American literature, identity formation often involves the negotiation of imposed stereotypes and cultural expectations. As Bhabha (1994) suggests, the notion of the "hybrid" identity—emerging from the interaction of cultures, histories, and narratives—challenges essentialist notions of race and nation. Despite the rich thematic content of Everett's novels, few studies have applied stylometric or corpus-based methods to analyze how identity is constructed linguistically. Scholars like Jockers (2013) and Hoover (2010) have illustrated the potential of computational analysis in revealing underlying stylistic choices, yet African American postmodern literature remains underexplored in this domain.

The present research study, therefore, seeks to address this gap by using Voyant Tools to perform a comparative lexical analysis of Erasure and I Am Not Sidney Poitier, focusing on vocabulary distribution, word repetition, and identity-linked keywords. By doing so, it offers a fresh, empirical perspective on how Everett constructs identity through style.

Methodology

This study uses a mixed-methods descriptive approach, integrating qualitative literary analysis with quantitative data from Voyant Tools. Digital copies of Percival Everett's books Erasure and I Am Not Sidney Poitier were used to create a small corpus. Word

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

clouds, total and unique word counts, vocabulary density, average sentence length, collocations, and the thirty most common terms from each novel were among the numerical data and visualizations that were extracted from the texts using Voyant's Summary, Cirrus, Terms, and Collocates tools. Voyant Tools was the main platform for lexical and stylistic analysis of the books, and the results were then qualitatively analyzed using the Hermeneutica theory as put out by Sinclair and Rockwell (2016).

Results and Analysis

The objective of the study was to identify and compare significant lexical patterns and stylistic features in Erasure and I Am Not Sidney Poitier. To realize the objective stylometry quantitative analysis of both the novels has been done through Voyant tools.

Figure 1: Erasure (2001) through Voyant Summary tool

This corpus has 1 document with 82,976 total words and 10,835 unique word forms. Created 32 seconds ago.

Vocabulary Density: 0.131

Readability Index: 7.849

Average Words Per Sentence: 10.5

Most frequent words in the corpus:

- say (721); said (481); like (291); don't (276); i'm (275); mother (271); know (255); just (247); look (177); got (158); asked (146); come (142); it's (141); shit (137); fuck (136); think (135); ain't (134); lorraine (132); house (132); man (130); good (126); gone (126); way (125); looked (125); didn't (118); right (114); ax (109); father (106); tell (104); want (103)

Figure 1 shows key lexical features of the novel Erasure. The data consist of 82,976 total words and 10,835 unique word forms. The vocabulary density is 0.131, suggesting that about 13.1% of the words in the text are unique, reflecting moderate lexical density. The readability index is 7.849, which reflects that the text is accessible at a mid-level reading difficulty. The average length per sentence is 10.5, which indicates that the writer composes short sentences. This paper has analyzed thirty most frequent words from the novel Erasure. The most frequent words play a dynamic role in reflecting the theme, writing style, and narration of any text.

Figure 2: Erasure through Voyant Cirus tool

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10(2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The data reported in Figure 2 was represented as a word cloud, placing emphasis on those words that were the most frequent, with larger fonts indicating greater frequency. Notably, say and said to demonstrate the author's dialogic style. Slang such as shit and fuck are singled out through the use of red to denote emotional frustration and anxiety of identity. Nouns such as mother and father indicate family relationships. Verbs such as asked and come add dynamism to the narrative, while negation words such as don't tend to show a transgression of racial norms
Personal pronouns such as I'm draw attention to the protagonist's struggle for identity.

Figure 3: Erasure through Voyant Term tool

	Term	Count
1	say	721
2	said	481
3	like	291
4	don't	276
5	i'm	275
6	mother	271
7	know	255
8	just	247
9	look	177
10	got	158
11	asked	146
12	come	142
13	it's	141
14	shit	137
15	fuck	136
16	think	135
17	ain't	134
18	lorraine	132
19	house	132
20	man	130
21	good	126
22	gone	126
23	way	125
24	looked	125
25	didn't	118
26	right	114
27	ax	109
28	father	106
29	tell	104
30	want	103

Figure 3 displays the figures in the form of frequency distribution of the common words and their respective frequencies together with their proper serial numbers, indicating the significance of the words. The order of the three commonest words is say (721), said (481), like (291), these being verbs for speech and signifying likeness

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10(2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

or preferences. Among the others are the pronouns and contractions, as don't (276), I'm (275), and mother (271), father (106), suggesting that many of the phrases are of personal or family significance. Among the verbs are know, ask, come and look, referring to action and the intelligence of thought. The use of such words as gonna (126), wanna (103), indicate the conversational nature of the text or of the tractate.

Figure 4: Erasure through Voyant Collocate tool

Term	Collocate	Count (context)
black	black	28
black	look	6
black	life	5
black	watch	4
black	turtleneck	4
black	stagg	4

Figure 4 shows the collocation of selected words. Here, the word black is selected and given to the collocate tool to find out the word black collocation from the text Erasure. The collocations also played a very crucial role in giving a deep insight into the text and its themes. As the collocations like black black (28), black look (6), black life (5), and black stagg (4) tell the readers about the identity formation of the black people in the novel.

I Am Not Sidney Poiter (2009)

Figure 4: I am not Sidney Poitier through Voyant Summary tool

This corpus has 1 document with 71,893 total words and 9,855 unique word forms. Created about 21 minutes ago.

Vocabulary Density: 0.137

Readability Index: 8.371

Average Words Per Sentence: 11.0

Most frequent words in the corpus:

- said (800); like (310); don't (236); know (234); asked (212); looked (192); just (175); i'm (174); didn't (146); say (141); sidney (137); money (129); that's (124); man (121); it's (121); poitier (118); ted (117); think (114); way (104); want (104); good (103); got (99); room (95); you're (94); head (92); maggie (91); everett (91); white (90); thought (90); face (89)

asked, and looked together with a number of character names like Sidney, Poitier, and Maggie, which indicates the narrative style.

Figure 6: I am not Sidney Poitier through Voyant Cirus tool

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10(2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

In Figure 6, the data are presented in the form of a word cloud, which highlights words that have been used often, with larger fonts signifying increased usage. Words such as said and asked illustrate the means by which the writer has chosen to convey his narrative bent. Other words displayed in larger print are like, know, asked, looked, Sidney, and Poitier which indicates that these words have been used frequently. Their inclusion would also completely indicate that the prose is rich in conversational interchange and direct speech. The use of personal pronouns and contractions like: don't, I'm, didn't, and that's also gives credence to the high dialogic nature of the text, which mingles narrative progression with an extra character type narration.

Figure 7: I am not Sidney Poitier through Voyant Term tool

According to Figure 7, the most frequent words like money (129), white (90), just

(175) suggest the importance of the power of social class and race. As in the beginning of the novel, the protagonist belongs to the lower class, but as the story proceeds, he becomes part of the upper class and gains power. The people mistake him for the real Sidney Poitier, and despite the misrecognition of his identity, he enjoys that status and power, which is an absurdity in itself.

Figure 8: Collocate through the Voyant Collocate tool

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10(2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Term	Collocate	Count (context)
white	black	12
white	man	10
white	men	9
white	looked	5
white	white	4

To find any writing style and themes of any writing, collocations play a very important role in the research domain. Here, from the novel, we select a frequently used word white (90), which plays a drastic role in race and identity formation. The word white collocates here with black (12), man (10), men (9), looked (5), and white (4). Therefore, these collocations reflect the supremacy of white and the power it has in American society.

Table 1: Quantitative data of both novels by Voyant Summary tools

Novels	**Total words**	**Unique words**	**Vocabulary density**	**Readability index**	**Average words per sentence**
Erasure	82,976	10835	0.131	7.849	10.5
I am not Sidney Poitier	71,893	9855	0.137	8.371	11.0

A number of features contribute to the readability of both Erasure and I Am Not Sidney Poitier. These features, however, reveal some important similarities and differences in the two works, as seen in Table 1. The total number of words in Erasure is 82,976, with the total number of unique words being 10,835. The density of vocabulary is then .131. The total number of words in I Am Not Sidney Poitier is 71,893 and the number of unique words is 9855, resulting in a vocabulary density of .137. The readability indices denote that Erasure is somewhat more readable than I Am Not Sidney Poitier at 7.849 compared to 8.371. The average length of sentence in Erasure is shorter, with an average of 10.5 words per sentence, while I Am Not Sidney Poitier has an average of 11.0 words per sentence. These indices suggest that while the two works have a somewhat similar degree of lexical richness they differ in their word density in that I Am Not Sidney Poitier utilizes a denser and more complicated type of sentence structure which would account for the better readability score.

The second objective of the study was to analyze how do variation in vocabulary and word frequency of the Erasure and I Am Not a Sidney Poitier contribute to identity representations. The qualitative data analysis is as under:

Theme of Racial Identity and Authenticity in Erasure

The most frequent words give us one of the dominant themes of the novel which is racial identity and authenticity. The protagonist struggles throughout the plot for this identity

formation as a black writer but could not succeed. Then he writes under a fake name

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

about black being enough black and gets succeeded. The novel shows the struggle of being a black but not being able to get fit into the society's expectations to being enough black.

Identity Formation in I am not Sidney Poitier

The most frequent words depict the main theme of the novel, which is misrepresentation of the identity and identity crises. The main character, whose name is 'not Sidney Poiter', fights with the identity formation, but as he gains status and wealth, people start to respect him. Due to the money and upper status in the society, he is mistaken for the real Sidney Poitier. Rather than disregarding this issue, he embraces it and takes pleasure in the respect he receives. He accepts it and enjoys the respect people give him because from his childhood he faced the identity issues of being Black.

Conclusion

The stylometric analysis conducted using Voyant Tools offers empirical support for the argument that language in Percival Everett's novels operates not just as a narrative medium but as a critical instrument for identity construction and deconstruction. Erasure, with its emotionally intense and linguistically complex lexicon, interrogates the publishing industry's demand for racially authentic narratives and challenges essentialist representations of Blackness. Conversely, I Am Not Sidney Poitier uses repetition, irony, and metafictional references to destabilize singular identity frameworks, highlighting the performative and often absurd nature of racial categorization. Through word frequency trends, collocational patterns, and vocabulary density, this study demonstrates that Everett's stylistic techniques are intricately tied to the thematic exploration of race, authorship, and selfhood. The integration of corpus stylistics tools like Voyant not only validates literary interpretations with data but also opens new pathways for understanding how identity is encoded in narrative form. To sum up, this research bridges digital textual analysis and literary criticism, affirming that stylometry is a powerful methodology for uncovering the deeper ideological work performed by language in postmodern African American literature.

References

Bhabha, H. K. (1994). The location of culture. Routledge.

Boukhaled, M.-A., Frontini, F., Bourgne, G., & Ganascia, J.-G. (2015). Computational Study of Stylistics: A Clustering-based Interestingness Measure for Extracting Relevant Syntactic Patterns. International Journal of Computational Linguistics and Applications, 6(1), 45-62. ISSN 0976-0962. Retrieved October 6, 2025, from https://www.gelbukh.com/ijcla/2015-1/IJCLA-2015-1-pp-045-062-Computational-Study-of-Stylistics.pdf

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267–287.

Christman, P. (2023, December 8). Percival Everett is having a moment. Commonwealth

Online ISSN 3007-3197
Print ISSN 3007-3189

Magazine.
Craig, H., and Kinney, A. (2009). Shakespeare, Computers, and the Mystery of Authorship.
Cambridge University Press.
Everett, P. (2001). Erasure. Hanover, NH: University Press of New England
Everett, P. (2009). I Am Not Sidney Poitier. Minneapolis, MN: Graywolf Press.
Hall, S. (1996). Introduction: Who needs "identity"? In S. Hall & P. du Gay (Eds.), Questions of
cultural identity (pp. 1–17). Sage Publications.
Hollingsworth, C. D. (2012). Syntactic Stylometry: Using Sentence Structure for Authorship
Attribution [Master's thesis, University of Georgia]. Retrieved October 6, 2025, from https://www.ai.uga.edu/sites/default/files/inline-files/hollingsworth_charles_d_201208_ms.pdf
Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. Literary and
linguistic computing, 13(3), 111–117. https://doi.org/10.1093/llc/13.3.111
Hoover, D. (2010). Corpus Stylistics, Stylometry, and the Styles of Henry James. Style 41
(2): 174–203.
Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of
Illinois Press.
King, L. (2011). Race, authorship, and meta-identity in the novels of Percival Everett.
African American Review, 44(1–2), 35–49. https://doi.org/10.1353/afa.2011.0007
Larrue, J. (2018). The pitfalls of parody: Melancholic satire in Percival Everett's Erasure.
In C. Larrue (Ed.), Le sens et ses fictions (pp. 219–236). Presses Universitaires François-Rabelais. https://books.openedition.org/pufr/5459
McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice.
Cambridge University Press.
Mirza, S. M. (2020). Stuart Hall: 'Cultural Identity and Diaspora'. Retrieved October 6, 2025,
from
https://www.lkouniv.ac.in/site/writereaddata/siteContent/20200422161333875 7smmirza
eng_Stuart_Hall.pdf
Moretti, F. (2013). Distant reading. Verso Books.
Mosteller, F., & Wallace, D. L. (1964). Inference and disputed authorship: The Federalist.
Addison-Wesley.
Raynaud, C. (2013). Naming, nonsense and not naming in I Am Not Sidney Poitier.
Transatlantica, 1. https://journals.openedition.org/transatlantica/6305

Online ISSN 3007-3197
Print ISSN 3007-3189

Schmidt, C. (2021). The parody of post-blackness in I Am Not Sidney Poitier and the end(s) of
African American literature. Journal of Contemporary African American Thought, 5(2),
87–105. https://www.academia.edu/82205279
Sinclair, S., & Rockwell, G. (2016). Voyant Tools [Computer software]. https://voyanttools.org/ Young, H. (2010). Rewriting racial authorship: The politics of identity in Everett's Erasure
Callaloo, 33(4), 1254–1272. https://doi.org/10.1353/cal.2010.0176