

# **Stylometric Analysis for Author Attribution**

## **“A Visit of St. Nicholas?”**

**Subject:** Vertiefung Digital Humanities

**Name of the Lecturer:** Prof. Dr. Christof Schöch

**Author:**

Muhammad Owais Aziz

Matriculation Nr: 1662971

Email: [s2muaziz@uni-trier.de](mailto:s2muaziz@uni-trier.de)

**Semester:** Sose 2024

**Submission Date:** October 15, 2024

## Declaration Of Authenticity

I, Muhammad Owais Aziz declare herewith, that this work is my own origin work. This term paper has never been previously submitted in its current or similar form in any other course and/or degree programmed. I have clearly referenced all sources used in the work in the bibliography or references.

Trier, 13-October-2024

Place, Date

A handwritten signature in black ink, appearing to read 'Muhammad Owais Aziz', written over a horizontal line.

Signature of the student

Title of the course

Digital Humanities

Name of the lecturer

Prof. Dr. Christoph Schöch

Family Name, First Name of Student

Aziz, Muhammad Owais

Github Repository Link:

<https://github.com/owaisazizi602/Stylometric-Analysis-for-Author-Attribution---A-Visit-of-St.-Nicholas->

Stylometric Analysis of 'A Visit from St. Nicholas'

This repository contains scripts, datasets, detailed documents, and results for the authorship analysis of 'A Visit from St. Nicholas', focusing on three potential authors: Clement Clarke Moore, Henry Livingston Jr., and Fitz-Greene Halleck.

Contents:

Scripts & Data:

Python and R scripts used for stylometric analysis (Advanced Feature Extraction, PCA, Bootstrap Consensus Trees, Rolling Delta).

Poem datasets from all three poets.

Results:

Visual outputs, including PCA plots and consensus trees, Rolling Delta Algorithm, reflecting authorship patterns.

Document:

A final PDF document will explain the methodology and findings in detail.

## Table of Contents

<b>1. Abstract.....</b>	<b>1</b>
<b>2. Background and Context .....</b>	<b>2</b>
<b>3. Data Selection and Sources .....</b>	<b>3</b>
<b>4. Research Design and Methodology .....</b>	<b>6</b>
<b>5. Results and Findings .....</b>	<b>10</b>
5.1 Rolling Delta Algorithm .....	10
5.3 Bootstrap Consensus Tree .....	31
5.4 Principal Component Analysis .....	33
<b>6. Conclusion .....</b>	<b>35</b>
<b>References.....</b>	<b>36</b>

## 1. Abstract

In the area of Digital Humanities, sophisticated computational tools are being applied to investigate intricate literary issues by examining extensive corpora beyond the capabilities of conventional humanistic approaches. This paper utilizes stylometric analysis alongside advanced feature extraction techniques using the NLTK library in Python—including complete sentence structure analysis, part-of-speech (POS), and punctuation usage—to investigate the long-standing authorship debate surrounding 'A Visit from St. Nicholas'. Traditionally attributed to either Clement Clarke Moore or Henry Livingston Jr., this study broadens the scope by introducing Fitz-Greene Halleck as a third, distractor author for comparative analysis. Employing the "stylo" package in RStudio, techniques such as Principal Component Analysis (PCA), bootstrap consensus trees, and rolling delta stylometry are applied across a carefully selected dataset of texts. The paper overcomes the dataset size limitations imposed by rolling delta analysis by selecting representative samples of each author's work. Based on computational findings and literary interpretation, the study concludes that Henry Livingston Jr. remains the most probable author of the poem, while Fitz-Greene Halleck serves to illustrate key stylistic contrasts. These results add to the expanding depth of information regarding authorship attribution and the usefulness of digital techniques for literary study.

## 2. Background and Context

The appealing illustration of Santa Claus has captivated the imaginations of both kids and adults, and it has come to represent the happiness of the Christmas season. However, this adored character represents the apex of an intricate historical development that spans across cultural divides. St. Nicholas is at the heart of this transformation, a revered figure known for his kindness and generosity, particularly towards the needy and children. One of the most enduring stories associated with him tells of his secret gift of gold, which saved three daughters from a life of deprivation. This act established him as a patron of the vulnerable. (Divineuk, how-st-nicholas-became-santa-claus, 2021)

The poem "A Visit from St. Nicholas," which was initially published anonymously in a New York newspaper on December 23, 1823, is essential to this growth. The modern image of Santa Claus is said to have been influenced by this poem, but it also reignited the long-running dispute over authorship between Henry Livingston Jr. and Clement Clarke Moore. While Livingston's children assert that their father recited the poem to them a year before its publication, Moore, a respected academic, claimed authorship when he included it in his 1844 anthology, *Poet*, citing his children's encouragement.

As an additional viewpoint to this conversation, Fitz-Greene Halleck appears as a potential third author. Being a well-known contemporary of Livingston and Moore, Halleck's poetry combines humorous and serious subjects, which makes him an interesting diversion in this study of authorship. Even if Moore and Livingston have their devoted followers, analyzing Halleck's poetry alongside with theirs will provide readers a deeper grasp of the features and intricacies of the poetry of the time.

This paper will give a complete examination of the works of Moore, Livingston, and Halleck through stylometric approaches, highlighting their unique styles and contributions. Through, the application of advanced computational methods, such as advanced feature extraction sentences, structure analysis, part-of-speech, and punctuation usage analysis, our goal is to provide an understanding of who wrote "A Visit from St. Nicholas." Moreover, a critical evaluation of these digital approaches' limits will be conducted to guarantee a comprehensive comprehension of the authorship debate and its implications in the broader discourse of literature.

### 3. Data Selection and Sources

This part includes the dataset utilized for the stylometric analysis in this study. I've carefully chosen poems by Fitz-Greene Halleck, Clement Clarke Moore, and Henry Livingston Jr. in addition to the debated piece A Visit from St. Nicholas. To ensure the analysis's transparency and reproducibility, a list of the text files poetry is provided below, along with their sources.

#### **Henry Livingston Jr. (22 Poems)**

1. 1819 New Years Carriers Address
2. A Tenant of Mrs Van Kleeck
3. Acknowledgement
4. Acrostic Eliza Hughes
5. An Elegy on the Death of Montgomery Tappen
6. Apostrophe
7. Careless Philosophers Soliloquy
8. Catharine Breese Livingston
9. Dialogue
10. Epithalamium A Marriage Poem
11. Hiding Place
12. Letter Sent to Master Timmy Dwight
13. On My Sister Joanna Entrance Into Her 33rd Year
14. The Crane & The Fox, a Fable
15. The Dance
16. The IX Ode to Horace
17. The Procession
18. The Vine & Oak A Fable
19. To My Little Niece Anne Duyckinck
20. To My Little Niece Sally Livingston
21. To the Memory of Henry Welles Livingston
22. To the Memory of Sarah Livingston

#### **Clement Clarke Moore (31 Poems)**

1. A Song
2. A Trip to Saratoga
3. Apology for not accepting an Invitation to a Ball

4. Cowper the Poet
5. Farewell - In answer to a young lady's invitation to join a party of pleasure on an excursion to the country
6. From a Husband to his Wife
7. Lines accompanying a Bunch of Flowers
8. Lines accompanying some Balls made for a Fragment Fair
9. Lines on the Sisters of Charity
10. Lines sent to a Young Lady, with a Pair of Gloves
11. Lines to a Young Lady for Valentine's Day
12. Lines written after a season of Yellow Fever
13. Old Dobbin
14. Old Santeclaus
15. On receiving from a friend a Caricature cast of Paganini
16. On seeing my Name written in the sand of the sea-shore
17. The Mischievous Muse
18. The Pig and the Rooster
19. The Wine Drinker
20. To a Lady
21. To a Young Lady, on her Birth-Day
22. To my Children, with my Potrait
23. To my Daughter, on her marriage
24. To Petrosa
25. To the Fashionable Part of my Young Countrywomen
26. To the Nymphs of Mount Harmony
27. To Young Ladies who attended Philosophical Lectures
28. Lines Written After a Snow-storm
29. The Organist
30. The Water Drinker
31. To Southey

**Fitz-Greene Halleck (6 Poems)**

1. Burns
2. Alnwick Castle
3. Red Jacket

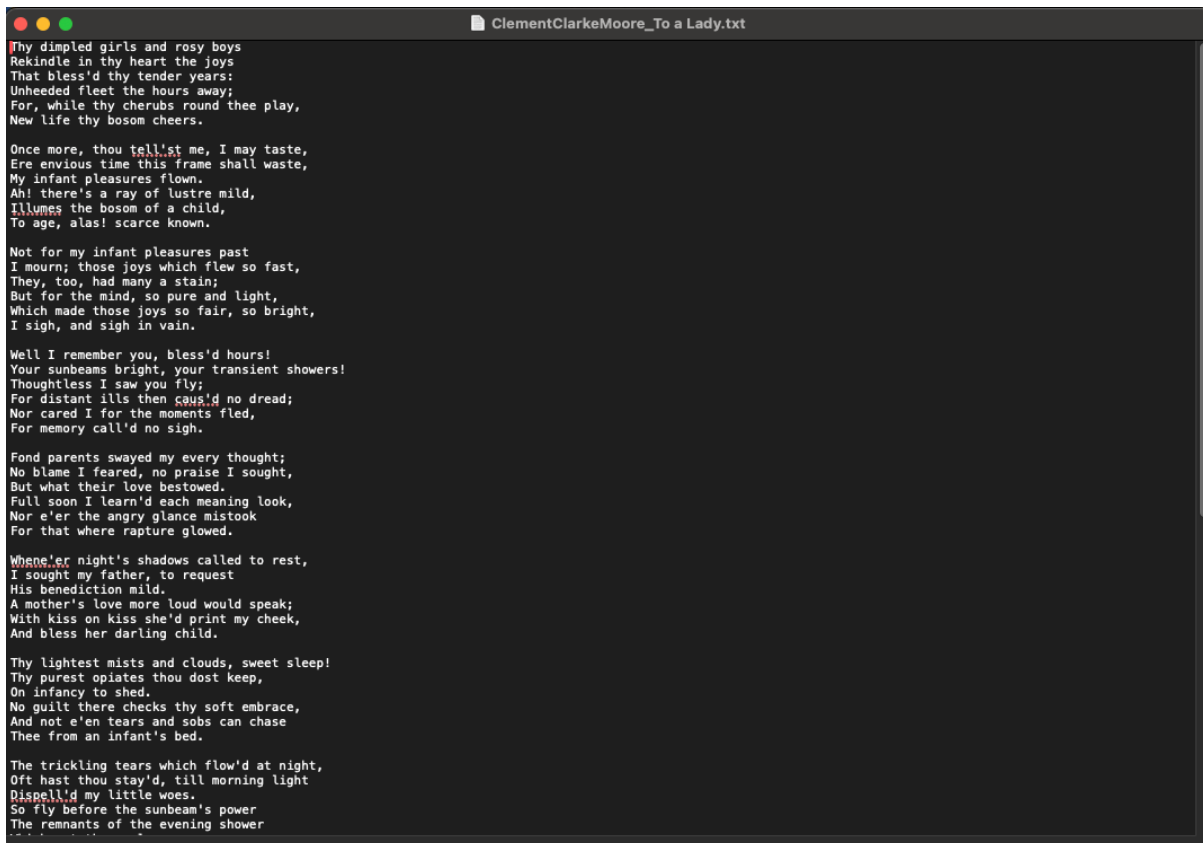


4. Young America
5. Fanny
6. Marco Bozzaris

### A Visit from St. Nicholas (1 Poem)

1. A Visit from St. Nicholas

A snapshot of the dataset, displaying the precise files used for the study, is also attached. There will be detailed references to all of the sources in the Appendix.



## 4. Research Design and Methodology

The main goal of this paper was to gather a corpus of writings from the three possible authors: Fitz-Greene Halleck, Henry Livingston Jr., and Clement Clarke Moore. To establish a balanced sample, we chose poetry from Livingston (22 poems), Moore (31 poems), and Halleck (6 poems). For every poem, a plain text format was created to make sure that stylometric tools would work with it. Since all of the poems were written in conventional English, no additional text cleaning or normalization procedures were needed. The poems were taken from reputable literary anthologies and collections to preserve data integrity.

### Stylometric Tools and Parameters

The "stylo" package in RStudio served as our main tool for doing the stylometric analysis. Stylo is a comprehensive R package that provides several ways for assessing language patterns and authorship through statistical testing, focusing especially on word frequency distributions and other stylistic markers.

The "Rolling Delta" technique was the first test run on the corpus. It allowed us to calculate the differences between "A Visit from St. Nicholas" and a chosen collection of poems by three different authors: Fitz-Greene Halleck, Henry Livingston Jr., and Clement Clarke Moore. The technique iterates through both the poem in question and a primary set of twelve selected reference poems, calculating their similarities and differences.

### Data Segmentation and Selecting:

The corpus for this test was split into two sections: a primary set of four poems by each of the three poets (12 total), and a second set called "A Visit from St. Nicholas." Every reference poem was divided into overlapping samples, each consisting of 10 words, which is about equivalent to a poem's sentence length. Every sample after that shifted by one word since the overlap between samples (step size) was set to one word. By comparing word frequencies, the algorithm was able to calculate "Delta" values, which indicate stylistic variances. To maintain continuity with earlier stylometric testing, we concentrated on the top 100 most commonly used words in this particular case. Based on these frequency profiles, a number of parameters were computed for each reference poem, yielding an overall "Delta Function." This Delta Function was then applied to the secondary text, "A Visit from St. Nicholas," which was split and sampled similarly to the reference poems.

## Frequency and Culling Parameters

The settings for this rolling delta analysis were:

Frequency Rank: Starting at 100, ranging up to 3000 broader analysis.

Increment: 100

Culling: The percentage of words discarded based on their frequency, ranging from 10% (minimum) to 80% (maximum), with an increment of 10%.

List Cutoff: 500 of the most frequently used words were included.

## Findings and Visualization

Plotting the "Delta" values of each reference poem in relation to "A Visit from St. Nicholas," was generated by the algorithm. Greater style differences are indicated by higher delta values, whilst closer stylistic similarities are suggested by lower delta values. We were able to track the differences between the contested poem and the chosen poems by Moore, Livingston, and Halleck at every stage due to this methodology. Importantly, as we go through the text, the evaluation tells us which author's style most closely resembles "A Visit from St. Nicholas."

This rolling delta analysis serves as the foundation for our inquiry into the authorship question, together with the cluster analysis, advanced feature extraction technique, and PCA tests. Using this approach, we can comprehend the stylistic components more deeply and determine whether Moore, Livingstone, or Fitz Halleck is the true author.

The second test I used an advanced feature extraction method with Python's Natural Language Toolkit (NLTK) package to improve the stylometric examination of the authorship question related to A Visit from St. Nicholas. With this method, we are able to extract particular linguistic features from each poem, like sentence length, the frequency of part-of-speech (POS), and the use of punctuation, which are then compared across the three authors in dispute.

### Preparing the Dataset:

I conducted three tests on three different poems from each of the three authors—Henry Livingston Jr., Fitz-Greene Halleck, and Clement Clarke Moore—for this examination. The dataset for comparison against the disputed poem was drawn from each poet's collection. These poems are picked from the available corpus dataset.

Advanced Feature Extraction Process:

Sentence Parsing and Tokenization: NLTK's `sent_tokenize` and `word_tokenize` functions were used to tokenize each poem into sentences and words. The average sentence length for each poem may then be determined by using this procedure.

Part-of-Speech (POS) Tagging: Every poem was tagged for part-of-speech using the `pos_tag` function of NLTK. I then calculated the frequency of each POS tag (adjectives, verbs, and nouns) to look for stylistic and linguistic patterns among the writers.

Punctuation Analysis: In order to gain additional insight into an author's style, the frequency of particular markings such as periods, commas, semicolons, and exclamation points was determined.

### Evaluation and Normalization

By averaging the outcomes over the three poems written by each author, the test results were normalized. This made sure that comparisons didn't rely too much on any one text and were instead based on consistent accurate measurements.

I computed:

Average Sentence Length: The disputed content was compared to each author's average sentence length.

POS Frequency: I determined stylistic patterns, such as the frequent usage of particular parts of speech, by normalizing the POS tag frequencies for each author and comparing them to the disputed text.

Punctuation Usage: To find potential indicators of a writer's unique style, the distribution of punctuation was also normalized across the texts written by each author.

The third test in author attribution for the three dispute poets—Clement Clarke Moore, Henry Livingston Jr., and Fitz-Greene Halleck—I use the bootstrap consensus tree analysis in this study. This approach is significant because it makes textual similarities and contrasts based on word frequencies across several poems visually apparent, so revealing stylistic characteristics unique to each author.

By executing several clustering rounds and integrating them into one, the consensus tree offers a more comprehensive analysis and a clearer representation of the most consistent groups. This is essential for author attribution since it reveals trends that show up in multiple tests, providing

information on whether a given text, such as "A Visit from St. Nicholas," is more associated with one poet than the others.

By running the analysis across these poets' texts, the consensus tree visually clusters their works, making it easier to observe how "A Visit from St. Nicholas" compares to their other poems, helping to narrow down its likely author based on linguistic style.

The fourth test, Principal Component Analysis (PCA) on the three contested texts, was to find and compare stylistic patterns in the writings of Fitz-Greene Halleck, Clement Clarke Moore, and Henry Livingston Jr. PCA improves in demonstrating of the connections between texts based on their linguistic characteristics by lowering the dimensionality of word frequency data. This gives me the ability to determine whether the content in question is more closely aligned with one author's body of work than another, providing a data-driven method for settling authorship disputes. This test is significant because it goes beyond subjective interpretation to objectively assess stylistic similarities and differences. When there is doubt over who wrote a work, like "A Visit from St. Nicholas," PCA offers statistical and visual proof to help establish which author's style most closely matches the disputed work. Establishing a more robust foundation for authorship claims is crucial as it enhances our comprehension of these writings from a literary and historical standpoint.

## 5. Results and Findings

### 5.1 Rolling Delta Algorithm

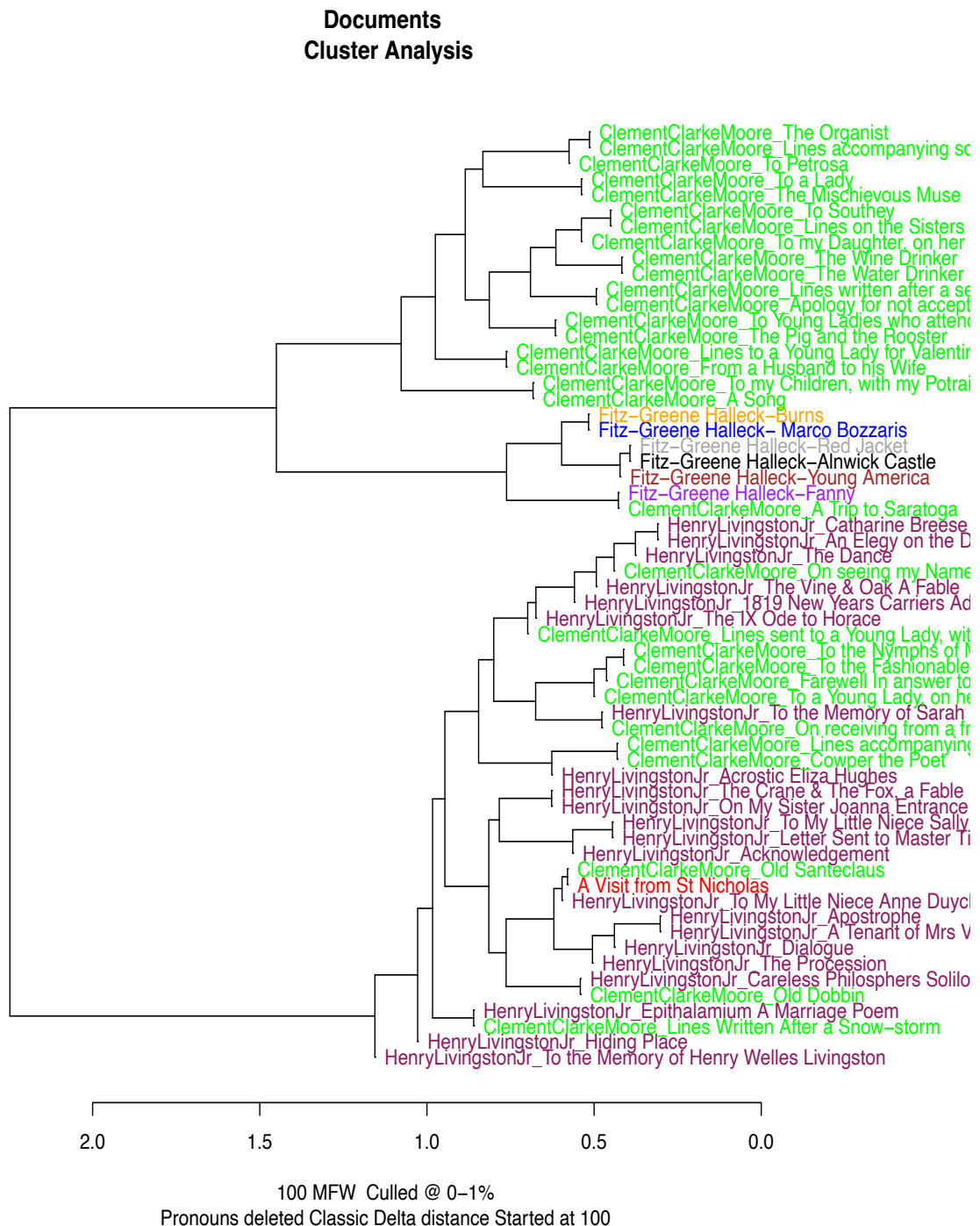


Fig 5.1

The analysis uses 100 Most Frequent Words (MFW) with pronouns deleted, applying the Classic Delta method. This method focuses on word usage patterns across the texts to determine stylistic similarities. The closer two texts are in the cluster, the more alike their writing styles are. The

analysis groups texts by Clement Clarke Moore, Henry Livingston Jr., and Fitz-Greene Halleck. As, the package only accept twelve poems against the poem 'A Visit to St. Nicholas' for the rolling delta algorithm. Therefore, a pre cluster analysis I conducted and pick four poems from each poet thus a total of twelve poems against the actual one.

Findings for Every Author:

Henry Livingston Jr. :

"A Visit from St. Nicholas" is strongly related to a number of Henry Livingston Jr.'s works.

"To My Little Niece Anne Duyckinck" and

"Acknowledgment" are two of these pieces.

"Letter Sent to Master Timmy Dwight" "To My Little Niece Sally Livingston"

The close clustering points to a considerable stylistic similarity between "A Visit from St. Nicholas" and Livingston's works. His compositions generally contain comparable word usage patterns, which means that Livingston's style might closely resemble that of the poetry. This connection is further strengthened by the fact that Livingston's other pieces, such "The Crane & The Fox" and "Epithalamium A Marriage Poem," also belong to a group that is closer to "A Visit from St. Nicholas."

Clement Clarke Moore:

Clement Clarke Moore's works, such as:

"Old Santeclaus"

"The Water Drinker"

"The Wine Drinker"

"To Young Ladies Who Attended Philosophical Lectures"

There is less stylistic similarity between these pieces with "A Visit from St. Nicholas" because they are placed together but not together. This indicates that Moore's writing style deviates more from "A Visit from St. Nicholas" as evidenced by these texts.

Although the poem is typically attributed to Moore, the cluster analysis based on the 100 MFW indicates that, in comparison to Livingston's style, Moore's style does not closely reflect the poem.

Fitz-Greene Halleck:

Halleck's works, such as:

"Fanny"

"Young America"

"Alnwick Castle"

"Red Jacket"

"Marco Bozzaris"

These pieces are set apart from Henry Livingston Jr.'s and "A Visit from St. Nicholas" in a distinct grouping.

Given his unique placement inside the cluster analysis, it is apparent that Halleck's writing style is very different from that of "A Visit from St. Nicholas" and the other authors' writings. There is minimal evidence from this research that would confirm Fitz-Greene Halleck as the author of the poem. The Classic Delta approach uses word frequency patterns to determine how much writing styles vary among texts. Greater similarity is indicated by lower delta scores. The texts are stylistically more similar to one another the closer they are in the clustering plot. The dendrogram result indicates that Henry Livingston Jr.'s writing style is a better fit for the poem "A Visit from St. Nicholas" than it is for Clement Clarke Moore's or Fitz-Greene Halleck's.



Fig 5.2

For the rolling delta analysis in stylometry, the tool has a limitation on the number of texts or files it can process simultaneously. In this case, it only accepts 12 files. To accommodate this, I strategically selected 4 texts files from each author (Clement Clarke Moore, Henry Livingston Jr., and Fitz-Greene Halleck), ensuring a balanced representation of the three authors in my analysis. Choosing these setting for my de

Start at Frequency Rank: 100



Maximum Frequency Rank: 2000

Culling Rate: 0.1

List Cutoff: 500

Slice Length: 10

Step Size: 1 word

These parameters should guarantee that the Rolling Delta algorithm can detect small discrepancies while offering an in-depth examination of the texts by the three disputed poets.

In most instances, Henry Livingston Jr. has the lowest delta values according to the Rolling Delta analysis, suggesting the greatest stylistic resemblance to "A Visit from St. Nicholas." The somewhat larger delta values for Clement Clarke Moore indicate a weaker stylistic link. Fitz-Greene Halleck's delta values are consistently the highest, suggesting that his style is the most dissimilar from the contested poetry. These results show that Livingston is the most likely author, and Halleck is most likely not the author.

## 5.2 Advanced Feature Extraction

The following Python code outlines the steps performed for feature extraction:

I'm just shring the test 1 code that I performed rest for the other tests the code structure is same.

```
import nltk
from nltk import pos_tag, word_tokenize, sent_tokenize
from collections import Counter
import statistics

# Ensure that NLTK resources are downloaded (if not already)
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')

actual_author_texts = [
'''
'Twas the night before Christmas, when all thro' the house,
Not a creature was stirring, not even a mouse;
The stockings were hung by the chimney with care,
In hopes that St. Nicholas soon would be there;
The children were nestled all snug in their beds,
While visions of sugar plums danc'd in their heads,
And Mama in her 'kerchief, and I in my cap,
Had just settled our brains for a long winter's nap -
When out on the lawn there arose such a clatter,
I sprung from the bed to see what was the matter.
Away to the window I flew like a flash,
Tore open the shutters, and threw up the sash.
The moon on the breast of the new fallen snow,
Gave the lustre of mid-day to objects below;
When, what to my wondering eyes should appear,
But a minature sleigh, and eight tiny rein-deer,
With a little old driver, so lively and quick,
I knew in a moment it must be St. Nick.
More rapid than eagles his coursers they came,
And he whistled, and shouted, and call'd them by name:
"Now! Dasher, now! Dancer, now! Prancer, and Vixen,
"On! Comet, on! Cupid, on! Dunder and Blixem;
"To the top of the porch! to the top of the wall!
"Now dash away! dash away! dash away all!"
As dry leaves before the wild hurricane fly,
When they meet with an obstacle, mount to the sky;
So up to the house-top the coursers they flew,
With the sleigh full of Toys - and St. Nicholas too:
```

And then in a twinkling, I heard on the roof  
 The prancing and pawing of each little hoof.  
 As I drew in my head, and was turning around,  
 Down the chimney St. Nicholas came with a bound:  
 He was dress'd all in fur, from his head to his foot,  
 And his clothes were all tarnish'd with ashes and soot;  
 A bundle of toys was flung on his back,  
 And he look'd like a peddler just opening his pack:  
 His eyes - how they twinkled! his dimples how merry,  
 His cheeks were like roses, his nose like a cherry;  
 His droll little mouth was drawn up like a bow,  
 And the beard of his chin was as white as the snow;  
 The stump of a pipe he held tight in his teeth,  
 And the smoke it encircled his head like a wreath.  
 He had a broad face, and a little round belly  
 That shook when he laugh'd, like a bowl full of jelly:  
 He was chubby and plump, a right jolly old elf,  
 And I laugh'd when I saw him in spite of myself;  
 A wink of his eye and a twist of his head  
 Soon gave me to know I had nothing to dread.  
 He spoke not a word, but went straight to his work,  
 And fill'd all the stockings; then turn'd with a jirk,  
 And laying his finger aside of his nose  
 And giving a nod, up the chimney he rose.  
 He sprung to his sleigh, to his team gave a whistle,  
 And away they all flew, like the down of a thistle:  
 But I heard him exclaim, ere he drove out of sight -  
 Happy Christmas to all, and to all a good night.

'''

]

# Disputed texts

disputed\_text\_1 = '''

Sweet Maid, could wealth or power  
 Thy heart to love incline,  
 I would not bless the hour,  
 The hour that calls thee mine.  
 Ah! no, beneath the Heaven  
 Blooms not so fair a flower  
 As love that's freely given.  
 Dear youth, have not these eyes,  
 To thine so oft returning,  
 Ah! say, have not these tell-tale sighs,  
 These cheeks with blushes burning,  
 My every thought bespoken?  
 Do these denote disguise?  
 Do these false love betoken?

Oh! bliss, all bliss transcending,  
When souls congenial blending,  
The sacred flame inspire  
Of love's etherial fire.  
Such love, from change secure,  
For ever shall endure.  
True love like this, of heavenly birth,  
Not here confin'd to mortal earth,  
Shall to immortal Heaven aspire.  
'''

disputed\_text\_2 = '''  
A vine from noblest lineage sprung  
And with the choicest clusters hung,  
In purple rob'd, reclining lay,  
And catch'd the noontide's fervid ray;  
The num'rous plants that deck the field  
Did all the palm of beauty yield;  
Pronounc'd her fairest of their train  
And hail'd her empress of the plain.  
A neighb'ring oak whose spiry height  
In low-hung clouds was hid from sight,  
Who dar'd a thousand howling storms;  
Conscious of worth, sublimely stood,  
The pride and glory of the wood.

He saw her all defenseless lay  
To each invading beast a prey,  
And wish'd to clasp her in his arms  
And bear her far away from harms.  
'Twas love -- 'twas tenderness -- 'twas all  
That men the tender passion call.

He urg'd his suit but urg'd in vain,  
The vine regardless of his pain  
Still flirted with each flippant green  
With seeing pleas'd, & being seen;  
And as the syren Flattery sang  
Would o'er the strains ecstatic hang;  
Enjoy'd the minutes as they rose  
Nor fears her bosom discompose.

But now the boding clouds arise  
And scowling darkness veils the skies;  
Harsh thunders roar -- red lightnings gleam,  
And rushing torrents close the scene.

The fawning, adulating crowd

Who late in thronged xx bow'd  
Now left their goddess of a day  
To the O'erwhelming flood a prey,  
which swell'd a deluge poured around  
& tore her helpless from the ground;  
Her rifled foliage floated wide  
And ruby nectar ting'd the tide.

With eager eyes and heart dismayed  
She look'd but look'd in vain for aid.  
"And are my lovers fled," she cry'd,  
"Who at my feet this morning sigh'd,  
"And swore my reign would never end  
"While youth and beauty had a friend?  
"I am unhappy who believ'd!  
"And they detested who deceived!  
"Curse on that whim call'd maiden pride  
"Which made me shun the name of bride  
"When yonder oak confessed his flame  
"And woo'd me in fair honor's name.  
"But now repentance comes too late  
"And all forlorn, I meet my fate."

The oak who safely wav'd above  
Look'd down once more with eyes of love  
(Love higher wrought with pity join'd  
True mark of an exalted mind,)  
Declared her coldness could suspend  
But not his gen'rous passion end.  
Beg'd to renew his am'rous plea,  
As warm for union now as he,  
To his embraces quick she flew  
And felt & gave sensations new.

Enrich'd & graced by the sweet prise  
He lifts her tendrils to the skies;  
Whilst she, protected and carest,  
Sinks in his arms completely blest.

'''

disputed\_text\_3 = '''

Home of the Percy's high-born race,  
Home of their beautiful and brave,  
Alike their birth and burial-place,  
Their cradle and their grave!  
Still sternly o'er the castle gate

Their house's Lion stands in state,  
As in his proud departed hours;  
And warriors frown in stone on high,  
And feudal banners "flout the sky"  
Above his princely towers.

A gentle hill its side inclines,  
Lovely in England's fadeless green,  
To meet the quiet stream which winds  
Through this romantic scene  
As silently and sweetly still,  
As when, at evening, on that hill,  
While summer's wind blew soft and low,  
Seated by gallant Hotspur's side,  
His Katherine was a happy bride,  
A thousand years ago.

Gaze on the Abbey's ruined pile:  
Does not the succoring ivy, keeping  
Her watch around it, seem to smile,  
As o'er a loved one sleeping?  
One solitary turret gray  
Still tells, in melancholy glory,  
The legend of the Cheviot day,  
The Percy's proudest border story.  
That day its roof was triumph's arch;  
Then rang, from aisle to pictured dome,  
The light step of the soldier's march,  
The music of the trump and drum;  
And babe, and sire, the old, the young,  
And the monk's hymn, and minstrel's song,  
And woman's pure kiss, sweet and long,  
Welcomed her warrior home.

Wild roses by the Abbey towers  
Are gay in their young bud and bloom:  
They were born of a race of funeral-flowers  
That garlanded, in long-gone hours,  
A templar's knightly tomb.  
He died, the sword in his mailed hand,  
On the holiest spot of the Blessed land,  
Where the Cross was damped with his dying breath,  
When blood ran free as festal wine,  
And the sainted air of Palestine  
Was thick with the darts of death.

Wise with the lore of centuries,

What tales, if there be "tongues in trees,"  
Those giant oaks could tell,  
Of beings born and buried here;  
Tales of the peasant and the peer,  
Tales of the bridal and the bier,  
The welcome and farewell,  
Since on their boughs the startled bird  
First, in her twilight slumbers, heard  
The Norman's curfew-bell!

I wandered through the lofty halls  
Trod by the Percys of old fame,  
And traced upon the chapel walls  
Each high, heroic name,  
From him<sup>3</sup> who once his standard set  
Where now, o'er mosque and minaret,  
Glitter the Sultan's crescent moons;  
To him who, when a younger son,  
Fought for King George at Lexington,<sup>4</sup>  
A major of dragoons.

That last half stanza—it has dashed  
From my warm lip the sparkling cup;  
The light that o'er my eyebeam flashed,  
The power that bore my spirit up  
Above this bank-note world—is gone;  
And Alnwick's but a market town,  
And this, alas! its market day,

And beasts and borderers throng the way;  
Oxen and bleating lambs in lots,  
Northumbrian boors and plaided Scots,  
Men in the coal and cattle line;  
From Teviot's bard and hero land,  
From royal Berwick's<sup>5</sup> beach of sand,  
From Wooller, Morpeth, Hexham, and  
Newcastle-upon-Tyne.

These are not the romantic times  
So beautiful in Spenser's rhymes,  
So dazzling to the dreaming boy:  
Ours are the days of fact, not fable,  
Of knights, but not of the round table,  
Of Bailie Jarvie, not Rob Roy:  
'Tis what "our President," Monroe,  
Has called "the era of good feeling:"  
The Highlander, the bitterest foe

To modern laws, has felt their blow,  
 Consented to be taxed, and vote,  
 And put on pantaloons and coat,  
 And leave off cattle-stealing:  
 Lord Stafford mines for coal and salt,  
 The Duke of Norfolk deals in malt,  
 The Douglass in red herrings;  
 And noble name and cultured land,  
 Palace, and park, and vassal-band,  
 Are powerless to the notes of hand  
 Of Rothschild or the Barings.

The age of bargaining, said Burke,  
 Has come: to-day the turbaned Turk  
 (Sleep, Richard of the lion heart!  
 Sleep on, nor from your cerements start)  
 Is England's friend and fast ally;  
 The Moslem tramples on the Greek,  
 And on the Cross and altar-stone,  
 And Christendom looks tamely on,  
 And hears the Christian maiden shriek,  
 And sees the Christian father die;  
 And not a sabre-blow is given  
 For Greece and fame, for faith and heaven,  
 By Europe's craven chivalry.

You'll ask if yet the Percy lives  
 In the armed pomp of feudal state?  
 The present representatives  
 Of Hotspur and his "gentle Kate,"  
 Are some half-dozen serving-men  
 In the drab coat of William Penn;  
 A chambermaid, whose lip and eye,  
 And cheek, and brown hair, bright and curling,  
 Spoke Nature's aristocracy;  
 And one, half groom, half seneschal,  
 Who bowed me through court, bower, and hall,  
 From donjon-keep to turret wall,  
 For ten-and-sixpence sterling.  
 '''

```
# Function to analyze text features
def analyze_text(text):
    sentences = sent_tokenize(text)
    avg_len = sum(len(word_tokenize(sent)) for sent in sentences) / len(sentences) if
sentences else 0
```



```

pos_tags = pos_tag(word_tokenize(text))
pos_freq = Counter(tag for word, tag in pos_tags)

punctuation = Counter(char for char in text if char in '.,;!?')

return avg_len, pos_freq, punctuation

# Analyze actual author's corpus
author_avg_lengths = []
author_pos_freqs = Counter()
author_punctuations = Counter()

for text in actual_author_texts:
    avg_len, pos_freq, punctuation = analyze_text(text)
    author_avg_lengths.append(avg_len)
    author_pos_freqs.update(pos_freq)
    author_punctuations.update(punctuation)

# Average statistics for the actual author
author_avg_length = statistics.mean(author_avg_lengths) if author_avg_lengths else 0
author_pos_freqs_normalized = {k: v / len(actual_author_texts) for k, v in
author_pos_freqs.items()}
author_punctuations_normalized = {k: v / len(actual_author_texts) for k, v in
author_punctuations.items()}

# Analyze the first disputed text
disputed_avg_length_1, disputed_pos_freqs_1, disputed_punctuations_1 =
analyze_text(disputed_text_1)

# Analyze the second disputed text
disputed_avg_length_2, disputed_pos_freqs_2, disputed_punctuations_2 =
analyze_text(disputed_text_2)

# Analyze the third disputed text
disputed_avg_length_3, disputed_pos_freqs_3, disputed_punctuations_3 =
analyze_text(disputed_text_3)

# Compare sentence lengths
print(f"Average Sentence Length - Actual Author: {author_avg_length}")
print(f"Disputed Text 1: {disputed_avg_length_1}, Disputed Text 2:
{disputed_avg_length_2}, Disputed Text 3: {disputed_avg_length_3}")

# Compare POS frequencies
print(f"POS Frequency - Actual Author: {author_pos_freqs_normalized}")
print(f"Disputed Text 1: {disputed_pos_freqs_1}")
print(f"Disputed Text 2: {disputed_pos_freqs_2}")
print(f"Disputed Text 3: {disputed_pos_freqs_3}")

```

```
# Compare punctuation usage
print(f"Punctuation - Actual Author: {author_punctuations_normalized}")
print(f"Punctuation - Disputed Text 1: {disputed_punctuations_1}")
print(f"Punctuation - Disputed Text 2: {disputed_punctuations_2}")
print(f"Punctuation - Disputed Text 3: {disputed_punctuations_3}")
```

Using this feature extraction method, I was able to evaluate the authors main stylistic elements completely. It became possible to find patterns that contradict or back up claims of authorship by comparing sentence length, POS tagging, and punctuation usage between the works of the real author and the opposed ones. This method offers insightful information that enhances the more comprehensive stylistometric and literary analyses in this work, even though it cannot conclusively identify the real author of "A Visit from St. Nicholas" on its own.

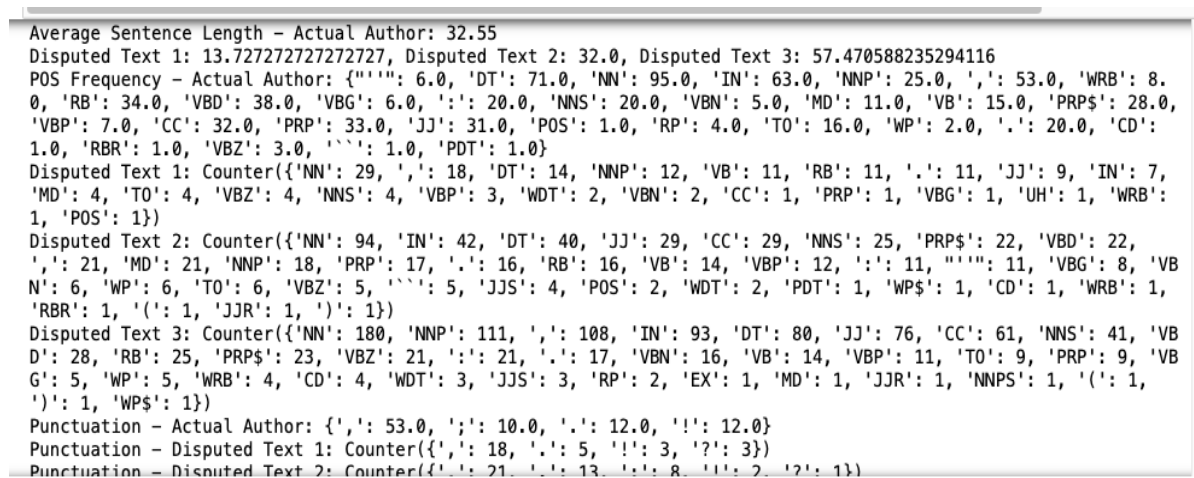
While performing the first test on the following poems I got the following result.

Actual author: A visit to St. Nicholas

Disputed text1: Clark Moore A song

Disputed Text 2: Henry Livingstone The vine & Oak a fable

Disputed Text 3: Fitz Alnmick Castle



```
Average Sentence Length - Actual Author: 32.55
Disputed Text 1: 13.7272727272727, Disputed Text 2: 32.0, Disputed Text 3: 57.470588235294116
POS Frequency - Actual Author: {'''': 6.0, 'DT': 71.0, 'NN': 95.0, 'IN': 63.0, 'NNP': 25.0, ',': 53.0, 'WRB': 8.0, 'RB': 34.0, 'VBD': 38.0, 'VBG': 6.0, ':': 20.0, 'NNS': 20.0, 'VBN': 5.0, 'MD': 11.0, 'VB': 15.0, 'PRP$': 28.0, 'VBP': 7.0, 'CC': 32.0, 'PRP': 33.0, 'JJ': 31.0, 'POS': 1.0, 'RP': 4.0, 'TO': 16.0, 'WP': 2.0, '.': 20.0, 'CD': 1.0, 'RBR': 1.0, 'VBZ': 3.0, '``': 1.0, 'PDT': 1.0}
Disputed Text 1: Counter({'NN': 29, ',': 18, 'DT': 14, 'NNP': 12, 'VB': 11, 'RB': 11, '.': 11, 'JJ': 9, 'IN': 7, 'MD': 4, 'TO': 4, 'VBZ': 4, 'NNS': 4, 'VBP': 3, 'WDT': 2, 'VBN': 2, 'CC': 1, 'PRP': 1, 'VBG': 1, 'UH': 1, 'WRB': 1, 'POS': 1})
Disputed Text 2: Counter({'NN': 94, 'IN': 42, 'DT': 40, 'JJ': 29, 'CC': 29, 'NNS': 25, 'PRP$': 22, 'VBD': 22, ',': 21, 'MD': 21, 'NNP': 18, 'PRP': 17, '.': 16, 'RB': 16, 'VB': 14, 'VBP': 12, ':': 11, '``': 11, 'VBG': 8, 'VBN': 6, 'WP': 6, 'TO': 6, 'VBZ': 5, '``': 5, 'JJS': 4, 'POS': 2, 'WDT': 2, 'PDT': 1, 'WP$': 1, 'CD': 1, 'WRB': 1, 'RBR': 1, '(': 1, 'JJR': 1, ')': 1})
Disputed Text 3: Counter({'NN': 180, 'NNP': 111, ',': 108, 'IN': 93, 'DT': 80, 'JJ': 76, 'CC': 61, 'NNS': 41, 'VBD': 28, 'RB': 25, 'PRP$': 23, 'VBZ': 21, '.': 21, ':': 17, 'VBN': 16, 'VB': 14, 'VBP': 11, 'TO': 9, 'PRP': 9, 'VBG': 5, 'WP': 5, 'WRB': 4, 'CD': 4, 'WDT': 3, 'JJS': 3, 'RP': 2, 'EX': 1, 'MD': 1, 'JJR': 1, 'NNPS': 1, '(': 1, ')': 1, 'WP$': 1})
Punctuation - Actual Author: {'': 53.0, ',': 10.0, '.': 12.0, '': 12.0}
Punctuation - Disputed Text 1: Counter({'': 18, '.': 5, '': 3, '?'': 3})
Punctuation - Disputed Text 2: Counter({'': 21, '.': 13, '': 8, '': 2, '?'': 1})
```

Fig 5.4

## 1. Average Sentence Length:

Actual Author: 32.55 words per sentence

The actual author, with a sentence length of around 32.55 words, shows a narrative with long, complex sentence structures, which is typical for descriptive or storytelling poetry.

Disputed Text 1: 13.73 words per sentence

Disputed Text 1 uses much shorter sentences than the actual author, which indicates a simpler, more direct writing style. This difference makes it unlikely that this poem matches the actual author's style, which tends toward longer sentences.

Disputed Text 2: 32.0 words per sentence

Disputed Text 2 is nearly identical to the actual author's average sentence length, suggesting a similar level of complexity. This similarity makes this text a closer match to the actual author's style in terms of sentence construction.

Disputed Text 3: 57.47 words per sentence

Disputed Text 3 has very long sentences, far exceeding the actual author's sentence length. This suggests a highly complex, possibly overly ornate style, making it distinct from the actual author's more moderately complex sentences.

## 2. Parts of Speech (POS) Frequency:

Actual Author:

Nouns (NN): 95 instances—demonstrates a focus on descriptive elements, typical of a narrative poem with strong visual imagery.

Pronouns (PRP): 33 instances—frequent use of pronouns indicates a more personal narrative involving multiple characters.

Verbs (VBD): 38 instances—many past-tense verbs, indicating a storytelling or reflective style.

Adjectives (JJ): 31 instances—showing a moderate level of description through adjectives.

Disputed Text 1:

Nouns (NN): 29 instances—substantially lower noun usage than the actual author, indicating a less descriptive style.

Pronouns (PRP): 1 instance—this text uses very few personal pronouns, which implies a less personal narrative.

Adjectives (JJ): 9 instances—much lower adjective use, suggesting a less detailed and descriptive style compared to the actual author.

Disputed Text 2:

Nouns (NN): 94 instances—almost identical noun usage to the actual author, indicating a similar descriptive focus.

Pronouns (PRP): 17 instances—similar use of pronouns, showing a somewhat personal narrative.

Adjectives (JJ): 29 instances—close to the actual author's use of adjectives, suggesting a similarly descriptive style.

Verbs (VBD): 22 instances—somewhat fewer past-tense verbs, but still comparable, indicating a narrative structure.

Disputed Text 3:

Nouns (NN): 180 instances—this text is extremely noun-heavy, suggesting a very descriptive and detailed style, possibly more so than the actual author.

Pronouns (PRP): 9 instances—fewer pronouns compared to the actual author, indicating less emphasis on personal elements.

Adjectives (JJ): 76 instances—extremely descriptive, perhaps too ornate compared to the actual author's moderate adjective use.

Verbs (VBD): 28 instances—this text has fewer past-tense verbs, which may indicate a more static, less dynamic narrative.

3. Punctuation Usage:

Actual Author:

Commas (','): 53 instances—indicates the frequent use of clauses, contributing to the complexity of sentences.

Exclamation Marks ('!'): 12 instances—suggests a lively, dramatic tone, which fits with the excitement of "Twas the Night Before Christmas."

Disputed Text 1:

Commas (','): 18 instances—fewer commas, reflecting the simpler, shorter sentence structure.

Exclamation Marks ('!'): 3 instances—less emphasis or excitement compared to the actual author.

Disputed Text 2:

Commas (','): 21 instances—more frequent commas compared to Disputed Text 1 but still fewer than the actual author, indicating moderately complex sentence structure.

Exclamation Marks ('!'): 2 instances—much less emphasis compared to the actual author's lively tone.

Disputed Text 3:

Commas (','): 109 instances—far more commas than the actual author, suggesting overly complex or fragmented sentence structures.

Exclamation Marks ('!'): 4 instances—less dramatic emphasis compared to the actual author.

Disputed Text 1:

There are significantly fewer individual elements, descriptive language, and longer sentences in this writing. It differs greatly from the real author's style due to its more restrained punctuation and simpler syntax. It is the the most distant from the actual author's work.

Disputed Text 2:

In terms of noun frequency, descriptive language, and sentence length (32.0 words), this text is the most similar to the real author. Overall, the style is more in line with the author's than the other texts; nevertheless, it uses less punctuation and fewer exclamation points than the author's, suggesting a more restricted approach.

Disputed Text 3:

The author's style is considerably more intricate and descriptive in this book than it is in the original. Despite the heavy reliance on nouns and adjectives, the sentence's length (57.47 words) and frequent comma usage give the impression that it is unduly elaborate. The more complex and sophisticated style of this text makes it improbable that it is by the same author as the poem itself.

Most Similar Text: Disputed Text 2 is the closest match to the actual author's style based on sentence length, descriptive language, and structure. It has a similar complexity to the actual author's poem, but with slightly less emphasis and punctuation. Disputed Text 3 is overly complex, and Disputed Text 1 is much simpler, making them less likely candidates.

The second test conducted on the following poems:

Actual author: A visit to St. Nicholas

Disputed text1: Clark Moore cowper the poet

Disputed Text 2: Henry Livingstone the crane & the fox

Disputed Text 3: Fitz Burns

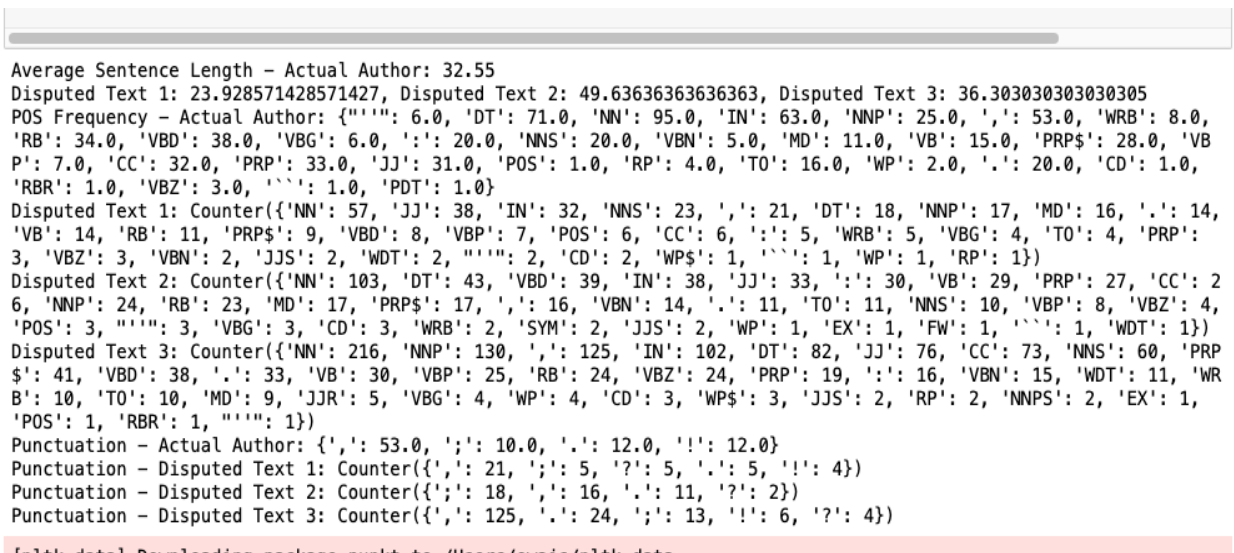


Fig 5.5

## 1. Average Sentence Length:

Actual Author: 32.55 words per sentence

The actual author's sentence length reflects moderately long and complex sentences, typical for narrative poetry with detailed descriptions.

Disputed Text 1: 23.93 words per sentence

The sentences in Disputed Text 1 are shorter than the actual author's, but still fairly complex. This indicates a simpler, less elaborate structure compared to the actual author, suggesting that it may not match as closely.

Disputed Text 2: 49.64 words per sentence

This text has significantly longer sentences than the actual author. The extended sentence length may indicate a more intricate, verbose style that differs from the more balanced complexity of the actual author's sentences.

Disputed Text 3: 36.30 words per sentence

Disputed Text 3 is the closest in sentence length to the actual author. Its sentences are slightly longer, but overall, it reflects a similar level of complexity, suggesting a more comparable writing style.

## 2. Parts of Speech (POS) Frequency:

Actual Author:

Nouns (NN): 95 instances—high frequency, reflecting a descriptive style centered around objects, people, and places.

Adjectives (JJ): 31 instances—moderate use of adjectives, indicating a descriptive but not overly ornate style.

Verbs (VBD): 38 instances—indicating a strong narrative focus, with many actions taking place.

Conjunctions (CC): 32 instances—frequent use of conjunctions to build complex sentences.

Punctuation (',' and '.'): Frequent use of punctuation suggests moderately complex sentence structures.

Disputed Text 1:

Nouns (NN): 57 instances—lower noun frequency compared to the actual author, which suggests less descriptive detail.

Adjectives (JJ): 38 instances—comparable to the actual author, suggesting a descriptive style.

Verbs (VBD): 8 instances—far fewer past-tense verbs, indicating a less action-oriented or narrative-driven style.

Conjunctions (CC): 6 instances—fewer conjunctions, indicating simpler sentence structures.

Disputed Text 2:

Nouns (NN): 103 instances—comparable to the actual author, indicating a similarly descriptive style.

Adjectives (JJ): 33 instances—similar to the actual author, suggesting comparable levels of description.

Verbs (VBD): 39 instances—almost identical to the actual author, indicating a similar narrative focus.

Conjunctions (CC): 26 instances—frequent use of conjunctions, contributing to complex sentence structures.

This POS distribution makes Disputed Text 2 very similar to the actual author.

Disputed Text 3:

Nouns (NN): 216 instances—extremely noun-heavy, indicating a very descriptive and detailed writing style, perhaps even more so than the actual author.

Adjectives (JJ): 76 instances—far more adjectives, indicating a highly ornate, possibly verbose style.

Verbs (VBD): 38 instances—similar to the actual author, which indicates a similar narrative focus.  
Conjunctions (CC): 73 instances—extremely high, suggesting highly complex, perhaps convoluted sentences.

This text has a far more elaborate style than the actual author, with excessive descriptions and longer sentence structures.

### 3. Punctuation Usage:

Actual Author:

Commas (','): 53 instances—frequent use of commas to break up clauses, contributing to moderately complex sentences.

Semicolons (;): 10 instances—used for more intricate sentence structures.

Periods ('.'): 12 instances each, indicating a lively, dynamic tone typical of the poem's narrative.

Disputed Text 1:

Commas (','): 21 instances—fewer commas, reflecting shorter, simpler sentences.

Exclamation Marks ('!'): 4 instances—less emphasis or excitement compared to the actual author.

Semicolons (;): 5 instances—somewhat similar in usage, but overall simpler punctuation patterns.

Disputed Text 2:

Commas (','): 16 instances—still fewer commas compared to the actual author, indicating slightly less complex sentence structures.

Semicolons (;): 18 instances—much higher use of semicolons, indicating more intricate sentences, possibly overusing them compared to the actual author.

Periods ('.'): 11 instances—comparable to the actual author.

Exclamation Marks ('!'): Only 2 instances—much less excitement or emphasis compared to the actual author's lively tone.

Disputed Text 3:

Commas (','): 125 instances—much more frequent use of commas, suggesting highly complex, perhaps overly fragmented sentences.

Periods ('.'): 24 instances—double the actual author's usage, indicating more frequent sentence termination.



Semicolons (;): 13 instances—slightly higher usage than the actual author.

Exclamation Marks (!): 6 instances—less emphasis compared to the actual author but more than the other disputed texts.

Disputed text 1:

less descriptive language and a simpler sentence structure than the real author. A more uncomplicated writing style is seen in the punctuation usage and POS frequency. Because of these variations, the likelihood that it is by the same author is reduced.

Disputed Text 2:

What most resembles the original author's writing is the length of sentences, the distribution of POS, and the use of punctuation. It is similar to the real author's writing style in many ways, such as sentence complexity, verb usage frequency, and noun frequency. The increased use of semicolons, however, points to a marginally more complex style.

Disputed Text 3: incredibly complex and elegant writing style, using many more conjunctions, adjectives, and nouns than the original author. Longer sentences and copious amounts of punctuation, particularly commas, imply that this work is too complex and verbose to be written by the real author. Even though it uses comparable verbs and has a similar narrative purpose, its more intricate and thorough language suggests that it is written by another author.

Most Similar Text: In terms of sentence length, descriptive language, and general organization, Disputed Text 2 most closely resembles the author's style. Disputed Text 1 is too simple to correspond with the real author, and Disputed Text 3 is unduly complicated.

Coming to the last test which is test third, got the following results on these poems.

Actual author: A visit to St. Nicholas

Disputed text 1: Clark Moore On seeing my Name written in the sand of the sea-shore

Disputed Text 2: Henry Livingstone n Elegy on the Death of Montgomery Tappen

Disputed Text 3: Fitz Red Jacket

```

Average Sentence Length - Actual Author: 32.55
Disputed Text 1: 18.3333333333332, Disputed Text 2: 18.5454545454547, Disputed Text 3: 46.0
POS Frequency - Actual Author: {'''': 6.0, 'DT': 71.0, 'NN': 95.0, 'IN': 63.0, 'NNP': 25.0, ',': 53.0, 'WRB': 8.0, 'RB': 34.0, 'VBD': 38.0, 'VBG': 6.0, '': 20.0, 'NNS': 20.0, 'VBN': 5.0, 'MD': 11.0, 'VB': 15.0, 'PRP$': 28.0, 'VBP': 7.0, 'CC': 32.0, 'PRP': 33.0, 'JJ': 31.0, 'POS': 1.0, 'RP': 4.0, 'TO': 16.0, 'WP': 2.0, '': 20.0, 'CD': 1.0, 'RBR': 1.0, 'VBZ': 3.0, '``': 1.0, 'PDT': 1.0}
Disputed Text 1: Counter({'NN': 13, 'NNP': 5, 'DT': 4, 'IN': 4, 'POS': 4, 'RB': 3, '': 3, 'MD': 3, 'JJ': 3, 'VBN': 2, ',': 2, 'CC': 2, 'VBP': 2, 'PRP$': 1, '': 1, 'VB': 1, 'VBG': 1, 'NNS': 1})
Disputed Text 2: Counter({'NN': 33, 'NNP': 19, 'JJ': 18, 'IN': 17, 'DT': 16, '': 11, ',': 11, 'NNS': 9, 'PRP$': 7, 'VBD': 7, 'VB': 7, 'CC': 7, 'VBZ': 6, 'MD': 5, 'JJ$': 4, 'POS': 4, 'WP': 3, 'RB': 3, 'VBN': 3, 'TO': 3, 'VBP': 3, 'CD': 2, 'PRP': 1, 'RBR': 1, '': 1, 'JJR': 1, 'WP$': 1, 'WRB': 1})
Disputed Text 3: Counter({'NN': 183, 'IN': 117, ',': 96, 'NNP': 94, 'DT': 69, 'JJ': 69, 'CC': 43, 'NNS': 42, '': 26, 'PRP$': 24, '': 23, 'VBZ': 23, 'VBD': 22, 'PRP': 20, 'RB': 17, 'VB': 15, 'VBP': 15, 'VBG': 15, 'VBN': 14, 'TO': 8, 'MD': 7, 'CD': 5, 'JJ$': 4, 'WP': 4, 'EX': 3, 'WP$': 2, 'WRB': 2, 'RP': 1, 'RBS': 1, 'RBR': 1, 'POS': 1})
Punctuation - Actual Author: {'': 53.0, ';': 10.0, '': 12.0, '!': 12.0}
Punctuation - Disputed Text 1: Counter({'': 2, '': 2, '!': 1})
Punctuation - Disputed Text 2: Counter({'': 11, '': 7, '!': 4, ';': 1, '?': 1})
Punctuation - Disputed Text 3: Counter({'': 96, ';': 19, '': 14, '!': 5, '?': 4})

```

Fig 5.6

Disputed Text 1 features a basic punctuation style that does not fit with the actual author.

The punctuation in Disputed Text 2 is more tastefully done, indicating a closer stylistic fit.

The extensive use of punctuation in Disputed Text 3 may indicate complexity, but it also runs the danger of making the text unclear.

Closest Match:

Disputed Text 2

Disputed Text 2 achieves a better balance with its moderate average sentence length, appropriate use of parts of speech, and variety of punctuation styles, but Disputed Text 1 is excessively simplistic and Disputed Text 3 is excessively elaborate. It is still simpler than Disputed Text 3 but has more descriptive features than Disputed Text 1. Despite having a rich descriptive quality, Disputed Text 3 is probably overly complex and verbose in comparison to the author's well-balanced narrative style.

## 5.3 Bootstrap Consensus Tree

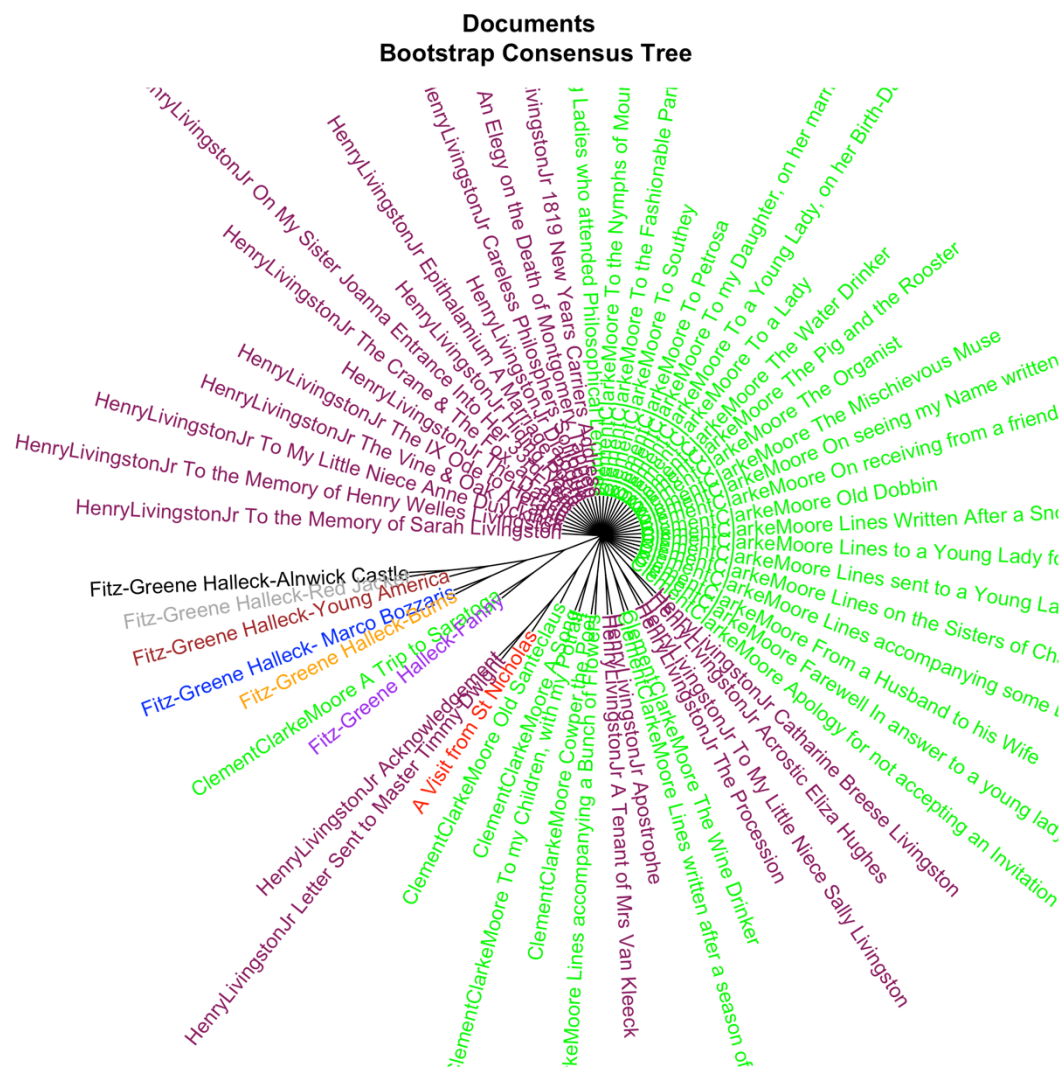


Fig 5.7

After pronouns are removed, the bootstrap consensus tree, which was created using settings for 100–300 MFWs and culling between 0-1%, graphically groups poems according to how similar they are based on word frequencies. The Classic Delta approach was applied, which captures small stylistic changes and is hence ideal for author identification.

Author-Clustered:

The poems by Henry Livingston Jr. (in purple) and Clement Clarke Moore (in green) mostly belong to different groupings, showing a noticeable difference in their word choice and style.

At the lowest point of the tree, Fitz-Greene Halleck's poetry (represented by orange and blue colors) creates a unique cluster that differs stylistically from the works of the other two writers.

Position of A Visit from St. Nicholas:

The red-highlighted poem A Visit from St. Nicholas is more closely aligned with Livingston's cluster and also shares some connection with Moore's poetry. This implies that A Visit from St. Nicholas does not entirely fit into either author's cluster, but rather shifts more toward Henry Livingston Jr.'s style in terms of word frequency and style.

Outliers & Overlap:

There are rare overlaps, such as the poems by Moore and Livingston being put close together, even though the majority of the poems are properly grouped by author. In authorship analysis, this is typical because authors may have stylistic similarities. Separate clusters' consistency, however, supports the idea that their overall writing styles are different.

Given the absence of clustering with the poem, Fitz-Greene Halleck's distinct separation from the other two authors serves to further support the idea that he is not the author of A Visit from St. Nicholas.

Based on the findings, A Visit from St. Nicholas is stylistically closer to the works of Henry Livingston Jr. than it is to those of Fitz-Greene Halleck or Clement Clarke Moore. Even while there is some stylistic resemblance, the consensus tree's closeness to Livingston's poems supports the theory that Livingston is the most likely author. Fitz-Greene Halleck appears as a unique stylistic outlier in our study, which supports the argument that he is not the author of the poem. This gives the argument between Moore and Livingston more support, as the word frequency clustering in the tree suggests a stronger case for Livingston.

## 5.4 Principal Component Analysis

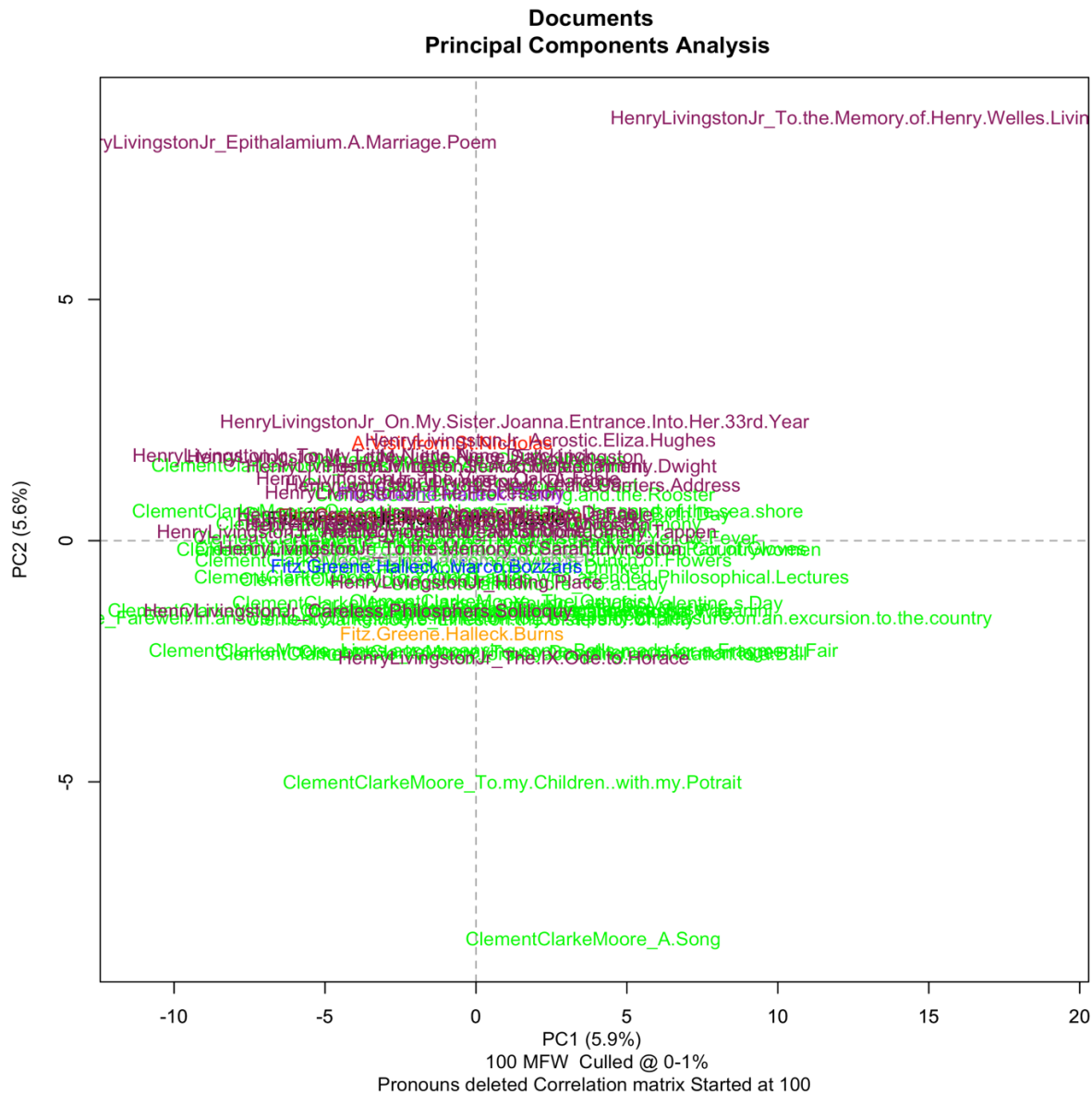


Fig 5.8

The following conclusions can be made from the PCA analysis I carried out as shown in fig. 8:

MFW : 100 to 2000 maximum

Increment : 100

Culling: 0-1%

Listcutoff:1000

Deleting Pronouns: Yes

In the upper-right quadrant of the graph, the texts written by Henry Livingston Jr. (shown in purple) form a distinct cluster, suggesting a distinctive and recognizable writing style. This distinction

between the other authors implies that Livingston's writings differ greatly from those of Fitz-Greene Halleck and Clement Clarke Moore.

Clarke, Clement Moore's writings (seen in green) are grouped primarily on the left, exhibiting an integrated and reliable style. This stands in contrast to Livingston's position, highlighting the distinct stylistic approaches taken by these two authors. Despite being fewer in number, Fitz-Greene Halleck's texts have a higher stylistic affinity with Moore as they more closely correspond with Moore's cluster than Livingston's.

The degree of separation is crucial in this case; the PCA graph indicates that Livingston's texts are constantly separated from Moore's, supporting the theory that their styles differ. Strong evidence suggests that Livingston is the more likely author if the contested text, like "A Visit from St. Nicholas," falls inside Livingston's purple cluster. The distinction in stylistic attributes, as collected by PCA, demonstrates that Livingston's works are stylistically unique enough to justify claims of authorship. Since Henry Livingston Jr.'s writing style clusters away from both Moore's and Halleck's, I can suggest with confidence that he is the likely author of the disputed text based on our PCA analysis.

## 6. Conclusion

By employing a range of advanced stylometric techniques to analyze the authorship of *A Visit from St. Nicholas* I reach to conclusion that Henry Livingston Jr. is the more plausible author of *A Visit from St. Nicholas*. Through the use of various stylometric techniques, such as cluster analysis, Principal Component Analysis (PCA), rolling delta analysis, and sophisticated feature extraction methods, I was able to identify stylistic markers that are more in line with Livingston's works than those of Fitz-Greene Halleck and Clement Clarke Moore. By contributing insightful information on sentence structure, part-of-speech usage, and punctuation patterns, the advanced feature extraction enhanced the case for Livingston's authorship. Looking ahead, exploring additional digital tools utilized by other researchers and incorporating expert opinions on the poetry of Moore, Livingston, and Halleck may provide further context. While my findings favor Livingston as the likely author, the complexities of authorship attribution necessitate ongoing investigation and discourse within the field.

## References

Stylometry with R: A Package for Computational Text Analysis by Maciej Eder, Jan Rybicki and Mike Kestemont (The R Journal Vol. 8/1, Aug. 2016)

<https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf>

The Story of St. Nicholas Published December 21, 2020

<https://www.divineuk.org/articles/how-st-nicholas-became-santa-claus/>

Clement Moore's Poetry

<https://www.henrylivingston.com/xmas/livingstonmoore/allmoorepoetry.htm>

Henry Livingston's Poetry

<https://www.henrylivingston.com/writing/poetry/allhenrypoetry.htm>

The Poetical Works of Fitz-Greene Halleck (1869) by Fitz-Greene Halleck

[https://en.wikisource.org/wiki/The\\_Poetical\\_Writings\\_of\\_Fitz-Greene\\_Halleck](https://en.wikisource.org/wiki/The_Poetical_Writings_of_Fitz-Greene_Halleck)