

A COMPARATIVE ANALYSIS OF WORD FREQUENCY IN “LOOKING FOR ALASKA” AND “PAPER TOWN” USING VOYANT TOOLS

Saffana Asfara Danna¹, Salma Kamelia²

¹²TBIG FTIK UIN K.H. Abdurrahman Wahid Pekalongan

¹saffana.asfara.danna24005@mhs.uingusdur.ac.id,

²salma.kamelia24006@mhs.uingusdur.ac.id

ABSTRACT

This study aims to compare the word frequency patterns in two novels by John Green, Looking for Alaska and Paper Towns, using Voyant Tools as a corpus analysis platform. The research focuses on identifying the most frequently used words, analyzing cross-document term distribution, exploring collocate patterns, and interpreting thematic implications based on the lexical data. Both novels were selected due to their thematic similarities and their popularity among young adult readers, making them suitable candidates for comparative corpus analysis. Data were collected through digital text processing via Voyant Tools, which produced visualizations in the form of Cirrus (word cloud), Trends graph, Document Terms table, and Collocates table. The analysis was conducted systematically by examining each visualization to identify recurring lexical patterns across both texts. Results show that the word "said" dominates both novels with relative frequencies of 8,856 and 7,960 respectively, indicating a strong reliance on dialogue as a narrative device. Character-specific terms such as "colonel," "alaska," "ben," "margo," and "radar" reflect distinct character configurations in each novel. These findings demonstrate how corpus-based tools like Voyant Tools can provide meaningful quantitative insights into literary texts, complementing qualitative literary analysis and opening new possibilities for computational approaches to literary studies

Keywords: frequency, corpus analysis, Voyant Tools

ABSTRAK

Penelitian ini bertujuan untuk membandingkan pola frekuensi kata dalam dua novel karya John Green, Looking for Alaska dan Paper Towns, dengan menggunakan Voyant Tools sebagai platform analisis korpus. Penelitian ini berfokus pada identifikasi kata-kata yang paling sering digunakan, analisis distribusi istilah lintas dokumen, eksplorasi pola kolokasi, serta interpretasi implikasi tematik berdasarkan data leksikal. Kedua novel tersebut dipilih karena kesamaan tematiknya dan popularitasnya di kalangan pembaca remaja, sehingga menjadikannya kandidat yang cocok untuk analisis korpus komparatif. Data dikumpulkan melalui pemrosesan teks digital menggunakan Voyant Tools, yang menghasilkan visualisasi dalam bentuk Cirrus (word cloud), grafik Trends, tabel Document Terms, dan tabel Collocates. Analisis dilakukan secara sistematis dengan memeriksa setiap visualisasi untuk mengidentifikasi pola leksikal yang berulang di kedua teks. Hasil menunjukkan bahwa kata “said” mendominasi kedua novel dengan frekuensi relatif masing-masing 8.856 dan 7.960, yang mengindikasikan ketergantungan yang kuat pada dialog sebagai perangkat naratif. Istilah-istilah yang khas bagi karakter

seperti “colonel,” “alaska,” “ben,” “margo,” dan “radar” mencerminkan konfigurasi karakter yang berbeda di masing-masing novel. Temuan ini menunjukkan bagaimana alat berbasis korpus seperti Voyant Tools dapat memberikan wawasan kuantitatif yang bermakna terhadap teks sastra, melengkapi analisis sastra kualitatif, dan membuka kemungkinan baru bagi pendekatan komputasional dalam studi sastra.

Kata Kunci: frekuensi, analisis corpus, Voyant Tools

A. Introduction

Language in literature is never random. Every word chosen by an author can reflect themes, characters, and writing style (Nosirova, 2025). This means that language is not simply a tool for telling a story, but also a way for authors to communicate deeper values and emotions to their readers. When certain words appear repeatedly in a literary text, they often show important patterns related to the story and the author's way of expressing ideas (Abdul, 2024). These patterns can serve as meaningful clues that help readers and researchers understand what the author truly wants to convey. In recent years, digital humanities have provided new ways to study literature through quantitative analysis, allowing researchers to identify patterns that may not be easily noticed through close reading alone (Jockers, 2013). This development is particularly valuable because it allows literary

analysis to become more systematic and less dependent on subjective interpretation. This shift toward computational approaches has opened new possibilities for analyzing literary language more systematically and objectively (Ramadhani, 2025). That is why, combining traditional literary reading with digital tools can lead to a more comprehensive understanding of a text.

Corpus linguistics is one of the most prominent approaches in this field because it allows researchers to examine large volumes of text through computational methods. According to Ramadhani (2025), corpus-based tools provide quantitative insights from real language use, making it possible to analyze vocabulary patterns, collocations, and thematic structures more systematically. Through corpus analysis, researchers can understand how specific words are distributed and repeated across a text, which in turn

helps reveal the author's stylistic tendencies and thematic priorities.

This study focuses on two novels by John Green, *Looking for Alaska* (2005) and *Paper Towns* (2008). John Green is one of the most well-known authors in contemporary young adult (YA) fiction, whose works are widely recognized for their emotional depth and relatable characters (Cart, 2016). His ability to capture the complexity of teenage life makes his novels particularly interesting to analyze from a linguistic perspective. His novels often explore themes such as teenage identity, love, loss, and the search for meaning, which are reflected through his distinctive use of language and vocabulary (Nosirova, 2025). These recurring themes suggest that Green's word choices are deliberate and closely connected to the emotional core of each story. Both novels have gained wide popularity and have been adapted into films, making them important works in modern YA literature (Cart, 2016). The cultural significance of these novels further justifies their selection as the corpus in this study, as they represent a consistent and recognizable writing style worth examining through corpus-based analysis.

Corpus-based approaches have been increasingly used to analyze literary texts. Ramadhani (2025) highlights that tools like Voyant Tools enable researchers to examine word frequency, collocational patterns, and term distribution across documents, offering a more objective and data-driven perspective on literary language. Similarly, Nosirova (2025) states that lexical analysis can reveal how words, symbols, and metaphors are used to communicate emotional themes and deeper meanings in literary texts. These approaches suggest that combining corpus methods with literary interpretation can produce richer and more comprehensive analyses.

However, despite the growing use of corpus-based tools in literary studies, comparative word frequency analysis specifically targeting John Green's young adult novels remains limited. Most existing corpus studies either focus on a single literary text or compare works across different authors and genres, leaving a gap in research that examines how lexical patterns vary across multiple novels by the same author (Ramadhani, 2025; Liang & Chen, 2023). Furthermore, no study has yet

systematically compared Looking for Alaska and Paper Towns using Voyant Tools to reveal how character-specific vocabulary, dialogue density, and colloquial register differ between the two works within the same authorial style. This study addresses that gap by offering a data-driven comparative analysis of both novels, demonstrating how corpus tools can uncover stylistic patterns and thematic differences that traditional close reading tends to overlook.

This study applies corpus-based text analysis using Voyant Tools, a web-based platform described as a "web-based reading and analysis environment for digital texts" (Sinclair & Rockwell, 2020). Voyant Tools offers more than 20 interactive features, including Cirrus (word cloud), Trends graph, Document Terms table, and Collocates table, all of which make it possible to visualize and compare lexical patterns across multiple texts (Ramadhani, 2025). By examining word frequency patterns, collocates, and word distribution in both novels, this research aims to identify similarities and differences in John Green's writing style and thematic focus, demonstrating how

corpus tools can complement and enrich traditional literary analysis.

Word Frequency Analysis in Literary Studies

Word frequency analysis is a foundational method in corpus linguistics and digital humanities. It involves counting how often specific words appear in a text, which can then be used to draw conclusions about a writer's style, thematic emphasis, and linguistic preferences (Biber et al., 1998). This method is particularly powerful because it transforms subjective impressions about a text into measurable, comparable data. Scholars have long used frequency data to compare literary works, identify authorship, and trace thematic development across a body of texts (Jockers, 2013). This approach has continued to grow in relevance, as recent corpus-based literary studies confirm that word frequency patterns can systematically reveal stylistic differences between texts from the same author or genre (Liang & Chen, 2023). The ability to track patterns across entire novels, rather than relying on selected passages, makes frequency analysis a reliable foundation for literary research.

In the context of literary analysis, high-frequency function words and content words serve different analytical purposes. Function words (e.g., "just," "like," "know") can reveal stylistic register and character voice, while content words (e.g., "alaska," "margo," "colonel") tend to reflect character prominence and thematic focus (Stubbs, 2001; Ramadhani, 2025). Understanding this distinction is important because it allows researchers to analyze both the surface style and the deeper thematic structure of a literary work. Together, they form a lexical portrait of a text that reflects the author's unique way of constructing meaning.

Corpus Tools in Literary Analysis

Voyant Tools, developed by Stéfan Sinclair and Geoffrey Rockwell, is a browser-based platform that offers a suite of text analysis tools for scholars working with digital corpora (Sinclair & Rockwell, 2020). Its accessibility and visual interface make it particularly suitable for classroom and introductory-level research. Among its features are the Cirrus word cloud, the Trends frequency graph, the Document Terms table, and the Collocates panel all of which were employed in this study. The availability

of these multiple visualization formats within a single platform makes Voyant Tools especially efficient for comparative corpus analysis.

Previous research has successfully used Voyant Tools to analyze literary texts across different genres and languages. Ramadhani (2025) demonstrated that Voyant Tools enables researchers to examine word frequency, collocational patterns, and term distribution across documents, offering a more objective and data-driven perspective on literary language. This shows that Voyant Tools is not only applicable in language learning contexts but also in literary and stylistic research. Furthermore, Moretti (2013) argued for a "distant reading" approach that uses quantitative data to study large bodies of literature, a perspective further supported by recent corpus-based literary studies showing that digital tools can uncover patterns invisible to traditional close reading (Liang & Chen, 2023). These frameworks support the use of digital tools not as replacements for close reading, but as complementary analytical lenses that strengthen the overall quality of literary interpretation.

Young Adult Fiction and John Green's Style

Young adult fiction is often characterized by first-person or close-third-person narration, heavy use of dialogue, and a focus on adolescent experiences such as identity formation, love, and loss (Hogan, 2003). These features make YA fiction a particularly interesting genre for corpus analysis, as the conversational and character-driven language tends to produce distinctive and consistent lexical patterns. John Green's writing specifically employs a conversational, witty narrative voice that frequently mimics spoken language. This is reflected in his frequent use of colloquial markers such as "like," "just," and "know" words that create a sense of immediacy and authenticity in his prose (Thomas, 2014). The high frequency of such markers in his novels suggests that Green deliberately constructs dialogue to reflect realistic teenage speech rather than formal literary language.

Both *Looking for Alaska* and *Paper Towns* follow a teenage male narrator who becomes fascinated with a mysterious female character Alaska Young and Margo Roth Spiegelman, respectively. This structural parallel

makes them ideal candidates for comparative analysis, as similarities and differences in word use can be mapped against their shared and divergent narrative patterns.

B. Method

This study uses a descriptive-comparative approach within the framework of corpus linguistics. The aim is not to test a hypothesis but to describe and compare the linguistic features of two texts using quantitative data. This kind of exploratory analysis is well-suited to understanding broad patterns in literary language (Conrad, 2002).

The primary data sources are the full digital texts of *Looking for Alaska* (Green, 2005) and *Paper Towns* (Green, 2008). These texts were processed through Voyant Tools, a free online platform for text analysis. Both novels are well within the YA genre and share thematic similarities that make them appropriate for comparative study.

The texts were uploaded to Voyant Tools and analyzed using four main features: (1) Cirrus, a word cloud visualization showing the most frequent terms; (2) Trends, a line graph showing relative word

frequencies across document segments; (3) Document Terms, a table listing each term with its raw count and relative frequency; and (4) Collocates, a table showing which words most frequently appear near each other. Common stop words (e.g., "the," "a," "and") were automatically filtered by the tool.

Data analysis was conducted in two stages. First, descriptive analysis was performed to identify the top-frequency terms in each novel and compare them across documents. Second, interpretive analysis was applied to discuss what these frequency patterns suggest about character roles, thematic content, and narrative style. This combination of quantitative description and qualitative interpretation aligns with the mixed methods approach commonly used in corpus-based literary studies (Baker, 2010).

C.Results and Findings

This section presents the findings from the corpus analysis of both novels using Voyant Tools, followed by a discussion of their meaning in relation to John Green's writing style and thematic focus. The data were obtained from four

visualization features: Cirrus (word cloud), Document Terms table, Trends graph, and Collocates table. Each subsection below reports the findings from one feature and discusses what those findings reveal about the language and themes of the texts. The analysis aims to answer two key questions: what are the most dominant lexical patterns in both novels, and what do those patterns suggest about Green's narrative style and thematic concerns.

Word Cloud (Cirrus)

The Cirrus visualization generated from the combined corpus of both novels. The size of each word in the word cloud represents its frequency the larger the word, the more often it appears in the texts. This visualization provides an immediate and accessible overview of the most dominant words across both novels before more detailed numerical analysis is conducted.



Picture 1. Cirrus Looking for Alaska and Paper Town

As shown in Figure 1, the word "said" appears at the center of the word cloud in the largest font size, making it visually the most dominant term across both novels. Other highly visible words include "know," "just," "like," "colonel," "margo," "ben," "alaska," and "radar." Words like "thought," "think," "yeah," "good," "going," and "room" also appear at moderate sizes, indicating their secondary but consistent presence across both texts.

The dominance of "said" is the most striking feature of the Cirrus visualization. In corpus-based literary analysis, a word that consistently appears at the top of the frequency list holds a central lexical and narrative role in the text (Biber et al., 1998). The word "said" is a dialogue attribution marker it appears every time a character's speech is reported in the narrative. Its overwhelming prominence in the word cloud therefore suggests that dialogue is not merely a supporting element in Green's writing, but a primary vehicle for storytelling. This finding is consistent with the general characteristics of young adult fiction, which relies heavily on dialogue to reflect authentic teenage speech

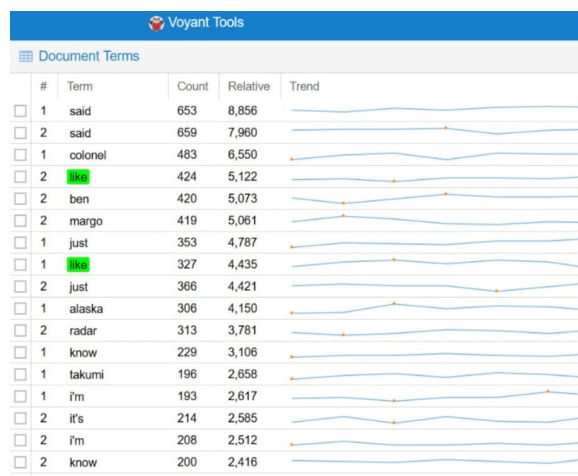
patterns and to develop character relationships (Hogan, 2003). Green's choice to use "said" rather than more expressive alternatives like "shouted," "whispered," or "replied" is also a deliberate stylistic decision. As Thomas (2014) notes, Green favors simple, transparent language that keeps the reader focused on the content of what is being said rather than the manner of its delivery.

Beyond "said," the word cloud also reveals two distinct clusters of vocabulary. The first cluster consists of character names: "colonel" and "alaska" for *Looking for Alaska*, and "ben," "margo," and "radar" for *Paper Towns*. The prominence of these names already signals at a glance that each novel revolves around a specific and tightly defined cast of characters. The second cluster consists of high-frequency function words such as "like," "just," "know," and "yeah." These words are markers of informal, conversational register, typical of teenage speech in everyday contexts (Stubbs, 2001). Their visual dominance in the word cloud confirms that Green's narrative voice consistently mimics spoken language across both novels, creating a sense of immediacy and relatability for young

adult readers. Taken together, the Cirrus visualization offers a first but already meaningful portrait of Green's lexical style: dialogue-heavy, character-centered, and conversationally informal.

Document Terms Table

The Table 1 below presents the top-10 most frequently used non-stopword terms in each novel, along with their raw count, relative frequency, and rank. Relative frequency is used instead of raw count to allow a fair comparison between the two novels, since they differ slightly in total word count.



Picture 2. Top-10 Most Frequent Terms in Both Novels

Term	Document	Raw Count	Relative Freq.	Rank
said	Looking for Alaska	653	8,856	1
said	Paper Towns	659	7,960	1

Term	Document	Raw Count	Relative Freq.	Rank
colonel	Looking for Alaska	483	6,550	2
like	Paper Towns	424	5,122	2
ben	Paper Towns	420	5,073	3
margo	Paper Towns	419	5,061	4
just	Looking for Alaska	353	4,787	3
like	Looking for Alaska	327	4,435	4
just	Paper Towns	366	4,421	5
alaska	Looking for Alaska	306	4,150	5

Table 1. Top-10 Most Frequent Terms in Both Novels

The data in Table 1 confirm what the Cirrus visualization suggested: "said" ranks first in both novels, appearing 653 times in *Looking for Alaska* (relative frequency: 8,856) and 659 times in *Paper Towns* (relative frequency: 7,960). Although the raw count of "said" is slightly higher in *Paper Towns*, its relative frequency is lower, which means that proportionally, dialogue attribution is slightly denser in *Looking for Alaska*. This difference may reflect the fact that *Looking for Alaska* is set largely within the confined social world of a boarding school, where face-to-face conversation between a small group of

characters dominates the narrative. In contrast, *Paper Towns* involves more movement, travel, and internal reflection in its second half, which may reduce the proportion of dialogue relative to narration (Green, 2005; Green, 2008).

The character name data in Table 1 are particularly revealing. In *Looking for Alaska*, "colonel" appears 483 times (rank 2), significantly more often than "alaska" itself, which appears 306 times (rank 5). This is a notable finding. The Colonel whose real name is Chip Martin functions as the narrator Miles's closest friend and constant companion throughout the novel. His name appears more frequently than the novel's title character, which suggests that the day-to-day social and emotional experience of the narrator is more closely tied to The Colonel than to Alaska Young herself. Alaska, by contrast, is a more episodic presence she appears in moments of intensity but is absent for long stretches of the narrative. The frequency data therefore reflect the actual structure of the novel's character relationships in a quantitatively measurable way (Stubbs, 2001).

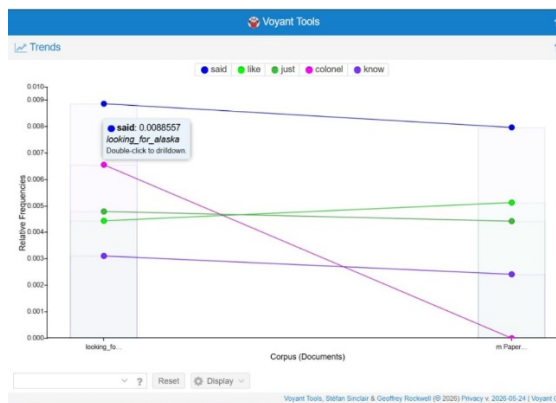
In *Paper Towns*, the pattern is different. "Ben" appears 420 times (rank 3), "margo" 419 times (rank 4), and "radar" appears further down the list. The near-equal frequency of "ben" and "margo" is significant: Margo is the central mystery of the novel, while Ben is the narrator's best friend and constant companion during the search. Their similar frequency suggests a narrative balance between the object of the protagonist's obsession (Margo) and the grounded social reality of his everyday life (Ben). This contrasts with *Looking for Alaska*, where The Colonel significantly outnumbers Alaska in frequency, reflecting a more socially embedded and dialogue-driven story structure.

Both novels also show high frequencies for the function words "like" (4,435 in *LFA*, 5,122 in *PT*) and "just" (4,787 in *LFA*, 4,421 in *PT*). The higher frequency of "like" in *Paper Towns* may reflect the slightly more colloquial and fast-paced dialogue style of that novel, while the higher frequency of "just" in *Looking for Alaska* may indicate a narrative voice that is more immediate and emphatic. As Liang and Chen (2023) note, even small differences in the frequency of common function words can reflect

meaningful stylistic variation between texts by the same author. These findings suggest that while both novels share a recognizable stylistic signature, each has its own subtle lexical identity shaped by its narrative structure and character dynamics.

Trends Graph

The Figure 3 below shows the Trends graph generated from Voyant Tools. The graph displays the relative frequency of five key terms "said," "like," "just," "colonel," and "know" across both novels, with each novel represented as a separate data point on the horizontal axis. This allows direct visual comparison of how each word behaves across the two texts.



Picture 3. Trends Graph from Voyant Tools

The most immediate pattern visible in Figure 2 is the consistent dominance of "said" across both novels, though its relative frequency is slightly higher in *Looking for Alaska* (approximately 0.0089) than in *Paper*

Towns (approximately 0.0080). This confirms the finding from the Document Terms table and reinforces the conclusion that dialogue is the central structural feature of Green's writing in both texts. The fact that "said" remains the highest-frequency content word across two different novels, with different plots and character casts, suggests that this is not an accident of individual story but a stable feature of Green's authorial style (Biber et al., 1998; Hoover, 2003).

The behavior of "colonel" in the Trends graph is the sharpest contrast visible in the data. In *Looking for Alaska*, "colonel" has a relative frequency of approximately 0.0065 the second highest among the five tracked terms. In *Paper Towns*, it drops to near zero, as expected, since The Colonel is a character exclusive to the first novel. This sharp drop serves as a clear visual demonstration of how character-specific vocabulary shapes the overall lexical profile of each novel. The Trends graph makes this character-driven lexical difference immediately visible in a way that raw frequency tables alone cannot (Ramadhani, 2025).

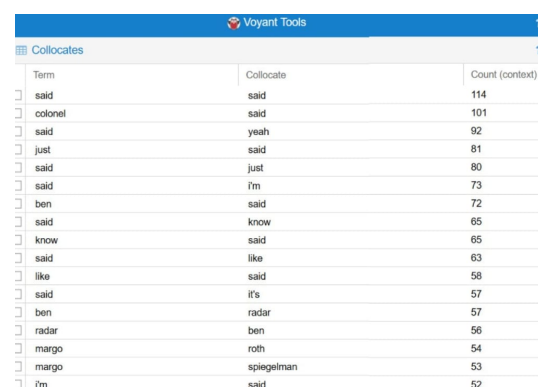
The function words "like" and "just" show a slight increase in *Paper Towns* relative to *Looking for Alaska*. "Like" rises from approximately 0.0044 to 0.0051, while "just" decreases slightly from 0.0048 to 0.0044. As Thomas (2014) argues, colloquial discourse markers like "like" and "just" are among the most reliable indicators of register in young adult fiction, functioning as hedges, fillers, and intensifiers in teenage dialogue. The slight increase of "like" in *Paper Towns* may reflect the more urban, free-roaming social world of that novel, where characters interact across a wider variety of social settings, producing slightly more varied and colloquial speech patterns.

Lastly, "know" remains relatively stable across both novels, with a relative frequency of approximately 0.0025 to 0.0031. This consistency is meaningful: "know" functions primarily as a cognitive verb in teenage discourse, often appearing in phrases like "I don't know," "you know," and "I know." Its stable frequency across both novels suggests that Green consistently uses this word to signal emotional uncertainty, interpersonal acknowledgment, and the reflective

inner voice of his narrators. This is consistent with what Nosirova (2025) describes as the use of lexical repetition to communicate emotional themes in literary texts in this case, the recurring uncertainty and self-questioning that characterize Green's teenage protagonists.

Collocates Table

The Figure 4 below shows the Collocates visualization from Voyant Tools. The Collocates panel identifies which words most frequently appear near each other within a defined context window in this study, five words to the left and right of the target word. A high collocate count indicates that two words tend to co-occur very consistently in the text, which can reveal patterns of language use that individual word frequencies alone cannot capture.



Term	Collocate	Count (context)
said	said	114
colonel	said	101
said	yeah	92
just	said	81
said	just	80
said	i'm	73
ben	said	72
said	know	65
know	said	65
said	like	63
like	said	58
said	it's	57
ben	radar	57
radar	ben	56
margo	roth	54
margo	spiegelman	53
i'm	said	52

Picture 4. Collocate

As shown in Figure 4, the most frequent collocate pairs are dominated by "said." The pair "said–said"

appears 114 times, meaning that the word "said" frequently appears within five words of another instance of "said." This pattern indicates that dialogue exchanges in both novels are very tightly clustered one character's speech is immediately followed by another character's speech, with very little narrative interruption in between. This rapid back-and-forth dialogue structure is a hallmark of Green's conversational writing style and contributes significantly to the fast, immersive reading experience characteristic of his novels (Thomas, 2014).

The collocate pairs "colonel-said" (101 occurrences) and "just-said" (81 occurrences) further confirm the centrality of dialogue in both novels. "Colonel-said" is the second most frequent pair, reflecting that The Colonel is not only the most frequently named character in *Looking for Alaska* but also the character most consistently associated with spoken language he speaks frequently, and his speech is consistently marked with "said." This supports the reading of The Colonel as the most verbally active character in the novel, whose constant conversation with the narrator drives much of the story's

momentum. Similarly, "just-said" suggests that dialogue in these novels is often prefaced or followed by the word "just," functioning as an informal intensifier or hedge in teenage speech (Stubbs, 2001).

The collocate pairs "margo-roth" (54 occurrences) and "margo-spiegelman" (53 occurrences) are particularly interesting. Together, they show that Margo's full name Margo Roth Spiegelman is repeatedly used as a single unit throughout *Paper Towns*. This is unusual; most characters in the novel are referred to by first name or nickname alone. The repeated use of Margo's full three-part name functions as a kind of incantation in the novel, emphasizing her larger-than-life quality and the narrator's sense of awe and mystery toward her. As Green himself constructs the character, Margo Roth Spiegelman is not simply a person but a legend, and the collocate data reflect this through the formal, complete naming pattern (Green, 2008). This is consistent with what Nosirova (2025) describes as the use of lexical choices to communicate deeper emotional and thematic meanings in literary texts.

Taken together, the collocate data reveal that both novels are

structured around dense dialogue exchanges involving a small, recurring cast of characters. The patterns of co-occurrence reinforce the findings from the Document Terms and Trends sections: "said" is the connective tissue of Green's narrative, character names cluster around it, and informal function words fill the spaces between. These collocational patterns are not random they reflect deliberate stylistic decisions that produce a consistent, recognizable narrative voice across both novels (Ramadhani, 2025; Liang & Chen, 2023).

D. Conclusion

This study has analyzed the word frequency patterns in two novels by John Green, *Looking for Alaska* and *Paper Towns*, using Voyant Tools as a corpus analysis platform. By examining four visualization features Cirrus, Document Terms table, Trends graph, and Collocates table this research has produced a set of converging findings that together offer a meaningful quantitative portrait of Green's writing style and thematic priorities.

Across all four features, the word "said" consistently emerges as the single most dominant lexical element

in both novels, with relative frequencies of 8,856 in *Looking for Alaska* and 7,960 in *Paper Towns*. This convergence of evidence confirms that dialogue is not simply a stylistic preference in Green's writing, but a structural foundation of his narrative approach. The collocate data further strengthen this conclusion, showing that dialogue exchanges are tightly clustered throughout both texts, with one character's speech consistently followed by another's in rapid succession. This pattern produces the fast, immersive reading experience that is characteristic of Green's novels and that makes them particularly appealing to young adult readers.

The character name data reveal that each novel constructs its own distinct lexical world. In *Looking for Alaska*, the vocabulary is shaped primarily by "colonel" (483 occurrences) and "alaska" (306 occurrences), while in *Paper Towns* it is shaped by "ben" (420 occurrences) and "margo" (419 occurrences). The fact that The Colonel appears more frequently than Alaska Young in the first novel reflects the actual structure of the narrator's social world a finding that word frequency analysis makes

visible in a measurable and objective way. Similarly, the near equal frequency of "ben" and "margo" in Paper Towns reflects the narrative balance between the protagonist's real friendship and his obsessive pursuit of Margo Roth Spiegelman. These different character configurations produce measurably different lexical profiles, even though both novels share the same author, genre, and general narrative structure.

The function words "like," "just," and "know" appear consistently at high frequencies across both novels and across all four visualization features. Their stable presence confirms that Green's narrative voice is deliberately calibrated to sound like authentic teenage speech informal, immediate, and emotionally direct. This is a key quality of his writing that makes young adult readers feel close to the narrator and invested in the story. The collocate data add a final layer to this picture, showing that Margo's full name Margo Roth Spiegelman functions as a repeated formal unit in Paper Towns, reflecting her status as a larger than life figure in the narrator's imagination rather than simply another character in the story.

Taken together, these findings demonstrate that corpus-based tools like Voyant Tools can provide meaningful and reliable quantitative insights into literary texts, revealing patterns of language use that would be difficult to identify through close reading alone. Digital corpus tools are not replacements for traditional literary analysis, but powerful complements to it tools that make the invisible visible by turning language into measurable and comparable data.

Despite its contributions, this study has several limitations that should be acknowledged. First, word frequency analysis does not distinguish the context in which a word appears. The word "said," for example, may occur in emotionally intense scenes as well as in routine exchanges, but all occurrences are counted equally without regard for their narrative significance. Second, the automatic stop word filtering in Voyant Tools removes common grammatical words such as "I," "she," and "he." In first-person novels like these, the pronoun "I" would almost certainly rank among the most frequent words, and its exclusion means that an important marker of narrative perspective is absent from

the analysis. Third, both novels were analyzed as complete texts without being divided by chapter or narrative section. As a result, it is not possible to identify whether frequency patterns shift across different parts of each novel a dimension that could reveal how the emotional tone and character focus change as each story develops. Future research could address these limitations by combining word frequency analysis with sentiment analysis or narrative segmentation to capture the emotional and structural dimensions of the texts. Such studies would further demonstrate the value of corpus-based approaches in literary research and contribute to the growing field of digital humanities.

REFERENCES

- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Cart, M. (2016). *Young adult literature: From romance to realism* (3rd ed.). ALA Editions.
- Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, 75–95. <https://doi.org/10.1017/S0267190502000045>
- Green, J. (2005). *Looking for Alaska*. Dutton Books.
- Green, J. (2008). *Paper towns*. Dutton Books.
- Hogan, W. (2003). *Humor in young adult literature: A time-honored truth*. Scarecrow Press.
- Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, 37(2), 151–178. <https://doi.org/10.1023/A:1022678806138>
- Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- Liang, Y., & Chen, H. (2023). A corpus-based stylistic study of Jane Eyre and Wide Sargasso Sea. *Journal of Social Science Humanities and Literature*, 6(1), 45–53.
- Moretti, F. (2013). *Distant reading*. Verso.
- Ramadhani, D. (2025). Corpus-based Voyant Tools for teaching IELTS writing: Design and implementation. *JETAL: Journal of English Teaching & Applied Linguistics*, 7(1), 72–79.
- Sinclair, S., & Rockwell, G. (2016). *Voyant tools*. <https://voyant-tools.org>
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.
- Thomas, E. E. (2014). Young adult literature and the spectrum of reading. *The ALAN Review*, 41(3), 29–36. <https://doi.org/10.21061/alan.v41i3.a.4>