Stylistic Variations in Neil Gaiman's Work: A Corpus-Driven Analysis

Hongxu Zhou h.zhou.16@student.rug.nl

Table of contents

1 Introduction

Stylistic variation in literature reflects how authors adjust their writing to suit different purposes, contexts, and audiences (Karlgren, 2010). This phenomenon becomes particularly interesting when exam-

ining authors who write for diverse age demographics, from young children to adults. Neil Gaiman is a compelling case study in this regard—his substantial body of work spans picture books for young readers and distinctly adult works.

Readers and critics often intuitively recognise differences between literature written for children versus adults. One example is the genre tags on Goodreads. However, the specific linguistic and stylistic elements that mark these differences remain less systematically documented. Traditional literary analysis typically approaches such questions through close reading and qualitative assessment. However, computational methods now offer opportunities to examine stylistic patterns across larger text collections with greater objectivity.

This study takes a data-driven approach to investigate stylistic variation across Gaiman's works targeted at different age groups. Rather than beginning with predetermined theories about how children's and adult literature should differ, this research allows patterns to emerge from the texts themselves. The analysis examines multiple dimensions of language use—from word-level features like vocabulary richness and complexity to sentence structure, narrative techniques, and emotional content.

This project aims to identify the key stylistic features that differentiate Gaiman's writing across audience demographics and explore to which extent these patterns align with theoretical expectations about children's versus adult literature. By analysing these features systematically, the study contributes to our understanding of how authors adapt their writing for different age groups while maintaining their distinctive authorial voice.

2 Literature Review

Stylistic variation in texts arises from authors making choices about how to organise material, select vocabulary, and craft their message for specific readers. These choices form patterns that scholars can detect, analyse, and interpret using both qualitative and quantitative methods. Defined by Karlgren (2010), style is "the information carried in a text when compared to other texts, or in a sense compared to language as a whole". These stylistic choices are not peripheral but form an important part of the intended communication.

Previous research has identified several key dimensions along which texts vary. Biber (1988) proposed dimensions such as "Involved vs Informed, Narration vs Argumentation, Personal vs Impersonal" as the foundation of textual variation. These dimensions help researchers move beyond surface-level observations to deeper patterns in how language functions across different contexts.

Hoover (2017) distinguishes between inter-textual variation (differences between texts by different authors or in different genres) and intra-textual variation (differences within the works of a single author). While authorship studies often focus on finding consistent markers of individual style across an author's works, Hoover argues that intra-textual variation offers equally valuable insights: "Style variation helps to create memorable styles and memorable characters and narrators, and it is often crucially linked to description, setting, theme, plot, and other textual features".

Karlgren (2010) further distinguishes between individual and situational factors that create stylistic variation. Individual variation stems from author preferences and personal style, while situational variation arises from "constraints imposed by situation the text is produced in". This distinction helps explain why the same author might write differently when targeting different audiences. Karlgren identifies three levels of stylistic constraints: highly formalised rule systems (like spelling and grammar), contextual conventions (lexical choices appropriate to genre), and textual organisation (where authors operate with minimal formal guidance). These levels provide a useful framework for examining how authors adapt their writing for different readerships.

Computational stylistics has evolved significantly over recent decades, moving from simple word frequency counts to sophisticated multi-dimensional analysis. Fialho & Zyngier (2023) trace this evolution through various historical branches of computational stylistic analysis. They note that early work in stylometry focused primarily on authorship attribution, while more recent approaches have expanded to consider genre, register, and other aspects of stylistic variation.

Despite the progress, the computational analysis of literary texts has faced criticism from traditional literary scholars. Hammond et al. (2013) describe a "two cultures" problem between computational linguists and literary scholars, noting fundamental differences in how each field approaches ambiguity: On the one hand, computational linguistics generally regards ambiguity as an obstacle to be resolved for the sake of clarity and precision; literary scholarship on the other hand —especially since the early twentieth century—has emphasised the value of ambiguity and polysemy, viewing them as integral to the literary experience. From this perspective, the goal of analysis is not to eliminate ambiguity, but to examine and interpret its various dimensions. This epistemological divergence complicates efforts to analyse literary style using computational methods.

Gius & Jacke (2022) address the criticism that computational literary studies are "structuralist" in their approach. They note that digital approaches to literature are sometimes characterised as reducing literary works to formalistically describable and objectively countable objects. This criticism emerges from concerns that computational methods prioritise linguistic patterns over interpretive depth. However, Gius and Jacke argue that while computational approaches do focus on formal textual features,

they do not need to reduce literary meaning to these features alone. Instead, they suggest computational methods can advance the breadth and depth of traditional interpretive approaches.

Despite growing interest in computational stylistics, few studies have directly examined how a single author varies their style when writing for different age demographics. Most existing research either compares different authors or focuses on specific linguistic features without a comprehensive approach to stylistic variation. This gap is significant because it raises an important question: are the differences between children's and adult literature a result of different authors writing in distinct ways, or do they reflect deliberate stylistic adjustments made by the same author for different audiences?

Prior research on children's books, such as Dawson et al. (2021), suggests that authors make specific linguistic adaptations when writing for younger audiences, though these adaptations are more nuanced than simple reduction in complexity. However, these observations have rarely been tested through systematic computational analysis of texts by authors who write for different age groups.

This study addresses these gaps by examining Neil Gaiman's works across the age-demographic spectrum, asking:

1. What linguistic features differentiate Gaiman's writing across works intended for different age demographics?
2. To what extent do these differences form coherent patterns that align with the intended audience age?
3. How do the observed patterns of stylistic variation compare with the linguistic characteristics of children's books identified by Dawson et al. (2021) ?

3 Methodology

This study uses a multi-level analysis approach. The corpus design enables examination of lexical, sentence, and discursive features to address RQ1, while the integrated analyses provide insights into the patterns of variation addressing RQ2. Throughout the analysis, findings are compared with expectations from the literature on children's books (RQ3).

3.1 Corpus

This study built a corpus comprising ten books by Neil Gaiman, selected to represent the author's rage across audience demographics[1]. The ten books are:

Table 1: Books in Gaiman Corpus

Book Title	Abbr.	Goodreads Genre Tags
American Gods	AG	Adult
Anansi Boys	AB	Adult
Neverwhere	NW	Adult; Young-Adult
Smoke and Mirrors	SM	Adult; Young-Adult
The Ocean at the End of the Lane	OCEAN	Young-Adult; Adult
Coraline	CL	Young-Adult; Children;
The Graveyard Book	TGB	Young-Adult; Children
Fortunately, the Milk	FTM	Children, Young-Adult
Odd the the Frost Giants	OFG	Children, Young-Adult
Stardust	ST	Young Adult, Adult

They span Gaiman's career and represent various genres within his work. Therefore, this corpus is a balanced representation for analysis of potential stylistic variations. Among the ten books, *Fortunately, the Milk* (FTM) and *Smoke and Mirrors* (SM) are special in different ways. *FTM* is a picture book, with its first chapter containing no text at all, while *SM* is the only short story collection, comprising 35 short stories and narrative poems.

Each books was obtained in EPUB format, which preserved textual content while maintaining structural elements that enables programming analysis. The `epubr` package (Leonawicz, 2025) was used to extract and parse content from these books while preserving metadata and chapter boundaries.

The corpus was designed with multi-tired structure to accelerate analysis at different levels of linguistic granularity:

1. **Document-level**: Each chapter/single story constitute a document. This layer of corpus keeps narrative unites as defined by the author, and so enables analysis of chapter-level stylistic patterns and narrative structure.

[1] Code can be found on GitHub.

2. **Sentence-level**: The text is segmented into individual sentences using the `tokenizers` package (Mullen, n.d.), which applies rule-based sentence boundary disambiguation. This intermediate level allows for analysis of syntactic patterns and sentence complexity.
3. **Token-level**: Individual words are extracted for lexical analysis. Tokens maintain references to their source sentences and documents, enabling multi-level analysis.

Additionally, two supplementary components are created:

1. **Metadata layer**: It contains corpus-level statistics (total size, unique vocabulary) and book-level metadata (lexical diversity, average sentence length).
2. **Vocabulary layer**: It provides frequency and distribution for each unique token in the corpus.

As a result, the corpus contains 711,385 words across 154 documents (chapter/story), with an average of 15.4 chapters per book. The corpus contains 60,127 sentences with an average length of 12.3 words per sentence. The vocabulary size (type) is 28, 968.

3.2 Analysis

3.2.1 lexical level

The lexical analysis examines word-level patterns that might distinguish text for different audience groups. I analysed word length distribution, lexical diversity, and distinctive vocabulary to investigate potential stylistic variations.

3.2.1.1 Word Length Distribution

Word length often correlates with text complexity — longer words typically appear in more advanced or specialised text (Biber, 1988). Figure 1 presents the percentage of words at each length across Gaiman's works. This visualisation reveals a remarkably consistent pattern across all books, with most words averaging around 4 characters regardless of the intended audience. This suggests that at the the basic word length level, Gaiman maintains a relatively stable style.

3.2.1.2 Lexical Diversity

Lexical diversity is measured using the type-token ratio (TTR), which represents vocabulary richness:

$$Lexical \quad Diversity = \frac{Number \quad of \quad Unique \quad Words}{Total \quad Word \quad Count}$$

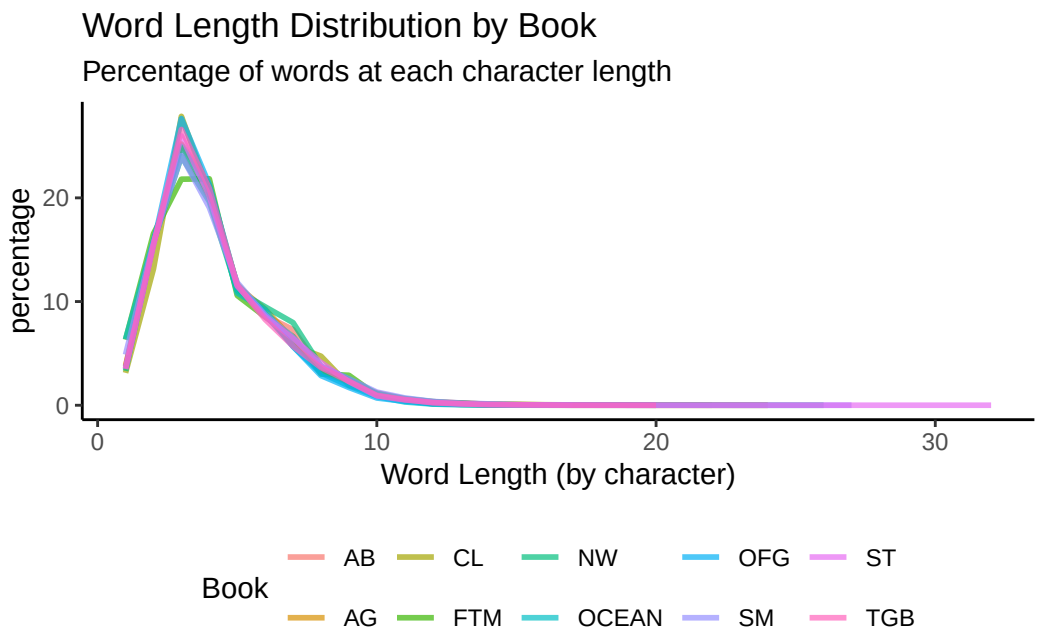

Figure 1: Word Length Distribution by Book

Figure 2 displays the lexical diversity values cross the 10 books. Interestingly, the TTR values do not clearly separate along audience lines. Even, Fortunately, the Milk, a typical picture book for children, has the highest TTR value. This finding may virtually challenge the assumption that books for younger readers use simpler vocabulary. However, it is acknowledged that TTR is sensitive to text length – longer texts naturally tend toward lower TTR values due to necessary repetition of function words and proper names. This limitation suggests the need for additional, more sophisticated linguistic measures beyond basic lexical metrics.

3.2.1.3 Keyword analysis

To identify vocabulary patterns distinctive to each book, I use a log-odds ratio analysis to compare word frequencies within each book against the rest of the corpus.

Figure 3 shows the top distinctive words for each book. The keyword analysis reveals thematic differences more than stylistic ones: distinctive words primarily reflect characters and plot elements rather than systematic vocabulary differences based on audience age.

The lexical results suggest that Gaiman does not significantly simplify his vocabulary for younger readers. This finding may contraindicate conventional assumptions about children's literature. The un-

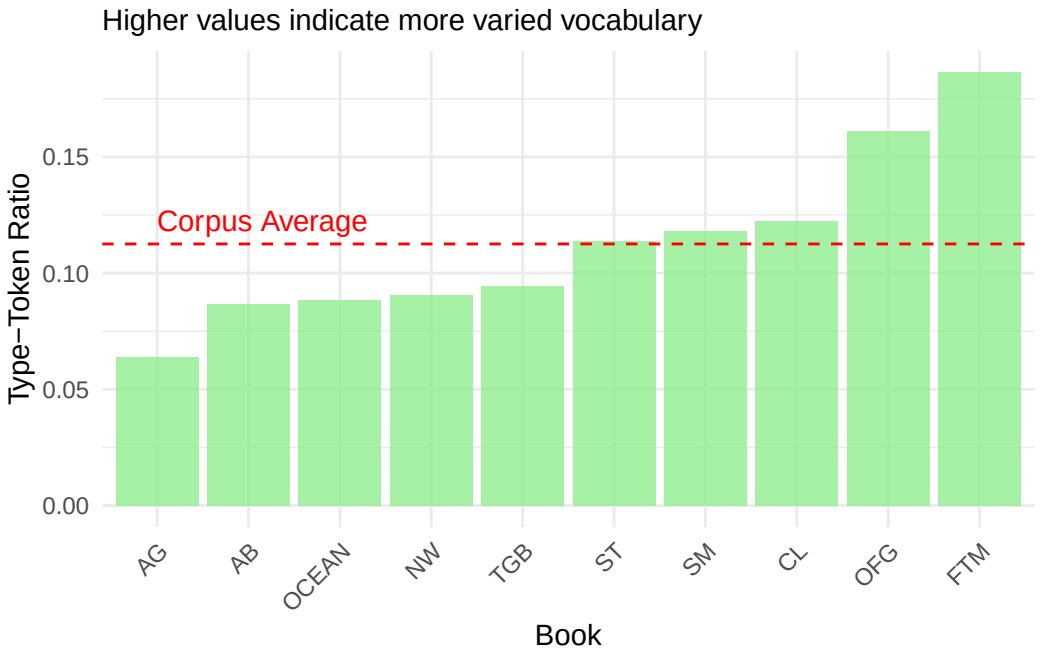

Figure 2: Lexical Diversity by Book

expected pattern prompts to look beyond word-level features toward sentence structure and complexity, where more pronounced stylistic variations might emerge.

3.2.2 Sentence Level

The sentence level analysis provides deeper insights into the stylistic variation in the corpus than word-level features alone. Based on sentence structure analysis, I examine complexity, readability and content characteristics across different works.

3.2.2.1 Sentence length analysis

```
Picking joint bandwidth of 1.35
```

As a fundamental metric, sentence length offers initial insights into text complexity. Figure 4 displays the distribution of sentence lengths across Gaiman's books. While all of them show similar central tendencies with peaks around 7-12 words per sentence, *Fortunately, the Milk* and *Odd and the Frost Giants* display narrower distributions with fewer long sentences. The more controlled sentence length

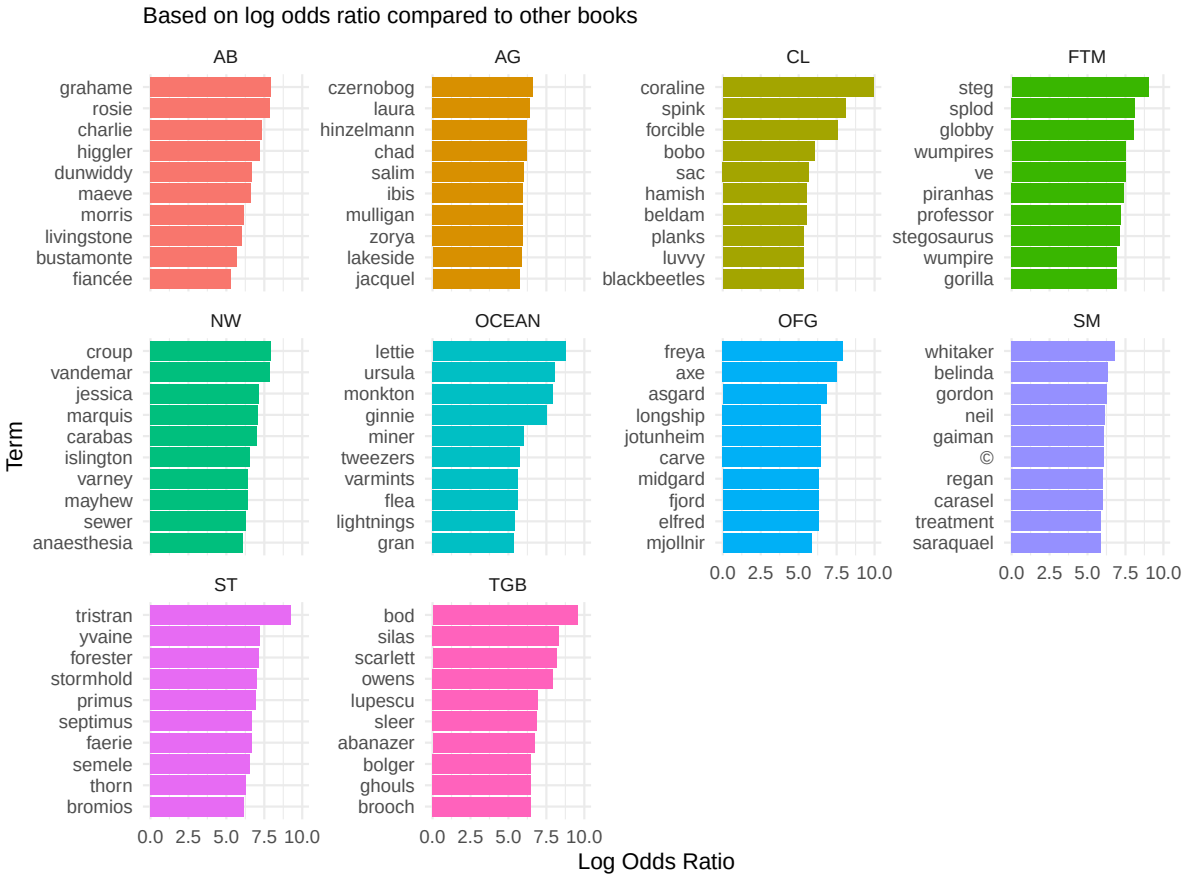

Figure 3: Top 10 Keywords by Book

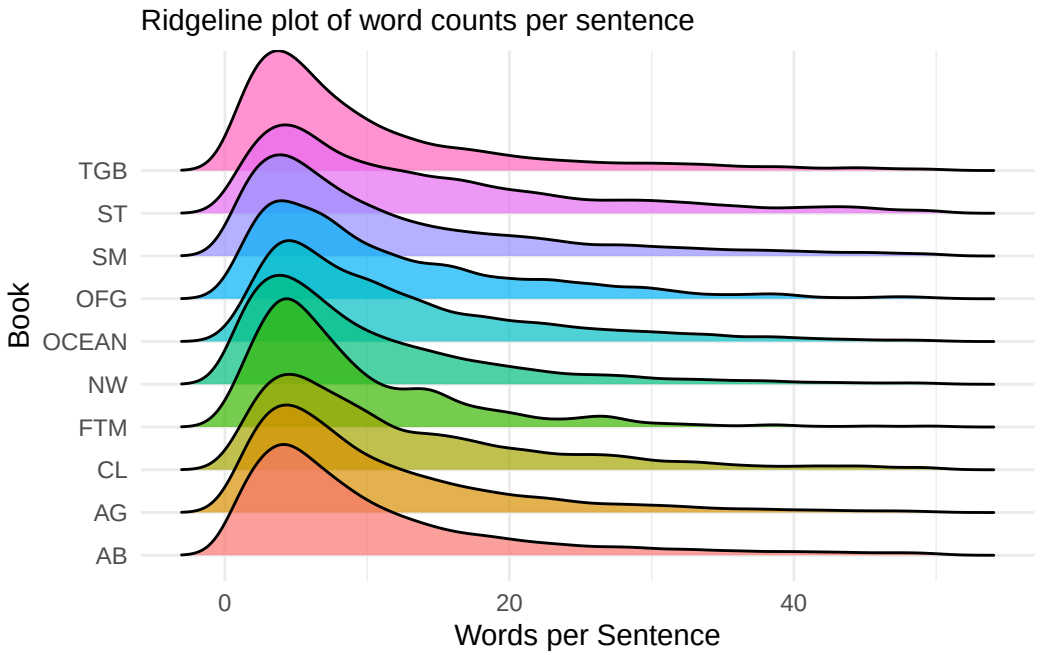

Figure 4: Sentence Length Distribution by Book

variance suggests that they are more likely for children, though the differences are subtle and require deeper syntactic analysis.

3.2.2.2 Syntactic Complexity

Syntactic complexity is measured through dependency parsing, which captures the relationships between words in a sentence. Mean dependency length (MDL) is a method that calculates the average distance between syntactically related words. It represents language comprehension difficulty (Haitao Liu, 2008). Higher MDL values indicate more complex sentence structures, for example, with embedded or non-adjacent relationships.

This analysis sampled 100 sentences from each book and parsed them using the Universal Dependency framework provided by `udpipe` (Straka et al., 2016). Figure 5 presents the mean dependency length for each work. Three books that are generally considered for adults, *American Gods, Neverwhere,* and *Anansi Boys* show consistently higher MDL values (from 3.5 to 4.2) compared to those typically for children and younger audiences (*Odd and the Frost Giants, Fortunately, the Milk*), which show lower values (from 2.4 to 3). This pattern suggests Gaiman systematically adjusts syntactic complexity based on intended audience.

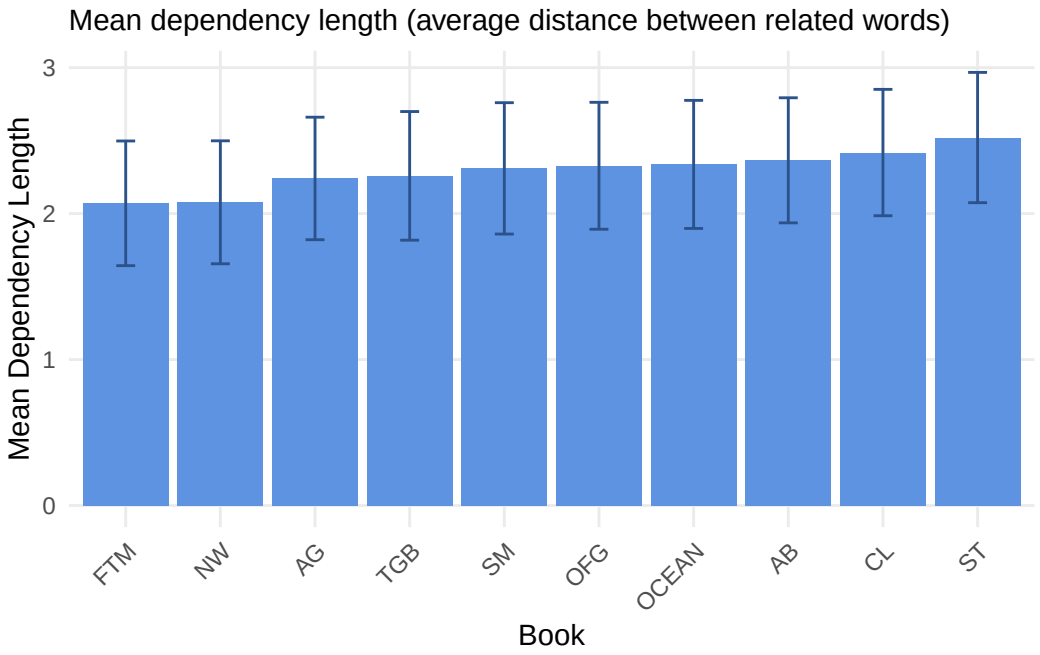

Figure 5: Syntactic Complexity in Gaiman's Works

3.2.2.3 Readability Analysis

The Flesch-Kincaid Grade Level scores provide standardised readability measurements based on word and sentence characteristics. Figure 6 shows these scores across Gaiman's works calculated by `quanteda` (Benoit et al., 2018) .

The results demonstrate a unexpected patterns that challenge conventional assumptions about audience targeting. All books are below 8th grade which fits Gaiman's niche as a writer of popular fiction. *American Gods*, often categorised as an adult fiction, scores lower than *The Graveyard Book*, which is tagged as "Middle Grade" and "Young Adult" on goodreads. This suggests that readability scores alone cannot reliably predict audience targeting in Gaiman's work.

This finding is echoed by recent criticism of readability metrics. Tanprasert & Kauchak (2021) argue that "Flesch-Kincaid Grade Level (FKGL) should not be used to evaluate text simplification systems," noting that the score can be easily manipulated with minor textual changes. FKGL formulas primarily rely on surface features such as word and sentence length, while fail to capture semantic complexity, narrative sophistication, thematic maturity, or cultural references that can better differentiate works for different audiences.

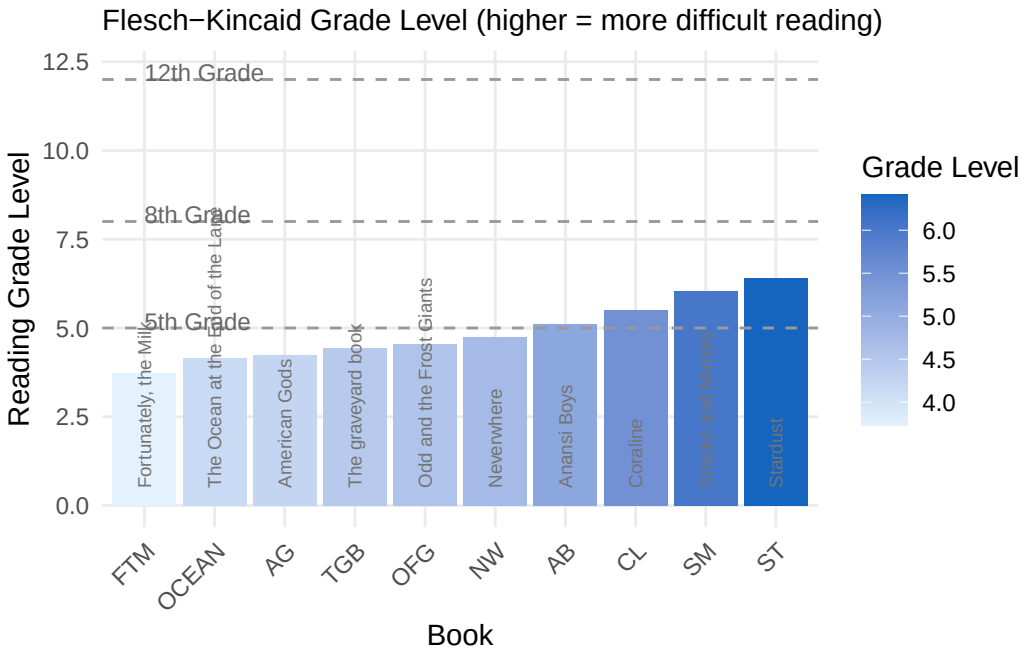

Figure 6: Reading Difficulty of Gaiman's Works

3.2.2.4 Profanity Analysis

A dictionary-based approach was used to detect and categorise profanity across words. Three categories were established:

- **mild profanity**: for example, "damn"
- **moderate profanity**: for example, "ass, shit"
- **strong profanity**: for example, "fuck, cunt"

The complete list of profanity can be found in Appendix. Figure 7 shows the frequency of profanity per 1,000 words for each book.

The data reveals notable variations in profanity usage across Gaiman corpus. *American Gods* shows the highest overall profanity rate (2.28 per 1,000 words), followed by *Smoke and Mirrors* (1.32 per 1,000 words), and *Neverwhere* (1.02). At the lower end, *Fortunately, the Milk* (0.13), *Coraline* (0.19), and *The Ocean at the End of the Lane* (0.24) contain minimal profanity. The figure also indicates the frequency and intensity of profanity are positively associated. The patterns of profanity distribution directly reflect the features that can influence demographic targeting such as thematic concerns, narrative tone, and publication context.

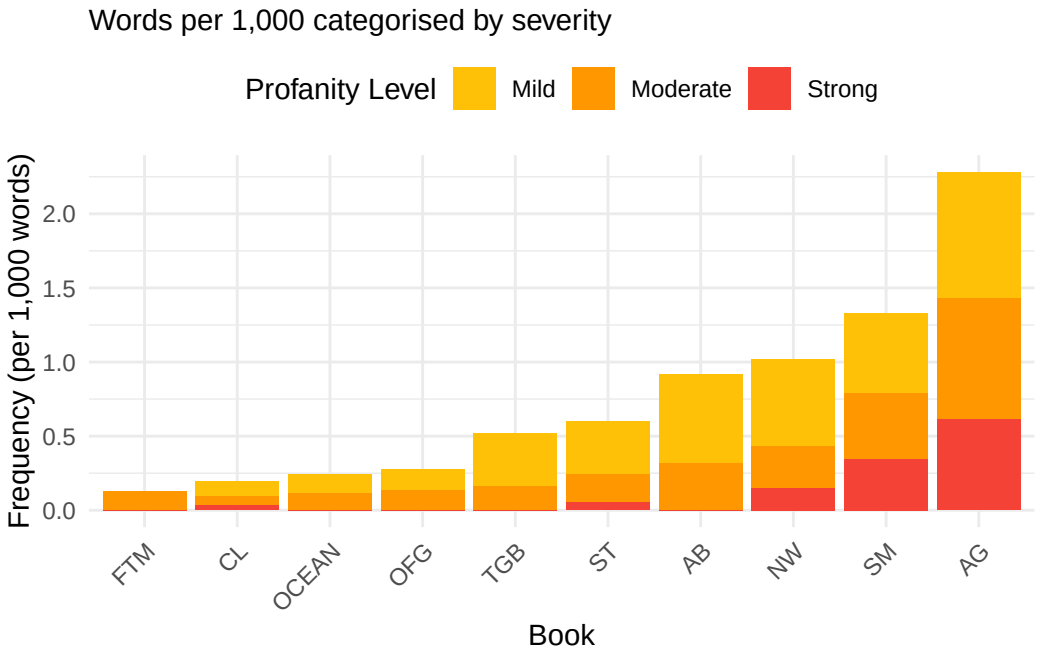

Figure 7: Profanity in Gaiman's Works by Intensity Level

3.2.2.5 Sentiment Analysis

I calculated sentence-level sentiment variation by `sentimentr` (Tyler W, 2021). Figure 8 shows the sentiment distribution across books.

The analysis demonstrates considerable consistency in sentiment distribution across Gaiman's works. Most books maintain a similar balance of emotional content (approximately 23 - 31% positive, 18 - 27% negative, and 43 - 52% neutral). *Odd and the Frost Giants* shows the highest proportion of negative sentences (33.45%) while *Stardust* shows the highest proportion of positive sentences (31%).

The emotional arcs — how sentiment progresses through the narrative – show some variation. Figure 9 illustrates these emotional trajectories. *Smoke and Mirrors* (SM) shows the highest sentiment range (0.46 with starting sentiment at -0.03 and ends at 0.23). However, this pattern likely stems from its nature as collection of short stories rather than representing a stylistic choice related to audience targeting. In general the sentiment arcs of the ten books vary from -0.23 to 0.24. The overall stable sentiment distribution across works suggests that one the one hand, Gaiman maintains a consistent emotional palette regardless of intended readership; on the other hand, `sentimentr`, as a dictionary-based approach, may fail to catch some subtle sentiments such as irony and satire. Unlike profanity usage and syntactic complexity, sentiment patterns do not appear to correlate meaningfully with potential audience

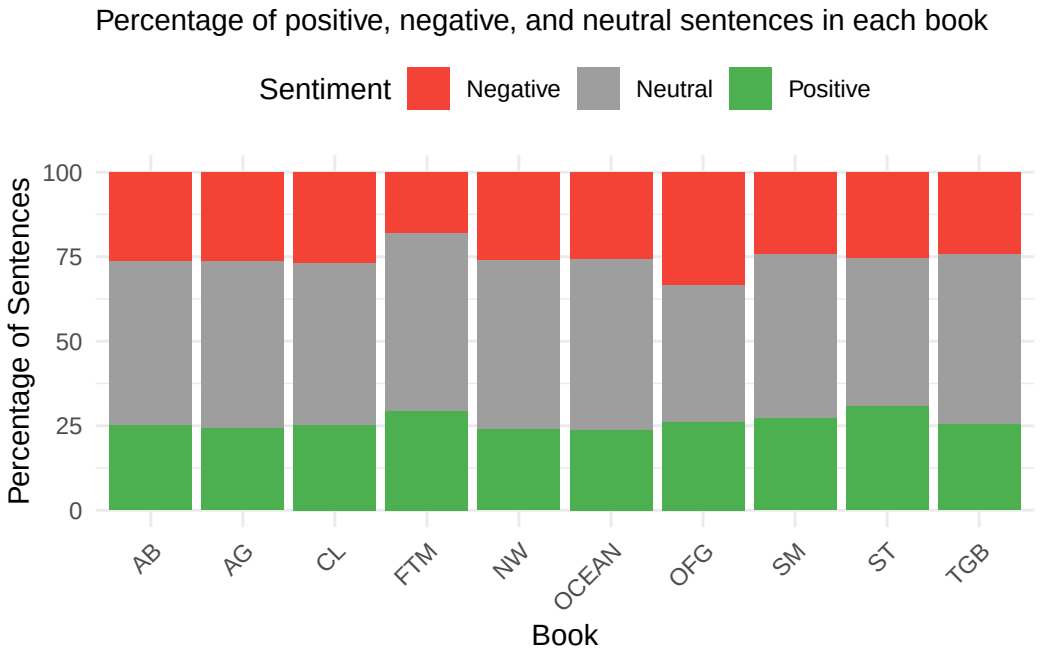

Figure 8: Sentiment Distribution in Gaiman's Works

demographics.

In sentence-level analysis, profanity usage appears to be the most distinctive content-based marker, with clear patterns of variation across different works. They create groupings that partially but not perfectly align with traditional audience categorisations. The sentence-level features provides more nuanced insights than lexical analysis alone, suggesting that Gaiman's stylistic adaptations for different audiences operate more at the structural and content-appropriateness levels than through vocabulary simplification. The findings so far pave the way to examine discourse-level patterns to develop a more comprehensive understanding of potential audience-based stylistic variations.

3.2.3 Discourse level

After examining word- and sentence-level patterns, this section moves further by exploring discourse-level characteristics in Gaiman's work. It focuses on how textual elements function together to create meaning across larger units of text. In literature stylistics, discourse features often reveal how authors construct narrative voice, manage perspective, and guide reader engagement – all of which may differ based on intended readership.

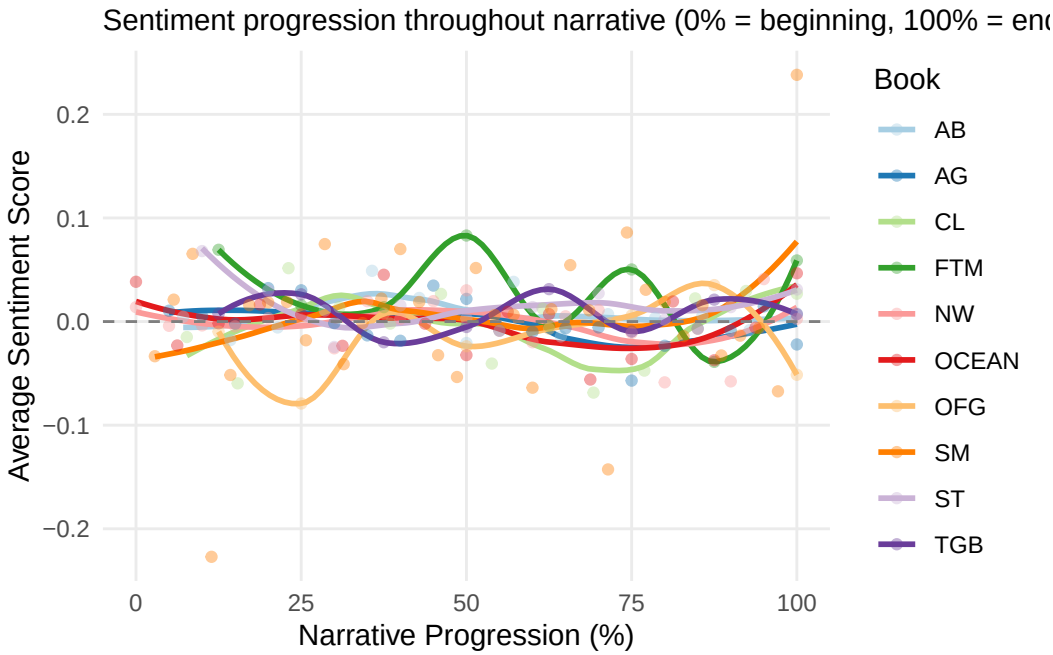

Figure 9: Emotional Arcs in Neil Gaiman's Works

3.2.3.1 Point of View analysis

Point of view (POV) represents a fundamental aspect of narrative discourse. The analysis tracked personal pronoun usage across the 10 books in the corpus to identify patterns in narrative perspective. Figure 11 shows the composition of first (I/me), second (you), and third (she/he/they) pronouns as percentages of total words across the corpus.

Books like *Fortunately, the Milk* (FTM) and *The Ocean at the End of the Lane* (OCEAN) show higher rates of first-person narration compared to books like *American Gods* (AM), *Coraline* (CL), and *Neverwhere* (NW). This suggest that POV alone is not a feature differentiating the target audience group of a book.

Perspective stability analysis, on the other hand, reveals clearer patterns that may related to audience considerations. Figure 10 displays the narrative perspective stability index. It measures how consistently each book maintains its dominant perspective without shifts. Books generally considered appropriate for younger readers (FTM and OFG) show higher stability indices (0.72 - 0.8), while books aimed at older audiences (AG, AB) demonstrate more perspective shifting (0.53 - 0.58). However, ST, a books that usually considered for young adults, has the highest shifting score.

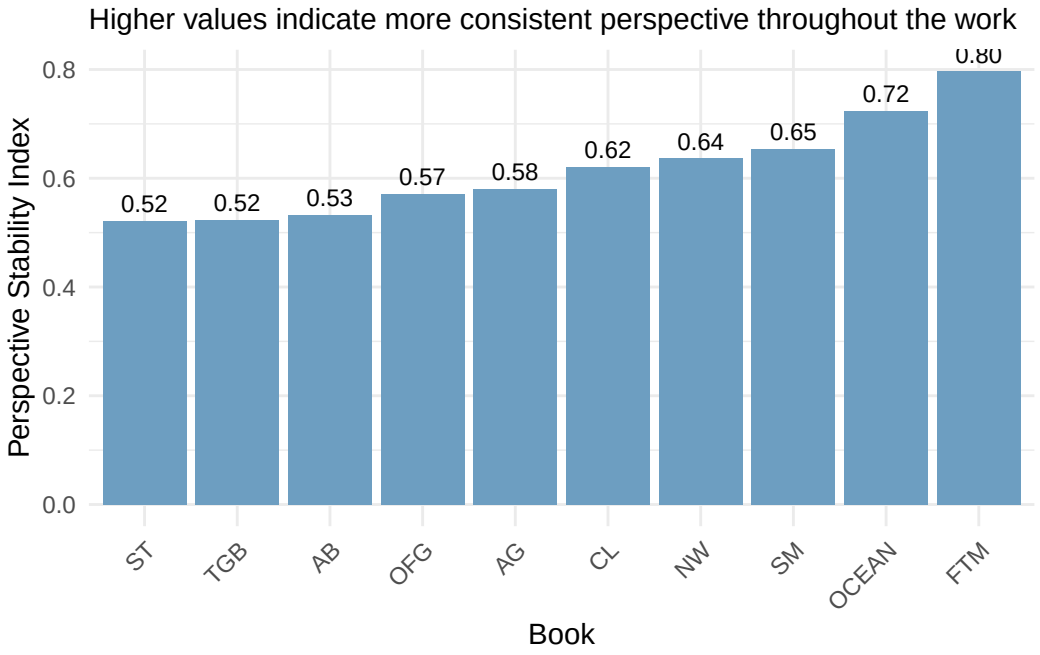

Figure 10: Narrative Perspective Stability in Gaiman's Works

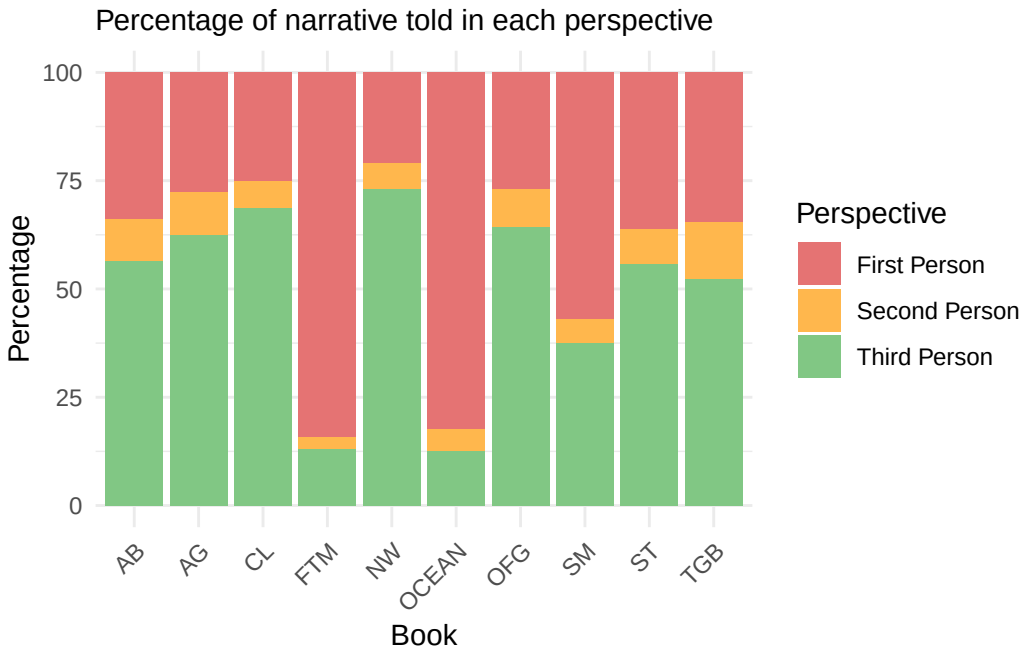

Figure 11: Narrative Perspective Composition in Gaiman's Works

3.2.3.2 Modal verbs analysis

Based on the framework of modal grammar (Simpson, 2003), I categorised modal verbs into four functional groups:

1. **World Building** (must, should, have to, always, never): These establish rules, obligations, and certainties within the narrative world.
2. **Possibility** (may, might, could, can, maybe): These express uncertainty and potential options.
3. **Hypothetical** (would, if, whether): These create conditional or speculative scenarios.
4. **Intent** (will, shall, going to, want to): These signal future actions and determinations.

The complete list can be found in Appendix. The distribution of the four groups of modals across the books is shown in Figure 12.

The most striking difference appears in the "authority ratio" (Figure 13) —the ratio of world-building and intent modals (groups 1 and 4, expressing certainty) to possibility and hypothetical modals (groups 2 and 3 which express uncertainty). *Fortunately, the Milk* (FTM) stands out with a distinctly high authority ratio (0.96), substantially higher than all other books which range between 0.42-0.53. This further challenges the attempt to categorise a work into a dichotomy of children/adults literature based on linguistic features alone. In fact, books like *Coraline* (0.45) and *The Graveyard Book* (0.45) have authority ratios comparable to or even lower than typically presumed adult-oriented works like *American Gods* (0.53) and *Smoke and Mirrors* (0.53).

This finding suggests that modal usage in Gaiman's work does not neatly align with presumed audience demographics. The exceptional authority ratio in *Fortunately, the Milk* may reflect its unique status as a humorous children's picture book rather than representing a general pattern for all children's literature. The similarity in authority ratios across most other books indicates that Gaiman maintains a relatively consistent approach to modal expression regardless of intended audience.

3.2.4 Integrated Analysis

The previous sections examined Gaiman's stylistic variations at multiple linguistic levels. While individual analyses revealed noteworthy patterns, they also highlighted the limitations of single-dimension approaches to categorising works by intended audience. This section integrates findings across lexical, sentence, and discourse levels to develop a more comprehensive understanding of stylistic variation in Gaiman's works.

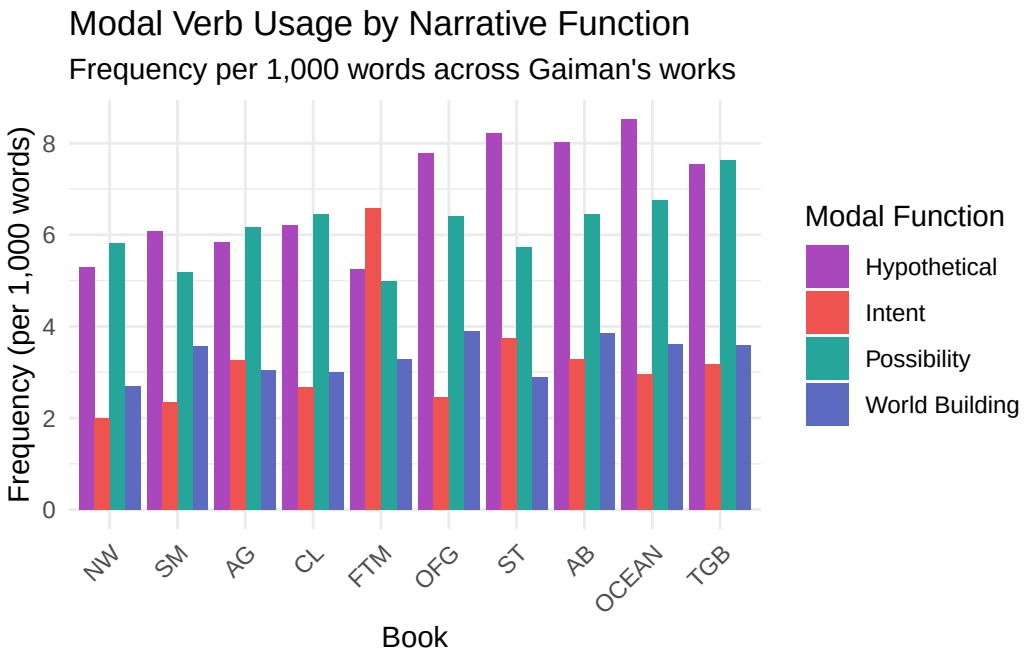

Figure 12: Modal Verb Usage by Narrative Function

To synthesise the multidimensional linguistic data, this study used Principal Component Analysis (PCA), a statistical technique that identifies patterns across multiple variables. PCA transforms potentially correlated variables into a smaller set of uncorrelated variables called principal components. Each component captures a portion of the total variation in the data, with the first component capturing the largest amount of variation, the second component capturing the second largest, and so on.

For this analysis, 12 key metrics from previous analyses were included:

- **Lexical features**: lexical diversity, average word length, percentages of short and long words
- **Sentence features**: average sentence length, percentages of short and long sentences, mean dependency length, readability scores
- **Content features**: sentiment percentages (positive/negative), profanity frequencies
- **Discourse features**: POV distributions, perspective stability, authority ratio

The PCA results reveal two distinct dimensions of stylistic variation in Gaiman's works. The first principal component, accounting for 42% of the total variance, corresponds primarily to linguistic complexity. This dimension shows substantial positive correlations with features that indicate more sophisticated language use: longer sentences (loading of 0.369), greater average word length (0.366), higher

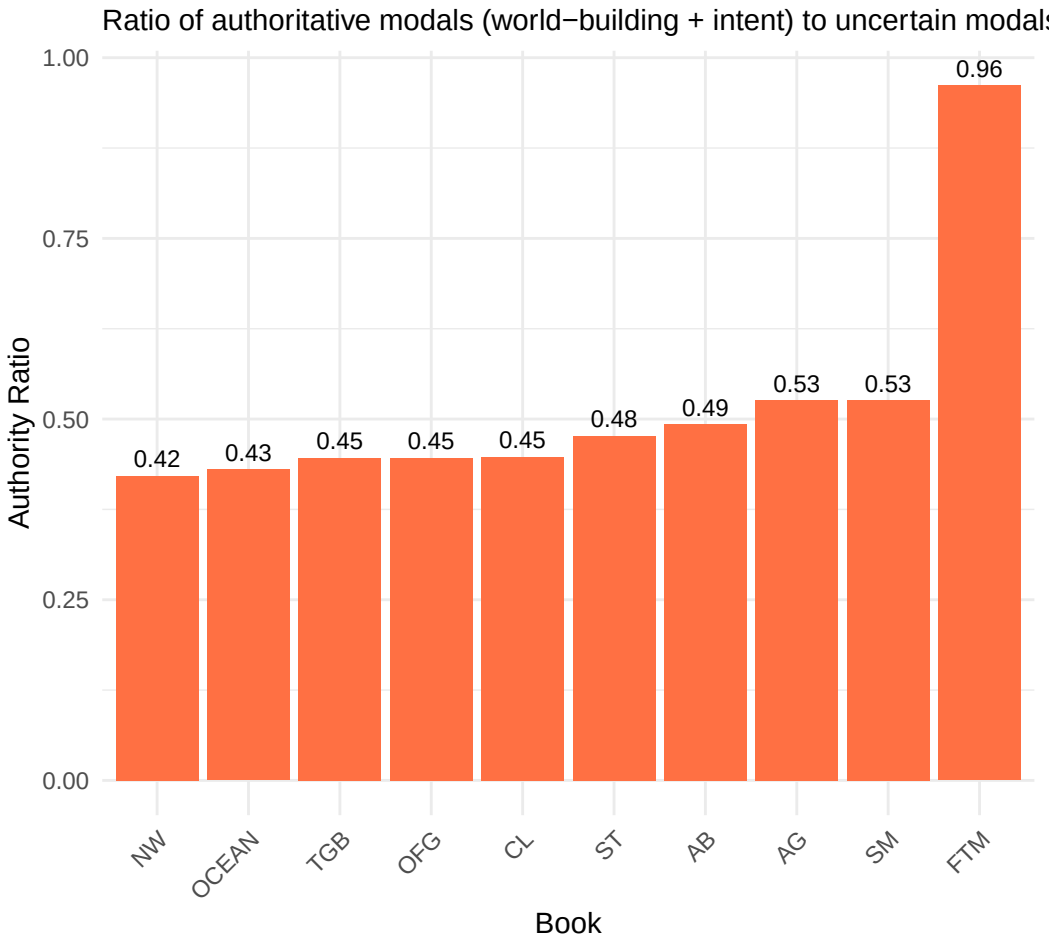

Figure 13: Narrative Voice Authority in Gaiman's Works

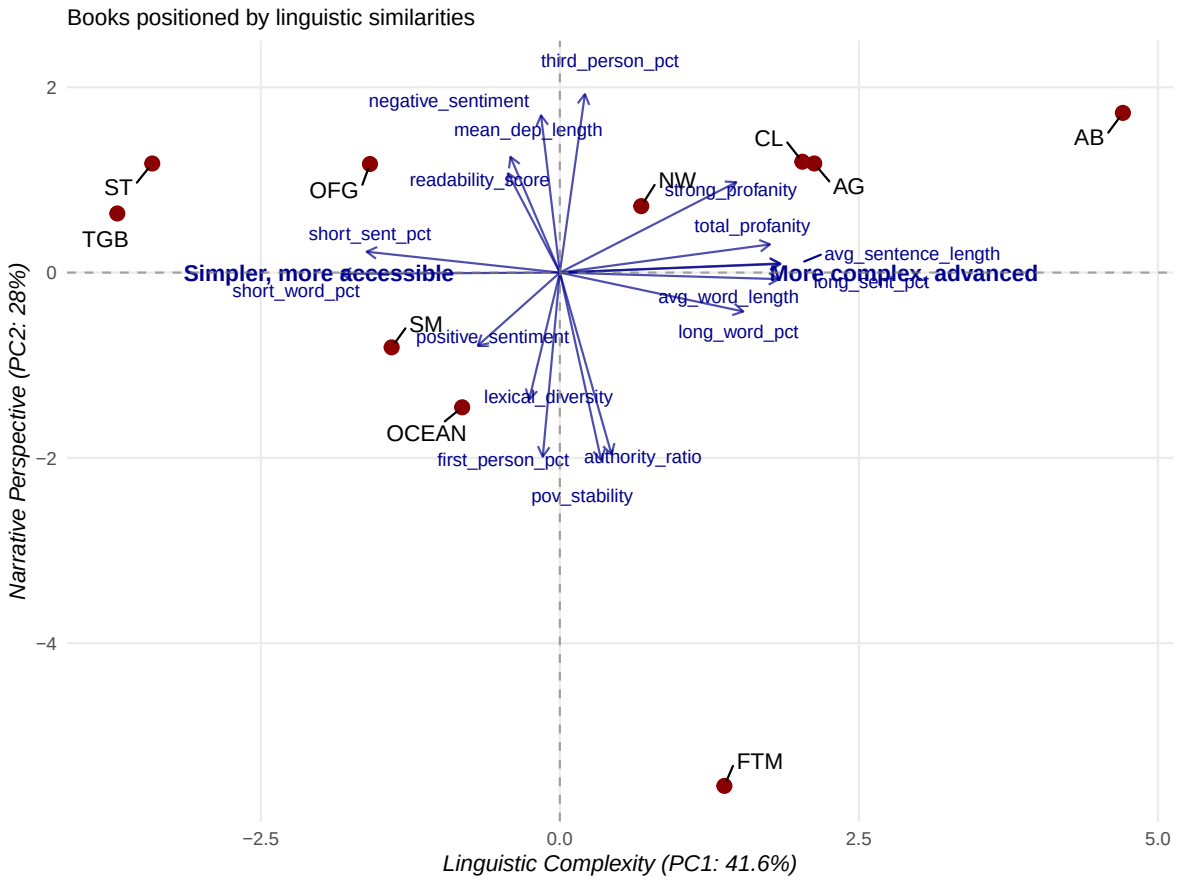

Figure 14: Linguistic Spectrum of Gaiman's Works

proportions of long words (0.307), and increased profanity usage (total: 0.351, strong: 0.295). Conversely, it correlates negatively with markers of simpler language, such as higher percentages of short words (-0.366) and shorter sentences (-0.323). These relationships suggest PC1 represents a spectrum from linguistically accessible text (negative values) to more complex expression (positive values).

The second principal component, explaining 28% of variance, captures variations in narrative perspective and approach. This dimension correlates positively with third-person narration (0.385), negative sentiment expression (0.339), lexical diversity (0.274), and syntactic complexity as measured by dependency length (0.250). In contrast, it shows negative associations with perspective stability (-0.405), first-person narration (-0.398), and higher authority ratios (-0.393). This pattern suggests PC2 distinguishes between two narrative approaches: at one end, stable first-person narration with higher narrative authority, and at the other, more varied third-person perspectives with greater lexical range and syntactic complexity.

4 Discussion

The computational analysis of Gaiman corpus across different audience demographics reveals patterns of stylistic variation that challenge conventional assumptions about children's and adult literature. Rather than showing a simple, formulaic progression from linguistic simplicity in children's books to complexity in adult-oriented works, Gaiman's texts exhibit multidimensional variation that appears more closely tied to narrative requirements than audience age, as shown in Figure 14.

For instance, *Fortunately, the Milk*, a children's picture book, contains relatively complex linguistic features within a highly stable narrative framework featuring consistent first-person narration and high authority markers. This combination may serve the book's fantastical multiverse-travel plot while maintaining accessibility for younger readers. Similarly, *The Ocean at the End of the Lane* uses simpler language to effectively convey its childlike perspective on adult themes, demonstrating how Gaiman strategically selects specific stylistic elements to serve narrative purposes.

These findings align with Gaiman's own statements about his writing process. In a 2015 interview Myatt (2015) , he explained that controversial or challenging content in his books "always arise naturally from the demands of the plot" rather than being included to shock readers of any age.

This multidimensional approach to stylistic analysis resolves apparent contradictions in individual metrics. When examined through single linguistic features, Gaiman's works often defied conventional categorisation by audience age. However, the integrated analysis revealed that Gaiman adapts specific stylistic dimensions based on narrative needs which in fact differentiates target audience groups. This

selective adaptation may explain why his works appeal across age boundaries while still maintaining distinct identities appropriate to their target demographics.

5 Conclusion

This study reveals that Neil Gaiman's stylistic variations are selective and purpose-driven rather than formulaic. At the lexical level, Gaiman maintains consistent vocabulary patterns regardless of audience age, while making more substantial adjustments to sentence complexity, profanity usage, and narrative perspective stability when writing for younger readers. The two-dimensional framework emerging from principal component analysis analysis illustrates how Gaiman balances creative and demographic considerations in his writing.

The computational methodology used in this research detected nuanced patterns across multiple works that traditional close reading might overlook, highlighting the complementary value of quantitative approaches to literary analysis. These insights could inform both the academic study of children's literature and practical approaches to writing across age boundaries.

6 Reference

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Biber, D. (1988). *Variation across speech and writing* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Dawson, N., Hsiao, Y., Tan, A., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*, *1.0*(1.0). https://doi.org/10.34842/5WE1-YK94

Fialho, O., & Zyngier, S. (2023). *Quantitative methodological approaches to stylistics* (2nd ed., pp. 349–366). Routledge. https://doi.org/10.4324/9780367568887-24

Gius, E., & Jacke, J. (2022). Are Computational Literary Studies Structuralist? *Journal of Cultural Analytics*, *7*(4). https://doi.org/10.22148/001c.46662

Haitao Liu. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191. https://doi.org/10.17791/JCS.2008.9.2.159

Hammond, A., Brooke, J., & Hirst, G. (2013). A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together. In D. Elson, A. Kazantseva, & S. Szpakowicz (Eds.), *Proceedings of the Workshop on Computational Linguistics for Literature* (pp. 1–8). Association for Computational Linguistics. https://aclanthology.org/W13-1401/

Hoover, D. L. (2017). The microanalysis of style variation. *Digital Scholarship in the Humanities*, *32*(suppl_2), ii17–ii30. https://doi.org/10.1093/llc/fqx022

Karlgren, J. (2010). *Textual Stylistic Variation: Choices, Genres and Individuals* (pp. 113–125). https://doi.org/10.1007/978-3-642-12337-5_6

Leonawicz, M. (2025). *Epubr: Read EPUB file metadata and text*. https://docs.ropensci.org/epubr/

Mullen, L. (n.d.). *tokenizers: Fast, Consistent Tokenization of Natural Language Text*. https://doi.org/10.32614/CRAN.package.tokenizers

Myatt, F. (2015). Neil Gaiman: 'My parents didn't have any kind of rules about what I couldn't read'. *The Guardian*. https://www.theguardian.com/childrens-books-site/2015/aug/29/neil-gaiman-banned-books-censorship-interview

Simpson, P. (2003). *Language, Ideology and Point of View* (0th ed.). Routledge. https://doi.org/10.4324/9780203136867

Straka, M., Hajič, J., & Straková, J. (2016). *LREC 2016* (N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis, Eds.; p. 42904297). European Language Resources Association (ELRA). https://aclanthology.org/L16-1680/

Tanprasert, T., & Kauchak, D. (2021). *Title = "flesch-kincaid is not a text simplification evaluation metric"* (A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, & W. Xu, Eds.; p. 114). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.1

Tyler W, R. (2021). *{Sentimentr}: Calculate text polarity sentiment*. https://github.com/trinker/sentimentr

7 Appendix

7.1 Profanity Word List

1. mild_profanity: "damn", "darn", "hell", "crap", "suck", "stupid", "idiot", "dumb", "moron", "fool", "jerk", "dork", "heck", "gosh", "jeez", "Christ", "God", "Jesus", "bloody", "blast", "drat", "freaking", "frigging", "poop", "butt", "arse", "bollocks", "turd", "bum"

2. moderate_profanity: “ass”, “asshole”, “bastard”, “shit”, “bullshit”, “piss”, “screw”, “screwed”, “dick”, “cock”, “prick”, “slut”, “whore”, “wanker”, “tosser”, “twat”, “crap”, “balls”, “nuts”, “pissed”, “fart”, “jackass”, “douche”, “douchebag”, “tits”, “boobs”
3. strong_profanity: “fuck”, “fucked”, “fucker”, “fucking”, “motherfucker”, “motherfucking”, “cunt”, “pussy”, “cock”, “bitch”, “cum”, “jizz”, “nigger”, “faggot”, “fag”, “spic”, “chink”, “kike”, “retard”, “whore”, “slut”

7.2 Modal Verbs

- world_building_pattern: must, should, have to, has to, had to, always, never
- possibility_pattern: may, might, could, can, maybe, perhaps
- hypothetical_pattern: would, if, whether
- intent_pattern : will, shall, going to, want to”

7.3 Justification for Reference

I used Zotero’s automatic reference insertion function, so it was not practical to provide a justification for each citation individually. Given that, I explain here the rationale behind the references used in this project. They can be grouped into three categories: (1) constructing the theoretical framework, (2) methodological support, and (3) references for R packages.

Karlgren (2010) serves as an anchor point, connecting the theoretical framework with the methodology. It not only provides a clear definition of *style* as used in this study, but also introduces the concept of *style variation* from the perspective of computational stylistics — which is central to this project. Biber (1988) is a foundational work in the field of stylistics and forms the broader theoretical basis for this research. My classification of modal verbs, on the other hand, draws on the framework by Simpson (2003) by simplifying his categorisation.

Dawson et al. (2021) is directly relevant to the objects of study in this project — ten books by Neil Gaiman, including children’s literature. Its findings, such as insights into syntactic complexity in children’s books, offer empirical grounding for this research.

Regarding methodology, I use Fialho & Zyngier (2023) and Hoover (2017) to address a long-standing debate in computational stylistics and literary studies, namely the concern explained by Hammond et

al. (2013) on whether computational methods can truly analyse and understand literary texts. I aim to provide a critical reflection through these sources and position this study in a stable and well-informed context. In doing so, I use Gius & Jacke (2022) to support that computational stylistic methods should be seen as meaningful contributions to literary research, rather than pseudoscience masked by statistical techniques.

During the analysis, Haitao Liu (2008) justifies the solution of using UD to measure syntactic complexity. Tanprasert & Kauchak (2021) provides a critical perspective indicating that Flesch-Kincaid Grade Level may not be very reliable.

Finally, I have cited the R packages used in this project — *sentimentr*, *udpipe*, and *quanteda* — according to academic conventions. In addition, I referred to an interview with Neil Gaiman to add the realistic factor of his diverse stylistic variations.

I believe the above explanation accounts for all the references used in this project and provides clear motivation for each.