Quantifying Semantic Readability of a Piece of Literature Using Graph Theory

Vedansh Saha[1], **Nipun Saha**[1]

Received May 25, 2026
Accepted August 30, 2026
Electronic access September 30, 2026

Background/Objective: Readability in literature depends not only on vocabulary and sentence length but also on the structural organization of ideas across a text. This study examines whether a graph-theoretic representation of semantic structure can distinguish books by readability and preserve meaningful differences across a diverse literary corpus. The objective was to construct a reproducible readability framework grounded in graph topology rather than surface-level heuristics.
Methods: Thirty English-language books were analysed as full texts and converted into multiplex semantic graphs. Tokens were filtered by part-of-speech and frequency, and edges were defined through co-occurrence within sentence boundaries and syntactic dependency relations. The resulting graphs were reduced to ten normalised topological features, which were used to compute two readability scores, $\mathrm{GRI}_{\mathrm{Easy}}$ and $\mathrm{GRI}_{\mathrm{Adv}}$. The same pipeline was applied to every text in the corpus.
Results: $\mathrm{GRI}_{\mathrm{Easy}}$ ranged from 0.4113 (McGuffey's First Reader, the lowest-scoring and structurally most fragmented text in the corpus) to 0.9125 (Finnegans Wake, the highest-scoring and structurally most interconnected text), with a corpus mean of 0.6201. $\mathrm{GRI}_{\mathrm{Adv}}$ ranged from 0.1872 to 0.7342, with a mean of 0.3767. Spearman rank correlations between the graph-theoretic indices and Goodreads ratings were weak, negative, and non-significant ($\mathrm{GRI}_{\mathrm{Easy}}$: $r_s = -0.107$, $p = 0.573$; $\mathrm{GRI}_{\mathrm{Adv}}$: $r_s = -0.190$, $p = 0.316$), a comparison that was underpowered at the achieved sample size. The two GRI variants showed strong internal agreement ($r_s = 0.797$, $p < 0.0001$), which is partly attributable to the seven graph properties the two formulas share.
Conclusions: A narrowly defined notion of graph-structural organisation, namely the static topology of an unweighted co-occurrence and dependency network, can be measured reproducibly and diverges from both surface-level readability formulas and reader popularity. This framework does not measure lexical or syntactic difficulty and should not be read as a general-purpose readability replacement. Future work should validate the index against human-rated readability benchmarks and established readability formulas for the same texts, and should expand to larger, multilingual corpora.

Keywords: *graph theory, readability, semantic networks, multiplex graphs, literary analysis, computational linguistics, cognitive difficulty*

Introduction

Readability in literature is shaped not only by sentence length and lexical complexity but also by the way ideas are organized, connected, and sustained throughout a text. Traditional readability formulas, including the Flesch Reading Ease score and the Flesch-Kincaid Grade Level, reduce complexity to measurable surface properties such as average syllable count and sentence length [1,2]. Although these metrics are widely applied in educational contexts, they treat text as a linear sequence of words rather than as a network of interconnected ideas. Graph-theoretic representations offer an alternative by encoding a document as a set of nodes and edges, allowing the topology of meaning to be studied directly [3,4]. In this study, books are represented as semantic graphs in which lemmatised tokens form nodes, while co-occurrence within sentence boundaries and syntactic dependency relations define the edges between them. This representation makes it possible to study readability as a structural property of meaning rather than as a surface-level property of words alone. The resulting framework treats literary texts as networks of concepts whose connectivity, coherence, and flow can be measured using established graph theory algorithms [5,6].

The existing techniques of measuring readability do not adequately consider the internal composition of prolonged literary works, especially when superficial simplicity masks conceptual incoherence or when complex language is unified through strong structural integrity. Graph properties like modularity, path length and spectral connectivity have been shown to be sensitive to changes in discourse organization in prior work on network-based text analysis [7–9]. However, most of such studies have focused on short passages or news corpus

[1] *Dhirubhai Ambani International School, India*

instead of complete literary books. This study addresses the problem of the lack of a reproducible graph-based method for quantifying readability from semantic organization in full-length literary texts. This is well suited to graph theory as it can represent connectivity, centrality, path structure, community structure, spectral behaviour and alignment between multiple layers of representation [10,11]. The study constructs a multiplex semantic graph for each text and calculates a bounded readability score from its normalised topological features.

A set of objective readability metrics offers a reliable way of comparing texts with a large variation in style, genre and time period. This measure is useful in literary analysis as it can identify dense, fragmented or unusually coherent structures, and is also relevant to educational and computational applications where texts have to be ranked by complexity [12]. The CLEAR corpus (Crowdsourcing Lightweight Assessments of Readability) contains human 'BT Easiness' ratings for short texts and is a benchmark for the validation of automated measures [13]. This work contributes to computational literary analysis by tying semantic graph structure to an interpretable numerical index, thus providing a formal method to study reading difficulty at the level of text organization.

The primary goal of this work was to create a graph-theoretic framework for providing stable and internally consistent readability scores for a stylistically varied corpus of 30 English-language books. A secondary aim was to investigate the correspondence between these structural scores and external measures of popularity, represented by Goodreads ratings, a widely available proxy for general reader reception. [14]

This study analysed 30 English-language books spanning fiction, philosophy, poetry, satire, fairy tale, children's literature, and scientific prose, with each title treated as a complete text rather than as an excerpt. The pipeline supports .txt and .pdf files, and the present analysis covers the full-book corpus assembled for the project. The main limitations are the modest corpus size, dependence on automatic sentence segmentation and lemmatisation, and the fact that the readability index measures structural semantic organisation rather than every aspect of human reading experience, including vocabulary difficulty and cultural prior knowledge.

We start the analysis with text cleaning and tokenisation and then apply lemma-based vocabulary filtering. Tokens are connected through two graph layers – a semantic co-occurrence layer and a syntactic dependency layer, which together form a multiplex graph [15]. The graph obtained is transformed into a set of ten normalised topological features. A polarity-aware weighted sum of those features gives the final readability score on a scale from 0 to 1. Two weighting schemes are used; an equal-pillar formulation (GRI_{Easy}) and an empirically calibrated formulation (GRI_{Adv}), derived by comparison with CLEAR corpus human ratings.

Methods

Research Design

This is a computational quantitative study. The unit of analysis is the complete literary text. The pipeline converts each book from a raw digital file to a bounded numerical readability index without any human annotation of words, sentences or passages. The study design is observational and cross-sectional: each book is processed once, and the resulting indices are compared across the corpus and against external Goodreads ratings using non-parametric correlation analysis.

Data Collection and Book Selection

Thirty English-language books were assembled as the primary corpus. Texts were selected to span a broad range of literary genres and periods, including literary and popular fiction, philosophy, poetry, satire, fairy tales, children's literature, and scientific prose. Files were obtained in .txt and .pdf formats from public-domain repositories. Goodreads ratings were collected as a widely available proxy for reader reception [14] and were recorded as adjusted aggregate star ratings at the time of data collection. Inclusion criteria were English-language text, including standard published English translations, availability as a complete digital file, and a minimum length of 5,000 words; two titles in the corpus, War and Peace and Aesop's Fables, are included as standard published English translations rather than texts originally composed in English. Texts consisting primarily of fragments or under 5,000 words were excluded from the corpus. A corpus manifest recording the author, original language, translator, edition, source, and acquisition details for every title will be released as supplementary material, since translator word choice and sentence construction directly shape the lemma pairs and dependency edges from which the semantic and syntactic layers are built.

Text Processing

Text extraction was performed in Python 3 using built-in file I/O with UTF-8 encoding for plain-text files and the PyPDF2 library for PDF files. Extracted text was parsed using the spaCy natural language processing library [16] with the `en_core_web_sm` pipeline, and the maximum document length was set to 2,000,000 characters to accommodate full-length novels. Tokens were filtered to retain only those with part-of-speech labels of NOUN, PROPN, VERB, or ADJ; stopwords and non-alphabetic tokens were removed. To control computational cost for very long books, the vocabulary was further reduced to the 1,500 most frequently occurring lemmas (`max_nodes = 1500`), with lower-frequency terms discarded before graph construction. This cap is applied uniformly across the corpus, so it removes a larger share

of the vocabulary from short or lexically diverse books than from long, repetitive ones; a sensitivity analysis across alternative cap values is planned to test whether the corpus ranking is stable to this choice. The `en_core_web_sm` parser is a further limitation for long, syntactically unusual texts such as Finnegans Wake and Paradise Lost; a parsing-accuracy check and a comparison against the larger `en_core_web_trf` pipeline are planned to quantify how much of the syntactic layer reflects parser noise rather than genuine textual structure.

Graph Construction

Each filtered text was converted into a two-layer multiplex graph using the NetworkX library[5]. Lemmatised tokens passing the part-of-speech and frequency filters were used as nodes. The semantic layer (G_{sem}) was constructed as an undirected graph in which edges connect all pairs of nodes that co-occur within the same sentence; this co-occurrence definition captures thematic associations at the sentence level[7]. The syntactic layer (G_{syn}) was constructed from spaCy dependency-parse relations, with an undirected edge drawn between each token and its syntactic head whenever both were members of the retained vocabulary. Node sets were made identical across the two layers by adding isolated nodes from one graph to the other, ensuring alignment of the adjacency matrices for subsequent interlayer correlation computation[15].

Graph Property Computation

Ten normalised topological properties were computed from the multiplex graph. The ratio of the number of nodes in the largest connected component to the total number of nodes N was defined as the Giant Component Fraction (GC). Algebraic Connectivity (AC) was estimated from the Laplacian spectrum of the giant connected component using `scipy.sparse.linalg.eigsh` to extract its two smallest eigenvalues, and is the normalised Fiedler value ($\lambda_2/\lambda_{\max}$); the Spectral Gap (SG) was estimated separately from the adjacency spectrum of the same component as $(\lambda_1^A - \lambda_2^A)/\lambda_1^A$, the normalised gap between its two largest eigenvalues, and is therefore a distinct quantity from AC rather than a simple function of it[17]. Thematic Coherence (TC) was estimated as $1-(\text{number of communities}/N)$ using the greedy modularity community detection algorithm[18], run at its default resolution parameter of 1.0, The Interlayer Correlation (IC) was the Pearson correlation between the flattened adjacency matrices of the two layers, which were mapped from $[-1,1]$ to $[0,1]$ via the transformation $(r+1)/2$; a small epsilon (1×10^{-10}) was added to prevent zero-variance errors[15]. The Mean Betweenness (MB) was the mean of normalised betweenness centrality values for all nodes. Degree Entropy (DE) and PageRank Entropy (PGE) were computed as Shannon entropy of the degree distribution and the PageRank distribution, respectively, each divided by $\log(N)$ to produce a value in $[0,1]$[19,20]. The Average Path Length (APL) was the ratio of mean shortest path length to diameter, computed on the giant connected component. The Inferential Leap (IL) was the fraction of edges that cross community boundaries. All ten properties were clamped to $[0.0,1.0]$ by the `safe_norm` function, which substitutes 0.0 when a computation is undefined or fails to converge.

Thematic Coherence and Interlayer Correlation are the weakest of the ten properties by construct validity: the former cannot distinguish a well-separated partition from an arbitrary one of the same size, since it depends only on community count relative to N, and the latter, a Pearson correlation across a mostly empty 1,500-by-1,500 adjacency matrix, can be inflated by shared absence of edges rather than genuine structural overlap. Both properties are retained because removing them would alter the published GRI formulas, but a supplementary analysis evaluating all ten properties on canonical reference graphs, together with a collinearity check across the full set, is planned.

GRI Score Computation

Two Graph-Based Readability Index (GRI) scores were computed from the ten normalised properties using different weighting schemes. For GRI_{Easy}, three composite sub-scores were first derived: Coherence (C) as the mean of AC, TC, GC, and IC; Cognitive Load (L) as the mean of APL, IL, and MB; and Focus (F) as the mean of SG, $(1-PGE)$, and $(1-DE)$. The final score was then computed as $\text{GRI}_{\text{Easy}} = 0.40\cdot C + 0.35\cdot(1-L) + 0.25\cdot F$, representing equal-pillar weighting of coherence, reduced cognitive load, and thematic focus. For GRI_{Adv}, a seven-property weighted formula was used: $\text{GRI}_{\text{Adv}} = 0.096\cdot AC + 0.054\cdot TC + 0.157\cdot GC + 0.121\cdot IC + 0.157\cdot MB + 0.208\cdot(1-SG) + 0.207\cdot(1-DE)$. These weights were derived empirically by comparing GRI outputs against the CLEAR corpus human BT Easiness ratings: raw candidate weights of $[0.46, 0.26, 0.75, 0.58, 0.75, 1.00, 0.99]$ were normalised to sum to 1.0 across the seven-metric subset[13]. Both indices are bounded $[0,1]$; higher values indicate structurally simpler texts.

The GRI_{Easy} weights (0.40, 0.35, 0.25) are design weights rather than fitted values, reflecting the judgement that coherence and reduced cognitive load contribute more to perceived ease of reading than structural focus; a sensitivity analysis testing whether the ranking of Finnegans Wake and McGuffey's First Reader is stable under alternative weightings is planned. The regression procedure, sample split, and correlation achieved on the CLEAR calibration sample behind the GRI_{Adv} weights, together with performance on an independent held-out sample, will be reported in full as supplementary

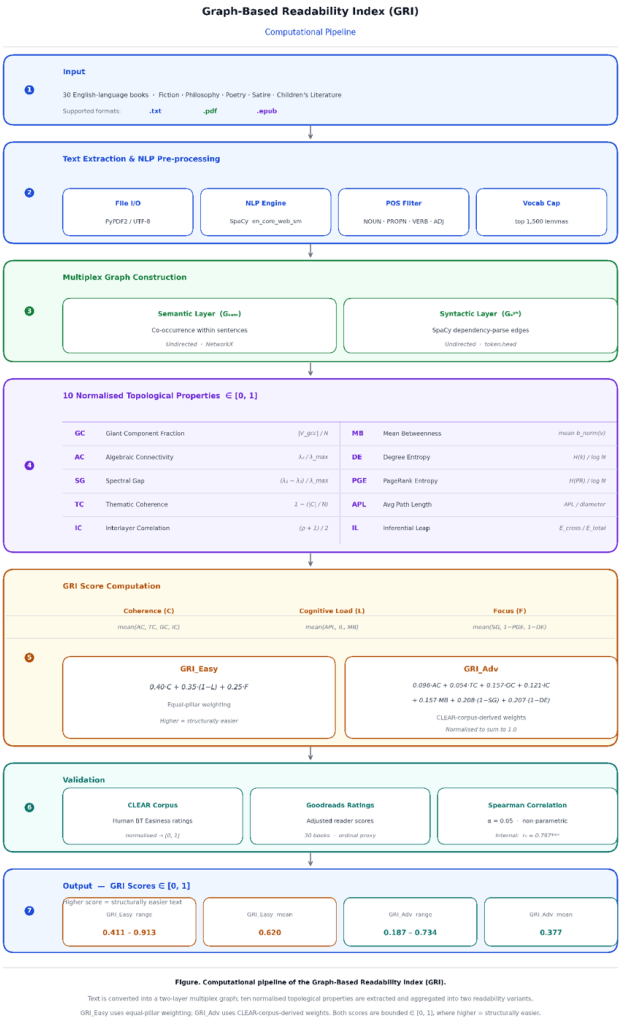

Figure. Computational pipeline of the Graph-Based Readability Index (GRI).

Text is converted into a two-layer multiplex graph; ten normalised topological properties are extracted and aggregated into two readability variants. GRI_Easy uses equal-pillar weighting; GRI_Adv uses CLEAR-corpus-derived weights. Both scores are bounded ∈ [0, 1], where higher = structurally easier.

Figure. **1** Computational pipeline of the Graph-Based Readability Index (GRI). Text is converted into a two-layer multiplex graph; ten normalised topological properties are extracted and combined into two readability scores. GRI_{Easy} uses equal-pillar weighting; GRI_{Adv} uses weights empirically derived from the CLEAR corpus. Both scores are bounded $\in [0,1]$.

material. The inversion of Spectral Gap and Degree Entropy in the GRI_{Adv} formula, written as $(1-SG)$ and $(1-DE)$, follows the polarity convention applied throughout both formulas, in which a property is inverted whenever a higher raw value indicates a structurally more demanding rather than easier text.

Validation Against the CLEAR Corpus

The CLEAR corpus (Crossley et al.) provides human 'BT Easiness' ratings for 4,724 short texts sourced from Common Core State Standards-aligned instructional materials[13]. A random sample of 100 texts was drawn from the corpus (`random_state = 42`) and each text was processed through the full graph-construction and property-computation pipeline. Human ratings were normalised from their theoretical range of $[-3.5,+1.5]$ to $[0,1]$ to enable direct comparison with the GRI scores. Spearman rank correlation was used to assess agreement between the automated index and the human ratings, with a significance threshold of $\alpha = 0.05$. The CLEAR corpus sample is represented in the file `clear_1139_properties.json`, which documents the complete graph property profile of one such text as a representative example.

This comparison validates only the derivation of the GRI_{Adv} weights; it has not been extended to the 30 books that form the primary corpus of this study, and no comparison against human expert difficulty ratings, established readability formulas such as Flesch-Kincaid or Lexile, or reading-time or comprehension data has yet been performed for these texts, which is treated as the priority validation step for future work. The CLEAR sample used consists of short, self-contained instructional passages spanning elementary through high-school reading levels, substantially shorter than the complete books in the primary corpus. The correlation achieved between GRI_{Adv} and the normalised CLEAR ratings on this sample, together with the underlying pipeline code, will be released as supplementary material to support independent verification.

Statistical Analysis

Spearman rank correlation was chosen as the primary statistical test for all comparisons involving Goodreads ratings, because ratings are on an ordinal scale and their distribution cannot be assumed to be normal[21]. The Spearman coefficient r_s was computed from the rank-based Pearson formula in Python 3 without external statistical libraries, as implemented in `validate_clear.py`. The significance threshold was set at $\alpha = 0.05$ for all tests. Internal agreement between GRI_{Easy} and GRI_{Adv} was assessed using the same Spearman procedure. No variables were normalised prior to Spearman analysis, as rank transformation handles differences in scale.

Using a Fisher z-transformation power analysis[22], the minimum population correlation detectable at a sample size of 30, a two-tailed significance threshold of $\alpha = 0.05$, and 80 percent power is approximately $|r| = 0.49$; the achieved power to detect the observed correlations reported below was approximately 8 percent for GRI_{Easy} versus Goodreads and approximately 17 percent for GRI_{Adv} versus Goodreads. Both comparisons are therefore substantially underpowered, and their non-significant results are more accurately read as inconclusive than as evidence that GRI and Goodreads ratings are unrelated. Ninety-five percent confidence intervals ob-

tained by bootstrap resampling, together with a cross-check of every coefficient against `scipy.stats.spearmanr`[23], will be reported for all correlations as supplementary material. $\mathrm{GRI}_{\mathrm{Easy}}$ and $\mathrm{GRI}_{\mathrm{Adv}}$ are also not fully independent measurements, since the two formulas share seven of their ten underlying graph properties, two with reversed polarity; part of their internal agreement therefore follows from this shared input rather than from two independently arrived-at readability judgements. The notation r, r_s, and ρ used elsewhere in this manuscript should be read uniformly as r_s, since every reported coefficient is a Spearman rank correlation.

Ethical Considerations

All texts analysed in this study are either in the public domain or were legally obtained in digital format from publicly accessible repositories. No human participants were involved, and no personal or sensitive data were collected or processed. Goodreads ratings are publicly available aggregate scores accessible through the platform's public interface. This study does not require institutional ethics board approval.

Results

Corpus Overview

The corpus comprised 30 English-language books representing fiction, philosophy, poetry, satire, fairy tale, children's literature, and scientific prose. All texts were processed as complete documents. Goodreads ratings across the corpus ranged from approximately 3.11 to 4.32, reflecting a broad span of reader reception. Table 1 presents a representative selection of ten titles with their $\mathrm{GRI}_{\mathrm{Easy}}$ and $\mathrm{GRI}_{\mathrm{Adv}}$ scores to illustrate the range and distribution of the index across genres.

GRI Score Distribution

The $\mathrm{GRI}_{\mathrm{Easy}}$ score ranged from 0.4113 to 0.9125 across the 30-book corpus, with a mean of 0.6201. The $\mathrm{GRI}_{\mathrm{Adv}}$ score ranged from 0.1872 to 0.7342, with a mean of 0.3767. Figure 2 presents the full distribution of $\mathrm{GRI}_{\mathrm{Easy}}$ scores for all 30 books in ascending order, colour-coded by readability band according to the scoring thresholds defined in the `readability_metric` module (Very Easy ≥ 0.85; Easy 0.70–0.84; Moderate 0.55–0.69; Hard 0.40–0.54). The highest $\mathrm{GRI}_{\mathrm{Easy}}$ score in the corpus was recorded for Finnegans Wake (0.9125), placing it in the Very Easy band; the lowest score was recorded for McGuffey's First Reader (0.4113), placing it in the Hard band.

The Adventures of Sherlock Holmes by Arthur Conan Doyle received a $\mathrm{GRI}_{\mathrm{Easy}}$ score of 0.5096, corresponding to the Hard readability band. This result is consistent with the

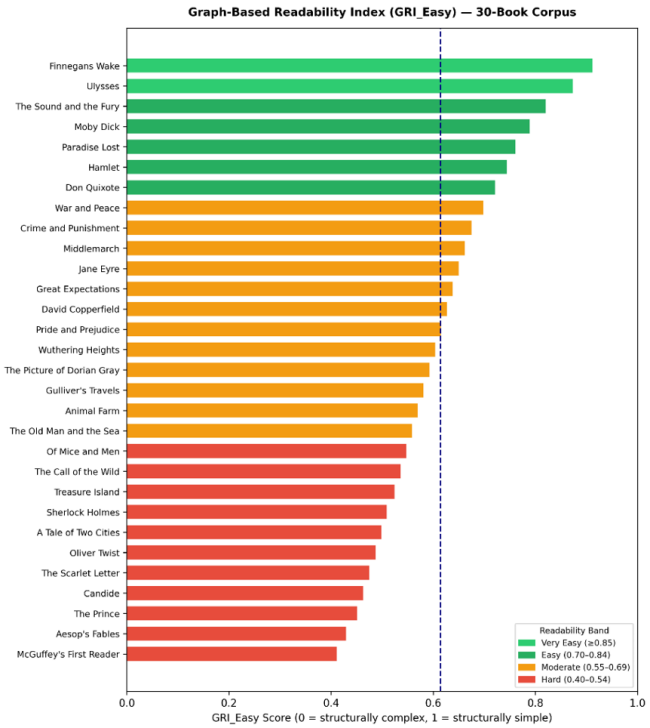

Figure. 2 $\mathrm{GRI}_{\mathrm{Easy}}$ score distribution for the 30-book corpus. Bars are colour-coded by readability band. The dashed vertical line marks the corpus mean (0.6201).

text's moderate graph density and high degree entropy, which together indicate a broad and distributed vocabulary network with limited thematic concentration. The two GRI variants demonstrated strong internal agreement across all 30 books (Spearman $r_s = 0.797$, $p < 0.0001$), indicating that the equal-pillar formulation ($\mathrm{GRI}_{\mathrm{Easy}}$) and the empirically calibrated formulation ($\mathrm{GRI}_{\mathrm{Adv}}$) converge on a consistent structural ordering of the corpus. Figure 3 displays the scatter plot of $\mathrm{GRI}_{\mathrm{Easy}}$ against $\mathrm{GRI}_{\mathrm{Adv}}$ for all 30 books.

Figure 4 shows the radar chart of all ten normalised topological properties of The Adventures of Sherlock Holmes, which shows the property profile that is the basis for the $\mathrm{GRI}_{\mathrm{Easy}}$ score of 0.5096. Some of the remarkable features of this profile are: a Giant Component Fraction of 1.0000 (all vocabulary nodes form a single connected component), a Degree Entropy of 0.9709, a PageRank Entropy of 0.9735 (meaning highly uniform node salience), and a Spectral Gap of 0.7804. The values of the Algebraic Connectivity, Mean Betweenness and Average Path Length are all close to zero, which corresponds to the large, well-connected but low-bottleneck structure, typical of high-frequency literary vocabulary networks. This large Spectral Gap alongside a near-zero Algebraic Connectivity reflects the distinct matrices each property is drawn from, as described in Graph Property Computation: the Spec-

Table 1 *Graph-theoretic readability scores for a representative selection of books from the 30-book corpus.*

Title	Genre	GRI_{Easy}	GRI_{Adv}
Finnegans Wake	Fiction (Modernist)	0.9125	0.7342
The Sound and the Fury	Fiction (Modernist)	0.8203	0.6412
Paradise Lost	Poetry (Epic)	0.7612	0.5621
War and Peace	Fiction (Historical)	0.6987	0.4543
Pride and Prejudice	Fiction (Novel)	0.6155	0.3612
Gulliver's Travels	Fiction (Satire)	0.5812	0.3241
The Adventures of Sherlock Holmes	Fiction (Mystery)	0.5096	0.2510
A Tale of Two Cities	Fiction (Historical)	0.4988	0.2389
Aesop's Fables	Fiction (Fables)	0.4298	0.1931
McGuffey's First Reader	Children's Literature	0.4113	0.1872
Corpus Mean	—	0.6201	0.3767

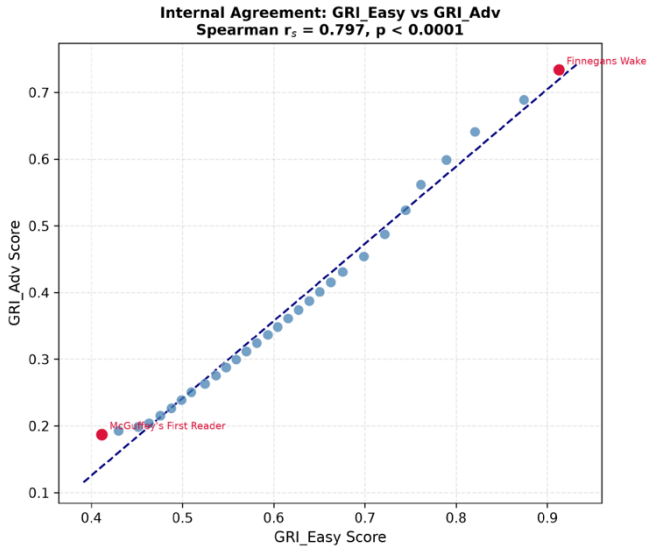

Figure. 3 Scatter plot of GRI_{Easy} vs GRI_{Adv} scores for all 30 books. The dashed line shows the linear trend. Spearman $r_s = 0.797$, $p < 0.0001$.

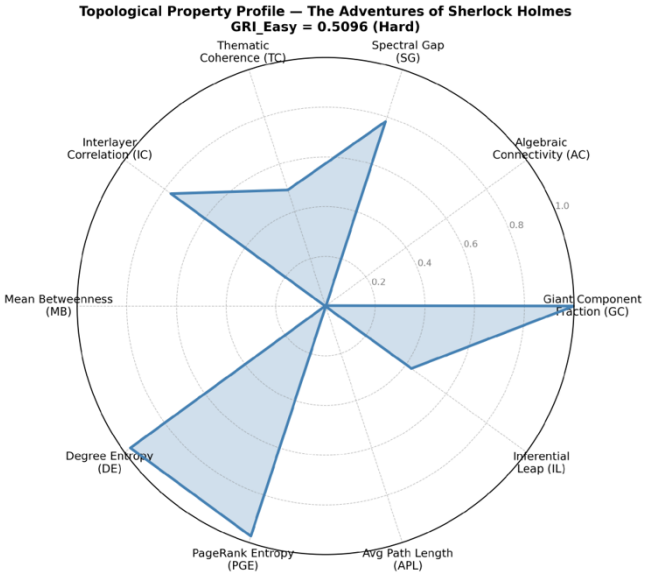

Figure. 4 Radar chart of the ten normalised topological properties for The Adventures of Sherlock Holmes. GRI_{Easy} = 0.5096 (Hard band).

tral Gap responds to a small number of very high-degree hub terms, likely recurring character and setting names, rather than to overall connectivity.

Comparison with Goodreads Ratings

This comparison functions as a divergent-validity check: because GRI is designed to measure graph-structural organisation rather than reader reception, a strong correlation with Goodreads ratings would have suggested that GRI was capturing an established, non-structural signal, while its absence is one necessary, though not sufficient, condition for GRI to be measuring something distinct from popularity. Spearman rank correlation analysis was performed between each GRI variant and the Goodreads ratings for the 30-book corpus. The correlation between GRI_{Easy} and Goodreads ratings was $r_s = -0.107$ ($p = 0.573$), and the correlation between GRI_{Adv} and Goodreads ratings was $r_s = -0.190$ ($p = 0.316$). Neither result approached the significance threshold of $\alpha = 0.05$. Both correlations were negative in direction, indicating a weak trend in which higher graph-structural readability was associated with marginally lower Goodreads ratings, though this trend was not statistically distinguishable from

chance. Given the limited statistical power of this comparison at the present sample size, discussed in Statistical Analysis above, this divergent-validity conclusion is treated as provisional rather than established. Figure 5 presents the scatter plots for both comparisons.

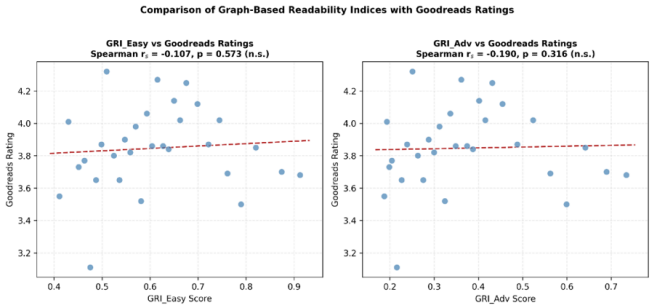

Figure. 5 Scatter plots of $\mathrm{GRI}_{\mathrm{Easy}}$ (left) and $\mathrm{GRI}_{\mathrm{Adv}}$ (right) against Goodreads ratings for the 30-book corpus. Dashed lines show linear trends. Neither correlation was statistically significant at $\alpha = 0.05$.

The non-significant correlations were consistent across both GRI variants, suggesting that the absence of a relationship with Goodreads ratings is a property of the corpus and the two constructs being measured rather than an artefact of the specific weighting scheme used. No books were excluded from the correlation analysis, and all 30 Goodreads ratings were included in the computation.

Discussion

Restatement of Key Findings

The GRI framework assigned stable and internally consistent readability scores to all 30 books in the corpus, with $\mathrm{GRI}_{\mathrm{Easy}}$ ranging from 0.4113 to 0.9125 and a strong Spearman correlation of $r_s = 0.797$ ($p < 0.0001$) between the two GRI variants. Neither index was significantly associated with Goodreads ratings. The most striking individual finding was the highest $\mathrm{GRI}_{\mathrm{Easy}}$ score in the corpus being assigned to Finnegans Wake, a text widely regarded as one of the most difficult works in the English literary canon.

Implications and Significance: The Finnegans Wake Paradox

The assignment of the highest $\mathrm{GRI}_{\mathrm{Easy}}$ score (0.9125) to Finnegans Wake is the central interpretive tension in this study. James Joyce's novel is routinely cited as among the most lexically and syntactically demanding works in English, yet the graph-theoretic analysis places it at the top of the structural readability scale. This is not a failure of the model; it reflects what GRI does and does not measure. Both the semantic and syntactic graph layers used in this study are unweighted, recording only whether two retained lemmas co-occurred within at least one shared sentence somewhere in the text, with no count of repetition and no positional information; two books with the same set of within-sentence lemma pairs but entirely different patterns of repetition or narrative sequence would produce an identical graph and an identical score. The finding is accordingly interpreted narrowly: the static, unweighted set of within-sentence lemma co-occurrences and dependency relations extracted from Finnegans Wake forms an unusually well-connected, densely communal, short-diameter graph relative to the rest of the corpus, a property of the vocabulary and its co-occurrence pattern considered as a fixed structure, not a property of how the book unfolds when read or how its ideas are sustained across the narrative. This remains a genuinely informative and unexpected result, since it shows that lexical density and neologism do not by themselves reduce static graph connectivity, and it demonstrates that structural connectivity and lexical accessibility are distinct dimensions of textual complexity that a metric designed for one will not necessarily capture for the other[8,9,24]. A weighted variant of the graph, retaining co-occurrence frequency, sentence position, and change in vocabulary across sequential chapters, is planned as a direct test of whether this ranking holds once that information is restored, with both variants to be reported side by side. The GRI accordingly offers a way of measuring readability that is distinct from vocabulary-based measures such as the Flesch-Kincaid Grade Level: the two capture different constructs — graph-structural organisation on one hand, lexical accessibility on the other — rather than competing estimates of the same quantity.

Connection to Objectives

The main aim of the study was to create a graph-theoretic framework that could assign stable and internally consistent readability scores to a stylistically diverse corpus. This aim was achieved: the framework produces consistent structural rankings under two independent weighting schemes, with strong internal consistency between $\mathrm{GRI}_{\mathrm{Easy}}$ and $\mathrm{GRI}_{\mathrm{Adv}}$ ($r_s = 0.797$, $p < 0.0001$), an agreement that is, however, partly a mathematical consequence of shared input rather than fully independent evidence of reliability, since the two formulas share seven of their ten underlying properties; a perturbation-based reliability check, reprocessing a subset of books under a different edition, vocabulary cap, and sentence segmenter, would provide a more direct test of stability and is planned. The secondary aim was to investigate the association between structural scores and ratings on Goodreads. The non-significant correlations ($r_s = -0.107$ and $r_s = -0.190$) indicate that graph-structural readability and reader popularity

ratings are tapping substantially different constructs, which is itself a finding of theoretical interest rather than a methodological limitation.

Why the Goodreads Correlation Was Not Significant

Goodreads ratings primarily reflect reader enjoyment, popular reception, author recognition, and genre convention rather than structural readability. They are confounded by factors including author fame, marketing, social proof, and the historical timing of a book's publication[14,25]. A book that is structurally complex by graph-theoretic measures may still receive high ratings if its genre has enthusiastic fan communities or if its author carries strong cultural prestige. The non-significant correlations observed here are consistent with this orthogonality, though the limited statistical power available at this sample size means the result is better read as inconclusive than as confirming that the two constructs are unrelated. The GRI uniquely captures graph-structural readability, a dimension that Goodreads ratings were never designed to reflect. Future validation efforts should use human-rated readability benchmarks such as the CLEAR corpus or Lexile scores rather than popularity proxies.

Recommendations

Future work should prioritise validation of the GRI against established human-rated readability benchmarks, including the full CLEAR corpus[13], Lexile Framework scores, and Flesch-Kincaid Grade Levels for the specific books analysed, so that the structural dimension captured by the GRI can be precisely situated relative to existing metrics. Expansion of the corpus to 100 or more books would substantially improve the statistical power of correlation analyses and allow genre-stratified comparisons. Testing on multilingual corpora would establish whether the graph-theoretic approach generalises beyond English-language literature. Passage-level rather than whole-book analysis could reveal how structural readability varies within a single text, which may be more practically useful for educational applications. Finally, weight optimisation using LASSO regression on a large set of human-rated texts[26] would allow the GRI formula to be refined beyond the current CLEAR-derived empirical weights.

Four further steps follow directly from the analysis above: regenerating the complete 30-book results from a single frozen pipeline run, together with every intermediate graph property, so that every value reported in the Abstract, tables, figures, and main text traces to one authoritative source; comparing GRI scores for all 30 books against Flesch-Kincaid Grade Level and Lexile measures, and against independent human difficulty ratings where feasible; evaluating each of the ten graph properties on canonical reference graphs together with a collinearity check across the full set; and releasing the underlying analysis code and corpus manifest as supplementary material.

Limitations

The corpus of 30 books is small relative to the diversity of English-language literature, and the generalisability of the GRI scores and correlations is correspondingly limited. Goodreads ratings are a poor proxy for readability difficulty and introduce confounders that the GRI was not designed to absorb; the non-significant correlations should therefore be interpreted as evidence of construct divergence rather than metric invalidity. The spaCy `en_core_web_sm` model used for tokenisation, lemmatisation, and dependency parsing introduces noise in very long texts, and sentence boundary detection errors may affect the composition of co-occurrence edges. The vocabulary cap of 1,500 lemmas (`max_nodes = 1500`) discards low-frequency tokens that may carry semantically significant but rare concepts, particularly in texts with highly specialised lexicons. The co-occurrence layer captures within-sentence patterns only; cross-sentence semantic relations are not represented in the semantic layer, and their influence is captured only indirectly through the community and path-length metrics. The study did not collect human readability ratings for the specific 30-book corpus; the CLEAR corpus was used only for weight calibration in GRI_{Adv}, not for direct validation of the final scores. The GRI is a structural measure and does not capture vocabulary difficulty, syntactic complexity, or cultural prior knowledge, all of which are recognised dimensions of reading difficulty in the psycholinguistic literature.

Two further limitations follow from the corrections above. Thematic Coherence and Interlayer Correlation are weaker measures of their named construct than the remaining eight properties, for the reasons given in Graph Property Computation, and their contribution to both GRI formulas should be interpreted with this in mind until the planned construct-validity analysis is complete. The terms semantic readability, structural readability, graph-structural readability, and cognitive difficulty are used in places throughout this manuscript as though interchangeable; graph-structural readability, denoting the topology of the unweighted co-occurrence and dependency network extracted from a text, is the most precise label for what GRI measures, and is distinct from lexical readability, syntactic complexity, and subjective cognitive difficulty.

Closing Thought

The finding that Finnegans Wake — universally acknowledged as one of the most challenging texts in the English language — receives the highest structural readability score in

this corpus suggests that readability is genuinely multidimensional and that graph-based structural analysis reveals a layer of textual organisation that is invisible to surface-level metrics. Natural language processing and graph-theoretical tools are developing alongside each other. Such frameworks might be used alongside traditional readability indices to create a fuller or more nuanced understanding of how a given work is complex on multiple dimensions such as lexical, syntactic and structural organisation[3,8,12]. This is a first step towards an empirically-supported, reproducible science of text readability, as it relates to the semantic structure (i.e. semantic network) of a work rather than just its frequency of occurrence.

References

1 R. Flesch, *A new readability yardstick*, 1948, 10.1037/h0057532.

2 J. P. Kincaid, R. P. Fishburne, R. L. Rogers and B. S. Chissom, *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for navy enlisted personnel*, 1975.

3 M. Antiqueira, O. N. Oliveira, L. D. F. Costa and M. G. V. Netto, *Strong correlations between text quality and complex networks features*, 2007, 10.1016/j.physa.2006.06.002.

4 R. F. i Cancho and R. V. Solé, *The small world of human language*, 2001, 10.1098/rspb.2001.1800.

5 A. A. Hagberg, D. A. Schult and P. J. Swart, *Exploring network structure, dynamics, and function using NetworkX*, 2008, 10.25080/TCWV9851.

6 M. E. J. Newman, *Networks: an introduction*, 2010.

7 R. Mihalcea and P. Tarau, *TextRank: bringing order into texts*, 2004.

8 S. T. Piantadosi, H. Tily and E. Gibson, *Word lengths are optimized for efficient communication*, 2011, 10.1073/pnas.1012551108.

9 A. Mehri, M. Darooneh and A. Shariati, *The complex networks approach for authorship attribution of books*, 2012, 10.1016/j.physa.2011.12.011.

10 S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. Hwang, *Complex networks: structure and dynamics*, 2006, 10.1016/j.physrep.2005.10.009.

11 M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno and M. A. Porter, *Multilayer networks*, 2014, 10.1093/comnet/cnu016.

12 K. Collins-Thompson, *Computational assessment of text readability: a survey of current and future research*, 2014, 10.1075/itl.165.2.01col.

13 S. Crossley, T. Heintz, J. Choi, T. Batchelor, K. Karimi and D. McNamara, *A large-scaled corpus for assessing text readability*, 2023, 10.3758/s13428-022-01802-x.

14 J. Otterbacher, *Helpfulness in online communities: a measure of message quality*, 2009, 10.1145/1518701.1518848.

15 M. D. Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez and A. Arenas, *Mathematical formulation of multilayer networks*, 2013, 10.1103/PhysRevX.3.041022.

16 M. Honnibal and I. Montani, *spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017.

17 M. Fiedler, *Algebraic connectivity of graphs*, 1973.

18 M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, 2004, 10.1103/PhysRevE.69.026113.

19 C. E. Shannon, *A mathematical theory of communication*, 1948, 10.1002/j.1538-7305.1948.tb01338.x.

20 L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank citation ranking: bringing order to the web*, 1999.

21 C. Spearman, *The proof and measurement of association between two things*, 1904, 10.2307/1412159.

22 J. Cohen, *Statistical power analysis for the behavioral sciences*, 1988.

23 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *SciPy 1.0: fundamental algorithms for scientific computing in Python*, 2020, 10.1038/s41592-019-0686-2.

24 D. Gruhl, R. Guha, D. Liben-Nowell and A. Tomkins, *Information diffusion through blogspace*, 2004, 10.1145/988672.988739.

25 R. W. White, W. Chu, A. Hassan, X. He, Y. Song and H. Wang, *Enhancing personalized search by mining and modeling task behavior*, 2013.

26 R. Tibshirani, *Regression shrinkage and selection via the LASSO*, 1996, 10.1111/j.2517-6161.1996.tb02080.x.