Feature-Level Decomposition of Text Complexity: Cross-Domain Empirical Evidence

Alexandr Parahonco, Liudmila Parahonco

Abstract

This study presents a feature-level analysis of text complexity using large language models (LLMs) in a two-phase design. Phase I operationalized six core features – lexical diversity, density, syntactic complexity, coherence, named entities, and readability – achieving Spearman correlations of 0.55–0.60 across domains. Phase II employed indirect prompting to surface additional qualitative dimensions (e.g., inferential load, rhetorical structure), yielding a mean correlation of 0.42 and revealing that the six features account for 40% of complexity variance. Domain dependencies were limited to named entities and lexical diversity. We propose a hybrid model combining normalization, root-based synergies, and newly quantified metrics with domain-tuned formulae for improved prediction.

Keywords: Text Complexity, Large Language Models, Feature Decomposition, Spearman Correlation, Prompting Strategy, Domain Dependency.

MSC 2020: 68T50, 91F20, 68T07.

1 Introduction

Text complexity is a fundamental factor that influences readability, comprehension, and effective communication in a wide range of fields, including education, linguistics, and natural language processing. A precise understanding of the multifaceted nature of text complexity is crucial for developing robust assessment tools that enable the creation of customized educational materials, reliable automated readability metrics, and advanced language models. Traditionally, text

doi:10.56415/csjm.v33.13

complexity has been treated as a holistic construct, often quantified by composite indices that aggregate lexical, syntactic, and discourse-level features [1]. However, such aggregation may obscure the distinct contributions and interactions of individual linguistic components that collectively shape complexity.

Our primary goal is to develop an intelligent content generation and search system tailored to university needs. The foundational concept was introduced in "E-course: Developing a Model for Content Generation" [2], and further expanded in "Elearning Content Processing Situations and Their Solutions" [3], which detailed the extraction processes for images, text, and video. As the project evolved, we noted the absence of a recommendation mechanism for ranking sources by quality.

We also addressed challenges in content presentation. Rather than merely reproducing original sources, our research turned to summarization techniques to enhance delivery. In [4], we evaluated extractive and abstractive approaches, selecting extractive summarization as the most effective. Graph-based methods, the Edmundson heuristic, and the TextRank algorithm emerged as top performers. This informed the development of our recommendation system (RS), presented in "The Design of a Recommendation System for Generating Content: Context of the Sources. Part I" [5].

The present study advances previous methodological frameworks by emphasizing the role of text complexity in enhancing content ranking within the RS. The module, Context of the Sources, classifies texts into academic, non-academic, and security domains – a crucial step for universities, which prioritize credible sources, though alternative content from blogs or forums may occasionally prove more informative. By integrating complexity metrics, our system aims to better match content to individual users' comprehension and preferences, thereby improving the overall impact of personalized content generation.

The dual **objectives** of this paper are: **(1)** to systematically analyze and classify the complexity features of text according to their relative importance, and **(2)** to examine the consistency of these dominant features in various textual domains. By addressing these goals, the research seeks to answer a critical question: Is text complexity a

stable construct with invariant defining features, or does it vary significantly depending on the domain and context?

2 Related Work

Recent studies emphasize the multidimensional nature of text complexity. Biber, Larsson, and Hancock argue that traditional "one-dimensional" metrics (e.g., sentence length, clause counts) are insufficient [6]. Empirical evidence shows that complexity spans grammar, lexis, and discourse structure, requiring both quantitative (e.g., frequency, length) and qualitative (e.g., meaning levels, organization, context) features to accurately predict reading outcomes [7].

Text Complexity refers to the inherent, objective features of a text. These features include vocabulary sophistication, sentence structure, grammatical patterns, discourse organization, and text structure. In essence, text complexity measures the intrinsic intricacy of the written language, independent of any reader's abilities or background.

A range of linguistic and cognitive features contribute to text complexity:

1. Lexical Complexity involves word frequency, abstractness, and variety; rare or abstract terms increase processing effort [8].

2. Syntactic Complexity includes sentence length, clause embedding, and variation – heightening difficulty as structural density grows [9].

3. Discourse and Cohesion reflect coherence across sentences and paragraphs; texts with nonlinear narratives or implicit meanings require active integration [10].

4. Qualitative Dimensions, such as inference demands, organizational depth, and figurative language, affect comprehension beyond what quantitative measures capture [11].

Text Difficulty, on the other hand, is a more subjective concept. It describes how challenging a text is for a particular reader or group of readers and depends on factors beyond the text itself, such as the

reader's prior knowledge, cognitive skills, motivation, and context. A text might be complex in its structure and vocabulary yet not be difficult for an experienced reader, while the same text could pose significant challenges for someone with less background in the subject [1]. Finally, while text complexity is about the static, measurable characteristics of the text, text difficulty reflects the dynamic interplay between these characteristics and the reader's individual context [12].

Recent progress in text complexity analysis highlights persistent gaps. Existing tools overemphasize formulaic readability scores (e.g., Flesch-Kincaid [13], SMOG [14]), while underrepresenting qualitative aspects like genre conventions and disciplinary demands [15], leading to mismatches—especially for multilingual learners and specialized texts [16]. In academic settings, systems rarely translate explainability gains (e.g., +3% precision) into meaningful insight [17], often reducing complexity to generic scores. Our research addresses these shortcomings by leveraging NLP and large language models to uncover the features that define complexity in academic discourse.

3 Feature-Level Decomposition of Text Complexity

Our research identifies six key features influencing text complexity: lexical diversity, lexical density, syntactic complexity, text coherence, named entities, and readability. These align with four overarching categories – lexical, syntactic, discourse, and semantic – enabling multidimensional analysis of textual structure and meaning.

- **Lexical Features**:
  - Lexical diversity (LD) reflects vocabulary range via indices like TTR, D, or MTLD.
  - Lexical density (LDen) measures the ratio of content words, indicating informational load and inferencing demand.
- **Syntactic Features**:

  - Syntactic complexity involves sentence length, clause embedding, and variation. Elevated syntactic sophistication correlates with higher cognitive load and overall complexity.

- **Discourse Features**:
  - Text coherence reflects logical and semantic connection across sentences and paragraphs. Coherence – local and global – enhances comprehension and reduces processing effort [18].

- **Semantic Features**:
  - Named entities (NE) add semantic richness by introducing domain-specific terms and requiring background knowledge, with growing emphasis in academic literature (2015–2024) [19].

- **Readability**:
  - A cross-cutting feature influenced by lexical, syntactic, semantic, and discourse aspects. Defined as an objective, text-based measure of comprehension ease, it aligns with metrics like Flesch-Kincaid and Dale-Chall [20].

4 Experimental Analysis of Text Complexity

4.1 Methodology

Research Objectives and Hypotheses

This study aims to (1) rank text complexity features by importance and (2) assess their consistency across domains. We hypothesize that, while all features contribute, semantic and discourse elements impose a higher cognitive load due to their role in meaning construction and organization.

Data Selection and Preparation

We compiled a corpus of 100 documents across five domains – Quantum Computing, Cloud Computing, Dark Web, Dark Romanticism,

and Semantic Web – to examine how subject matter shapes text complexity. Alongside academic articles, we included encyclopedias, web pages, and forum posts to capture broader complexity patterns. Preprocessing steps (tokenization, part-of-speech tagging, normalization) were applied using NLTK [21], spaCy [22], and Pylexitext [23].

Given the novelty of the research, we explored text complexity through a **mathematical lens**, testing several prompts and ultimately adopting **"Prompt_1_Formula_Generation"** [24], based on the **RCTC framework** (Role, Context, Task, Constraints): "*Role: You are an expert NLP researcher specializing in text-complexity modeling. Context: I am designing an empirical study on text complexity. My corpus is mixed-domain, and I have already extracted six feature families:*

- *Lexical diversity (e.g., type–token ratio variants)*
- *Lexical density (content-word ratio, information density)*
- *Syntactic complexity (e.g., mean clause length, subordination index)*
- *Text coherence (discourse-connective density, entity-grid scores)*
- *Named-entity load (NER counts, % of tokens that are NE)*
- *Readability metrics (Flesch, SMOG, etc.)*

Task 1 – Feature grouping*. Logically cluster these six families into higher-level dimensions (max 3 groups) and justify each cluster in 1-2 sentences.* ***Task 2 – Formula design****. 1. Propose 10 distinct, mathematically explicit formulas for a composite Text-Complexity Score (TCS). 2. For each formula, list: (a) the normalized feature terms it uses; (b) the weighting scheme (constant weights, learned weights, log-scaling, etc.); (c) a one-line rationale (e.g., "emphasises syntactic difficulty for academic prose")...".*

Using this prompt, we employed diverse AI models to generate 161 formulas centered on six key features (see file **"161_formulas"** in [24]). LLMs were selected prior to April based on *accuracy*, *accessibility* (open source or subscription type), and *computational efficiency* (compatible

with 32 GB RAM). This enabled a robust assessment of feature interactions shaping text complexity (see Table 1).

Table 1. AI models contributing to formula generation

Model name	Nr. of parameters	Notes
ChatGPT o1	Not disclosed	paid subscription
ChatGPT o3-mini	Not disclosed	paid subscription
ChatGPT o3-mini-high	Not disclosed	paid subscription
Copilot Pro	Not disclosed	paid subscription
GROK v.2	Not disclosed	paid subscription
Claude Haiku	$\approx$ 20 billion parameters	paid subscription
Llama 3.1	405 billion parameters	paid subscription
Gemini 1.5 Pro	Not disclosed	free subscription
Gemini 2.0 Flash	Not disclosed	free subscription
Mistral Large 2	123 billion parameters	free subscription
DeepSeek v3	671 billion total parameters	free subscription
DeepSeek R1	671 billion total parameters	free subscription
Aria (Opera)	Not disclosed	browser free AI assistant
Qwen 2.5 MAX	Not disclosed	free subscription
Qwen 2.5 Instruct	14 billion parameters	open source LLM
PHI 3.1-MINI	3.8 billion parameters	open source LLM

For the expert role in our experiment, we selected only models equipped with reasoning capabilities, meaning they can perform complex thought processes (see Table 2).

Table 2. AI models acting as experts

Model name	Notes
ChatGPT 4o	multimodal model that incorporates reasoning elements
ChatGPT o3	most advanced reasoning model of OpenAI
ChatGPT o4-mini	uses normal inference effort
ChatGPT o4-mini-high	uses increased inference effort
Mistral Large 2	optimized for high-complexity reasoning tasks
DeepSeek v3	emphasizes efficient reasoning through selective parameter activation
DeepSeek R1	is a reinforcement learning-optimized variant of v3
Qwen 2.5 MAX	versatile reasoning (not reasoning mode)
Qwen 2.5 MAX Reasoning	versatile reasoning (with reasoning mode)

Experimental Design and Grouping Strategy

Recent advances in prompt engineering emphasize two core strategies for interacting with large language models (LLMs): **direct (explicit) prompting**, which provides explicit instructions for precision-driven tasks, and **indirect (implicit) prompting**, which relies on minimal guidance and the model's inferential capabilities. While direct methods excel in structured, technical domains, indirect ones often perform better in creative or cross-disciplinary contexts. Effective prompt design requires balancing both approaches to align with task goals and model strengths.

In our study, we were initially uncertain which approach would be most effective. To address this, we designed an experiment with two parallel phases: **Phase 1** — LLMs operating under explicit prompting (see file **"Prompt_2_Direct_Approach"** in [24]): "*I want you to act as a text complexity analyzer and ranker. Your task is to analyze a set*

of text files and rank them based on their text complexity. Text complexity should be evaluated as a generalization of the following features: 1. Lexical Diversity: Measure the variety of words used in the text. (Consider using metrics like Type-Token Ratio or more sophisticated measures like Moving-Average Type-Token Ratio – MATTR). 2. ...", and **Phase 2** – LLMs engaged through implicit prompting (see file **"Prompt_3_Indirect_Prompting"** in [24]): "*You are given a set of texts intended for use as sources for university lectures, targeting an audience of students and professors. Your task is to rank these texts from most complex to least complex based on the following features of text complexity: 1. Vocabulary Sophistication and Diversity: Presence of rare or specialized terms. 2. ...*". Given the methodological contrast, *the results generated by LLMs through implicit prompting could potentially validate those obtained from explicit prompting or uncover distinct insights not accessible through direct instruction.*

All phases followed the same processing pipeline and shared a common objective: to estimate text complexity for each document within a thematic group (Quantum Computing, Cloud Computing, Dark Web, Dark Romanticism, and Semantic Web) and to rank the files in descending order of complexity. The only distinction lies in the prompting strategy applied to the LLMs – implicit versus explicit.

Procedure and Implementation

This study adopts a structured, two-tiered methodology that combines formula-based analysis with expert evaluations to assess text complexity across diverse domains.

1. Formula Generation. An ensemble of advanced AI models generated 161 unique analytical formulas (see file "161_formulas" in [24]), each incorporating six key text complexity features. This ensured broad analytical diversity and multiple perspectives on feature interaction.

2. Automated Document Ranking. Within each thematic domain (e.g., Cloud Computing, Dark Web), every document was evaluated using a distinct formula. This yielded 161 ranked lists of file names, ordered in descending complexity according to the respective formula.

3. Expert-Based Document Ranking. A set of high-

performing LLMs (see Table 2) served as expert agents. Each model analyzed documents within a specific thematic group and independently produced a complexity-based ranking. These domain-sensitive evaluations were treated as benchmark standards for comparison.

4. Aggregation via External Merge Sort. To consolidate multiple expert rankings into a single consensus sequence, we employed the External Merge Sort algorithm [25]. Each document, identified by a filename prefixed with a numeric ID (e.g., "1-article.txt"), appeared in ordered lists from different expert models. External Merge Sort efficiently merged these sorted runs – even when they exceeded memory capacity – into one unified ranking. This approach minimized disk I/O, ensured scalability, and enhanced robustness by capturing the collective expert consensus while filtering individual variation.

5. Comparative Evaluation. The aggregated expert sequence was then compared against each of the 161 formula-generated rankings. Spearman's rank correlation coefficient was used to quantify the degree of alignment between expert judgments and automated outputs, measuring how closely the orderings matched [26].

6. Identifying Feature Importance. Correlation scores were analyzed to identify which formulas—and by extension, which text features—most closely aligned with expert assessments. This enabled a data-driven evaluation of the relative importance of each feature in predicting text complexity.

By following this systematic framework, the study ensured a rigorous, scalable, and interpretable evaluation of how different textual features contribute to perceived complexity across thematic domains.

4.2 Results and Discussion

The analysis of the obtained results was conducted across three dimensions: by domain, by feature, and by formula.

To begin the analysis, we selected the top 10 formulas from each thematic domain and compiled a table summarizing the coefficients assigned to the six core text features. As an illustrative example, Table 3 displays the top three formulas from Phase I.

The 161 formulas differ not only in their coefficient values but also

Table 3. Summary table for phase 1. Top 3 formulas

LexDiv	LexDen	SynCplx	Coher	NE	Read	Dom	Formula
1	1	1	1	1	0	Cloud comp	34
1	0	1	0	1	0	Cloud comp	27
1	1	0	1	1	0	Cloud comp	54
1	1	0	1	1	1	Dark rom	7
0.15	0.15	0.15	0.25	0.1	0.2	Dark rom	15
0	0	0.4	0.4	0.2	0	Dark rom	48
0.25	0.2	0.25	0.2	0.1	0	Dark web	104
1	1	1	1	1	1	Dark web	40
1	1	1	1	1	1	Dark web	69
1	1	1	1	1	1	Quant comp	8
1	1	1	1	0.1	0.1	Quant comp	74
1	0	1	0	0	1	Quant comp	28
0.2	0.2	0.1	0.15	0.15	0.1	Sem web	152
0.2	0.2	0.2	0.2	0	0.2	Sem web	17
1	1	1	1	1	1	Sem web	47

in their mathematical structure and the presence or absence of specific features. Some formulas assign equal weights (e.g., coefficient = 1) to all features, suggesting equal influence on text complexity. To enable consistent comparison, we standardized the data by assigning a value of 1 to each feature included in a formula, regardless of its actual coefficient. This binary representation allowed for uniform quantification of feature presence, and column-wise aggregation supported a structured analysis of feature importance across domains.

Analysis by domain

The corpus is organized into two principal thematic categories: Computing & Web Technologies and Literary & Philosophical Movements. The former includes texts on Quantum Computing, Cloud Computing, the Dark Web, and the Semantic Web, while the latter is represented primarily by works associated with the Dark Romanticism movement. Phase I results, shown in Fig. 1 and Fig. 2, illustrate the relative contributions of six key text complexity features across these thematic domains.

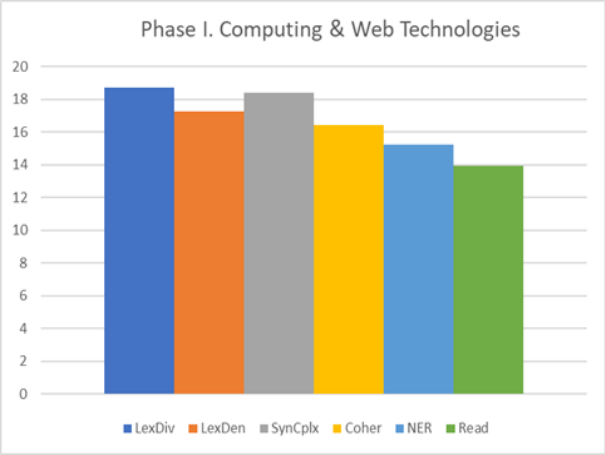

Figure 1. Phase I. Computing & Web Technologies

Figure 1 shows that within the Computing & Web Technologies domain, Lexical Diversity (19%) and Syntactic Complexity (18%) are the most influential features, reflecting the importance of vocabulary

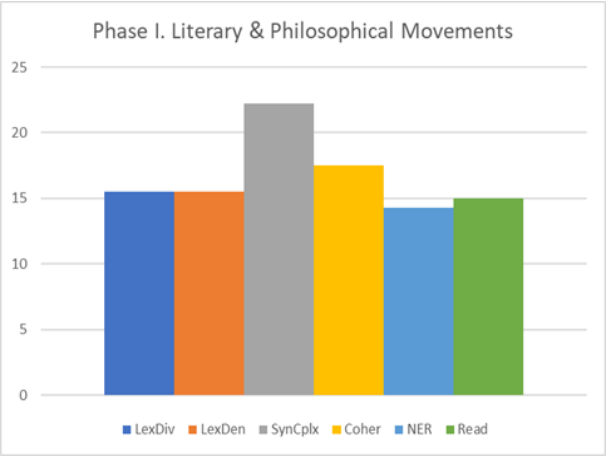

Figure 2. Phase I. Literary & Philosophical Movements

richness and structural complexity in technical texts. Lexical Density and Coherence contribute equally (17%), emphasizing content concentration and logical progression. In contrast, Named Entities (15%) and Readability (14%) play relatively smaller roles, indicating their reduced relevance in this context.

By comparison, Figure 2 reveals a distinct distribution for Literary & Philosophical Movements. Here, Syntactic Complexity leads (22%), highlighting the prominence of elaborate sentence structures. Coherence follows (18%), underscoring the role of narrative flow, while Lexical Density (16%) remains moderately significant. Lexical Diversity and Readability contribute equally (15%), and Named Entities have the lowest impact (14%), consistent with the genre's limited use of specialized terminology.

Phase II further investigates text complexity features across the same two thematic categories: Computing & Web Technologies and Literary & Philosophical Movements. The results are presented in Fig. 3 and Fig. 4, respectively.

In the Computing & Web Technologies domain (Fig. 3), Lexical Density and Syntactic Complexity emerge as the most influential features, each contributing 20% – a shift from Phase I, where Lexical

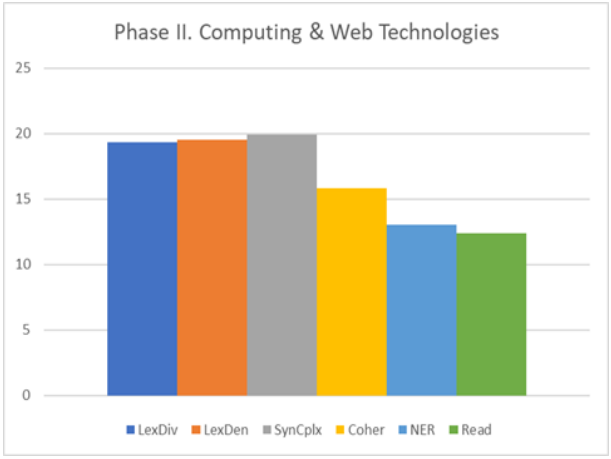

Figure 3. Phase II. Computing & Web Technologies

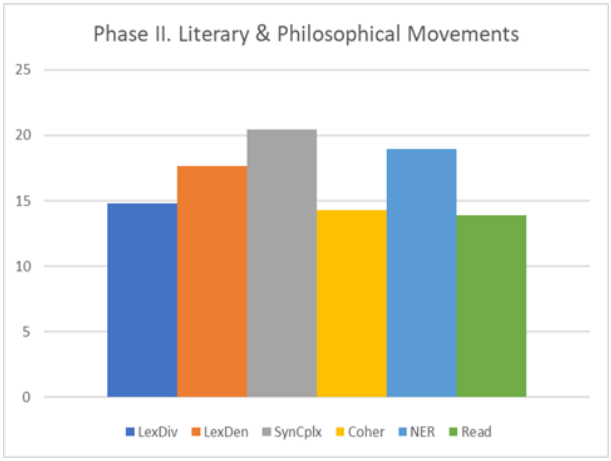

Figure 4. Phase II. Literary & Philosophical Movements

Diversity was dominant. Lexical Diversity remains important at 19%, followed by Coherence (16%). Named Entities and Readability maintain lower influence at 13% and 12%, consistent with Phase I patterns.

In the Literary & Philosophical Movements domain (Fig. 4), Syntactic Complexity continues to lead (20%), reaffirming the importance of complex sentence structures. Named Entities rise sharply to 19%, indicating that under indirect prompting, AI models assign greater weight to semantic and contextual cues. Lexical Density remains stable (18%), while Lexical Diversity and Readability contribute moderately (15% and 14%). Coherence decreases to 14%, reflecting a subtle shift in how text flow is evaluated.

Cross-phase comparisons highlight key trends. In Computing & Web Technologies, the emphasis moves from Lexical Diversity toward Lexical Density and Syntactic Complexity, suggesting a deeper content-level focus under indirect prompting. In Literary & Philosophical Movements, the growing role of Named Entities points to an expanded semantic interpretation influenced by genre-specific content and Phase II strategies.

Overall, most features exhibit stable contributions across phases, with the most notable variations observed in **Lexical Diversity** and **Named Entities**. These shifts underscore the cognitive demands associated with meaning construction and structural organization. The observed fluctuations within the Literary & Philosophical category may stem from corpus imbalance, suggesting the need for further cross-domain validation using broader textual datasets.

Analysis by feature

This section evaluates text complexity features independently of the thematic domain. For example, lexical complexity is assessed uniformly across Dark Romanticism, Quantum Computing, and Dark Web texts. The goal is to reduce domain-specific bias and highlight feature behavior across the full corpus. Fig. 5 presents aggregated results from both experimental phases.

The analysis reveals a stable hierarchy of feature influence. Syntactic Complexity dominates, followed by Lexical Diversity and Lexical Density, which together account for 56% of overall impact. Coherence and Named Entities contribute moderately, while Readability plays a

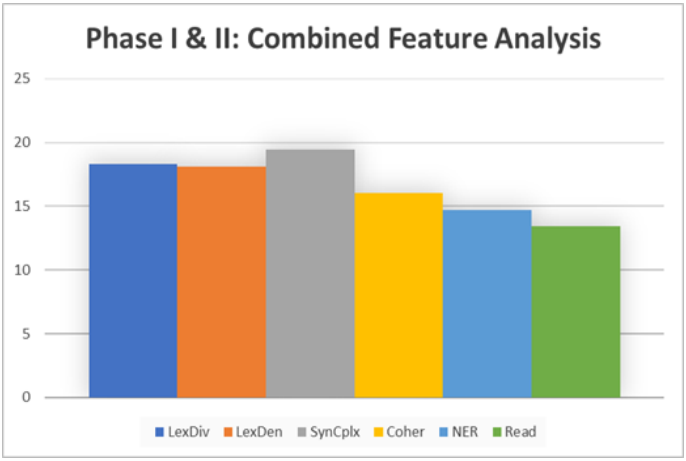

Figure 5. Feature Analysis per Feature: Overview

minor role. These findings confirm that structural and lexical dimensions are key drivers of text complexity across all domains.

Analysis by formula

During Phase I, we analyzed the ten formulas with the highest Spearman rank correlation coefficients (SRCC) for each thematic domain. Table 4 illustrates the top three formulas per domain, using the notation: LD = Lexical Diversity, D = Lexical Density, SC = Syntactic Complexity, Coh = Coherence, NE = Named Entities, R = Readability.

The highest correlations were observed in the Semantic Web domain (e.g., Formulas 152, 17, 47 at 0.812, 0.708, 0.690), outperforming the cross-domain mean (0.56). Dark Romanticism and Dark Web formulas performed moderately (0.49–0.67), while Cloud Computing yielded the weakest results (down to 0.406). These findings suggest that some formulas overlook domain-specific nuances or underweight key features, motivating the proposal of a generalized feature template for complexity modeling.

A review of all 50 top-performing formulas reveals a recurring structure: a weighted linear combination of six features, occasionally adjusted by mild non-linear operations. Three main patterns emerge:

1. Feature Grouping: - *Lexical Block*: *LD* and *D* emphasizes

Table 4. Top 3 formulas representation.

F.No.	Representation	SRCC	Dom
34	$(LD + D) \times SC - \frac{Coh}{NE^2}$	63.45	Cloud comp
27	$\big(LD \times D + SC\big) \times \ln\big(1 + NE\big)$	42.25	Cloud comp
54	$NE \times Coh - (LD + D)$	40.60	Cloud comp
7	$\frac{LD \times SC \; + \; D \times Coh \; + \; NE \times R}{3}$	67.36	Dark rom
15	$0.15\,LD \; + \; 0.15\,D \; + \; 0.15\,SC \; + \; 0.25\,Coh \; + \; 0.10\,NE \; + \; 0.20\,R$	61.65	Dark rom
48	$0.4\,SC \; + \; 0.4\,Coh \; - \; 0.2\,NE$	55.33	Dark rom
104	$0.3\,LD \; + \; 0.2\,D \; + \; 0.3\,SC \; + \; 0.2\,Coh$	57.74	Dark web
40	$(SC \times Coh) \; - \; NE^2 \; + \; \frac{R^3}{LD \; \times \; D}$	53.23	Dark web
69	$\sqrt{LD^2 + D^2 + SC^2 + Coh^2 + NE^2 + R^2}$	49.02	Dark web
8	$0.1\,LD \; + \; 0.15\,D \; + \; 0.2\,SC \; + \; 0.25\,Coh \; + \; 0.15\,NE \; + \; 0.15\,R$	57.89	Quant comp
74	$LD \times D \times SC \times Coh \times (1 + 0.1\,NE) \times (1 + 0.1\,R)$	54.13	Quant comp
28	$\frac{(LD \times D) \; + \; (SC \times D)}{R}$	52.48	Quant comp
152	$0.2\ln(1+LD) \; + \; 0.2\ln(1+D) \; + \; 0.2\ln(1+SC) \; + \; 0.15\ln(1+Coh) \; + \; 0.15\ln(1+NE) \; + \; 0.1\ln(1+R)$	81.20	Sem web
17	$0.2\,LD \; + \; 0.2\,D \; + \; 0.2\,SC \; + \; 0.2\,Coh \; + \; 0.2\,R$	70.82	Sem web
47	$0.4\,(LD + D) \; + \; 0.3\,SC \; + \; 0.1\,Coh \; - \; 0.1\,NE \; + \; 0.1\,R$	69.02	Sem web

vocabulary richness. - *Structural Block*: *SC* and *Coh* reflects syntactic and logical flow. - *Informational vs. Readability*: *NE* and *R* typically carry smaller or compensatory weights.

2. Coefficient Patterns: - *Equal Weights*: (e.g. $\frac{1}{6}$ per feature) dominate. - *Skewed Weights*: often prioritize lexical and structural features (0.2–0.3), with NE and R lower (0.10–0.15).

3. Mathematical Operations: - *Addition (+)* is most common. - *Multiplication (×)* is used within feature blocks (e.g., $LD \times D$). - *Normalization/log* (30%) compresses feature scales. - *Roots/norms* (20%) promote synergy but are less frequent.

An archetypal formula summarizing these patterns is:

$$F_{arch} = wLD + wD + wSC + wCoh + wNER + wR$$

Feature Interactions. Among the 50 complexity formulas, two interaction patterns emerge. **Group 1** uses simple operations (+, ×, −), combining lexical (LD, D) and structural (SC, Coh) features – often to express synergy or trade-offs with readability (R) and entity coverage (NE). **Group 2** applies non-linear transformations: logarithms temper scale, exponentials emphasize outliers, and norm-based operations balance all six features. These patterns reflect varied hypotheses on how linguistic and structural elements co-influence text complexity.

Our analysis identified robust formulaic patterns across domains. Key outcomes include: (1) the centrality of lexical–structural interactions, (2) the strategic use of additive, multiplicative, and non-linear transformations, and (3) domain-specific variations in feature emphasis. These insights inform Phase II, where formula families will be empirically tested on cross-domain corpora, refining complexity metrics through the lens of reader-effort theory.

Phase II Results

Experimental Design. In Phase II, we adopted an indirect prompting strategy: instead of explicitly presenting the six predefined features (LD, D, SC, Coh, NE, R), we embedded twelve broader complexity dimensions into the instructions (e.g., vocabulary sophistication, semantic density, inferential load, rhetorical structure, etc.). This approach aimed to evaluate whether our formulas align with how LLMs perceive and assess complexity when reasoning autonomously across

domain-specific texts. The experiment offers valuable insight into the broader phenomenon of text complexity.

Performance Overview. The 50 formulas tested in Phase II achieved a lower average Spearman rank correlation coefficient (SRCC) of 0.42, compared to Phase I's $\sim$ 0.55–0.60.

Domain-level performance: 1) Dark Romanticism: 54.1 (avg) / 67.36 (max); 2) Semantic Web: 43.1 (avg) / 57.89 (max); 3) Quantum Computing: 41.6 (avg) / 54.58 (max); 4) Dark Web: 40.5 (avg) / 54.73 (max); 5) Cloud Computing: 32.8 (avg) / 63.45 (max).

No fundamentally new mathematical archetype emerged; formulas remained grounded in additive-synergy structures, though modified through varied normalization schemes.

Formulas in Phase II underperformed compared to the domain-tuned models of Phase I, with an SRCC drop of 0.13–0.18. The domain ranking pattern held steady (highest: Dark Romanticism, lowest: Cloud Computing), but Semantic Web and Cloud Computing saw the sharpest declines ($\approx$30–35 points).

These results point to a gap between the qualitative complexity dimensions the LLMs can detect, such as inferential depth and rhetorical structure, and the quantitative features currently embedded in our models. The six original features appear to capture only part of the complexity landscape.

Nevertheless, the experiment successfully identified new feature candidates for inclusion. Despite reduced performance, the consistent $\approx$ 40–50% correlation affirms the core relevance of our six foundational features, while also signaling the need to evolve the general formula to integrate emerging complexity indicators.

5 Conclusion

This study presented a feature-level analysis of text complexity using large language models. In Phase I, explicit prompting of six core features – lexical diversity, lexical density, syntactic complexity, coherence, named entities, and readability – produced robust Spearman correlations (0.55–0.60), with domain-tuned, additive-synergistic formulas outperforming generalized ones. Phase II adopted an indirect

prompting approach, embedding twelve higher-order complexity traits (e.g., inferential load, rhetorical structure, narrative complexity) to explore whether LLMs could surface indicators beyond the original six.

While Phase II formulas yielded a lower average SRCC (0.42), the drop offered valuable insight: the initial six features account for approximately 40% of perceived complexity. Indirect prompting successfully prompted LLMs to detect latent complexity dimensions, highlighting the need for an expanded feature set (see Fig. 6).

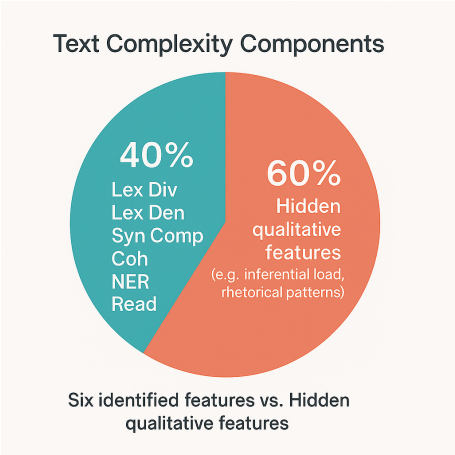

Figure 6. Text complexity features

Overall, our findings suggest three main takeaways: (1) The six foundational features remain essential, capturing the bulk of measurable complexity interactions. (2) Text complexity is partially domain-dependent; only lexical diversity and named entities showed statistically significant domain variance. (3) Indirect prompting can expose richer complexity cues, though these must be translated into quantifiable metrics to recover predictive accuracy.

For future work, we propose a **hybrid modeling framework** that (a) quantifies newly surfaced dimensions using corpus-based metrics (e.g., type-token ratios, discourse markers, cohesion indices), and (b) integrates them with the six core features through log-normalized weights and synergy-based aggregation. We also envision using LLMs

not only for feature discovery but for real-time complexity scoring, combining deterministic NLP tools with LLM-inferred signals in an ensemble architecture.

Acknowledgments. This research is co-funded by the European Union, Marie Skłodowska-Curie Actions (MSCA), HORIZON-MSCA-2021-SE-01 program, as part of the "Elevating Higher Education public policies: an empowering SPRIngboard" (HESPRI) project.

References

[1] M. Solnyshkina, R. Zamaletdinov, L. Gorodetskaya, and A. Gabitov, "Evaluating Text Complexity and Flesch–Kincaid Grade Level," *J. Soc. Stud. Educ. Res.*, vol. 8, no. 3, pp. 238–248, 2017.

[2] A. Parahonco and M. Petic, "E-course: developing a model for content generation," in *Logic & Artificial Intelligence*, S. Cojocaru *et al.*, Eds. Chişinău, Moldova: Vladimir Andrunachievici Institute of Mathematics and Computer Science, MSU, 2023, pp. 199–207. ISBN: 978-9975-68-484-2.

[3] A. Parahonco and M. Petic, "Elearning Content Processing Situations and Their Solutions," in *Proc. Workshop on Intell. Inf. Syst. (WIIS2022)*, (Chişinău, Republic of Moldova), 2022, pp. 154–159. ISBN: 978-9975-68-461-3.

[4] A. Parahonco and M. Petic, "Educational text content generation for elearning systems," in *Proc. 18th Int. Conf. Linguistic Resources and Tools for Natural Language Processing (ConsILR2023)*, (Brașov, Romania), Dec. 11–14, 2023, pp. 43–55. DOI: 10.47743/ConsILR2023.04.

[5] A. Parahonco and L. Parahonco, "The design of a recommendation system for generating content. Context of the sources. Part I," in *Smart Innovation, Systems and Technologies*, Springer, 2025, to be published.

[6] D. Biber, "Conversation text types: A multi-dimensional analysis," in *Le poids des mots: Proceedings of the 7th Int. Conf. on the Statistical Analysis of Textual Data*, (Louvain-la-Neuve, Belgium), 2004, pp. 15–34.

[7] S. J. Amendum, K. Conradi, and E. Hiebert, "Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension," *Educ. Psychol. Rev.*, vol. 30, no. 1, pp. 121–151, Mar. 2018. DOI: 10.1007/s10648-017-9398-2.

[8] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 English words", *Behav. Res. Methods*, vol. 44, no. 4, pp. 978–990, Dec. 2012. DOI: 10.3758/s13428-012-0210-4.

[9] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *Int. J. Corpus Linguist.*, vol. 15, no. 4, pp. 474–496, 2010. DOI: 10.1075/ijcl.15.4.02lu.

[10] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-Metrix: Analysis of text on cohesion and language", *Behav. Res. Methods Instrum. Comput.*, vol. 36, no. 2, pp. 193–202, May 2004. DOI: 10.3758/BF03195564.

[11] Common Core State Standards Initiative. (2010). *Measuring Text Complexity: Qualitative measures—levels of meaning, structure, language conventionality and clarity, and knowledge demands*. [Online]. Available: https://learning.ccsso.org/wp-content/uploads/2022/11/ELA_Standards1.pdf. Accessed on: May 28, 2025.

[12] Massachusetts Department of Elementary and Secondary Education. (2017, June). *Quick Reference Guide: Text Complexity and the Growth of Reading Comprehension*. [Online]. Available: https://www.doe.mass.edu/frameworks/ela/2017-06QRG-ReadingComp.pdf. Accessed on: May 28, 2025.

[13] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas for Navy enlisted personnel," Naval Air Station, Millington, TN, Rep. Research Branch Report 8-75, 1975.

[14] G. H. McLaughlin, "SMOG grading: A new readability formula," *J. Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[15] Common Core State Standards Initiative. (2010). *Qualitative Features of Text Complexity Explained: Companion to the Qualitative Dimensions Scale*. [Online]. Avail-

able: https://www.corestandards.org/assets/Qualitative_Features_of_Text_Complexity_Explained.pdf. Accessed on: May 28, 2025.

[16] New York State Education Department. (n.d.). *Demystifying Complex Texts: What Are Complex Texts and How Can We Ensure Students Can Access Them?* [Online]. Available: https://www.nysed.gov/sites/default/files/programs/bilingual-ed/de-mystifying_complex_texts-2.pdf. Accessed on: May 28, 2025.

[17] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems," *Front. Artif. Intell.*, vol. 7, Article ID: 1410790, Sep. 2024. DOI: 10.3389/frai.2024.1410790.

[18] R. G. Benjamin, "Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty," *Educ. Psychol. Rev.*, vol. 24, no. 1, pp. 63–88, Jan. 2012. DOI: 10.1007/s10648-011-9181-8.

[19] A. Caputo, P. Basile, and G. Semeraro, "Integrating Named Entities in a Semantic Search Engine," in *Proc. 1st Italian Inf. Retr. Workshop (IIR 2010)*, (Padua, Italy), Jan. 27–28, 2010, pp. 15–16. [Online]. Available: https://ceur-ws.org/Vol-560/paper4.pdf. Accessed on: July 15, 2025.

[20] K. Dziuk Lameira, "Zur Komplexität von Texten. Von der Lesbarkeitsformel zur textlinguistischen Komplexität," in *Textkomplexität und Textverstehen. Studien zur Verständlichkeit von Texten*, Berlin/Boston, Germany: De Gruyter, 2023, pp. 69–98. DOI: 10.1515/9783111041551-003.

[21] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural Language Processing: Python and NLTK*, Birmingham, UK: Packt Publishing Ltd, 2016, 687 p.

[22] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python*, Zenodo, 2020. DOI: 10.5281/zenodo.1212303.

[23] V. Bona, *Pylexitext: A Python library for NLP methods and text analysis*, v0.3.2. [Online]. Available: https://github.com/vicotrbb/Pylexitext. Accessed on: Jul. 22, 2025.

[24] A. Parahonco, *Text-Complexity: Research data repository for text complexity analysis*. [Online]. Available: `https://github.com/AlexPex/Text-Complexity`. Accessed on: Jul. 22, 2025.

[25] D. E. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching* (Addison-Wesley Series in Computer Science and Technology), 2nd ed., Reading, MA, USA: Addison-Wesley, 1998, pp. 426–458.

[26] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, pp. 72–101, 1904.

Alexandr Parahonco, Liudmila Parahonco

Received May 29, 2025
Revised 1 July 23, 2025
Revised 2 August 5, 2025
Accepted August 6, 2025

Alexandr Parahonco
ORCID: `https://orcid.org/0009-0007-3486-5597`
Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University, Moldova;
Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, 700506, România
E–mail: `alexandr.parahonco@math.md, alexandr.parahonco@usarb.md`

Liudmila Parahonco
ORCID: `https://orcid.org/0000-0002-7010-3107`
Faculty of Letters, Alecu Russo State University of Bălți, MD-3100, Moldova
E–mail: `liudmila.parahonco@usarb.md`