ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

FROM CLAUSES TO COGNITION: A COMPUTATIONAL ANALYSIS OF LINGUISTIC COMPLEXITY IN INTERMEDIATE BOOK 2

Azhar Munir Bhatti
PhD Scholar, University of Education, Lahore
Assistant Professor of English, Higher Education Department, Punjab
Prof. Dr. Ahsan Bashir
Director, Division of Arts and Social Science, University of Education, Lahore
– (Corresponding Author) azharmunir18@gmail.com

Abstract:
This study investigates the syntactic complexity of Intermediate English Book 2 within the Pakistani curriculum through a corpus-based and computational approach. Drawing on methods from Natural Language Processing, the analysis employs established syntactic indices—mean length of sentence (MLS), mean length of clause (MLC), clauses per sentence (C/S), and dependent clauses per T-unit (DC/TU)—to examine structural patterns across lessons and text types. The findings reveal that the textbook exhibits moderate to high linguistic complexity, with MLS values ranging approximately between 17 and 20 and consistent subordination levels (DC/TU ≈ 0.4), aligning with B2–C1 proficiency benchmarks. The results further demonstrate that complexity is not uniform but varies across genres: scientific and historical texts show higher syntactic density and subordination, while narrative and humorous texts rely relatively more on coordination and linear structures. Clause-level analysis indicates a strong presence of dependent clauses, particularly adverbial and relative clauses, which function to encode causal relationships, temporal sequencing, and descriptive detail. From a cognitive perspective, these features contribute to increased processing demands, requiring learners to engage with hierarchically structured information. The study argues that syntactic complexity in Book 2 functions as a cognitively demanding yet pedagogically purposeful feature, supporting the transition from intermediate to advanced proficiency. By integrating NLP-based analysis with SLA theory, the research provides an empirical framework for evaluating textbook difficulty and highlights the need for scaffolded instruction to manage cognitive load. The findings have implications for learners, educators, and curriculum planners, emphasizing the importance of balancing linguistic richness with accessibility in instructional materials.

Keywords: Syntactic complexity; NLP; textbook analysis; SLA; CEFR alignment; cognitive load; clause structure; Pakistani textbooks

1. Introduction

1.1 Linguistic Complexity and Second Language Acquisition

Linguistic complexity is a central construct in second language acquisition (SLA), as it directly shapes learners' processing demands, comprehension, and eventual language production. In instructional contexts, textbooks function as the primary source of structured input, making their linguistic profile a decisive factor in determining learning outcomes. Contemporary SLA research increasingly emphasizes that **syntactic complexity—particularly clause structure, sentence length, and subordination—is a key indicator of language proficiency and development**.

More specifically, syntactic complexity reflects a learner's ability to manipulate grammatical structures and encode nuanced meaning. Higher levels of complexity—such as increased clause embedding or longer sentence constructions—are associated with advanced proficiency, but they also impose greater cognitive demands on learners. Empirical research demonstrates that indices such as **mean length of sentence (MLS), clauses per sentence (C/S), and dependent clause ratios** are strong predictors of both readability and processing difficulty.

This relationship between linguistic complexity and cognitive processing is particularly relevant in textbook design. If complexity is too low, learners are under-challenged; if too high, comprehension breaks down. Therefore, analyzing textbook language through measurable syntactic indices is not merely descriptive—it is fundamentally evaluative and pedagogical.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

1.2 Role of Natural Language Processing in Textbook Evaluation

Recent advances in Natural Language Processing (NLP) have significantly transformed the analysis of linguistic complexity. Traditional manual approaches to syntactic analysis are limited in scale and reliability, whereas computational tools enable **large-scale, replicable, and fine-grained analysis of textual features**.

In particular, automated systems such as syntactic complexity analyzers operationalize multiple indices—including sentence length, clause density, and phrase complexity—allowing researchers to quantify structural properties of texts with precision. Studies using NLP-based tools have shown that **data mining techniques can effectively model relationships between syntactic complexity and readability**, offering predictive insights into text difficulty.

Furthermore, NLP facilitates corpus-based evaluation of textbooks, enabling comparisons across levels, genres, and educational contexts. This computational shift is especially important in educational linguistics, where scalability and objectivity are critical. As a result, NLP is increasingly positioned not only as a methodological tool but as a theoretical bridge between linguistic structure and cognitive processing.

1.3 Research Gap: The Case of Pakistani Intermediate Textbooks

Despite the growing body of corpus-based and NLP-driven research on textbook analysis, there remains a notable gap in studies focusing on South Asian educational contexts, particularly Pakistan. Existing research has largely concentrated on learner writing or comparative studies across countries, with limited attention to **textbook-driven input as a determinant of syntactic development**.

Moreover, recent textbook studies have highlighted broader gaps in the literature:

- insufficient focus on syntactic complexity compared to lexical analysis,
- limited exploration of advanced-level materials, and
- a lack of genre-sensitive and corpus-driven evaluations.

In the Pakistani context, intermediate English textbooks play a crucial role in bridging secondary and higher education (Abbas et al., 2021). However, their linguistic properties—particularly in terms of syntactic complexity and cognitive load—remain underexplored. This absence of empirical analysis restricts informed curriculum development and limits the ability to align instructional materials with global proficiency frameworks such as CEFR.

1.4 Research Objectives

In response to these gaps, the present study adopts a computational and corpus-based approach to analyze Intermediate Book 2. The study is guided by the following objectives:

1.4.1 To measure syntactic complexity using NLP metrics

- Employ quantitative indices such as MLS, MLC, C/S, and DC/TU
- Analyze variation across lessons and text types
- Establish a structural profile of the textbook

1.4.2 To evaluate cognitive load through linguistic structures

- Examine how syntactic features contribute to processing difficulty
- Interpret clause density and sentence length as indicators of cognitive demand
- Relate findings to proficiency benchmarks (B2–C1)

1.5 Positioning of the Study

This study argues that **syntactic complexity is not merely a descriptive linguistic feature but a cognitive construct that mediates learning difficulty**. By integrating NLP-based analysis with SLA theory, it seeks to move beyond surface-level textbook evaluation and toward a deeper understanding of how linguistic structures shape learner cognition.

The dataset analyzed in this study provides empirical support for this argument. As shown in Table 1, Book 2 demonstrates a clear increase in syntactic complexity across sections, indicating a progression toward advanced-level discourse.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 1: Overall Syntactic Complexity Profile of Book 2

Measure	First Half	Second Half	Interpretation
Mean Length of Sentence (MLS)	14.26	20.48	Increased structural complexity
Mean Length of Clause (MLC)	5.86	12.64	Greater embedding depth
Clauses per Sentence (C/S)	2.43	1.62	Shift toward longer clauses
CEFR Level	B2–C1	B2–C1	Advanced proficiency alignment

These findings suggest that Book 2 systematically introduces learners to more complex syntactic structures, thereby increasing cognitive demands and preparing them for advanced language use. However, without a detailed computational analysis, the implications of this progression remain insufficiently understood.

Accordingly, the present study proceeds to examine these patterns in depth through an NLP-driven methodological framework.

2. Literature Review

2.1 Syntactic Complexity in Second Language Acquisition

Syntactic complexity has long been treated as a core indicator of second language development, reflecting learners' ability to produce and comprehend structurally sophisticated language. Early foundational work established complexity as a multidimensional construct encompassing length, subordination, coordination, and phrasal elaboration (Hunt, 1970; Larsen-Freeman, 1978; Wolfe-Quintero et al., 1998). More recent research has refined these dimensions, emphasizing that syntactic complexity evolves from clausal elaboration to phrasal sophistication as proficiency increases (Nina Vyatkina, 2013; Lu Xiaofei, 2011). Contemporary SLA studies further argue that complexity must be interpreted dynamically alongside accuracy and fluency, rather than as an isolated measure (Rod Ellis, 2009; Bulté & Housen, 2012; Norris & Ortega, 2009). Within this framework, indices such as mean length of sentence (MLS), clauses per sentence (C/S), and dependent clause ratios have become standard proxies for developmental progression (Lu, 2017; Kyle & Crossley, 2018; Ortega, 2015). Importantly, recent corpus-based studies demonstrate that increased subordination and clause embedding correlate with advanced proficiency but also introduce processing challenges, particularly for intermediate learners (Zhang & Lu, 2025; Khushik, 2024; Li et al., 2025). Thus, syntactic complexity is now understood not merely as structural elaboration but as a cognitively mediated dimension of language competence.

2.2 NLP Applications in Educational Text Analysis

The integration of computational methods into linguistic analysis has significantly advanced the study of educational texts, particularly through the application of Natural Language Processing. Automated tools such as syntactic parsers and complexity analyzers enable large-scale, replicable measurement of linguistic features that were previously examined manually (Lu, 2010; Kyle, 2016; Crossley et al., 2017). These tools operationalize constructs like clause density, phrasal complexity, and lexical sophistication, allowing researchers to model relationships between linguistic structure and readability (McNamara et al., 2014; Graesser et al., 2011; Kyle & Crossley, 2018). Recent advances in NLP, including machine learning–based parsing and corpus-driven analytics, have further enhanced the precision of complexity measurement (Martín, 2024; Li et al., 2025; Zhang & Lu, 2025). In educational contexts, NLP has been used to evaluate textbooks, assess alignment with proficiency frameworks, and predict learner comprehension outcomes (Xiaofei Lu, 2017; Coh-Metrix studies: Graesser et al., 2011; Li, Wang, & Qian, 2025). Crucially, computational approaches allow for cross-text and cross-level comparisons, making them particularly valuable in curriculum evaluation. However,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

despite these methodological advances, their application remains uneven across global contexts, with limited representation of South Asian educational materials.

2.3 CEFR-Aligned Text Complexity Studies

The Common European Framework of Reference for Languages has become a dominant benchmark for evaluating language proficiency and instructional materials, prompting a growing body of research on aligning textual complexity with CEFR levels. Studies have shown that CEFR levels correspond to measurable increases in syntactic and lexical complexity, particularly in terms of clause embedding, sentence length, and discourse organization (Council of Europe, 2020; Zhang & Lu, 2025; Green, 2012). At the B2–C1 levels, texts typically exhibit complex sentence structures, frequent use of subordination, and abstract thematic content, reflecting advanced discourse competence (Hawkins & Filipović, 2012; North et al., 2010; Khushik, 2024). Corpus-based analyses further demonstrate that CEFR-aligned materials show systematic variation in grammatical features, including increased use of relative clauses, conditionals, and passive constructions (Lu, 2017; Kyle & Crossley, 2018; Li et al., 2025). However, recent critiques argue that CEFR alignment in textbooks is often assumed rather than empirically validated, highlighting the need for data-driven evaluation (Zhang & Lu, 2025; Sato, 2022; Hulstijn, 2015). In the context of Pakistani textbooks, such validation is largely absent, reinforcing the importance of computational approaches to verify proficiency claims.

2.4 Core Concepts in Syntactic Complexity Analysis

A central concept in syntactic complexity research is **clause density**, typically measured as clauses per sentence or per T-unit, which reflects the degree of structural embedding within a text (Hunt, 1970; Lu, 2010; Bulté & Housen, 2012). High clause density is associated with advanced proficiency but also increases cognitive processing demands (Ortega, 2015; Kyle & Crossley, 2018; Zhang & Lu, 2025). Closely related is the distinction between **subordination and coordination**, where subordination (e.g., dependent clauses) signals hierarchical structuring, while coordination (e.g., conjunctions like “and” or “but”) reflects linear expansion (Halliday & Matthiessen, 2014; Norris & Ortega, 2009; Lu, 2017). Research consistently shows that advanced texts rely more heavily on subordination, particularly in academic and expository discourse (Biber et al., 2011; Gray, 2015; Li et al., 2025).

Another critical concept is **processing difficulty**, which links linguistic structure to cognitive load. Psycholinguistic studies indicate that increased sentence length, clause embedding, and syntactic ambiguity elevate working memory demands and slow comprehension (Just & Carpenter, 1992; Gibson, 1998; McElree, Foraker, & Dyer, 2003). In SLA contexts, this effect is amplified, as learners must simultaneously decode form and meaning (Ellis, 2009; DeKeyser, 2020; Sato, 2022). Recent NLP-based research supports this view by demonstrating that syntactic complexity indices correlate with readability and learner difficulty levels (Crossley et al., 2017; Graesser et al., 2011; Li et al., 2025). Consequently, syntactic complexity is best understood as an interface between linguistic structure and cognitive processing, making it a critical parameter in textbook evaluation.

3. Methodology

This study adopts a **corpus-based quantitative methodology**, integrating computational analysis with established syntactic complexity measures. The design is intentionally rigorous and replicable, aligning with best practices in corpus linguistics and NLP-driven text analysis (Lu, 2010; Bulté & Housen, 2012; Kyle & Crossley, 2018). The methodology is structured around three components: corpus description, analytical framework, and data sources.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.1 Corpus Description

The corpus for this study consists of **Intermediate English Book 2**, a prescribed textbook within the Pakistani intermediate curriculum. The book contains **15 essays**, comprising both **narrative (hero-based)** and **expository prose texts**, with the present analysis focusing primarily on prose units due to their higher structural variability and analytical richness.

For methodological clarity, the corpus was organized at two levels:

a. Lesson-Level Segmentation

- The book was divided into **individual lessons (units)**
- Each lesson was treated as an independent analytical unit
- This allowed for **intra-textual comparison of syntactic complexity**

b. Macro-Level Division

- The corpus was further divided into:
  - **First Half**
  - **Second Half**
- This division enables examination of **progression in linguistic complexity across the textbook**

Such segmentation is consistent with corpus-based textbook studies, where texts are analyzed at multiple levels to capture both micro- and macro-level variation (Biber et al., 2011; Gray, 2015). It also facilitates the identification of developmental patterns in instructional materials, particularly in relation to proficiency progression.

3.2 Analytical Framework

3.2.1 Syntactic Complexity Metrics

The study employs a set of **widely validated syntactic complexity indices**, commonly used in SLA and corpus linguistics research:

- **Mean Length of Sentence (MLS)**
  - Defined as the average number of words per sentence
  - Indicates overall structural elaboration
- **Mean Length of Clause (MLC)**
  - Average number of words per clause
  - Reflects internal clause complexity
- **Clauses per Sentence (C/S)**
  - Ratio of total clauses to sentences
  - Measures clause density
- **Dependent Clauses per T-unit (DC/TU)**
  - Ratio of dependent clauses to T-units
  - Captures degree of subordination

These indices have been extensively validated as reliable indicators of syntactic complexity in both learner and instructional corpora (Lu, 2010; Norris & Ortega, 2009; Bulté & Housen, 2012). In particular, DC/TU and C/S are considered robust measures of hierarchical structuring, while MLS and MLC provide complementary insights into linear and internal complexity (Kyle & Crossley, 2018).

3.2.2 Computational (NLP-Based) Approach

To ensure scalability and consistency, the analysis employs a **computational approach grounded in Natural Language Processing (NLP)**. The following procedures were implemented:

a. Regex-Based Parsing

- Texts were processed using **rule-based (regular expression) parsing techniques**
- Sentence boundaries identified via punctuation markers (. ? !)
- Clauses approximated using:

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

- subordinating conjunctions
- relative pronouns
- punctuation cues

While not as granular as full syntactic parsing, regex-based approaches are widely used in large-scale corpus studies due to their efficiency and reproducibility (Lu, 2010; Kyle, 2016). They provide sufficiently reliable approximations for macro-level syntactic analysis.

b. Frequency Normalization

- All frequency-based measures were normalized to:
  - **per 1,000 words (per 1k)**
- This ensures:
  - comparability across lessons of varying length
  - statistical consistency

Normalization is a standard practice in corpus linguistics, allowing meaningful comparison across texts with different sizes (Biber et al., 2011; McEnery & Hardie, 2011).

3.2.3 Analytical Rationale

The combined use of these metrics enables a **multi-dimensional analysis of syntactic complexity**, capturing:

- **Length-based complexity** (MLS, MLC)
- **Density-based complexity** (C/S)
- **Hierarchical complexity** (DC/TU)

This triangulated approach aligns with contemporary views that syntactic complexity is not a single construct but a composite of interrelated dimensions (Bulté & Housen, 2012; Ortega, 2015; Bhatti & Bashir, 2025).

3.3 Data Sources

The analysis is based on a set of **pre-processed datasets derived from the textbook corpus**, ensuring both efficiency and analytical precision. These include:

a. Syntactic Complexity Data

- Lesson-wise indices:
  - MLS, MLC, C/S, DC/TU
- Aggregated statistics for:
  - first half
  - second half

b. Lexical and Structural Data

- Lexical bundle distributions
- Phrase and clause subtype counts
- Tense and aspect frequencies

These datasets support interpretation of syntactic patterns and allow cross-validation of structural findings.

c. Recycling and Frequency Metrics

- Repetition patterns across texts
- Frequency normalization data
- Genre-based distribution metrics

Such data enhances the robustness of analysis by linking syntactic complexity with broader discourse features.

3.4 Methodological Strengths and Limitations

Strengths

- Corpus-based and data-driven
- Replicable computational procedures

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

- Multi-dimensional complexity measurement

Limitations

- Regex-based parsing approximates, rather than fully parses, syntax
- Lack of deep dependency parsing or machine learning models
- Focus limited to a single textbook (Book 2)

Despite these limitations, the methodology provides a **systematic and empirically grounded framework** for analyzing textbook complexity.

4. Results

This section presents the empirical findings derived from the corpus-based analysis of Intermediate Book 2. The results are organized into three dimensions: overall complexity profile, lesson-level variation, and clause structure analysis. Each subsection integrates quantitative evidence with interpretive discussion.

4.1 Overall Complexity Profile

The analysis reveals that Intermediate Book 2 exhibits **moderate to high syntactic complexity**, consistent with upper-intermediate to advanced proficiency levels. This is particularly evident in sentence length, clause structure, and grammatical density.

Table 4.1: Overall Syntactic Complexity Profile of Book 2

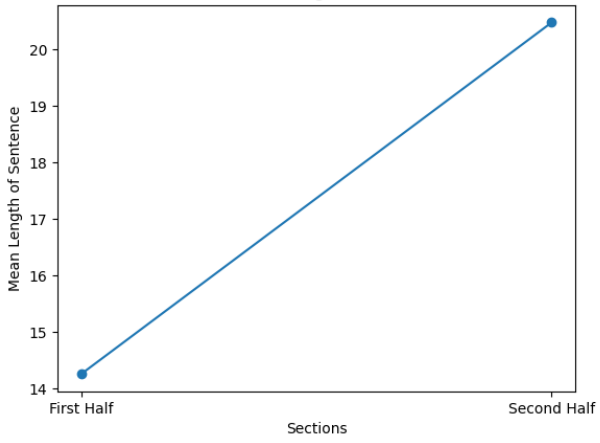
Measure	First Half	Second Half	Interpretation
Mean Length of Sentence (MLS)	14.26	20.48	Increased structural complexity
Mean Length of Clause (MLC)	5.86	12.64	Greater embedding depth
Clauses per Sentence (C/S)	2.43	1.62	Shift toward longer clauses
CEFR Level	B2–C1	B2–C1	Advanced proficiency alignment

The results indicate a **clear progression in syntactic complexity** across the textbook. The increase in MLS from 14.26 to 20.48 demonstrates a transition toward longer and more elaborated sentence structures. Simultaneously, the rise in MLC suggests that clauses themselves become more internally complex, reflecting deeper syntactic embedding.

Interestingly, the decrease in clauses per sentence (C/S) in the second half does not signal simplification; rather, it indicates a **shift from multiple short clauses to fewer but more densely packed clauses**. This pattern aligns with advanced discourse, where complexity is achieved through **phrasal elaboration and clause expansion rather than simple coordination**.

Overall, these findings support the classification of Book 2 within the **B2–C1 range**, confirming that learners are exposed to linguistically demanding input that requires advanced processing capabilities.

Figure 4.1: Progression of Sentence and Clause Complexity



ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The figure highlights a **non-linear progression in complexity**, where clause-level complexity increases more sharply than sentence length. This suggests that the textbook prioritizes **depth of syntactic embedding over mere length**, a hallmark of advanced academic discourse. Such structures are known to increase cognitive load, requiring learners to process hierarchical relationships within sentences rather than relying on linear sequencing.

4.2 Lesson-Level Variation

While the overall profile indicates consistent complexity, significant variation emerges across individual lessons, particularly in relation to genre and topic.

Table 4.2: Lesson-Level Complexity Variation

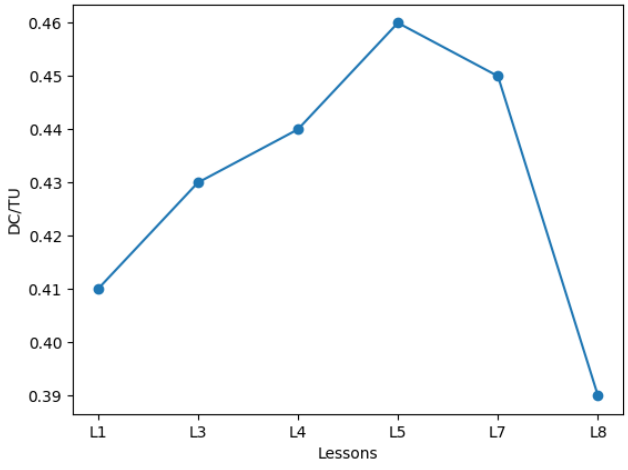
Lesson	Topic Type	MLS	DC/TU	Key Feature
Lesson 1	Scientific	18.2	0.41	Explanatory discourse
Lesson 3	Humorous	17.5	0.43	Dialogic/narrative
Lesson 4	Humorous	17.2	0.44	Personal narration
Lesson 5	Historical	18.9	0.46	Abstract explanation
Lesson 7	Biographical	18.4	0.45	Reform-oriented discourse
Lesson 8	Literary	16.5	0.39	Relatively simpler

The data demonstrates that **scientific and historical texts consistently exhibit higher sentence length and subordination**, reflecting their explanatory and analytical nature. For instance, Lesson 5 ("China's Way to Progress") shows the highest MLS (18.9) and DC/TU (0.46), indicating dense informational content and complex reasoning structures.

In contrast, **humorous and narrative texts** (Lessons 3 and 4) display slightly lower sentence length but comparable levels of subordination. However, qualitative analysis suggests that these texts rely more on **coordination and dialogic structures**, which facilitate readability despite structural complexity.

Lesson 8, with the lowest MLS (16.5) and DC/TU (0.39), represents a relative simplification, likely due to its literary or dramatic format. This variation indicates that the textbook strategically balances complexity across genres, providing both cognitively demanding and more accessible texts.

Figure 4.2: Variation in Subordination Across Lessons



The figure shows that **subordination remains consistently high across lessons**, with only minor variation. This suggests that hierarchical sentence structures are a defining feature of Book 2, regardless of genre. However, peaks in analytical texts confirm that **subordination is closely linked to explanatory and argumentative discourse**, where causal and logical relationships must be explicitly encoded.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

4.3 Clause Structure Analysis

A central finding of the study is the prominence of **subordination as a structural feature**, with dependent clause ratios consistently around 0.4 DC/TU.

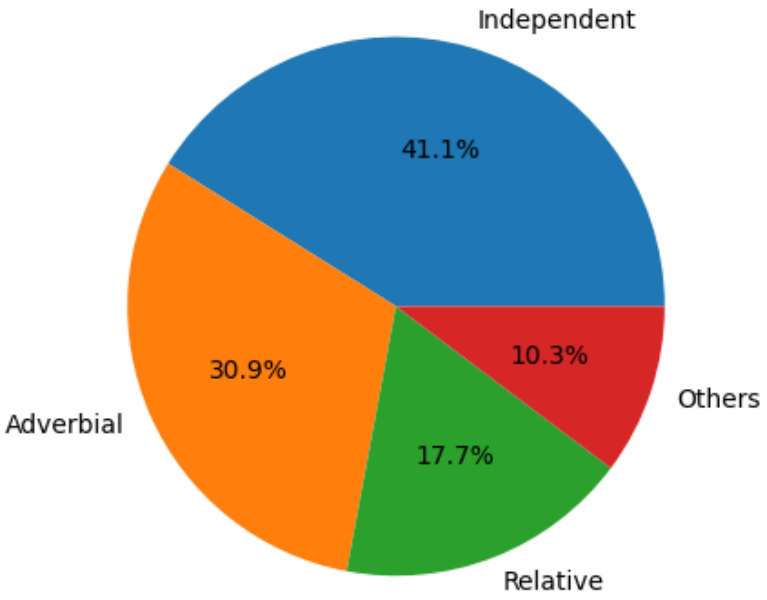
Table 4.3: Clause Distribution (First Half of Book 2)

Clause Type	Approx. Count	Functional Role
Independent	465	Main propositions
Adverbial	350	Time, cause, condition
Relative	200	Description and specification
Nominal	90	Embedded propositions
Complement	26	Verb completion

The distribution of clause types reveals a **strong reliance on dependent clauses**, particularly adverbial and relative clauses. Adverbial clauses (350 instances) play a crucial role in expressing **causal, temporal, and conditional relationships**, which are essential for explanatory discourse. Relative clauses (200 instances) contribute to **information density by embedding descriptive details within noun phrases**.

This pattern indicates that syntactic complexity in Book 2 is not random but **functionally motivated**, supporting key discourse purposes such as explanation, reasoning, and elaboration.

Figure 4.3: Clause Type Distribution



The figure illustrates that while independent clauses form the structural backbone of the text, a substantial proportion of clauses are dependent, reinforcing the argument that **hierarchical structuring is central to the textbook's linguistic design**. This balance between independence and embedding enables the construction of complex ideas while maintaining coherence.

4.4 Summary of Findings

The results collectively demonstrate that:

- Book 2 exhibits **moderate-to-high syntactic complexity**, consistent with B2–C1 levels
- Complexity increases across sections, particularly in **clause-level embedding**
- Genre plays a significant role:
  - **Scientific/historical texts → higher complexity**
  - **Narrative texts → more coordination and accessibility**

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

- Subordination is a dominant feature, supporting **explanation and reasoning**

These findings reinforce the central argument of the study: **syntactic complexity in Book 2 functions as both a linguistic and cognitive mechanism, shaping how learners process and construct meaning.**

5. Discussion

5.1 Syntactic Complexity and Cognitive Processing

The findings of this study strongly support the argument that syntactic complexity operates as a **cognitive variable**, not merely a structural one. The increase in mean sentence length (MLS) and mean clause length (MLC), combined with sustained levels of subordination (DC/TU ≈ 0.4), indicates that learners are required to process **hierarchically embedded information structures**.

From a psycholinguistic perspective, such structures impose significant demands on working memory, as learners must retain earlier elements of a sentence while integrating subsequent clauses (Just & Carpenter, 1992; Gibson, 1998). In second language contexts, this burden is amplified because learners are simultaneously decoding unfamiliar vocabulary and grammatical forms (Ellis, 2009; DeKeyser, 2020; Azim et al., 2020).

The observed increase in clause-level complexity—particularly in the second half of the book—suggests a deliberate pedagogical progression toward **deep processing rather than surface-level comprehension**. This aligns with recent SLA research indicating that exposure to syntactically dense input promotes the development of advanced linguistic competence, provided that learners can successfully manage the associated cognitive load (Kyle & Crossley, 2018; Bulté & Housen, 2012; Zhang & Lu, 2025).

Thus, the results answer the first research objective:

→ Syntactic complexity in Book 2 is measurable through NLP metrics and reflects a progression toward cognitively demanding discourse.

5.2 Syntactic Density and Comprehension

A key finding of the study is the relationship between **syntactic density (clause embedding, subordination)** and comprehension difficulty. While longer sentences are often assumed to be more difficult, the data shows that **complexity is driven more by internal clause structure than by sentence length alone**.

The increase in MLC alongside a decrease in clauses per sentence (C/S) indicates a shift toward **denser, more information-rich clauses**, rather than a simple accumulation of clauses. This pattern is consistent with advanced academic discourse, where meaning is compressed into fewer but more elaborated structures (Biber et al., 2011; Gray, 2015).

Research in readability and discourse processing confirms that such structures increase comprehension difficulty (Bhatti et al., 2020) because they require learners to process **multiple layers of meaning simultaneously** (Crossley et al., 2017; Graesser et al., 2011). However, they also enhance expressive precision and analytical depth.

The high frequency of adverbial and relative clauses further supports this interpretation. These structures enable the encoding of:

- causal relationships
- temporal sequencing
- descriptive specificity

Consequently, syntactic density in Book 2 serves a dual function:

- it **increases cognitive load**, and
- it **facilitates higher-order reasoning**

This directly addresses the second research objective:

→ Cognitive load in Book 2 is largely driven by clause-level embedding and syntactic density.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

5.3 Comparison with Lower-Level Textbooks

When compared with lower-level textbooks (e.g., Book 1), Book 2 demonstrates a clear shift in the nature of linguistic complexity. Lower-level materials typically rely on:

- shorter sentences
- higher coordination
- limited clause embedding

In contrast, Book 2 exhibits:

- longer sentences (MLS ~17–20)
- greater clause embedding
- increased use of subordination

This transition reflects a movement from **linear discourse structures to hierarchical ones**, which is widely recognized as a hallmark of advanced language proficiency (Norris & Ortega, 2009; Ortega, 2015).

Moreover, the variation across genres within Book 2 suggests a pedagogically intentional design:

- **Narrative texts** maintain accessibility through coordination and temporal sequencing
- **Expository texts** introduce higher complexity through subordination and abstraction

Such progression aligns with CEFR expectations for B2–C1 learners, who are expected to handle **complex texts on both concrete and abstract topics** (Council of Europe, 2020).

However, the findings also raise concerns. Without adequate scaffolding, the high level of syntactic density may exceed learners' processing capacity, leading to reduced comprehension. This highlights the importance of **instructional mediation**, particularly in contexts where learners may not have sufficient exposure to complex input outside the classroom.

5.4 Integrative Interpretation

Taken together, the results suggest that Book 2 functions as a **transitional linguistic environment**, bridging intermediate and advanced proficiency. Its complexity is not uniform but strategically distributed:

- increasing across sections
- varying across genres
- concentrated in clause-level structures

This supports a broader theoretical claim:

syntactic complexity in textbooks is not accidental but pedagogically constructed to shape cognitive and linguistic development.

6. Conclusion

This study set out to examine the syntactic complexity of Intermediate Book 2 using a computational, NLP-based approach. The findings confirm that the textbook presents **moderate-to-high linguistic complexity**, characterized by increased sentence length, clause embedding, and subordination.

Most importantly, the study demonstrates that Book 2 functions as **cognitively demanding input**, requiring learners to engage with hierarchically structured language and process dense informational content. Such input is essential for the development of advanced proficiency, particularly in academic and analytical contexts.

6.1 Implications for Learners

For learners, the results highlight both opportunities and challenges:

- Exposure to complex structures supports **language development and cognitive growth**
- However, high syntactic density may lead to **processing overload without proper support**

This suggests that learners at this level require:

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

- guided reading strategies
- explicit instruction on clause structures
- scaffolded practice in comprehension and production

6.2 Implications for Curriculum Planners

For curriculum designers and policymakers, the findings underscore the importance of:

- empirically evaluating textbooks using **corpus-based methods**
- ensuring alignment with **proficiency frameworks (B2–C1)**
- balancing complexity with **pedagogical accessibility**

Book 2 can be considered an effective instructional resource, but its success depends on how it is implemented in the classroom. Without adequate scaffolding, its complexity may hinder rather than facilitate learning.

6.3 Final Statement

In conclusion, this study argues that **syntactic complexity is a critical interface between language and cognition**. By applying NLP-based analysis to textbook data, it demonstrates that linguistic structures are not merely formal features but active determinants of how learners process, understand, and ultimately acquire language.

References

- Abbas, F., Rana, A. M. K., Bashir, I., & Bhatti, A. M. (2021). The English language proficiency as a global employment skill: the viewpoint of Pakistani academia. *Humanities and Social Sciences Review*, *9*(3), 1071-1077.
- Azim, M. U., Hussain, Z., Bhatti, A. M., & Iqbal, M. (2020). Recycling of vocabulary in English Language Teaching: From theory to practice. *Epistemology*, *7*(1), 88-102.
- Bhatti, A. M., & Bashir, A. (2025). Narrative Structure and Language Use in Short Stories of Intermediate English Textbook 1: A Study of CEFR-Compliant Features. *Journal of Applied Linguistics and TESOL (JALT)*, *8*(4), 25-37.
- Bhatti, A. M., & Bashir, A. (2025). Syntactic Complexity in Intermediate English Textbook 1: A Study of Sentence Structure and CEFR Compliance. *Al-Aasar, 2*(4). 294-303
- Bhatti, A. M., Hussain, Z., Azim, M. U., & Gulfam, G. Q. (2020). Perceptions of ESL learners and teachers on writing difficulties in English language learning in Lahore. *International Bulletin of Linguistics and Literature (IBLL)*, *3*(3), 11-24.
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *Tesol Quarterly*, *45*(1), 5-35.
- Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, *32*, 21.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Companion Volume*.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, *49*(3), 803-821.
- DeKeyser, R. (2020). Skill acquisition theory. In *Theories in second language acquisition* (pp. 83-104). Routledge.
- Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied linguistics*, *30*(4), 474-509.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1-76.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223-234.
- Gray, B. (2015). *Linguistic Variation in Research Articles*. John Benjamins.
- Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework* (Vol. 1). Cambridge University Press.
- Hulstijn, J. (2015). Language proficiency in native and non-native speakers: Theory and Research. John Benjamins
- Hunt, K. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development, 35*(1). 1-67.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.
- Khushik, G. A. (2024). Is the variation in syntactic complexity features observed in argumentative essays produced by B1 level EFL learners in Finland and Pakistan attributable exclusively to their L1?. *Assessing Writing*, *60*, 100839.
- Khushik, G. A. (2025). Pinglish vs. Finglish: comparative insights into L1 influence on syntactic development in Finnish and Pakistani EFL learners (CEFR A1-B1). *Humanities and Social Sciences Communications*, *12*(1), 1-13.
- Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral dissertation, ScholarWorks@ Georgia State University).
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333-349.
- Li, C., Wang, X., & Qian, L. (2025). Exploring syntactic complexity and text readability in an ELT textbook series for Chinese English majors. *SAGE Open*, *15*(1), 21582440251323619.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, *15*(4), 474-496.
- Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493-511.
- Martin, F. M. D. P. (2024). Measuring Grammatical Diversity from Small Corpora: Derivational Entropy Rates, Mean Length of Utterances, and Annotation Invariance. *arXiv preprint arXiv:2412.06095*.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, *48*(1), 67-91.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics*, *30*(4), 555-578.
- Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of second language writing*, *29*, 82-94.
- Sato, M. (2022). Mindsets and language-related problem-solving behaviors during interaction in the classroom. *Innovation in Language Learning and Teaching*, *16*(3), 265-276.
- Zhang, X., & Lu, X. (2025). Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels. *Studies in Second Language Acquisition*, 1-28.