北海道言語文化研究
No. 24, 1-14, 2026.
北海道言語研究会

Harry Potter and the Maturity of Language: Lexical and Syntactic Complexity Across the Series

Robert J. Ashcroft

ハリー・ポッターと言語の成熟 ― シリーズにおける語彙的・統語的複雑性の分析

アシュクロフト　ロバート　ジョン

Abstract: This study examines the progression of lexical and syntactic complexity across the seven Harry Potter novels by J. K. Rowling and investigates how this progression aligns with patterns of linguistic development during adolescence. Using a corpus-linguistic approach, the seven texts were analyzed with a range of computer-based measures, including lexical diversity (TTR, hapax legomena, Guiraud's Index), lexical density and coverage (GSL, AWL, total coverage), syntactic complexity (average words per sentence, verbal elements per sentence), and readability and processing difficulty (syllables per word, polysyllabic words, readability indices). The results reveal a clear trend of increasing lexical and grammatical complexity across the series, corresponding closely to developmental trajectories documented in prior research on adolescent vocabulary and syntactic growth. These findings suggest that popular adolescent literature may reflect graded linguistic development and offer empirically grounded guidance for pedagogical sequencing. In EFL contexts such as Japan, these results indicate that Harry Potter Books 1–3 may be more suitable for intermediate learners (GFI 7–9), while Books 4–7 may be better suited to advanced learners (GFI 8.5–10.5), balancing increasing lexical diversity with relatively stable high-frequency vocabulary coverage.

Keywords: Corpus linguistics; Lexical complexity; Syntactic complexity; Readability; Adolescent language development; Young adult literature; EFL pedagogy, Harry Potter

1. Introduction

The Harry Potter series is one of the most popular collections of books of all time. J.K. Rowling's fictional stories about the life of a young wizard have been translated into more than 80 languages and are dearly loved by children and adults alike. In addition to their commercial and cultural success, these books also represent a unique opportunity for linguistic research. Upon reading the books, the author perceived a steady increase in language complexity from one book to the next. Tracing Harry's life from age eleven to seventeen, it seemed to make sense that the seven books should follow a steady progression of linguistic complexity, narrative sophistication, and increasing cognitive demands placed on a maturing readership. This study uses a corpus linguistics approach to systematically analyze the texts of the Harry Potter books to confirm this progression intuited by the author.

Existing research in developmental psychology and applied linguistics shows that there are important stages in the development of children's language as they mature. These changes can be seen in the increasing lexical variety, lexical density, sentence-level syntax complexity, and readability difficulty of texts that adolescents can process as they mature (Anglin, 1993; Berman & Slobin, 1994; Berman & Verhoeven, 2002; Chall, 1983; Nippold, 2007).

There have been numerous stylistic and literary studies of the Harry Potter series, including analyses of narrative techniques and discourse features (Medan et al., 2024), stylistic registers (Jamal & Nasrum, 2018), functional stylistics in children's literature (Zhao & Wang, 2018), and comparative stylistic portraits (Glinka et al., 2021). However, this body of research has tended to adopt qualitative, interpretive, or text-internal approaches and has not systematically examined developmental changes in linguistic complexity across the series using corpus-based methods. Consequently, there is relatively little corpus-based research investigating whether the language of the Harry Potter series reflects a trajectory of linguistic development comparable to that observed in adolescent readers. The current study aims to partially address this gap by providing corpus-based evidence that may help inform the pedagogical use of popular adolescent literature, particularly in EFL contexts such as Japan.

In EFL contexts such as Japan, the Harry Potter series is frequently perceived by teachers and learners as an accessible and motivating source of extended reading due to its global popularity and ready availability. However, decisions about when and how to introduce individual volumes of the series are often based on intuition or general proficiency assumptions rather than systematic linguistic evidence. Therefore, understanding how linguistic complexity develops across a widely read adolescent series of books has clear pedagogical advantages.

This research used each of the seven Harry Potter books as an individual corpus. A battery of computer-based lexical and surface-level syntactic measures was employed to conduct a comparative analysis across the series. The aim was to investigate whether linguistic development through the books reflects patterns already identified in adolescent language development. To this end, the following research questions are addressed:

RQ1: Do the Harry Potter books show evidence of increasing lexical and syntactic complexity across the series?

RQ2: Do the findings for RQ1 align with the developmental patterns described in the current literature about adolescents' ability to process language of increasing complexity as they mature?

The value of this study lies in its focus on language development as reflected in one of the most widely read young adult series of the past three decades. While research on adolescent language growth has extensively examined spoken and written data from educational and experimental contexts (e.g., Berman & Slobin, 1994; Nippold, 2007; Crossley et al., 2011), less attention has been given to how popular literature may mirror the developmental patterns of its intended readership, particularly from a corpus-based perspective.

This study analyzes the complete text of the novels (including narration and dialogue from characters of all ages) rather than isolating the language produced by adolescent characters. The approach allows for a comprehensive analysis of the texts as they are read by adolescents. The corpora used in this study represent Rowling's authorial voice and literary style, and not authentic adolescent speech. This study proceeds from the premise that literature written for adolescent audiences tends to align with readers' developing linguistic abilities, even if such alignment is not explicitly planned or systematically engineered. Thus, the Harry Potter series provides a unique corpus spanning seven books published over a decade and aimed at an audience that matured alongside the texts. By analyzing these novels using established corpus linguistic tools, this study contributes to our understanding of how popular literature for adolescent readers can reflect, and potentially shape, the linguistic development trajectory of its readership.

2. Linguistic Development During Adolescence

Linguistic development during adolescence is a complex process that involves changes in vocabulary, syntax, readability, and word-level processing demands. Research in this area examines the receptive and productive language abilities of adolescent learners, which directly inform our understanding of what level of linguistic complexity is appropriate for age-graded texts. While children usually acquire the basic foundations of grammar and high-frequency vocabulary before puberty (Ricketts et al., 2020), teenagers continue to expand and refine their linguistic skills throughout their adolescence. This ongoing development can be seen in their increasing use of sophisticated vocabulary, more complex sentence structures, and ability to process longer and more demanding texts. The following sections review relevant research in each of these areas and explain how the present study measures lexical and syntactic development across the Harry Potter volumes.

2.1 Lexical Diversity

Studies indicate that lexical development proceeds unabated during adolescence, as young people increase the scope and depth of the words they understand and use. Children acquire a foundation of basic words and the most frequently used vocabulary, and in adolescence, teenagers increasingly use less frequent, more specialized, and more abstract words and terms as they age (Corson, 1997; Nippold, 2007). These changes represent a gradual expansion of lexical diversity during adolescence. The current study uses Guiraud's Index, a modified type-token ratio (TTR) measure, and hapax legomena measures to chart lexical diversity in the Harry

Potter books. Taken together, these two metrics provide a nuanced understanding of vocabulary richness across the series.

2.2 Lexical Density & Coverage

Throughout childhood and adolescence, spoken and written lexical density tends to increase with age, although it remains lower in speech than in writing at every age. A cross-sectional study of 10-, 13-, and 17-year-olds, plus adults, found significant age effects for lexical density in speech (Johansson, 2009). As they mature, teenagers gradually move away from reliance on high-frequency words and simple constructions towards increased content word use (Halliday, 1985; Ure, 1971). This progression means that teenagers show a growing ability to produce more information-rich discourse, a trend that is especially noticeable in academic and expository language contexts (Biber, 1995). These findings point to the ability of children, teenagers, and young adults to produce and understand language with a steadily increasing lexical density as they age.

In addition to vocabulary density, research using vocabulary coverage measures also shows how teenagers extend beyond the most frequent 1000 words to use less common and more specialized vocabulary (Ricketts et al., 2020). Increasing lexical density and spreading coverage into less frequent vocabulary throughout adolescence are both clear signals that the older children become, the more able they are to not only produce but also comprehend lexically broader and richer language.

Lexical density and coverage were measured in the present study using overall lexical density, the proportion of words from the first and second 1,000-word bands of the General Service List (GSL), Academic Word List (AWL) coverage, and total coverage. The GSL was originally developed to represent the most frequent and widely useful words in English (West, 1953), while the AWL was later derived from academic corpora to capture high-frequency academic vocabulary not included in the GSL (Coxhead, 2000). Together, these lists are widely used in applied linguistics to estimate vocabulary difficulty and lexical sophistication. In the present study, these metrics chart the balance of content and functional words and the degree to which the Harry Potter books employ frequent versus less frequent vocabulary, allowing for comparison of information load and lexical range across the series.

2.3 Sentence-level Syntax and Word-level Complexity

According to several studies on language development, syntactic complexity continues to develop throughout adolescence. While children master the fundamentals of grammar structure before puberty, adolescents' ability to produce and understand increasingly complex sentences continues to develop during their teenage years. Specifically, adolescents demonstrate an increasing ability to understand and use subordinate clauses, embedded structures, and longer sentences (Nippold, 1998, 2000). Later studies show that metrics such as T-unit length, clausal density, and the ratio of subordinate to coordinate structures all increase throughout adolescence (Nippold et al., 2008). These studies point to ongoing syntactic maturation from late childhood to early adulthood. While comprehensive syntactic analysis requires sophisticated parsing tools to examine features such as subordination ratios, clause embedding depth, and T-unit complexity, surface-level metrics like sentence length and verbal density provide accessible indicators of sentence elaboration that have been shown to correlate with more detailed measures of syntactic complexity (Biber et al., 2011; Lu, 2010). The present study employs such surface-level measures to trace syntactic development across the Harry Potter series.

Reading and language development research has also shown that word-level features significantly affect processing difficulty. Greater cognitive effort is required to decode longer words, which can lead to slower reading rates and higher comprehension demands (Just & Carpenter, 1987; Rayner et al., 2012). In teenagers, the ability to process longer words develops in conjunction with broader lexical and syntactic growth (Nippold, 2007; Perfetti, 2007). These results highlight the role of word-level complexity as a key indicator of linguistic difficulty in texts.

The present study uses two types of sentence-level measures: verbal elements per sentence as a metric of syntactic elaboration, and the percentage of words with more than two syllables as a measure of word-level complexity. The former serves as an indirect proxy for sentence elaboration and clausal density, while the latter reflects the cognitive processing demands associated with decoding longer, less frequent words. Both these features are identified by Nippold (2000) and others as significant aspects of linguistic maturation during the

teenage years. Although these surface-level measures do not capture the full range of syntactic complexity, they provide systematic and replicable indicators of increasing linguistic demands across the Harry Potter books.

2.4 Readability

Readability is a measure of the difficulty of reading a text. In other words, it measures how accessible a text is to a reader as a function of their linguistic development. Readability formulas aim to predict the difficulty of written texts by analyzing surface features such as word and sentence length and syllable counts (Crossley et al., 2011). While readability indices do not capture a complete view of comprehension difficulty, they remain valuable as broad, surface-level indicators of text complexity (Crossley et al., 2011). As children's and adolescents' reading skills improve, they can comprehend texts with increasingly high readability scores. The current study uses the Gunning Fog Index to measure how the readability of the series develops across the seven volumes, serving as an indicator of the changing linguistic demands placed on Harry Potter readers.

The above strands of research highlight the ongoing development of adolescent language, from vocabulary use and syntactic growth to text readability and word-level processing demands. The present study applies a range of established metrics to the Harry Potter series to trace how these linguistic features shift across the seven books. The following section outlines the methods and tools employed.

3. Methodology

3.1 Corpora

This research is based on the complete seven-volume Harry Potter series by J. K. Rowling. To prepare the texts for analysis, the author purchased the official Kindle editions of all seven novels from Amazon Japan. The editions used were the 2012 Kindle releases of the original works published between 1997 and 2007 (Rowling). The e-books were obtained in their standard Kindle format (AZW/KF8) and subsequently converted to plain text format using Calibre (Goyal, 2024), an open-source e-book management application with built-in format conversion capabilities.

The converted text files were then carefully reviewed using word-processing software to remove paratextual elements such as publisher information, copyright notices, and formatting artifacts introduced during the conversion process. This procedure ensured each corpus contained only the narrative text suitable for linguistic analysis. Each volume was processed as a separate file and verified for textual accuracy by comparison with sample passages from the original Kindle editions.

Under Japanese copyright law, computational analysis of copyrighted works for non-commercial research purposes falls within permissible use for information analysis, provided the works are legally obtained and not redistributed (Copyright Act, 1970, Article 47-7). The digital corpora assembled for this study were used exclusively for the analytical purposes described and were not shared or distributed. This procedure produced accurate and legally obtained corpora for subsequent linguistic analyses.

Thus, the seven corpora included *all the text* from the Harry Potter novels. This naturally includes third-person narration, dialogue from characters of all ages (children, adolescents, and adults), and descriptive passages. The present study does not focus on the language produced by adolescents; rather, it investigates the linguistic complexity of texts written for an adolescent readership. The study aims to look at how Rowling's writing as a whole increases in lexical and syntactic sophistication to match the developing reading abilities of her target audience. Thus, when this study refers to *adolescent language development*, it refers to the *linguistic competencies that adolescent readers are developing*, and not to the speech patterns of the adolescent characters within the novels.

In addition to the seven-book corpora, two smaller corpora were created for each book: one for the first 10,000 words of each book and one for the last 10,000 words of each book. These were used with the Text Inspector online analysis tool, which is limited to texts of up to 10,000 words. The next section examines each analytical tool in detail, including the Text Inspector app. The following sections include a detailed description of the Text Inspector application, along with the other tools used in the analysis.

3.2 Tools

3.2.1 AntWord Profiler

AntWord Profiler, created by Laurence Anthony (2014), is a free downloadable vocabulary analysis program. It is used to measure the extent to which a text is covered by high-frequency and academic vocabulary. The tool compares words in a text with standard lists and calculates coverage percentages. In this study, the General Service List (GSL) (1st 1–1000, and GSL 2nd 1001–2000) and Academic Word List 570 (AWL 570) settings were used. The software was downloaded and installed on a personal computer for analysis. Each of the seven Harry Potter books was loaded into the program in plain text format. The program reported the percentage of words in each book that matched the first 1,000 and second 1,000 bands of the GSL, and the AWL. The total coverage of words in the text covered by these three lists was also calculated.

3.2.2 AntConc

AntConc, also developed by Laurence Anthony (2014), is a free concordance and text analysis program for analyzing word frequency, concordances, collocations, and keyword distributions. In this study, it was used to calculate the number of tokens (total words) and types (unique words) in each of the seven Harry Potter books. AntConc was installed on a personal computer, and each Harry Potter book was loaded into the program in plain text format. The word list tool was used to generate the token and type counts.

3.2.3 Lexical Diversity Analysis Tool (LDAT)

LDAT (Reuneker, 2017) is a free web-based application that calculates a number of lexical diversity and surface-level text metrics. For the present study, LDAT was used to calculate the following measures: token–type ratio (TTR), hapax legomena as a percentage of types, and Guiraud's Index. In addition, LDAT was used to obtain the average number of words per sentence, and the mean word length.

3.2.4 Text Inspector

Text Inspector (Lexical Computing Ltd, 2025) is an online text analysis tool that generates lexical, syntactic, and readability metrics. In this study, the paid version of Text Inspector was used, allowing uploads of up to 10,000 words per file. Due to this limitation, two separate corpora were created for each book: one consisting of the first 10,000 words and another of the last 10,000 words. These were uploaded to the platform in plain text format for analysis. Text Inspector was used to calculate the verbal elements per sentence, the percentage of words with more than two syllables, and the Gunning Fog Index. Therefore, two sub-corpora were used for these three metrics, corresponding to the first and last 10,000 words of each book. The result of each measure was calculated as the mean of the results from the two sub-corpora. The choice of first and last 10,000 words provides insight into how each book opens and concludes, potentially capturing Rowling's most accessible prose (opening) and most complex passages (climax/resolution). While this sampling strategy introduces a degree of limitation, it does allow for consistent cross-book comparison.

3.2.5 Voyant Tools

Voyant Tools (Sinclair, S., & Rockwell, G., 2016) is a free, web-based text analysis platform that is widely used in digital humanities research and language education. In this study, Voyant was used to calculate Lexical Density and the Coleman–Liau Readability Index. The text samples were uploaded directly into the Voyant web interface, and the results were automatically generated in the "Summary" and "Readability" panels.

3.3 Measures

3.3.1 TTR & Guiraud's Index

The type-token ratio (TTR) is the proportion of unique words (types) in a text and is intended to measure vocabulary richness. However, as the size of the text under scrutiny increases, the TTR is skewed downwards. Because the total number of words in each Harry Potter book fluctuates so dramatically (approximately 77,000 in book 1 to approximately 197,000 in book 7), a more refined measure than raw TTR was needed. Therefore, Guiraud's Index was used to reduce the distortion caused by the text length. This measure mitigates the effect of varying text lengths by using the square root of the number of tokens. It is calculated using the formula below (Fig. 1).

Figure 1
Guiraud's Index Formula.

$$\text{Guiraud's Index} = \frac{\text{Number of Unique Words (Types)}}{\sqrt{\text{Total Number of Words (Tokens)}}}$$

3.3.2 Hapax Legomena (% of types)

Hapax legomena (from Greek meaning "read only once") are words that occur only once in a text. Expressing them as a percentage of the total number of types indicates the proportion of the vocabulary made up of these unique items. A higher proportion of hapax legomena suggests a greater lexical variety. This measure offers additional insight into the lexical diversity of the books. While Guiraud's Index captures the breadth of vocabulary, hapax legomena highlight how frequently new or distinctive words are introduced.

3.3.3 Lexical Density

Lexical density refers to the proportion of content words (nouns, verbs, adjectives, and adverbs) to function words (such as articles, pronouns, and prepositions) in a text. A high lexical density score means that a text carries more information through its vocabulary, whereas a low score indicates a greater reliance on grammatical words and a lighter information load. By examining lexical density across the seven volumes of Harry Potter, this study can examine how Rowling's writing varies in terms of information richness. Lexical density is calculated as follows:

Figure 2
Lexical Density Formula

$$\text{Lexical Density (LD)} = \frac{\text{Number of Lexical Words}}{\text{Total Number of Words (Tokens)}} \times 100$$

3.3.4 Lexical Coverage

Lexical Coverage is a measure of the proportion (%) of words in a text that come from established frequency lists. It shows how much of a text is made up of high-frequency, everyday vocabulary compared to less common or more academic words. Lexical Coverage was measured in the present study using the first and second 1,000 words of the General Service List (GSL), the Academic Word List (AWL), and the total coverage across these three categories. By identifying trends in lexical coverage across the Harry Potter books, the aim was to find out whether they reflect the same patterns identified in adolescent language development toward increased lexical sophistication (see Section 2.2).

3.3.5 Percentage of Words with More than Two Syllables

This measure reflects the proportion of polysyllabic words, defined here as words containing more than two syllables in a text. Longer multisyllabic words tend to be less frequent and more abstract, thereby increasing the linguistic challenge for readers (Kearns & Hiebert, 2022). Using this measure provides an additional metric for tracking lexical complexity across the series. While this measure reflects word-level phonological and morphological complexity rather than syntactic structure, it is relevant to overall text difficulty and reading

development, as longer words increase processing demands and are typically associated with more sophisticated vocabulary (Rayner et al., 2012).

3.3.6 Verbal Elements per Sentence

This is a measure of the average number of verbs or verb phrases in each sentence of the text. Writing with a higher number of verbal elements per sentence typically uses more subordinate or coordinated clauses, reflecting syntactic sophistication and a denser expression of ideas. Examining verbal elements per sentence helps assess how Rowling's writing develops across the Harry Potter series in terms of the structural richness of her sentences, which has direct implications for text difficulty.

3.3.7 Gunning Fog Index (GFI)

The GFI is a readability measure that estimates the years of formal education a reader would need to easily understand a given text. It is calculated using the average sentence length and the proportion of complex words, defined as those containing three or more syllables. Other indices, such as the Flesch Reading Ease and the Flesch–Kincaid Grade Level, also combine sentence length and word complexity; however, they rely primarily on syllable counts and produce more general measures of readability. In contrast, the Fog Index is more sensitive to the introduction and increasing frequency of longer, less common words, which is directly relevant to tracking the growth in lexical sophistication observed in the Harry Potter texts.

4. Results & Discussion

The following sections include a discussion of the results of the linguistic analysis of the Harry Potter books, organized according to the key measures outlined in the earlier methodology section: lexical diversity, lexical density and coverage, syntactic complexity, and readability.

Before examining specific measures of lexical and syntactic complexity, it is important to establish the fundamental quantitative statistics for each Harry Potter volume. These are shown in Table 1 below, including the total tokens (word count), total types (unique words), and the raw type-token ratio (TTR) for each book. The size of each corpus increases substantially across the series, with Book 1 containing around 77,860 words, and Book 5, the longest volume, containing over 257,000 words. Although Book 7 is shorter than Book 5, it still contains almost 197,500 words. The number of unique words (types) also increases from 6,150 in Book 1 to 12,909 in Book 5. This is indicative of the increased text length and wider vocabulary range. Type-token ratio (TTR) declines across the volumes from 0.08 in Book 1 to 0.05-0.06 in the later books, reflecting the inverse relationship between text length and TTR (see Section 3.3.1).

Table 1
Basic Corpus Statistics for the Harry Potter Series

Book	Total Tokens	Total Types	Type-Token ratio (TTR)
1	77,680	6,150	0.08
2	84,955	7,267	0.09
3	104,848	7,737	0.07
4	190,926	10,788	0.06
5	257,213	12,909	0.05
6	175,866	10,472	0.06
7	197,456	11,873	0.06

4.1 Lexical Diversity

The range and scope of words in a text represent its lexical diversity. A high lexical diversity score indicates a wide-ranging vocabulary, whereas a low score suggests a narrower range with greater repetition of the same words. Examining lexical diversity across the series shows how Rowling's writing varies book-by-book in terms of varied word choice. The following sections discuss the results of the Guiraud's index and hapax legomena measures of lexical diversity.

4.1.1 TTR & Guiraud's Index

The results for Guiraud's Index for all seven books are presented in Table 2 below:

Table 2
Results for Guiraud's Index

Book	1	2	3	4	5	6	7
Guiraud's Index	22.07	24.93	23.89	24.69	25.45	26.86	26.72

Guiraud's Index values show a steady upward trend across all seven books. Book 1 has a score of 22.07, which rises to 26.72 by Book 7, representing an increase of about 4.5 points, or approximately 20%. Although there is a small dip in Book 3 (23.89) compared to Book 2 (24.93), and again from Book 6 (26.86) to Book 7 (26.72), a Pearson product–moment correlation analysis revealed a very strong and positive relationship between book number and Guiraud's Index ($r = 0.93$, $p < .01$), indicating a statistically significant upward trend in lexical diversity across the series.

This pattern suggests that the vocabulary in the books becomes more varied as the series progresses, which may reflect both the increasing narrative complexity of the books and the gradual shift toward more mature themes and readership expectations as the series progresses. Although an increase of approximately 4.5 points is not a dramatic leap, it is large enough to support claims of growing lexical diversity across the series. The relatively stable rise in Guiraud's index is consistent with the idea that later books demand a higher level of lexical knowledge from readers.

4.1.2 Hapax Legomena (% of types)

The results for hapax legomena as a percentage of types can be seen in Table 3 below:

Table 3
Results for Hapax Legomena (% of types)

Book	1	2	3	4	5	6	7
Hapax Legomena	44.11	44.82	42.98	41.5	39.45	43.89	43.83

The percentage of hapax legomena as a proportion of total types ranges between approximately 39% and 45% across the seven books. The relatively high proportion is consistent with the Zipfian distribution of lexical frequency in natural language (Zipf, 1949), whereby a small number of high-frequency words coexist with a long tail of low-frequency items. The results show that the Harry Potter books also make extensive use of unique vocabulary items.

4.2 Lexical Density

Table 4
Results for Lexical Density

Book	1	2	3	4	5	6	7
Lexical Density	0.078	0.086	0.073	0.055	0.048	0.063	0.059

The lexical density values for the seven books range from 0.048 to 0.086, with the highest density appearing in Book 2 (0.086) and the lowest in Book 5 (0.048). The results suggest some fluctuation in the proportion of content words across the series rather than a steady developmental trend. The relatively high density in the earlier books may reflect a more straightforward narrative style with the frequent introduction of new objects, characters, and settings, requiring more nouns and descriptive content. The dip in the middle books, especially Book 5, may reflect a shift toward a greater reliance on function words and the discourse of interpersonal interactions, leading to a lighter informational load per sentence. The lower density likely reflects stylistic changes rather than a simple decrease in reader difficulty. The later books show a slight increase again, though they do not reach the levels of the first two. Overall, the lexical density of the series appears variable, probably reflecting stylistic differences between books rather than a simple linear increase in informational load.

4.3 Lexical Coverage

Table 5
Results for Lexical Coverage (percentage values)

Book	1	2	3	4	5	6	7
GSL 1st 1-1000	78.28	75.9	75.74	76.84	76.96	77.72	77.79
GSL 2nd 1001-2000	6.81	6.83	7.02	6.49	6.46	5.99	6.07
AWL 570	0.43	0.66	0.89	1.04	1.11	1.2	1.12
Total COVERAGE	85.52	83.4	83.65	84.37	84.53	84.91	84.98

The results show a high and consistent reliance on the most frequent words in English, with the GSL 1st 1,000 accounting for around three-quarters of all tokens in every book. The GSL 2nd 1,000 words add a further 6–7%, while the AWL contributes a small but gradually increasing share, from 0.43% in Book 1 to just over 1% by Books 6 and 7. Taken together, the total coverage remained stable across the series, hovering between 83% and 85%.

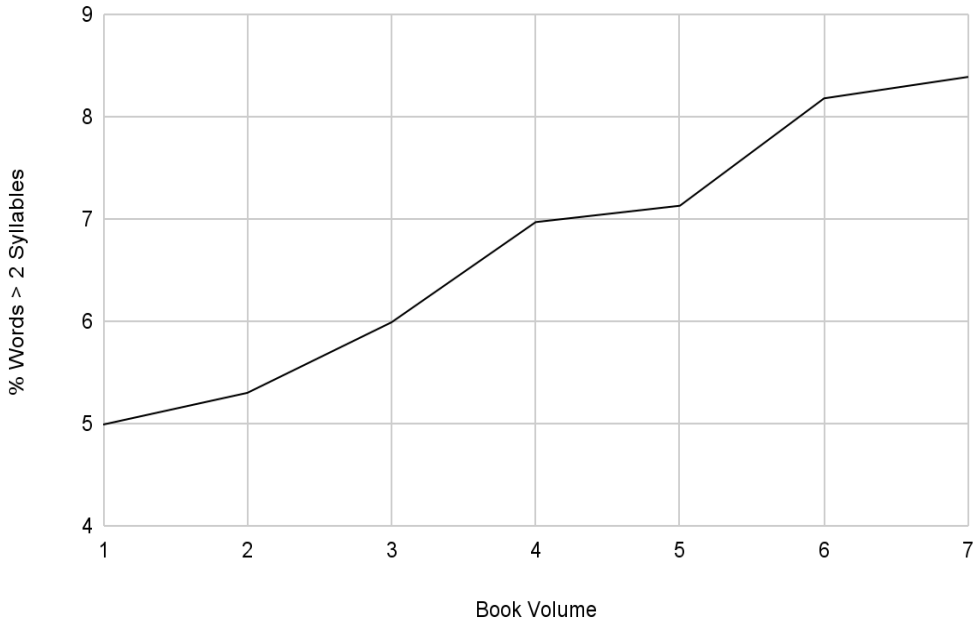
These findings suggest that Rowling's writing consistently draws on core high-frequency vocabulary, making the texts broadly accessible to readers, including younger adolescents. Simultaneously, the gradual rise in AWL coverage indicates a modest but tangible enrichment of vocabulary as the series progresses.

4.4 Sentence-level and Word-level Complexity

This section examines two complementary aspects of linguistic complexity: sentence-level syntactic features and word-level processing demands. Verbal elements per sentence provides an indicator of syntactic elaboration, while polysyllabic word percentage reflects the lexical and phonological complexity that affects reading difficulty. Together with the Gunning Fog Index, these measures chart the increasing cognitive demands placed on readers across the series. Due to the limitations of the computational tool used (Text Inspector), the following results are based on two 10,000-word sub-corpora for each book (see Section 3.2.4).

4.4.1 Words with more than two Syllables (%)

Figure 3
Results for words with more than two Syllables (%)



The percentage of words with more than two syllables shows a clear upward trend across the Harry Potter series, rising from 4.99% in Book 1 to 8.39% in Book 7. The correlation between book number and percentage of polysyllabic words was extremely strong and positive ($r = 0.99$, $p < 0.001$), representing the most robust linear trend observed in this study and confirming a clear progression toward greater word-level complexity. As the series progresses, vocabulary becomes more sophisticated, thus demanding greater word-level processing skill from the reader.

4.4.2 Verbal Elements per Sentence

Table 6
Results for Verbal Elements per Sentence

Book	1	2	3	4	5	6	7
Verbal Elements per Sentence	1.23	0.99	0.98	1.15	1.21	1.3	1.58

The results for Verbal Elements per sentence show a gradual increase across the series, from 1.23 in Book 1 to 1.58 in Book 7. Despite the initial dip in Books 2 and 3, the correlation between book number and verbal elements per sentence was strong and positive ($r = 0.79$, $p < 0.05$), indicating an overall statistically significant trend toward greater syntactic elaboration. From a readability perspective, this progression reflects a shift from simpler sentence structures in the earlier novels, which are more accessible to younger readers, towards more sophisticated and information-dense sentences in the later volumes.

4.4.3 Gunning Fog Index (GFI)

Table 7
Results for Gunning Fog Index (GFI)

Book	1	2	3	4	5	6	7
GFI	8.93	6.95	7.41	8.74	8.58	9.72	10.45

The GFI results show a general upward trend in readability difficulty across the volumes, rising from 8.93 in the first book to 10.45 in the last one. Although there are fluctuations between books, most notably a dip in Book 2 (6.95) and a rise again in Book 4 (8.74), the correlation between book number and the Gunning Fog Index was strong and positive ($r = 0.86$, $p < 0.05$), demonstrating a statistically significant increase in readability difficulty across the series. Scores in the range of 7–9 are commonly interpreted as corresponding to middle to early high school reading levels, while scores above 10 are generally associated with late high school or early postsecondary reading demands (Chall, 1983). The final book, with the highest score (10.45), indicates a noticeable increase in linguistic complexity compared to the beginning of the series. The Gunning Fog Index results highlight how the series introduces progressively more sophisticated language, supporting the interpretation that Rowling's vocabulary choice and style evolve in step with the increasing linguistic competence of her audience.

4.5 Limitations

This study has several limitations. Firstly, the Text Inspector's 10,000-word upload limit meant that only portions of the books could be analyzed in some cases, rather than the complete texts. Second, the syntactic measures employed in this study are surface-level and focus on sentence length and verbal density rather than deep structural features such as subordination ratios, clause embedding depth, or T-unit complexity. While more sophisticated syntactic parsing tools (e.g., TAASSC, Coh-Metrix) would provide a more comprehensive analysis of syntactic complexity, the surface-level measures used here (average words per sentence and verbal elements per sentence) offer systematic, replicable indicators that correlate with deeper syntactic elaboration (Nippold et al., 2008). It should also be noted that the percentage of polysyllabic words is a measure of word-level phonological complexity and processing difficulty, not syntactic complexity per se, though it contributes to overall text difficulty as reflected in readability indices. The choice to use surface-level metrics is the result of both practical constraints (availability of tools for large-scale corpus analysis) and the study's focus on general trends across the series rather than detailed structural analysis. Future research employing more sophisticated syntactic analysis tools would complement the present findings. Finally, it should be acknowledged that the books were written and published over a decade (1997-2007), meaning later books may reflect not only target audience maturation but also Rowling's own development as a writer and increasing editorial influence from publishers due to the success of the series.

5. Conclusion

5.1 Summary of Findings

This study examined the Harry Potter series using a range of lexical and syntactic measures to investigate whether the progression of the books reflects patterns of adolescent language development. The analysis considered lexical diversity, lexical density and coverage, sentence-level syntax, and readability. The results provide strong empirical support for the hypothesis that Rowling's language becomes more complex as the series progresses. Statistical analysis of the relationship between book number and key complexity measures revealed significant positive correlations across multiple dimensions of linguistic sophistication. The strongest correlation was observed for polysyllabic word percentage ($r = 0.99$, $p < 0.001$), representing an almost perfectly linear developmental trajectory in word-level complexity. Guiraud's Index, measuring lexical diversity, showed a very strong positive correlation ($r = 0.93$, $p < 0.01$), confirming genuine vocabulary diversification across the series. The Gunning Fog Index demonstrated a strong positive correlation ($r = 0.86$, $p < 0.05$), indicating systematic increases in readability difficulty. Finally, verbal elements per sentence showed a strong positive correlation ($r = 0.79$, $p < 0.05$), reflecting growing syntactic elaboration despite initial fluctuations in Books 2 and 3. These

correlation coefficients provide robust quantitative evidence that the observed trends represent statistically meaningful developmental progression rather than random variation.

Beyond these correlations, the findings also reveal several complementary indicators of linguistic maturation. Lexical diversity increases gradually but meaningfully across the series, with Guiraud's Index rising approximately 20% from Book 1 to Book 7, while hapax legomena remain consistently high (39-45%), indicating sustained lexical creativity throughout. Lexical coverage patterns show subtle but important shifts, with the proportion of academic vocabulary (AWL) more than doubling from 0.43% in Book 1 to 1.12% in Book 7, while maintaining stable coverage of high-frequency vocabulary (GSL 1st 1000: 75-78% across all books). This balance suggests that Rowling progressively introduces more sophisticated vocabulary while preserving accessibility for developing readers. Sentence-level complexity also increases systematically, with verbal elements per sentence rising from 1.23 to 1.58, and the proportion of polysyllabic words nearly doubling from 4.99% to 8.39%. The Gunning Fog Index rises from 8.93 to 10.45, corresponding to a shift from middle school to late high school reading levels.

Taken together, these findings align closely with established research on adolescent language development, which highlights continuing growth in the capacity to process longer, more demanding texts throughout the teenage years. The near-perfect linear progression in polysyllabic word usage is particularly striking and is compatible with the idea that Rowling may have calibrated linguistic difficulty to her readers' developing abilities, but this remains speculative. The Harry Potter series thus reflects a trajectory of linguistic complexity corresponding to the developmental path observed in adolescent learners, maintaining accessibility in the early volumes while gradually introducing greater lexical and syntactic sophistication.

5.2 Implications and Practical Applications

The findings of this study demonstrate that linguistic development in popular literature can be measured systematically through corpus-based methods, offering insights relevant to linguistics, education, and literary studies. The results show how lexical diversity, density, and syntactic complexity evolve in texts designed for adolescent readers, aligning with established developmental trajectories. For educators, the gradual rise in complexity across the series underscores the potential of popular fiction to scaffold literacy development, exposing readers to increasingly sophisticated vocabulary and syntax. For literary scholarship, the analysis provides evidence that Rowling's writing style matured in parallel with her audience, contributing to the cultural impact of the series.

The results also have practical value for English as a Foreign Language (EFL) educators. The Harry Potter series illustrates how vocabulary diversity, syntactic complexity, and readability can increase gradually across a sequence of texts while still maintaining engagement. This progression suggests that carefully chosen young adult literature may serve as a scaffold for language learners, offering accessible entry points in earlier volumes while introducing more advanced structures and less frequent vocabulary in later books. Teachers could use graded exposure to such series to support extensive reading programs, vocabulary expansion, and the development of reading fluency, making authentic literature both motivating and pedagogically effective.

5.3 Further Research

Future studies could build on this research in several ways. Firstly, more advanced syntactic tools such as TAASSC or Coh-Metrix could provide more detailed analyses of subordination, clause embedding, and syntactic variety. Such tools would address the limitations of the surface-level syntactic measures employed in the present study and could reveal more nuanced patterns of structural complexity across the series, including the ratio of coordinate to subordinate clauses, the depth of syntactic embedding, and the variety of sentence structures employed. Second, semantic or thematic approaches could examine how shifts in subject matter (e.g., death, morality, relationships) interact with linguistic complexity. Finally, future research could compare the present study's findings to other popular young adult series (e.g., The Hunger Games, Percy Jackson) to determine whether similar developmental trajectories exist across the genre or whether Harry Potter's pattern reflects Rowling's unique authorial choices.

REFERENCES

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. Monographs of the Society for Research in Child Development, 58(10), 1–186. https://doi.org/10.2307/1166112

Anthony, L. (2014). AntWordProfiler (Version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Berman, R. A., & Slobin, D. I. (1994). Relating events in narrative: A crosslinguistic developmental study. Lawrence Erlbaum Associates.

Berman, R. A., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities in speech and writing. Written Language & Literacy, 5(1), 1–44. https://doi.org/10.1075/wll.5.1.02ber

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press.

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. https://doi.org/10.5054/tq.2011.244483

Chall, J. S. (1983). Stages of reading development. McGraw-Hill.

Copyright Act, Act No. 48 of 1970, Article 47-7 (Japan). https://www.japaneselawtranslation.go.jp/en/laws/view/3848

Corson, D. (1997). The learning and use of academic English words. Language Learning, 47(4), 671–718. https://doi.org/10.1111/0023-8333.00025

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213–238. https://doi.org/10.2307/3587951

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. Modern Language Journal, 95(1), 1–20. https://doi.org/10.1111/j.1540-4781.2010.01138.x

Glinka, N., Zaichenko, Y., & Machulianska, A. (2021). Stylistic portrait of English fantasy texts including Harry Potter. Arab World English Journal.

Goyal, K. (2024). Calibre (Version 7.20) [Computer software]. https://calibre-ebook.com/

Halliday, M. A. K. (1985). Spoken and written language. Deakin University Press.

Jamal, R. F., & Nasrum, N. (2018). Language style used in J.K. Rowling's Harry Potter and the Cursed Child. Elite: English and Literature Journal, 5(2), 190-200.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. Working Papers in Linguistics, 53, 61–79.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Allyn & Bacon.

Kearns, D. M., & Hiebert, E. H. (2022). The word complexity of primary-level texts. TextProject. https://textproject.org/wp-content/uploads/2022/07/Kearns-and-Hiebert.pdf

Lexical Computing Ltd. (n.d.). Text Inspector [Computer software]. Retrieved September 22, 2025, from https://textinspector.com

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Medan, P. A., Safani, I., Permatasari, J., & Silitonga, D. (2024). Discourse analysis of character development in J.K. Rowling's Harry Potter series: Linguistic and narrative techniques. International Journal of Society Reviews.

Nippold, M. A. (1998). Later language development: The school-age and adolescent years. Pro-Ed.

Nippold, M. A. (2000). Language development during the adolescent years: Aspects of pragmatics, syntax, and semantics. In M. L. Rice & S. F. Warren (Eds.), Developmental language disorders: From phenotypes to etiologies (pp. 111–139). Erlbaum.

Nippold, M. A. (2007). Later language development: School-age children, adolescents, and young adults (3rd ed.). Pro-Ed.

Nippold, M. A., Mansfield, T. C., & Billow, J. L. (2008). Development of complex syntax: More than a growth of length. Journal of Speech, Language, and Hearing Research, 51(3), 685–698.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. Scientific Studies of Reading, 11(4), 357–383. https://doi.org/10.1080/10888430701530730

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. Scientific Studies of Reading, 10(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Reuneker, A. (2017). Lexical Diversity Measurements. Retrieved 17 September 2025, from https://www.reuneker.nl/files/ld.

Ricketts, J., Lervåg, A., Dawson, N., Taylor, L. A., & Hulme, C. (2020). Reading and Oral Vocabulary Development in Early Adolescence. Scientific Studies of Reading, 24(5), 380-396. https://doi.org/10.1080/10888438.2019.1689244

Rowling, J. K. (2012). Harry Potter and the philosopher's stone. Pottermore Publishing. (Original work published 1997). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the chamber of secrets. Pottermore Publishing. (Original work published 1998). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the Prisoner of Azkaban. Pottermore Publishing. (Original work published 1999). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the goblet of fire. Pottermore Publishing. (Original work published 2000). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the Order of the Phoenix. Pottermore Publishing. (Original work published 2003). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the half-blood prince. Pottermore Publishing. (Original work published 2005). Kindle edition.

Rowling, J. K. (2012). Harry Potter and the Deathly Hallows. Pottermore Publishing. (Original work published 2007). Kindle edition.

Sinclair, S., & Rockwell, G. (2016). Voyant Tools [Computer software]. Retrieved September 22, 2025, from https://voyant-tools.org/

Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), Applications of linguistics (pp. 443–452). Cambridge University Press.

West, M. (1953). A general service list of English words. Longman.

Zhao, W., & Wang, Y. (2018). Owls in Harry Potter: A functional stylistic study in children's literature. In Advances in Social Science, Education and Humanities Research.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley.

Author's Information:
Name: Robert John Ashcroft
Faculty: Tokai University Sapporo Campus
Email: bob.ashcroft@tokai.ac.jp