A Stylometric Analysis of Stylistic and Linguistic Evolution in Children's Literature (1700–Present)

Topics in Natural Language Processing | Ben Gurion University of the Negev | Spring 2025

Students Names:
Taymaa Rfaii, ID:212605737
Mas Wattad, ID:318975539
Ghazal Nobany, ID:211401245

Abstract

This study examines the stylistic and linguistic evolution of children's literature from the 18th to the 21st century. By analyzing a curated corpus of fifteen influential books, we investigate how sentence structure, vocabulary, and narrative style have changed over time. Using natural language processing and stylometric techniques, we measure shifts in average sentence length, syntactic complexity, lexical richness, and thematic emphasis. The results reveal a clear transition from didactic, formal language in earlier works to simpler, clearer, and more accessible storytelling in modern texts. While moral instruction has declined, emotional tone has remained relatively stable, with a growing emphasis on inclusive themes such as identity, imagination, and empathy. These changes reflect a broader societal shift toward making children's literature more engaging, balanced, and supportive for young readers.

Introduction

Children's literature has evolved significantly over the centuries. Initially used as a tool for moral and religious instruction and became a diverse and dynamic literary genre. This shift reflects changes in how society views childhood, education, and the role of storytelling.

In the 18th century, works such as A Little Pretty Pocketbook (1744) marked the emergence of literature specifically written for young readers. Influenced by Enlightenment ideals that emphasized rationality, self-discipline, and social virtue. The language in these books matched the ideas of the time—they used long and difficult sentences, hard words, and a serious tone. Authors usually wrote in third person, reinforcing adult authority and aiming to teach children proper behavior and moral values.

By the 19th century, particularly during the Victorian era, children's books began incorporating elements of fantasy and imagination. Lewis Carroll's *Alice's Adventures in Wonderland* (1866) exemplifies this shift, portraying childhood as a time of creativity, wonder, and play. The language became more playful, using jokes, made-up words, and experimental sentence structures that matched children's imagination and the way they think about language.

In the 20th century, the changes in children's books became even more noticeable. Stories were no longer formal or focused only on teaching rules and morals. Writers began using simpler language ,shorter sentences, everyday words, and a more personal tone which helped children understand the stories better and connect with the characters. At the same time, the topics became deeper and more emotional. Books started to explore ideas like identity, feelings, and growing up. Authors like Roald Dahl and J.K. Rowling wrote stories that were fun and full of adventure, but also meaningful and easy for children to relate to.

The 21st century continues this trend. Modern books such as *Wonder* (2012) explore topics like disability, bullying, and belonging, using inclusive and culturally aware language. These stories often combine simple sentence structures with meaningful vocabulary to support themes like empathy, kindness, and social inclusion.

Overall, the development of children's literature both in style and language shows how it has adapted over time to reflect changing cultural, educational, and social priorities.

Based on this background, our main research question is:
How has children's literature evolved over time in the way it's written, the words it uses, and the themes it explores?

Our hypothesis is that children's literature has progressively become simpler in language, more emotionally expressive, and increasingly inclusive in thematic focus over time. Specifically, we expect to observe shorter and clearer sentences, easier and more repetitive vocabulary, and a thematic shift away from moral instruction toward topics emphasizing identity, empathy, and belonging. These anticipated changes reflect the evolving educational priorities and cultural values aimed at making literature more accessible, engaging, and relevant to contemporary young readers.

Corpus

This study is based on a selected corpus of 15 well-known and widely read children's books published between the 18th and 21st centuries. The collection begins with A Little Pretty Pocket-Book (1744), often recognized as the first book specifically aimed at children, and ends with Wonder (2012), a contemporary novel centered on empathy and inclusion. The selected texts reflect a wide range of genres, styles, and historical contexts, including moral tales, adventure stories, fantasy, and modern realistic fiction. Each book was chosen for its cultural impact and popularity in its time, providing meaningful insight into the ways children's literature has developed across different historical periods.

List of Works in the Corpus

- **A Little Pretty Pocket-Book** (1744): Considered the first children's book, it combines rhymes, games, and moral lessons to encourage good behavior in young readers.
- **Goody Two-Shoes** (1766): A moral story about a poor orphan girl who rises through virtue and education, reflecting the didactic style of early children's books.
- **The History of Sandford and Merton** (1783): Presents contrasting characters to promote moral education and social responsibility, common in Enlightenment-era children's writing.
- **The King of the Golden River** (1841, John Ruskin): An early fantasy tale with moral undertones, showing the triumph of kindness and generosity over greed.
- **Alice's Adventures in Wonderland** (1866, Lewis Carroll): A surreal fantasy that plays with logic and language, capturing Victorian fascination with childhood imagination and nonsense.
- **Heidi** (1880, Johanna Spyri): Follows a young girl's life in the Swiss Alps, emphasizing nature, kindness, and the emotional development of children.
- **The Adventures of Pinocchio** (1883, Carlo Collodi): A classic Italian tale about a wooden puppet's journey to become a real boy, rich with moral lessons and transformation.
- **Treasure Island** (1888, Robert Louis Stevenson): A coming-of-age adventure involving pirates and treasure, widely credited with shaping the modern adventure genre for young readers.
- **The Wonderful Wizard of Oz** (1900, L. Frank Baum) :A beloved American fantasy featuring magical lands and characters, celebrating imagination and self-discovery.
- **Peter and Wendy** (1911, J.M. Barrie): Explores themes of childhood, escapism, and the fear of growing up through the adventures of Peter Pan and the Darling children.
- **Charlie and the Chocolate Factory** (1965, Roald Dahl): A whimsical story with dark humor, following a poor boy's visit to a magical chocolate factory, offering satirical critiques of greed and misbehavior.
- **Matilda** (1988, Roald Dahl): Tells the story of a gifted girl who overcomes neglect and injustice, focusing on intelligence, strength, and empowerment.
- **The Philosopher's Stone** (1997, J.K. Rowling): Begins the Harry Potter series, blending fantasy and school life into a magical world that captivated a generation of young readers.
- **Coraline** (2002, Neil Gaiman): A dark fantasy in which a brave girl faces a sinister parallel world, reflecting modern themes of identity, fear, and courage.
- **Wonder** (2012, R.J. Palacio): A contemporary novel about a boy with a facial difference navigating school life, promoting empathy, kindness, and inclusion.

Methodology

To study how children's literature has changed over time, we used a combination of natural language processing (NLP) and stylometric techniques. The analysis focused on measuring specific features across 15 children's books published between the 18th and 21st centuries.

Text Collection and Preparation

The texts were collected from public domain sources such as Project Gutenberg. All files were converted into plain text format and cleaned to remove extra formatting, page numbers, and non-story content (like introductions or tables of contents). Each book was labeled by its title and publication year.

Grouping by Period

To better observe trends over time, the books were grouped into historical periods:

- **18th century** (1701–1799)
- **19th century** (1800–1899)
- **20th century** (1900–1999)
- **21st century** (2000+)

This helped us compare language and style changes across different eras.

Tools and Libraries

The analysis was done using Python and several widely used libraries for natural language processing and data analysis:

- **NLTK**: was used for core NLP tasks including tokenization, POS tagging, lemmatization, named entity recognition, sentiment analysis, and analyzing word usage patterns.
- **pandas** – for organizing and processing text data in tables.
- **matplotlib** and **seaborn** – for creating graphs and visualizing trends.
- **scikit-learn (sklearn)** – used for various analytical and modeling tasks, including:
  - **TfidfVectorizer** to convert text into numerical representations for stylistic comparison.
  - **CountVectorizer** to prepare raw text for topic modeling.
  - **Latent Dirichlet Allocation (LDA)** to extract dominant themes across books.
  - **cosine_similarity** to measure similarity between texts or topic distributions.
- **textstat :** Readability metrics like Flesch score, Dale-Chall score, complex words.

Features Analyzed

The following stylometric and linguistic features were analyzed to track the evolution of writing style in children's literature from the 18th to 21st centuries:

Word-Level Features:

This study conducted word-level analysis across four historical periods to uncover lexical and stylistic trends in children’s literature. The selected features reflect vocabulary complexity, diversity, and emotional or moral emphasis:

- **Average Word Length** – Used as a proxy for lexical complexity, reflecting how dense or sophisticated the vocabulary is.
- **Flesch Reading Ease Score** – Measures text readability based on sentence and word length, with higher scores indicating easier texts.
- **Dale–Chall Score** – Assesses readability by factoring in sentence length and the use of unfamiliar words, with lower scores indicating more accessible vocabulary.
- **Complex Word Ratio** – Indicates the proportion of multisyllabic words, reflecting lexical and cognitive difficulty.
- **Unique Word Ratio** – Measures vocabulary diversity by calculating the proportion of distinct words used in the text.
- **Moral Word Ratio** – Quantifies the use of ethically themed words, indicating didactic or instructional intent in the narrative.
- **Emotion Word Ratio** – Measures the frequency of emotionally expressive words, capturing affective tone and engagement.
- **Sentiment Scores (positive, negative, neutral)** – Analyze the emotional polarity of the text to assess its overall tone and affective balance.
- **Most Frequent Words (nouns, verbs, adjectives)** – Identifies recurring thematic elements and stylistic patterns by analyzing word usage frequency.
- **Noun Categories: Person-related Nouns** – Measures the frequency of human-centric references, indicating character focus.
- **Noun Categories: Artifact-related Nouns** – Reflects the prevalence of references to human-made objects, signaling material and setting details.
- **Adjective Frequency** – Captures descriptive density and narrative vividness by measuring how often adjectives are used.

Note: To capture thematic and stylistic differences across eras, emotional and moral content was measured using custom lexicons. Two manually compiled word lists one for emotional terms and one for moral terms, were used to calculate their relative frequency in each text. While not exhaustive, the lists include a broad range of relevant expressions, expanded with context-specific additions from children’s literature, providing a consistent basis for tracking effective and ethical language across the corpus.

Sentence-Level Features:

This study analyzed sentence structures across four historical periods to identify stylistic changes in children’s literature. These features were chosen for their ability to reflect writing complexity, tone, and readability:

- **Average Sentence Length** – Indicates overall syntactic complexity.
- **Long Sentence Ratio (≥15 words)** – Shows use of extended sentence structures.
- **Short Sentence Ratio (≤7 words)** – Highlights preference for brevity and directness.
- **Complex Sentence Ratio** – Measures how complicated sentences are, based on long words and multiple parts in a sentence.
- **Exclamatory Sentence Count** – Measures emotional tone and expressive style.
- **Question Sentence Count** – Reflects interactive tone, dialogue, or direct address.

Using average values and percentage-based measures allowed for fair comparisons between texts of different lengths and volumes.

Topic Modeling Features:

To explore how themes in children’s books have changed over time, we used a technique called topic modeling to identify common patterns in what the stories are about. Specifically, we applied Latent Dirichlet Allocation (LDA), which automatically groups related words and ideas into topics based on how they appear across the texts.
Here’s what we looked at:

- **Main Topic per Book** – Identified dominant theme (e.g., magic, morality, friendship).
- **Topics Changes Over Time** – We compared the most common themes in each decade to see how the focus of children’s stories changed through history.
- **Similarities Between Books** – By comparing books’ topics, we measured how close or different they are from one another, showing which stories share similar ideas.
- **Theme Shifts Across Eras** – We tracked how themes like “nature” or “family” faded or became more popular as time went on.
- **Themes by Era** – For each time period, we counted how many books focused on each theme. This helped us understand what mattered most to authors (and readers) during different times.
- **Story Categories** – Each theme was grouped into broader story types (like “Adventure Stories” or “Moral/Educational Stories”) to see how types of storytelling evolved.

This helped us map the evolution of storytelling trends and cultural values in children's literature.

Besides just looking at these features to see how writing changed over time, we also used them for some of the analysis we did later in the project. For example, we used cosine similarity to compare how similar the writing was between different time periods. We also used the features to train machine learning models, like SVM, to predict which era a book belongs to based on its writing style. More details about these steps are explained in the following sections.

Results

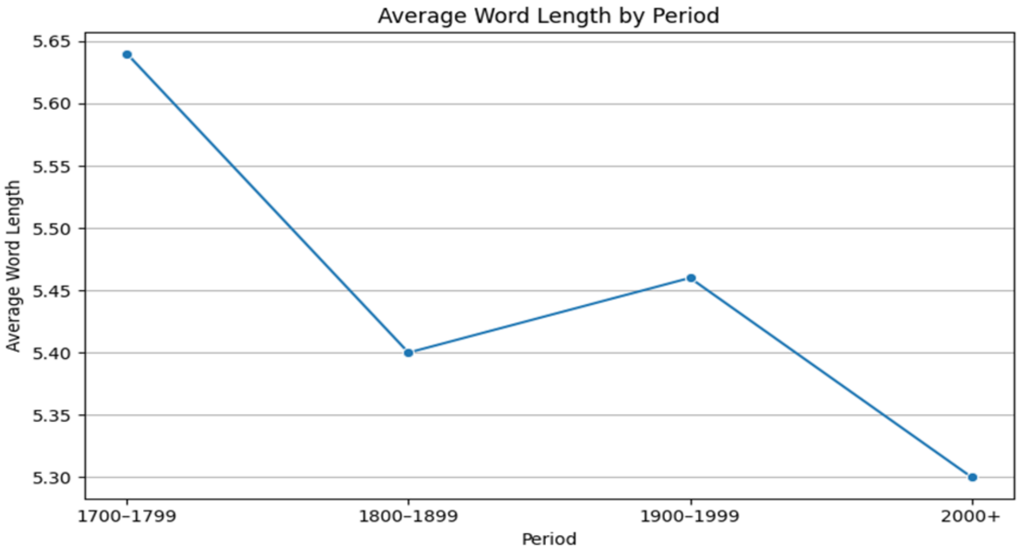
In this section, we present the main findings of our study using a series of charts and tables. The results are divided into three main parts: sentence analysis, vocabulary analysis, and topic modeling. Each part shows how the examined features have changed across four historical periods (1700–1799, 1800–1899, 1900–1999, and 2000 and beyond), allowing us to trace the evolution of writing style and language in children's literature over time.

Word Analysis:

Average word length:

average word length	year	book
5.353126	1744	A little pretty pocket-Book
5.529951	1766	Goody two-shoes
6.049411	1783	The history of Sandford and merton
5.711008	1841	The king of the golden river
5.235214	1866	Alices Adventures in Wonderland
5.461995	1880	HEIDI
5.368095	1883	The Adventure of Pinocchio
5.362873	1888	Treasure island
5.288504	1900	THE WONDEFUL WIZRD OF OZ
5.434122	1911	Peter and Wendy
5.402528	1965	Charlie And the Chocolate Factory
5.515968	1988	matilda
5.488014	1997	The Philosopher Stone
5.436798	2002	coraline
5.163619	2012	Wonder

Figure 1 : average word length by period



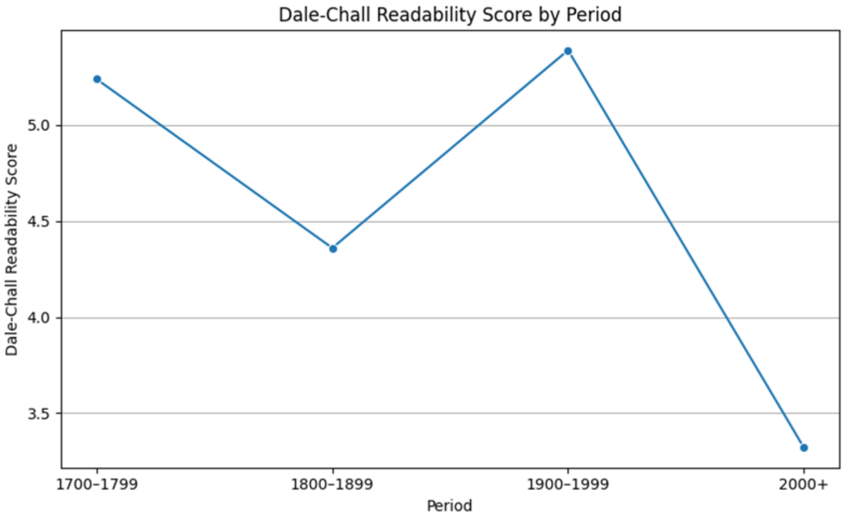
The average word length in the selected children's books ranges from 5.16 to 6.05 characters. Higher averages are found in earlier texts, while lower values appear in more recent ones. Although individual variation is present, a gradual decrease in average word length is visible over time.

Readability scores:

flesch score	dale chall	year	book
78.89	6.59	1744	A little pretty pocket-Book
67.93	6.78	1766	Goody two-shoes
55.81	2.34	1783	The history of Sandford and merton
67.59	6.67	1841	The king of the golden river
78.99	5.79	1866	Alice's Adventures in Wonderland
82.75	1.44	1880	HEIDI
83.56	5.09	1883	The Adventure of Pinocchio
77.57	5.59	1888	Treasure island
77.67	1.57	1900	THE WONDERFUL WIZARD OF OZ
81.12	5.62	1911	Peter and Wendy
85.49	5.33	1965	Charlie And the Chocolate Factory
84.57	5.45	1988	Matilda
75.5	5.15	1997	The Philosopher Stone
84.37	5.27	2002	Coraline
85.18	1.36	2012	Wonder

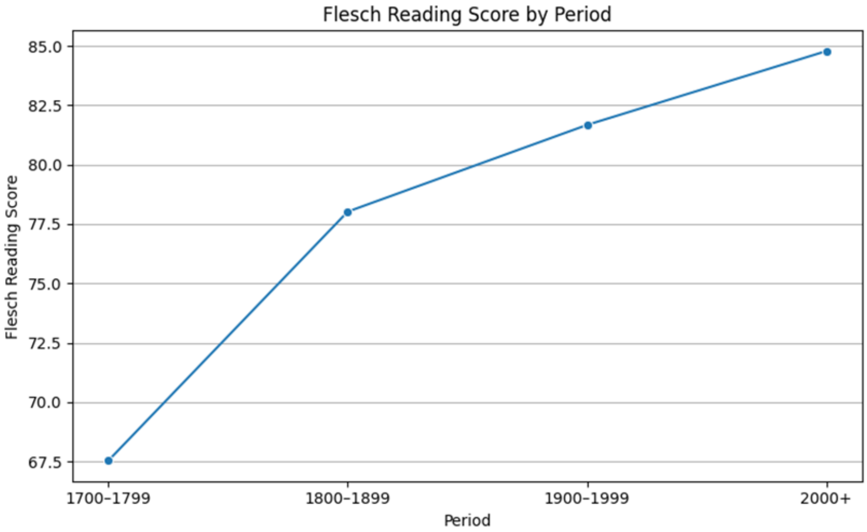
The table presents readability scores for a selection of children’s books published between 1744 and 2012, using the Dale–Chall and Flesch Reading Ease formulas. Dale–Chall scores in the dataset range from 1.36 to 6.78, while Flesch scores range from 55.50 to 85.56. Earlier texts generally show higher Dale–Chall values and more varied Flesch scores, while later books tend to have lower Dale–Chall scores and higher Flesch Reading Ease values.

Figure 2: Dale-Chall Readability score by period



The figure presents the average per period of the Dale-Chall Readability score. We can see that some variation is present, but we can see that there is a decrease in average of the Dale-Chall Readability score especially in recent years .

Figure 3: Flesch Reading Score by period



The figure presents the average per period of Flesch Reading Score but we can see a steady increase of the average of Flesch Reading Score overtime .

Complex and unique word:

complex word ratio	unique word ratio	year	book
0.16	0.355	1744	A little pretty pocket-Book
0.163	0.305	1766	Goody two-shoes
0.082	0.115	1783	The history of Sandford and merton
0.186	0.372	1841	The king of the golden river
0.094	0.214	1866	Alice's Adventures in Wonderland
0.074	0.139	1880	HEIDI
0.075	0.161	1883	The Adventure of Pinocchio
0.096	0.181	1888	Treasure island
0.058	0.14	1900	THE WONDERFUL WIZARD OF OZ
0.116	0.202	1911	Peter and Wendy
0.089	0.176	1965	Charlie And the Chocolate Factory
0.106	0.197	1988	matilda
0.067	0.121	1997	The Philosopher Stone
0.092	0.194	2002	coraline
0.066	0.118	2012	Wonder

Figure 4: Average per period of the ratio of complex words from total words

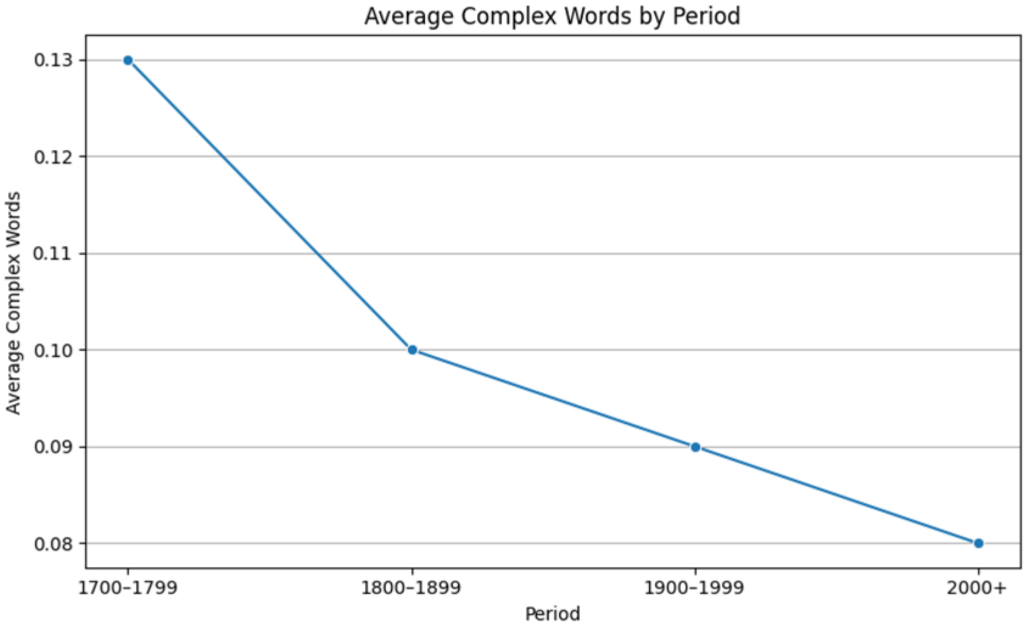
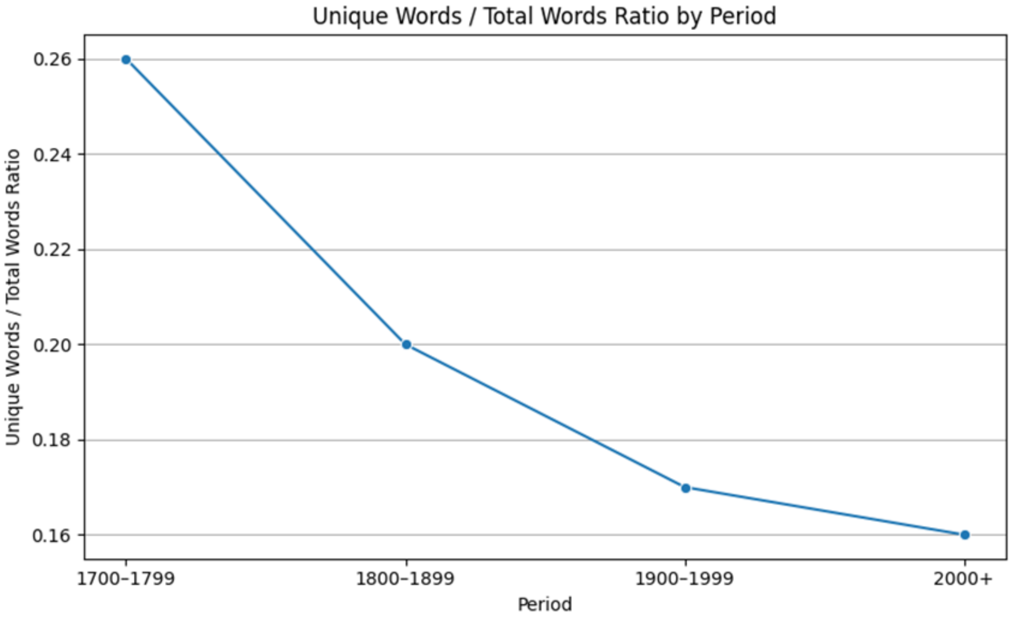


Figure 5: Average per period of the ratio of unique words from total words

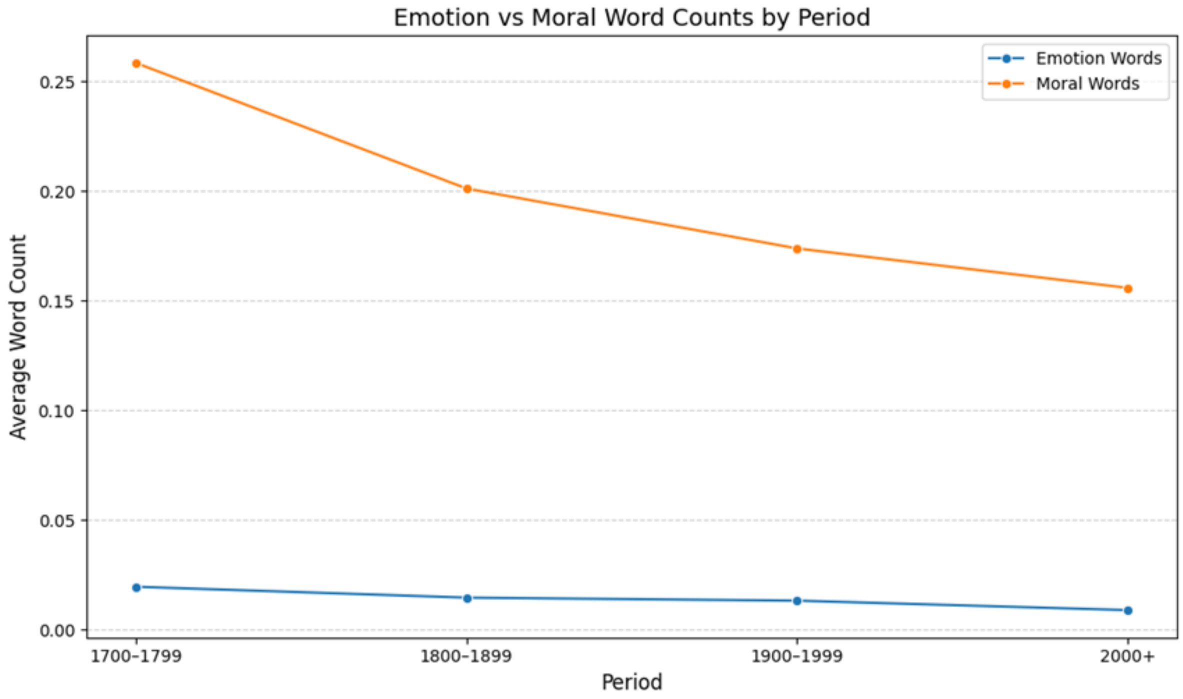


The figures show a clear decline in both the unique word ratio and the proportion of complex words in children’s books from the 18th century to the present.

Moral and emotion words:

ratio of words of morals	ratio of words of emotion	year	book
0.355	0.018	1744	A little pretty pocket-Book
0.305	0.025	1766	Goody two-shoes.
0.115	0.016	1783	The history of Sandford and merton
0.372	0.008	1841	The king of the golden river
0.214	0.014	1866	Alice's Adventures in Wonderland
0.139	0.019	1880	HEIDI
0.161	0.022	1883	The Adventure of Pinocchio
0.181	0.011	1888	Treasure island
0.14	0.015	1900	THE WONDERFUL WIZARD OF OZ
0.202	0.019	1911	Peter and Wendy
0.176	0.016	1965	Charlie And the Chocolate Factory
0.197	0.009	1988	matilda
0.121	0.009	1997	The Philosopher Stone
0.194	0.008	2002	coraline
0.118	0.01	2012	Wonder

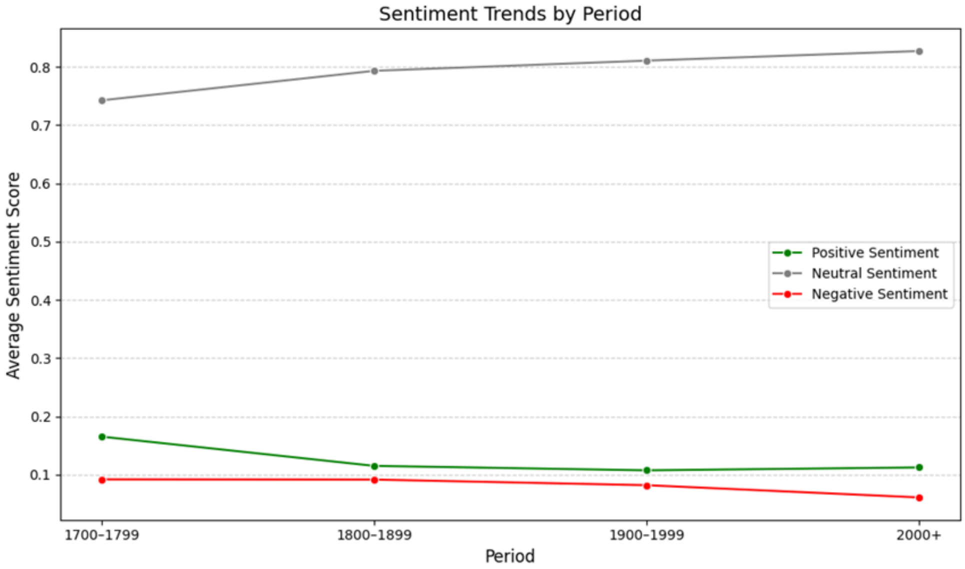
Figure 6: Average per period of the ratio of moral words from total words and the ratio of emotional words from total



The figure presents average word frequencies for emotion and moral terms across four historical periods in children's literature. Moral word frequency is highest in the 1700–1799 period and shows a steady decline across subsequent periods, reaching the lowest average in the 2000+ category. In contrast, emotional word use remains consistently low but relatively stable, with a slight decline in the 21st century.

sentiment:

Figure 7: Average per period of sentiment trends



The figure presents average sentiment scores positive, negative, and neutral across four historical periods in children's literature. Neutral sentiment consistently records the highest values, increasing steadily from around 0.74 in the 1700–1799 period to over 0.82 in the 2000+ period. Positive sentiment starts near 0.17 in the earliest period, decreases through the 1800s and 1900s, and then shows a slight rise in the most recent period, reaching just above 0.11. Negative sentiment remains low across all periods, with minor fluctuations and a small decrease in the 2000+ category to approximately 0.06.

Most recurring words:

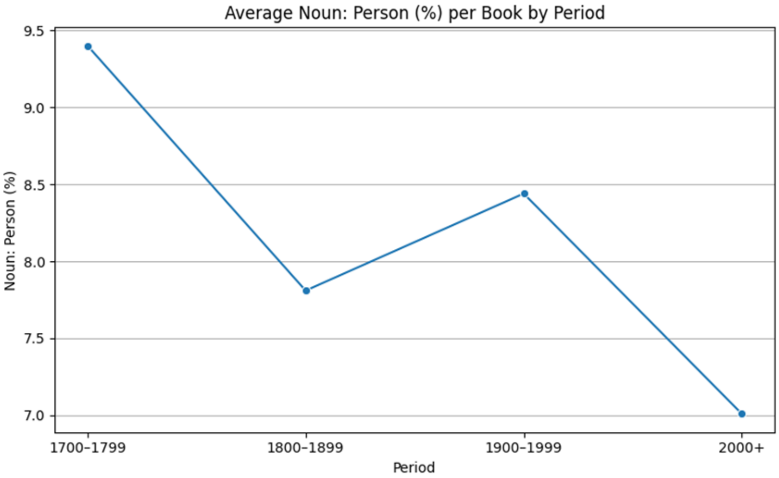
3 Most frequent adjectives	3 Most frequent verbs	3 Most frequent nouns	Year	Book
Little - 132 Thy - 107 Pretty - 91	Make-23 Take - 17 go-17	Pocketbook-87 Play - 39 Thing- 37	1744	A little pretty pocket-Book
Little- 90 Good- 57 Poor - 47	Made – 44 Say - 41 came- 35	Mr. - 64 Man - 50 Margery- 49	1766	Goody two-shoes
Little - 510 Great - 297 Poor - 233	Said-482 Made- 176 Found- 151	Mr. -417 time-330 man- 322	1783	The history of Sandford and merton
Old - 45 little- 38 golden - 32	Said - 68 Went- 23 looked - 21	Gluck- 52 Water- 44 River- 42	1841	The king of the golden river

Little-133 Great-39 large: 30	said, 459 went, 82 know, 57	Alice- 217 Thing- 85 Time- 75	1866	Alice's Adventures in Wonderland
Old- 185 Little- 168 Many- 87	Said-314 Come- 157 Go- 156	Heidi- 471 Child- 216 Peter- 214	1880	HEIDI
Little- 168 Poor-122 Good- 100	Said- 225 Go- 85 Want- 75	Pinocchio- 316 marionette- 196 Boy- 107	1883	The Adventure of Pinocchio
Good - 121 old - 116 last - 106	Said- 341 Say-140 See- 109	Man- 251 Hand- 221 Doctor- 202	1888	Treasure island
Great- 143 Little- 133 Green- 104	Said- 335 Asked- 112 Came- 103	Lion- 174 Illustration- 146 scarecrow- 136	1900	THE WONDERFUL WIZARD OF OZ
Little- 100 Last- 59 first: 52	Said- 356 Cried-116 Darling- 98	Peter-336 Wendy- 150 hook- 147	1911	Peter and Wendy
Little- 103 Old-61 Golden-54	Said- 339 Cried- 108 Go- 95	Wonka- 266 Mr.- 240 Charlie-158	1965	Charlie And the Chocolate Factory
Little- 92 Small- 60 Good- 54	Said- 572 Asked - 82 got - 76	Honey- 381 Matilda-358 Trunchbull- 144	1988	matilda
Last-83 First - 81 Good - 79	Said- 794 Rowling- 342 Got-199	Harry-1319 Potter- 430 stone- 414	1997	The Philosopher Stone
Black- 76 Old- 63 Little-60	Said-391 Went- 99 Looked- 74	Coraline- 451 Mother-136 Hand- 105	2002	coraline
Little-173 Good- 102 Big- 88	Said – 907 Know-272 Say-179	Mom- 402 School- 261 Jack- 217	2012	Wonder

Word Categories:

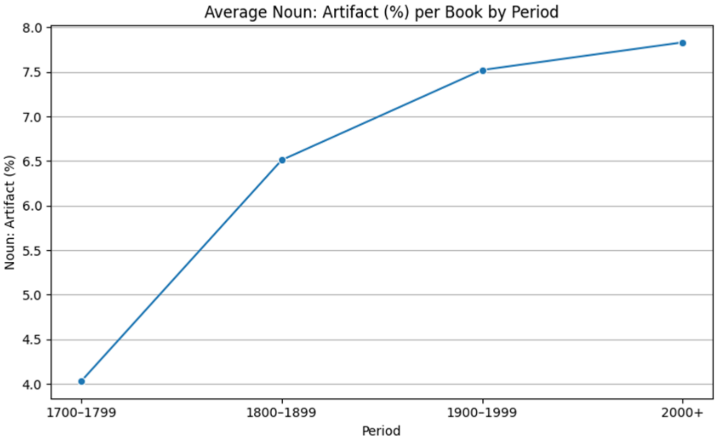
adv.	adj	noun Artifact	noun Person	Total words	year	book
207	370	200	422	4783	1744	A little pretty pocket-Book
340	495	341	914	8180	1766	Goody two-shoes
3931	5476	2658	5805	70835	1783	The history of Sandford and merton
207	331	302	431	5187	1841	The king of the golden river.txt
678	542	680	736	13137	1866	Alice's Adventures in Wonderland
1268	1517	1376	2463	24405	1880	HEIDI
792	1014	1336	1333	18699	1883	The Adventure of Pinocchio
1455	1673	2489	2606	32375	1888	Treasure island
909	1187	1393	1414	18381	1900	THE WONDERFUL WIZARD OF OZ
1515	1212	1636	2286	21904	1911	Peter and Wendy
723	990	1397	1070	16061	1965	Charlie And the Chocolate Factory
971	1071	1320	1867	19602	1988	Matilda
1867	1899	3084	3062	42966	1997	Book 1 - The Philosopher's Stone
620	722	1535	860	14865	2002	Coralineב
1816	1613	1968	3039	36909	2012	Wonder

Figure 8: Average per period of the ratio of nouns from the total word number



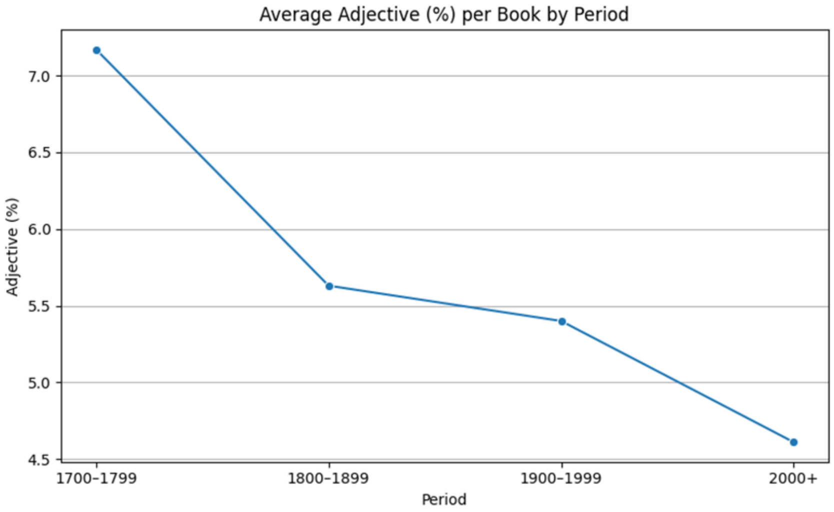
The graph shows the average percentage of person-related nouns per book across four time periods. In the 1700–1799 period, the percentage was highest at approximately 9.4%. This value dropped significantly in the 1800–1899 period to about 7.8%. In the 1900–1999 period, there was a moderate increase to around 8.4%. Finally, in the 2000+ period, the percentage decreased again to the lowest point of about 7.0%.

Figure 9: Average per period of the ratio of artifact nouns from the total word number



The graph shows the average percentage of artifact-related nouns per book across four time periods. Artifact nouns refer to words that denote man-made objects such as tools, clothing, machines, or other physical items created by humans. In the 1700–1799 period, these nouns appeared least frequently, at around 4.0%. The percentage rose sharply in the 1800–1899 period to about 6.5%, continuing increasing to approximately 7.5% in the 1900–1999 period, and reached its highest point in the 2000+ period at around 7.8%.

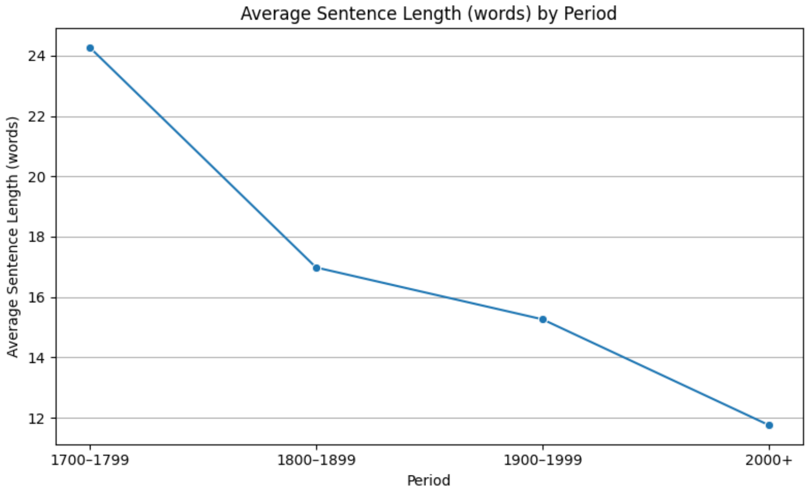
Figure 10: Average per period of the ratio of adjectives from the total word number



The graph shows the average percentage of adjectives per book across four time periods. In the 1700–1799 period, adjectives appeared most frequently, with a percentage slightly above 7.0%. This percentage dropped sharply in the 1800–1899 period to around 5.6%. A smaller decline continues into the 1900–1999 period, reaching approximately 5.4%. The lowest usage is observed in the 2000+ period, when the percentage of adjectives falls to around 4.6%.

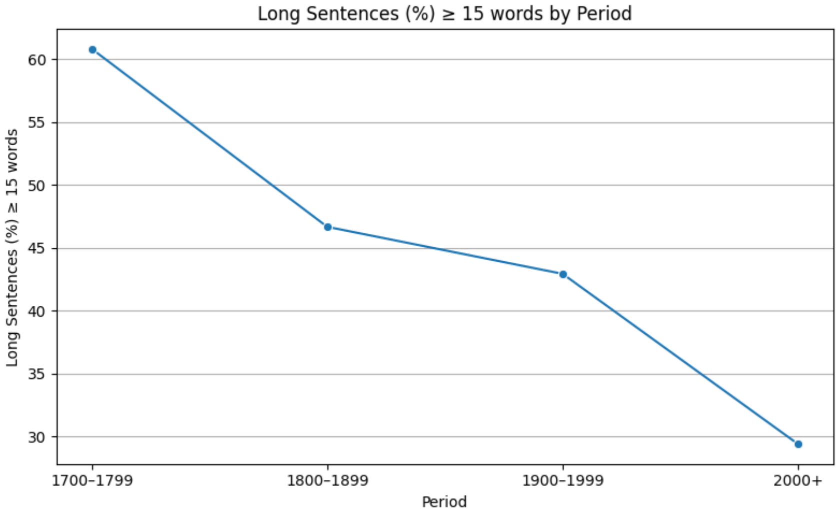
Sentence Analysis:

Figure 11: Average sentence length by period



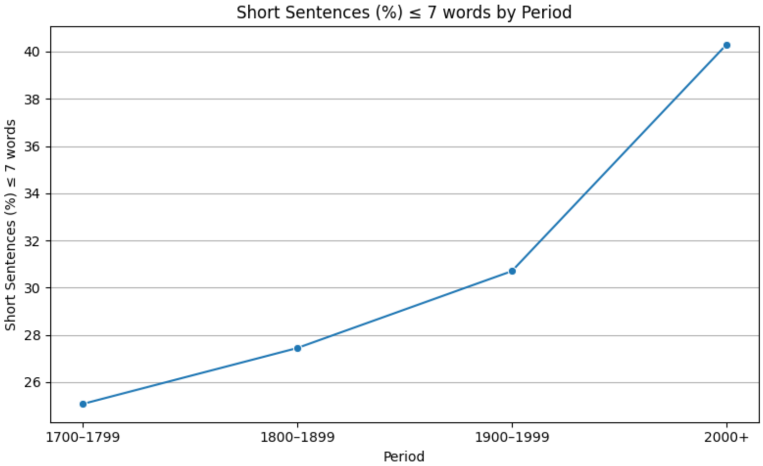
The figure shows a decline in average sentences across historical periods. Sentences averaged approximately 24 words in the 1700–1799 period, decreasing steadily to about 12 words in works published after 2000.

Figure 12: Proportion of long sentences (≥15 words) by period



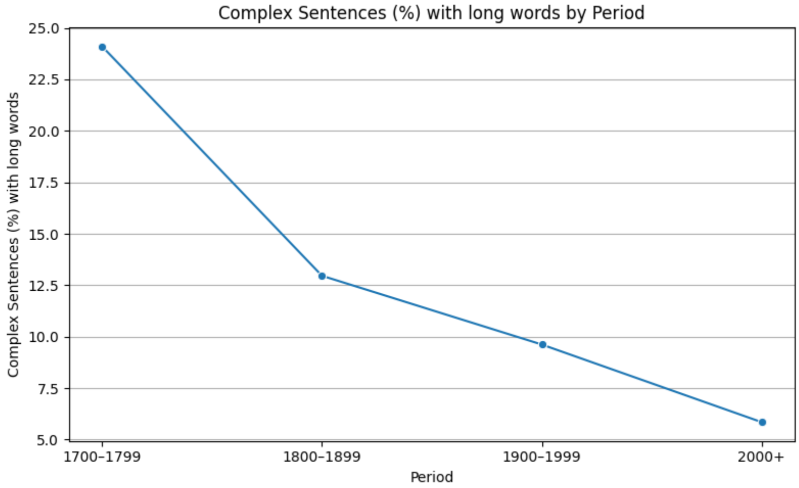
The figure shows a decline in long sentences from approximately 61% in 1700–1799 to approximately 29% in 2000+.

Figure 13: Proportion of short sentences (≤7 words) by period



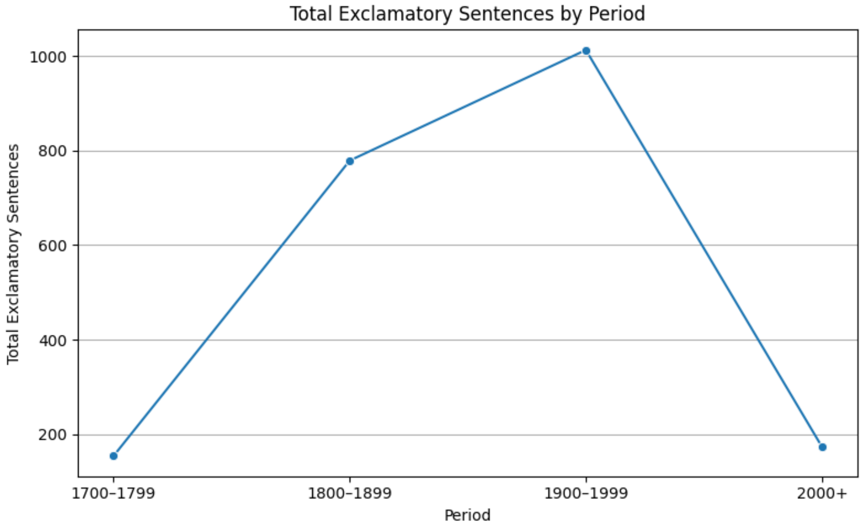
The figure shows an increase in short sentences from approximately 25% in 1700–1799 to approximately 40% in 2000+.

Figure 14: Proportion of complex sentences by period



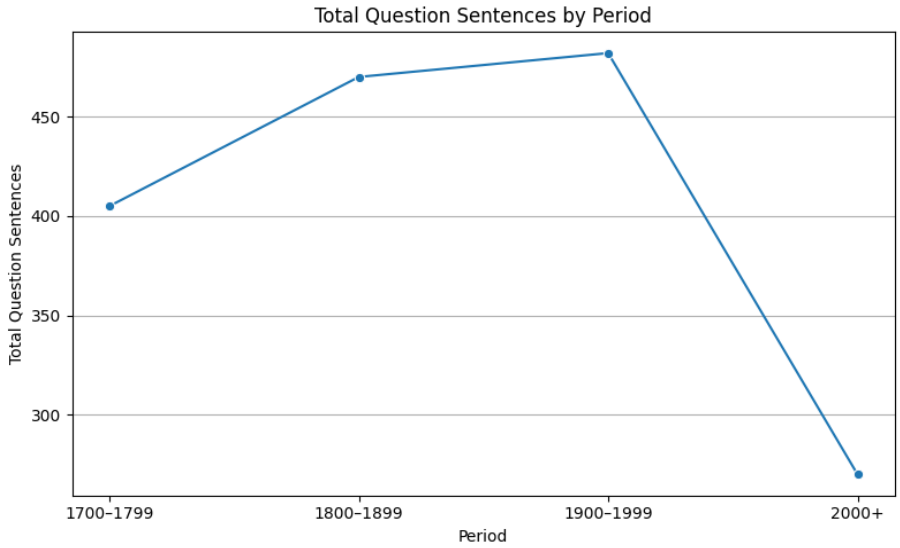
The figure shows a decline in complex sentences from approximately 24% in 1700–1799 to approximately 6% in 2000+.

Figure 15: Total exclamatory sentences by period



The figure shows an increase in exclamatory sentences, peaking at approximately 1,000 in 1900–1999, followed by a sharp decline in 2000+.

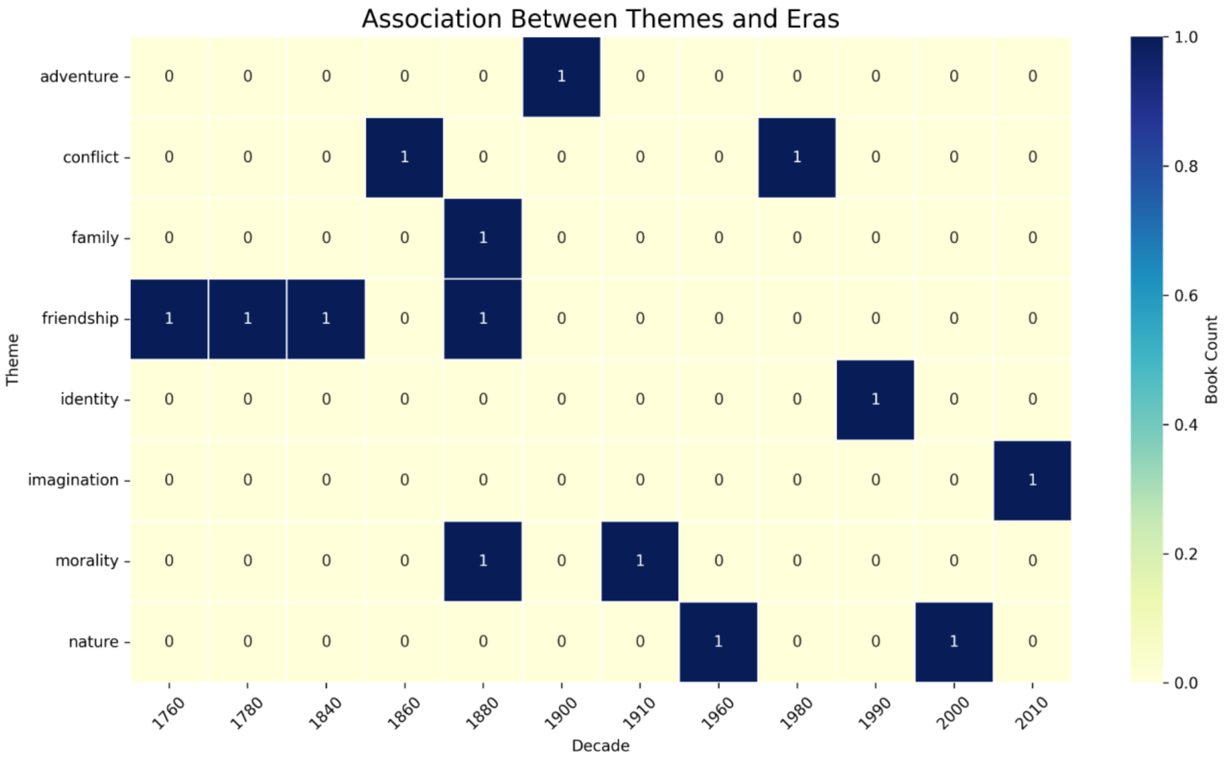
Figure 16: Total question sentences by period



The figure shows a rise in question sentences, peaking at approximately 480 in 1900–1999, followed by a sharp decline to around 270 in 2000+.

Topic Modeling:

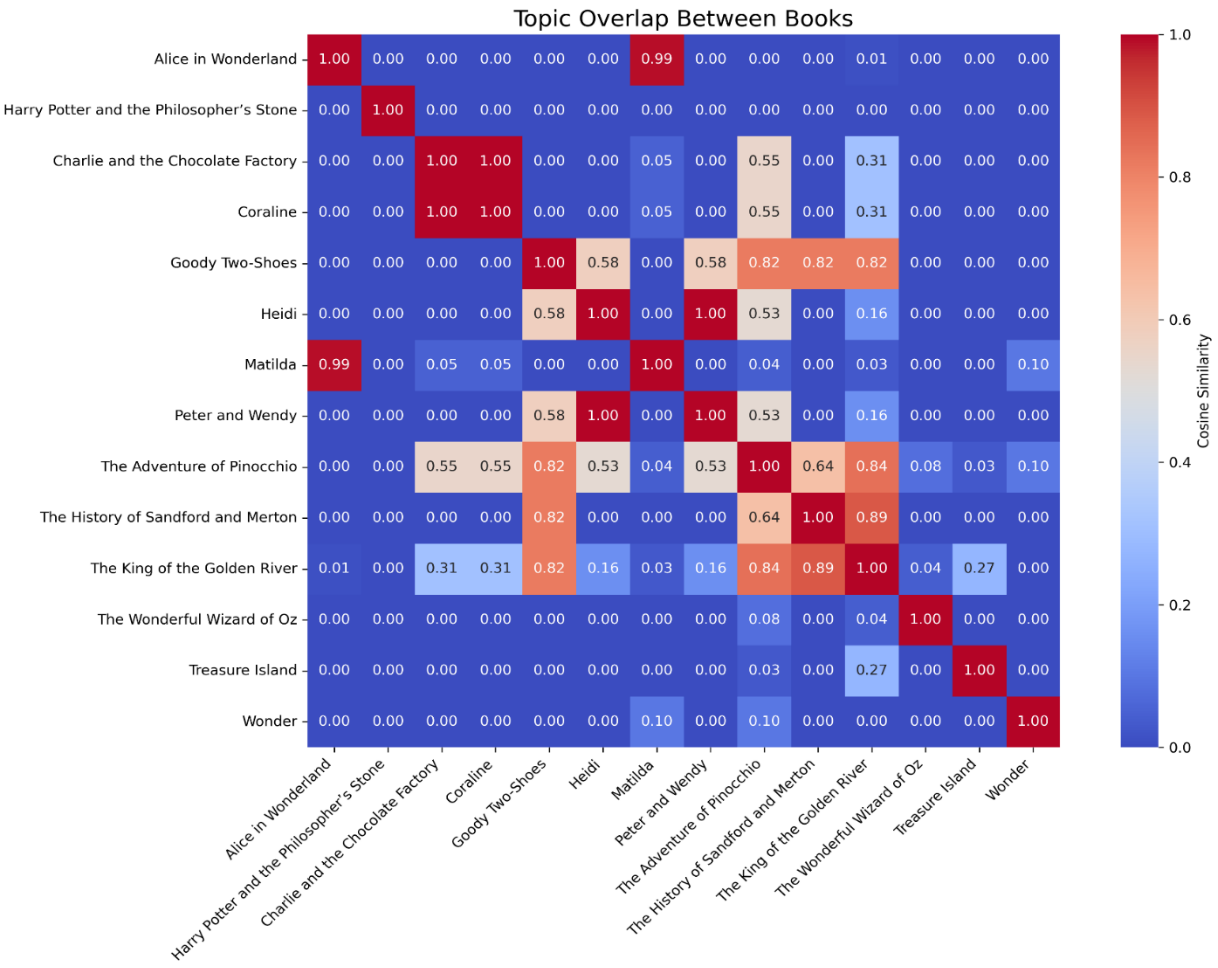
Figure 17: Association Between Themes and Eras



Theme	1760	1780	1840	1860	1880	1900	1910	1960	1980	1990	2000	2010
adventure	0	0	0	0	0	1	0	0	0	0	0	0
conflict	0	0	0	1	0	0	0	0	1	0	0	0
family	0	0	0	0	1	0	0	0	0	0	0	0
friendship	1	1	1	0	1	0	0	0	0	0	0	0
identity	0	0	0	0	0	0	0	0	0	1	0	0
imagination	0	0	0	0	0	0	0	0	0	0	0	1
morality	0	0	0	0	1	0	1	0	0	0	0	0
nature	0	0	0	0	0	0	0	1	0	0	1	0

The chart shows which themes were most common in children's books across different decades. In the 1700s and 1800s, friendship was a very popular theme, showing up in many books. The 1880s had a wide mix of themes like adventure, family, morality, and conflict, making it one of the most diverse periods. In the early 1900s, morality continued to be an important theme. Later, in the 1960s and 2000s, stories with nature became more common. In the 1990s, conflict was a key theme, and by the 2010s, imagination stood out as the main focus. This shows how children's stories changed over time, moving from teaching lessons and relationships to more creative and emotional themes.

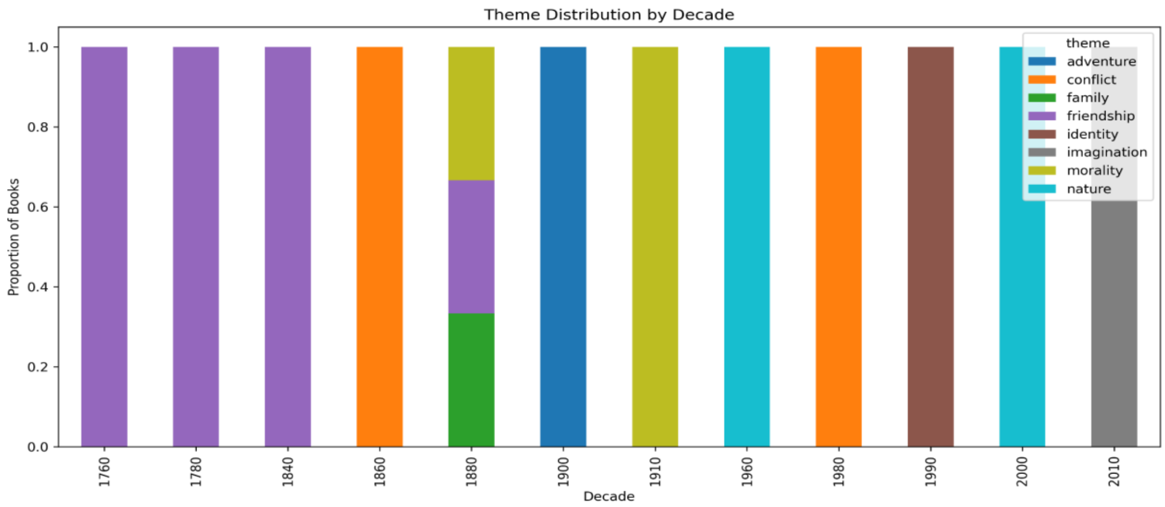
Figure 18: Topic Overlap Between Books



	Alice in Wonderland	Harry Potter and the Philosopher's Stone	Charlie and the Chocolate Factory	Coraline	Goody Two-Shoes	Heidi	Matilda	Peter and Wendy	The Adventure of Pinocchio	The History of Sandford and Merton	The King of the Golden River	The Wonderful Wizard of Oz	Treasure Island	Wonder
Alice in Wonderland	1.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Harry Potter and the Philosopher's Stone	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Charlie and the Chocolate Factory	0.00	0.00	1.00	1.00	0.00	0.00	0.05	0.00	0.55	0.00	0.31	0.00	0.00	0.00
Coraline	0.00	0.00	1.00	1.00	0.00	0.00	0.05	0.00	0.55	0.00	0.31	0.00	0.00	0.00
Goody Two-Shoes	0.00	0.00	0.00	0.00	1.00	0.58	0.00	0.58	0.82	0.82	0.82	0.00	0.00	0.00
Heidi	0.00	0.00	0.00	0.00	0.58	1.00	0.00	1.00	0.53	0.00	0.16	0.00	0.00	0.00
Matilda	0.99	0.00	0.05	0.05	0.00	0.00	1.00	0.00	0.04	0.00	0.03	0.00	0.00	0.10
Peter and Wendy	0.00	0.00	0.00	0.00	0.58	1.00	0.00	1.00	0.53	0.00	0.16	0.00	0.00	0.00
The Adventure of Pinocchio	0.00	0.00	0.55	0.55	0.82	0.53	0.04	0.53	1.00	0.64	0.84	0.08	0.03	0.10
The History of Sandford and Merton	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.64	1.00	0.89	0.00	0.00	0.00
The King of the Golden River	0.01	0.00	0.31	0.31	0.82	0.16	0.03	0.16	0.84	0.89	1.00	0.04	0.27	0.00
The Wonderful Wizard of Oz	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.04	1.00	0.00	0.00
Treasure Island	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.27	0.00	1.00	0.00
Wonder	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.10	0.00	0.00	0.00	0.00	1.00

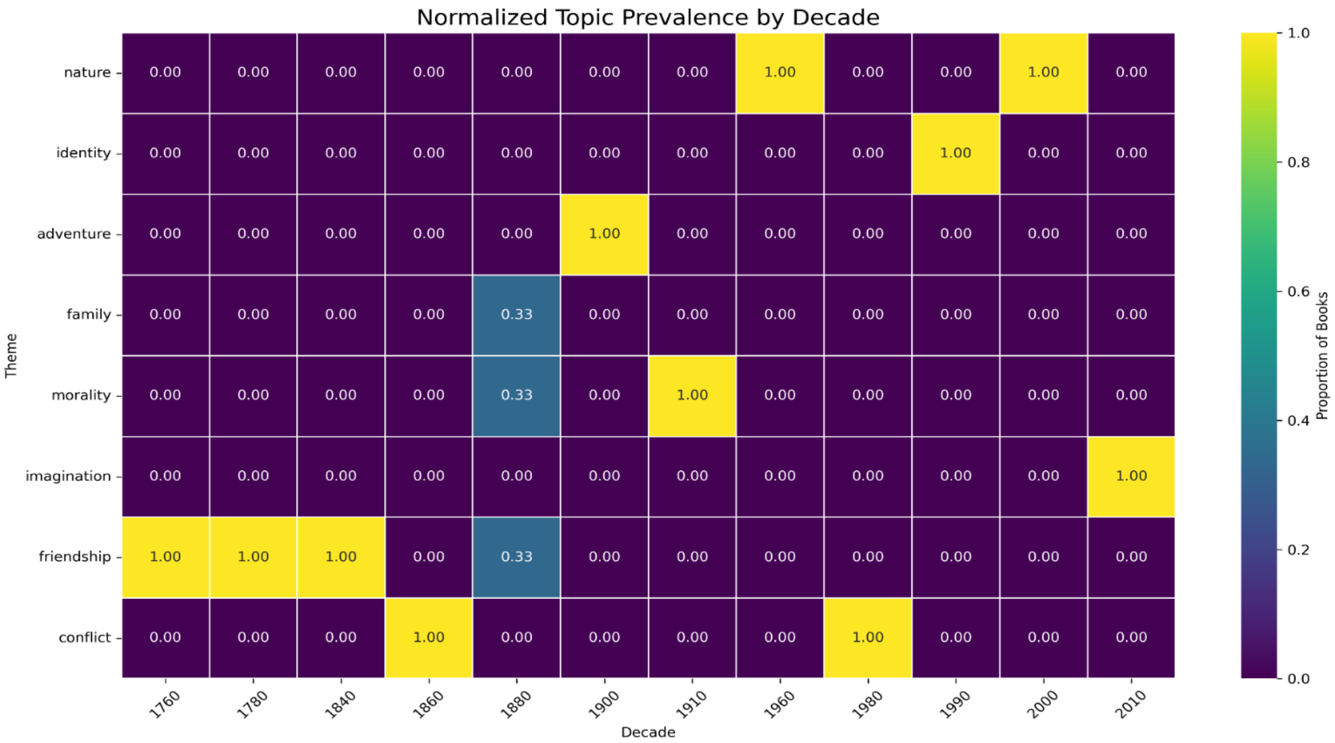
This heatmap shows how similar the books are in terms of topics. Red squares mean the books share similar themes, while blue means they are different. For example, *Matilda* and *Alice in Wonderland* are very similar (0.99), likely sharing imaginative themes. In contrast, *Wonder* and *Harry Potter* have almost no overlap, meaning they focus on different ideas.

Figure 19: Theme distribution by Decade



This chart shows how book themes changed over time. In the 1700s and early 1800s, most books focused on friendship. By the late 1800s, themes became more mixed, including family and morality. In the 1900s, adventure and morality stood out. From 1980 onward, conflict, identity, nature, and imagination became more common, showing a shift in the kinds of stories told to children.

Figure 20: Normalized Topic Prevalence by decade



Theme	1760	1780	1840	1860	1880	1900	1910	1960	1980	1990	2000	2010
nature	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
identity	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
adventure	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
family	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
morality	0.00	0.00	0.00	0.00	0.33	0.00	1.00	0.00	0.00	0.00	0.00	0.00
imagination	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
friendship	1.00	1.00	1.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
conflict	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00

This chart shows which themes were most common in each decade. In the 1700s and early 1800s, friendship was the top theme. In the late 1800s, family and morality became more present. The 1900s featured adventure and morality, while the 1960s and 2000s saw a rise in nature. Identity appeared strongly in the 1990s, and imagination peaked in the 2010s.

Book Themes and Story Categories

Title	Year	Dominant Topic	Theme	Story Category
Alice in Wonderland	1866	9	conflict	Fantasy/Fairy Tale Style
Harry Potter and the Philosopher’s Stone	1997	2	identity	Emotional/Realistic Stories
Charlie and the Chocolate Factory	1965	0	nature	Moral/Educational Stories
Coraline	2002	0	nature	Moral/Educational Stories
Goody Two-Shoes	1766	8	friendship	Fantasy/Fairy Tale Style
Heidi	1880	6	morality	Moral/Educational Stories
Matilda	1988	9	conflict	Fantasy/Fairy Tale Style
Peter and Wendy	1911	6	morality	Moral/Educational Stories
The Adventure of Pinocchio	1883	8	friendship	Fantasy/Fairy Tale Style
The History of Sandford and Merton	1783	8	friendship	Fantasy/Fairy Tale Style
The King of the Golden River	1841	8	friendship	Fantasy/Fairy Tale Style
The Wonderful Wizard of Oz	1900	4	adventure	Adventure Stories
Treasure Island	1888	5	family	Emotional/Realistic Stories

Wonder	2012	7	imagination	Fantasy/Fairy Tale Style

Book Similarity Table

	Alice in Wonderland	Harry Potter and the Philosopher’ s Stone	Charlie and the Chocolate Factory	Coraline	Goody Two-Shoes	Heidi	Matilda	Peter and Wendy	The Adventure of Pinocchio	The History of Sandford and Merton	The King of the Golden River	The Wonderful Wizard of Oz	Treasure Island	Wonder
Alice in Wonderland	1.0	0.0	0.0	0.0	0.0	0.0	0.99	0.0	0.0	0.0	0.01	0.0	0.0	0.0
Harry Potter and the Philosopher’s Stone	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Charlie and the Chocolate Factory	0.0	0.0	1.0	1.0	0.0	0.0	0.05	0.0	0.55	0.0	0.31	0.0	0.0	0.0
Coraline	0.0	0.0	1.0	1.0	0.0	0.0	0.05	0.0	0.55	0.0	0.31	0.0	0.0	0.0
Goody Two-Shoes	0.0	0.0	0.0	0.0	1.0	0.58	0.0	0.58	0.82	0.82	0.82	0.0	0.0	0.0
Heidi	0.0	0.0	0.0	0.0	0.58	1.0	0.0	1.0	0.53	0.0	0.16	0.0	0.0	0.0
Matilda	0.99	0.0	0.05	0.05	0.0	0.0	1.0	0.0	0.04	0.0	0.03	0.0	0.0	0.1
Peter and Wendy	0.0	0.0	0.0	0.0	0.58	1.0	0.0	1.0	0.53	0.0	0.16	0.0	0.0	0.0
The Adventure of Pinocchio	0.0	0.0	0.55	0.55	0.82	0.53	0.04	0.53	1.0	0.64	0.84	0.08	0.03	0.1

The History of Sandford and Merton	0.0	0.0	0.0	0.0	0.82	0.0	0.0	0.0	0.64	1.0	0.89	0.0	0.0	0.0
The King of the Golden River	0.01	0.0	0.31	0.31	0.82	0.16	0.03	0.16	0.84	0.89	1.0	0.04	0.27	0.0
The Wonderful Wizard of Oz	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.08	0.0	0.04	1.0	0.0	0.0
Treasure Island	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	0.27	0.0	1.0	0.0
Wonder	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	1.0

Application of Article-Based Methods

As part of the project, we explored two stylometry studies and applied methods from them to our own corpus.

1. Corpus Periodization Framework

To support the historical structure of our analysis, we used a method from the article *“Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus”* by Alsudais and Tchalian (2016). The idea behind this method is to check whether there are big differences in writing style between different time periods, based on the words used. It uses TF-IDF and cosine similarity to compare neighboring time segments and decide whether they should be merged or kept as separate periods.

We applied a simplified version of this method to our dataset by dividing the 15 books into four time periods based on their publication years:

- 1700–1799
- 1800–1899
- 1900–1999
- 2000 and later.

For each period, the text content of all books was concatenated into a single document. Then, we applied **TF-IDF vectorization on bigrams** to create feature vectors, and calculated **Cosine Similarity** between each pair of **consecutive time periods**.

Periods Compared	Cosine Similarity	Action
1700–1799 and 1800–1899	0.032	Keep
1800–1899 and 1900–1999	0.045	Keep
1900–1999 and 2000+	0.051	Keep

Here are the results:

The similarity scores between the different time periods were all much lower than 0.7. This means that the writing style and vocabulary were clearly different from one period to the next. Because of that, we didn’t combine any of the periods. These results support our idea that the style of children’s literature has changed significantly over time, and that it makes sense to treat each historical period as a separate and unique stage of writing.

2. Cross-Validated Evaluation of Machine Learning Models Trained on Stylometric Features for Classifying Children's Literature by Historical Era

The second article we applied in our project is *"Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts"* (Gómez-Adorno et al., 2018).

In this study, a stylometric approach was used to explore the evolution of children's literature from the 18th to the 21st century. Rather than dividing texts strictly by century, the corpus was grouped into three custom eras early (pre-1880), middle (1880–1964), and late (1965 onward) based on widely recognized shifts in children's publishing. This partition reflects recognized historical phases in children's literature: the early era, marked by moralistic and didactic prose; the middle era, shaped by the expansion of mass-market publishing and standardized storytelling; and the late era, defined by child-centered narratives and genre diversity. These phases correspond to stylometric differences in readability, function word usage, lexical richness, and syntactic complexity. The choice of eras was also supported by earlier sentence- and word-level analyses, which revealed patterns aligning with these divisions. Although the overall corpus was limited in size, the era-based grouping helped balance the dataset and enabled more meaningful stylistic comparisons. It also contributed to more consistent results in training and testing the machine learning models. (For example, if the partition were based strictly on centuries, the 21st century would include only two books, making it unreliable for training, as a single book per class would not provide sufficient data for accurate classification).

To evaluate these changes quantitatively, a set of stylometric features was extracted from a curated corpus of fifteen children's books. These features were selected for their ability to capture structural, lexical, and readability-related attributes of writing styles. The following features were extracted for each book:

- **Average sentence length:** Measures syntactic complexity by calculating the mean number of words per sentence.
- **Hapax legomena ratio:** The proportion of words that occur only once in the text, serving as an indicator of lexical diversity.
- **Type-token ratio:** A traditional measure of lexical richness, calculated as the number of unique words divided by the total number of words.
- **Function word ratio:** The proportion of stopwords (like the, is, and) relative to all words, which reflect syntactic and stylistic patterns independent of content.
- **Flesch reading ease score:** A readability metric that estimates how easy a text is to read, based on sentence length and word syllable count.
- **Dale-Chall readability score:** Another readability measure that factors in the use of complex words are not found on a familiar word list.
- **Complex word ratio:** The proportion of words identified as difficult by the TextStat library, used to assess vocabulary complexity.
- **Long sentence ratio:** The proportion of sentences is longer than 15 words, providing additional granularity in syntactic complexity.
- **Short sentence ratio:** The proportion of sentences shorter than 7 words, which may relate to simplicity or pacing of narrative.

- **Complex sentence ratio:** An estimated measure of syntactic intricacy, based on the number of sentences containing multiple conjunctions (like because and although).

These features were chosen for their relevance to stylometry, which quantifies linguistic style. Syntactic patterns are reflected in sentence length and complex sentence ratios, while lexical diversity is captured by type-token and hapax legomena ratios. Readability scores like Flesch and Dale-Chall assess cognitive accessibility shaped by stylistic choices. Function word usage and sentence length distributions further highlight stylistic consistency across eras, making these features effective for tracing historical shifts in children's literature.

All texts were preprocessed using standard natural language processing techniques, including sentence segmentation, tokenization, and removal of non-alphabetic tokens. Custom markers were employed to isolate the main textual content, helping ensure consistency across books and minimizing noise that could interfere with feature extraction.

Machine Learning Results and Analysis

To assess the discriminative power of stylometric features across eras, three supervised classification models Logistic Regression, Linear Support Vector Machine (SVM), and Random Forest were evaluated using stratified 5-fold cross-validation. This setup helped ensure that each era was evenly represented in both training and testing subsets.

The Linear SVM achieved the highest overall classification accuracy (73%), outperforming both Logistic Regression (67%) and Random Forest (60%). Notably, the SVM model exhibited balanced precision and recall across the three classes, with particularly strong performance on early and late texts (80% F1-score for each). Logistic Regression demonstrated high precision for early texts but lower recall, indicating it was more conservative in assigning this class. The Random Forest classifier, while occasionally effective, showed the greatest variability and weakest performance on middle-era texts, likely due to overfitting and a lack of distinctive decision boundaries for that category.

Model	**Accuracy**	**Early F1**	**Middle F1**	**Late F1**
Logistic Regression	0.67	0.75	0.55	0.73
Linear SVM	0.73	0.80	0.60	0.80
Random Forest	0.60	0.75	0.50	0.60

Feature Weight Interpretation on (SVM) classifier

Feature	Early	Middle	Late
Avg. Sentence Length	0.3902	-0.1411	0.5306
Hapax Legomena Ratio	0.3055	0.2935	0.2440
Type-Token Ratio	0.3160	0.2209	0.2542
Function Word Ratio	0.5770	-0.9224	-0.2225
Flesch Reading Score	-0.7257	0.4772	-0.5580
Dale-Chall Score	0.2213	0.7900	-0.0246
Complex Words	-0.0447	0.0283	-0.2062
Long Sentences Ratio	-0.0500	-0.3747	0.0612
Short Sentences Ratio	0.6759	0.1557	0.6269
Complex Sentences Ratio	0.5466	-0.3417	0.1579

Support Vector Machine (SVM) was chosen for feature interpretation as it achieved the highest classification accuracy among all tested models. Analyzing the feature weights offers insight into the linguistic characteristics that most distinguish the literary eras.

For the early era, higher values for function_word_ratio, short_sentences_ratio, and complex_sentences_ratio were strong positive indicators. This suggests early children's books relied heavily on function words and often used short but syntactically rich sentences. Additionally, a low flesch_reading score further points to reduced readability, reflecting more complex sentence structures or archaic vocabulary.

The middle era showed a distinct profile, with dale_chall and flesch_reading scores contributing most positively, indicating texts that were generally more readable and age appropriate. A notably low function_word_ratio helped distinguish this period, possibly reflecting a stylistic shift toward clearer, more direct narrative styles.

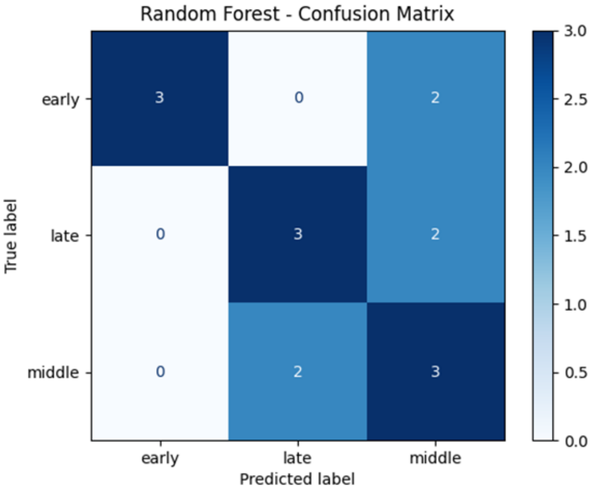
In the late era, higher avg_sentence_length and short_sentences_ratio emerged as key features, indicating a tendency to alternate between longer narrative passages and brief, simple sentences. A lower function_word_ratio and complex_words score also marked this era, suggesting simplified vocabulary and grammar, consistent with modern trends in children's literature aiming for accessibility and engagement.

These findings highlight how distinct stylometric patterns align with historical shifts in writing style and how SVM effectively captures these differences due to its high discriminative power.

Conclusion

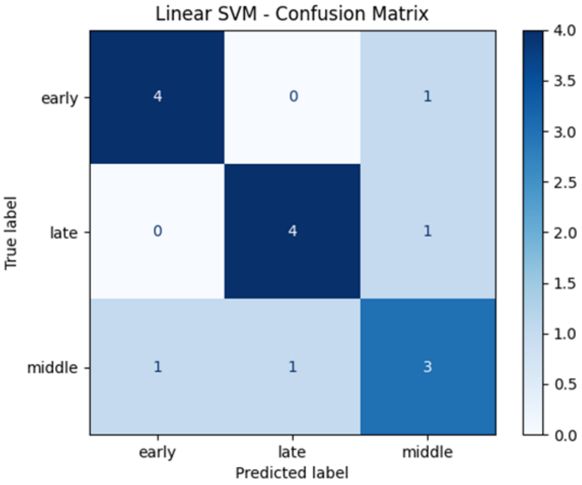
This study demonstrates that stylometric features capturing syntactic complexity, lexical diversity, and readability effectively reveal meaningful stylistic differences across early, middle, and late eras of children's literature. Despite the limited size of the corpus, clear distinctions were observed, particularly between early and late texts, while middle-era texts reflected a transitional profile. Although a larger dataset would allow for more nuanced generalization, the results indicate that even with a small but curated sample, era-based classification grounded in stylistic patterns offers valuable insights into the historical evolution of literary style. This supports the broader applicability of computational stylistics in literary analysis.

Random Forest – Confusion Matrix



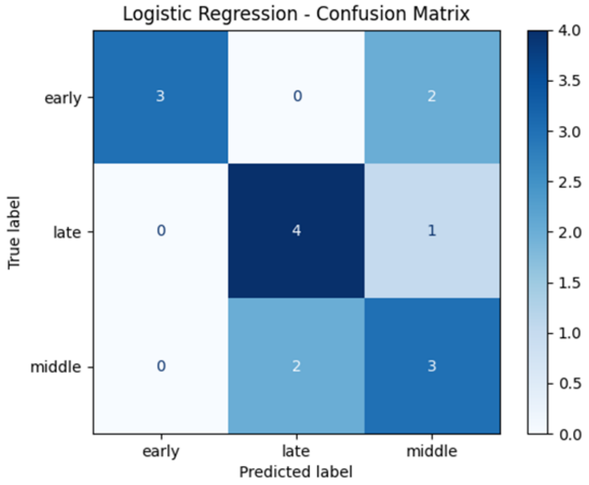
The confusion matrix for the Random Forest model highlights the gradual transition in linguistic style. While the model correctly identifies several early, middle, and late books, there is notable confusion between adjacent eras, especially between middle and both early and late. This pattern suggests that the stylistic features learned by the model do not have sharply defined boundaries consistent with a continuous stylistic shift over time. The model's errors are not random but reflect stylistic proximity, further reinforcing the idea that children's literature evolved incrementally in style rather than through discrete leaps.

Linear SVM – Confusion Matrix



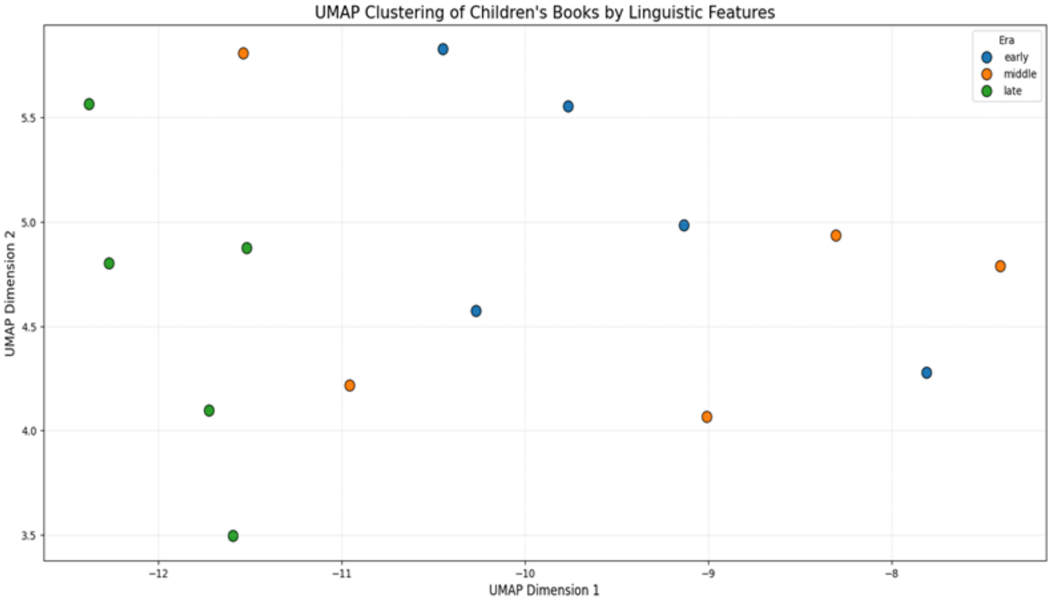
The Linear SVM model achieves high accuracy for early and late texts, with some confusion in the middle era. This suggests that the stylistic characteristics of early and late periods are more distinctive, while the middle era shares overlapping features with both. The middle books' misclassification as early or late is telling that it reflects a blended style consistent with a transitional phase in literary history. This result aligns well with the UMAP visualization and reinforces the notion that stylistic evolution in children's literature is progressive, with middle-era texts serving as a bridge between older and modern styles.

Logistic Regression – Confusion Matrix



Similar to the SVM, Logistic Regression captures the stylistic distinctions between early and late texts reasonably well but shows greater confusion around middle-era books. These results underscore a key finding: while early and late styles appear relatively stable and separable, the middle period exhibits more fluid and overlapping stylometric traits. This pattern reflects a gradual evolution in writing style, with the middle era displaying transitional features that span both earlier and later periods. The classification errors are informative; they reveal where stylistic boundaries are less defined and illustrate the progressive nature of stylistic change .

UMAP Clustering Children's Books by Stylometric Features Colored by Era:



This plot uses UMAP to reduce the dimensionality of stylometric features extracted from children’s books, projecting them into two dimensions for visualization. Each point represents a book, colored by its respective era: early (blue), middle (orange), or late (green). The visualization reveals that books from the late era form a relatively cohesive cluster, suggesting a shift toward more distinct stylometric patterns in recent decades. In contrast, early and middle era books show greater dispersion and partial overlap, reflecting a more gradual stylistic evolution over time. The middle era occupies an intermediate position, bridging the earlier and later styles. This supports the hypothesis that children's literature evolved gradually over time, with the middle era serving as a stylometric bridge.

Discussion:

This section discusses the key findings from the linguistic analysis of children's literature across several centuries. By interpreting our results, we aim to understand how children's books have changed over time in both language and thematic focus, reflecting shifts in educational priorities, cultural values, and the evolving needs of young readers.

Across the four centuries surveyed, a general decrease in average word length (Figure 1) indicates a shift toward simpler, more accessible syntax that improves readability and eases cognitive processing. Alongside this, a steady downward slope in unique-word ratio and complex-word ratio (Figures 4 and 5) reflects a move toward simpler, more repetitive vocabulary, which enhances clarity and reduces lexical difficulty. This trend is further confirmed by a general decrease in the Dale–Chall readability score (Figure 2) paired with a steady increase in the Flesch Reading Ease score (Figure 3), indicating that texts have become simpler through the use of more familiar words and clearer sentence structures, collectively improving overall accessibility and ease of reading. Moreover, across the historical periods surveyed, there is a steady decline in average sentence length (Figure 11) and in the proportion of long and complex sentences (Figures 12 and 14), which reveals a clear shift toward simpler, more concise sentence structures that enhance readability and reduce cognitive load. Concurrently, the rise in short sentences (Figure 13) reflects a growing preference for brevity and clarity, while fluctuations in exclamatory and question sentences (Figures 15 and 16) point to evolving narrative styles and changing patterns of emotional expression. Taken together, these linguistic trends demonstrate a broader move toward age-appropriate, accessible language and narrative clarity, which align closely with educational priorities and publishing practices focused on engaging young readers effectively with clear, accessible, and relatable texts.

When it comes to emotional and moral tone, our results show a steady decrease in the frequency of moral words(figure 6), while emotional word use remains consistently low but relatively stable. This pattern indicates a shift from the strong moral instruction of early children's literature toward a reduced emphasis on ethics. Overall, it reflects a broader transition from didactic storytelling to narratives that prioritize emotional engagement and empathy.

Sentiment analysis shows a steady rise in neutral tone, a slight decline in positive sentiment followed by modest recovery, and a gradual decrease in negative sentiment(figure7). These trends suggest a shift in children's literature toward emotionally balanced storytelling, reducing distress while maintaining a calm, supportive tone. This reflects a broader move away from emotionally intense or morally charged narratives, aligning with the modern tendency to create stories that are gentle, reassuring, and emotionally appropriate for young readers.

the recurring words show an increased use of specific character names and a strong Presence of verbs the analysis also shows a decline in the use of person-related nouns over time(figure8), while artifact-related nouns have steadily increased(figure9), and adjective usage has gradually decreased(figure10). These features reveal shifts in narrative emphasis and style: specific character names reflect personalized storytelling and stronger connections with young readers, person-nouns reflect social and character-driven storytelling, artifact-nouns highlight growing attention to material settings, and adjectives signal the richness of description. The observed trends suggest that children’s literature has moved away from socially focused, richly described narratives toward more object-driven, action-oriented storytelling with simpler, leaner language.

Building on the linguistic and stylistic shifts identified earlier, topic modeling reveals a clear evolution in the themes of children’s literature. Early works predominantly focused on friendship and moral lessons, reflecting didactic intentions. By the late 19th and early 20th centuries, themes diversified to include family, adventure, and conflict, indicating broader narrative complexity. More recent decades show a shift toward themes of nature, identity, and imagination, emphasizing creativity and emotional exploration. The thematic overlap analysis highlights distinct clusters, with books like *Matilda* and *Alice in Wonderland* sharing imaginative fantasy elements, while others like *Wonder* and *Harry Potter* focus on different emotional or identity-driven narratives. These changes suggest a movement from traditional, lesson-oriented storytelling toward richer, more varied themes that engage young readers’ imagination and personal development, aligning with evolving cultural and educational values in children’s literature. This thematic progression complements the linguistic and stylistic trends observed earlier, collectively illustrating how children’s literature has evolved not only in language use but also in the stories and messages it conveys to young readers.

Building on our observations of stylistic change, computational approaches have been used to better quantify these shifts in children’s literature. The first article (2016) corpus periodization framework applies statistical measures of vocabulary and style to segment texts into distinct historical periods, showing through cosine similarity of TF-IDF vectors that children’s literature evolves in clearly distinguishable phases rather than gradually blending over time. This supports our decision to analyze the corpus by separate eras and underscores the importance of viewing stylistic change as occurring in discrete stages. Complementing this, the second article (2018) uses stylometric features alongside machine learning classification to identify distinct linguistic signatures in early and late eras, while revealing the middle era as a transitional phase blending characteristics of both. Their findings of a gradual, cumulative evolution align well with our thematic and linguistic analyses.

Conclusion

This study sets out to explore how children’s literature has evolved from the 18th to the 21st century in terms of language, style, and thematic focus. Our hypothesis was that children’s books have gradually shifted toward simpler language, more emotionally expressive content, and increasingly inclusive and imaginative themes. The results of our analysis strongly support this hypothesis.

By analyzing 15 well-known children's books from different time periods, we found that writing has changed a lot. Sentences have gotten shorter, the words are easier, and the books are more readable overall. We also saw that books today use fewer moral lessons and more emotional language, which helps readers connect with the characters. The tone has become calmer and more balanced, showing a move toward stories that feel more supportive and less strict.

These changes match how people’s ideas about childhood have changed over the years. In the past, books were mostly about teaching kids how to behave. Today, books help kids explore their feelings, identity, and imagination. Topic modeling also showed how themes have changed—from friendship and morality to nature, imagination, and personal experiences.

Even though we only analyzed 15 books, we were able to see clear patterns and trends. Using text analysis and machine learning helped us track how writing styles have shifted over time. It also showed us that children’s books are more than just stories – they reflect what society thinks children need to learn and feel at different points in history.

In the future, it would be interesting to expand this research by analyzing a larger number of books from each time period to get more accurate and general conclusions. We could also compare books from different countries or cultures to see if the same changes happen in other parts of the world or if different societies have different trends in how they write for children. Another direction could be to focus on specific genres, like fantasy or realistic fiction, and study how style and themes change within those categories. Finally, it could be valuable to look at how illustrations, formatting, or even digital media affect modern children’s literature, especially as more kids read stories online or in apps.

References

Primary Texts

- Project Gutenberg. (n.d.). *Project Gutenberg.* https://www.gutenberg.org/
- Additional texts were retrieved as publicly available PDFs from online archives.

Background and Theoretical Context

- ScienceDirect. (2018). *The length of words reflects their conceptual complexity*.
  link: The length of words reflects their conceptual complexity - ScienceDirect
- Wikipedia contributors. (n.d.). *Readability*. Wikipedia.
  link: Readability - Wikipedia
- Wikipedia contributors. (n.d.). *Dale–Chall readability formula*. Wikipedia.
  link:Dale–Chall readability formula - Wikipedia
- Readable. (n.d.). *About readability*.
  link:About Readability – Readable, the home of readability
- Britannica. (n.d.). *Children's literature: Fairy tales, classics, adventure*.
  link:Children's literature - Fairy Tales, Classics, Adventure | Britannica
- Yılmaz, E. (2021). *The historical development of children's literature and reflections of child/hood in literature*. [PDF]. ResearchGate.
  link:(PDF) The historical development of children's literature and reflections of child/hood in literature

Methodological Framework

- Gómez-Adorno, H., Markov, I., Posadas-Durán, J.-P., Sidorov, G., & Gelbukh, A. (2018). *Stylometry-based approach for detecting writing style changes in literary texts*. ResearchGate.
  link: (PDF) Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts

- Alsudais, A., & Tchalian, H. (2016). *Corpus periodization framework to periodize a temporally ordered text corpus*. ResearchGate.
  link :
  https://www.researchgate.net/profile/Abdulkareem_Alsudais/publication/305566743_Corpus_Periodization_Framework_to_Periodize_a_Temporally_Ordered_Text_Corpus/links/579395d008aed51475bf344d/Corpus-Periodization-Framework-to-Periodize-a-Temporally-Ordered-Text-Corpus.pdf