Intro/Background
Our research delves into the fascinating realm of children’s literature, focusing on the significant contributions of Caroline Hewins and Anne Carroll Moore in shaping this genre. These two influential figures published essential lists of recommended books for children:

- Hewins’ list published in 1882 with ~1000 books.
- Moore’s list published in 1903 with ~500 books.

These books played a pivotal role in defining and codifying what is now recognized as children’s literature.

The core of our research revolves around comparing and contrasting the stylistic elements and thematic connections between these two lists. We aim to understand how Moore’s compilation might have evolved from Hewins’ groundwork, examining the changes, consistencies, and the lasting impact of Hewins’ selections.

Methodology

Data Preparation:

Data Acquisition and Digitization: The books in the lists were compiled from reputable archival sources such as HathiTrust. These texts were digitized using OCR technology to convert the scanned documents into editable and analyzable text formats.

Data Cleaning and Normalization: Post-digitization, the data underwent extensive cleaning to correct:
- OCR errors, remove non-ASCII characters, and normalize textual inconsistencies;
- Filtering out irrelevant information such as page numbers and headings, and fixing spelling mistakes.

Data Tagging: To enhance the data’s analytical value, we tagged each word with relevant metadata, including part-of-speech tagging. This step was critical for differentiating homophones and context-specific interpretations of words (e.g., distinguishing between ‘bat’ as an animal and ‘bat’ used in sports).

Data Analysis Techniques:

Topic Modeling: We used an unsupervised machine learning technique, topic modeling, to identify and categorize thematic patterns in the texts. This process involved assigning words to various topics based on their distribution and frequency, and iteratively refining these assignments to develop coherent thematic clusters.

Thematic Distribution Analysis: Each book was modeled as a weighted combination of topics. We examined the distribution of these topics within individual books and across the entire dataset.

Genre Correlation: Recognizing the potential influence of book genres on thematic content, we mapped topics to genres to understand the genre-specific thematic patterns.

Results

Genre to Genre Mapping: After getting the topic to genre mappings, we are interested in seeing in a Moore genre to Hewins genre mapping. To do this, we embed the topic correlation information into the genre to genre mapping. This way we get a clear visual representation of the correlation between the two genre lists.

Data Cleaning and Normalization

Filtering out irrelevant information such as page numbers and OCR errors, remove non-ASCII characters, and normalize textual inconsistencies.

Data Tagging

To enhance the data’s analytical value, we tagged each word with relevant metadata, including part-of-speech tagging. This step was critical for differentiating homophones and context-specific interpretations of words (e.g., distinguishing between ‘bat’ as an animal and ‘bat’ used in sports).

Data Analysis Techniques

Topic Modeling: We used an unsupervised machine learning technique, topic modeling, to identify and categorize thematic patterns in the texts. This process involved assigning words to various topics based on their distribution and frequency, and iteratively refining these assignments to develop coherent thematic clusters.

Thematic Distribution Analysis: Each book was modeled as a weighted combination of topics. We examined the distribution of these topics within individual books and across the entire dataset.

Genre Correlation: Recognizing the potential influence of book genres on thematic content, we mapped topics to genres to understand the genre-specific thematic patterns.

Conclusion

Topic shifts: Through our analysis, we find evidence that when controlling for genre, books in Anne Carroll Moore’s list are on average more likely to contain topics relating to government and history. We also find that books in Caroline Hewins’ list are on average more likely to contain books about family and rural life. We also find several topics which are equally prevalent in the Hewins and Moore list. In particular, books in the Hewins list on average have the same likelihood as books in the Moore list to display topics relating to science, exploration, and nature.

Hewins to Moore shift: In graphs 2 and 3, remnants of genre distributions from Hewins are evident in Moore’s work. However, graph 4 illustrates that Moore refined her focus on specific topics, in contrast to the broader range observed in Hewins’ work.

PCA for Genre Embeddings: In our preliminary model, we represented genres as a simply a weighted combination of topics to predict a one hot encoded vector for genre. While effective in certain contexts, this approach did not account for the possible relationships between within list genres. To address this, we developed genre embeddings by employing a bag-of-words approach, allowing us to capture the semantic richness of each genre. The PCA then reduced the dimensionality of these embeddings, preserving as much of the variance as possible. The resulting plot reveals the proximity of genres to one another, offering a visual representation of genre similarity based on their thematic content.

References


