Analyzing the Evolution of Children's Literature: A Comparative Study of Lists by Caroline Hewins and Anne Carroll Moore

Vernon Luk, Yasemin Rees, Patrick Phelan

Client: Rebekah Fitzsimmons Advisor: Cosma Shalizi

Carnegie Mellon University

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.

Anne Carroll Moore

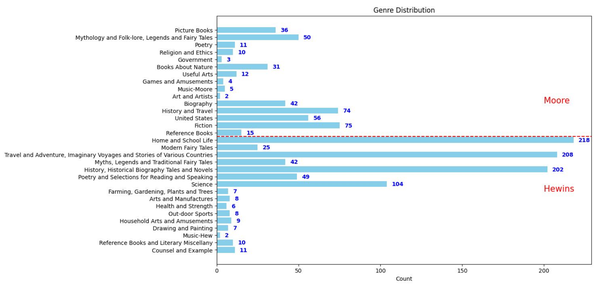
Caroline Hewins

Data

Our dataset comprises titles from the two children's literature lists by Hewins and Moore. We digitized these lists, initially gathered from historical sources, using Optical Character Recognition (OCR). For each book, the dataset includes the full English text, the title, author, publication date, and category. Additionally, Hewins' list provides specific recommendations for gender and age range.

	Title	Author	Illustrator	Publisher	Publisher Location	Date of Publication	Rec. Gender	Rec. Age Range	Category
Hewins List	✔	✔	✔	✔	✔	✔	✔	✔	✔
Moore List	✔	✔	✔	✔	✔	✔	✘	✘	✔

Chart 1: The difference between high-level Moore and Hewins data



Graph 1: The distribution of the number of books for Moore and Hewins

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.

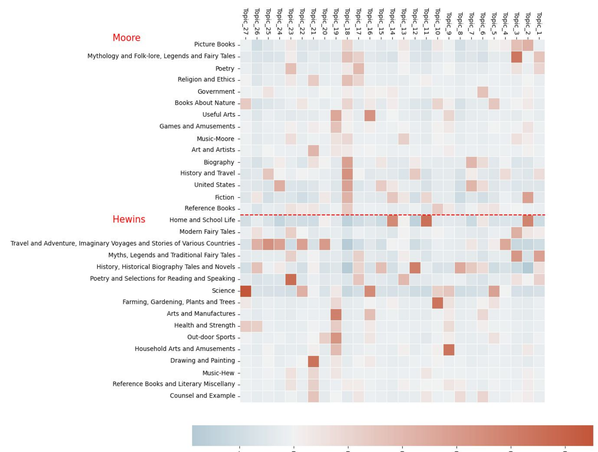
Data Tagging: To enhance the data's analytical value, we tagged each word with relevant metadata, including part-of-speech tagging. This step was critical for differentiating homophones and context-specific interpretations of words (e.g., distinguishing between 'bat' as an animal and 'bat' used in sports).

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

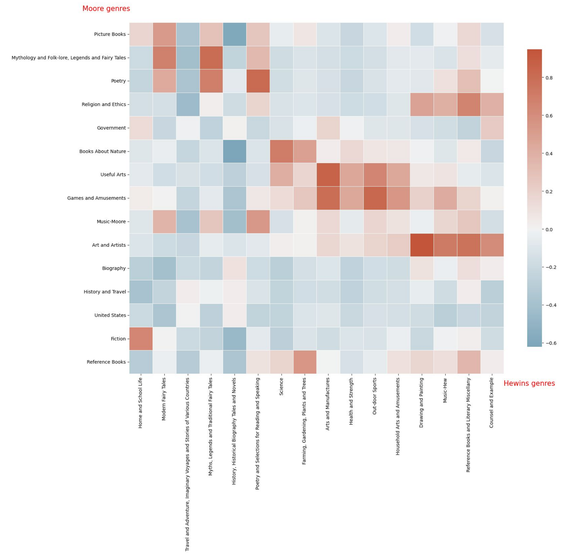
Genre Correlation: Recognizing the potential influence of book genres on thematic content, we mapped topics to genres to understand the genre-specific thematic patterns.



Graph 2: The correlation matrix of 27 topics against the labeled genres of Hewins and Moore. Higher numbers indicate how well each topic predicts the labeled genre.

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.



Graph 3: The correlation of correlations of the labeled genres of Hewins by Moore

From graph 3, we can see there is a spread of correlated and uncorrelated genres.

- Higher Correlations: Fiction and Home and School Life; Mythology from both lists; Art and Artists and Drawing and Painting.
- Lower Correlations: Picture Books and History; Books about Nature and History.

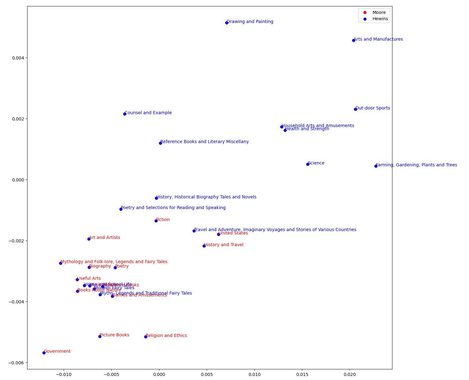
It's important to recognize that when transitioning from graph 2 to graph 3, the correlation matrix is derived by utilizing topic correlations as predictors to infer new correlations. This process results in a loss of information, particularly when dealing with topics that exhibit similar low topic correlations.

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.

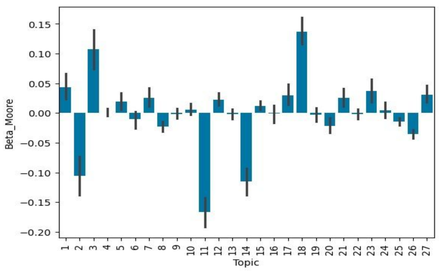
PCA for Genre Embeddings: In our preliminary model, we represented genres as a simply a weighted combination of topics to predict a one hot encoded vector for genre. While effective in certain contexts, this approach did not account for the possible relationships between within list genres. To address 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.



Graph 4: PCA projection of the average word distribution of each genre

Predicting topic weights from embeddings:

- A regression analysis where we compare PCA and the Moore Indicator variable with various topics. After determining the regression parameters, we assess whether the Moore Indicator is non-zero for each topic.
- Non-zero value indicates that the Moore Indicator is a weak predictor and that there was no significant shift in the distribution of topics from Moore to Hewins.
- We obtain confidence intervals for our beta parameters by bootstrap resampling and rerunning our regression.



Graph 5: The predictive power of the Moore indicator by topic

References

HathiTrust Digital Library – millions of books online. HathiTrust Digital Library – Millions of books online. (n.d.). https://www.hathitrust.org/

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3:993–1022. http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

George A. Smathers libraries " UF Libraries " University of Florida. UF monogram. (n.d.). https://uflib.ufl.edu/

David Brown — Associate Teaching Professor of English, Associate Director of First-Year Writing for Research and Assessment