Analyzing Gender and Topical Readership in Children's Books

By Gisele Wu, Lawrence Jang, Matthew Song Project Advisors: David Brown, Rebekah Fitzsimmons Project Supervisors: Jamie McGovern, Peter Freeman

Background

Books for the Young (1882) by Caroline Hewins was a guide for children's literature that contained a list of 1005 titles recommended for developing children. It was assembled before the idea of "children's literature" being culturally codified as an accepted genre. Hewins and her work is largely considered to be one of the pioneers of children's literature and continues to stand as an influential figure in the American Library Association (ALA).

The goal of the study is to analyze Hewins' classification of the books and ultimately compare the underlying social standards in 19th-century literature vs now. Her list included not only the titles she considered suitable as "children's literature" but also other metadata. One additional piece of information that she included was the gender she imagined as appropriate: "Boys and GIrls", "Boys" or "Girls". This poster will thus primarily focus on analyzing gender readership with regards to text topic and type.

Data

Our dataset consists of 1092 individual book text files along with a spreadsheet of 1) characteristics such as gender of author(s), year of publication, etc. 2) Hewins' classification of each text file, such as readership regarding children of different genders, genre, etc.

Preprocessing Pipeline

1) **Text Cleaning** or pre-processing of corpus, including omitting front- and back-matter, and correcting OCR errors;
2) **DocuScope Tagging** to investigate text-type variation;
3) **Topic modeling**, which followed Jockers' procedure of part-of-speech tagging texts, splitting them into 1000-word chunks, and applying Latent Dirichlet Allocation (LDA) to each;
4) **PCA for dimension reduction** on both extracted topics (by means for each text) and DocuScope frequencies (normalized per 100 tokens); and
5) **Agglomerative Hierarchical Clustering (AHC)** to further explore how texts group by DS frequencies.

Conclusion

Our results show that books with an imagined audience of exclusively girls had less of a variety in terms of text type variation and content. We can generally see that books imagined for an audience of exclusively "Girls" largely consist of household-related topics and narrative text-types (with notable outliers), while books for "Boys" and "Boys and Girls" have greater variation in their topics and text-types. These include topics related to adventure and informational text-types. Our data suggest that girls' readership was imagined with stricter social conformities than boys'.

Analysis and Results

Given that metadata our team set out to explore the relationships among those categories and variables extracted using 2 techniques: **Topic Modeling** and **Docuscope** (a tagger that assigns words and phrases into rhetorically oriented categories.

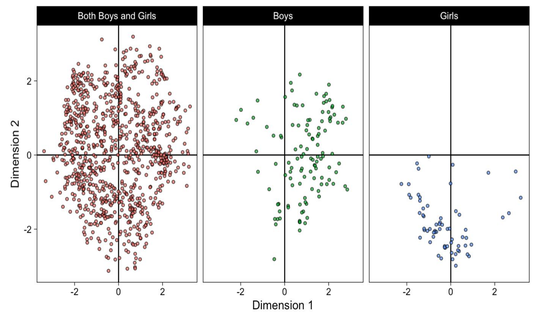

Figure 1: Scatterplot PC1 against PC2 by Gender Readership (Topical)

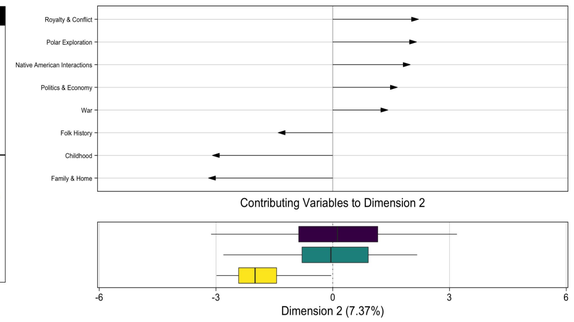

Figure 2: Boxplot of Top 10 Variable Contributions to PC2 (Topical)

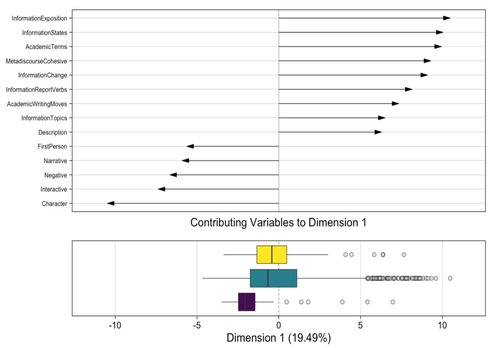

Figure 3: Boxplot of Top 10 Variable Contributions to PC1 (Docuscope)

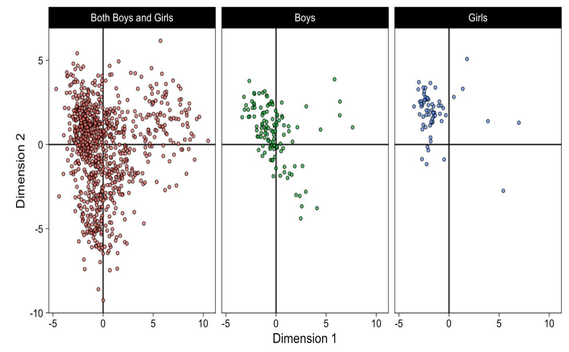

Figure 4: Scatterplot PC1 against PC2 by Gender Readership (DocuScope)

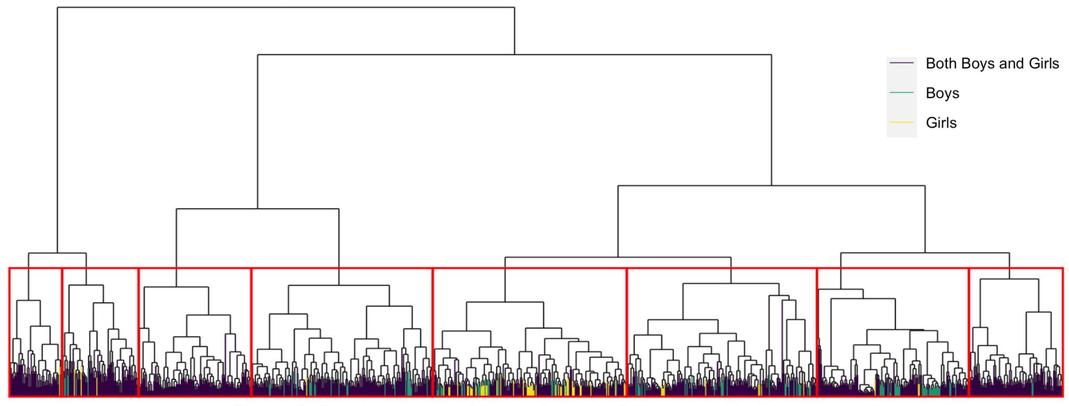

Figure 5: Dendrogram on Corpus Data via Docuscope Variables

I. Topical Modeling

- Notable Findings:
  - Books imagined for girls are all at the negative end of PC2, which is comprised of topics related to domesticity, and none of the positive end of PC2 which is made up of topics related to exploration and conquest.
  - Books imagined for boys contain more variation in the topics.

II. DocuScope Tagging

- **Principal Component Analysis (PCA)**
  - Notable Findings:
    - PC1: Books for girls comprised more **Character, Interactive, Negative** text types
    - PC2: Books for girl comprised more **PublicTerms, InformationPlace, InformationTopics** text types
    - Texts for other audiences include more Interactive, Reasoning Information, Confidence, First person text types, **largely absent from books for girls**
- **Hierarchical Clustering**
  - A **top-down approach** to group the data by analyzing the clustering behavior across all the DocuScope variables
    - While articles for boys (green) were spread out among different clusters, articles solely for girls (yellow) centered around the 5th cluster.
    - **Cluster 5 - the majority of files with Readership as Girls**
      ~ 63% high in **Interactive, Description,** and **Positive**
      ~ 37% high in **Interactive** and **Reasoning**, low in **Description**, and **InformationChangePositive**
  - It shows that in Hewins' corpus, for texts imagined for an audience of exclusively girls, the text-types have **less variation** in their cluster locations.

"In general it is harder to find good books for girls than for boys, although a girl who has not a precocious appetite for love-stories often enjoys the same stories that please her brother. " - Caroline Hewins

Citations:

- Ishizaki, S., & Kaufer, D. (2012). Computer-aided rhetorical analysis. In *Applied natural language processing: Identification, investigation and resolution* (pp. 276-296). IGI Global.
- Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.