Implicit Gender Inequality in Children's Picture Books: Evidence from a Text Mining Analysis of 200 Bestselling Chinese and British Titles

Li, Yi
s2127837@ed.ac.uk
School of Literatures, Languages and Cultures, University of Edinburgh, United Kingdom

Terras, Melissa
m.terras@ed.ac.uk
College of Arts, Humanities and Social Sciences, University of Edinburgh, United Kingdom

Li, Yongning
lee9512@163.com
School of Systems Science, Beijing Normal University

Background

As the primary resource for preschool children, picture books, and their gender narratives, can unconsciously shape and change children's perceptions of sex roles and gender identity (Bleakley et al., 1988; Connor & Serbin, 1978; Latima, 2020). However, existing studies show concerning trends in the representation of gender inequality in modern picture books, such as the overwhelming number of male main characters and traditional gender stereotypes of vocations, personalities and habits (Casey et al., 2021; Hamilton et al., 2006; Lee & Chin, 2019; Terras, 2018). It is therefore important for children's picture books to have diverse gender descriptions and improved equal gender representations.

Since the second Feminist Movement in the 1960s, gender equality in UK children's picture books have been continuously examined yet slowly improved (Adams et al., 2011; Allen et al., 1993; Capuzza, 2020; Hamilton et al., 2006). Similar studies have been far less common in China, as the Chinese picture book market only developed from the start of the 21st Century (Xiao, 2021). One study has shown the existence of the traditional gender biases in Chinese picture books (Liu & Chen, 2018). Based on the research gap between these two countries, this study will (1) investigate gender representations and narratives in picture books, (2) compare the similarities and differences between bestselling British and Chinese picture books texts from 2010 to 2020. We do so by applying text mining techniques to analyse gender narratives within picture book texts themselves. This follows on from our 2022 study where we analysed publisher's descriptions of texts, rather than full text mining of the book's content (Li et al., 2022).

Method

Data Corpus

We collected lists of bestselling books from *The Bookseller* and *The Publisher*: publishing trade magazines in the UK and China. We ascertained the best-selling 10 British and Chinese titles in 2010- 2020, to compile a 200-title booklist. We then procured a physical copy of all titles, and manually transcribed machine processable text of all title content to enable text analysis. Our resulting corpora contain 310,000 Chinese characters and 80,000 English words. All data and texts in this process were only used for analysis and research, complying with the text and data mining copyright exceptions (Kelly, 2016).

Table 1. **Table of British and Chinese Bestselling Children's Book Data**

Data source	**Data description**	**Manually Transcribed Data Details**
The Bookseller	Weekly top10 children's pre-school or picture book list in English from 2011 to 2015. Top 20 children's pre-school or picture books list in weekly issue since 2016	316,458 Chinese characters [1]; 10,598 Chinese words
The Publisher (Open Book)/ Dangdang.com	Weekly top10 children's pre-school or picture book list in Chinese from 2013 to 2015. Monthly top 20 children's books list since then; No data available on 2010 - 2012	79,352 English words

Full Text Data Analysis

We examined the popularity of gendered words by using established text mining techniques – including word segmentation and term frequency - in the transcribed texts of the 200 picture books, comparing gender-related words. First, we split the Chinese corpus and English corpus into two wordlists using word segmentation. We did not use a stopwords list as they normally include pronouns that indicate gender (Rao & Taboada, 2021). In English, we split sentences by recognising the space between words, while in Chinese text, we applied Jieba package (precise mode) to split words (BREEZEGEOGRAPHY, 2018). This generated two wordlists (Chinese and English) of all unique words with their frequencies, and we ranked the words' popularity by frequency to produce the top 100 wordlist. Second, we manually marked gendered words (those which contains gender features) from two wordlists and manually classified all gender words into four groups (see Table 2). We finally calculated the number of categories and frequencies of all masculine and feminine words and presented them as below, as well as the number of gender words in the top 100 wordlists, to compare how linguistic narratives indicate gender separately in Chinese and English texts.

Table 2. **Gendered Words Classification for Text Mining (both English and Chinese)**

	Female	**Male**
Pronouns	She/her/她	He/his/him/他
Nouns	Mrs/女士, witch/女巫, princess/公主,etc.,	Mr/先生, Captain/队长, King/国王,etc.,
Family identities	Mum (妈妈), grandma (奶奶/姥姥), Aunt(阿姨), etc.,	Dad (爸爸), grandpa (爷爷/姥爷), uncle (叔叔), etc.,
Names (of characters)	Peppa, Spinderella, 卡门(Carmen), 歪歪兔(Wobbly Rabbit) , etc.,	Wally, Alfie, 嘎嘎(Gaga), 约瑟(Joseph), etc.,

Note: We classified all gendered words in two corpora into five groups. This table provides some examples.

Previous studies have tested sentiment analysis to examine the gender biases in social media, such as film synopsis (Bhaskaran & Bhallamudi, 2019; Pair et al., 2021; Park & Woo, 2019; Ramadi et al., 2022; Xu et al., 2019; Zhang et al., 2022). We applied sentiment analysis on our texts to detect emotional differences in gender narratives. Firstly, we separated the corpora into a list of individual sentences by grammar, such as full stops, question marks, exclamation etc., and applied TextBlob [2]and Paddlehub [3] on each sentence. The two packages are based on machine learning techniques, providing English and Chinese sentiment analysis. We then categorised all sentences by gender into four groups - female sentences, male sentences, mixed gender sentences (including both male and female) and gender-neutral sentences. A final average and median value of sentences in each group are presented and compared between two corpora below.

Results

Figure 1 shows the gender of authors and illustrators of 200 best-selling titles. The preference for male words in the book texts of the top 200 picture books exists in both British and Chinese titles (see in Table 3&4). However, the contrast between gender is sharper in the Chinese corpus than the British.

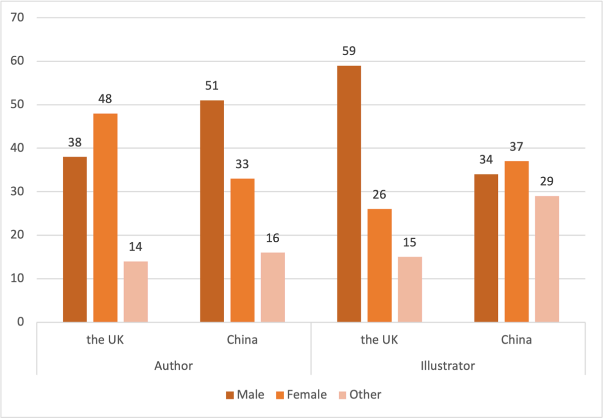

Figure 1. **Gender Coded for Authors and Illustrators Top 200 Picture Books**

Table 3. **Categories and Total Frequency of Gendered Words**

	China	**The UK**		
	Male	Female	Male	Female
Pronoun	3 [4]/1,195	1/241	3/1,404	3/1,370
Identity	24/465	25/627	7/167	11/222
Nouns	43/303	31/93	24/382	14/438
Names	151/1,894	33/460	100/935	49/474
Animals	3/52	1/9	1/1	2/56

Note: This table provides statistics with (1) the number of gendered words in that category (left) and (2) the frequency of all words in that category (right); the red mark is the female dominance in that category.

Table 4. **Top30 Gender Words in British and Chinese Corpora (respectively)**

Rank	**The UK titles**	**Chinese titles**	**Rank**	**The UK titles**	**Chinese titles**
1	he	他 (he)	16	Scrooge (cha*)	哥哥 (big brother)
2	she	妈妈 (mum)	17	Danny (cha*)	迪克 (Dickie, cha*)
3	his	她 (she)	18	grandpa	佩罗 (Peiluo, cha*)
4	her	霸王龙 (Bawanglong, cha*)	19	girl	大脚怪 (Dajiao-guai,cha*)
5	mum	爸爸 (dad)	20	Mickey (cha*)	小兔 (Xiaotu,cha*)
6	Peppa	卡梅利多 (Camilido, cha*)	21	granny	小弟 (little brother)
7	him	卡门 (Carmen, cha*)	22	Elsa (cha*)	罗西娜 (Rosina, cha*)
8	George	豌豆射手 (Pea shooter, cha*)	23	princess	布瓦 (Bu-wa,cha*)
9	Mr	贝里奥 (Beleo)	24	sister	朗朗 (Langlang, cha*)
10	Wally (cha*)	歪歪兔 (Waiwaitu, cha*)	25	Anna (cha*)	维克托(Victor, cha*)
11	Miss	威威龙 (Weiweilong,cha*)	26	Minion (cha*)	英雄 (hero)
12	witch	卡梅拉 (Carmela, cha*)	27	fairies	大嘴花 (Dazui-hua,cha*)
13	dad	爷爷 (Grandpa)	28	Harry (cha*)	小熊 (Xiaoxiong, cha*)
14	Queen	柯尔克 (Kolk, cha*)	29	man	甲龙 (Jialong, cha*)
15	wizard	先生 (Sir)	30	Minnie (cha*)	小黑 (Xiao-hei,cha*)

Note: 'Cha*' represents the word is a name of character in the book texts, all names in the corpora have been manually coded and aligned with the gender in the books. Red marks as female words.

Table 5. **Common Gender Words in Top 30 Gender Wordlists**

Word in Common	**Rank in English Corpus**	**Rank in Chinese Corpus**
He (him/himself/his)	1/1	1/1
She (her/herself)	2/2	3/9
Mum	5/43	2/3
Dad	13/96	5/13
Mr	9/78	15/108
Grandpa	18/156	13/94

Note: The first number in each box is the rank of the word in the gender wordlist; the second number is the rank of the word in the whole wordlist, it represents the importance of the gender words in the full texts.

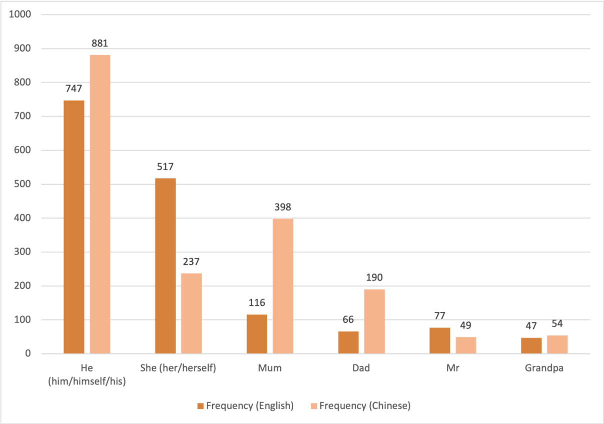

Figure 2. **Frequencies of Common Top Gender Words in the UK and Chinese Corpora**

Table 6 shows the sentiment change by gender narratives in British and Chinese picture book texts. In general, the British corpus is more gender neutral than the Chinese corpus, and the narratives are more positive when the sentence includes both male and female words. However, the algorithm and model in the packages we used are black boxes, and the results require more explanations with further qualitative analysis.

Table 6. **Sentiment Score for Gender Narratives the UK and Chinese Titles**

	The UK	**Chinese**				
	Median	Average	Number	Median	Average	Number
Male	0	0.04	1680	0.052	0.058	2302
Female	0	0.06	1192	0.2462	0.15	655
Male and female	0	0.097	1864	0.4594	0.24	347
Gender Neutral	0	0.04	3421	-0.0642	-0.04	6951

Note: This table provide the results for sentiment analysis from the packages used, the scores are between [-1;1] [5]. *Median means the median value of sentiment score in that group, average means the average sentiment score of sentences in that group, number represents the number of sentences with the gender category in the left box.*

Discussion

This study indicates that gender inequality exists in both publishing contexts, with the male dominance being more pronounced in Chinese bestselling picture-book titles than in British. There are not only more female words (in more word categories) of British picture books, but the overall gap between two genders is smaller than in the Chinese texts. However, this study is only an experimental analysis of 200 best-selling books and cannot represent the whole picture book market in two countries. Besides, the sentiment analysis technique is based on deep learning with hidden algorithms, meaning the results are dependable on the package we used and may cause classification problems. We still hope the final sentiment scores can provide additional evidence for a deeper discourse analysis. Further research will expand the dataset and add publication dates as a variable, as well as applying other methods such as relationship extraction, text similarity and network analysis.

Conclusion

The computational analysis of fully transcribed texts analysed the implicit gender stereotypes and the preference towards male representation in British and Chinese bestselling children's picture books in the last decade. Although male dominance still exists in both corpora, the British titles showed areas of equality, such as the ratio of female words in the British corpus. Future work will include applying these text mining methods to other languages and different children's book categories, to reflect gender inequality within published texts from a data-based perspective. Our research provides a method which will be applicable to others wishing to compare and contrast gender-related differences in individual book markets, which will also be useful to translation studies.

Notes

1. Words are made of characters in Chinese and are not fixed as English words, normally, a Chinese word includes two or three characters.
2. Available online: https://github.com/sloria/TextBlob
3. Available online: https://github.com/PaddlePaddle/PaddleHub
4. The pronouns are three words like she, her and herself, rather than one, but they all represent women.
5. TextBlob sentiment package reports sentiment score for English text between [-1:1], from negative to positive; while Paddle hub reports the results for Chinese text between [0:1] from negative to positive. In this study, we standardised the results from Paddle hub in Chinese to [-1:1] to better compare the results in two corpora.

Bibliography

Adams, M., Walker, C., & O'connell, P. (2011). Invisible or Involved Fathers? A Content Analysis of Representations of Parenting in Young Children's Picturebooks in the UK. Sex Roles, 65(3–4), 259–270. http://dx.doi.org.ezproxy.is.ed.ac.uk/10.1007/s11199-011-0011-8

Allen, A. M., Allen, D. N., & Sigler, G. (1993). Changes in Sex-Role Stereotyping in Caldecott Medal Award Picture Books 1938—1988. Journal of Research in Childhood Education, 7(2), 67–73. https://doi.org/10.1080/02568549309594842

Bhaskaran, J., & Bhallamudi, I. (2019). Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. ArXiv:1906.10256 [Cs]. http://arxiv.org/abs/1906.10256

Bleakley, M. E., Westerberg, V., & Hopkins, K. D. (1988). The Effect of Character Sex on Story Interest and Comprehension in Children. American Educational Research Journal, 25(1), 145–155. https://doi.org/10.2307/1163164

Breezegeography. (2018, January 25). How to Segment Chinese Texts: Putting in Spaces with Jieba. VB Geography. https://breezegeography.wordpress.com/2018/01/25/how-to-segment-chinese-texts-putting-in-spaces-with-jieba/

Capuzza, J. C. (2020). "T" is for "transgender": An analysis of children's picture books featuring transgender protagonists and narrators. Journal of Children and Media, 14(3), 324–342. https://doi.org/10.1080/17482798.2019.1705866

Casey, K., Novick, K., & Lourenco, S. F. (2021). Sixty years of gender representation in children's books: Conditions associated with overrepresentation of male versus female protagonists. PLOS ONE, 16(12), e0260566. https://doi.org/10.1371/journal.pone.0260566

Connor, J. M., & Serbin, L. A. (1978). Children's responses to stories with male and female characters. Sex Roles, 4(5), 637–645. https://doi.org/10.1007/BF00287329

Hamilton, M. C., Anderson, D., Broaddus, M., & Young, K. (2006). Gender Stereotyping and Under-representation of Female Characters in 200 Popular Children's Picture Books: A Twenty-first Century Update. Sex Roles, 55(11), 757–765. https://doi.org/10.1007/s11199-006-9128-6

Kelly, J. (2016). The text and data mining copyright exception: Benefits and implications for UK higher education. Jisc. https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception

Latima, M. (2020). Picture Books Should Have More Diversity. Canadian Children's Book News, 43(2), 21–22.

Lee, J. F. K., & Chin, A. C. O. (2019). Are females and males equitably represented? A study of early readers. Linguistics and Education, 49, 52–61. https://doi.org/10.1016/j.linged.2018.12.003

Li, Y., Terras, M., & Li, Y. (2022, July 22). Gender and Cultural Diversity in Chinese Children’s Picture Books: A Data-led Analysis of Bestselling Modern Titles. Alliance of Digital Humanities Organizations. https://confit.atlas.jp/guide/event/dh2022/subject/SP4-01-160/detail

Liu, X., & Chen, S. (2018). The phenomenon and analysis of gender bias in early reading promotion——Analysis of ‘must read picture books for boys and girls’ 早期阅读推广中的性别偏见现象及分析——对“男孩/女孩必读绘本” 的分析. Library Theory and Practice, 12, 17–20.

Pair, E., Vicas, N., Weber, A. M., Meausoone, V., Zou, J., Njuguna, A., & Darmstadt, G. L. (2021). Quantification of Gender Bias and Sentiment Toward Political Leaders Over 20 Years of Kenyan News Using Natural Language Processing. Frontiers in Psychology, 12. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.712646

Park, S., & Woo, J. (2019). Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum. Applied Sciences, 9(6), 1249-. https://doi.org/10.3390/app906124

Ramadi, K. B., Mehta, R., He, D., Chao, S., Chu, Z., Atun, R., & Nguyen, F. T. (2022). Grass-roots entrepreneurship complements traditional top-down innovation in lung and breast cancer. Npj Digital Medicine, 5(1), Article 1. https://doi.org/10.1038/s41746-021-00545-x

Rao, P., & Taboada, M. (2021). Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. Frontiers in Artificial Intelligence, 4. https://www.frontiersin.org/articles/10.3389/frai.2021.664737

Terras, M. (2018). Picture-Book Professors: Academia and Children’s Literature. Cambridge University Press.

Xiao, ci. (2021). 蒲蒲兰: 突破 “黄金十年” Poplar Picture Books: Breakth-rough the Golden decade. The Publisher, 02, 30–31.

Xu, H., Zhang, Z., Wu, L., & Wang, C.-J. (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. PLOS ONE, 14(11), e0225385. https://doi.org/10.1371/journal.pone.0225385

Zhang, L., Li, Y.-N., Peng, T.-Q., & Wu, Y. (2022). Dynamics of the social construction of knowledge: An empirical study of Zhihu in China. EPJ Data Science, 11(1), Article 1. https://doi.or-g/10.1140/epjds/s13688-022-00346-6