A Replication of "Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction" by Authors Eve Kraicer and Andrew Piper, Published in the Journal of Cultural Analytics on January 30, 2019

Hannah Hardenbergh
Harvard University
Cambridge, MA
hhardenbergh@college.harvard.edu

ABSTRACT

UPDATED—13 May 2019. This paper engages with the work of authors Eve Kraicer and Andrew Piper in their research on gender-biased decisions in character structure in contemporary fiction as an extension of a broader topic on the underrepresentation and decentralization of women across various cultural fields. They explain the importance of this research in their paper:

> "Fiction, speaking very generally, is about the individual in society, about the expectations and conflicts that color a life when an obdurate reality stands in the way of one's self-image or desires...The invisible centerpiece of every great novel is the protagonist's rebellion or coming to terms with his or her place in the scheme of things."

Kraicer and Piper consider how today's modern societal gender constructs and standards work against the promotion of women. Their research is a deep dive into the analysis within novels themselves. They are concerned that text-language within novels is itself biased towards male-dominant spaces, individuals, and relationships.

Author Keywords

Fiction, contemporary; gender bias; heteronormativity; visibility; social balance theory; character prominence; character interactions; gender pairs.

INTRODUCTION

I show herein how authors Kraicer and Piper can improve upon their impressive and extensive research on gender hierarchy as found in the analysis of language in contemporary novels. I begin briefly by contextualizing the replicated paper in the current cultural climate of problems with gender-normative standards in American society. Then, with the goal of simplifying their graphics, I analyze the first four figures and improve upon the author's visualizations.[1] This improves the strength of their claim that gender bias is visible in contemporary literature through their quantitative analysis as the more simplified graphics are more appropriate for their intended audience within the digital humanities field and beyond. In a field with both statisticians and literary scholars, it is imperative to have graphics that are easy to read and intriguing to consider.

Additionally, I mention the benefits of having scholars like Kraicer and Piper conducting research that asks readers to consider an important convergence between the humanities and the social sciences fields. Not only was Kraicer and Piper's research difficult to find in the Harvard Dataverse, but also theirs was one of very few papers related to the literary arts and impacts of societal behaviors on cultural material (rather than political science). This topic interests me as I continue to learn the many structures that academic research can take, be it a study on the political leanings of persons related to their exposure of racial threat or a computational study of culture. It interests me to consider literature and the arts as a topic worthy of quantitative analysis. An expansion of this intersection of two unique fields could lead to studies such as the analysis of the perceived value of artworks in relation to the artist's race, for example. The digital humanities is emerging as somewhat new and its limitations are yet to be discussed. This work has been in part pioneered by the Department of Languages, Literatures, and Cultures at McGill University, a group that publishes the Journal of Cultural Analytics.[2] Their team "explores the use of computational and quantitative approaches towards understanding literature and culture in both the past and present." They use data science to "promote a

[1] Note that this is not an analysis of every graphic in Kraicer and Piper's report.

[2] More information on the Journal of Cultural Analytics can be found on their website, here: https://culturalanalytics.org/about/about-ca/ and here: https://txtlab.org/2019/02/social-characters-new-lab-collaboration-by-eve-kraicer/.

more inclusive understanding of culture and creativity." They hope to improve the found gender bias in the publishing industry and promote more gender fluidity among writers.

FINDINGS

The authors find the following to be true: first, the ratio of women to men characters in novels is estimated to be 40:60, which remains the same in all genres observed. Interestingly, one outlier in this finding is that main characters in all novels had a 50.2/49.8 ratio of men to women. Then, Kraicer and Piper find that women authors write more women characters than men do but are still placing male characters in more prominent roles overall. The authors also find that only 19% of novels from all seven genres showed women in the top two most mentioned character positions in a 100,000-word analysis. Then, they analyze relationships within the texts and discuss how authors favor pairing men and women characters together – i.e. making heteronormative relationships - much more than they would a same-gender pair.

CONTEXT AND PREVIOUS RESEARCH

Kraicer and Piper's work solidifies research that has confirmed the marginalization of women in various facets within culture, including book reviewing, filmmaking, acting, and journalism.[3] They consider the work that has been done on how words themselves can be viewed as gendered. Take, for example, the word *chairman*, or *manpower*.

METHODS

BookNLP, an analytical tool created by David Bramman.[4] The package is a natural language processing tool, or NLP, strives to understand how verbal and non-verbal communication affect the human brain.[5] These connections refer to the language of human behavior.

Kraicer and Piper use bootstrapping, chi-square tests, assortativity, and prevalence tests to look at the details of gendered language and character relationships.

VISUALIZATIONS AND GRAPHICS

For the first set of figures, the authors simply display how their character rankings method correlates to the number of times a character is mentioned within 100,000 words of a selected text in any of their novels. The figures are replicated in this project. However, it would be much simpler for whoever is reading the original paper just to get a clearer visual of something more interesting other than a black line that simply shows the nature of a ranked list. The correlation between a ranked list and the number of times a character is mentioned in a novel would be easy to understand with a short explanation included in the paper. Therefore, if this paper were to be revised, I would include simple bar graphs to show the imbalance of gender in characters who are mentioned more often within text. It becomes clear that male characters are more prominent. This would be a great addition to contextualizing the type of gender bias that the authors begin to work with in their later tests to find more detailed conclusions about the effect of gender in contemporary fiction.

The authors describe their "manual" process of assigning gender ID's to each character once they acquired the data. This process included an error based on how their methods failed to assign some (about fifteen) characters either male or female. Below is a table and another graphic to visualize this error. It is easier to understand that, by this bar graph, should the number of unassigned (marked with a "?" in the data) gendered characters be female, the number of female character mentions would be greater. Additionally, visualizing this error also calls more attention to the question of using a gender binary in contemporary fiction. What if there were other categories besides male and female?

LIMITATIONS

Kraicer and Piper address some of the constraints their research confronted, but not all. First, though a small limitation, they noted the fact that they estimated predicted-gender assignments rather than asserting confidence of gender identities across all novels in their dataset. They manually assigned each character (of which there are over 26,000 total), which inevitably produces errors.

QUESTIONS / PROBLEMS

What causes a 50:50 ratio in protagonists, but not in major characters overall? And what is the definition of a character "interaction?" Can it be defined as a definitive relationship in the novel, or is it merely the proximity of a character name to another's in the text of the novel itself? In other words, are "interactions"

[3] Kraicer and Piper, 1-2.

[4] See Bramman's Github on BookNLP program: https://github.com/dbamman/book-nlp.

[5] https://www.neurolinguisticprogramming.com/

equivalent to or alluding to a level of intimacy between characters? If not, the data and natural language processing method feels less substantial to claim that certain relationships exist more often than others, e.g. heteronormative "pairs." And, what about pronouns, do they count as identifiers of gendered characters?

I am equally curious in the authors' decision to build on a 60-year-old theory about social balances.[6]

CONCLUSION

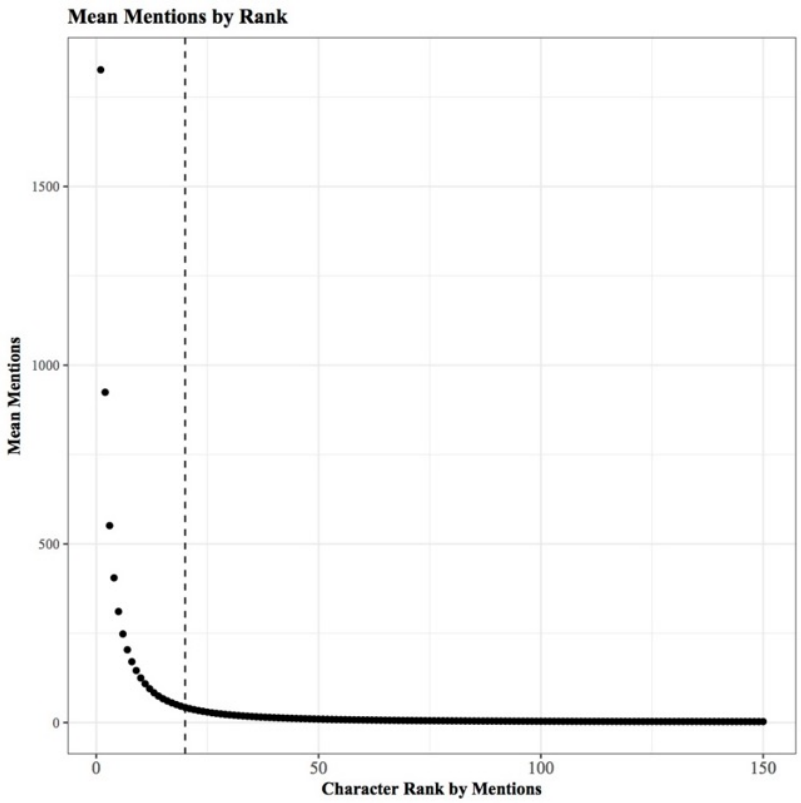
The first thing someone does when skimming a published paper is look at the images. The graphics therefore serve a huge purpose in allowing the reader to quickly understand what initially the research will entail. When I first read Kraicer and Piper's publication, it simply took too much time to understand their graphics, especially the first two, which are simply trying to represent that the more often a character was mentioned in a novel, the higher of a ranking that character received in their dataset.

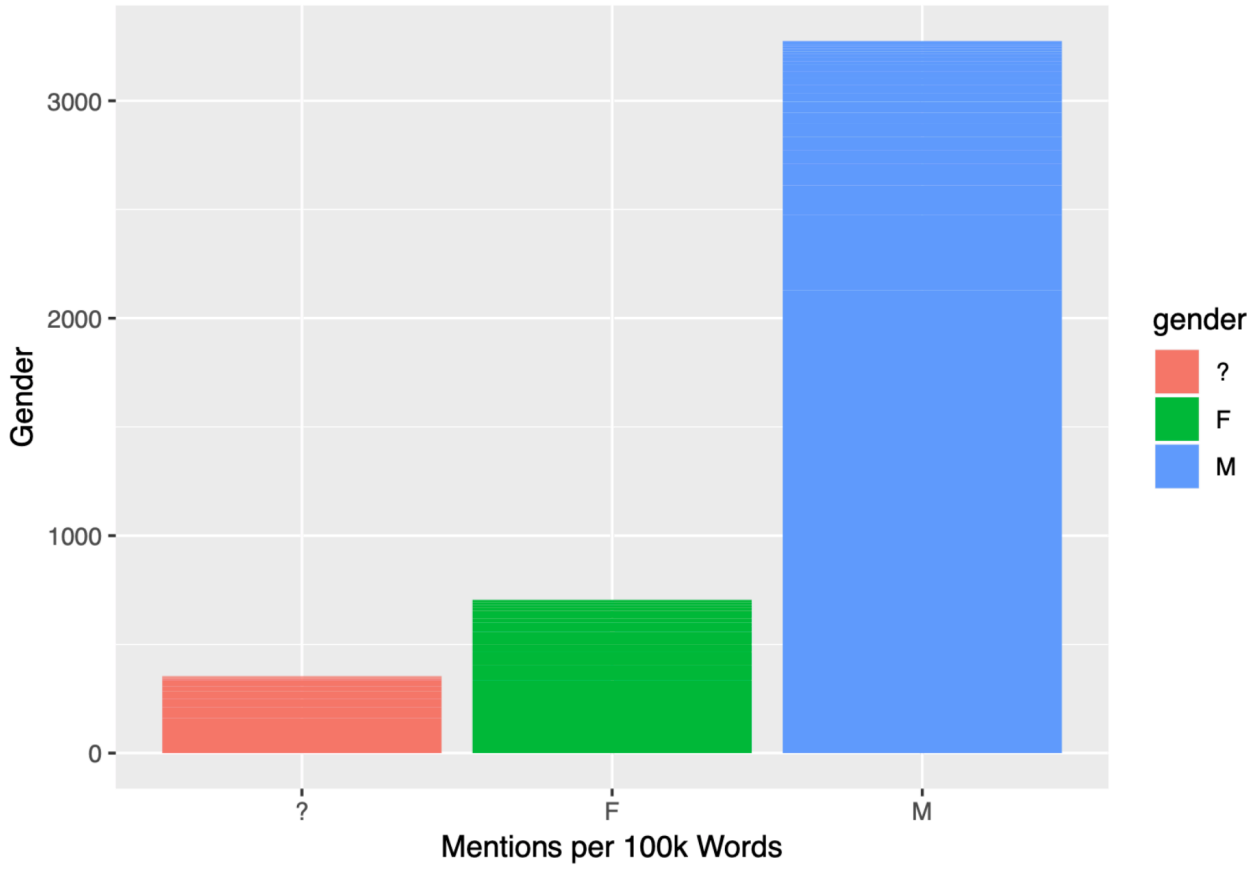
REFERENCES

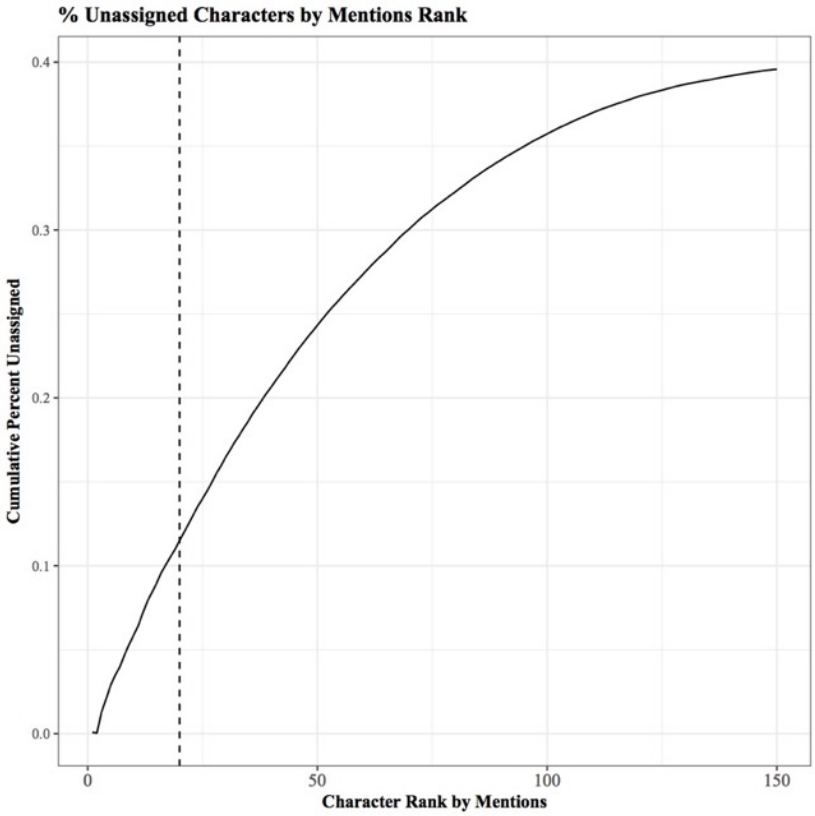
1. David Bamman, Ted Underwood and Noah Smith, "A Bayesian Mixed Effects Model of Literary Character," ACL 2014.

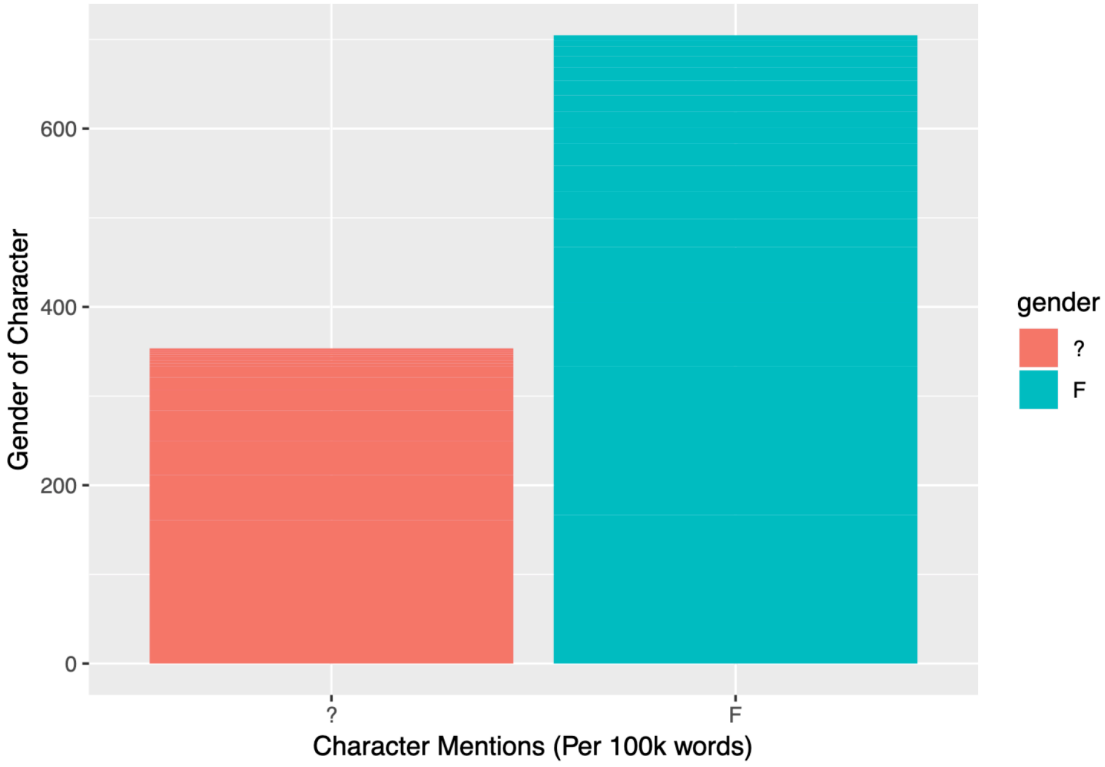
2. Eve Kraicer and Andrew Piper, "Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction," Journal of Cultural Analytics. January 30, 2019.
3. Fritz Heider, "Attitudes and Cognitive Organization," The Journal of Psychology21 (1946): 107- 112 and Dorwin Cartwright and Frank Harary, "Structural Balance: A Generalization of Heider's Theory," Psychological Review 63 (1956): 277-293.

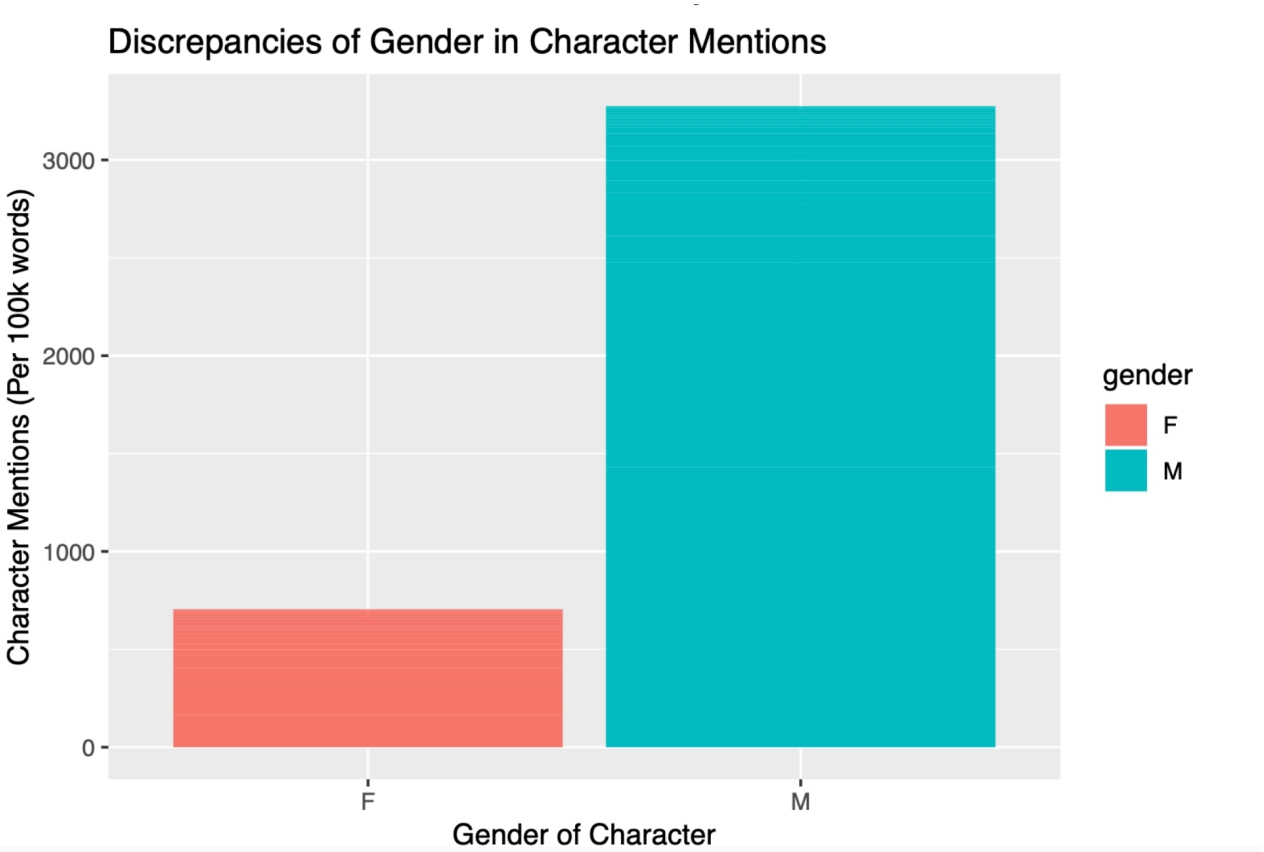
[6] See Fritz Heider, "Attitudes and Cognitive Organization," The Journal of Psychology21 (1946): 107- 112 and Dorwin Cartwright and Frank Harary, "Structural Balance: A Generalization of Heider's Theory," Psychological Review 63 (1956): 277-293.

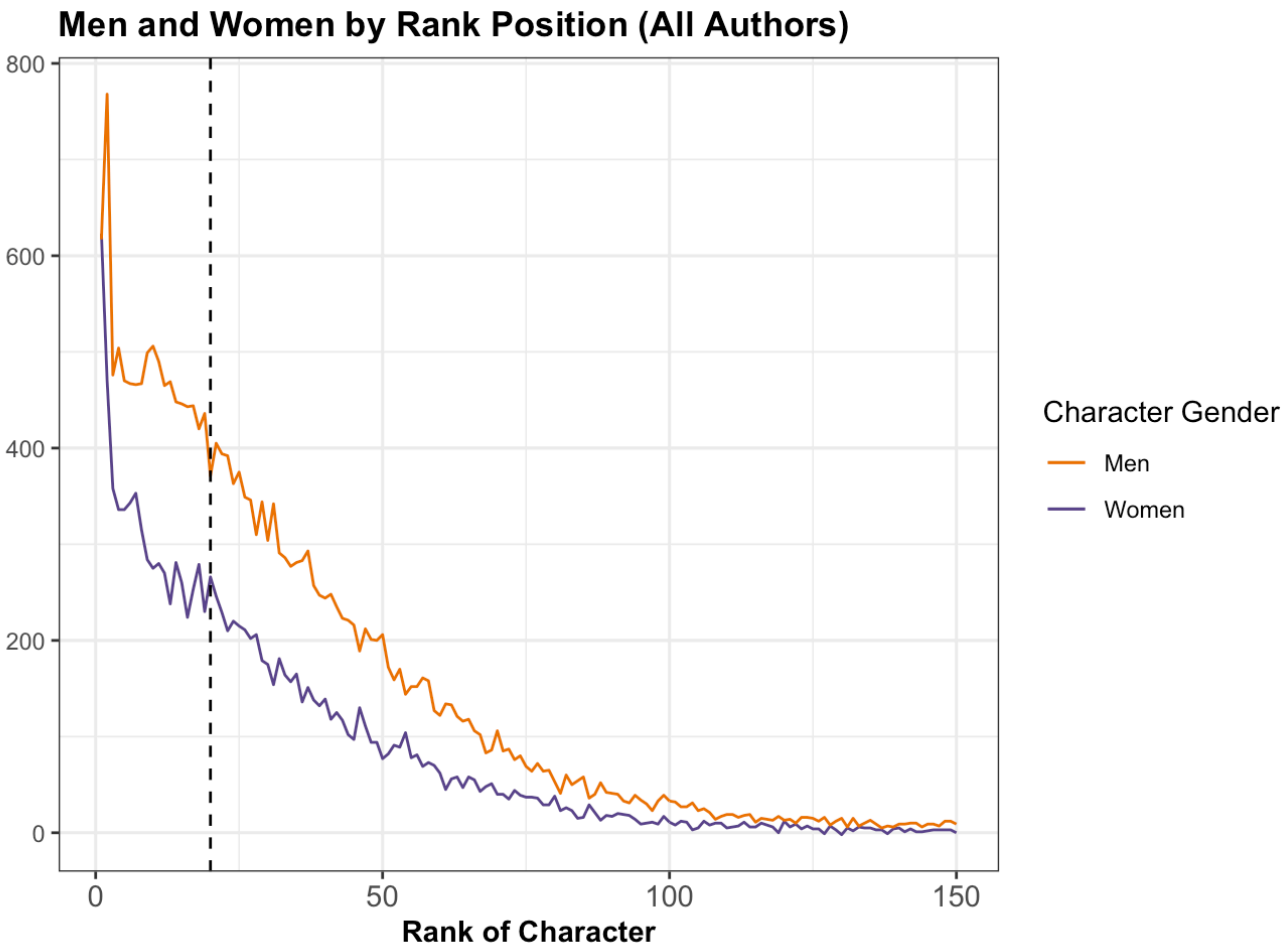

Mean Mentions by Rank
Mean Mentions
1500
1000
500
0
0
50
100
150
Character Rank by Mentions

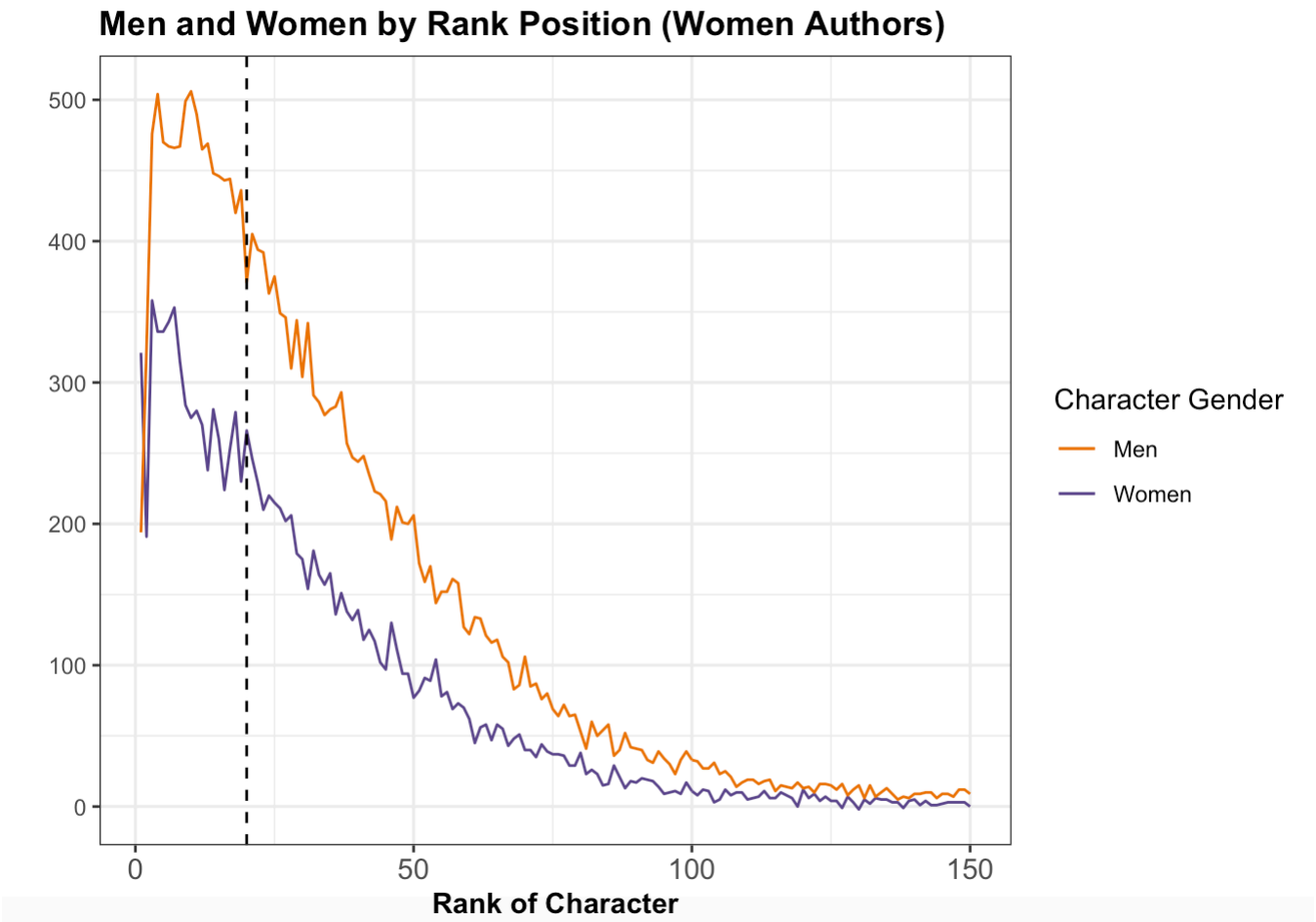

Gender
3000
2000
1000
0
?
F
M
Mentions per 100k Words
gender
?
F
M


% Unassigned Characters by Mentions Rank
Cumulative Percent Unassigned
0.4
0.3
0.2
0.1
0.0
0
50
100
150
Character Rank by Mentions


Gender of Character
600
400
200
0
?
F
Character Mentions (Per 100k words)
gender
?
F


Discrepancies of Gender in Character Mentions
Character Mentions (Per 100k words)
3000
2000
1000
0
F
M
Gender of Character
gender
F
M


Men and Women by Rank Position (All Authors)
800
600
400
200
0
0
50
100
150
Rank of Character
Character Gender
Men
Women


Men and Women by Rank Position (Women Authors)
500
400
300
200
100
0
0
50
100
150
Rank of Character
Character Gender
Men
Women