Gender and verbs across 100,000 stories: a tidy analysis

Previously in this series

- Examining the arc of 100,000 stories

I was fascinated by my colleague Julia Silge's recent blog post on what verbs tend to occur after "he" or "she" in several novels, and what they might imply about gender roles within fictional work. This made me wonder what trends could be found across a larger dataset of stories.

Mark Riedl's Wikipedia plots dataset that I examined in yesterday's post offers a great opportunity to analyze this question. The dataset contains over 100,000 descriptions of plots from films, novels, TV shows, and video games. The stories span centuries and come from tens of thousands of authors, but the descriptions are written by a modern audience, which means we can quantify gender roles across a wide variety of genres. Since the dataset contains plot descriptions rather than primary sources, it's also more about what **happens** at than how an author describes the work: we're less likely to see "thinks" or "says", but more likely to see "shoots" or "escapes".

As I usually do for text analysis, I'll be using the tidytext package Julia and I developed last year. To learn more about analyzing datasets like this, see our online book Text Mining with R: A Tidy Approach, soon to be published by O'Reilly. I'll provide code for the text mining sections so you can follow along. I don't show the code for most of the visualizations to keep the post concise, but as with all of my posts the code can be found here on GitHub.

Setup

We'd start with the same code from the last post, that read in the `plot_text` variable from the raw dataset. Just as Julia did, we then tokenize the text into bigrams, or consecutive pairs of words, with the tidytext package, then filter for cases where a word occurred after "he" or "she".[1]

```
library(dplyr)
library(tidytext)

bigrams <- plot_text %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2, collapse = FALSE)

bigrams_separated <- bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")
```

```
he_she_words <- bigrams_separated %>%
  filter(word1 %in% c("he", "she"))

he_she_words
```

```
## # A tibble: 797,388 × 4
##   story_number                 title word1      word2
##          <dbl>                 <chr> <chr>      <chr>
## 1            1          Animal Farm    he    refers
## 2            1          Animal Farm   she   accuses
## 3            1          Animal Farm    he  collapses
## 4            1          Animal Farm   he  celebrates
## 5            1          Animal Farm    he  abolishes
## 6            2 A Clockwork Orange (novel)    he         is
## 7            2 A Clockwork Orange (novel)    he  describes
## 8            2 A Clockwork Orange (novel)    he    meets
## 9            2 A Clockwork Orange (novel)    he   invites
## 10           2 A Clockwork Orange (novel)    he      drugs
## # ... with 797,378 more rows
```

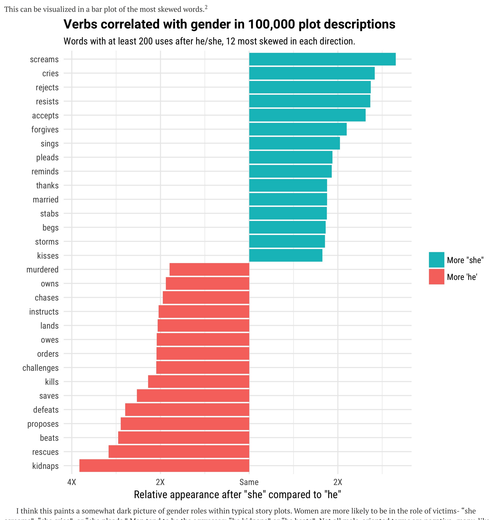
For example, we see the plot description for "Animal Farm" has five uses of a verb after "he", such as "he refers" and "he accuses". (Note that throughout this post I'll refer to these after-pronoun words as as "verbs" since the vast majority are, but some are conjunctions like "and" or adjectives like "quickly").

Gender-associated verbs

Which words were most shifted towards occurring after "he" or "she"? We'll filter for words that appeared at least 200 times.

```
he_she_counts <- he_she_words %>%
  count(word1, word2) %>%
  spread(word1, n, fill = 0) %>%
  mutate(total = he + she,
         he = (he + 1) / sum(he + 1),
         she = (she + 1) / sum(she + 1),
         log_ratio = log2(she / he),
         abs_ratio = abs(log_ratio)) %>%
  arrange(desc(log_ratio))
```

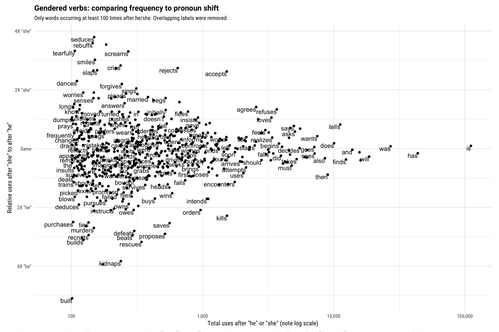
This can be visualized in a bar plot of the most skewed words.[2]



I think this paints a somewhat dark picture of gender roles within typical story plots. Women are more likely to be in the role of victims- "she screams", "she cries", or "she pleads." Men tend to be the aggressor: "he kidnaps" or "he beats". Not all male-oriented terms are negative- many, like "he saves","he rescues" are distinctly positive- but almost all are active rather than receptive.

We could alternatively visualize the data by comparing the total number of words to the difference in association with "he" and "she". This helps find common words that show a large shift.

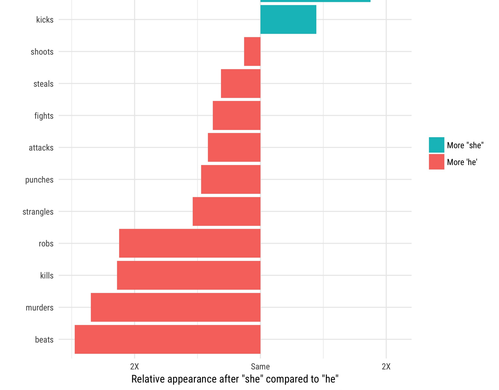
```
he_she_counts %>%
  filter(word1 %in% c("himself", "herself", "she"),
         total >= 100) %>%
  ggplot(aes(total, log_ratio)) +
  geom_point() +
  scale_x_log10(breaks = c(100, 1000, 10000, 1e5),
                labels = comma_format()) +
  geom_text(aes(label = word2), vjust = 1, hjust = 1,
            check_overlap = TRUE) +
  scale_y_continuous(breaks = seq(-3, 3),
                     labels = c("4X 'he'", '2X "he"', "Same", "2X \"she\"", '4X "she"')) +
  labs(x = "Total uses after \"he\" or \"she\" (note log scale)",
       y = "Relative uses after \"she\" to after \"he\",
       title = "Gendered verbs: comparing frequency to pronoun shift",
       subtitle = "Only words occurring at least 100 times after he/she. Overlapping labels were removed.") +
  expand_limits(x = 75)
```



There are a number of very common words ("is", "has", "was") that occur equally often after "he" or "she", but also some fairly common ones ("agrees", "loves", "tells") that are shifted. "She accepts" and "He kills" are the two most shifted verbs that occurred at least a thousand times, as well as the most frequent words with more than a twofold shift.

"Poison is a woman's weapon": terms related to violence

Women in storylines are not always passive victims: the fact that the verb "stabs" is shifted towards female characters is interesting. What does the shift look like for other words related to violence or crime?



There's an old stereotype (that's appeared in works like *Game of Thrones* and *Sherlock Holmes*) that "poison is a woman's weapon", and this is supported in our analysis. Female characters are more likely to "poison", "stab", or "kick"; male characters are more likely to "beat", "strangle", or simply "murder" or "kill". Men are moderately more likely to "steal", but much more likely to "rob".

It's interesting to compare this to an analysis from the Washington Post of real murders in America. Based on this text analysis, a fictional murderer is about 2.5X as likely to be male than female, but in America (and likely elsewhere) murderers are about 9X more likely to be male than female. This means female murderers may be *overrepresented* in fiction relative to reality.

The fact that men are only slightly more likely to "shoot" in fiction is also notable, since the article noted that men are considerably more likely to choose guns as a murder weapon than women are.

Try it yourself

This data shows a shift in what verbs are used after "he" and "she", and therefore what roles male and female characters tend to have within stories. However, it's only scratching the surface of the questions that can be examined with this data.

- **Is the shift stronger in some formats or genre than another?** We could split the works into films, novels, and TV series, and ask whether these gender roles are equally strong in each.

- **Is the shift different between male- and female- created works?**
- **Has the difference changed over time**? Some examination indicates the vast majority of these plots come from stories written in the last century, and most of them from the last few decades (not surprising since many are movies or television episodes, and since Wikipedia users are more likely to describe contemporary work).

I'd also note that we could expand the analysis to include not only pronouns but first names (e.g. not only "she tells", but "Mary tells" or "Susan tells"), which would probably improve the accuracy of the analysis.

Again, the full code for this post is available here, and I hope others explore this data more deeply.

1. Gender is not a binary, so please note that this analysis is examining how the Wikipedia description's author refers to the character. For example, we miss cases of singular they, but it would be challenging to separate it from the more common plural. ↩
2. I'm also skipping the words "himself" and "herself", which are the most gender-shifted words but aren't interesting for our purposes. ↩

David Robinson

Data Scientist at Stack Overflow, works in R and Python.

Email Twitter Github Stack Overflow

Gender and verbs across 100,000 stories: a tidy analysis was published on April 27, 2017.

YOU MIGHT ALSO ENJOY (VIEW ALL POSTS)