Cathy Liu

STAT 153

21820688

The Use of Gender-Specific Pronouns in English-Language Books

I. Introduction

In this project, the question we are studying is: Have the rise and growth of feminism and the progress towards gender equality have affected the use of gender-specific pronouns in English literature? We will examine the usage of the gender-specific pronouns 'he' and 'she' in English-language books over time, and determine whether there are any declines in the use of 'he' or increases in the use of 'she' that correspond to the first-wave (1848-1920), second-wave (1963-1982), and third-wave (1990-2008) feminist movements.

II. Data

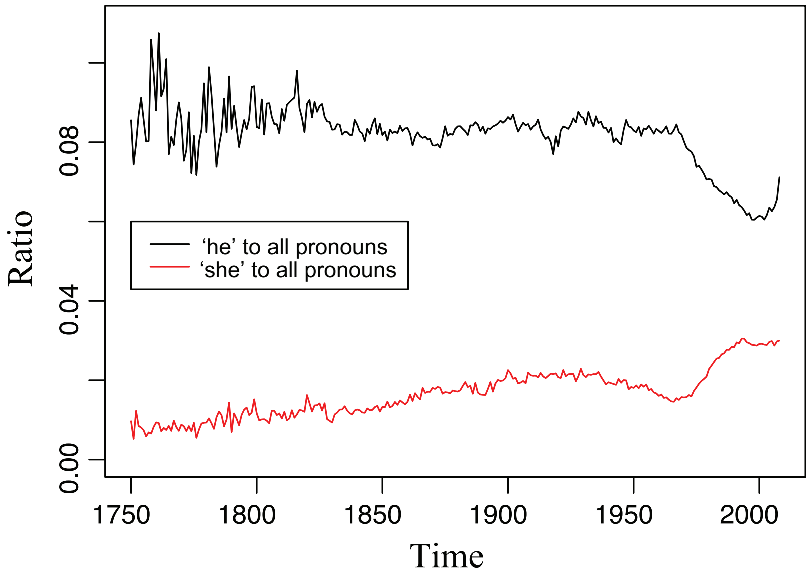
The data to conduct this project was retrieved from Google Books' Ngram data. The data includes the number of occurrences of specific words or phrases in English language books throughout each year. I extracted this data for 'he', 'she', and all pronouns from the year 1750 throughout the year 2008, the most recent year in which Google Books had data. There is data available as early as in the 1500s, but in the earlier years, there is not data available every year and the data is not very consistent, so I will be looking at the data starting from the year 1801 until the year 2008. Because I would have to work with extremely large numbers if I were to use the data as it was, I decided to use ratios for the project so the data would be easier to manage. I then proceeded to calculate ratios: the number of occurrences of ‘he’ to the number of occurrences of all pronouns, and the number of occurrences of ‘she’ to the number of occurrences of all pronouns.

III. Analysis

Preliminary Analysis

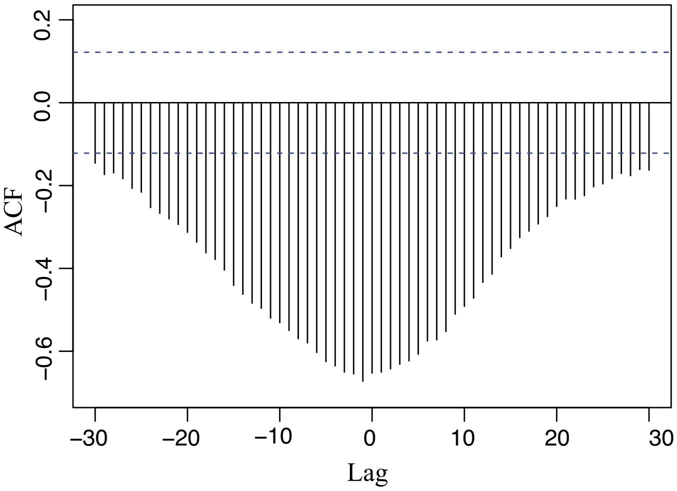
Figure 1 on the next page contains the plots for both the ‘he’ ratio and ‘she’ ratio time series. As we can see from the figure, before approximately 1830, the ‘he’ ratio series is rather random, with relatively constant mean and generally large variance. The ‘she’ ratio series appears to be rather random as well, but it seems to be displaying a slight increase, so the mean is not very constant. After 1830, the ‘she’ ratio series generally increases, while the ‘he’ ratio seems to fluctuate, and it experiences a steep decrease starting in the 1960s, at what seems to be the same rate at which the ‘she’ ratio increases.

Figure 1. Usage of ‘he’, ‘she’ over time



Next, I looked at the cross-correlation function of the ‘he’ ratio and ‘she’ ratio series, which is shown in Figure 2 below. As we can see from the figure, the ‘he’ series and the ‘she’ series are negatively correlated, showing that the two series are perhaps inversely related.

Figure 2. Correlation between ‘he’ & ‘she’



Choosing a model

First, I examined the periodograms, displayed in Figure 3 on the next page, for both the ‘he’ and ‘she’ series. The frequencies as shown in the periodograms proved to be extremely small values, and as such I determined that a harmonic model would not be a good fit for either of the series.

Figure 3a. Periodogram for ‘he’ series

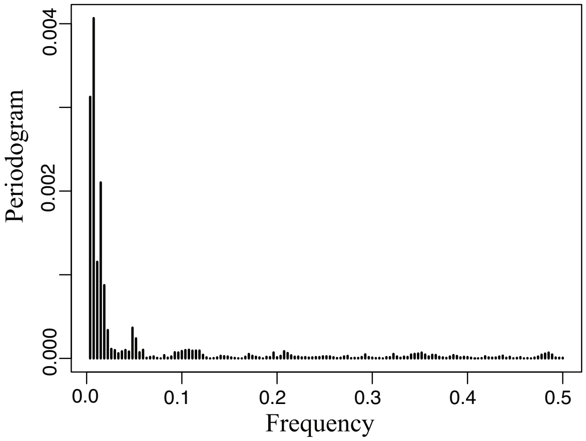
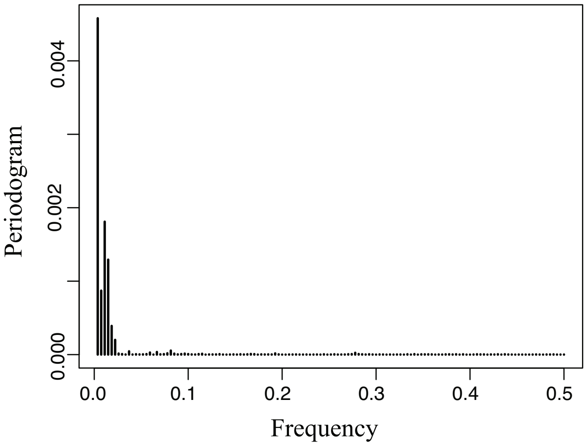


Figure 3b. Periodogram for ‘she’ series



Next, I examined the autocorrelation and partial autocorrelation functions for both series, shown below in Figure 4.

Figure 4a. ACF of ‘he’ series

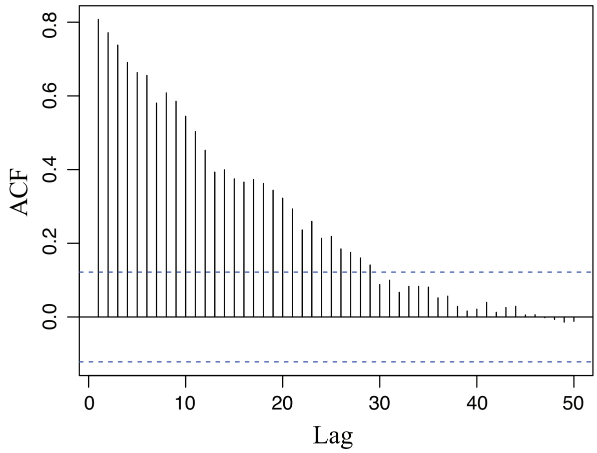


Figure 4b. Partial ACF of ‘he’ series

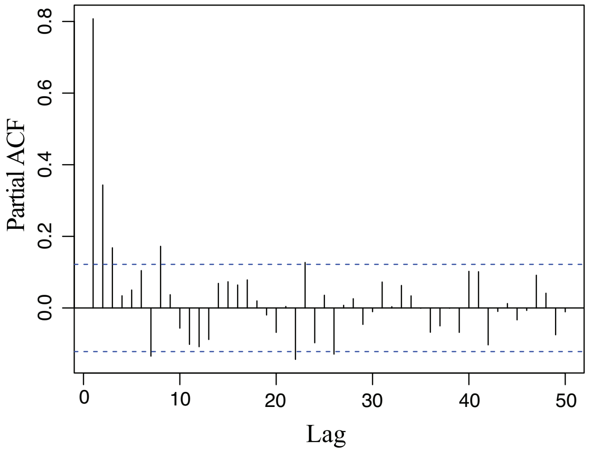


Figure 4c. ACF for ‘she’ series

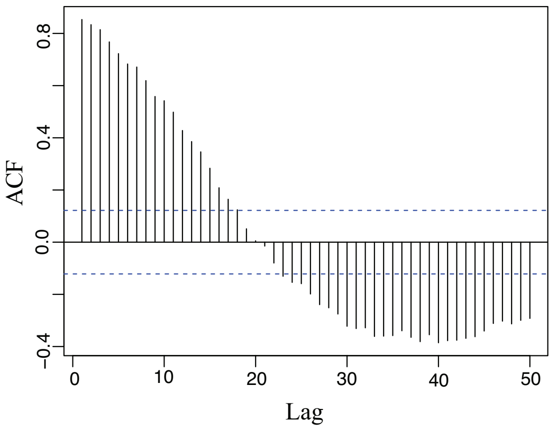
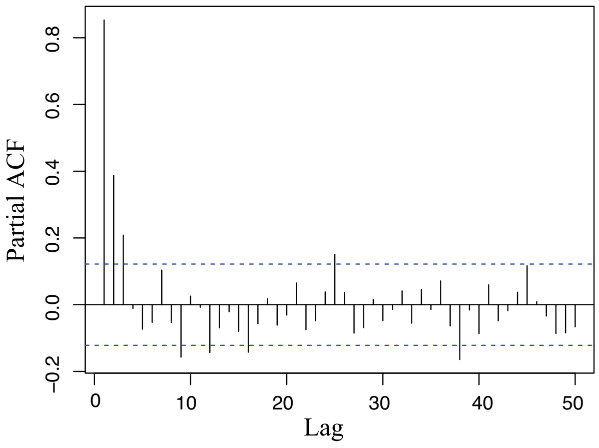


Figure 4d. Partial ACF of ‘she’ series



As seen in Figure 4, both the ‘he’ and ‘she’ series have many significant autocorrelations, as well as a few significant autocorrelations. From this, I determined that an ARMA or ARIMA model is probably best. Since we want to account for the affect that the feminism movements may have on the data, an ARIMA with an *xreg* component. I included in the *xreg* a component for the linear trend and a series *x* of indicator variables – in each year in the duration of a feminist movement, the value of the variable would be 1, and the value of the variable would be 0 otherwise.

To determine which order ARIMA, I experimented with different values for *p*, *q*, and *i* to determine which combination would generate a model with the lowest or most negative BIC, which indicates that that specific model is the best fit for the data. For the 'he' series, the most negative BIC came from the ARIMA model where p = 1, i = 0, and q = 1. The most negative BIC came from the ARIMA model where p = 4, i = 0, and q = 1. As a result, I decided that an ARIMA(1,0,1) would be the best fit for the 'he' series, and an ARIMA(4,0,1) model would be the best fit for the 'she' series. The models with their estimated coefficients are:

ARIMA(1,0,1):

$$Y_t = 0.9406Y_{t-1} + e_t - 0.5528e_{t-1} + 0.1901 - 0.0001t - 0.0012x$$

ARIMA(4,0,1):

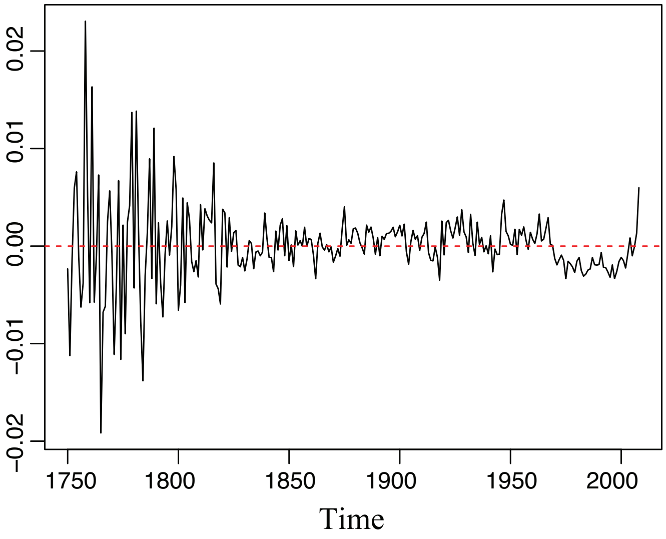
$$Y_t = 1.3822Y_{t-1} - 0.1236Y_{t-2} + 0.0220Y_{t-3} - 0.2883Y_{t-4} + e_t - e_{t-1} - 0.1132 + 0.0001t - 0.0003x$$

(***Note**: Despite the fact that the periodogram exhibited extremely small frequencies, I devised a harmonic model for each of the series. However, the BIC showed that an ARIMA was a better model. See Appendix.*)

Diagnostics

Figure 5 shows the diagnostic plots for the ARIMA(1,0,1) model used to explain the data for 'he'. When we look at Figure 5a at the top of the next page, we see that while the residuals seem to be pretty random, it seems as if the data before and after approximately 1825 are the residuals for two different series – the variance of the residuals before 1825 is on average much larger than that of the residuals after 1825. The mean does seem to be rather constant, however, until approximately 1960; after 1960, the mean of the residuals becomes more negative, and the residuals are no longer white noise, indicating that another model maybe be better to explain the data after 1960.

Figure 5a. Residuals for ARIMA(1,0,1)



Next, examining Figure 5b, the histogram for the residuals of the ARIMA(1,0,1) seems to follow a normal distribution. The QQ plot in Figure 5c further shows that the residuals follow a normal distribution, but the tails are very skinny. Figures 5d and 5e show that the residuals of the model have very few significant correlations, indicating that the model is a good fit.

Figure 5b. Histogram of ARIMA(1,0,1) residuals

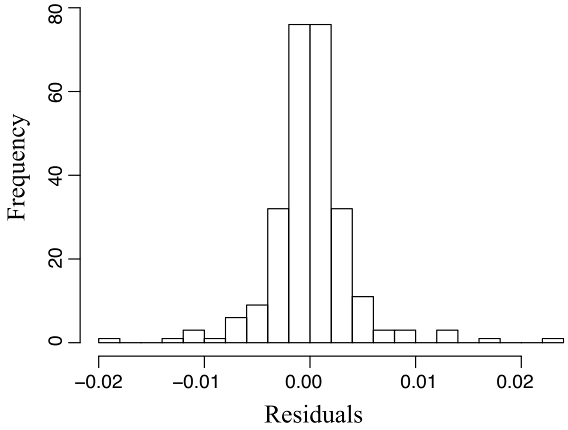


Figure 5c. QQ plot for ARIMA(1,0,1) residuals

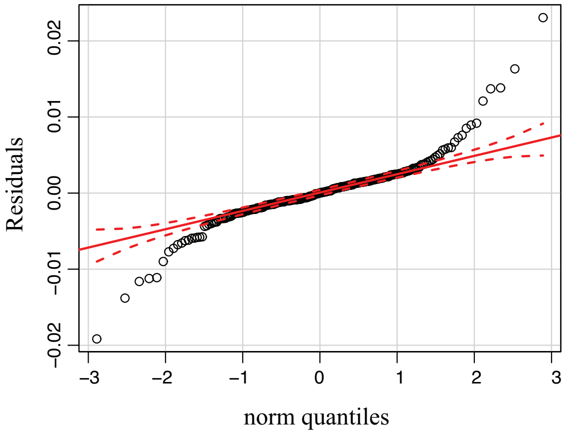


Figure 5d. ACF of ARIMA(1,0,1) Residuals

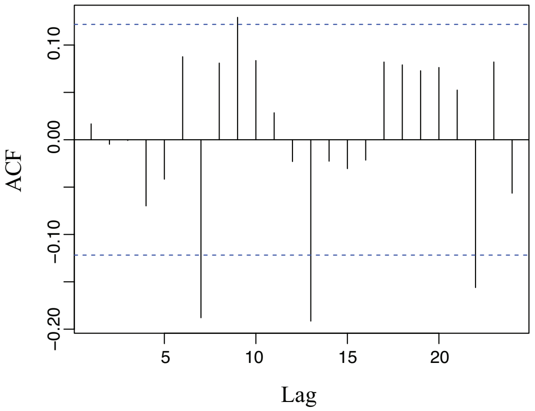


Figure 5e. Partial ACF of ARIMA(1,0,1)

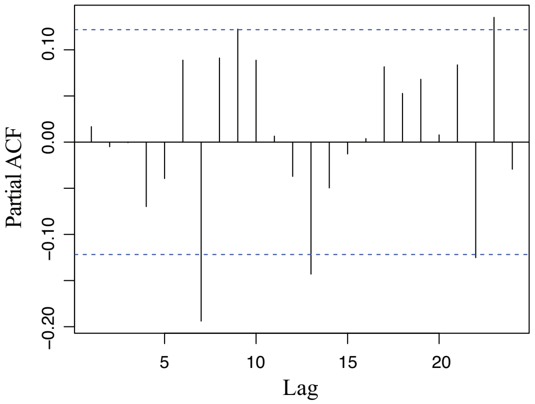


Figure 6 shows the diagnostic plots for the ARIMA(4,0,1) model used to explain the data for ‘she’. Up until approximately the year 1960, the residuals seem to be random and stationary. This suggests that the model is a good fit for the data prior to 1960, but another model may be better for after 1960.

Figure 6a. Residuals for ARIMA(4,0,1)

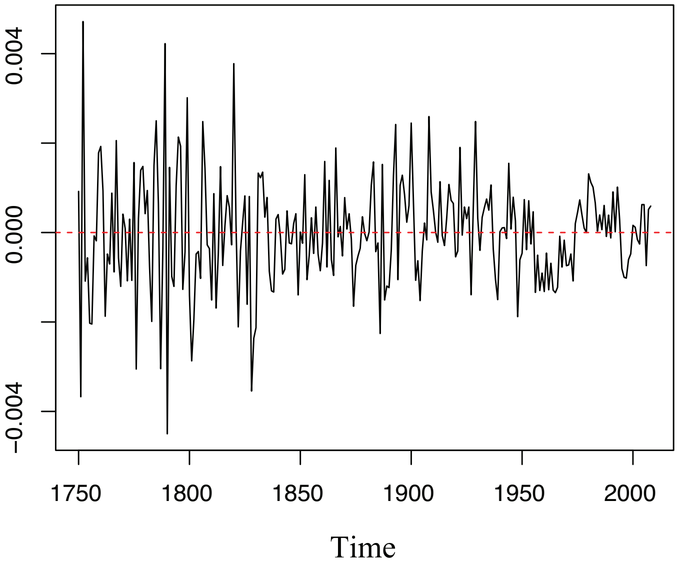


Figure 6b on the next page displays a histogram of the residuals for the ARIMA(4,0,1) model. From the histogram, it seems that the model's residuals follow a normal distribution, though the data seems to be skewed slightly to the right. The QQ plot in Figure 6c also shows that the model's residuals follow a normal distribution, but the tails are quite skinny, and the right tail is not very straight. Very few autocorrelations and partial autocorrelations of the residuals are significant, indicating that the ARIMA(4,0,1) is a satisfactory model for the data.

Figure 6b. Histogram of ARIMA(4,0,1) residuals

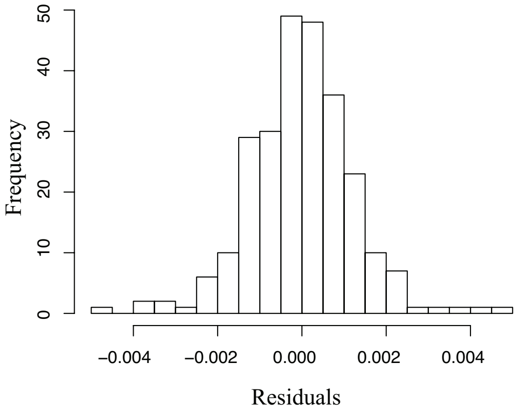


Figure 6c. QQ plot for ARIMA(4,0,1) residuals

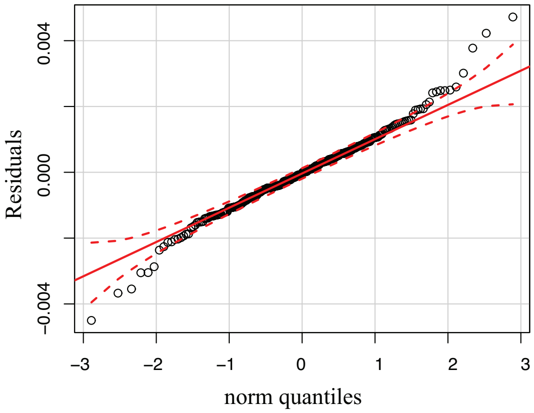


Figure 6d. ACF of ARIMA(4,0,1) residuals

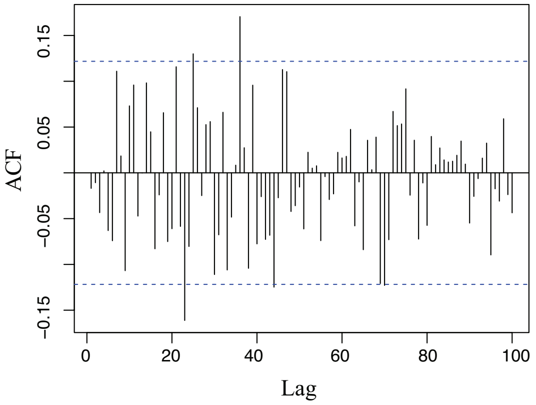
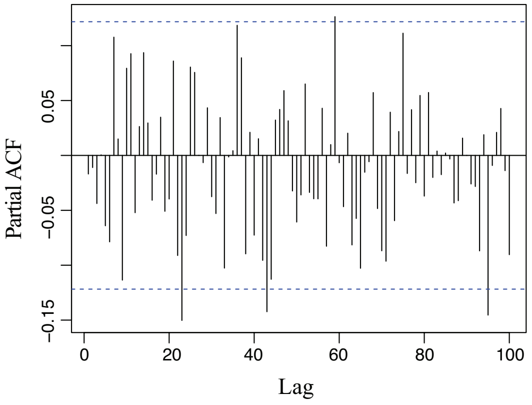


Figure 6e. Partial ACF of ARIMA(4,0,1) residuals



IV. Conclusion

The answer to my question is: from the analyses I have conducted, it seems that the first-wave, second-wave, and third-wave feminist movements have indeed affected the usage of ‘he’ and ‘she’ in English language books.

For the duration of each of these movements, it seems that there are general decreases in the use of ‘he’, and general increases in the use of ‘she’ in comparison. There are, however, also some increases in the usage of ‘he’ and decreases in the usage of ‘she’ during these time periods. As such, we cannot say for sure that the feminism movements are the only reason for any changes of usage in these two pronouns. It also seems that the feminist movements have affected the usage of ‘she’ more than the usage of ‘he’, and the later part of the third-wave feminist movement did not have as much effect on the usage of either of these pronouns.

Also, it is interesting to note that post-feminism, before the start of the nineteenth century, the usage of ‘he’ is a rather stationary series, but the ‘she’ series already exhibits a gradual increase, which can be attributed to the introduction of feminist ideals during the Enlightenment.

V. References

"First-wave Feminism." *Wikipedia*. <http://en.wikipedia.org/wiki/First-wave_feminism>.

"Second-wave Feminism." *Wikipedia*. <http://en.wikipedia.org/wiki/Second-wave_feminism>.

"Third-wave Feminism." *Wikipedia*. <http://en.wikipedia.org/wiki/Third-wave_feminism>.

“Protofeminism.” *Wikipedia*. <http://en.wikipedia.org/wiki/Protofeminism>.

Google Books Ngram. <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>.

Cryer, Jonathan D., and Kung-sik Chan. *Time Series Analysis: With Applications in R*. New York: Springer, 2008.