

# Multivariate Analysis Applied to Differences in Literary Style

## *GSO Statistics Talk*

Roger D. Peng

November 1, 2000

## Previous Work

- Mendenhall's characteristic curves in 1901 (reproduced by C. B. Williams)
- F. Mosteller and D. Wallace: *The Federalist*
  - 77 essays, 5 written by John Jay, 43 written by Alex Hamilton, 14 written by James Madison, 12 disputed between Hamilton and Madison, 3 joint papers between Hamilton and Madison
  - Used function words and many different discrimination techniques
- A. Q. Morton: Greek prose
- B. Efron and R. Thisted: The Taylor poem (1985)
  - Estimated the rate of discovery of new words
- D. I. Holmes: Mormon scripture
- Don Foster: The Night Before Christmas

## Problems

- Data are not in a manageable format – often took months to process
- Mosteller and Wallace used grad students and secretaries to count words
- Subjectivity
  - “We know a work is genuine because we can see in them the mind and style of the author and the external evidence agrees with this judgement.”
  - “We know the mind and style of the author because we see them in the genuine works.”

## Assumptions

- The genuine works of an author form a single and stable population within which works differ from each other only by the expected differences of random sampling
- Validity depends on the unit of analysis

## Basics

- What should be the unit of analysis?
  - Word counts (function words)
  - Syllables
  - Parts of speech – verb/noun ratio
  - Word length
  - Sentence length
- I used 69 function words from the Miller-Newman-Friedman list
- Justification: Authors don't think about the way they use words like "and" or "the" so the counts should be relatively stable within each author.
- a all also an and any are as at be been but by can do down even every for from had has have her his if in into is it its may more must my no not now of on one only or our should so some such than that the their then there things this to up upon was were what when which who will with would your

## Data

- Documents are divided into blocks of 1700 words
- In each block, counts of each word on the word list are tallied

Author	Dates Lived	# of Blocks
Marlowe	1564-1593	56
Shakespeare	1564-1616	179
Milton	1608-1674	56
Austen	1775-1817	437
Dickens	1812-1870	598
Doyle	1859-1930	552
Kipling	1865-1936	157
Cather	1873-1947	237
London	1876-1916	299

## Data Structures

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \Rightarrow \begin{bmatrix} \text{a} & \text{all} & \text{also} & \text{an} & \text{and} & \dots \\ 40 & 34 & 54 & 20 & 78 & \dots \\ 35 & 10 & 0 & 7 & 44 & \dots \\ 46 & 2 & 0 & 3 & 4 & \dots \\ 40 & 12 & 31 & 6 & 14 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} = X$$

$$\text{Between Groups Covariance Matrix} = \frac{1}{n} X^T G (G^T G)^{-1} G^T X$$

$$\text{Total Covariance Matrix} = \frac{1}{n} X^T X$$

## Canonical Discriminant Analysis

- Find linear combinations of the predictor variables (word counts) which maximize the between groups variance, subject to a constraint on the total variance. Between and Within!
- We get the following eigenvalue decomposition

$$[\text{Between Groups Covariance}] \beta = \lambda [\text{Total Covariance}] \beta$$

$$X^T G (G^T G)^{-1} G^T X \beta_i = \lambda_i X^T X \beta_i$$

- The  $\beta_i$ 's are the *discriminant functions*. There are  $L = \min(p, g - 1)$  non-trivial eigenvalues
- If  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_L$  are the non-trivial eigenvalues, then

$$\frac{\lambda_i}{\sum_{j=1}^L \lambda_j}$$

is the proportion of variance explained by the  $i$ th discriminant function.

## Canonical Vector Plots

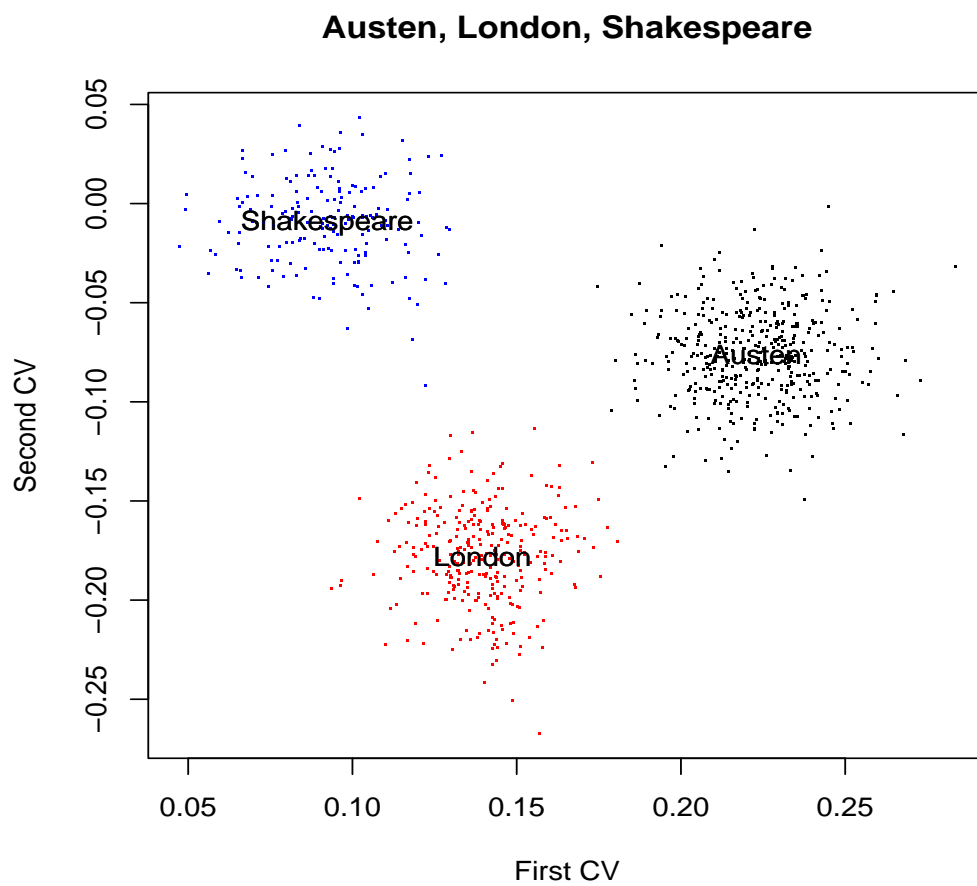
- If  $\beta_1, \beta_2, \dots, \beta_L$  are the discriminant functions and  $X$  is the data matrix of word counts, then  $Y_i = X\beta_i$  is the  $i$ th canonical vector/ivariate.
- Plot  $Y_1$  vs.  $Y_2$  and observe separation.

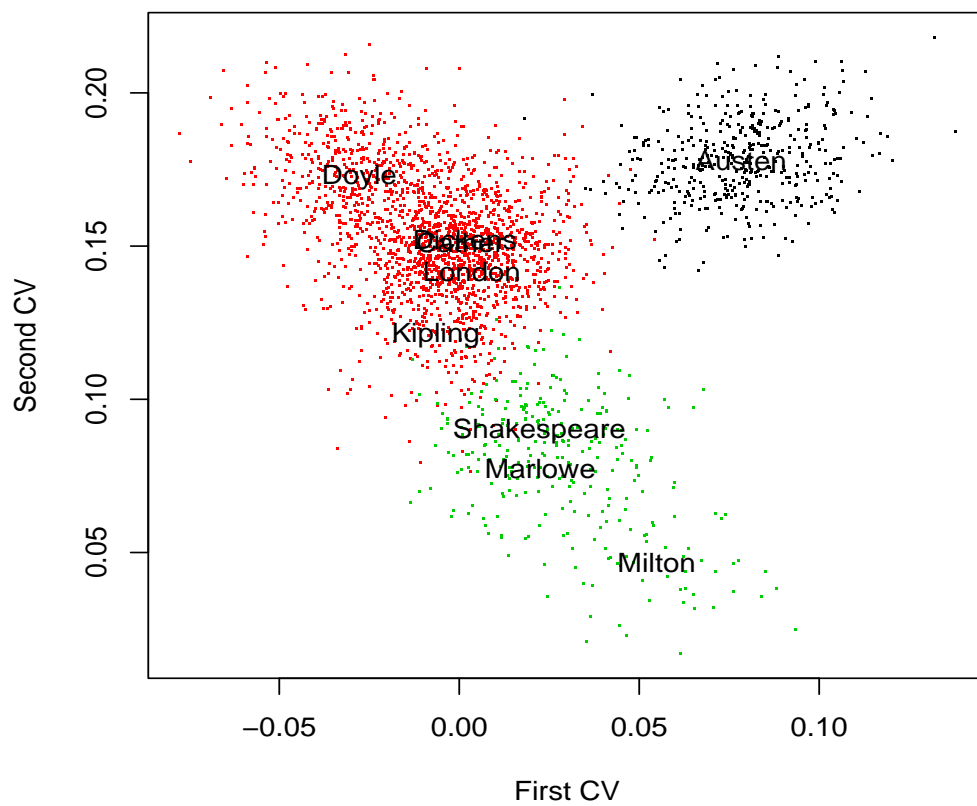
## Loadings

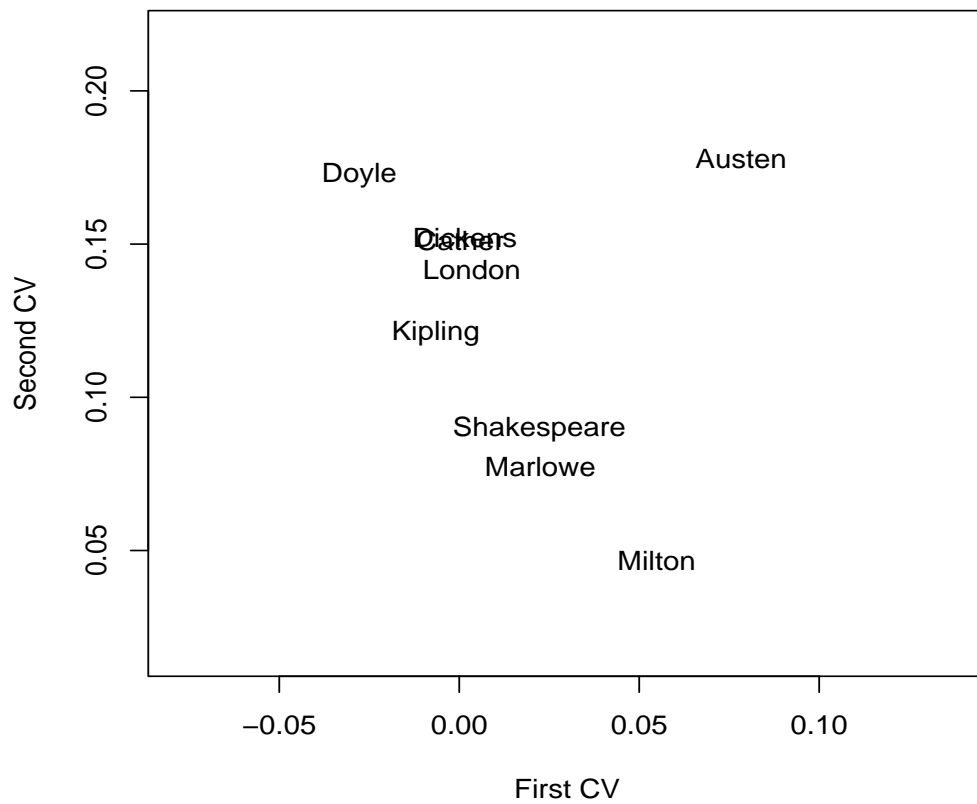
- The *loadings* are the correlations between the columns of  $X$  and columns of  $XB$ , where

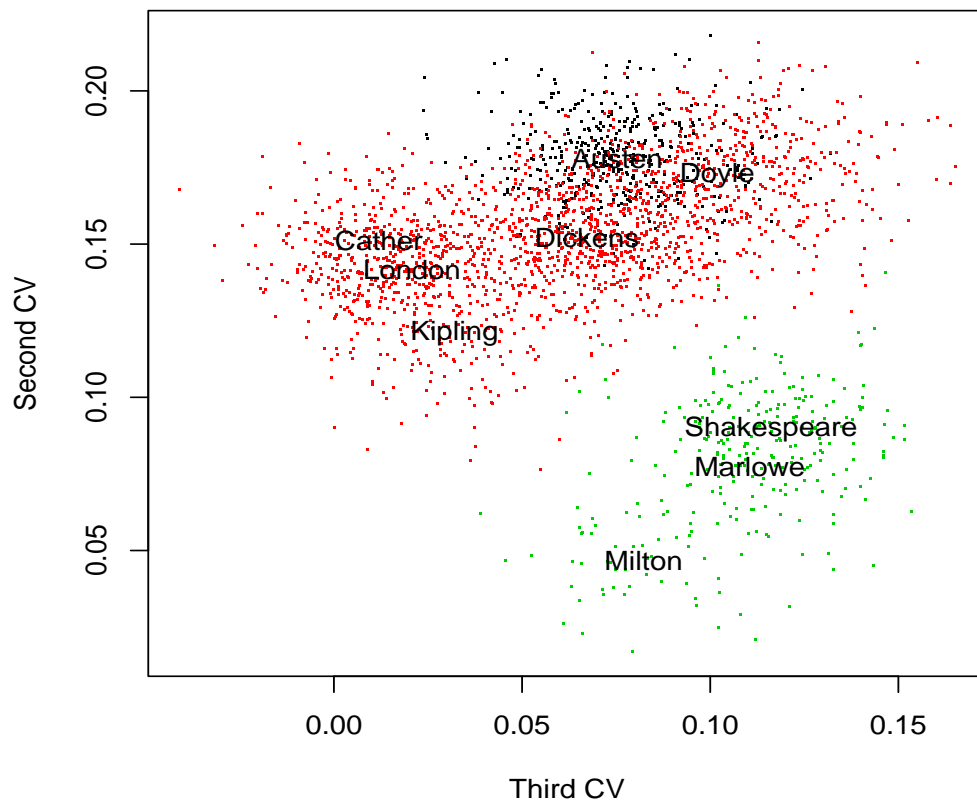
$$XB = X \begin{bmatrix} \vdots & \vdots & & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_L \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

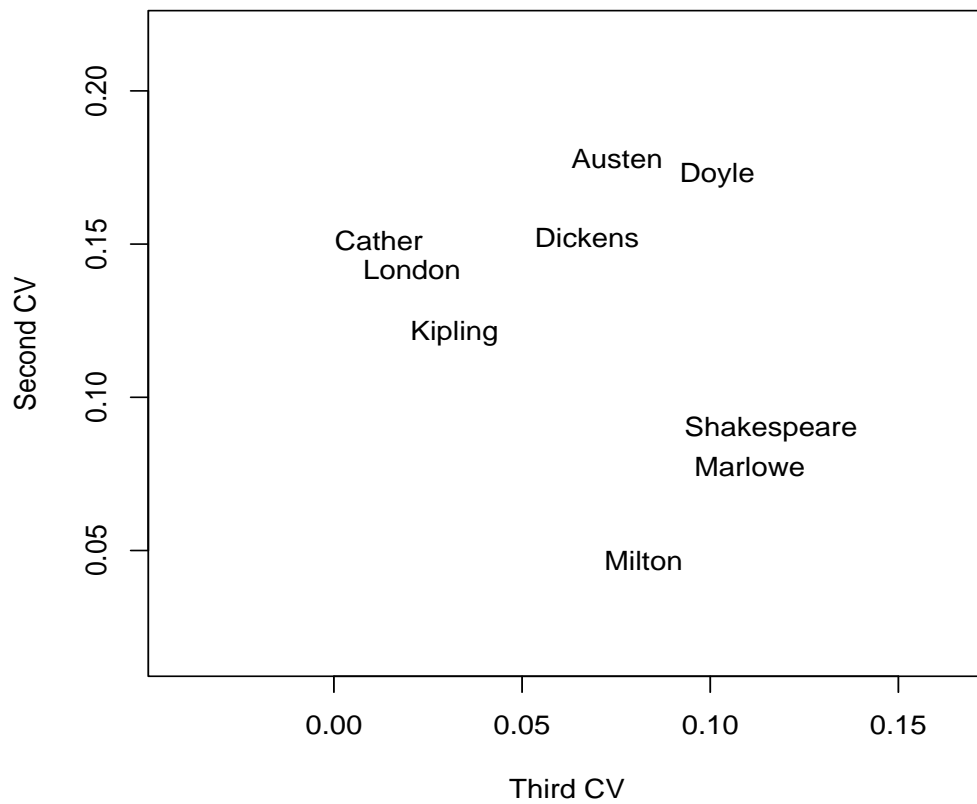
- The loadings can identify each canonical vector with a small set of variables.

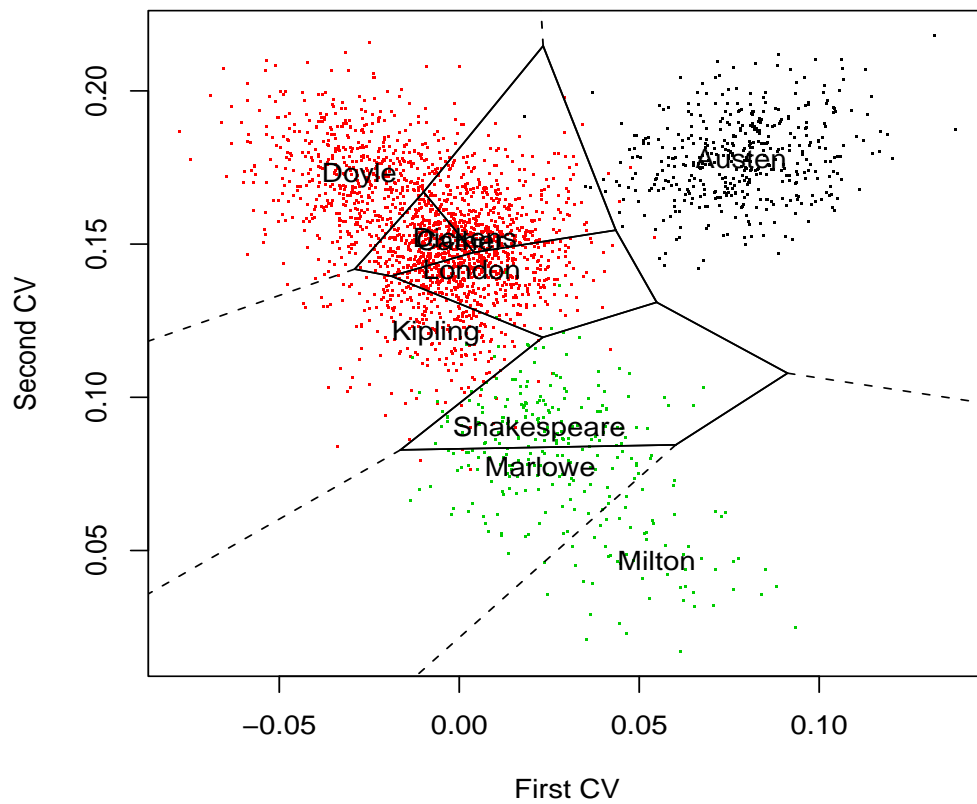


**All Authors**

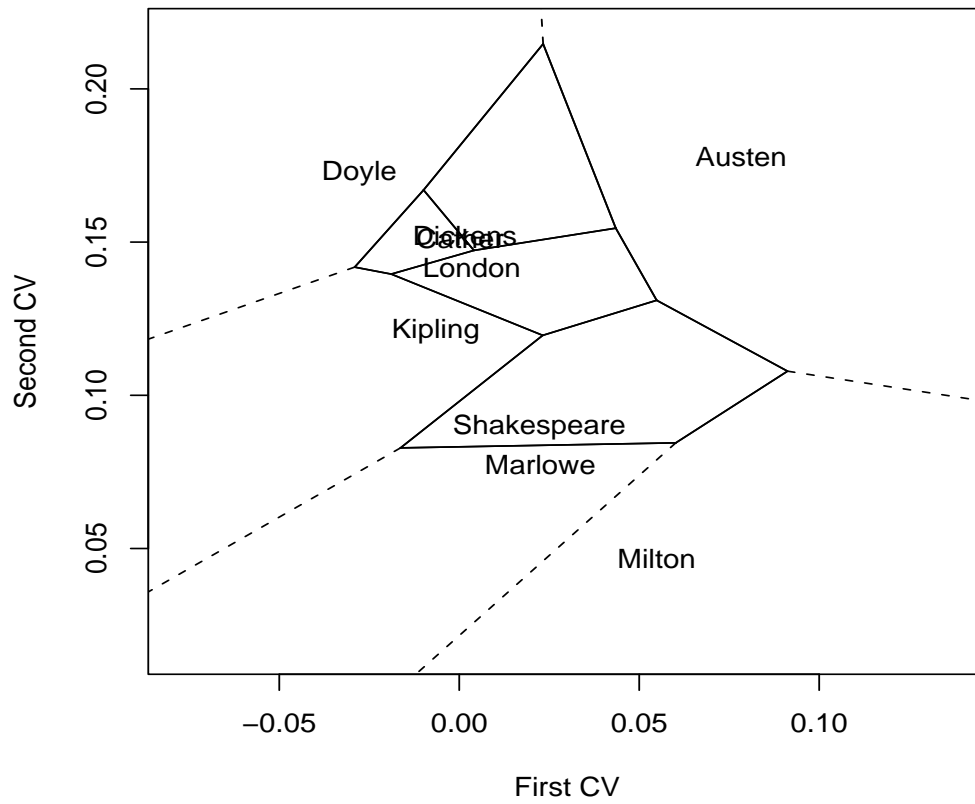
**All Authors**

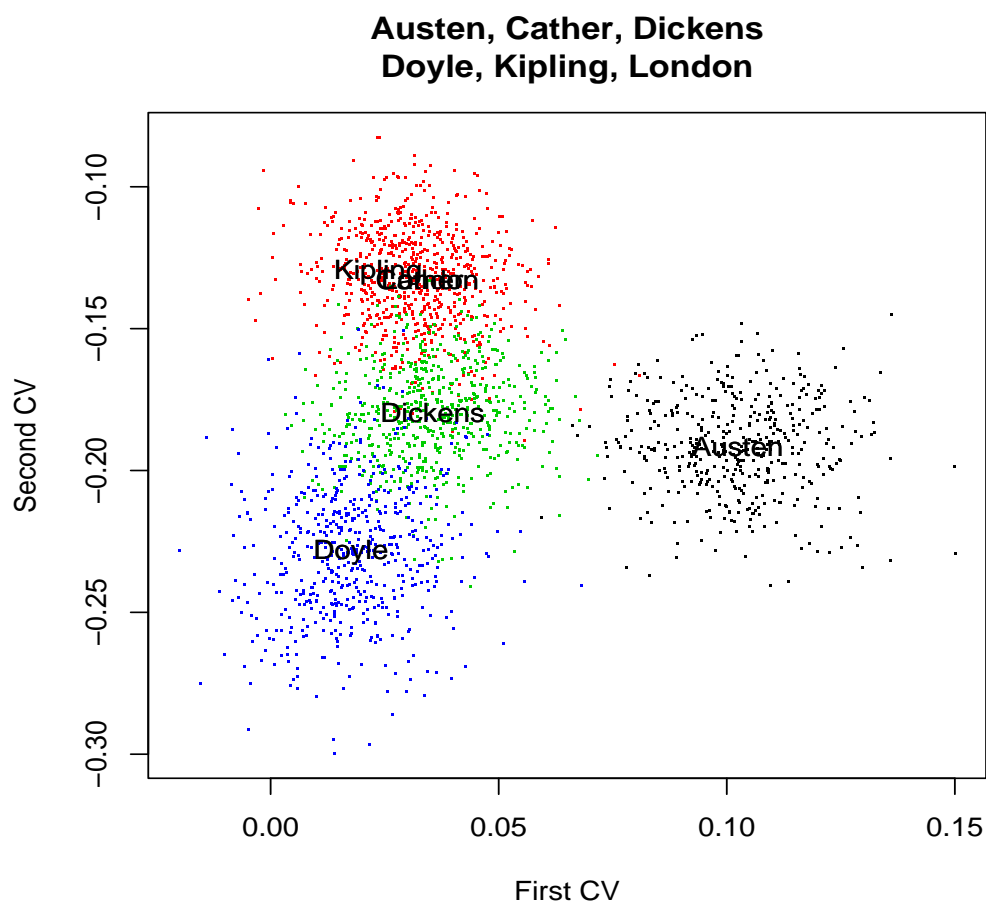
**All Authors**

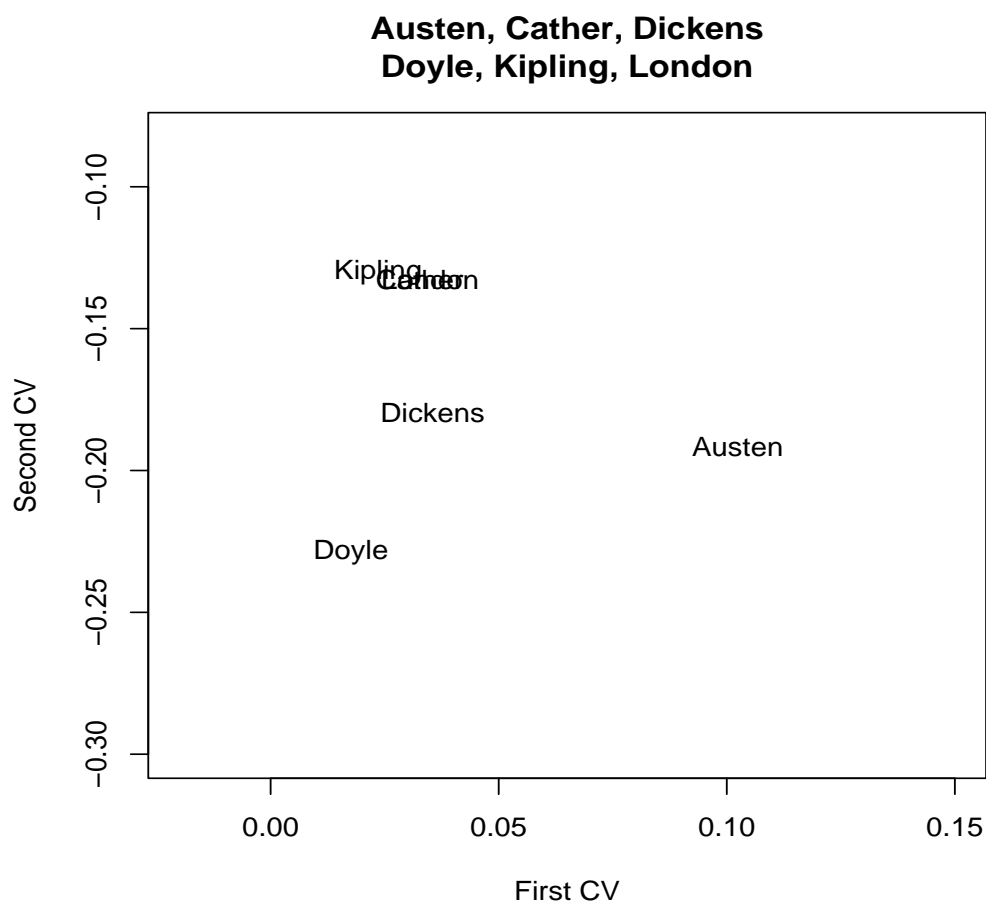
**All Authors**

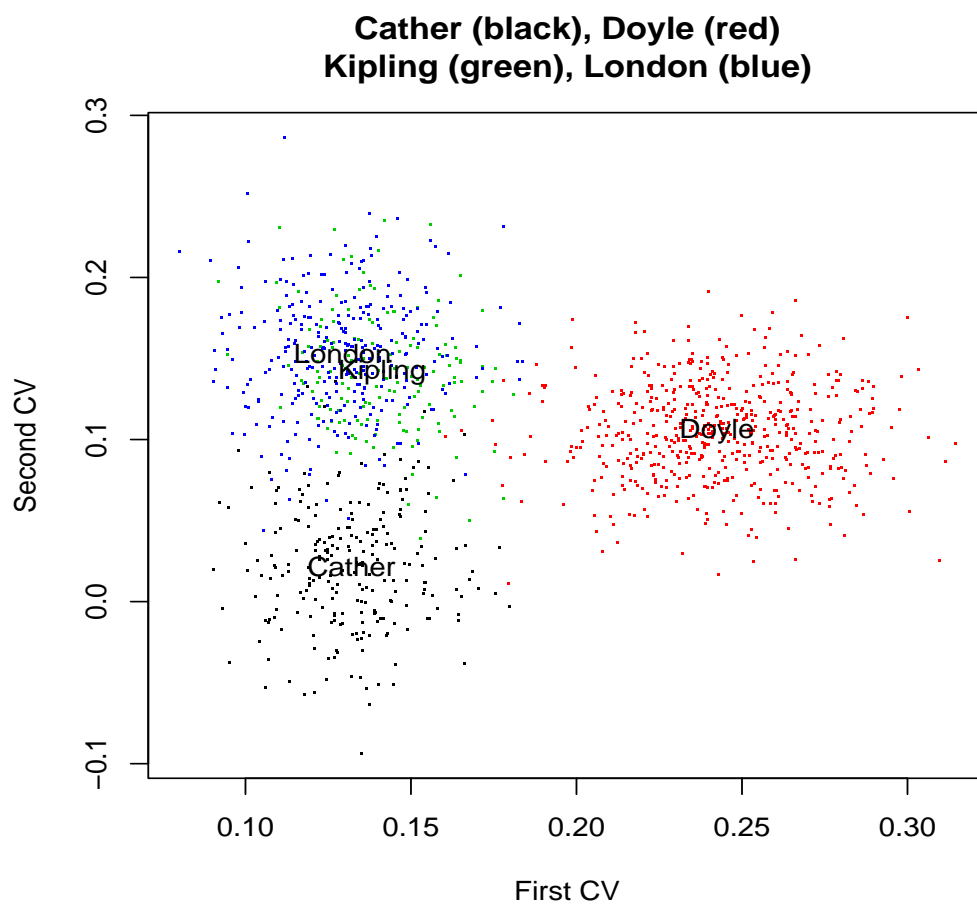
**First Two Canonical Vectors for All Authors**

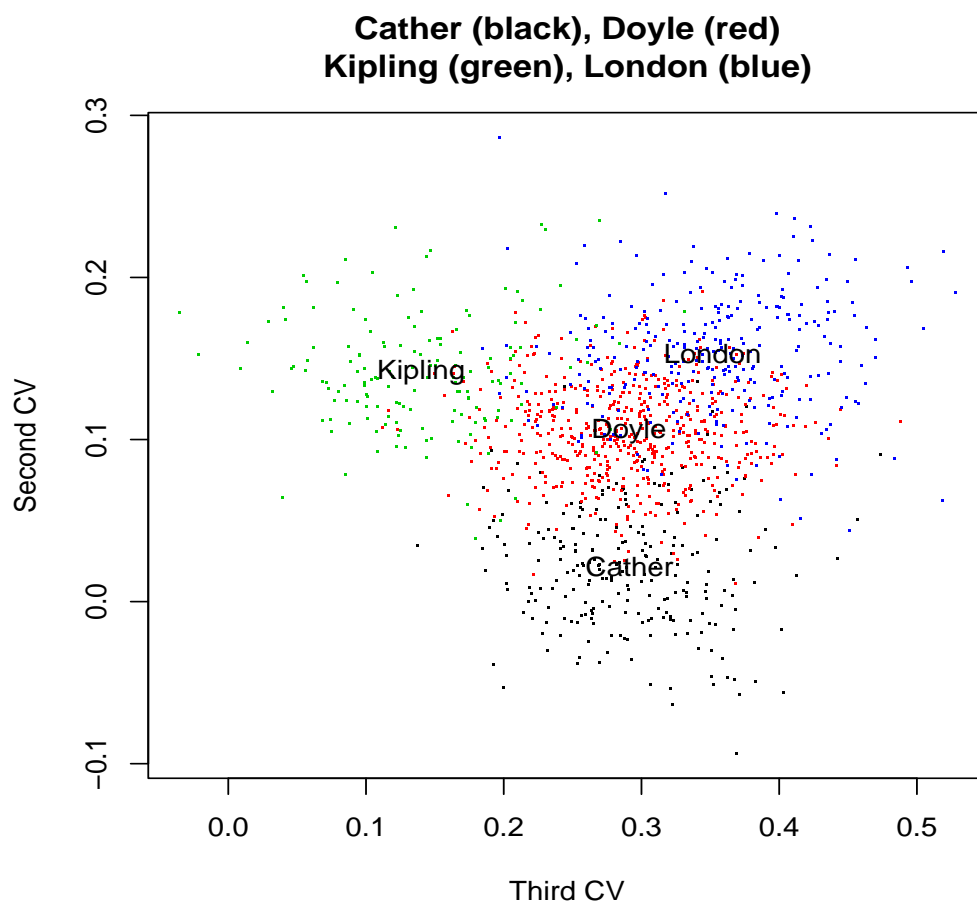
### First Two Canonical Vectors for All Authors

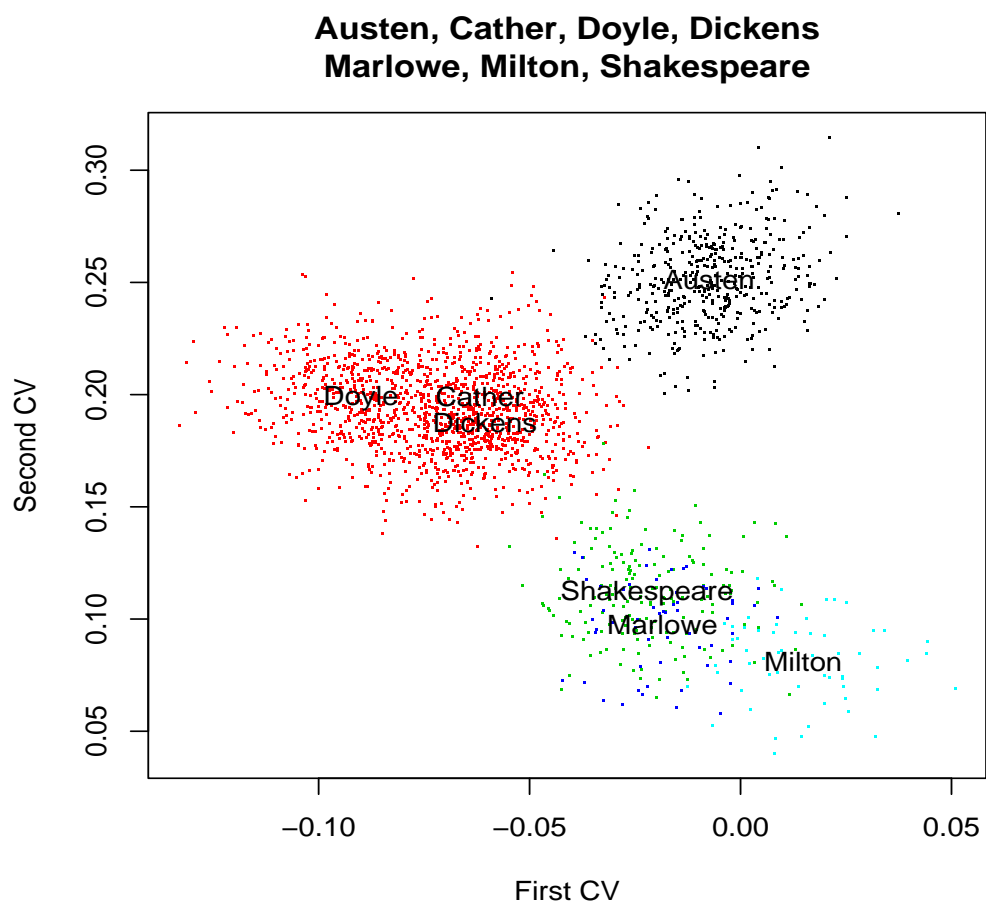


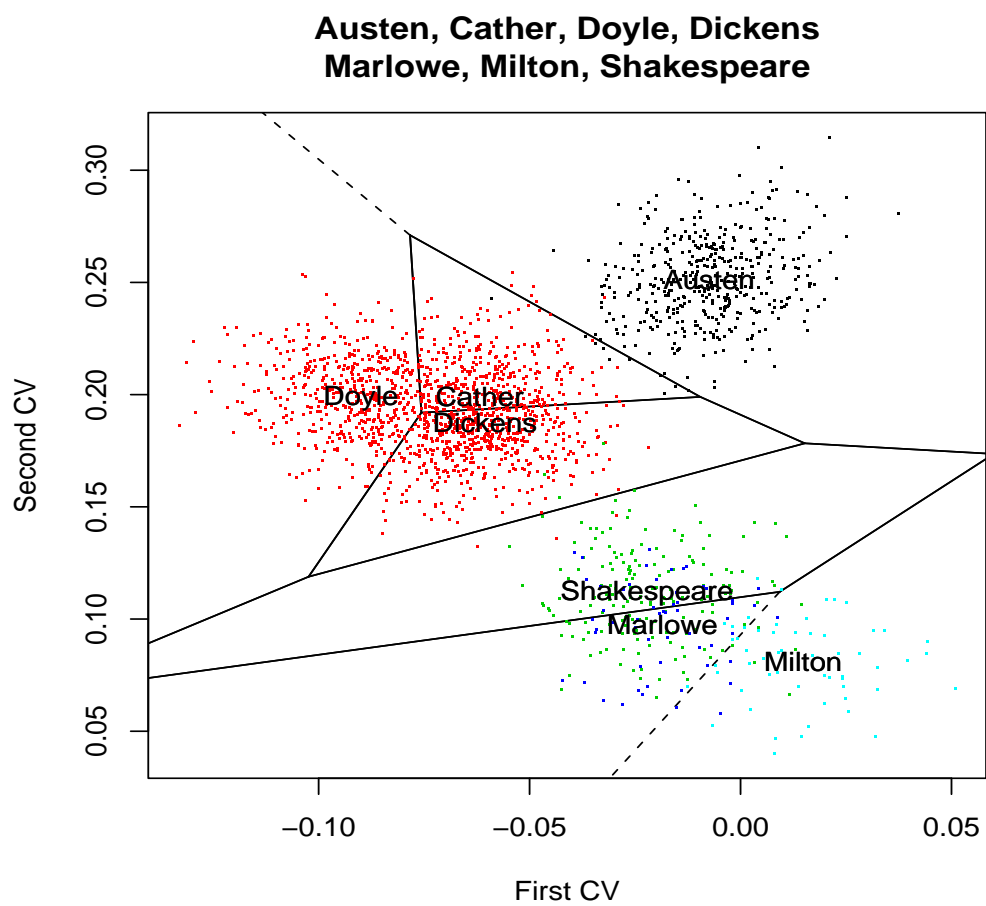




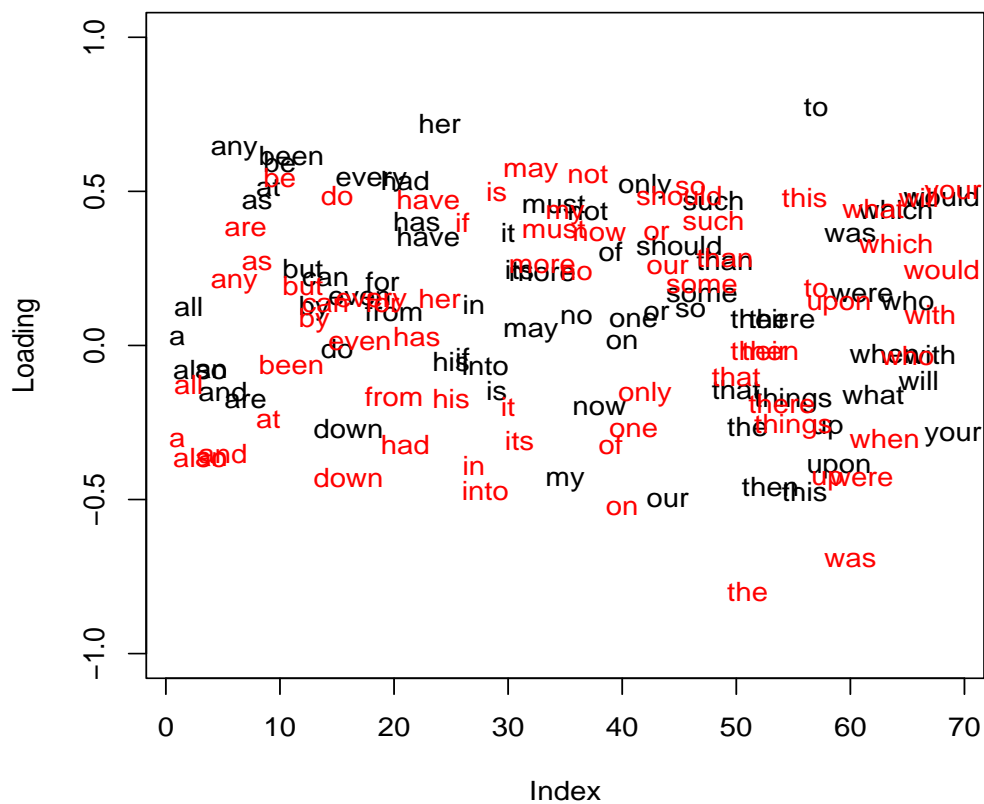






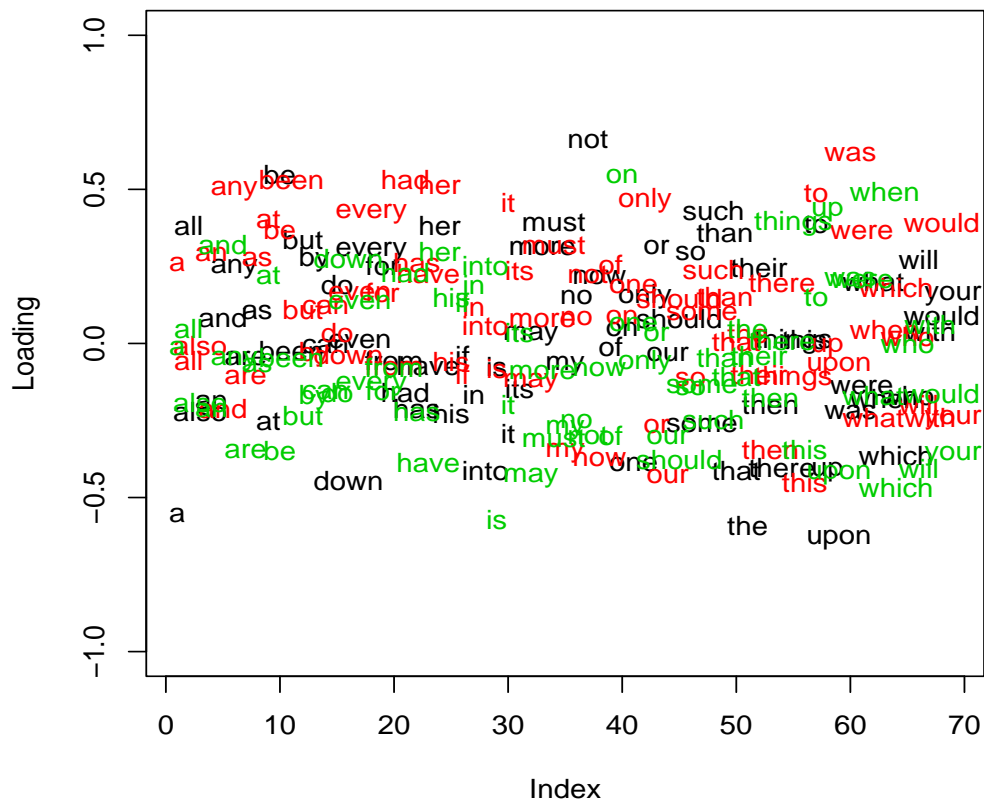


### Loadings for Austen, London, Shakespeare

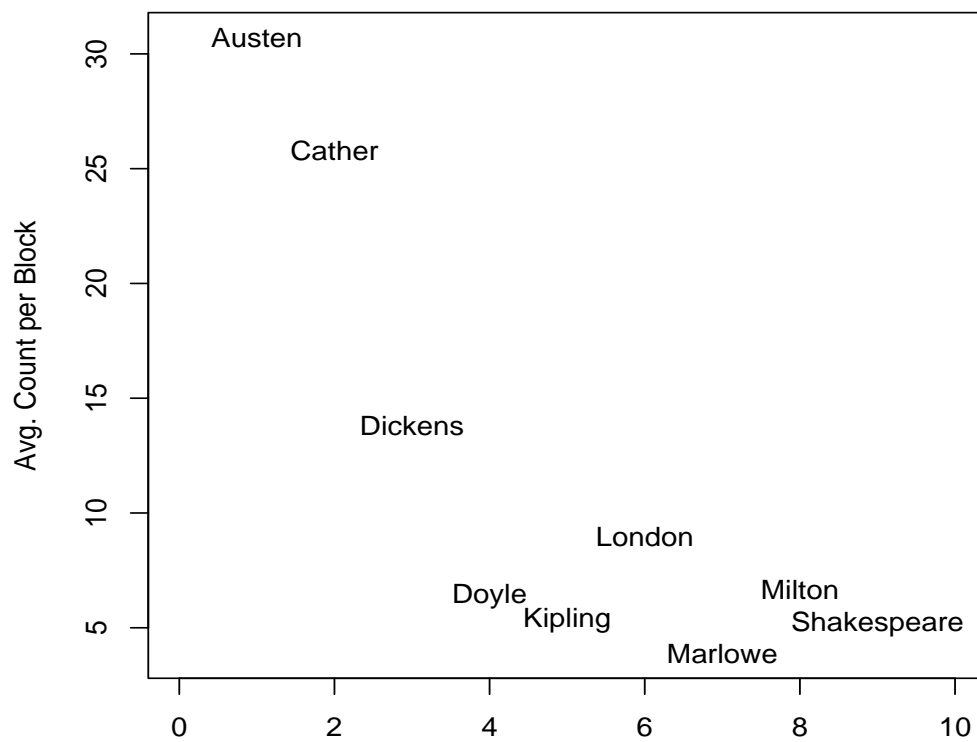




### Loadings for Austen, Cather Dickens, Doyle Marlowe, Milton, Shakespeare

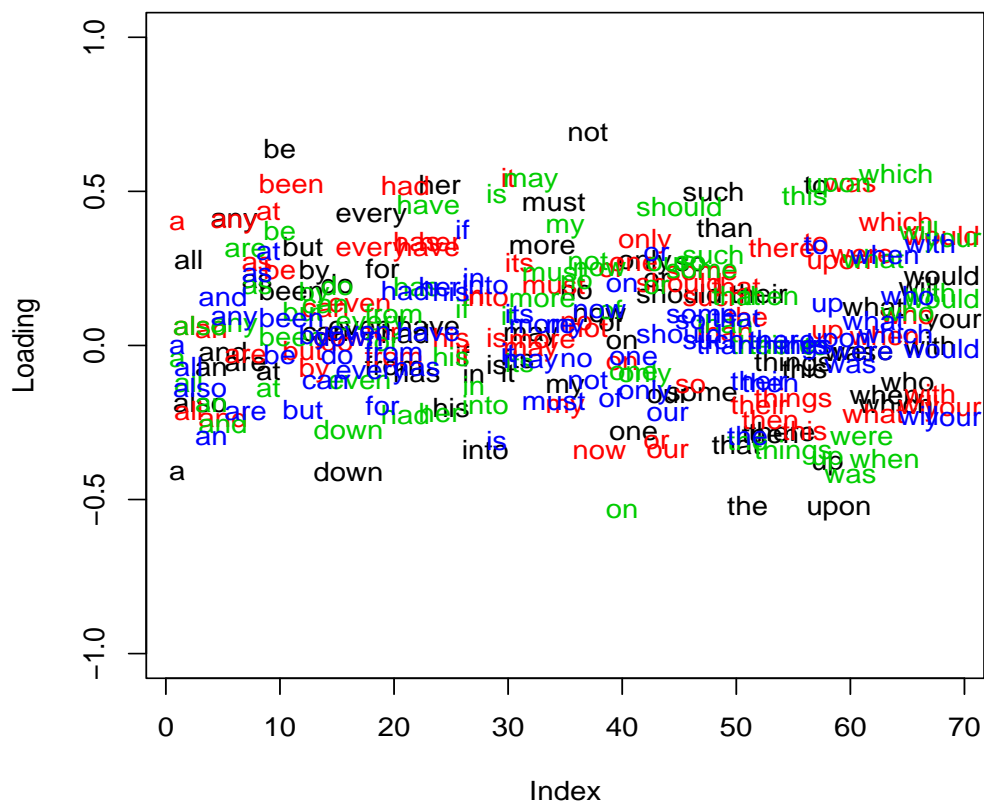


### Usage of the word 'her'





## Loadings for All Authors



## Other Applications

- Text processing
- Document indexing
- Searching and querying