Multivariate Analysis and the Study of Style Variation

by David Hoover (English Department, NY University)
Authorship Attribution & Stylistics

- **Authorship Attribution** is classifying texts by author.
- **Stylistics** is the study of style and its variations.

Question: can methods of authorship attribution be used to track stylistic variations within the writing(s) of a single author?

- Method in question: multivariate analysis of counts of the top x frequent words
- Doesn’t AA assume that these variables are consistent across the author’s writings?
Books to be abused

- **1984** by George Orwell
  - chapter by Emmanuel Goldstein

- **The Inheritors** by William Golding
  - chapter by Tuami

- The Good Soldier by Ford
- Howards End by Forster
- Sons and Lovers by Lawrence
- Of Human Bondage by Maugham
Authorship Attribution Testing

- removed dialogue, personal pronouns, *any words for which a single section provides more than 80 % of occurrences* i.e. names, places, etc.
- split the novels into 18 25,000-word sections
- applied clustering and PCA with the number of top frequent words (of all POS) used varying over 50, 100, 200, 300, 400, 500, 600, 700, 800
Authorship Attribution Testing 1: clustering

- clustering does best with 800 most frequent words, but succeeds with as little as 10
Authorship Attribution Testing 2: PCA

- succeeds with as little as 20 most frequent words
- does best with 600: the first two PCs account for 42% of the total variance
Sounds promising for stylistics:

1984

- large # of features used → Goldstein’s book correctly attributed to Orwell
- small # of features used → clustering and PCA fail
- Focusing on 1984 alone, it’s clearly separated:
Sounds promising for stylistics: Inheritors

- 3 languages: A (Lok), B (transition), C (Tuami)
- Differences: syntactic patterns and vocabulary
- Sections s.t. every one contains only one language results in promising clustering using 500-800 features:
Modification of the Technique

Instead of finding the $x$ most frequent words in the whole corpus, either don’t ignore the words with 80% of occurrences in a single section, or:

- count frequency of every word in every section
- for each word, keep only the highest count
- retain $x$ words with the highest counts

Difference:

- in addition to the most frequent words overall, retain more oddly distributed words
- words with higher frequency variation get higher ranks
Results on 1984 and Inheritors

- slight improvement for 1984 when not ignoring 80%-in-a-section words
- big improvements in separating language C and slight for language B
Another case: The Picture of Dorian Gray

- chapter about studies of spiritual mysteries
- Results of the standard AA method improve slightly by not ignoring the 80%-in-one-section words, but the modified method with 300+ most fr. words does best
Conclusion

- Hoover came up with a modification of the AA clustering method that allows to study the style variations within texts.
- Difficulty in future work: lack of known instances of style variation.
- Is this really style variations?
  - different genre in 1984
  - different narrator in *The Inheritors*
  - unusual topic in *Dorian Gray*