SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES

Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans Sòisealta na h-Alba

Authorship
and
Stylometry

0930 Wednesday 18 April
marc.alexander@glasgow.ac.uk

Stylometry

The "measurement" of style

- Sometimes called computational stylistics or computational text analysis

Used for:

- genre classification
- diachronic linguistics
- literary analysis
- authorship attribution

Discriminators

Criteria which make a style distinguishable

- And, for our purposes, that computers can count

Need *closed* questions, not open ones

- 'If X wrote A, did X also write B?'
- 'Who of X, Y, or Z is more likely to have written A?'
- **NOT** 'Who wrote X?'

Style

Must begin with an understanding of style

Authors' styles change over time and with their purpose:

- style of author
- style of author at a particular time
- style of author in a particular way
- style of author when writing for a particular person
- style of author when writing in a particular genre

Style

Style can be:

- subconscious ("real" style)
- conscious (pastiche and parody)

So questions are better if they don't ask:

- Is this like Jane Austen's style?

And instead ask:

- Is this like Jane Austen's style of writing the speech of female characters in her middle novels?

Stylometry examples

Federalist Papers

- Mostellller and Wallace, 1964; closed set of authors (Hamilton, Madison, Jay)

Primary Colors

- *Roman à clef* about US politics and a Bill-Clinton-like governor

The Book of Mormon, Jockers *et al*.

The style of Henry James; David Hoover

Douglas Biber *et al.* on register analysis

Stylometry

Word/sentence length

Frequencies of letter pairs

Vocabulary richness

Word frequency

Selected sets of words

- Mosteller and Wallace: *upon, whilst, there, on, while, by, consequently, would,* etc

Most frequent words (MFW)

MFW

Best technique

Focuses on function words

Very successful (often very good in 500-4000 range)

Usually needs *culling*

Once you have your data...

You have multivariate data

- data with lots of variables

You then need the 'distance' between each variable

- We use Burrows' *Delta*, the most often used measure

Don't *really* need to understand the inner workings of Delta, but there are references at the end of the slides…

Analysing Multivariate Data

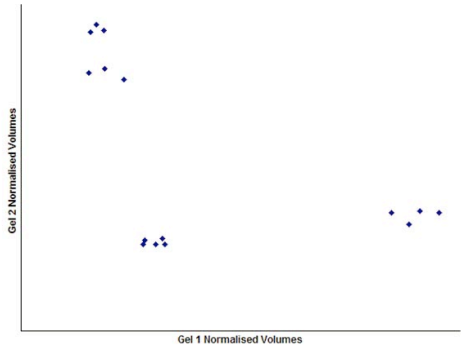
Principal Component Analysis (PCA)

Cluster Analysis

Bootstrap consensus

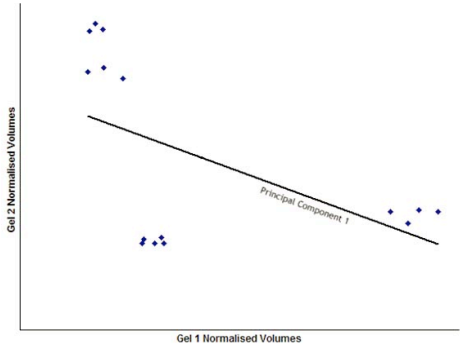
PCA

PCA

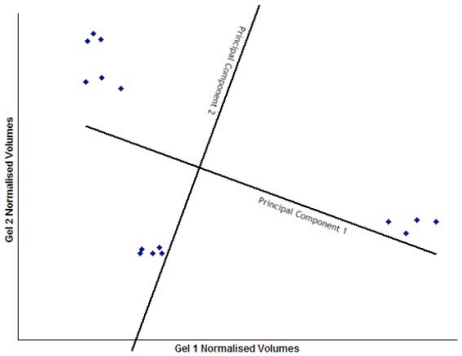


Graphs from http://www.nonlinear.com/support/progenesis/samespots/faq/pca.aspx

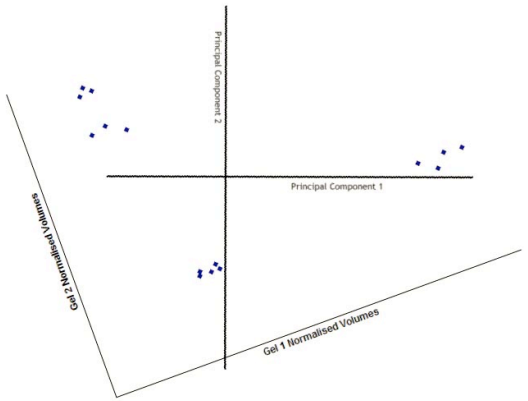
PCA



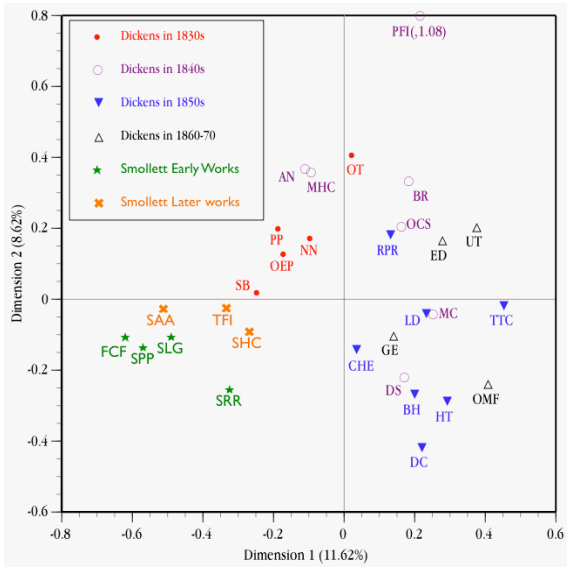
PCA



PCA



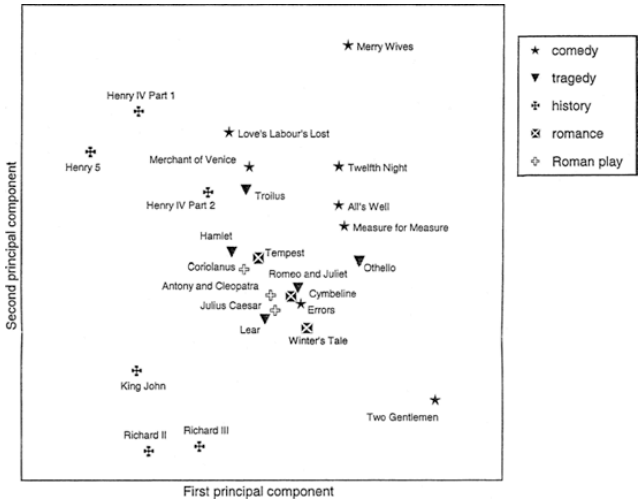
PCA

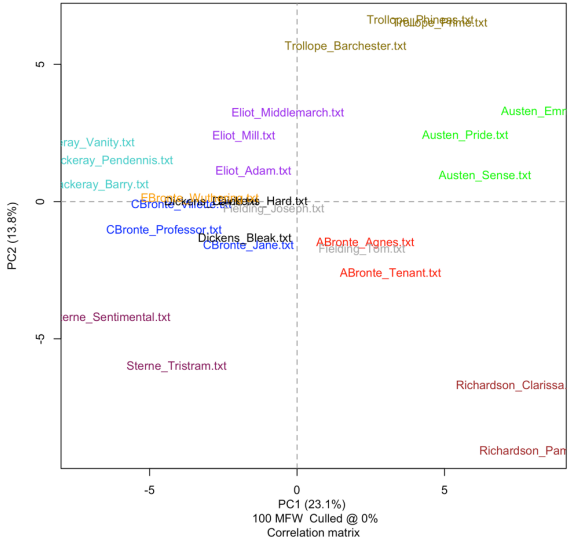


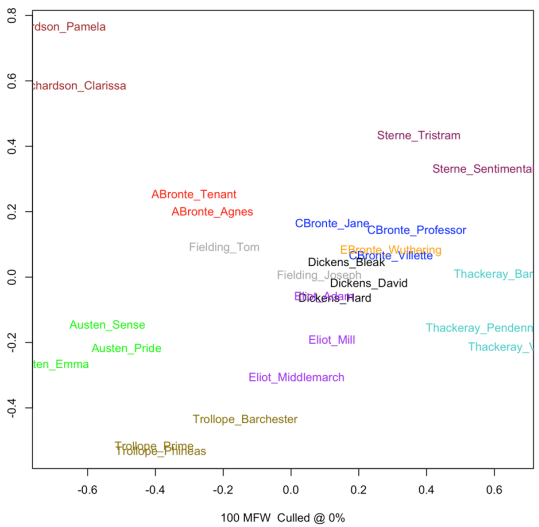
Tomoji Tabata

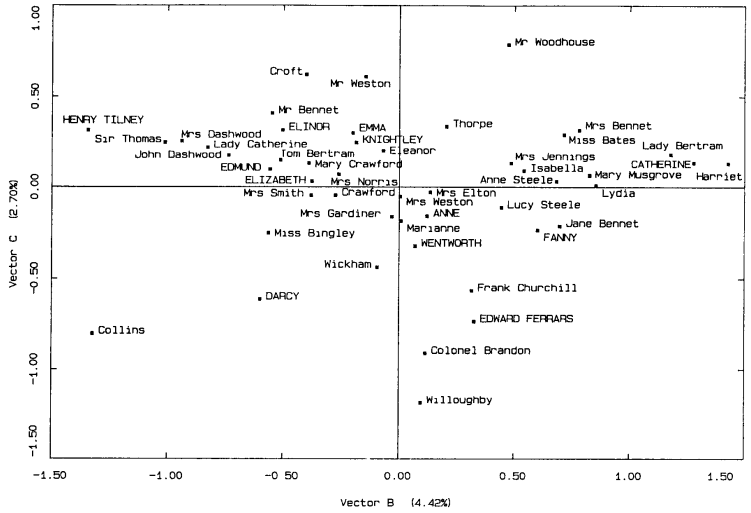
PCA

Hugh Craig









Graph 1. Jane Austen's major characters.

John Burrows

Henry James: The “Early” Style

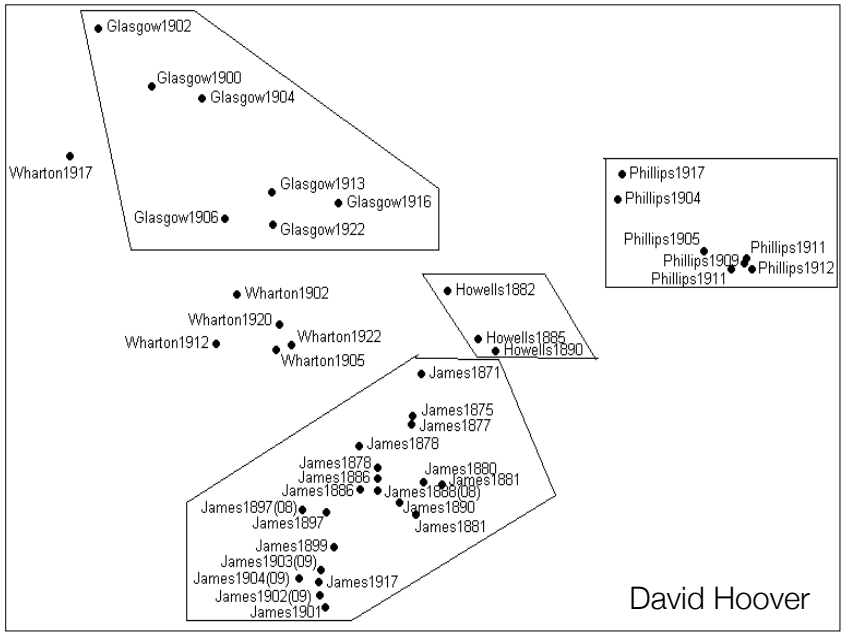
Newman looked at her a moment; he saw that she was pretty, but he was not in the least dazzled. He remembered poor M. Nioche's solicitude for her ‘innocence,’ and he laughed out again as his eyes met hers. Her face was the oddest mixture of youth and maturity, and beneath her candid brow her searching little smile seemed to contain a world of ambiguous intentions. She was pretty enough, certainly, to make her father nervous; but, as regards her innocence, Newman felt ready on the spot to affirm that she had never parted with it. She had simply never had any; she had been looking at the world since she was ten years old, and he would have been a wise man who could tell her any secrets.

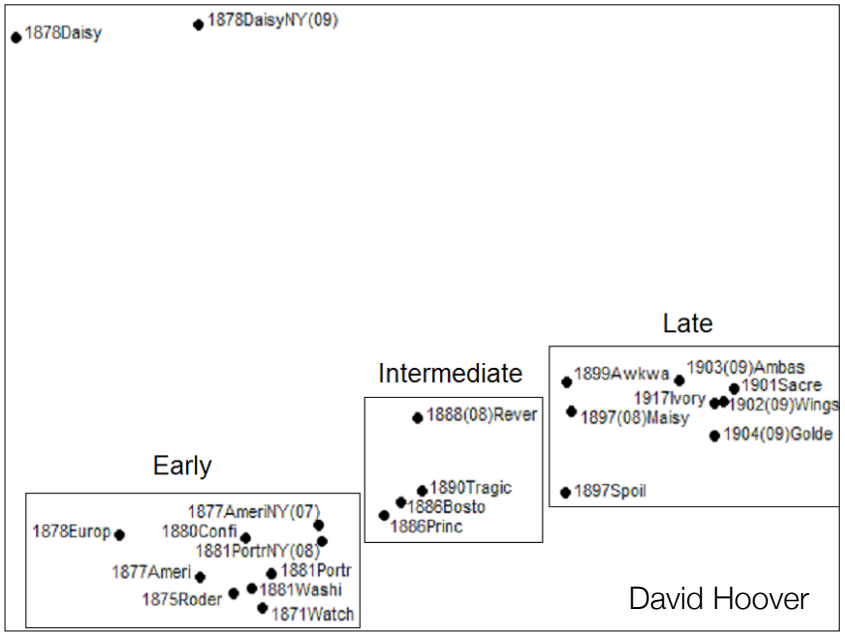
The American (1877 [1879 edition])

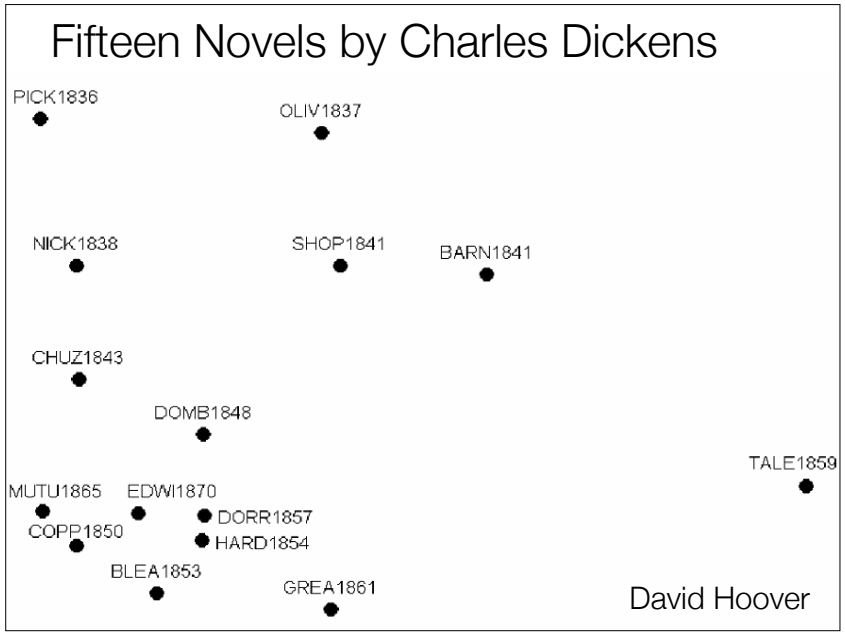
Henry James: The “Early” Style

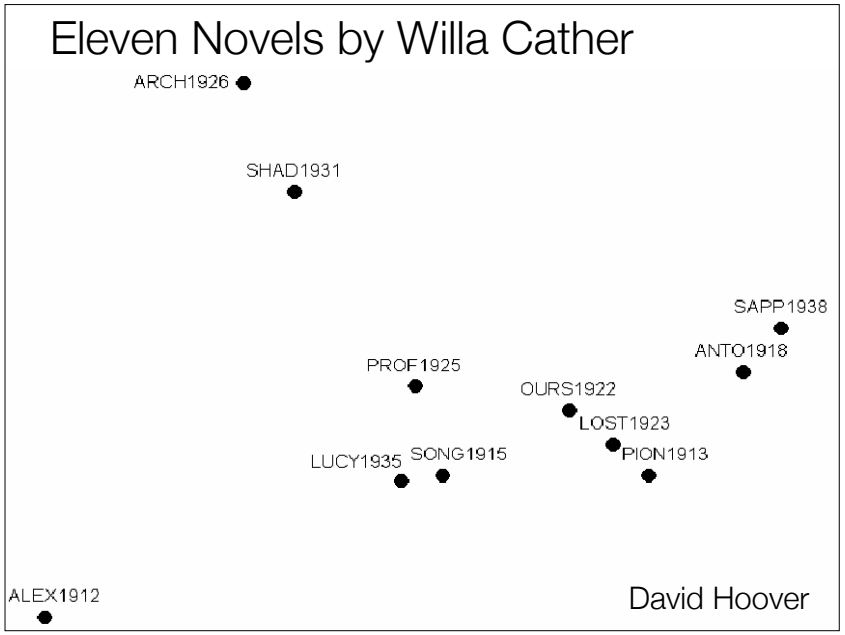
That brought back to Maisie--it was a roundabout way--the beauty and antiquity of her connexion with the flower of the Overmores as well as that lady's own grace and charm, her peculiar prettiness and cleverness and even her peculiar tribulations. A hundred things hummed at the back of her head, but two of these were simple enough. Mrs. Beale was by the way, after all, just her stepmother and her relative. She was just--and partly for that very reason--Sir Claude's greatest intimate (‘lady-intimate’ was Maisie's term) so that what together they were on Mrs. Wix's prescription to give up and break short off with was for one of them his particular favourite and for the other her father's wife.

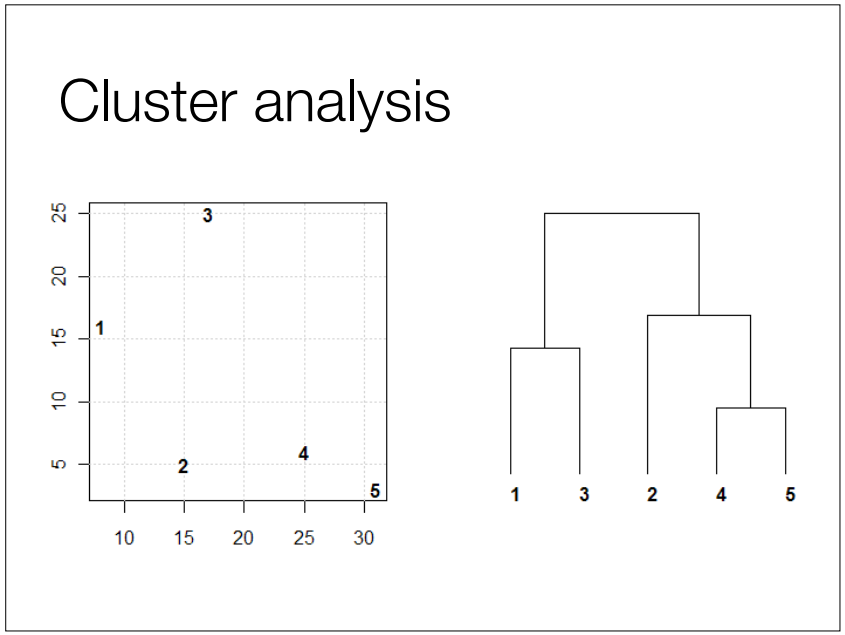
What Maisie Knew (1897: NYE,1908)

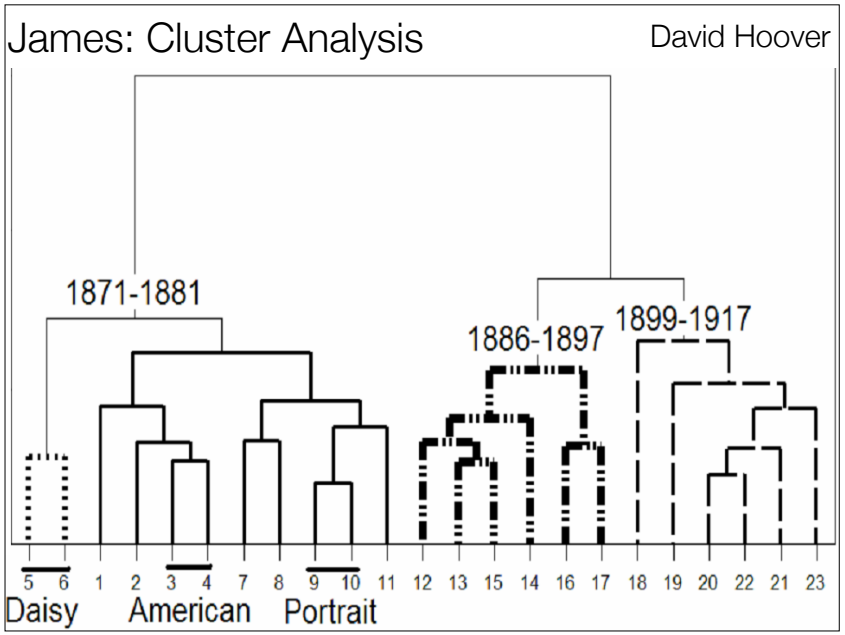


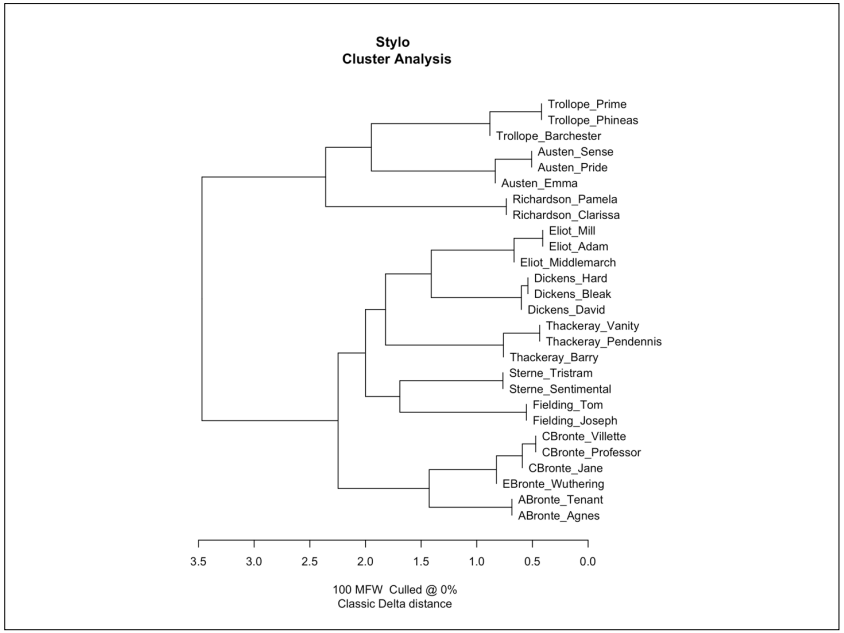


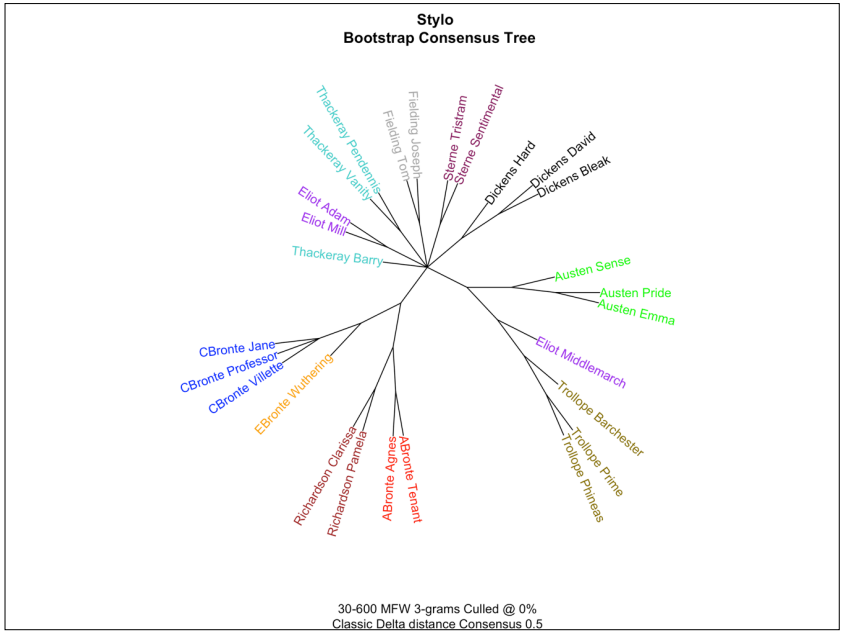

Fifteen Novels by Charles Dickens
PICK1836
OLIV1837
NICK1838
SHOP1841
BARN1841
CHUZ1843
DOMB1848
TALE1859
MUTU1865
EDWI1870
DORR1857
COPP1850
HARD1854
BLEA1853
GREA1861
David Hoover


Eleven Novels by Willa Cather
ARCH1926
SHAD1931
SAPP1938
ANTO1918
PROF1925
OURS1922
LOST1923
LUCY1935
SONG1915
PION1913
ALEX1912
David Hoover


Cluster analysis
25
20
15
10
5
3
1
2
4
5
10
15
20
25
30
1
3
2
4
5


James: Cluster Analysis
David Hoover
1871-1881
1886-1897
1899-1917
5 6 1 2 3 4 7 8 9 10 11 12 13 15 14 16 17 18 19 20 22 21 23
Daisy
American
Portrait





Bibliography

Highly recommended:

Hoover, D. L. (2007). Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41: 174-203.

Eder, M., and J. Rybicki (2011). *Computational Stylistics.* https://sites.google.com/site/computationalstylistics/. (Program used in the workshop.)

Eder, M., and J. Rybicki (2012). Do birds of a feather really flock together, or how to choose test samples for authorship attribution. *Literary and Linguistic Computing* 27.

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing* 22: 27-47.

Classic *Federalist Papers* work:

Mosteller, F., and D. L. Wallace (1964). *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. New York: Springer 1984. CSLI Publications published a reprint of the second edition in 2007, titled “Inference and Disputed Authorship”, with a new introduction by John Nerbonne.

Extended Bibliography

Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burrows, J. (2002). ‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17: 267-87.

Craig, H. (1999). Authorial attribution and computational stylistics: if you tell authors apart, have you learned anything about them? *Literary and Linguistic Computing* 14(1): 103-113.

Craig, H., and A. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010: Conference Abstracts*. King’s College London, pp. 132-135.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing* 22: 251-70.

Hoover, D. L. (2004). Testing Burrows’s Delta. *Literary and Linguistic Computing* 19(4): 453-475.

Jockers, M., D.M. Witten and C.S. Criddle (2007). Reassessing authorship of the *Book of Mormon* using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing* 23(4): 465-491.

Juola, P. (2008). *Authorship Attribution*. Foundations and Trends in Information Retrieval 1: 233-334.

Merriam, Th. (1989). An Experiment with the *Federalist Papers*. *Computers and the Humanities* 23(3): 251-254.

Rudman, J. (1998). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* 31: 351-365.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance* 16(2): 26-32.

Rybicki, J., and M. Eder (2011). Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing* 26(3): 315-321.