

PCA Online

The Shakespeare Computational Stylistics Facility

The CSF is an initiative of the Australian e-Humanities Network. It was created by the Centre for Literary and Linguistic Computing at the University of Newcastle, NSW, Australia. The project was funded by the Australian Research Council through a Learned Academies Special Projects grant. Design and programming is by Russell Whipp.

∞ CONTENTS ∞

• Introduction	2
• Walk-throughs	3
▶ Walk-through A	3
▶ Walk-through B	4
▶ Walk-through C	5
▶ Walk-through D	6
• More about PCA	7
• Play groupings	8
• Play names	8
• Character names	9
• Character attributes	9
• Contractions	9
• Limitations	10
• References	11
▶ (1) PCA as a technique	11
▶ (2) PCA in stylistic analysis	12

Introduction

The Computational Stylistics Facility presents a set of Shakespeare play texts with a ready-made apparatus for computational-stylistics exploration. Within its parameters, users can define any number of variations on what is analysed and how. The system has been designed for use by those with no experience in computational stylistics, and is set out so as to work intuitively as far as possible.

The texts can be analysed as whole plays, as blocks, i.e. sequential segments of plays, or as character parts. Word-variables for the analyses can be typed or pasted in, or the system can calculate for you the 20, 50, 100 most common words of the whole set or the sub-set of texts you are using. The best way to start is with a simple walk-through. Four are offered below. You may like to start by printing one of them out and use the printed sheet to guide you through a test run.

Walk-throughs

There are four walk-throughs (A,B,C, and D). There is also a poster presentation of the CSF you may like to look through. This is available as a .pdf download from:

<http://www.newcastle.edu.au/centre/cllc/pcaonline/index.html>

.....

Walk-through A

(Note that when navigating back to re-run any analysis, you need to “Remove all” of your play or character selections, and “Clear” your word selections in order to start afresh.)

Aim: to explore broad differences between the three genres

1. Click on “Start CSF”.
2. From opening “Text Segmentation Method” screen, select “By Play”. Click “Select Plays”.
3. From “Play Selection” screen, choose “Select All”. Click “Select Characters”.
4. From “Character Selection” screen, choose “Select All”. Click “Select Words”.
5. On the “Word Variable Selection” screen, go to the “Auto Populate Word List” area, select “Entire Corpus” and for “Word List Size” 100. Click on “Populate” and once the list is populated click on “Display Results”.
6. On the “Word Frequencies by Play” screen click on “PCA Plot”.
7. On the “Principal Component Plot by Play” screen, toggle symbol labels off by clicking on the graph, to see the genre pattern, and on again to identify individual plays. Click on “Component Scores”.
8. On the Principal Component Scores by Play” screen click on “PCA Word Plot”.
9. Note the words at the extremes of PC1, i.e. *our* and *their* (low values, to the left) and *you* and *I* (high values, to the right). Click on “Component Scores”.
10. On the “Principal Component Scores by Word” screen, click on the “First Component” to sort the words in order of First Component score and get the full picture of the heavily weighted words on this component. Scroll down to see the words at the low-scoring end. Words with high scores are evidently associated (among other things) with speakers speaking as individuals, those with low scores with speakers speaking for groups.
11. Click on the “Back” button until the “PCA plot” screen reappears and compare the distribution of genres and plays across the First Component with the words appearing at the two extremes of the Component.

Walk-through B

(Note that when navigating back to re-run any analysis, you need to “Remove all” of your play or character selections, and “Clear” your word selections in order to start afresh.)

Aim: to explore differences between the three Falstaffs of *Henry IV Part 1*, *Henry IV Part 2* and *The Merry Wives of Windsor*

1. Click on “Start CSF”.
3. From opening screen, select “By Character”. Click on “Select Plays”.
4. From the “Play Selection” screen, choose “Select by Criteria”. Tick “Comedy” and “History”. Click “Add” and “Done”. Click on “Select Characters”.
5. From the “Character Selection” screen, choose “Select by Criteria”, and then “Words Spoken” “More than” and “3000”. Click on “Add” and “Done” and then “Select Words”.
6. On the “Select Word Variables” screen, go to the “Auto Populate Word List” area, select “Entire Corpus” and for “Word List Size” 100. Click on “Populate” and once the list is populated click on “Display Results”.
7. On the “Word Frequencies by Character” screen, click on “PCA Plot”.
8. On the “Principal Component Analysis by Characters” screen, click on the graph to toggle labels on and off. Note that while two of the Falstaffs (from the colour of the symbol, one from a comedy and one from a history) occupy much the same area of the graph, a third Falstaff is separated from them on the Second Component. Click on “Component Scores”.
9. On the “Principal Component Scores by Character” screen, click on “Second Component”. This will sort the entries with the highest Second Component score at the top. Scroll down to find the Falstaff of *Henry IV Part 2*, then some way down the Falstaffs of *Henry IV Part 1* and *Merry Wives*. Click on “PCA Word Plot”.
10. On the “Principal Component Analysis by Word” screen, note the words to the top and bottom, which are those contributing most to the Second Component. Characters to the upper end of the Second Component on the whole use more of the words appearing to the upper end, and rather fewer of those appearing at the lower end. The extremes are *which* (at the top) and *come* (at the bottom). The Falstaff of *Henry IV Part 2* is closer to the *which, that, but, by, your* end than the other two, and further from the *come, say, go, shall, well* end than they are.
11. To look at some instances in context, click the “Back” button until you get to the “Word Frequencies by Character”. Go to “Select Play” and go down to *Henry IV Part 2*. Scroll down to *which* and double-click on it to read some instances (navigating by clicking the “Next” button). One hypothesis for the distinctiveness of the Falstaff of *Henry IV Part 2* that this suggests is his mock speechifying, his teasing adoption of the forms of academic disputation. This would of course need further exploration through instances of other words, including those on which the other two Falstaffs tend to score more highly than the *Henry IV Part 2* one.

Walk-through C

(Note that when navigating back to re-run any analysis, you need to “Remove all” of your play or character selections, and “Clear” your word selections in order to start afresh.)

Aim: to explore larger characters’ use of the various forms of the second person pronoun

1. Select and copy the following list using the Edit menu of your browser:

you
your
yours
thou
thee
thy
thine
ye

2. Click on “Start CSF”.
3. From opening screen, select “By Character”. Click on “Select Plays”.
4. From the “Play Selection” screen, choose “Select All”. Click on “Select Characters”.
5. From the “Character Selection” screen, choose “Select by Criteria”, and then “Words Spoken” “More than” and “3000”. Click on “Add” and “Done” and then “Select Words”.
6. On “Word Variable Selection” screen, click on the “Word List” space and paste in the copied list of seven pronouns. (This may be by right-clicking on your mouse. You can also simply type them in here.) Click on “Display Results”.
7. On the “Word Frequencies by Character” screen, click on “PCA Plot”.
8. On the “Principal Component Analysis by Characters” screen, click on the graph to toggle labels on and off. Note the way the characters arrange themselves along the First Component, from Prospero to Menenius. Click on “Component Scores”.
9. On the “Principal Component Scores by Character” screen, click on “PCA Word Plot”.
10. On the “Principal Component Analysis by Word” screen, note that the First Component is an opposition between *your* and *you* and the “thou” forms. Go “Back” (twice) to the “Principal Component Analysis Plot by Character” screen. Toggle the characters’ names off by clicking on the graph area and try colouring the symbols by “Genre” and then by “Gender”.
11. Go “Back” to the “Word Frequencies by Character” screen. Go to “Select Play” and select *The Tempest* from the list. Click on “Show Frequencies”. Note that *thou* is just over 2% of Prospero’s dialogue, and *you* just over 0.9%.
12. On the same screen go to “Select Play” again and choose *Coriolanus* from the list. Note that *thou* is just over 0.1% of Menenius’ dialogue, and *you* 3.7%.

Walk-through D

(Note that when navigating back to re-run any analysis, you need to “Remove all” of your play or character selections, and “Clear” your word selections in order to start afresh.)

Aim: to explore the consistency in the contrast between some representative tragedies and comedies of Shakespeare’s middle period.

1. Click on “Start CSF”.
2. From opening screen, select “By Block” and enter a “Block Size” of 4000 words. Click on “Select Plays”.
3. From the “Play Selection” screen, select and then “Add” *All’s Well that Ends Well*, *As You Like It*, *Hamlet*, *King Lear*, *Macbeth* and *Merchant of Venice*. Click on “Select Characters”.
4. From “Character Selection” screen, choose “Select All”. Click on “Select Words”.
5. On “Word Variables Selection” screen, go to the “Auto Populate Word List” area, select “Entire Corpus” and for “Word List Size” 30. Click on “Populate” and once the list is populated click on “Display Results”.
6. On the “Word Frequencies by Character” screen, click on “PCA Plot”.
7. On the “Principal Component Analysis by Blocks” screen, click on the graph to toggle labels on and off. Go to the “Colour by” menu and select “Genre”. Note that there is a broad separation along the First Component: tragedy blocks to the left, comedy blocks to the right. Go back to the “Colour by” menu and click on “Play” to identify the plays. Note that of the tragedy blocks it is the *King Lear* ones (in particular blocks 6, 7 and 10 – the second, third and last sections of the play) that stray towards the low-value end of the Component which is mainly populated by blocks from the comedies.

More about PCA

The use of PCA in computational stylistics was pioneered by John Burrows. Burrows' method is to use frequencies of the most common thirty, fifty, or 100 words as variables and count them in texts, or parts of texts. Though these words are mostly "function words", i.e. have a role in syntax rather than lexis, their rate of occurrence turns out to be rich in information about style. Studies with the Burrows method have shown that patterning in the distribution of very common words corresponds with authorship, period, genre, gender, character, and almost any other category of interest to literary scholars.

PCA is a multivariate, multifactor technique. Through it the investigator can identify new latent composite variables which account for a great deal of the variation in the original variables. In this sense it is a data reduction method. It involves a mathematical procedure that transforms a number of (possibly) correlated variables into a smaller number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The aim is to reduce the dimensionality of the data set and to identify new meaningful underlying variables.

Once the principal components have been isolated one can inspect the loadings for each one to see which variables have been most important in forming them, i.e. are most heavily weighted. One can also multiply weightings and the scores for individual observations to see where the observations fall on each component.

A general rule is that for reliability, twenty-five or more observations should be used. In general, the potential divergence between the true coordinates of a given variable or observation and the actual one calculated reduces as the sample size increases.

Play Groupings

Dates are ranges, taken from *The Norton Shakespeare: Based on the Oxford Edition*, ed. Stephen Greenblatt, Walter Cohen, Jean E. Howard and Katharine Eisaman Maus. (New York: Norton, 1997). Plays are also classified by genre (“tragedy”, “comedy”, and “history”) and authorship (“single author”, “collaborative”, and “modified” -- the latter for plays where there is evidence of some lesser participation by a second author after the original act of composition).

Play Names

These are abbreviated in the graphs according to the MLA list:

Ado	<i>Much Ado about Nothing</i>
Ant	<i>Antony and Cleopatra</i>
AWW	<i>All's Well that Ends Well</i>
AYL	<i>As You Like It</i>
Cor.	<i>Coriolanus</i>
Cym.	<i>Cymbeline</i>
Err.	<i>Comedy of Errors</i>
Ham.	<i>Hamlet</i>
1H4	<i>Henry IV, Part 1</i>
2H4	<i>Henry IV, Part 2</i>
H5	<i>Henry V</i>
1H6	<i>Henry VI, Part 1</i>
2H6	<i>Henry VI, Part 2</i>
3H6	<i>Henry VI, Part 3</i>
H8	<i>Henry VIII</i>
JC	<i>Julius Caesar</i>
Jn.	<i>King John</i>
LLL	<i>Love's Labours Lost</i>
Lr.	<i>King Lear</i>
Mac.	<i>Macbeth</i>
MM	<i>Measure for Measure</i>
MND	<i>A Midsummer Night's Dream</i>
MV	<i>Merchant of Venice</i>
Oth.	<i>Othello</i>
Per.	<i>Pericles</i>
R2	<i>Richard II</i>
R3	<i>Richard III</i>
Rom.	<i>Romeo and Juliet</i>
Shr.	<i>The Taming of the Shrew</i>
TGV	<i>Two Gentlemen of Verona</i>
Tim.	<i>Timon of Athens</i>
Tit.	<i>Titus Andronicus</i>
Tmp.	<i>The Tempest</i>
TN	<i>Twelfth Night</i>
Tro.	<i>Troilus and Cressida</i>
Wiv.	<i>Merry Wives of Windsor</i>
WT	<i>The Winter's Tale</i>

Character Names

These follow the Moby Shakespeare, with rare changes to enable characters to be distinguished across plays. Where characters change names in the course of a play (Prince Henry to Henry IV, Marcius to Coriolanus), these are treated as separate characters.

Character attributes

Characters have been tagged with various attributes to allow users to select them in categories. Some of the categories are also appear on the symbol colouring menu so that they can be identified on the PCA text graphs.

Two of the classification schemes, gender and social class, are broadly comprehensive. All characters are assigned to either the male or female gender, even where this is a matter of some discussion, e.g. Ariel in *A Midsummer Night's Dream*. Social class is a little more complicated. Most characters are labelled either “ordinary”, “bourgeois”, “gentle”, “noble”, or “royal”, though there are also categories outside this hierarchy such as “spirit”, “allegorical figure”, and “god”. The social classifications are often quite difficult to do and must be used with caution. Characters in the Roman plays in particular do not always fit into a scheme based on Early Modern English society. Users can of course add or remove characters from an analysis individually where they find the classification unsatisfactory. The classifications will form a starting point, and, true to the principles of exploratory data analysis, users can modify their selections as they please in succeeding analyses.

There are also some other kinds of attributes which are by their natures more sporadic. Most characters are labelled according to occupation, ranging from “servant”, with many examples, to “bellows mender”, with only one. Some characters are also labeled according to dramatic function, as “protagonist”, “antagonist”, or “villain.”

Contractions

There are a great many contracted forms like “I'll” and “he's” in Shakespeare dialogue. In the CSF all these have been expanded. Each instance of “I'll”, for instance, is counted as one instance of *I* and one of *will*. Depending on the sense, each instance of “he's” is counted either as one instance of *he* and one of *is*, or as one instance of *he* and one of *has*. The base text used for instances in context, however, retains the original contractions.

Limitations

- The CSF uses a single base text, the Moby Shakespeare. It was adopted as the only public-domain complete modern-spelling Shakespeare currently available. The Moby text derives from a transcription of the Cambridge “Globe” edition of 1863-6, edited by William George Clark, John Glover, and William Aldis Wright, and will vary from other modern-spelling editions and from unedited early printed versions like the Folio.
- To print any of the results provided by the CSF users will have to capture a screen shot which can then be saved or printed.
- The CSF offers word frequency data from the plays of Shakespeare, with segmentation by play, character, and block, and PCAs using this data. All sorts of extensions and variations in procedures will occur to users. (The most obvious is to include comparable plays by other writers of the period.) The CSF is a primer in computational stylistics and the hope is that some users, once they see the possibilities of the methods, may go on either to create their own systems, or to use other existing software (word-counting, spreadsheet, and statistics programs) to run analyses to explore in different ways.
- The Moby Shakespeare does not include *The Two Noble Kinsmen*, so the play does not appear in the CSF.

References

(1) PCA as a technique

- Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.
- Harman, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1967.
- Kim, J. O., & Mueller, C. W. *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage Publications, 1978.
- Kim, J. O., & Mueller, C. W. *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage Publications, 1978.
- Lawley, D. N., & Maxwell, A. E. *Factor analysis as a statistical method* (2nd. ed.). London: Butterworth & Company, 1971.
- Lindeman, R. H., Merenda, P. F., & Gold, R. *Introduction to bivariate and multivariate analysis*. New York: Scott, Foresman, & Co., 1980.
- Morrison, D. F. *Multivariate statistical methods*. (3rd Ed.). New York: McGraw-Hill, 1990.
- Mulaik, S. A. *The foundations of factor analysis*. New York: McGraw Hill, 1972.
- Stevens, J. *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum, 1986.

(2) PCA in stylistic analysis

- Binongo, J.N.G. "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution." *Chance: A Magazine of the American Statistical Association* 16 (2003): 9-17.
- Binongo, J.N.G., and M. W. A. Smith. "The Application of Principal Component Analysis to Stylometry." *Literary and Linguistic Computing* 14 (1999): 445-65.
- Binongo, J.N.G., and M. W. A. Smith. "A Bridge between Statistics and Literature: The Graphs of Oscar Wilde's Literary Genres." *Journal of Applied Statistics* 26 (1999): 781-87.
- Burrows, J. F. *Computation into Criticism. An Essay in Jane Austen and an Experiment in Method.* Oxford: Clarendon Press, 1987.
- Burrows, J. F. "'A Vision' as a Revision." *Eighteenth Century Studies*, 22 (1989), 551-65.
- Burrows, John F. "Computers and the Study of Literature." *Computers and Written Texts.* Ed. Christopher S. Butler. Oxford: Blackwell, 1992. 167-204.
- Burrows, J. F. and A.J. Hassall. "Anna Boleyn and the Authenticity of Fielding's Feminine Narratives." *Eighteenth Century Studies*, 21 (1988), 427-453.
- Burrows, John. "A Computational Approach to the Rochester Canon." *The Complete Works of John Wilmot, Earl of Rochester.* Ed. Harold Love. Oxford: Clarendon, 1999. 681-95.
- Burrows, John, and D H Craig. "Lyrical Drama and the 'Turbid Mountebanks': Styles of Dialogue in Romantic and Restoration Tragedy." *Computers and the Humanities* 28 (1994): 63-86.
- Burrows, John, and Hugh Craig. "Lucy Hutchinson and the Authorship of Two Seventeenth Century Poems: A Computational Approach." *The Seventeenth Century* 16 (2001): 259-82.
- Craig, Hugh. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" *Literary and Linguistic Computing* 14 (1999), 103-13.
- Craig, Hugh. "Contrast and Change in the Idiolects of Ben Jonson Characters." *Computers and the Humanities* 33:221-40, 1999.
- Craig, Hugh. "The Weight of Numbers: Common Words and Jonson's Dramatic Style." *Ben Jonson Journal* 6 (1999): 243-59.
- Forsyth, R S, D I Holmes, and Emily Tse. "Cicero, Sigonio, and Burrows: Investigating the Authenticity of the *Consolatio*." *Literary and Linguistic Computing* 14 (1999): 375-400.
- Holmes, D. I. "A Stylometric Analysis of Mormon Scripture and Related Texts." *Journal of the Royal Statistical Society* A155 (1992), 91-120.
- Holmes, D. I. 1994. "Authorship Attribution." *Computers and the Humanities* 28(2) (1994), 87-106.
- Holmes, David I, Lesley J Gordon, and Christine Wilson. "A Widow and Her Soldier: Stylometry and the American Civil War." *Literary and Linguistic Computing* 16 (2001): 403-20.
- Holmes, D. I. and R. S. Forsyth. 1995. "The Federalist Revisited: New Directions in Authorship Attribution." *Literary and Linguistic Computing* 10(2), 111-127.
- McKenna, Wayne, and Alexis Antonia. "Intertextuality and Joyce's "Oxen of the Sun" Episode in *Ulysses*: The Relation between Literary and Computational Evidence." *Revue Informatique et Statistique dans les Sciences humaines* 30 (1994): 75-90.
- McKenna, W, and A Antonia. "'a Few Simple Words' of Interior Monologue in *Ulysses*: Reconfiguring the Evidence." *Literary and Linguistic Computing* 11 (1996): 55-66.
- Tabata, Tomoji. "The Language of Dickens and Its Computer-Based Evidence: A Step Towards a Chronological Study." *Kumamoto Studies in English Language and Literature* 36 (1993): 116-34.
- Tweedie, F. J., D. I. Holmes, and Thomas N. Corns. 1998. "The Provenance of *De Doctrina Christiana*, Attributed to John Milton: A Statistical Investigation." *Literary and Linguistic Computing* 13(2), 77-87.