Computer Studies of Shakespearean Authorship

Michael P. Oakes
University of Wolverhampton

Canon, Dubitanda, Apocrypha

- Elliott and Valenza (1996) classified the plays which are or have been associated with William Shakespeare according to the strength and reliability of the association.
- Canon: 35 plays identified by Donald Foster as having certainly been written by Shakespeare, such as “Hamlet” and “Romeo and Juliet”. Most of these are in the “First Folio”, and have been confirmed by Elliott and Valenza’s own tests.
- Dubitanda: plays like “Titus Andronicus” or “Two Noble Kinsmen” where the evidence for Shakespearean authorship is weaker than for the canon, or plays thought to have been collaboratively written.
- Apocrypha: Not generally accepted as being by Shakespeare, but some tenuous association e.g. “Arden of Faversham” was very popular in Shakespeare’s day, but we have no idea who wrote it. “King Leir” which may have inspired “King Lear”.

Shakespeare, Wilkins and “Pericles”

- There is a marked difference in style between the first two and the last three acts.
- The later, better written, part is generally attributed to Shakespeare.
- “Pericles” has been linked with Wilkins since he wrote a novel purporting to be a report of a play on the same topic in 1608; also Wilkins and Shakespeare once lodged together.
- Wilkins is known to have written “The Miseries of Enforced Marriage”, and part of “The Travels of Three English Brothers” (other parts by Day and Rowley).
- Smith (1987) used a variant of the chi-squared test to find which words best discriminate between the works of one dramatist and another.
- Pearson’s residuals for each cell (next slide) = $(O - E) / \sqrt{E}$
- The residuals are a form of z-score, so if $PR > 1.96$, $p < 0.05$.

Use of Pearson residuals to find words more typical of Shakespeare's "Cymbeline" or Wilkins' "Miseries"

	Cymbeline	Miseries	Row total
A	455 (518.0) [-2.77]	518 (455.0) [2.95]	973
For	238 (275.3) [-2.25]	279 (241.7) [2.40]	517
It	409 (351.4) [3.07]	251 (308.6) [-3.28]	660
Man	75 (113.4) [-3.61]	138 (99.6) [3.85]	213
More	103 (77.7) [2.87]	43 (68.3) [-3.06]	146
Now	61 (82.5) [-2.37]	94 (72.5) [2.53]	155
Sir	79 (113.4) [-3.23]	134 (99.6) [3.45]	213
Than	88 (67.6) [2.48]	39 (59.4) [-2.65]	127
The	897 (748.0) [5.45]	508 (657.0) [-5.81]	1405
To	651 (708.6) [-2.16]	680 (622.4) [2.31]	1331
Column Total:	3056	2684	5740

Plays in rank order of chi-squared closeness of fit to "Pericles" III-V when samples of 13 function words are compared (Smith 1987).

Rank	Title	Author	Chi-squared
1	Cymbeline	Shakespeare	11.172
2	The Family of Love 1	(Dekker et al.)	11.624
3	The Tempest	Shakespeare	12.791
4	The Family of Love 3	(Dekker et al.)	13.324
5	The Maid's Tragedy	Beaumont-Fletcher	14.301
6	Othello	Shakespeare	16.013
7	Travels of the 3 English Brothers (non-W)	Day-Rowley	16.610
8	The Traitor	Shirley	17.080
9	Blurt Master Constable	(Dekker)	18.612
10	Royal King and the Loyal Subject	Heywood	19.176

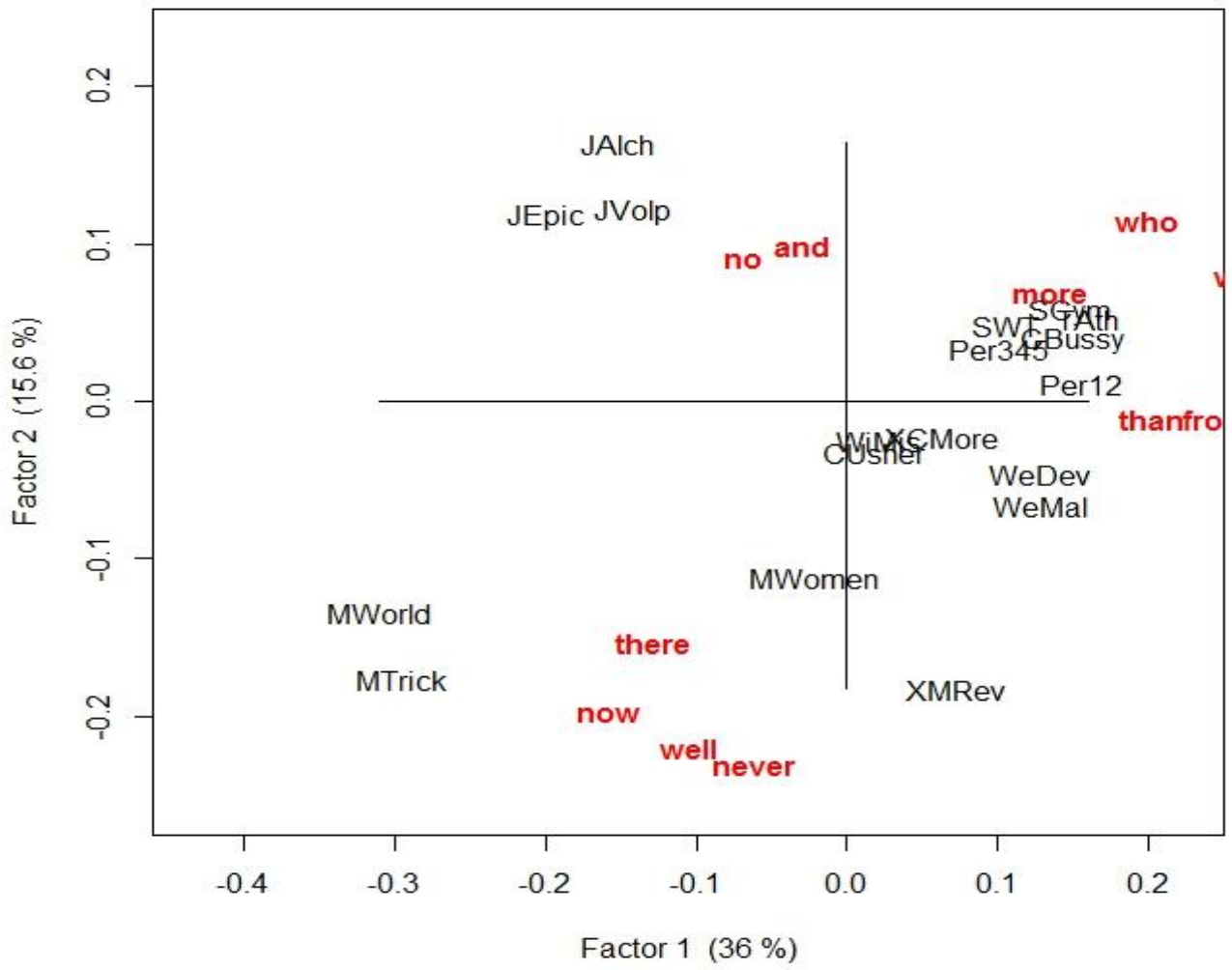
Plays in rank order of chi-squared closeness of fit to "Pericles" I-II when samples of 13 function words are compared

Rank	Title	Author	Chi-squared
1	Miseries of Enforced Marriage 2	Wilkins	10.322
2	Philaster	Beaumont-Fletcher	14.069
3	Travels of the 3 English Brothers (W1)	Wilkins	15.433
4	Family of Love 1	(Dekker et al.)	17.644
5	Travels of the 3 English Brothers (W2)	Wilkins	18.685
6	Woman Killed with Kindness	Heywood	19.532
7	Royal King and the Loyal Subject	Heywood	20.240
8	Woman Never Vexed IV-V	(Wilkins)	21.448
9	The Traitor	Shirley	22.138
10	The Bondman	Massinger	22.154

Comparison between Acts I-II and Acts III-V of Pericles for Jackson's 13 function words

- R commands:
- early = c(146, 235, 61, 49, 90, 35, 99, 106, 135, 104, 240, 257, 49)
- late = c(182, 251, 87, 45, 94, 27, 114, 105, 166, 110, 339, 234, 85)
- table = cbind(early, late)
- chisq.test(table)
- This gives a chi-squared value of 26.798, which for 12 degrees of freedom, gives p = 0.008.
- So the two "halves" are significantly different.

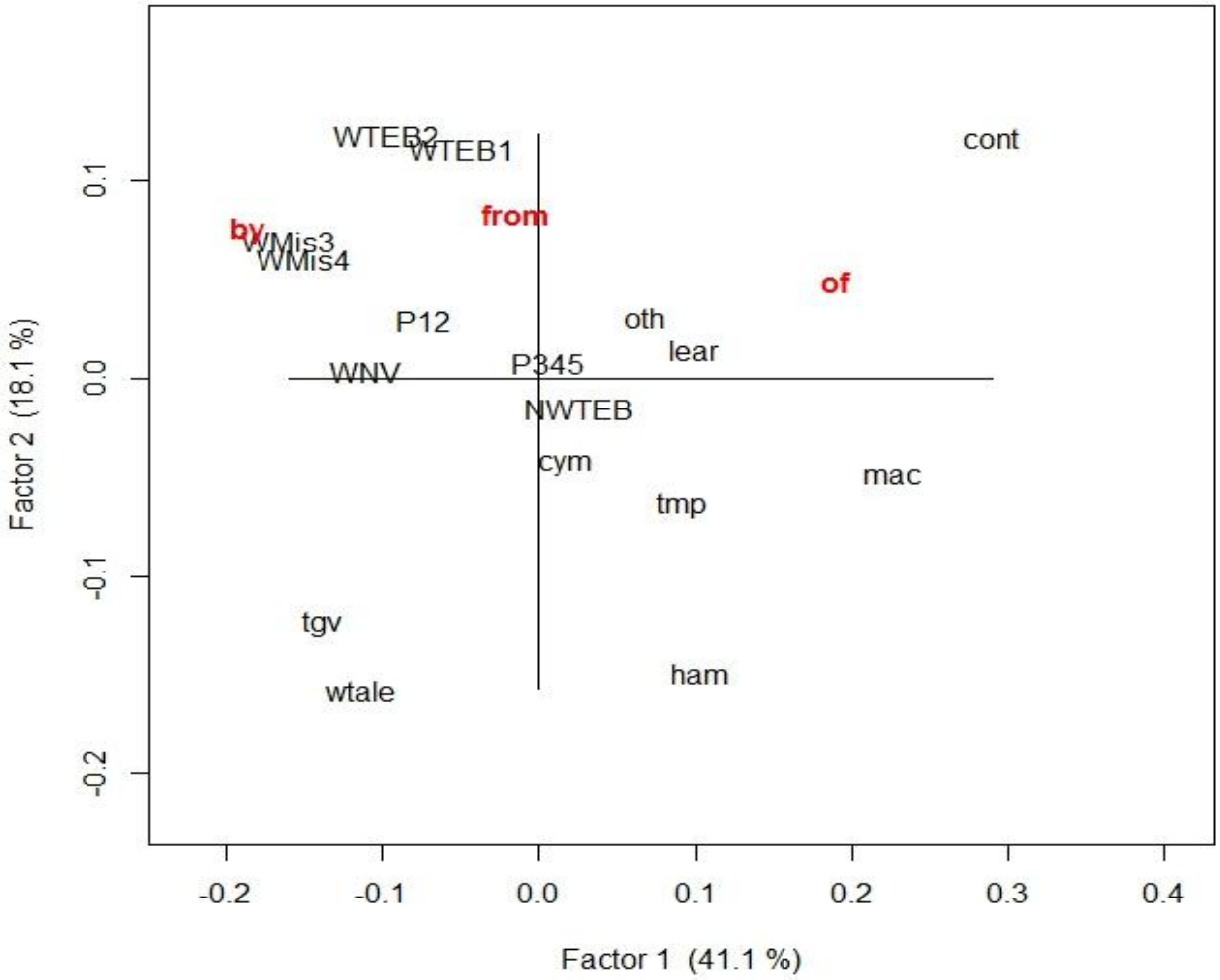
Correspondence Analysis for Pericles and related texts (Smith's data: 46 common words).



Keys to previous slide

- Shakespeare: Cymbeline, A Winter's Tale
- Wilkins: Miseries of Enforced Marriage
- Webster: Devil's Law Case; Duchess of Malfi
- Chapman: Bussy D'Ambois; Gentleman Usher; Sir Thomas More (?)
- Middleton: A Trick to Catch the Old One; A Mad World, My Master; Women Beware Women; Revenger's Tragedy (?)
- Jonson: Alchemist; Epicoene; Volpone
- Unknown: Timon of Athens, Pericles I-II, Pericles III-V.

Correspondence Analysis for plays associated with Pericles (data of Jackson 1979, 1991)



Keys to previous slide

- Wilkins: Miseries of Enforced Marriage (WMIS3, WMIS4); Travels of the Three English Brothers (WTEB1, WTEB2); A Woman Never Vexed (?)
- Day and Rowley: Travels of the Three English Brothers (NWTEB)
- Shakespeare: The Contention: Henry VI; Two Gentlemen of Verona, Hamlet, Othello, King Lear, Macbeth, Cymbeline, A Winter's Tale, Tempest.
- Seems to confirm that A Woman Never Vexed is in Wilkins' style.
- It may be that the division between S and W's contributions is not so clear cut, and that the two parts each contain fragments of each other's writing.

“King John”

- Stylometric studies have shown two writing styles in the Play “King John”, one certainly Shakespeare, the other possibly Marlowe.
- Thomas Merriam (1996, 2004) found the following to be good discriminators: and, I, is, it, not, of, their, with, you, BoB1, BoB2, Bob5, ‘ll, -ish, ne’re.
- z-scores for each these variables were found for various text samples (28 Shakespeare Plays, Marlowe’s “Tamburlaine” and “King John”).
- If $z > 1.96$, the samples were deemed outliers.
- On most of these tests, both “Tamburlaine” and Marlowe’s portion of “King John” were found to be outliers.
- Merriam’s second experiment used Principal Component Analysis (PCA).
- The input was a matrix where the rows were the plays and the columns were the frequencies of frequent words In King John.
- Best Discrimination was seen on the 2nd PC.

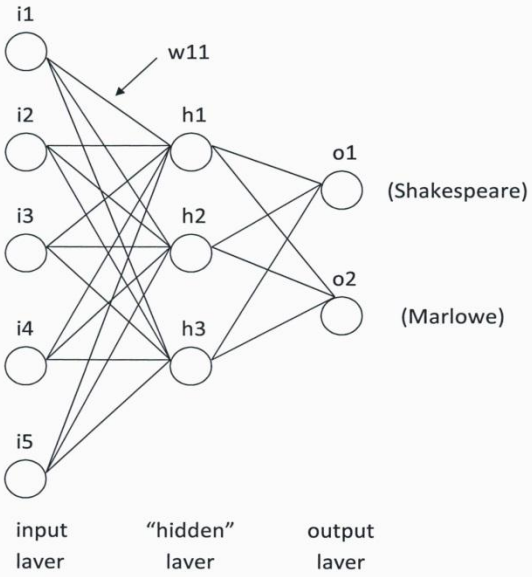
Parts of “King John” distinguished by Principal Components Analysis.

Play	PC2
1 Tamburlaine	-1.98
King John (non-Shakespeare)	-1.98
2 Tamburlaine	-1.91
Shakespeare core canon (28 plays)	Range -1.02 to 1.55
King John (Shakespeare)	3.46

“The Raigne of King Edward III”

- “Edward III” is the anonymous play most likely to have been written (or partly written) by Shakespeare.
- Alfred Hart in 1934 found that “Edward III” was similar in word frequencies to the rest of the Shakespeare canon.
- Using a variant of the chi-squared distance he used for “Pericles” Smith (1991) found that two samples of “King Edward III” (one considered more likely than the other to be Shakespeare’s part) were both more similar to two undisputed Shakespeare plays than to any of 8 other playwrights.
- Most striking is the abrupt change at the end of Act II when Edward leaves the Countess of Salisbury and returns to the war in France.

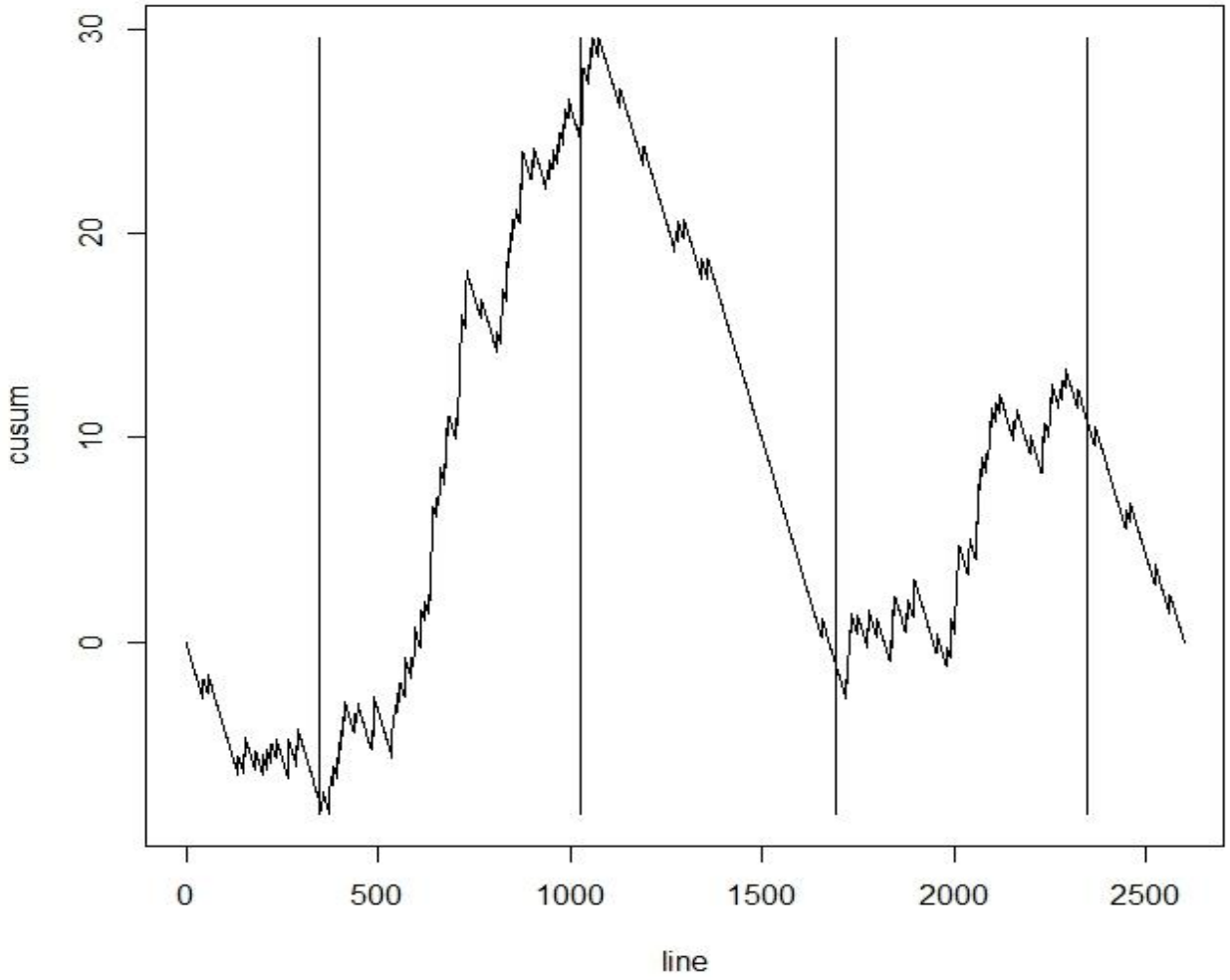
Neural Network to distinguish Shakespeare and Marlowe (Merriam and Matthews, 1994)



The controversial “Cusum” technique

- Originally derived from process control, but has not been successful in stylometry.
- Ideally the technique would show variations in the frequencies in different parts of a text, such that abrupt changes in the slope of the plot will occur where there is a change in authorship.
- Divide the total occurrences of linguistic feature x in the text and divide by the number of lines in the play.
- For each line, the difference between the number of x and the average number of x is found.
- The sum of these differences from the start of the text (the cumulative sum) to the line in question is recorded.
- The cumulative sum at each point is plotted on the y axis, and the line of the play is on the x axis.
- The cumulative sum starts and ends at 0.

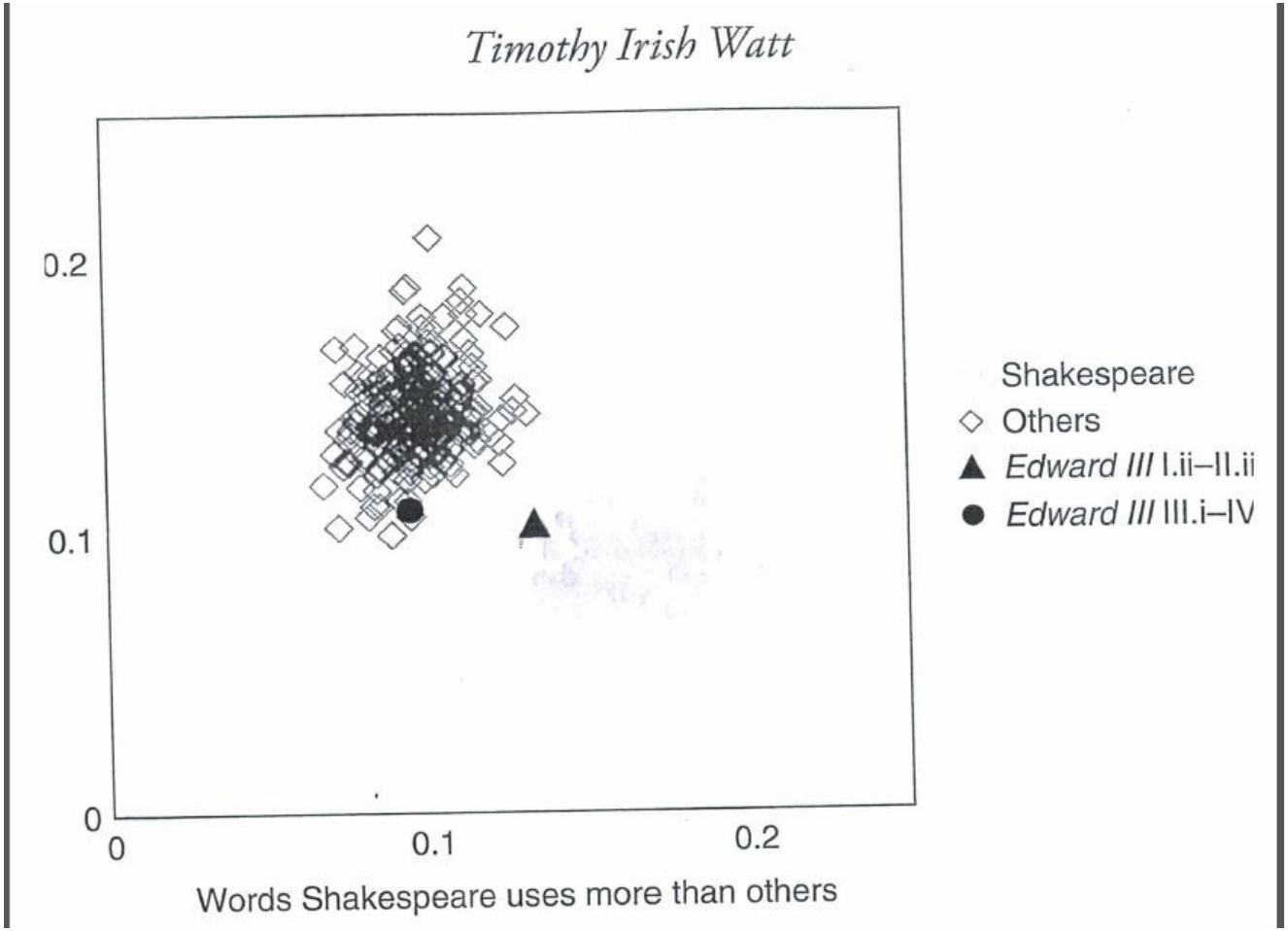
Cusum chart of feminine endings in Edward III (Merriam, 2000)



Burrows' Zeta

- Burrows' Delta is a form of z-score for high frequency words
- Burrows' Zeta compares the frequencies of content words. Watt (2009) found words more typical of Shakespeare than other playwrights such as "spoke", and words more typical of the other playwrights such as "hopes".
- X axis is the proportion of the text made up of Shakespeare markers, y the proportion made up of "others" markers.
- Countess scenes were placed in the Shakespeare cluster, although on the periphery, while French Campaign scenes were in the "other" cluster.

Burrows' Zeta (2)



Burrows Iota

- Here Watt (2009) considered the relatively rare words which occur in the “Countess” scenes, i.e. those which appear in the “Countess” scenes but do not appear in 60% of the other samples.
- Divide by the total number of words to produce a vector of relative frequencies.
- Produce analogous vectors for 16 different writers.
- Find correlation coefficient between vector for “Countess” scenes and each writer in turn.
- The experiment was repeated for the “French Campaign” scenes.

Correlation of authors' vocabulary lists and sections from Edward III

Author	Correlation with "Countess"	Correlation with "French Campaign"
Shakespeare	0.26	0
Marlowe	0.17	0.32
Lyly	0.09	0.08
Greene	0.08	0.13
Chapman	0.05	0.02
Peele	0.04	0.22
Kyd	0.04	0.19
Heywood	0.03	0.06
Fletcher	0.01	-0.02
Wilson	-0.04	0.07
Marston	-0.05	0.01
Haughton	-0.05	-0.04
Webster	-0.05	-0.03
Dekker	-0.06	-0.07
Middleton	-0.20	-0.20

Elliott and Valenza

- The “Shakespeare Clinic” at Claremont McKenna College in California.
- Were able to distinguish the poetry of Shakespeare and the Earl of Oxford (Edward de Vere), “long time favourite among those who believe that Shakespeare was not Shakespeare”.

Preferred and dispreferred words in two of the Bundle of Badges tests

Test	Badges	Flukes
Bundle of Badges 5	the is to you he is your we him as an	a Sir I now I'll 'tis all come her
Bundle of Badges 7	is	it's there's I'm here's she's that's what's

Data for samples of Poetry from Shakespeare and Oxford

Test	Shakespeare poems (mean)	Shakespeare (std. dev)	Oxford poems (mean)	Difference in std. dev.
Grade Level	11.79	1.08	7*	4.435
HCW/20K	112.09	34.75	32	2.305
Fem. Endings (%)	12.00	34.75	32	2.305
Open lines (%)	14.93	4.13	7	1.920
Enclitics / 1000 lines	54.07	17.24	13*	2.382
Proclitics / 1000 lines	329.29	50.74	115*	4.223
With (21ws)	13.43	7.89	5	1.068
No / (No + Not)	359.36	97.32	500	-1.445
BoB5	320.14	126.44	290	0.238
BoB7	754.93	136.18	1000*	-1.799
BoB8	-512.14	90.43	-301	-2.335
T-E slope test	-0.08	0.08	0.12	-0.875
T-E New Word	-7.33	12.23	-10.6	0.137

Rejections

- The Oxford texts received 6 rejections out of 14 tests. What would the probability of this if Shakespeare were the author?
- The R function `pbinom(x,n,p)` is the cumulative binomial distribution function, and gives the probability of receiving fewer than x rejections in a series of n tests, where the probability of rejection is p (approx 0.026 for Shakespeare texts).
- Thus `pbinom(5, 14, 0.026)` would give the probability of 5 rejections or fewer in 14 tests, so `1 - pbinom(5, 14, 0.026)` is the probability of 6 or more rejections = 7.749 e-07.
- Thus the writing style of Oxford's poetry differs significantly from that of Shakespeare, suggesting that they are unlikely to be one and the same person.
- Similar tests were done for plays. All 29 of their "core" Shakespeare plays had 2 or fewer rejections.
- None of the apocrypha has fewer than 7 rejections.
- Of the dubitanda, closest were "Henry VI part 2", "Pericles III-V", "Henry 5" (3 rejections, $p > 0.05$).

“Hand D” in “Sir Thomas More”

- Very little of Shakespeare’s own handwriting survives – just six signatures, all with different spellings.
- However, it is just possible that a handwritten section of the play “Sir Thomas More” may be in Shakespeare’s hand.
- The spelling “scilens” (silence) appears in no other play of the era except the 1600 quarto edition of “Henry VI part II” where it appears 18 times as the surname of the character Justice Silence.
- The Latin word “ergo” (therefore) is mispronounced “argo” in both Hand D and “Henry VI part 2”, and as “argal” in Hamlet. There is only one other contemporary example of this error (Middleton’s “The Phoenix” of 1607).
- May be examined by applying two Bayesian approaches (“discrete” and “continuous”) to the raw data of Elliott and Valenza.

Discrete Bayesian approach

- Jackson (2007) was the first to advocate a Bayesian approach to questions of Shakespearean authorship, although Mosteller and Wallace (1964) successfully applied Bayes' theorem to the Federalist papers.
- A great strength of Bayesian analysis is that it can be used to combine many individual pieces of evidence.
- Assuming beforehand that Hand D is equally likely to have or have not been written by Shakespeare, we start with prior odds of 0.5/0.5 = 1.
- Posterior odds = prior odds x likelihood ratio.
- Likelihood ratio = probability of failing the grade level test when the text is not by Shakespeare / probability of failing the grade level test when the text is by Shakespeare
- LR = (18/51) / (0 / 35).
- This involves division by zero, so use the Laplacian correction:
- LR = (19/52) / (1/37) = 13.264
- Posterior odds = 1 x 13.264 = 13.264
- The posterior odds become the new prior odds, and we move to the next test (Feminine endings).

Discrete Approach: Numbers of Shakespeare and Non-Shakespeare samples accepted and rejected by each of 5 tests

Test	S accepted	S rejected	NS accepted	NS rejected
Grade level	35	0	33	18
Feminine endings	35	0	29	22
Open lines	35	0	33	18
BoB5	34	1	30	21
BoB7	34	1	29	22

Hand D: combination of evidence using a discrete Bayesian analysis

Individual test	Hand D outcome	Prior Odds	Likelihood ratio	Posterior odds	Probability not by Shakespeare
Grade Level	Rejected	1	(19/53) / (1/37)	13.264	0.930
Feminine endings	Accepted	13.264	(30/53) / (36/37)	7.720	0.875
Open lines	Accepted	7.720	(34/53) / (36/37)	4.819	0.828
BoB5	Rejected	4.819	(22/53) / (2/37)	37.006	0.974
BoB7	accepted	37.006	(29/53) / (35/37)	21.406	0.955

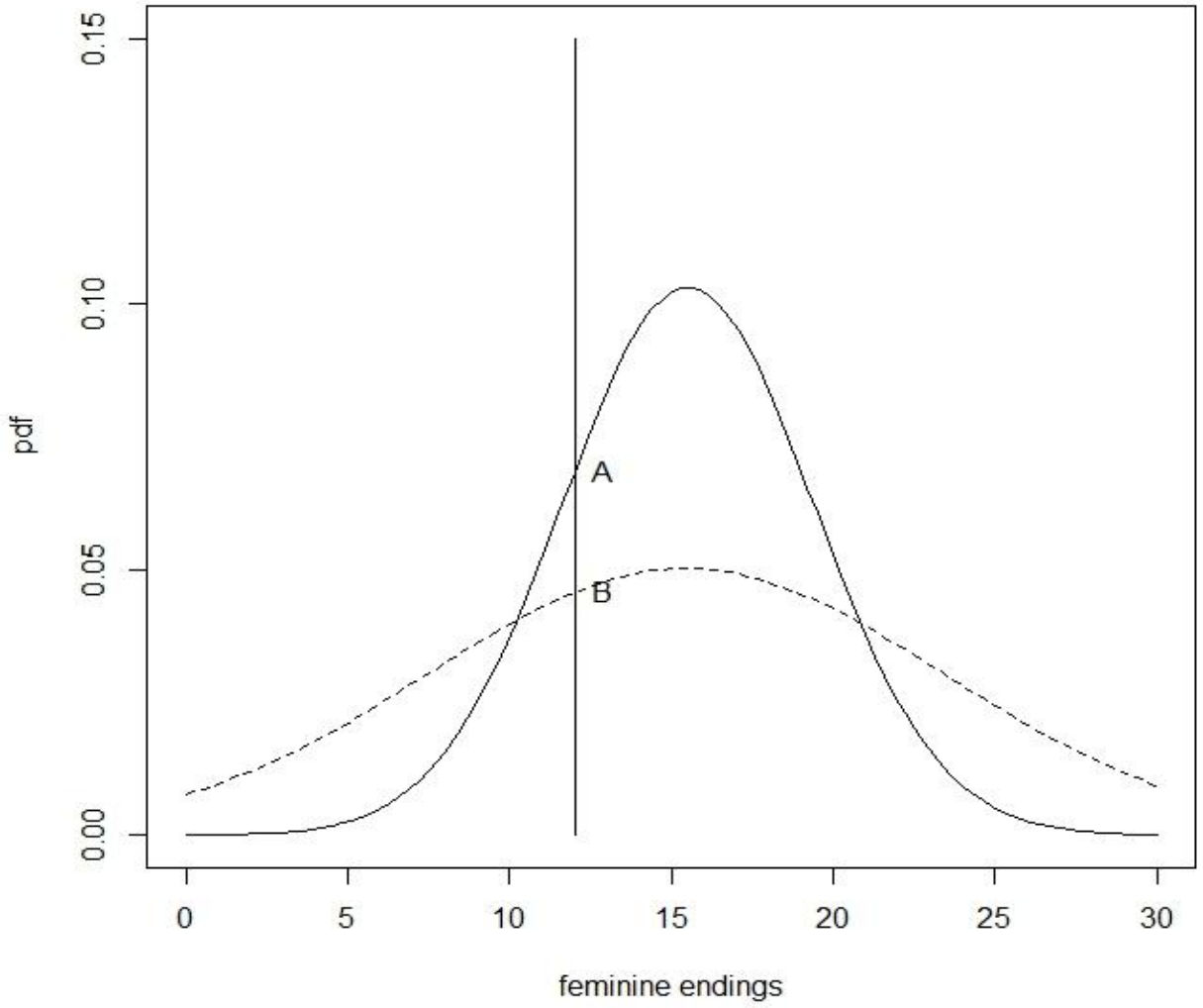
Continuous Bayesian approach

- Taking Feminine Endings as an example.
- Mean number of FE over all non-Shakespeare plays was 15.45 and standard deviation was 7.95.
- From this data, it is possible to draw the corresponding normal distribution curve:
- curve(dnorm(x,15.45, 7.95), add=TRUE)
- This gives the flatter curve (dotted line)
- For the set of Shakespeare plays, the mean was similar (15.51), but the standard deviation was much less (3.87).
- The normal distribution curve (solid line) for this is much sharper.
- The heights of the two curves are the probability density functions (pdf)
- The ratio between the heights of the two curves is a measure of the relative likelihood of a text of unknown authorship with a certain number of feminine endings belonging either to the Shakespeare or non-Shakespeare set.
- The number of FE in Hand D was found to be 12, as shown by the vertical line.

Continuous Bayesian Analysis: Hand D data and summary data for reference samples according to 5 tests.

	Grade level	Feminine endings	Open lines	BoB5	BoB7
Hand D	12	12	43	762	556
NS mean	6.00	15.45	22.58	243.00	301.05
MS std. dev	2.41	7.95	8.91	161.60	409.62
NS pdf	.00742	.0456	.0221	.000889	.000759
S mean	4.83	15.51	26.29	294.11	520.54
S std dev	0.82	3.87	11.44	92.32	154.76
S pdf	1.43 e-17	.0683	.0160	4.05 e-05	1.165 e-05
LR = NS pdf / S pdf	5.17 e 14	.669	1.384	21.92	65.16

Probability density functions for feminine endings in both Shakespearean and non-Shakespearean samples



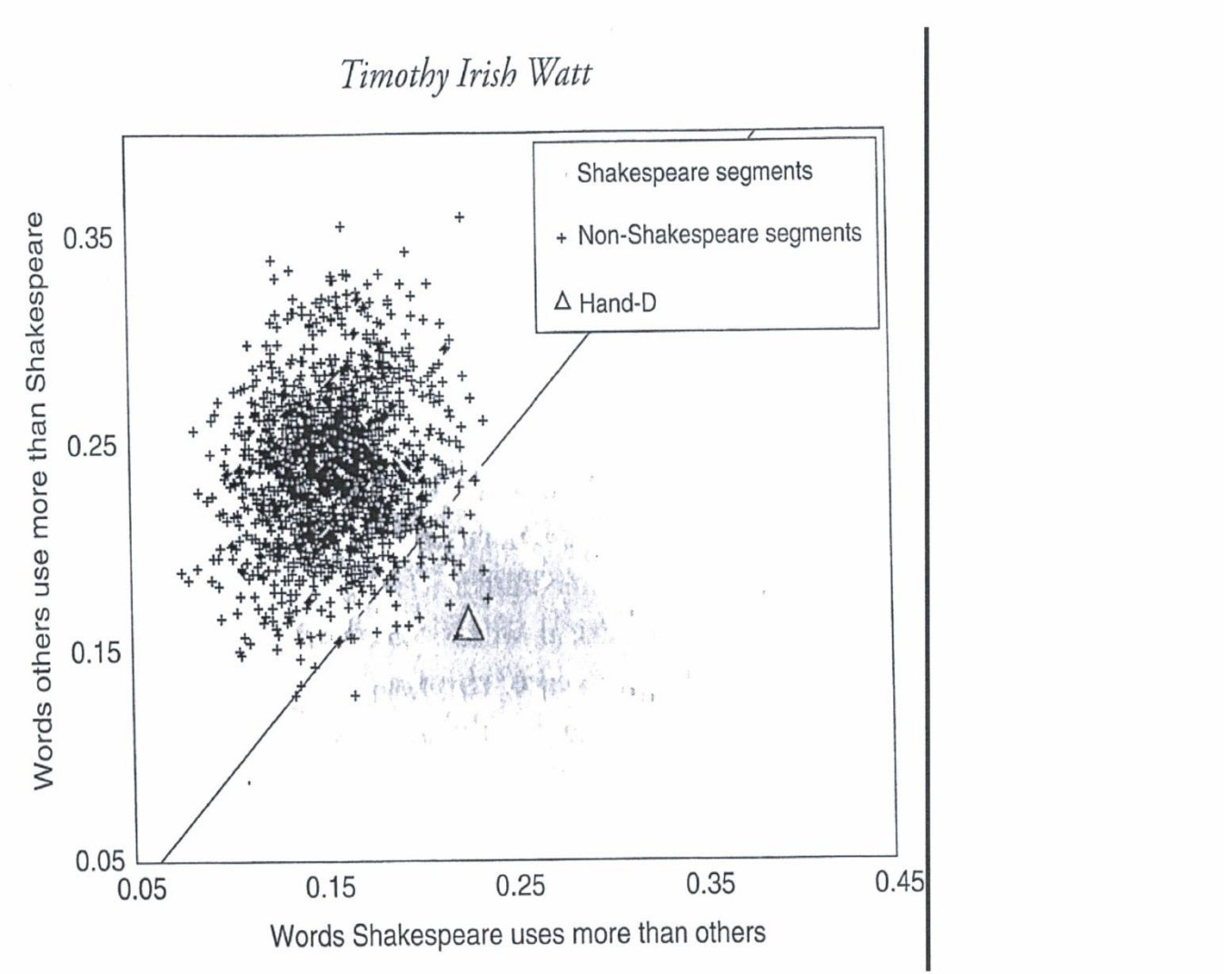
Continuous Bayesian approach (2)

- The heights of the curves at any point can be found using the command dnorm(value, mean, std dev), e.g. dnorm(12, 15.51, 3.87) for the Shakespeare curve.
- The ratio of the pdf values at these two points is 0.0683 / 0.0456 = 1.498. This is the likelihood ratio.
- Since this value is more than 1, the evidence of the FE is that the Hand D text is more likely to be by Shakespeare.
- The evidence of all the tests is taken into account by multiplying together the likelihood ratios.

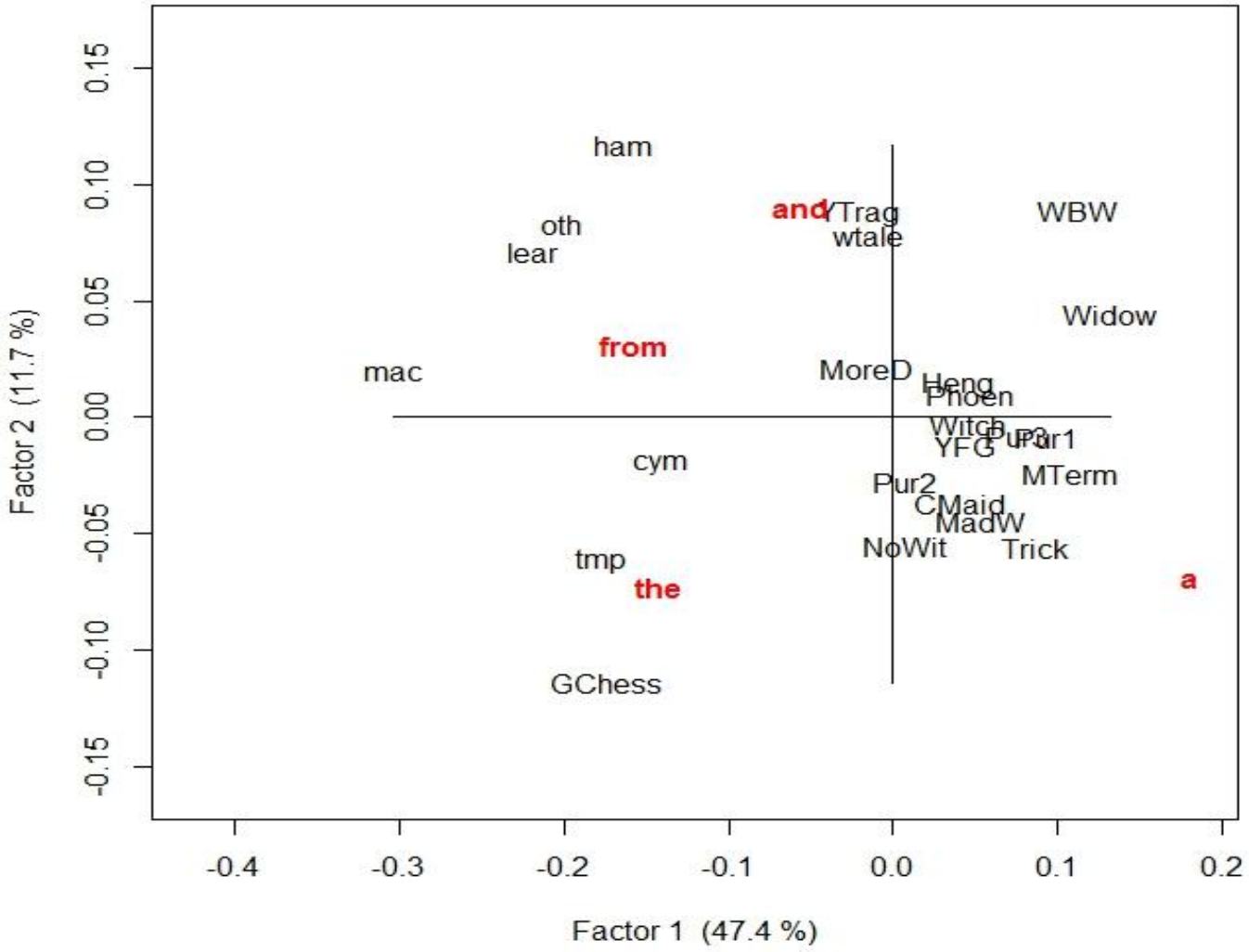
Distance between the Hand D centroid and the centroids of the base and counter sets for 6 candidate authors (Watt, 2009) – Zeta method

Base set	plays	Counter set	plays	Distance (base)	Distance (counter)
Dekker	5	Non-Dekker	107	0.637	0.172
Heywood	5	Non-Heywood	107	0.146	0.033
Jonson	12	Non-Jonson	100	0.08	0.006
Middleton	10	Non-Middleton	102	0.096	0.013
Shakespeare	27	Non-Shakespeare	85	0.019	0.054
Webster	3	Non-Webster	109	0.146	0.012

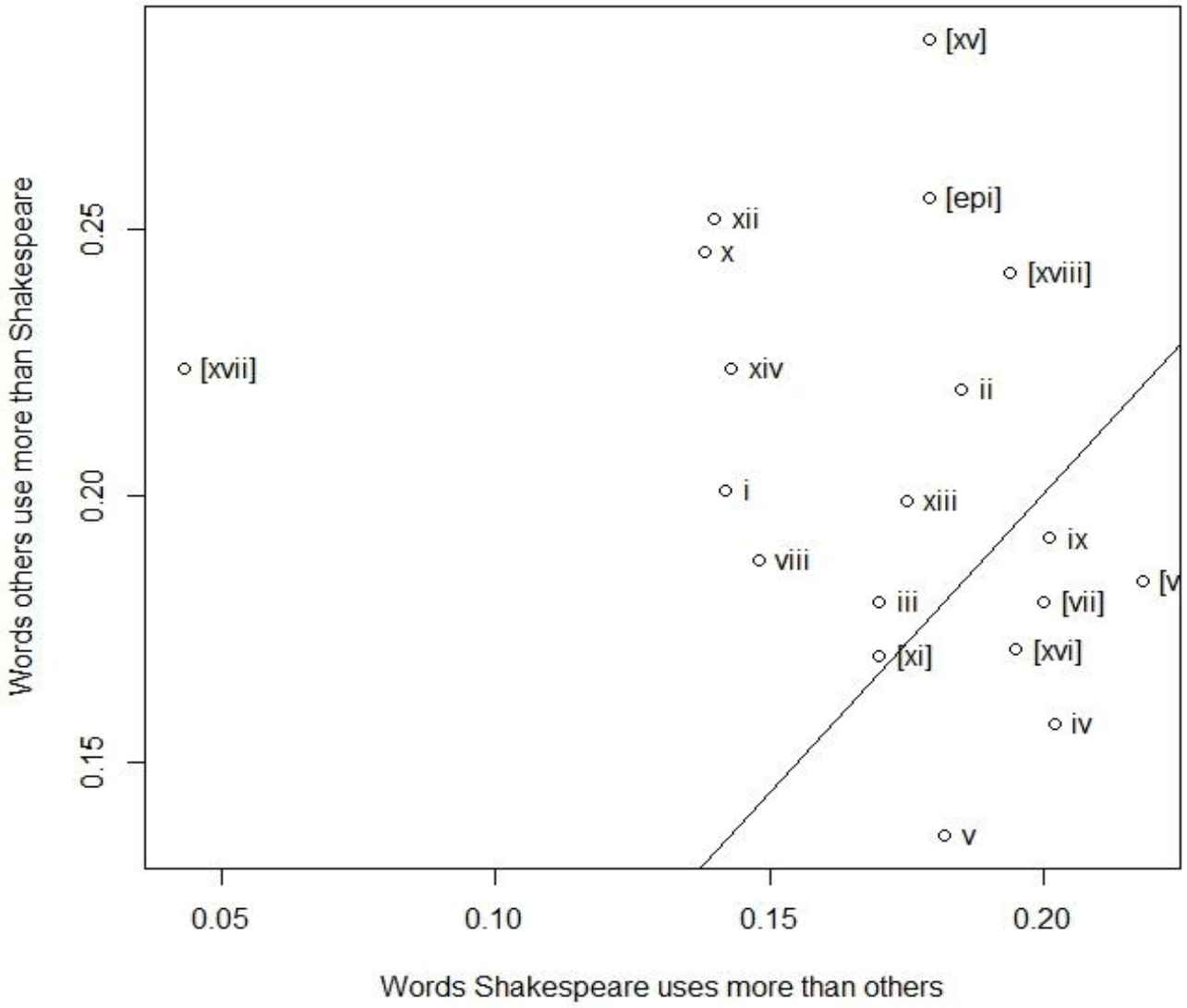
Watt (2009): Burrows' Zeta for Hand D. Note one method of distance between centroids in Euclidean distance. Different conclusion to Elliott and Valenza.



Correspondence Analysis for Shakespeare and Middleton (Ytrag = Yorkshire Tragedy, Pur1 & 2 = Puritan) (data of Jackson 1979)



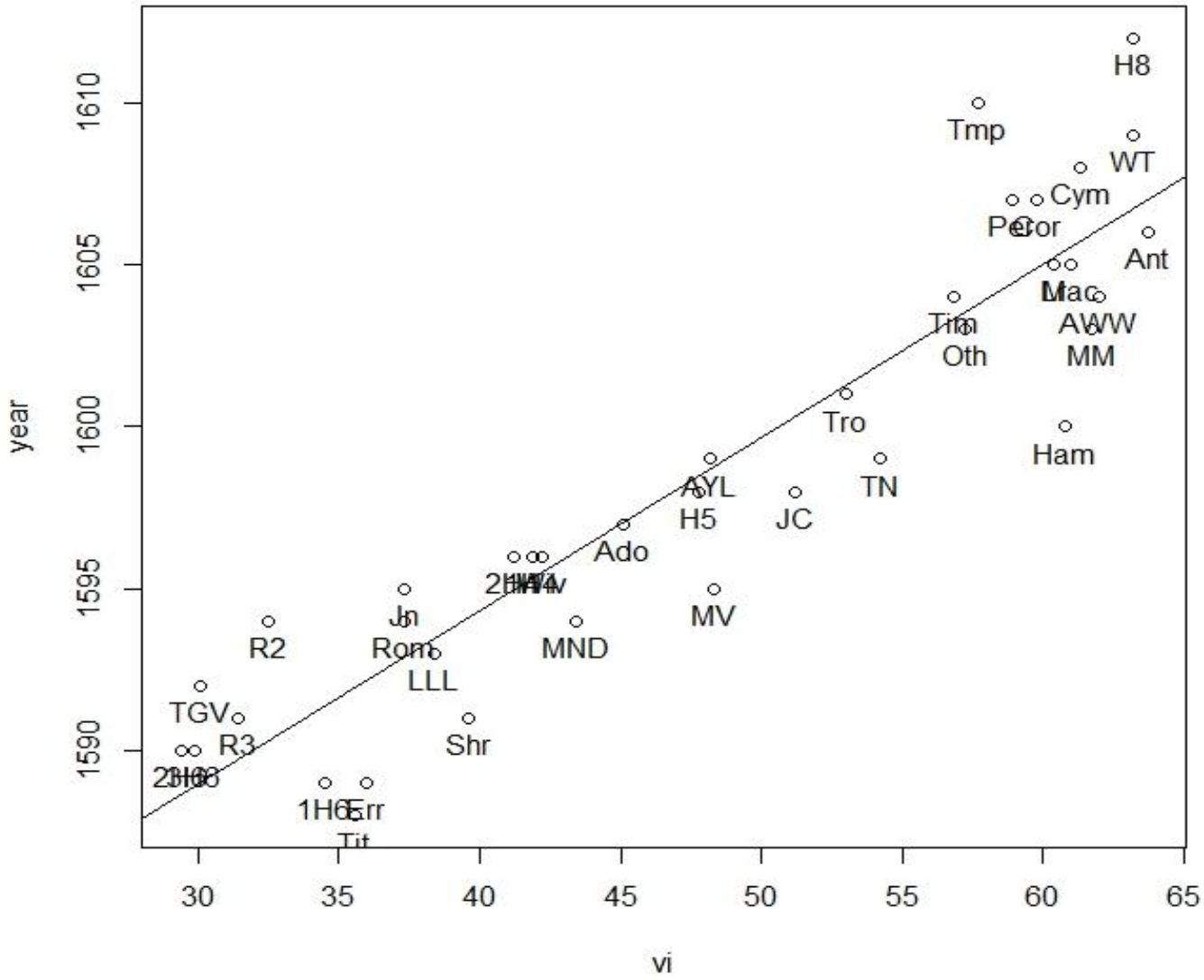
Analysis of "Arden of Faversham" by Burrows' Zeta method (Kinney, 2009)



Thisted and Efron (1987)

- Paxton's method for estimating how many species of sea monsters are "out there".
- Bennett et al. (2002) analogy – a party ("first sample") where you meet 12 Swedish, 9 Chinese, 6 French, 4 Israeli, 3 Korean and 1 Iranian. Even if you meet no-one else, from this you can estimate the total number of nationalities represented by the party.
- Used by Efron and Thisted (1976) to estimate Shakespeare's total vocabulary – canon is "first sample" (about 66,000 words).
- Gary Taylor discovered a sonnet "Shall I die" – word frequencies "reasonably consistent" with a "second sample" of Shakespeare.

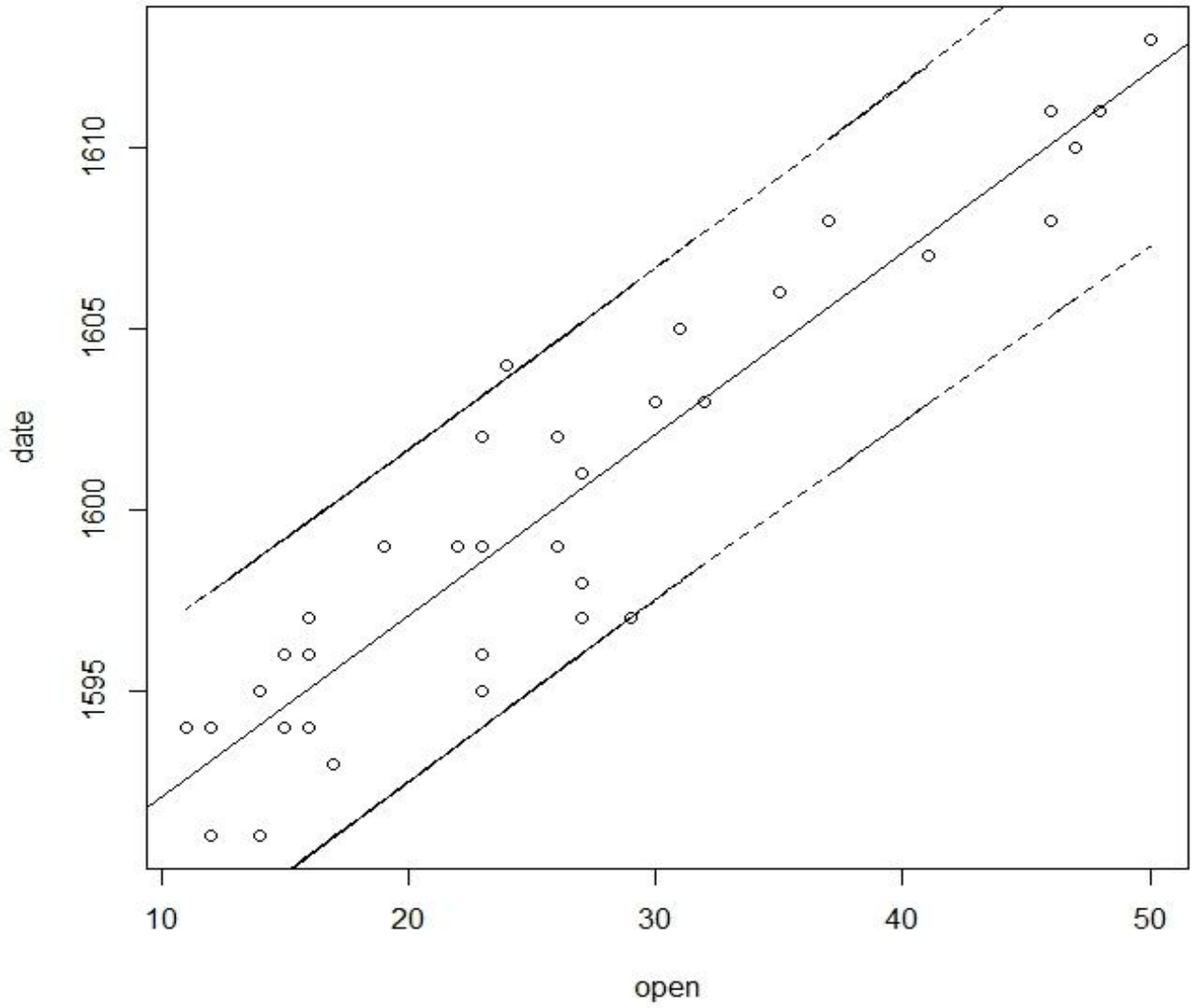
Jackson's chronology of Shakespeare



Notes on Regression

- The dates of Shakespeare's plays and the corresponding "open line" counts can be placed into lists (or "arrays") using R:
- date = c(1591, 1591, 1593, 1594, 1594, 1594, 1594, 1595, 1595, 1596, 1596, 1596, 1597, 1597, 1597, 1598, 1599, 1599, 1599, 1599, 1601, 1602, 1602, 1603, 1603, 1604, 1605, 1606, 1607, 1608, 1608, 1610, 1611, 1611, 1613)
- open = c(14, 12, 17, 15, 11, 16, 12, 23, 14, 23, 15, 16, 29, 16, 27, 27,22, 19, 26, 23, 27, 23, 26, 30, 32,24, 31, 35,41, 37, 46, 47, 46, 48, 50)
- cor.test(open,date)
- res = lm(date ~ open)
- res
- will then show that the intercept (where the line cuts the vertical or "y" axis) is 1587.1 and the slope is 0.5012. This means that we can estimate the date of a "new" Shakespeare play using the formula date = 1587.1 + (0.5012 * percentage of open lines).
- resid(res) shows that in no case are we more than five years from the true values. We can also calculate a quantity called the "prediction intervals" (95% confidence bounds)
- pred.res = predict(res, int="pred")
- pred.res
- The original regression line and the prediction intervals can be plotted out as follows:
- plot(date ~ open)
- abline(res)
- lines(open,pred.res[,2],lty=2)
- lines(open,pred.res[,3],lty=2)

Regression line for proportion of open lines and date



Correlation Coefficients for 7 predictor variables and date

feature	Correlation coefficient	p	Intercept	Slope
Fem	.727	7.3 e-07	1582.3	1.154
Open	.935	2.2 e-16	1587.1	0.5012
Total7	.847	1.3 e-10	1594.4	0.1182
Ll	.317	0.06		
d/'ld	.765	8.7 e-08	1596.6	1.495
Very	.220	0.20		
Most	.668	1.2 e-05	1592.9	0.2886

Conclusion

- We started by reviewing a list of 35 plays considered to have certainly been written by Shakespeare, which contains all of his best-known works.
- There is no convincing computer evidence to suggest that anyone else wrote these.
- Collaborative works: computer studies can shed light on which parts of these plays each author was involved in.
- Apocrypha: possibly “Arden of Faversham” and “A Yorkshire Tragedy” were written by Shakespeare.
- Researchers are divided on the authorship of “Hand D”
- Stylochronometry for establishing a timeline for Shakespeare’s plays.