

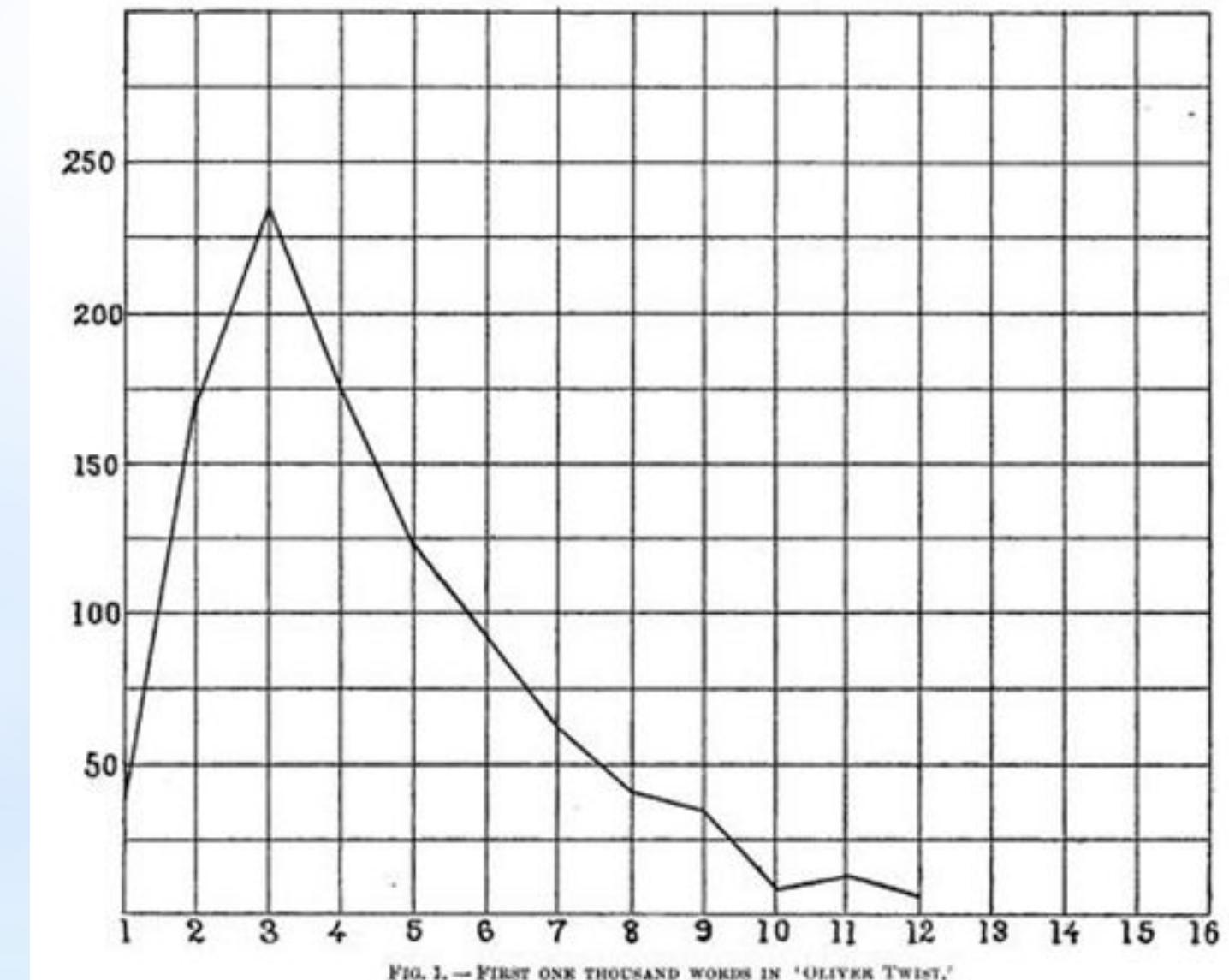


STYLOMETRY

THE IDEA

«Measuring» authorial style

(Mendenhall, 1887)



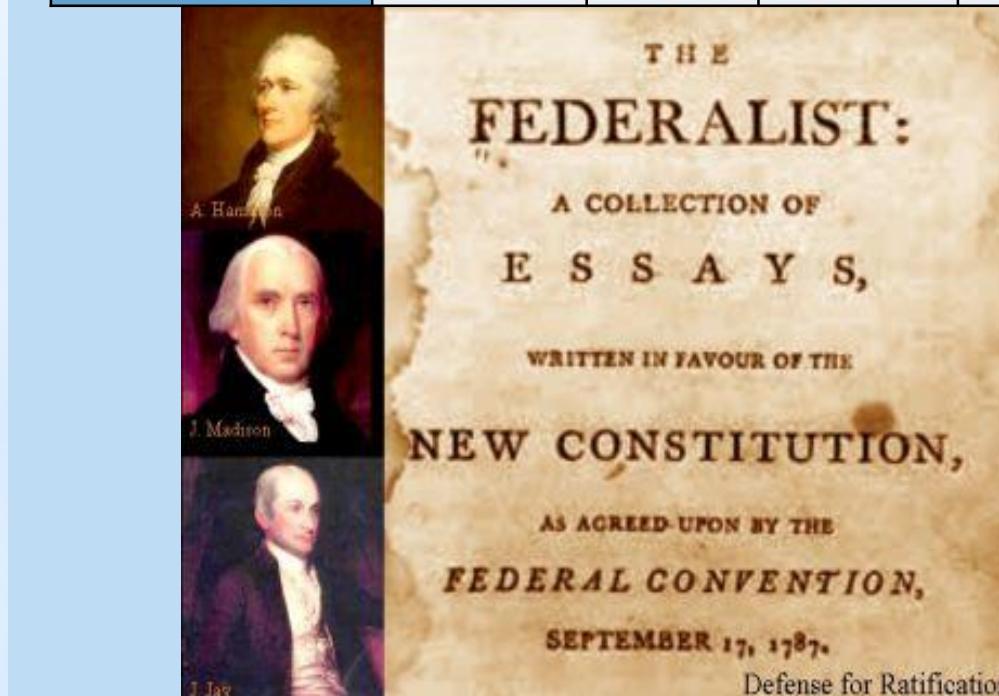
THE IDEA

«Measuring» authorial style

Successes

(Mosteller and Wallace 1964)

	enough	while	whilst	upon
Hamilton	0.59	0.26	0	2.93
Madison	0	0	0.47	0.16
Disputed texts	0	0	0.34	0.08
Co-authored texts	0.18	0	0.36	0.36



Defense for Ratification

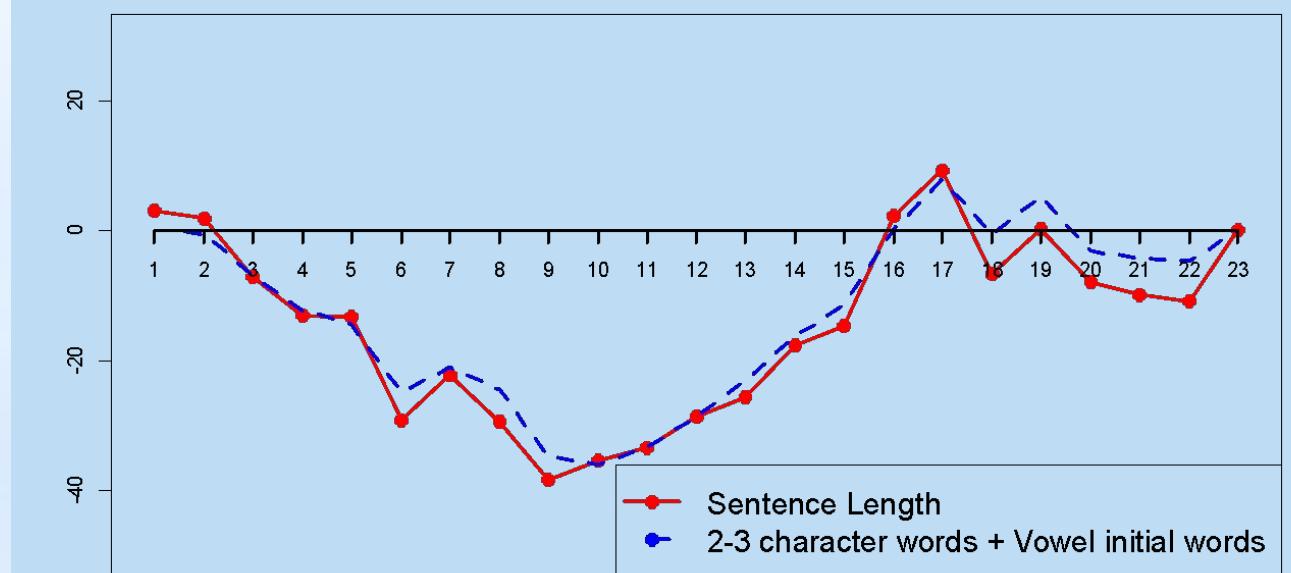
THE IDEA

«Measuring» authorial style

Failures

Andrew Morton in the early '60 adapted **Cumulative Sum – CUSUM or QSUM**

During a BBC live show (1993):
Documents of convicted criminals were attributed to ... the
Secretary of State for Justice!!!



THE (PLETHORA OF) METHODS FOR STYLOMETRY AND AUTHORSHIP ATTRIBUTION

- Character-level analysis
- Syntax-level analysis
- Multi-method analysis (e.g. JGAAP, PAN competition software...)
- ...and many others
- In this lesson, **just two methods:**
 - **Delta method** (for authorship attribution)
 - **Zeta method** (for the quantitative analysis of style)

WORD-FREQUENCY BASED STYLOMETRY

'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship¹

"Literary and Linguistic Computing"
17, no. 3
(2002): 267–87

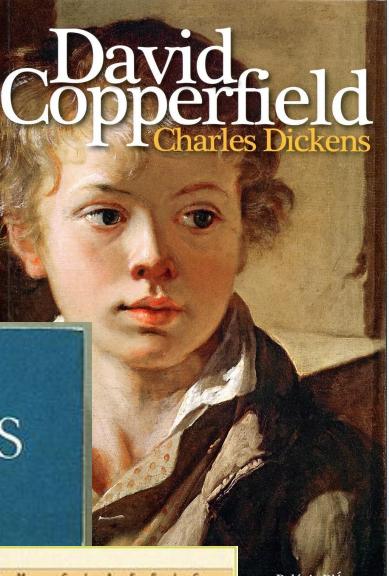
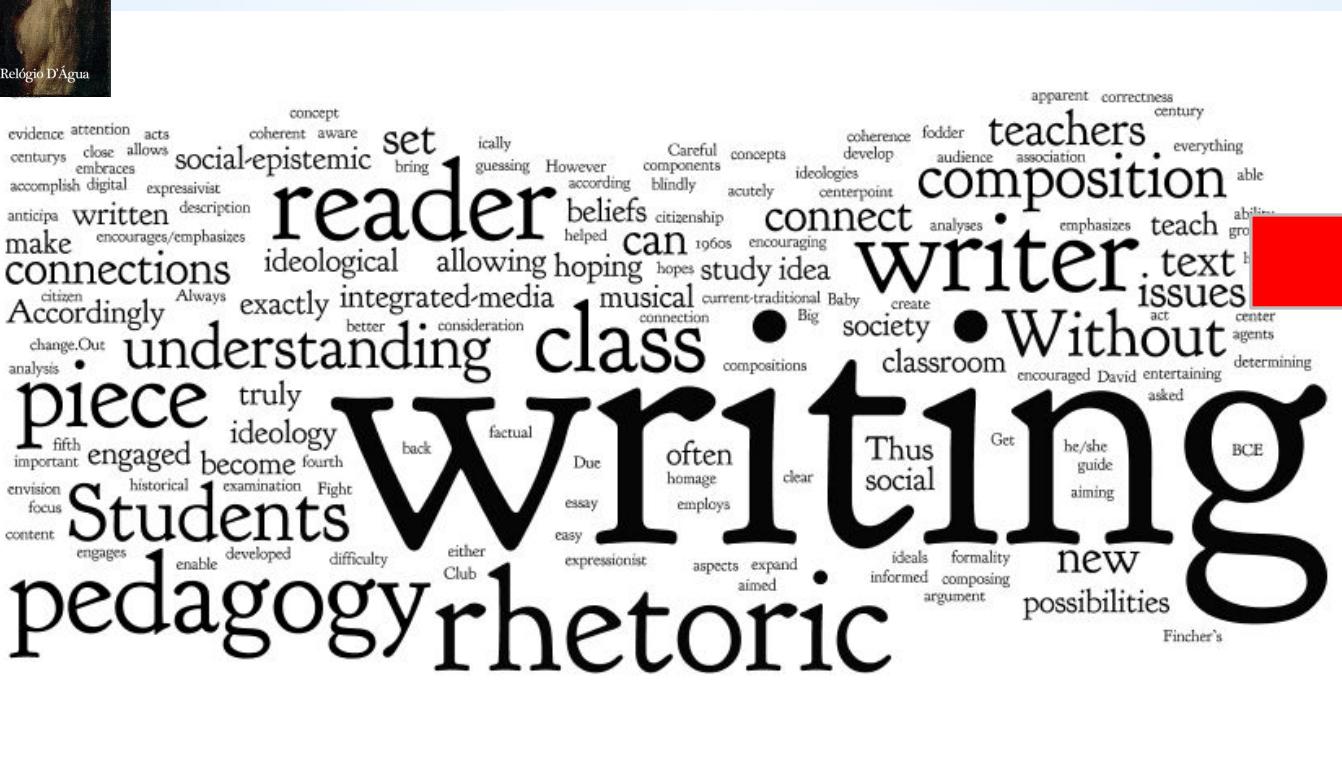
John Burrows
University of Newcastle, Australia

Abstract

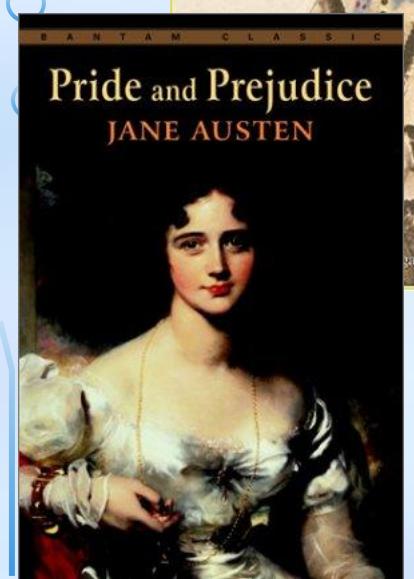
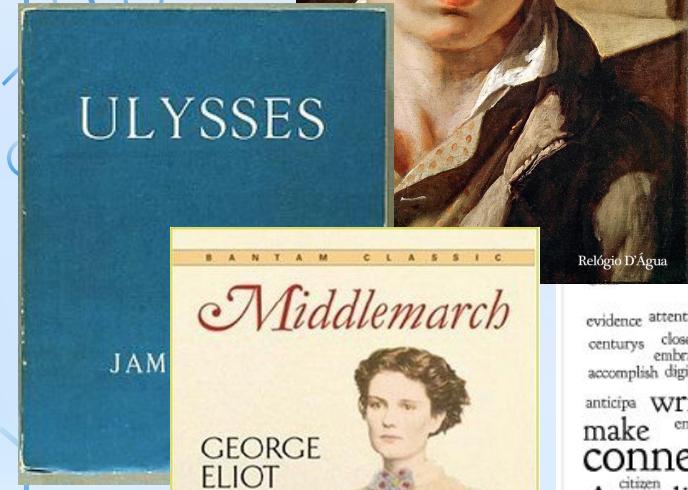
This paper is a companion to my 'Questions of authorship: attribution and beyond', in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself; to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it

- the
- and
- of
- to
- a
- i
- in
- he
- was
- it
- that
- you
- his
- her
- with
- as
- had
- she
- for

DELTA DISTANCE



Relógio D'Água



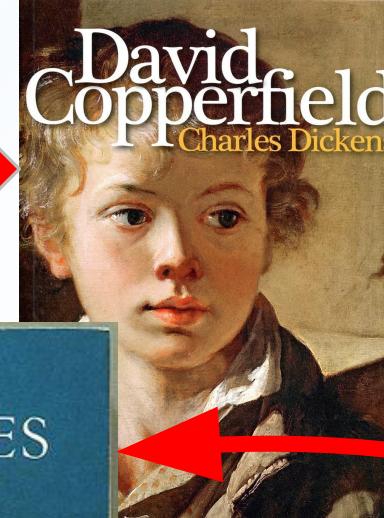
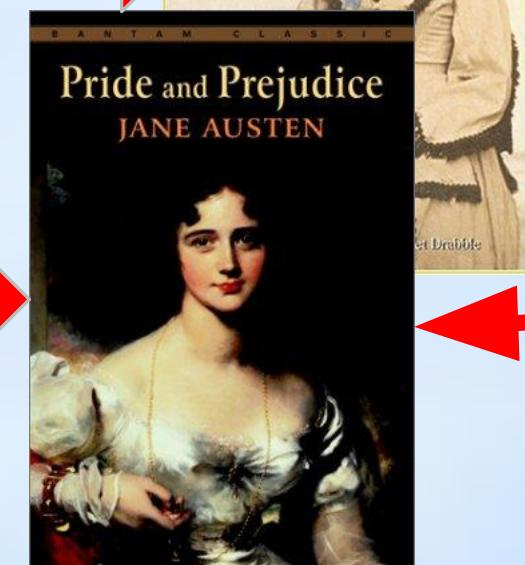
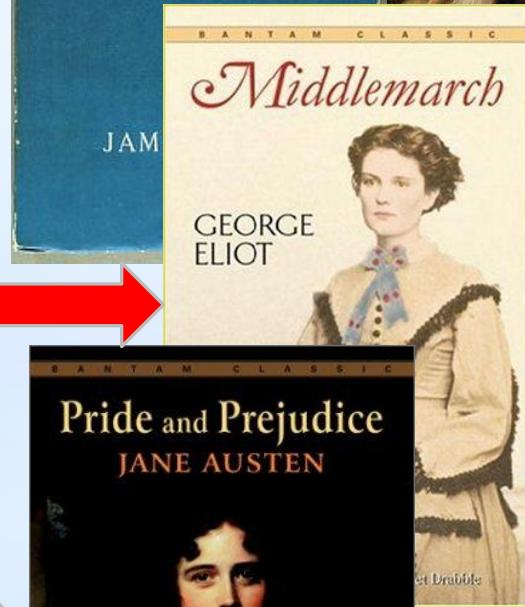
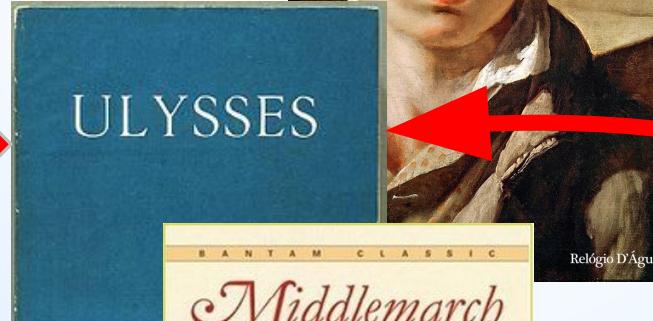
1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

5.1%
3.2%
2.4%
2.5%

4.1%
3.3%
2.2%
2.7%

3.1%
4.2%
1.4%
1.2%

5.2%
3.2%
2.4%
2.5%



$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|$$

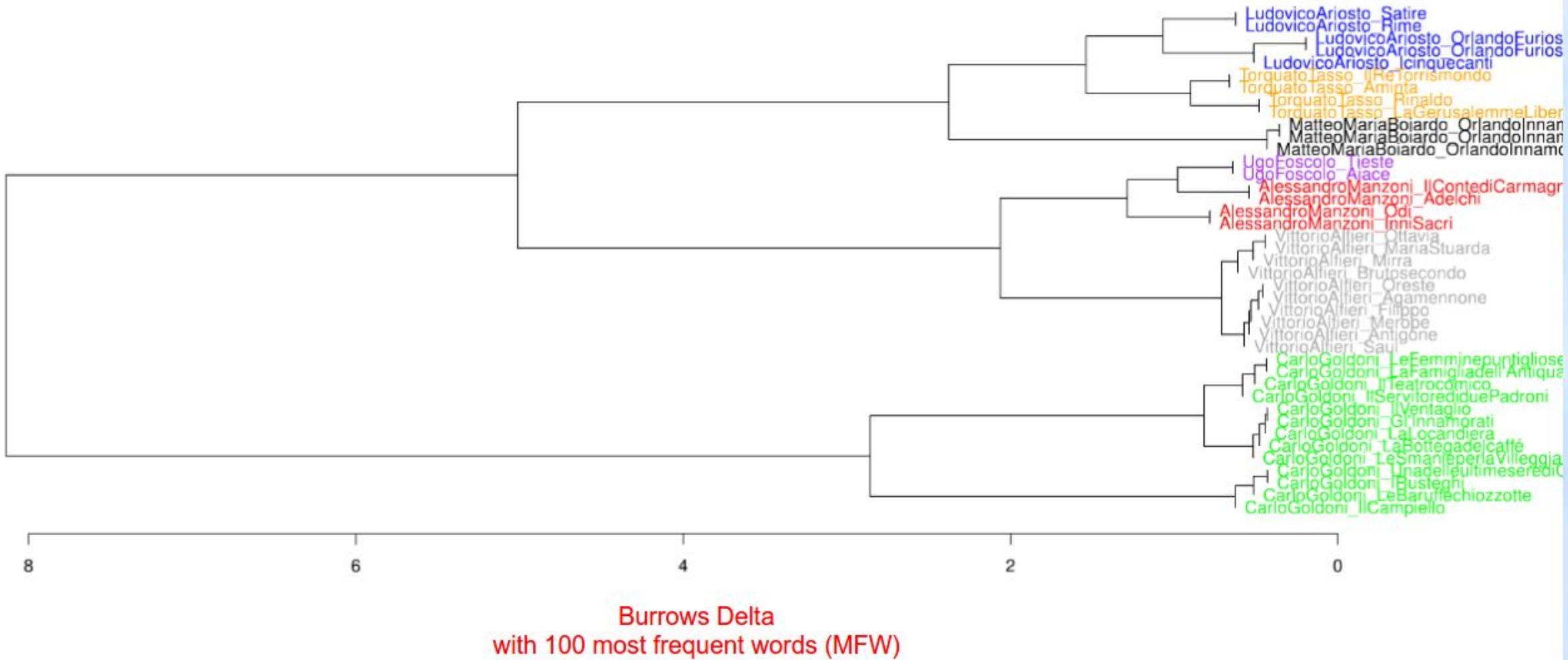
A	B	C	D	E	F
	AlessandroManzoni_Adelchi	AlessandroManzoni_IIContediCarmagnola	AlessandroManzoni_InniSacri	AlessandroManzoni_Odi	AlessandroManzoni_Poesiegio
2	AlessandroManzoni_Adelchi	0	0,481290655	0,666926925	0,738545533
3	AlessandroManzoni_IIContediCarmagnola	0,481290655	0	0,746348745	0,814261157
4	AlessandroManzoni_InniSacri	0,666926925	0,746348745	0	0,633663965
5	AlessandroManzoni_Odi	0,738545533	0,814261157	0,633663965	0,7338
6	AlessandroManzoni_Poesiegiovanili	0,568820863	0,654375023	0,634854567	0,7338
7	CarloGoldoni_GlInnamorati	0,980786338	0,936018177	1,013723738	1,101305203
8	CarloGoldoni_IICampiello	1,016924762	1,031300757	1,018625104	1,092680684
9	CarloGoldoni_IIServitorediduePadroni	0,94860233	0,9266662976	0,976288639	1,080804722
10	CarloGoldoni_IITeatrocomico	0,915941412	0,896367382	0,971870697	1,085346366
11	CarloGoldoni_IIVentaglio	1,011953514	1,00041649	1,074888328	1,131792245
12	CarloGoldoni_IRusteghi	1,089096895	1,124315967	1,047451935	1,1240649
13	CarloGoldoni_LaBottegadelcaffé	0,997940632	0,980781404	1,069965126	1,139058754
14	CarloGoldoni_LaFamigliadell'Antiquario	0,97647637	0,968110166	1,038499373	1,080510085
15	CarloGoldoni_LaLocandiera	0,97946604	0,952399004	1,052505983	1,110322738
16	CarloGoldoni_LeBaruffechiozzotte	1,051753673	1,103993387	1,018834132	1,082447143
17	CarloGoldoni_LeFemminepuntigliose	0,940334542	0,938723973	1,008461186	1,076438004
18	CarloGoldoni_LeSmanieperlaVilleggiatura	1,023938091	0,964832878	1,056736183	1,148650567
19	CarloGoldoni_UnadelleultimeserediCarnovale	1,045847956	1,085480986	1,047945641	1,10681856
20	VittorioAlfieri_Agamennone	0,684514153	0,743793265	0,829452563	0,905939302
21	VittorioAlfieri_Antigone	0,73781244	0,801189414	0,824156384	0,91495815
22	VittorioAlfieri_Brutosecondo	0,675393312	0,675937144	0,830722082	0,910174086
23	VittorioAlfieri_Filippo	0,69672213	0,73856813	0,806194725	0,93419818
24	VittorioAlfieri_MariaStuarda	0,693145931	0,715015202	0,806081448	0,948928306
25	VittorioAlfieri_Merope	0,735463235	0,783055974	0,855979157	0,971583955
26	VittorioAlfieri_Mirra	0,76329317	0,819104452	0,864045202	0,9659327
27	VittorioAlfieri_Oreste	0,70530237	0,777981376	0,829335057	0,930970217
28	VittorioAlfieri_Ottavia	0,762895099	0,791949819	0,874379901	0,96265065
29	VittorioAlfieri_Saul	0,645417404	0,735038238	0,760393582	0,871007648

VISUALIZATIONS

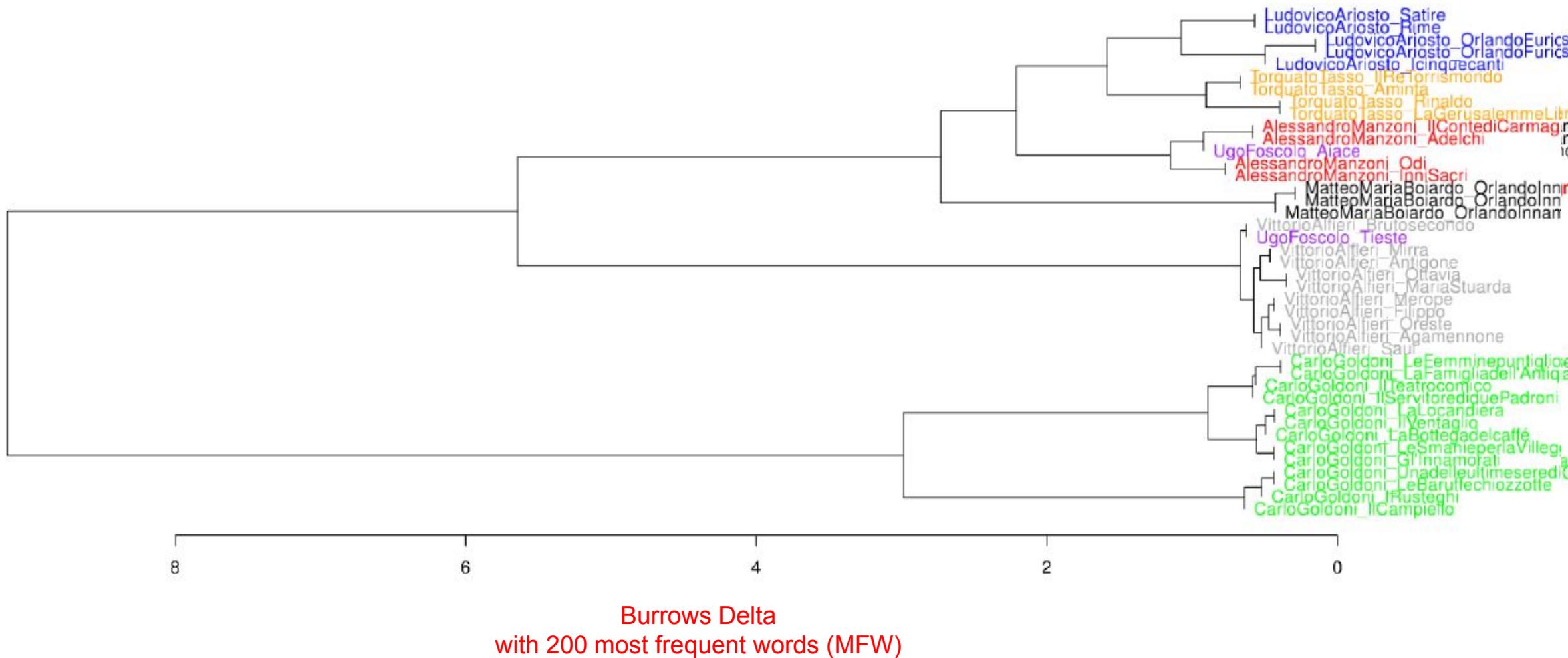
1. Dendograms

Ward's clustering algorithm (Ward, 1963)

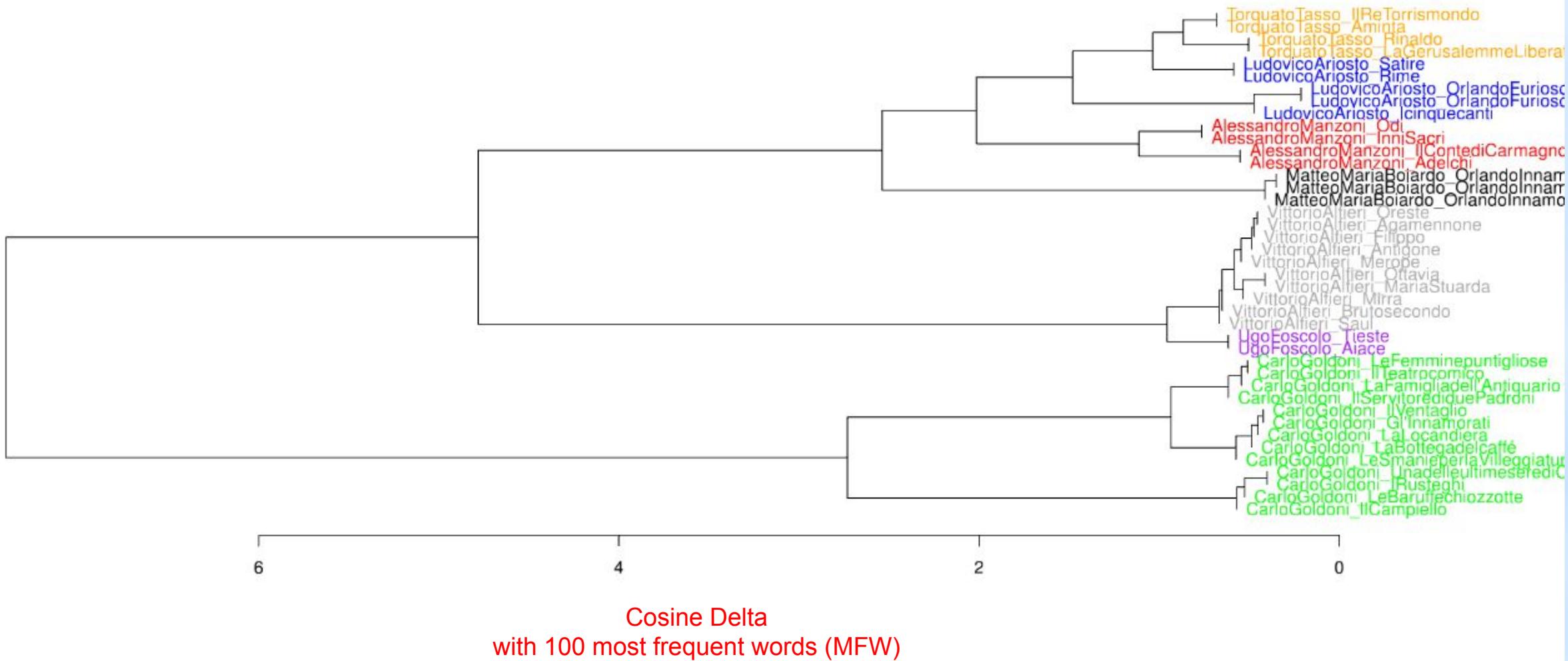
Letteratura Italiana
Cluster Analysis



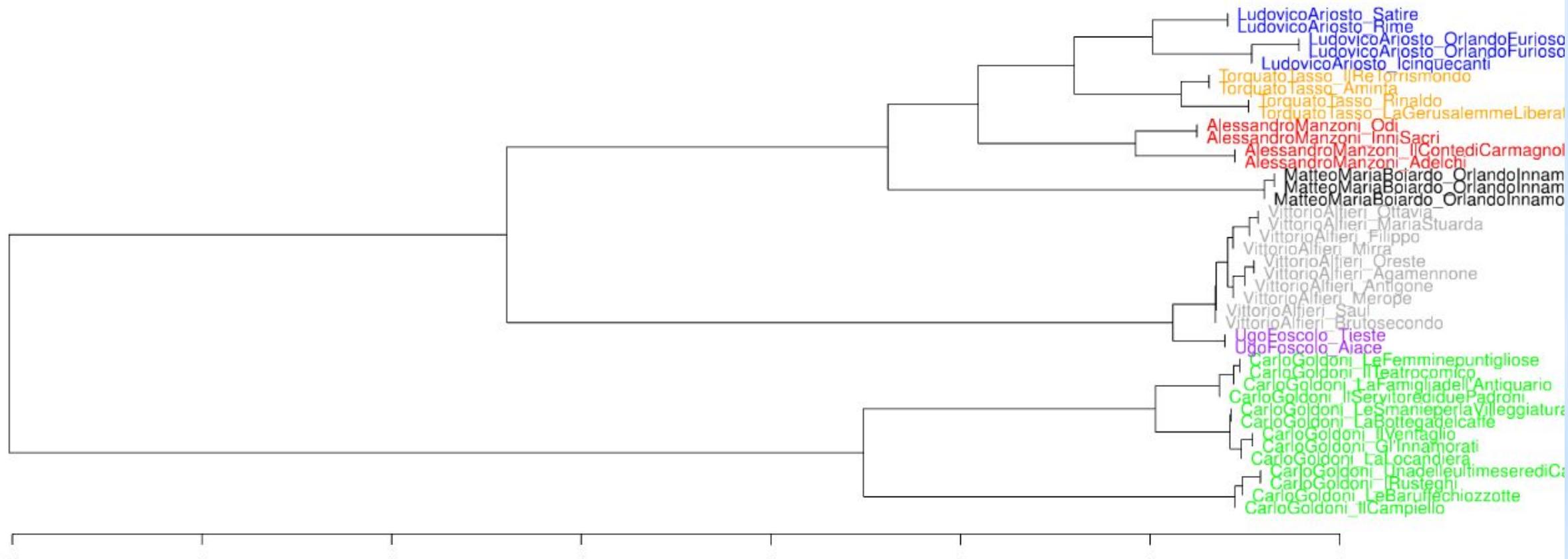
Letteratura Italiana Cluster Analysis



Letteratura Italiana Cluster Analysis



Letteratura Italiana Cluster Analysis



My Weird Distance Measure
with 1,000,000 most frequent words (MFW)

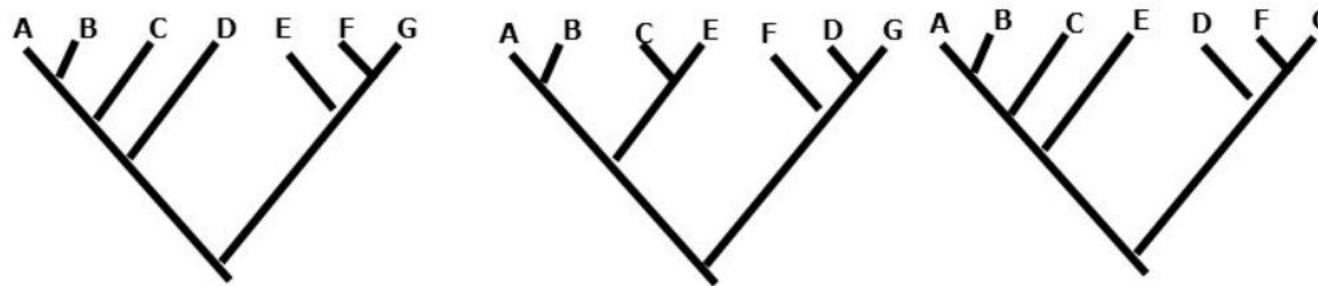
VISUALIZATIONS

2. Consensus Trees

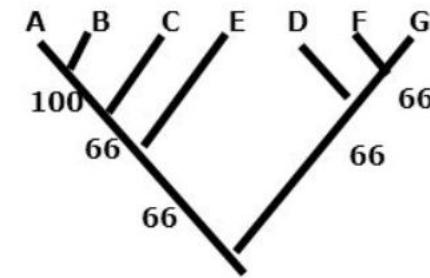
Method developed in phylogenetics
(see Paradis et al. 2004)

Consensus Trees

Majority rule consensus

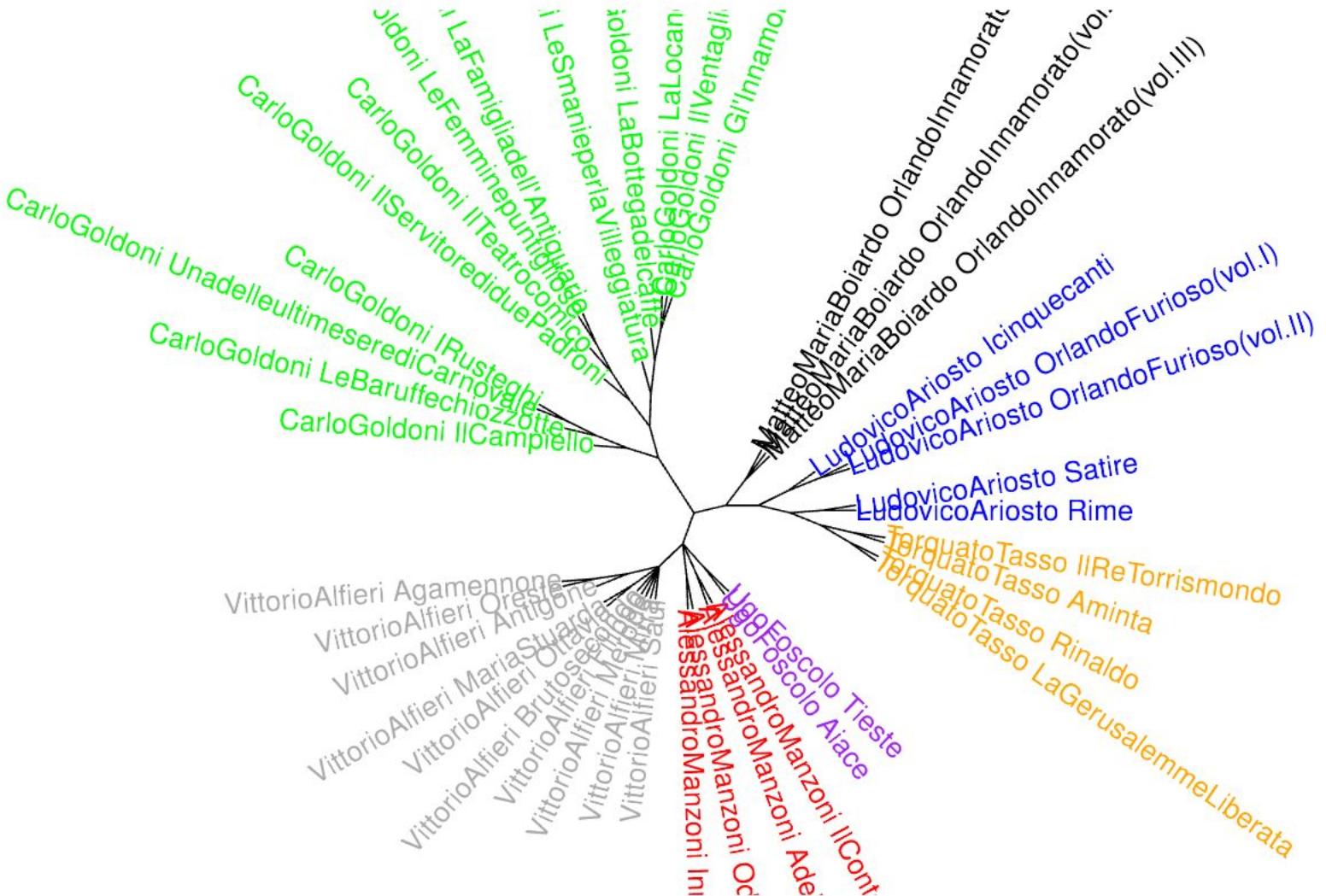


Numbers indicate frequency of
clades in the fundamental trees

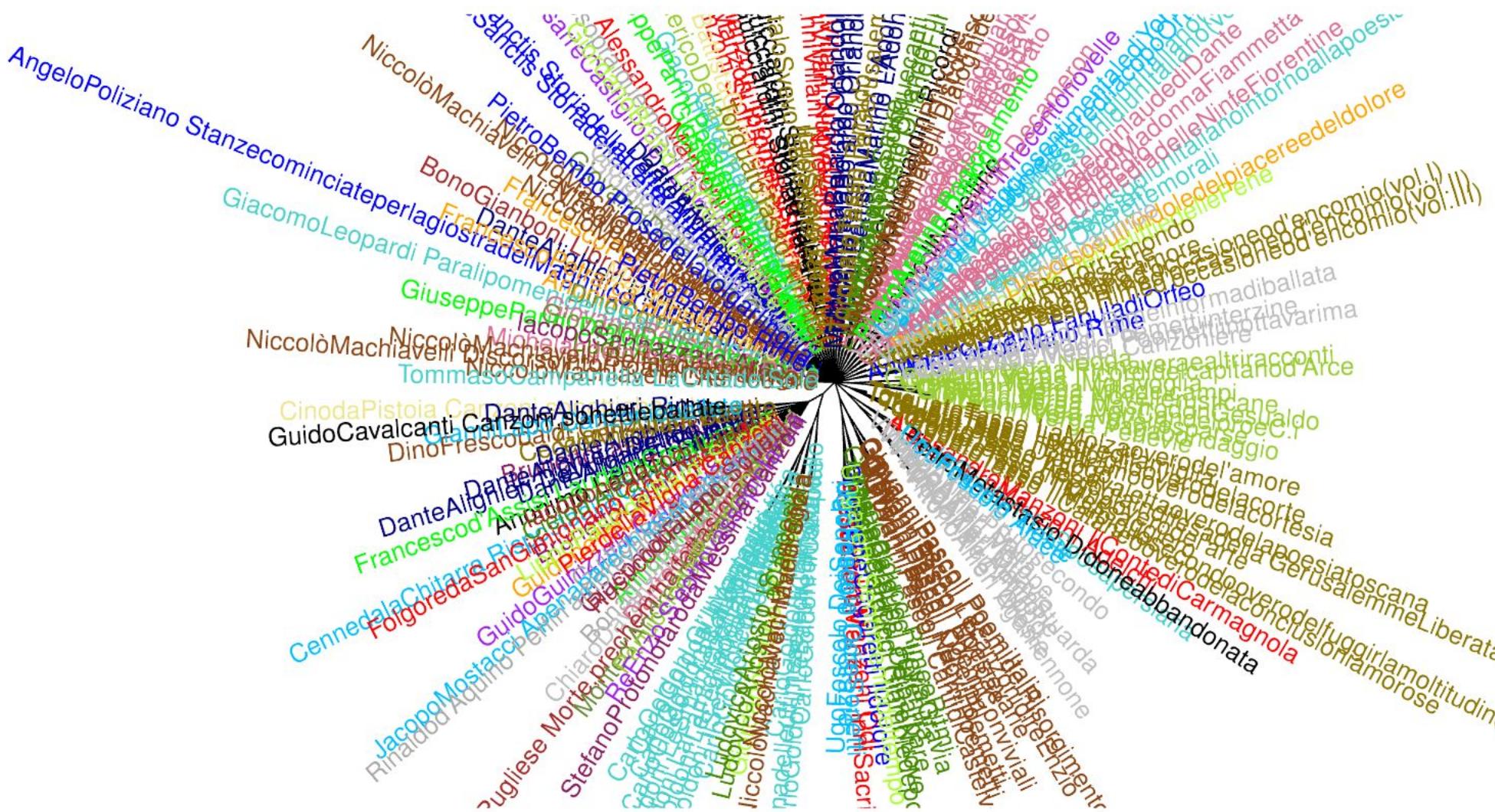


MAJORITY-RULE CONSENSUS TREE

Letteratura Italiana Bootstrap Consensus Tree



Letteratura Italiana Bootstrap Consensus Tree



100–1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

VISUALIZATIONS

3. Network Graphs

See Eder, 2017

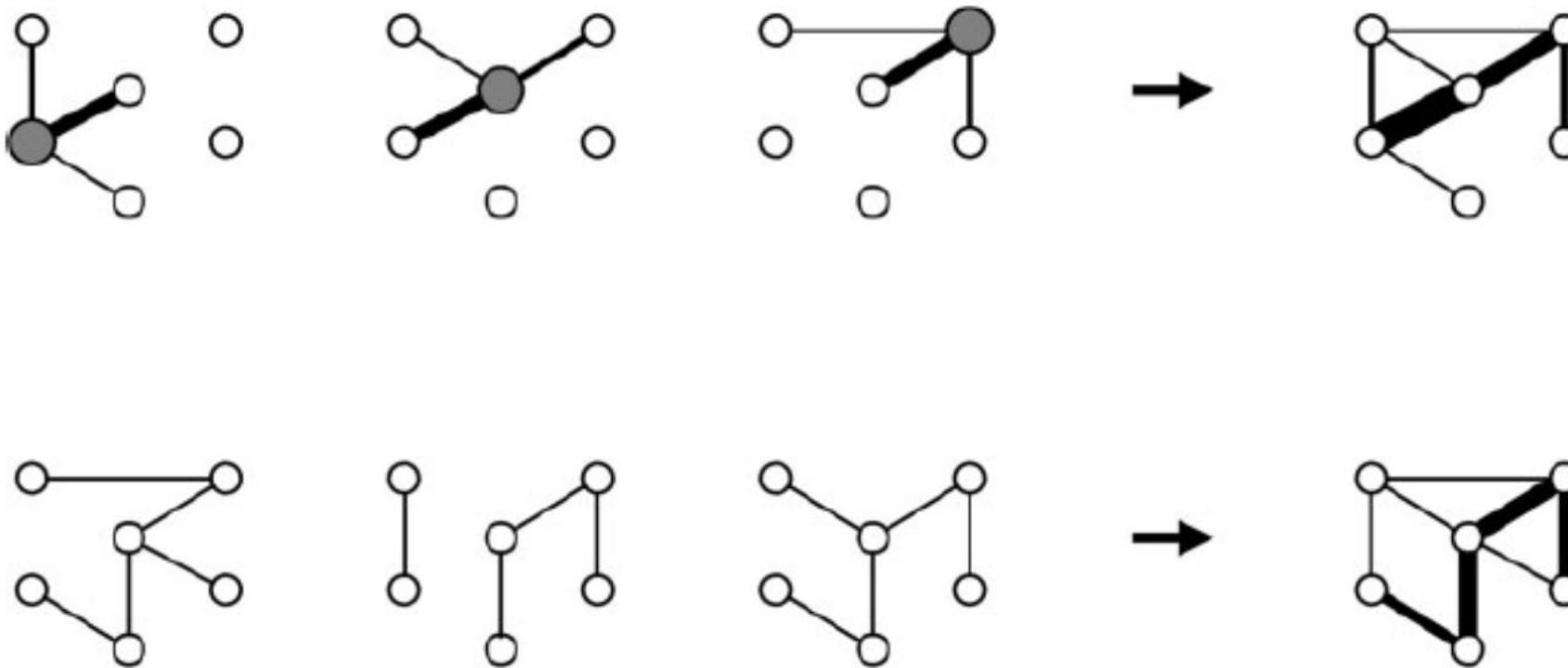
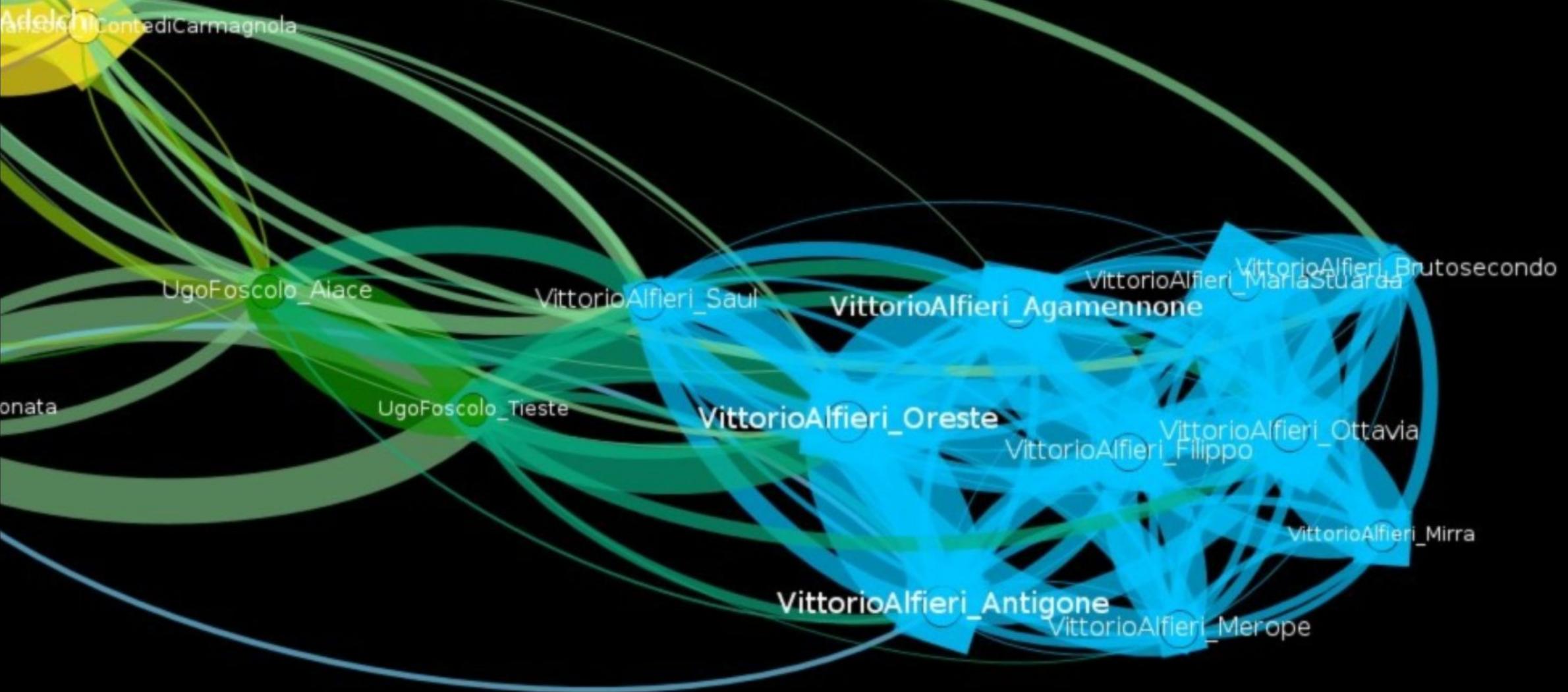


Fig. 6. Two algorithms of mapping textual relations: establishing weighted links to a nearest neighbor and two runners-up (top); producing a consensus network (bottom).



WHY DOES IT WORK?



object they describe. Hence, the old Venetian proverb: "chi guarda cartelo, no magna vedelo," who looks at labels, eats no veal (comes to grief). That Hebrew inscription, however (if it really means Magister (?) Laurentius Costa), is contradicted by the picture itself, which so plainly bears on its face the stamp of Tura, that it might well be set before the tyro as a type of his manner. Again, as this figure of St. Sebastian, excellent in its way, was the occasion of Cosmè being taken for his pupil Costa; so in another famous picture (at present in the house Strozzi at Ferrara) Costa himself has been confounded with his pupil Ercole Grandi di Giulio Cesare. One must, however, admit, that here the scholar has come so close to the manner of the master, that it would not perhaps be too bold to assume, that the composition of the picture comes from Costa, and only the execution belongs to Grandi.³

For the instruction of my young friends, I will here set

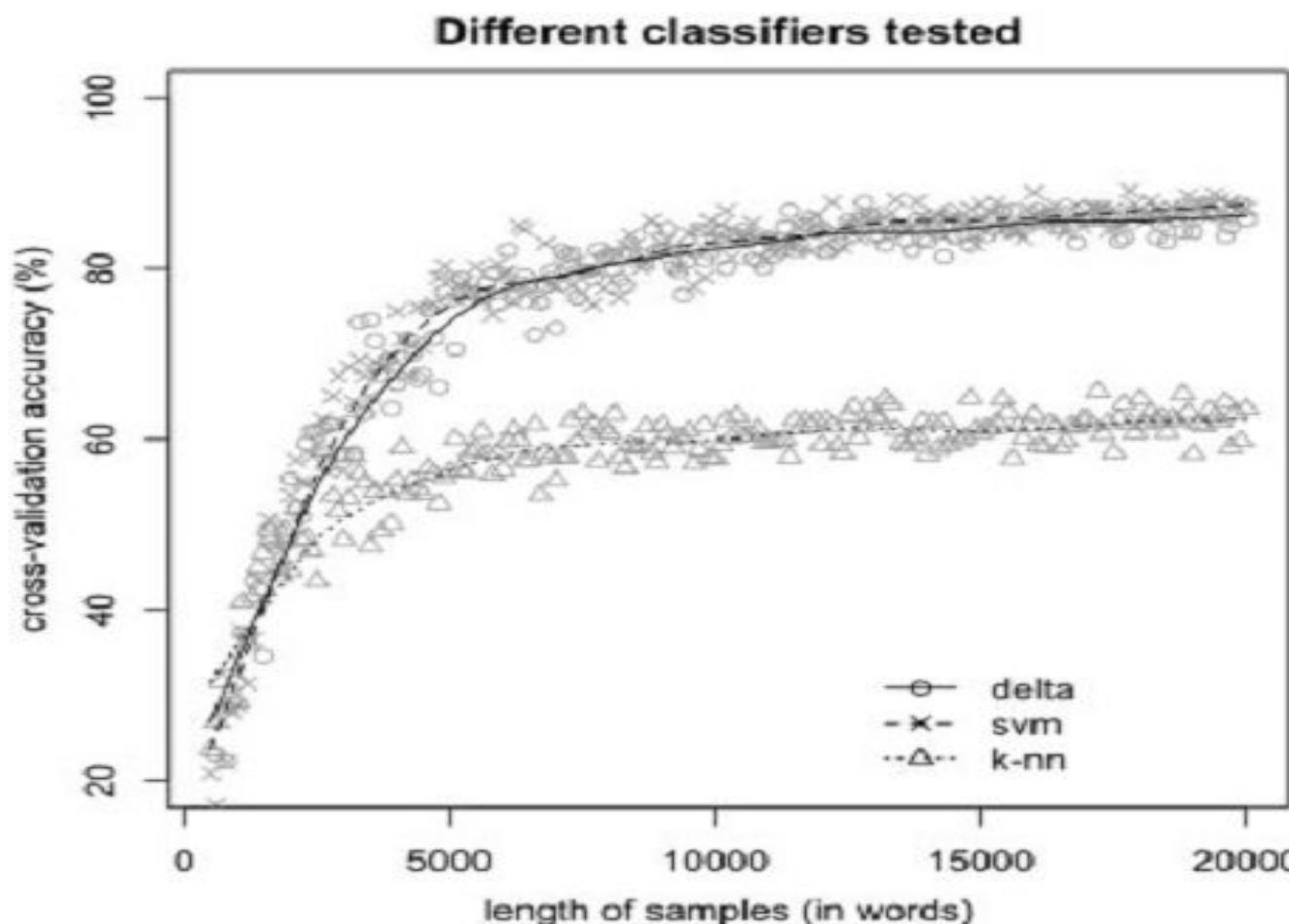


before their eyes a facsimile of the shapes of ear'nd hand in
Cosimo Tura and in Lorenzo Costa, that they may the



"It has been noted that the switch from content words to function words in authorship attribution studies has **an interesting historic parallel in art-historic research.** [...] Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a Quattrocento painting to some Italian master, could not happen based on 'content' [...] Morelli thought it better **to restrict an authorship analysis to discrete details such as ears, hands and feet**" (Kestemont 2014)

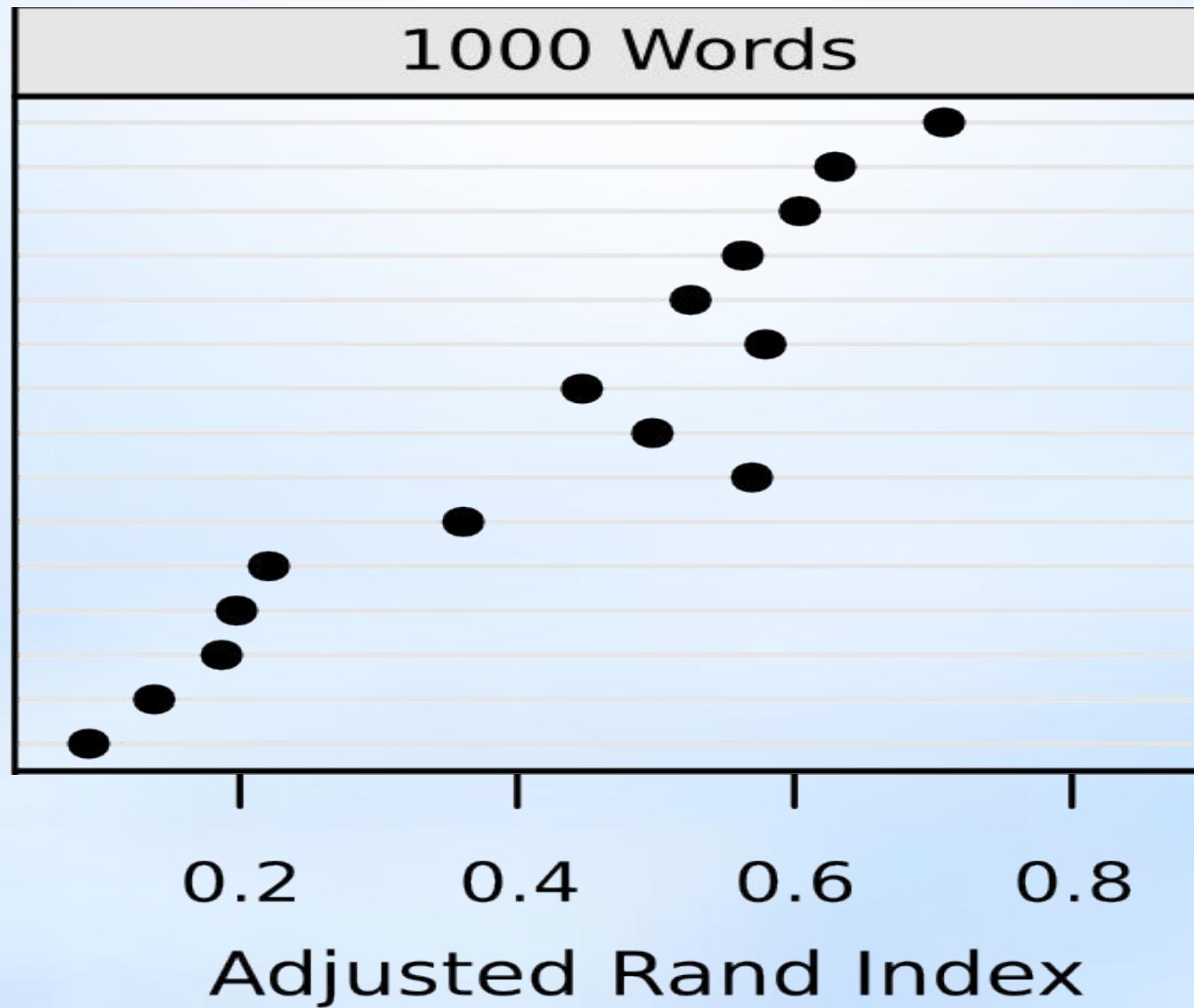
CORPUS SELECTION (TEXT DIMENSIONS)



Minimum text length for a reliable stylometric analysis is **about 5,000 words** (Eder 2015)

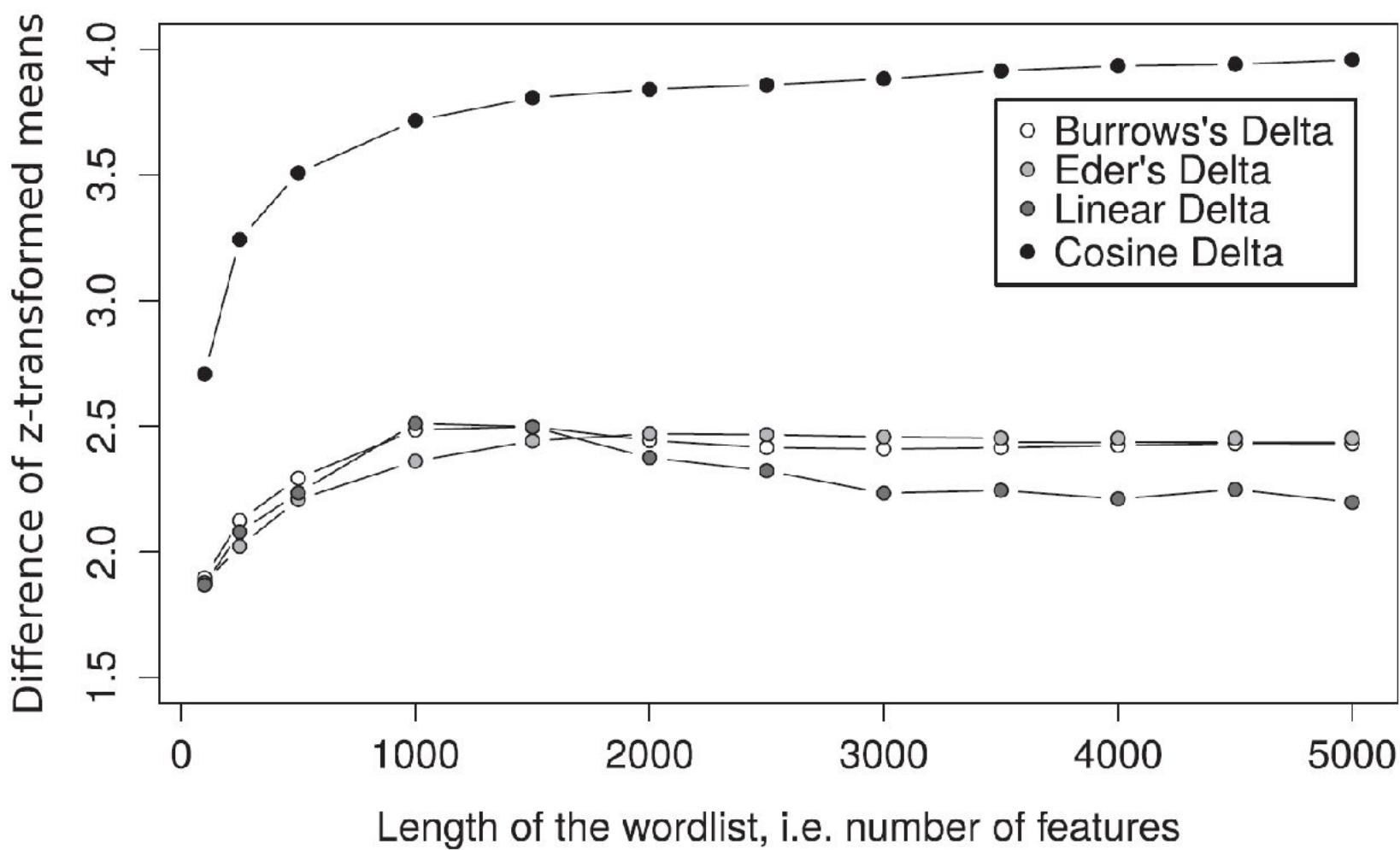
FEATURES SELECTION (DISTANCE MEASURE)

Cosine Delta
Burrows's Delta
Eder's Delta
Hoover's Delta P1
Linear Delta
Eder's Simple Delta
Bray-Curtis
Canberra
Manhattan
Quadratic Delta
Euclidean
Correlation
Cosine
Chebyshev
Rotated Delta



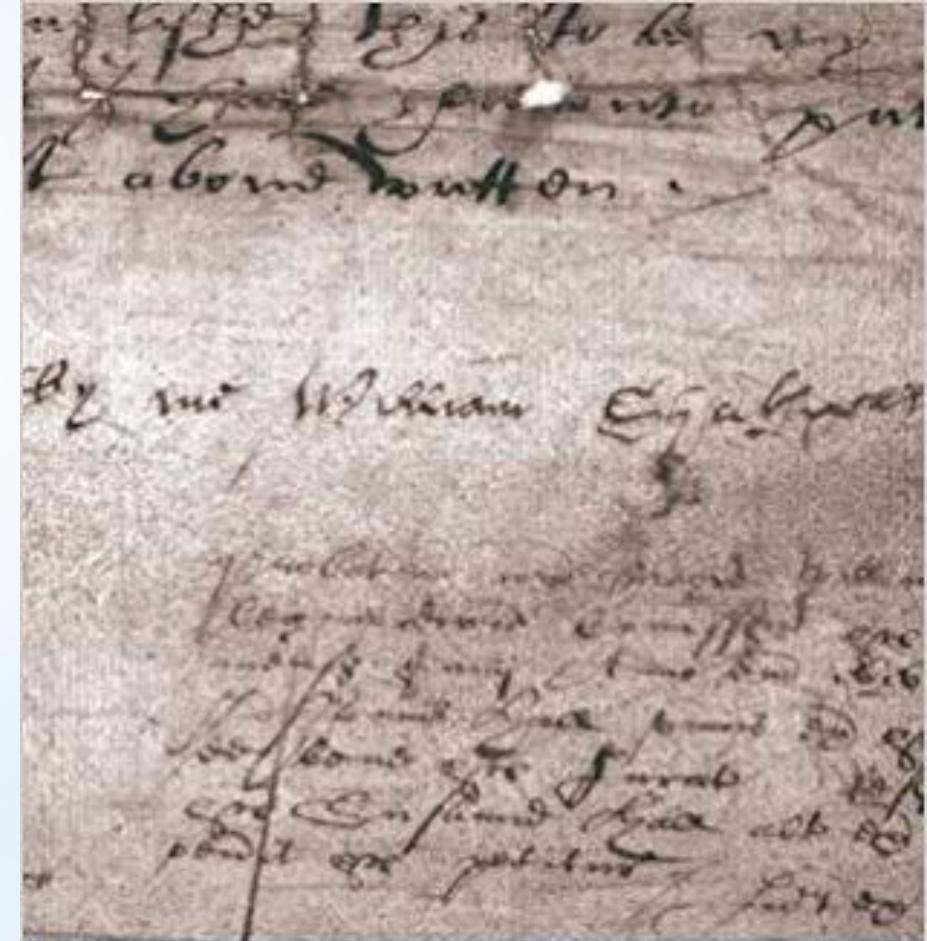
Cosine Delta is the best performing distance (Evert et al. 2017)

FEATURES SELECTION (MFW)



About
1,000-2,000
MFW produce the
best results
(Evert et al. 2017)

THE «ZETA» METHOD



Shakespeare, Computers, and the Mystery of Authorship

EDITED BY
Hugh Craig and Arthur F. Kinney

CAMBRIDGE

WHAT MAKES THESE THREE SENTENCES «SHAKESPEAREAN»?

I meant indeed to pay you with this, which if like an ill venture it come unluckily home, I break, and you, my gentle creditors, lose. Here I promised you I would be, and here I commit my body to your mercies. Bate me some, and I will pay you some, and (as most debtors do) promise you infinitely.

2 Henry IV

But since you have made the days and nights as one,
To wear your gentle limbs in my affairs,
Be bold you do so grow in my requital
As nothing can unroot you.

All's Well that Ends Well

Julius Caesar

This is a sleepy tune. O murd'rous slumber!
Layest thou thy leaden mace upon my boy,
That plays thee music? Gentle knave, good night;
I will not do thee so much wrong to wake thee.

WHAT MAKES THESE THREE SENTENCES «SHAKESPEAREAN»?

I meant indeed to pay you with this, which if like an ill venture it come unluckily home, I break, and you, **my gentle creditors**, lose. Here I promised you I would be, and here I commit my body to your mercies. Bate me some, and I will pay you some, and (as most debtors do) promise you infinitely.

2 Henry IV

But since you have made the days and nights as one,
To wear **your gentle limbs** in my affairs,
Be bold you do so grow in my requital
As nothing can unroot you.

All's Well that Ends Well

Julius Caesar

This is a sleepy tune. O murd'rous slumber!
Layest thou thy leaden mace upon my boy,
That plays thee music? **Gentle knave**, good night;
I will not do thee so much wrong to wake thee.

STYLOMETRY AND SHAKESPEARE

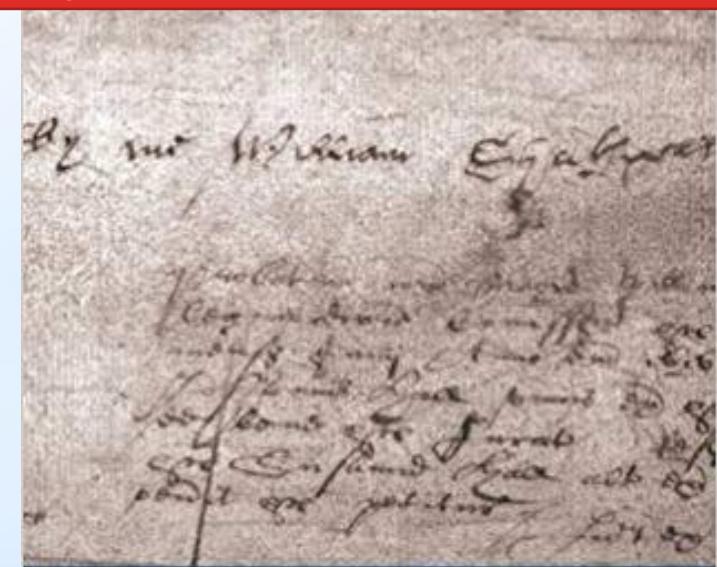
▪ ▪ ▪ Shakespeare's works (as a single string of text)

3,000 words

3,000 words

3,000 words
etc.

How many of the slices contain the word «Gentle»?

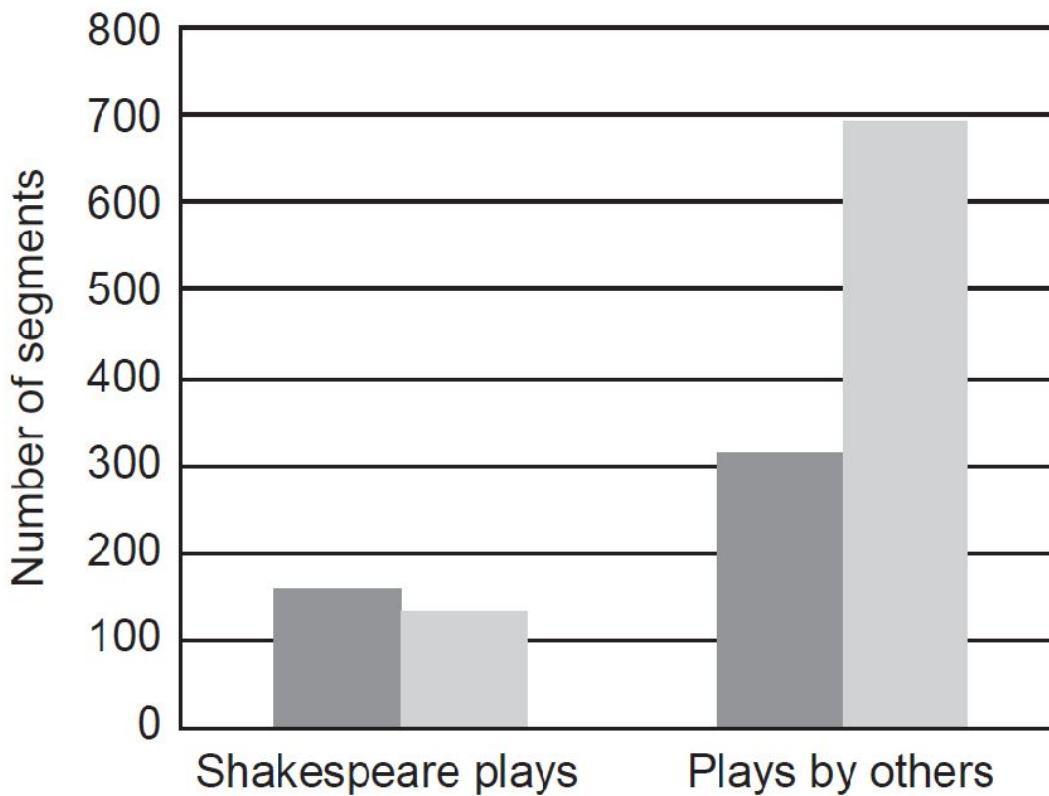


Shakespeare,
Computers, and the
Mystery of Authorship

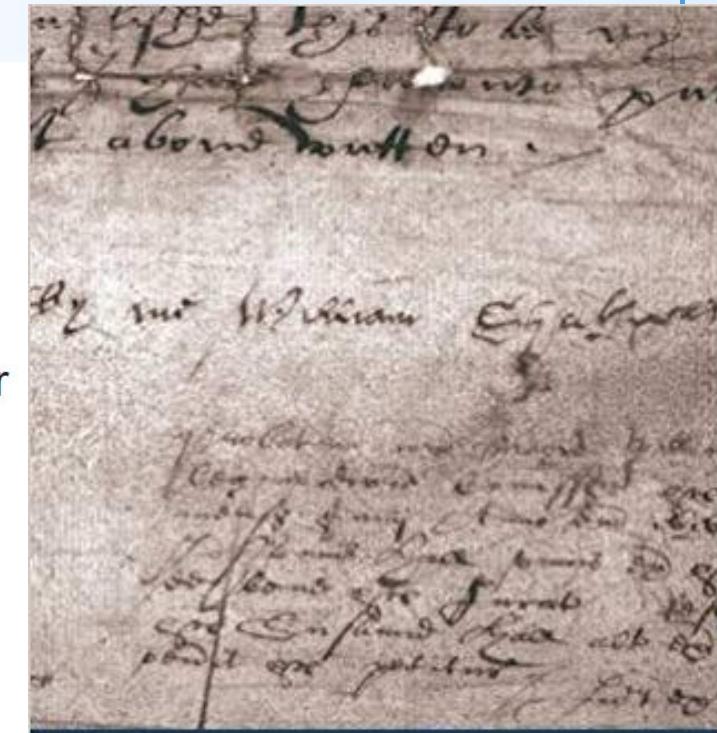
EDITED BY
Hugh Craig and Arthur F. Kinney

CAMBRIDGE

STYLOMETRY AND SHAKESPEARE



- Segments in which 'gentle' appears
- Segments in which 'gentle' does not appear



**Shakespeare,
Computers, and the
Mystery of Authorship**

EDITED BY
Hugh Craig and Arthur F. Kinney

CAMBRIDGE

THE “ZETA” METHOD

Pick up a word:
«gentle» (for example)

Text A



Text B

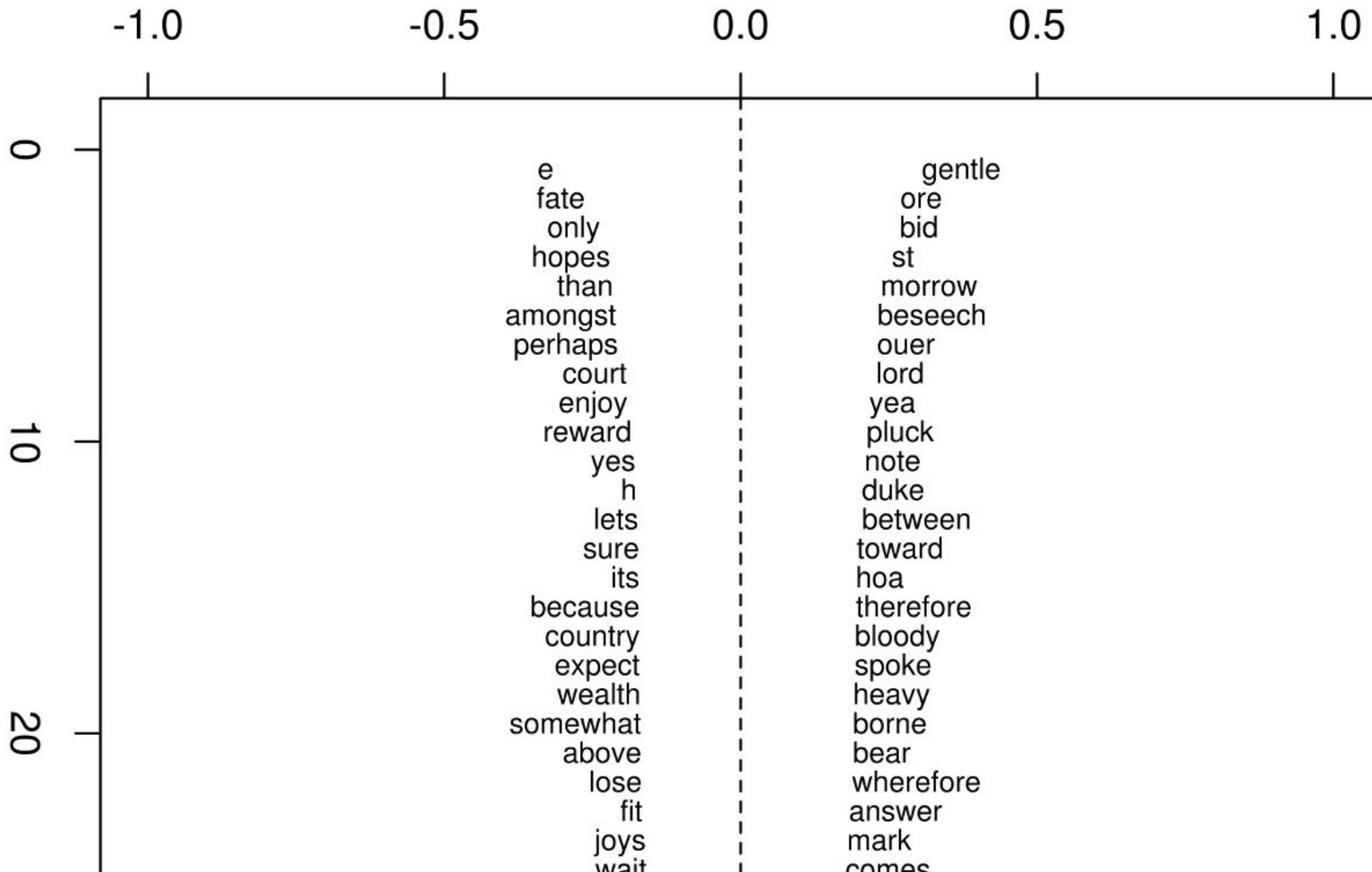


3,000 words 3,000 words ...

...

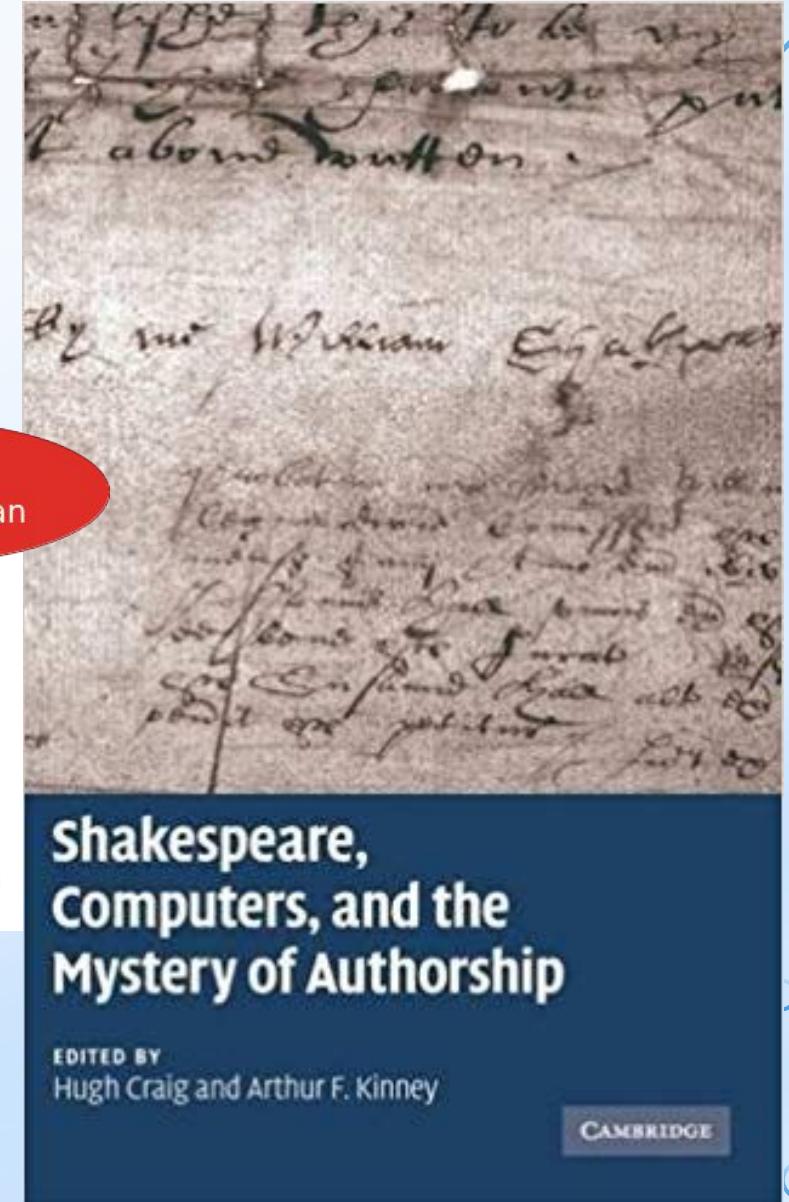
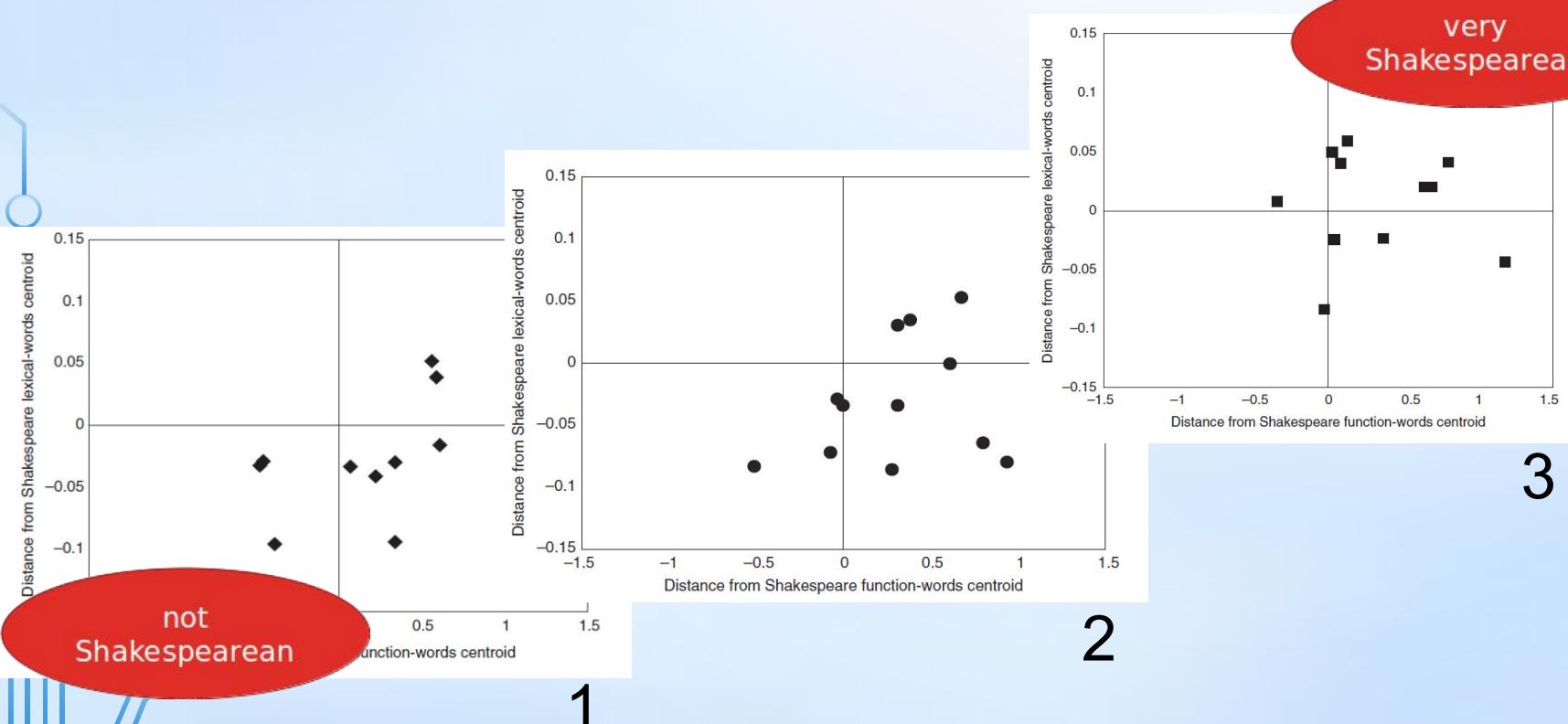
- **Count** in how many slices of the text the word «gentle» appears
- **Calculate** the proportion
Text A: 1 (100%); text B: 0.33 (33%)
- **Subtract** the two values
(so the word «gentle» has Zeta = 0.66 for Text A)
- **Repeat** the operation for all the words in the two texts

Score



APPLICATIONS IN LITERARY STUDIES

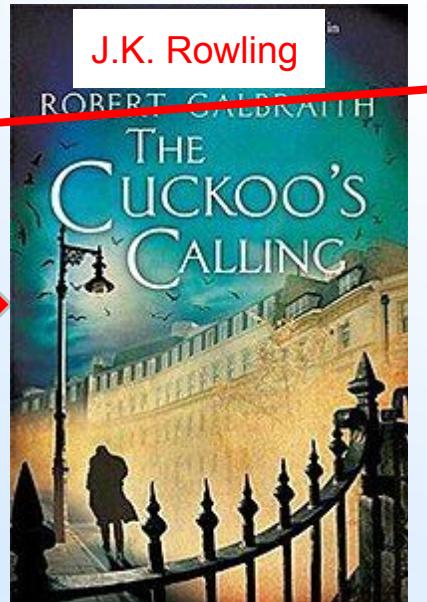
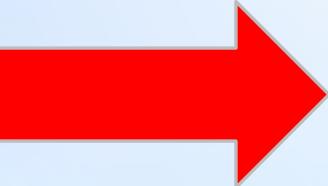
Stylometry and the three parts of *Henry VI*
(Craig and Kinney 2009)



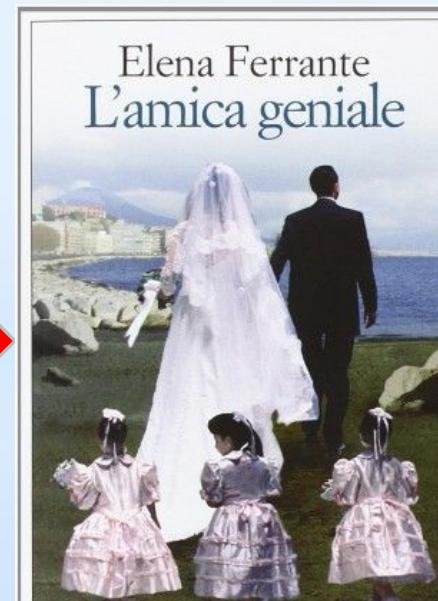
APPLICATIONS IN LITERARY STUDIES



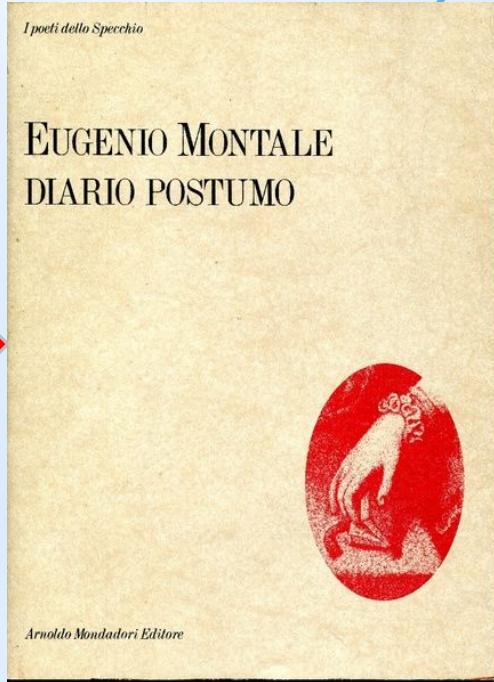
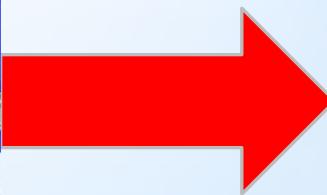
(Patrick Juola)



(Arjuna Tuzzi)

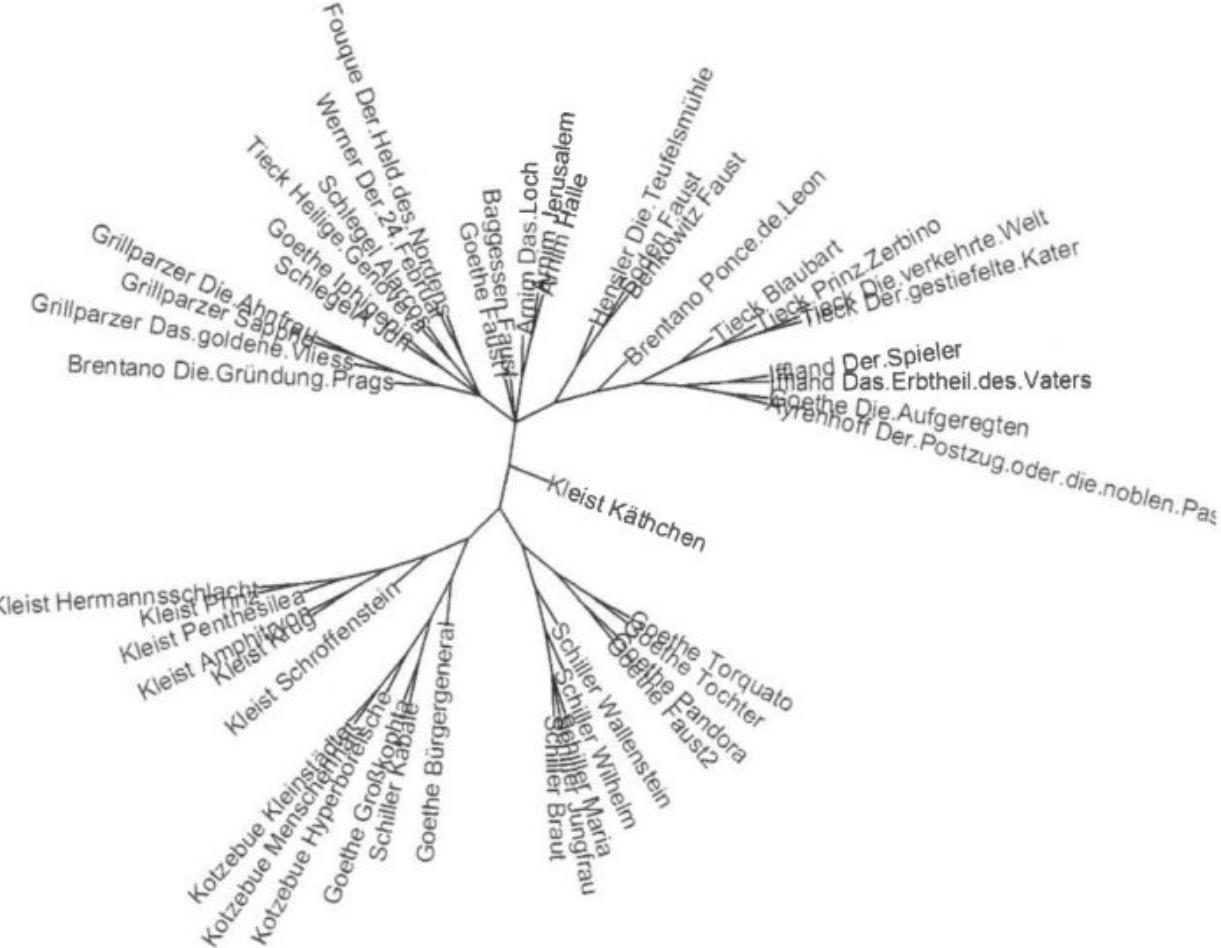


(Paolo Canettieri)



APPLICATIONS IN LITERARY STUDIES

Is Kleist a
classicist or
a romantic?



(Jannidis and
Lauer, 2014)

APPLICATIONS IN LITERARY STUDIES

ON LATE STYLE

MUSIC AND LITERATURE
AGAINST THE GRAIN

EDWARD W. SAID

"These studies . . . buzz with excitement and intelligence and demonstrate what his admirers already knew, the extraordinary range of Said's intellectual interests."
—Frank Kermode, *London Review of Books*



Does Late
Style Exist?

APPLICATIONS IN LITERARY STUDIES

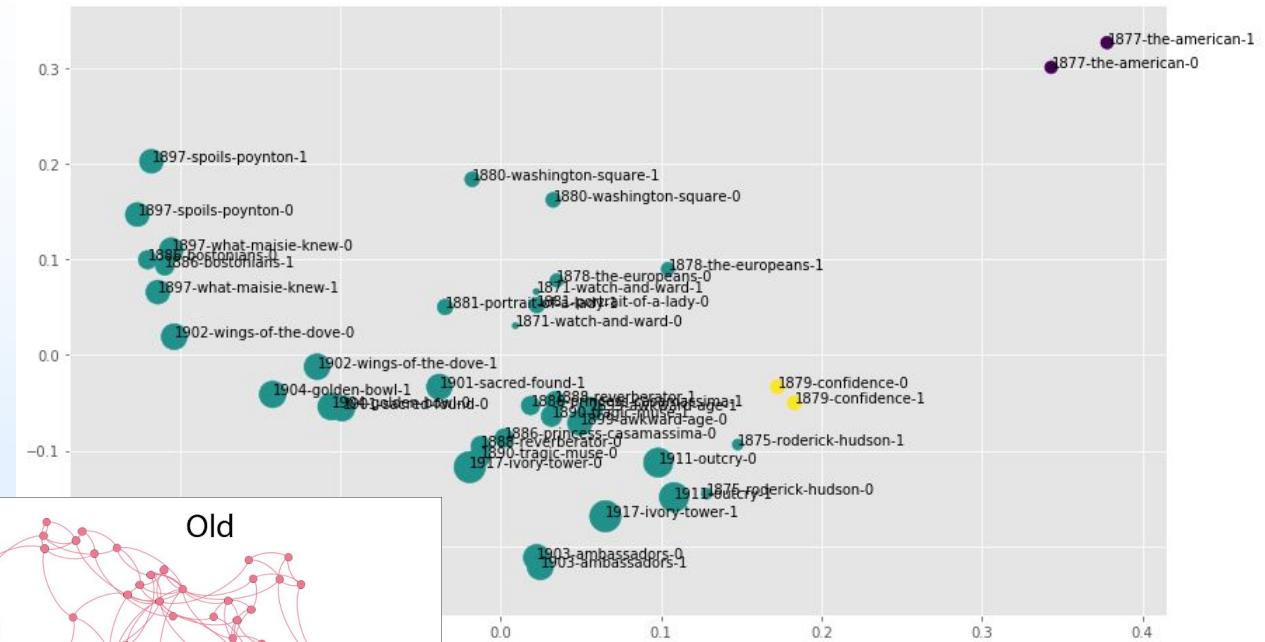
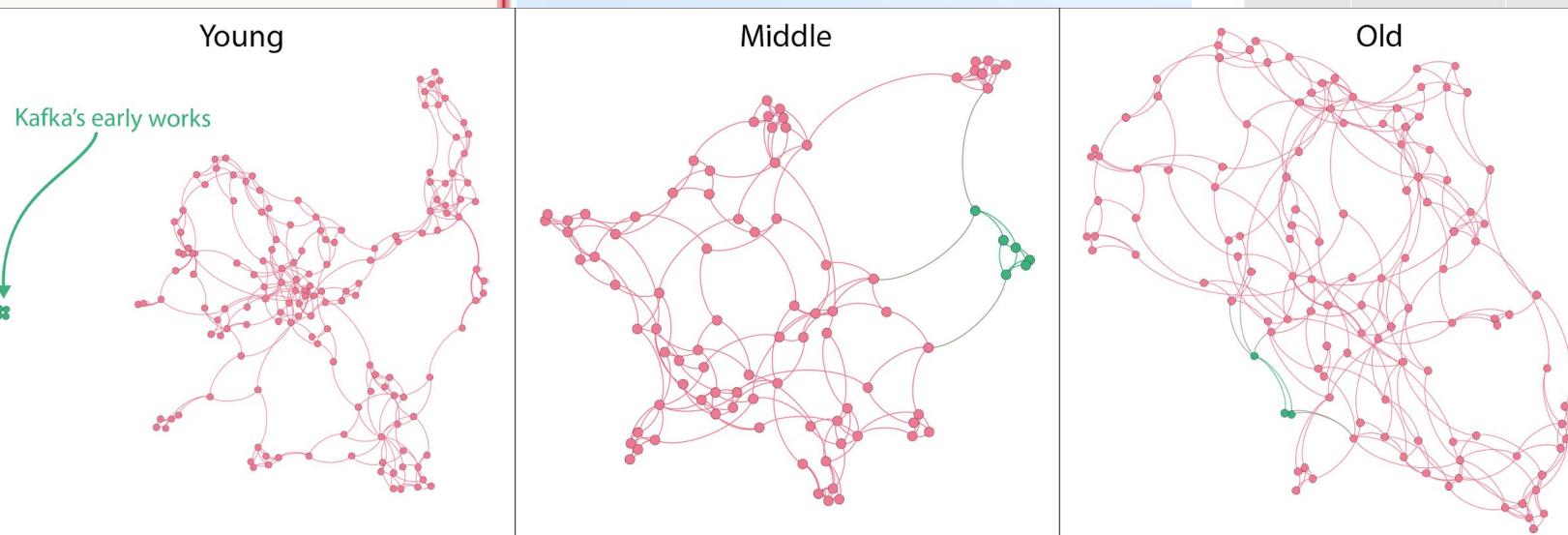
ON LATE
STYLE

MUSIC AND LITERATURE
AGAINST THE GRAIN

EDWARD W. SAID

Does Late
Style Exist?

(Reeve,
2018)



(Rebora and
Salgaro, 2018)

The background of the slide features a subtle, abstract pattern composed of thin blue lines and small circular nodes. These elements create a sense of depth and motion, resembling a network or a stylized landscape. The lines vary in length and thickness, while the nodes are uniform blue circles.

HANDS-ON!!!