Computational stylistics: tools and methods

Maciej Eder

Pedagogical University in Krakow
& Polish Academy of Sciences

29/10/2012

Contents

1. Stylometry: measuring word frequencies (counting words)
2. Multidimensional approaches to capturing authorial style
3. Grammatical structures: do they improve stylometric performance?

Some research questions of stylometry

- What is *common* in the language and what is related to cultural contexts and/or writer's *individuality*?
- What elements of style are affected by literary period, genre, topic?
- What is unconsciously incorporated by the author and reflects his/her education, gender, religious background, social or historical conditions?
- Which features of a written text can betray the person who wrote it *despite* his/her aesthetic, social, or historical conditions?

chapter 1

Intuitively, style is determined by meaningful words

Concordances, attributions, ...

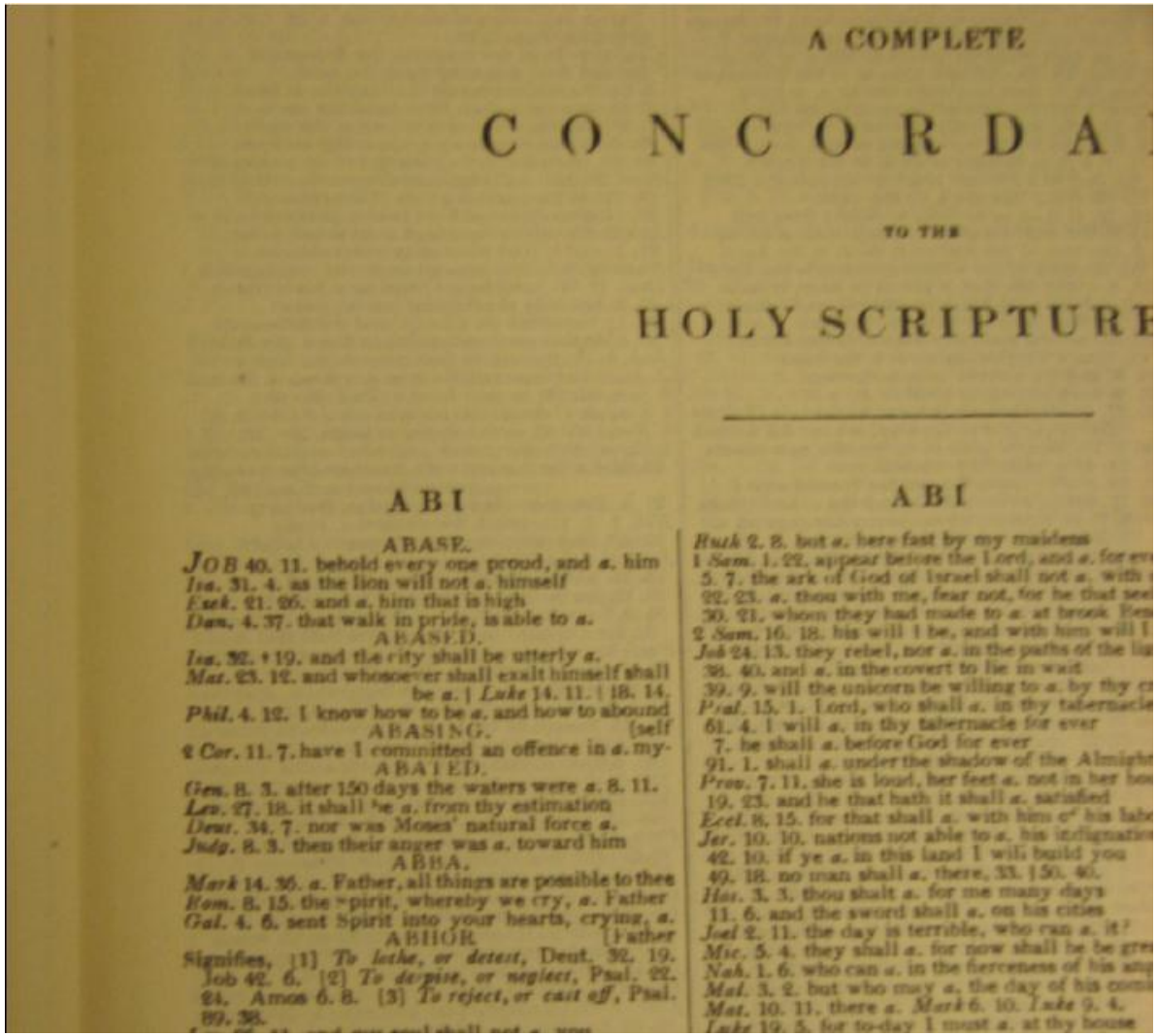
- abbot Hugo de Saint-Cher: first concordance (ca 1230)
- Lorenzo Valla: authorship attribution of *The Donation of Constantine*
- Erasmus Roterodamus: attribution of St. Paul's and Seneca's letters
- Roberto Busa: *Index Tomisticus* (1949–1970)

Concordances, attributions, ...

- abbot Hugo de Saint-Cher: first concordance (ca 1230)
- Lorenzo Valla: authorship attribution of *The Donation of Constantine*
- Erasmus Roterodamus: attribution of St. Paul's and Seneca's letters
- Roberto Busa: *Index Tomisticus* (1949–1970)

Mostly friars and monks! (stylometry requires patience...)

Cruden's Concordance to the *King James Bible* (1737)



A COMPLETE

CONCORDA

TO THE

HOLY SCRIPTUR

ABI

ABASE.

Job 40. 11. behold every one proud, and *a.* him
Isa. 31. 4. as the lion will not *a.* himself
Ezek. 21. 26. and *a.* him that is high
Dan. 4. 37. that walk in pride, is able to *a.*

ABASED.

Isa. 32. †19. and the city shall be utterly *a.*
Mat. 23. 12. and whosoever shall exalt himself shall be *a.* | *Luke* 14. 11. | 18. 14.
Phil. 4. 12. I know how to be *a.* and how to abound

ABASING. [self

2 *Cor.* 11. 7. have I committed an offence in *a.* my-

ABATED.

Gen. 8. 3. after 150 days the waters were *a.* 8. 11.
Lev. 27. 18. it shall be *a.* from thy estimation
Deut. 34. 7. nor was Moses' natural force *a.*
Judg. 8. 3. then their anger was *a.* toward him

ABBA.

Mark 14. 36. *a.* Father, all things are possible to thee
Rom. 8. 15. the Spirit, whereby we cry, *a.* Father
Gal. 4. 6. sent Spirit into your hearts, crying, *a.*

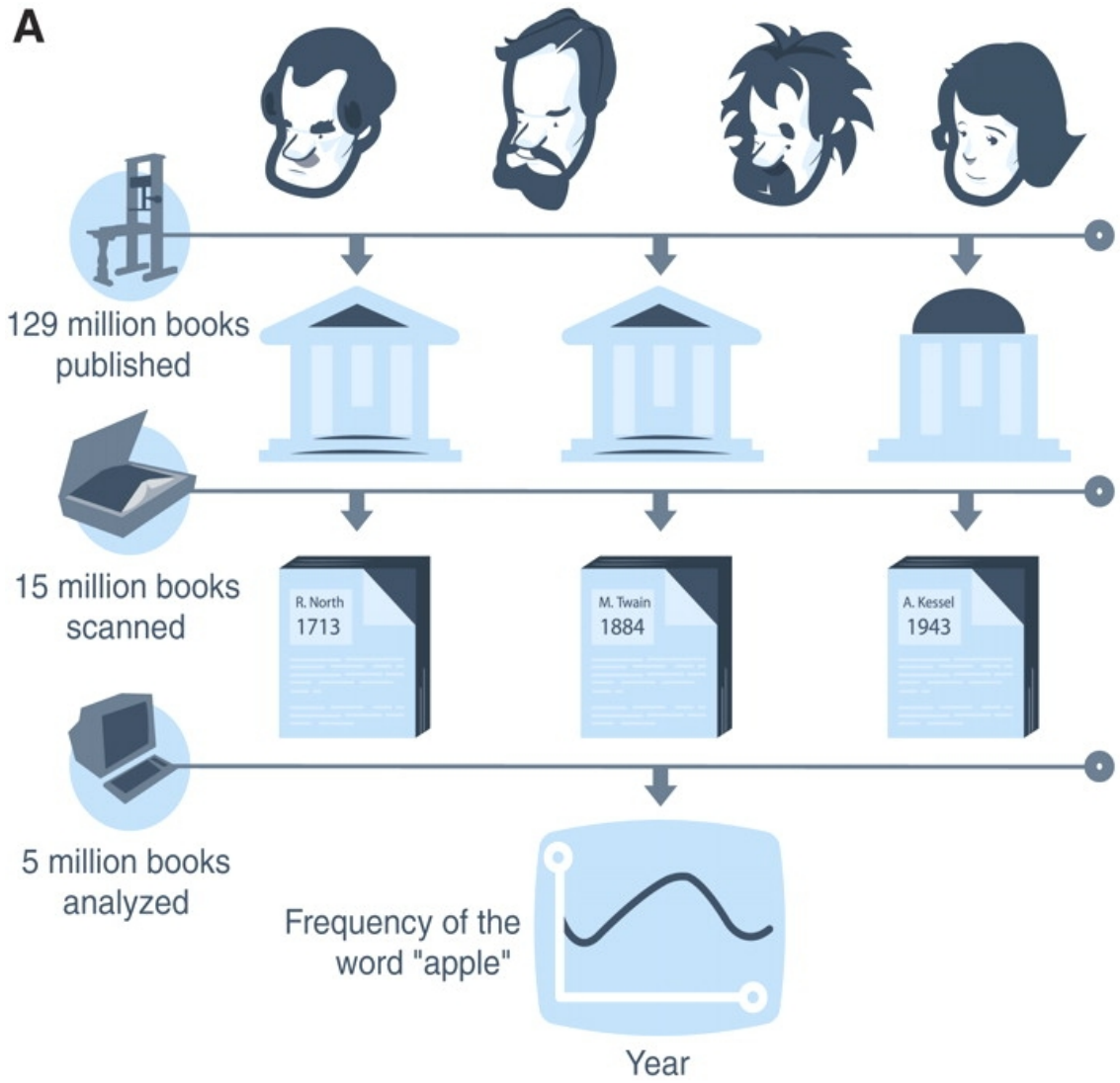
ABHOR [Father

Signifies, [1] *To loathe, or detest*, Deut. 32. 19. Job 42. 6. [2] *To despise, or neglect*, Psal. 22. 24. Amos 6. 8. [3] *To reject, or cast off*, Psal. 89. 38.

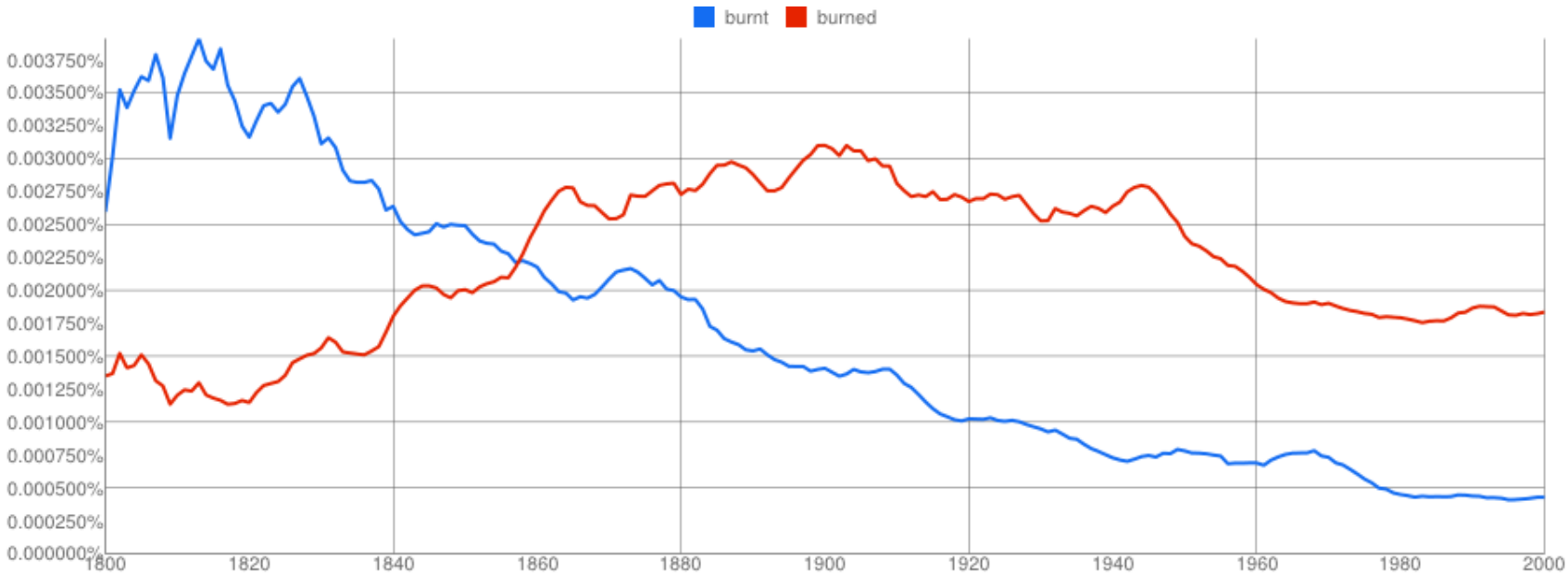
ABI

Ruth 2. 8. but *a.* here fast by my maidens
1 *Sam.* 1. 22. appear before the Lord, and *a.* for ev
5. 7. the ark of God of Israel shall not *a.* with
22. 23. *a.* thou with me, fear not, for he that see
30. 21. whom they had made to *a.* at brook Be
2 *Sam.* 16. 18. his will I be, and with him will I
Job 24. 13. they rebel, not *a.* in the paths of the li
38. 40. and *a.* in the covert to lie in wait
39. 9. will the unicorn be willing to *a.* by thy c
Psal. 15. 1. Lord, who shall *a.* in thy tabernacle
61. 4. I will *a.* in thy tabernacle for ever
7. he shall *a.* before God for ever
91. 1. shall *a.* under the shadow of the Almigh
Prov. 7. 11. she is loud, her feet *a.* not in her ho
19. 23. and he that hath it shall *a.* satisfied
Eccl. 8. 15. for that shall *a.* with him of his labo
Jer. 10. 10. nations not able to *a.* his indignatio
42. 10. if ye *a.* in this land I will build you
49. 18. no man shall *a.* there, 33. | 50. 40.
Hos. 3. 3. thou shalt *a.* for me many days
11. 6. and the sword shall *a.* on his cities
Joel 2. 11. the day is terrible, who can *a.* it?
Mic. 5. 4. they shall *a.* for now shall he be gre
Nah. 1. 6. who can *a.* in the fierceness of his an
Mal. 3. 2. but who may *a.* the day of his com
Mat. 10. 11. there *a.* *Mark* 6. 10. *Luke* 9. 4.
Luke 19. 5. for to-day I must *a.* at thy house

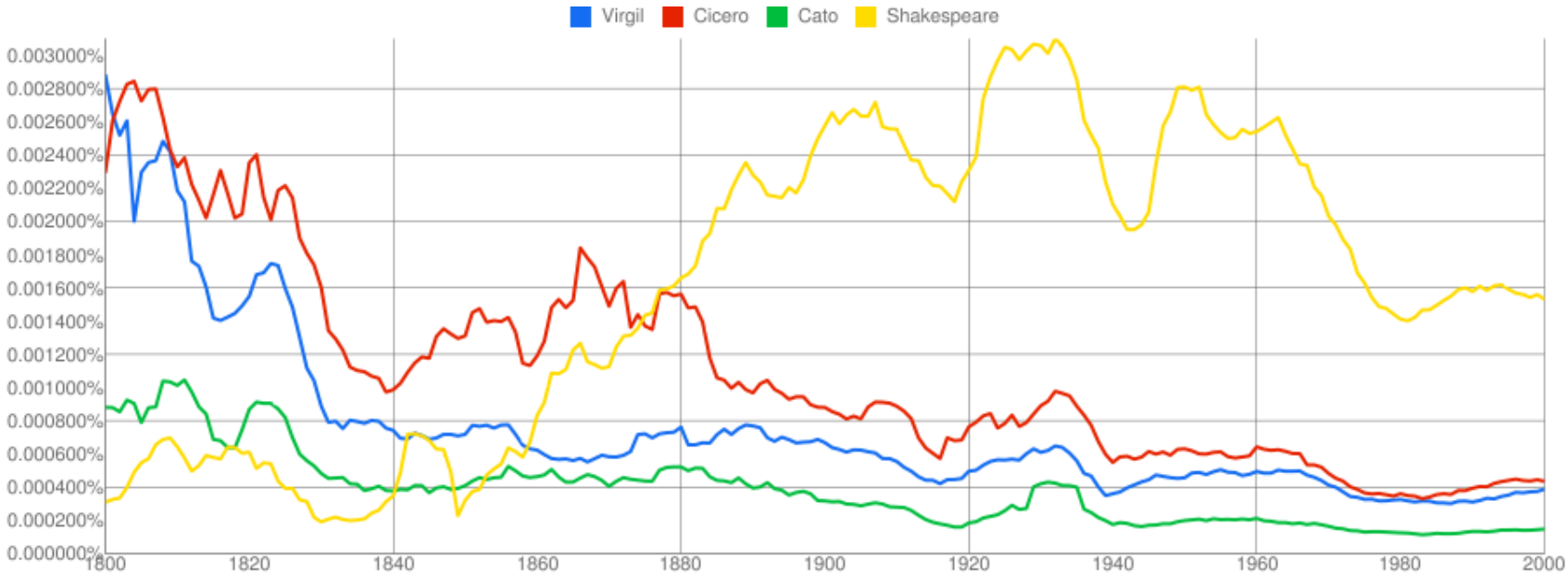
How to read 5 million books? (Google ngram viewer)



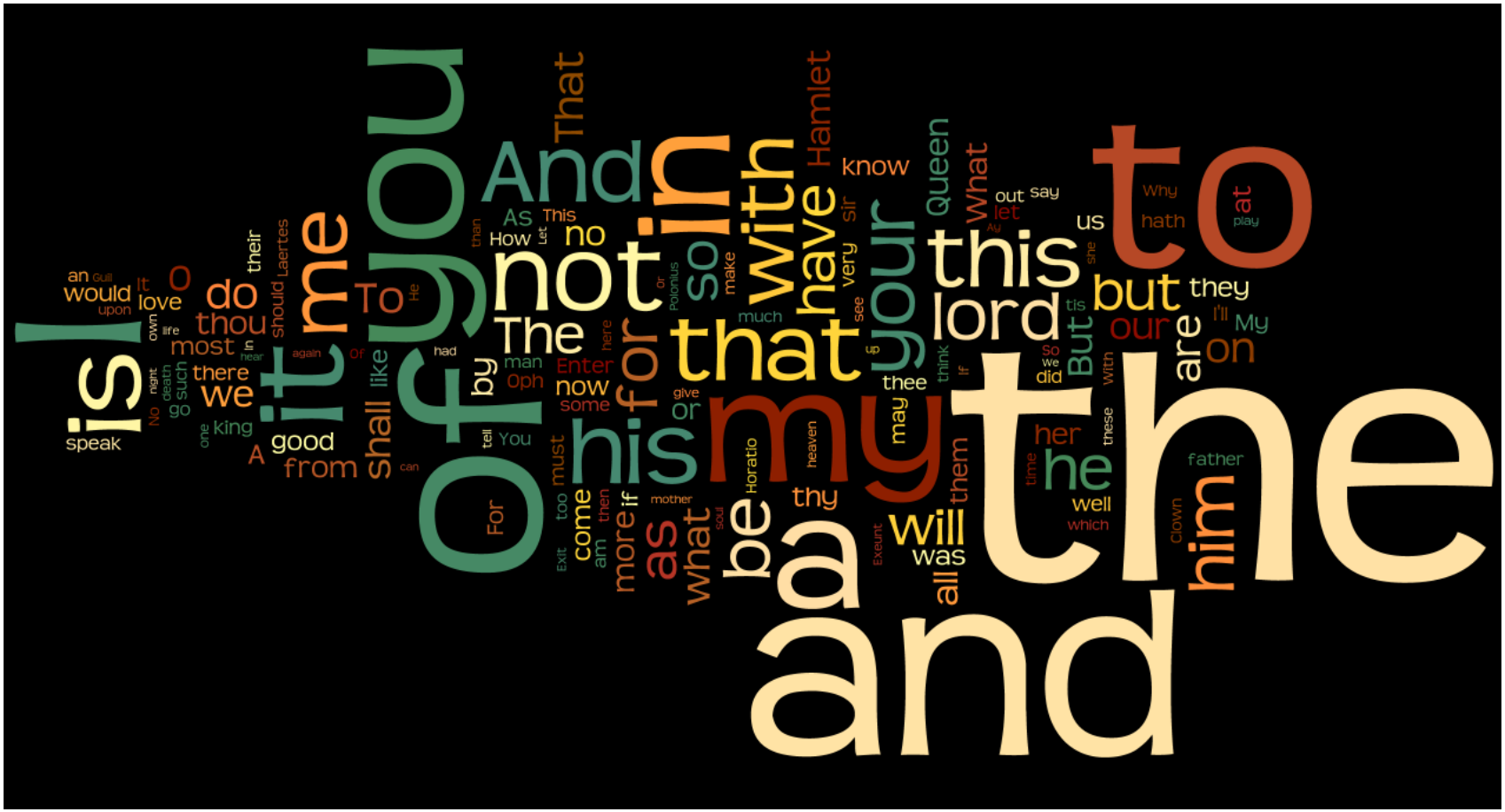
Evolution of language: burnt vs. burned



Decline of classical education



Hamlet: a word cloud (www.wordle.net)

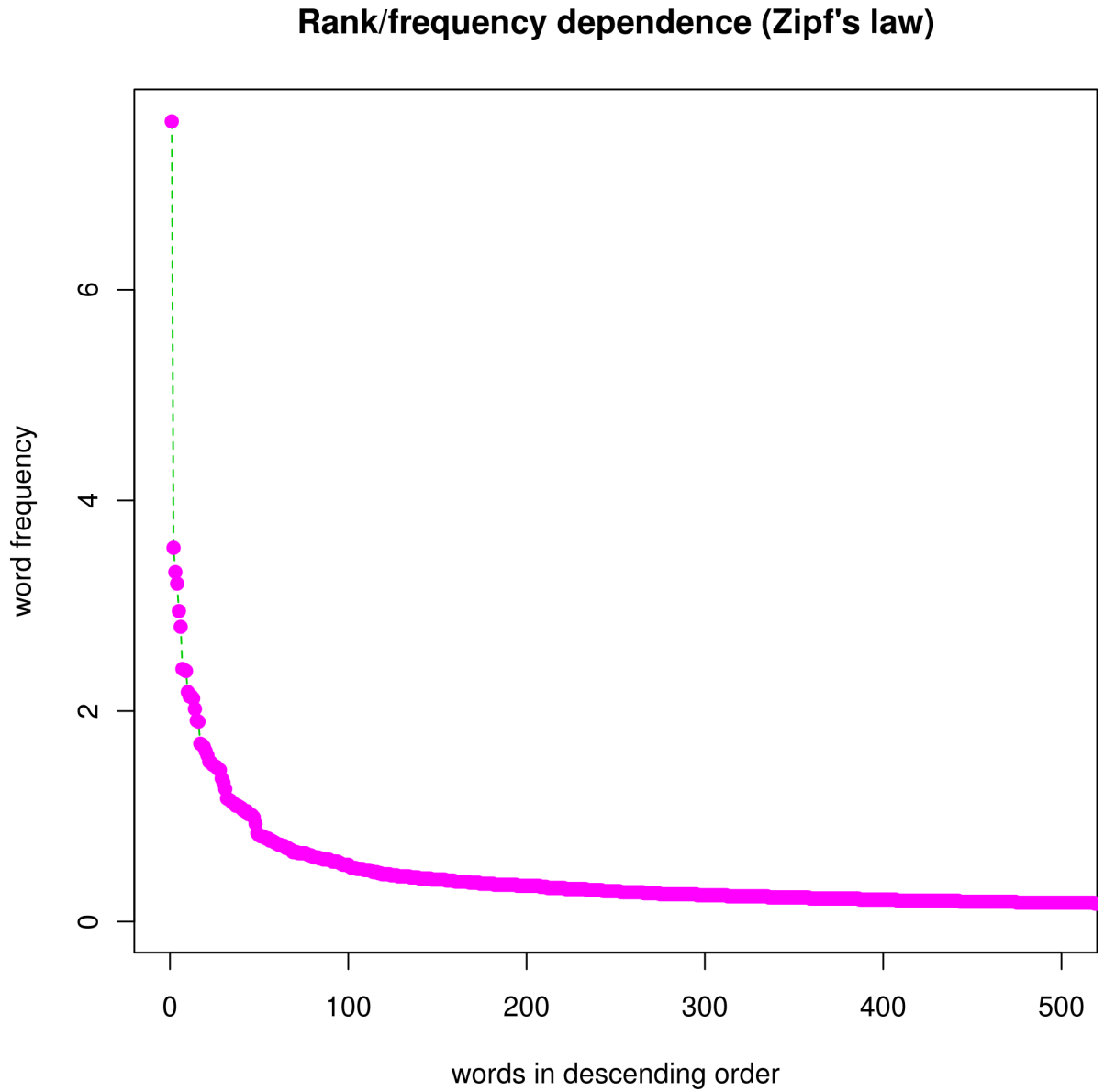


Hamlet: a word cloud (function words excluded)

(1) Sterne, *Sentimental*, (2) Hor. *Ars*, (3) *Bartachom.*

1.	the	5.7384	et	4.69104	δ’	4.3154
2.	i	3.2138	non	1.26173	καὶ	3.09004
3.	and	3.1292	si	1.22938	ἐς	2.18434
4.	of	3.0518	in	1.13232	δὲ	1.97123
5.	to	2.7640	aut	0.905856	ἦν	1.86468
6.	a	2.4061	qui	0.841152	ὁ	1.17208
7.	it	1.9104	ut	0.776448	οὐ	1.06553
8.	in	1.8838	quid	0.744096	ἐπὶ	1.01225
9.	had	1.1583	nec	0.711744	τὸν	0.745871
10.	was	1.1462	est	0.711744	κατὰ	0.745871
11.	as	1.1100	an	0.420576	τε	0.692595
12.	my	1.0834	ad	0.420576	ἄπ’	0.639318
13.	his	1.0108	quae	0.388224	μὲν	0.639318
14.	he	0.9673	ego	0.388224	ἐπ’	0.586042
...	...	...	...	...	...	...

Zipf's law


Rank/frequency dependence (Zipf's law)
word frequency
0
2
4
6
0
100
200
300
400
500
words in descending order

chapter 2

the, *in*, *of* , *or*, *I*, *is*, ...
or the mystery of bare function words

Wincenty Lutosławski (1863–1954)

John Burrows

Non-traditional authorship attribution

Given:

- a text of uncertain or anonymous authorship, and
- a comparison corpus of texts by known authors,

is it possible to find the 'nearest neighbor' among the available candidates?

What should be analyzed? *What* is the style?

- An idea of stylistic fingerprint
  - indiscoverable with the naked eye
  - beyond authorial control
  - resistant to imitation, plagiarism and parody
  - popular solution: usage of function words
- Is the style is determined by an individual? (cf. human DNA code, fingerprint, or patterns in one's iris)

Stylistic fingerprint

'If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting'

(Lutosławski 1897: 66)

Stylistic fingerprint

'If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting'

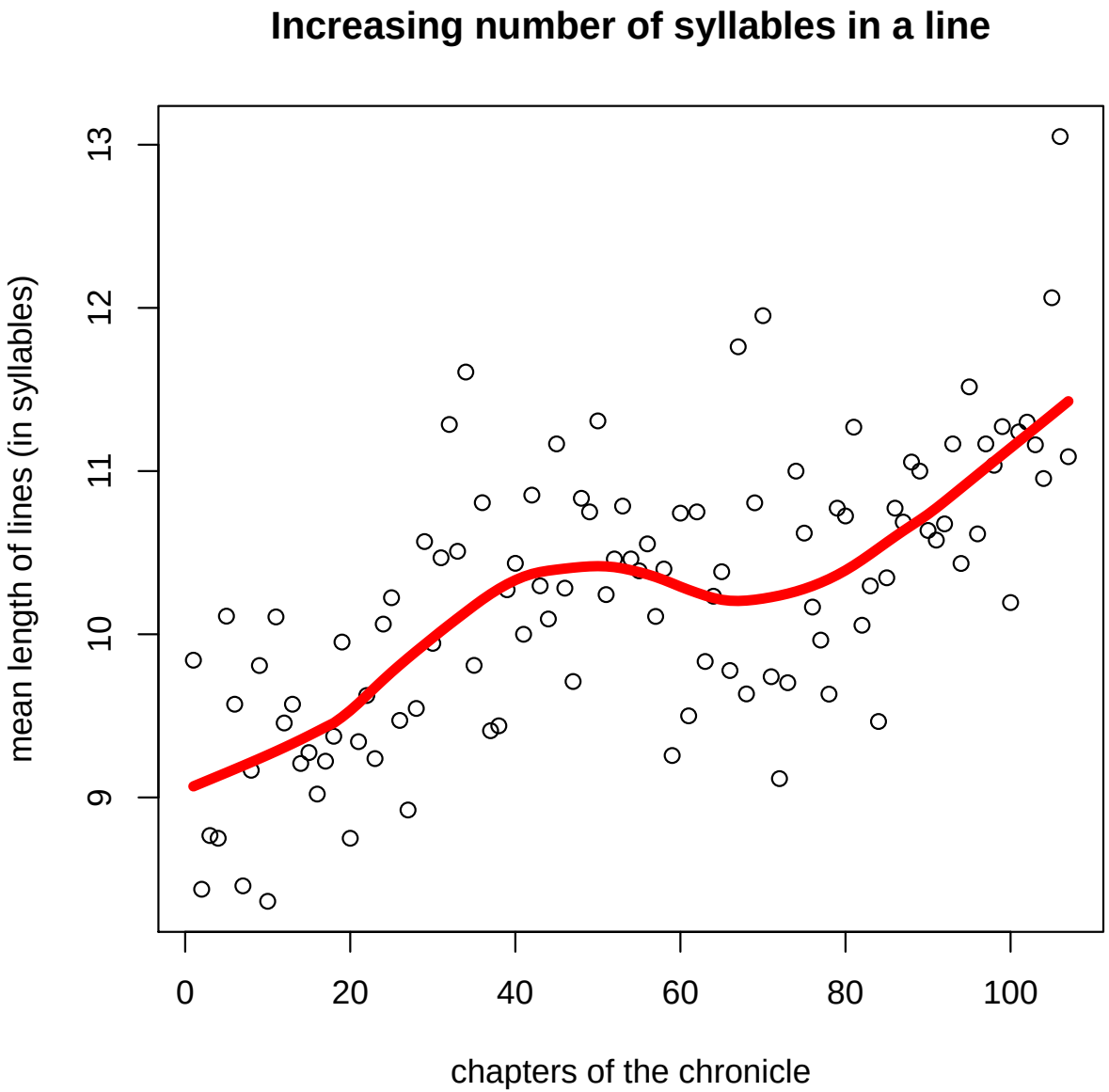
(Lutosławski 1897: 66)

However, nowadays we rather seek for some statistically significant regularities rather than for a determined uniqueness in style

How to measure differencies between texts?

- One-dimensional methods
  - indexes of lexical density
  - mean word length, mean sentence length
  - degree of rhythmicity
  - ...
- Multidimensional methods
  - several features (dozens, or hundreds) measured at once
  - they aggregate weak discriminative strength of single features

Dalimil's Chonicle – stylistic change



Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

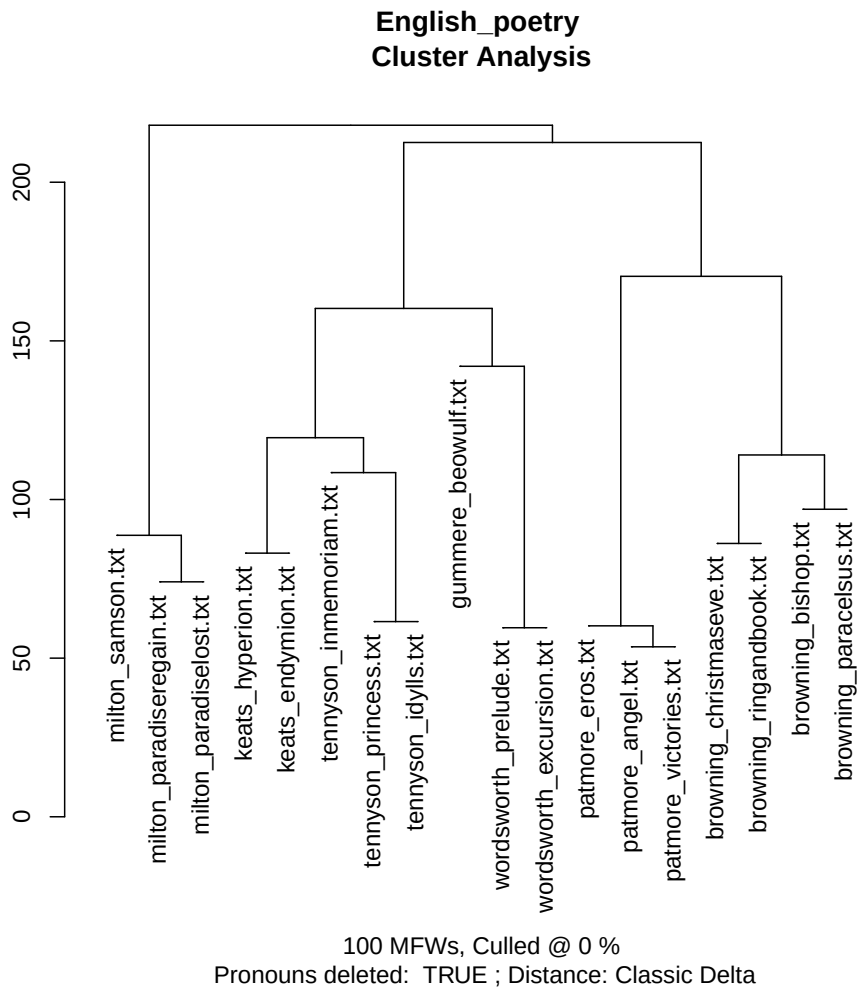
. . . . ?

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

. . . . ? . . . . . ???

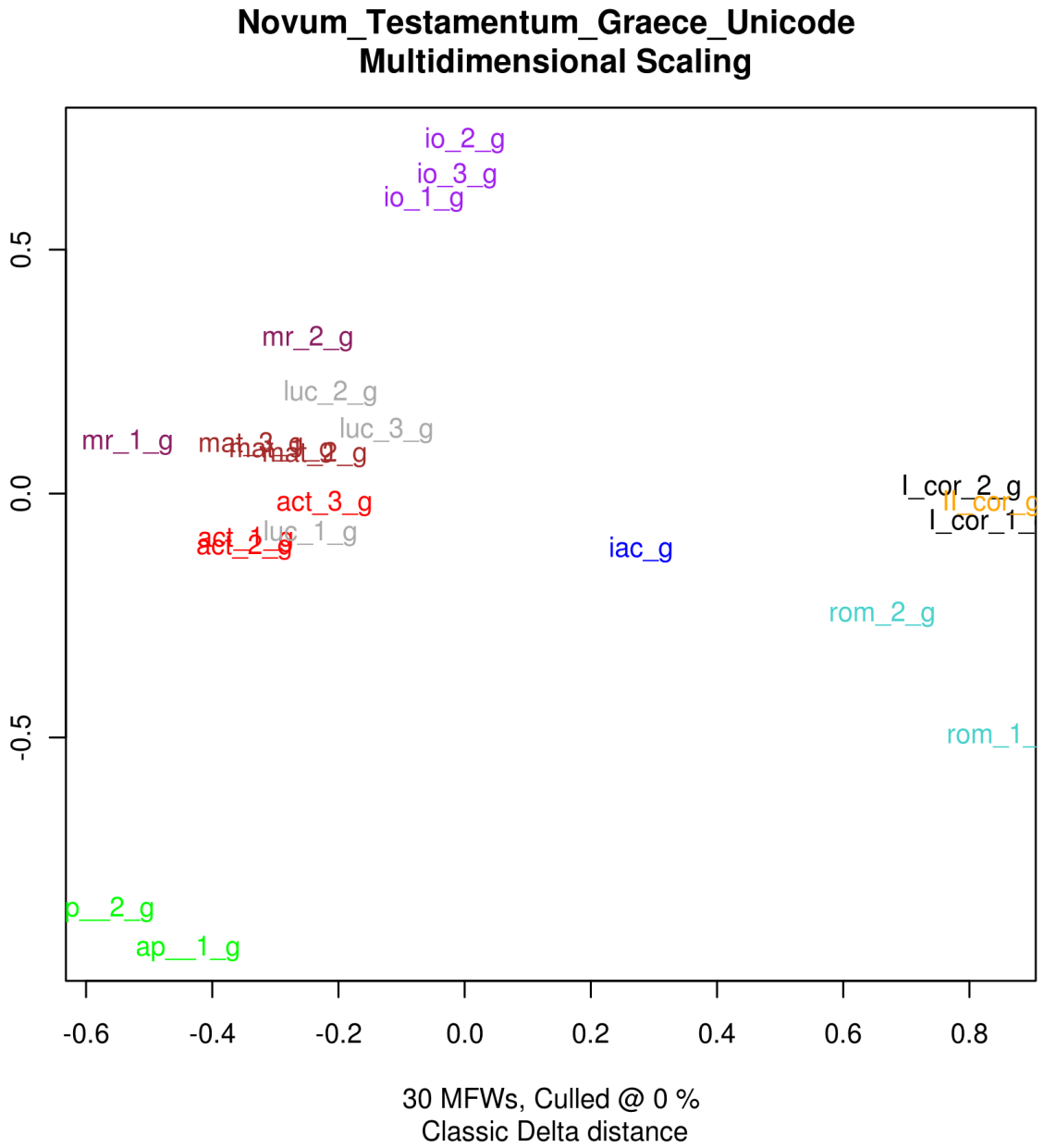
Explanatory approach: Cluster Analysis



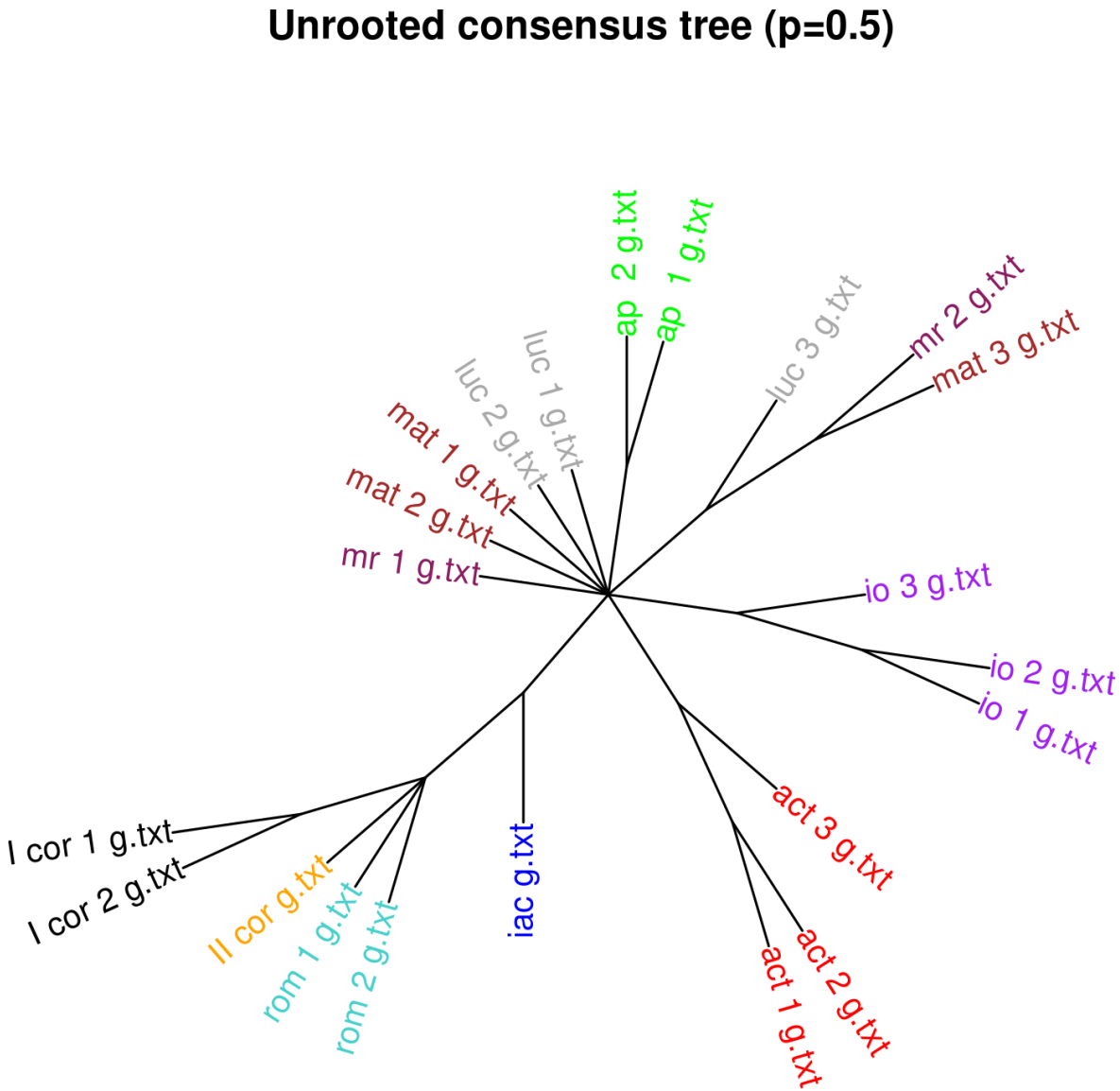
Greek New Testament and its authors

- Synoptic Gospels – strong similarities, numerous identical sentences, c. 50% of common material
  - Mark: written c. 68–73
  - Matthew: c. 70–100
  - Luke: c. 80–100
- John: c. 90–100 – gospel developed from a Johannine circle
- Acts: c. 60–64 – traditionally ascribed to St. Luke
- Epistles – Paul, James, John, Peter, Jude
- Revelation – traditionally ascribed to St. John, but also to John the Presbyter (Dionysius of Alexandria, 3rd cent.)

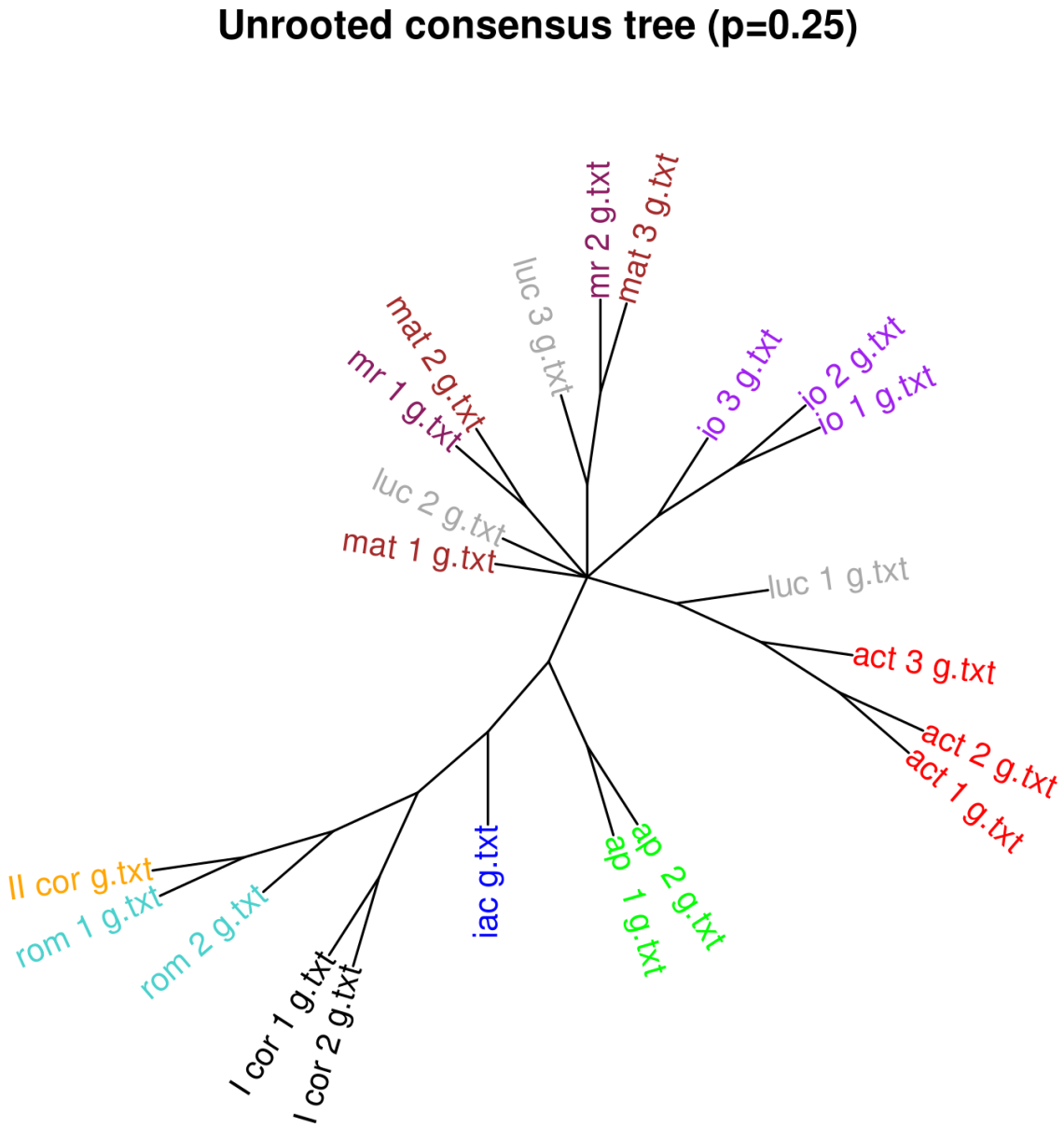
Greek New Testament: textual similarities



Different method, similar results



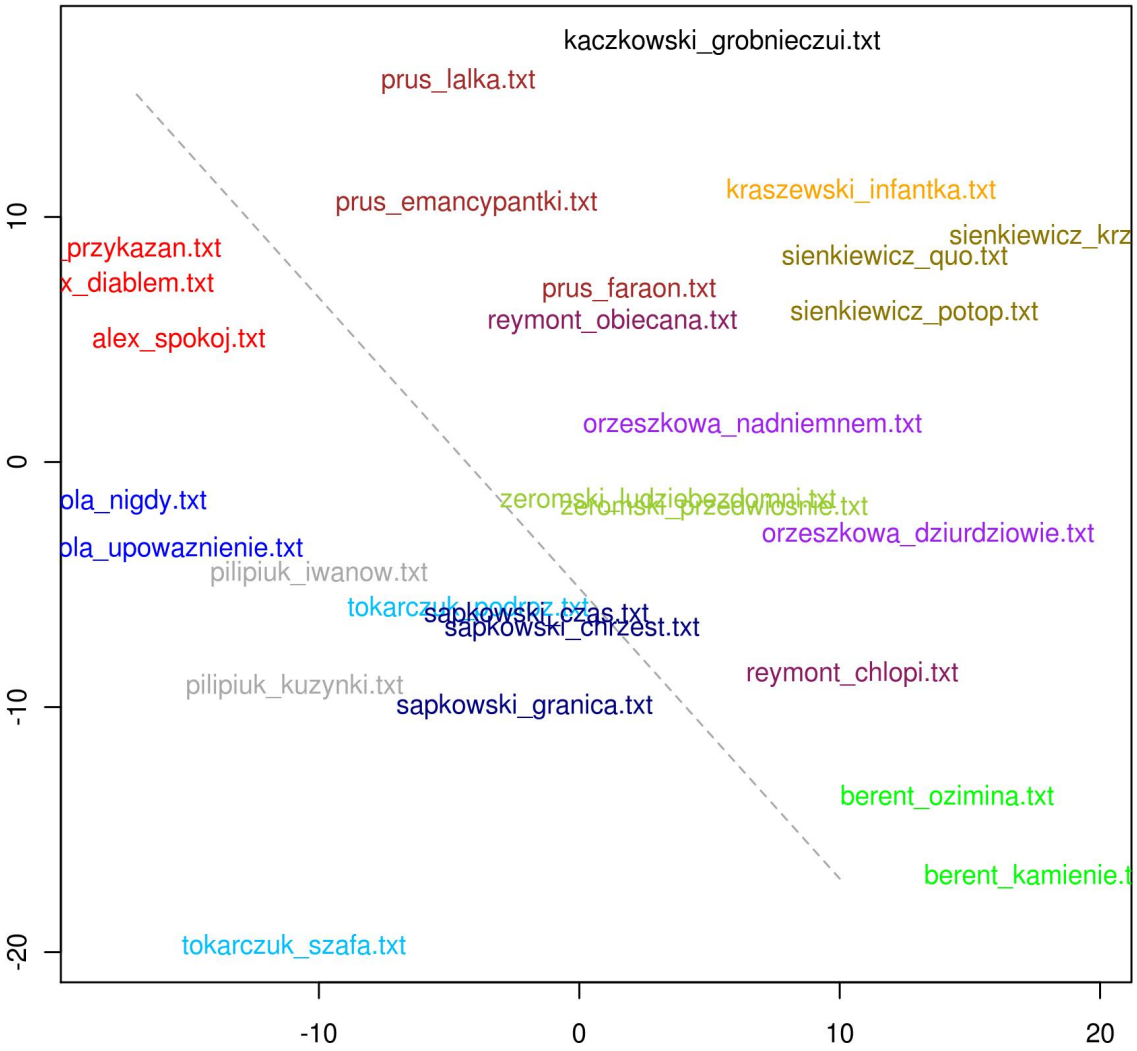
Another approach (less rigorous)



Applications of multidimensional methods

- authorship attribution
- genre recognition
- gender recognition
- chronology
- ...

A break in style after World War II



chapter 3

How far can we go without any words?

Parts-of-speech tagging: grammatical labels

My father was a clergyman.

PRP$	NN	VBD	DT	NN
(poss. pron.)	(noun)	(verb)	(determiner)	(noun)

Different output formats, similar results

Stanford POS Tagger

My_PRP$ father_NN was_VBD a_DT clergyman_NN ...

TreeTagger

My	PP$	my
father	NN	father
was	VBD	be
a	DT	a
clergyman	NN	clergyman
⋮	⋮	⋮

What about getting rid of original words?

It_PRP was_VBD by_IN your_PRP$ desire_NN that_IN I_PRP first_RB thought_VBD of_IN such_JJ a_DT composition_NN ._. So_RB many_JJ years_NNS have_VBP since_IN past_NN ,_, that_IN you_PRP may_MD have_VB ,_, perhaps_RB ,_, forgotten_VBN this_DT circumstance_NN :_: but_CC your_PRP$ desires_NNS are_VBP to_TO me_PRP in_IN the_DT nature_NN of_IN commands_NNS ;_: and_CC the_DT impression_NN of_IN them_PRP is_VBZ never_RB to_TO be_VB erased_VBN from_IN my_PRP$ memory_NN ._. Again_RB ,_, Sir_NNP ,_, without_IN your_PRP$ assistance_NN this_DT history_NN had_VBD never_RB been_VBN completed_VBN ._. Be_VB not_RB startled_VBN at_IN the_DT assertion_NN ._. I_PRP do_VBP not_RB intend_VB to_TO draw_VB on_IN you_PRP the_DT suspicion_NN of_IN being_VBG a_DT romance_NN writer_NN I_PRP mean_VBP no_DT more_JJR than_IN that_IN ...

What about getting rid of original words?

It_PRP was_VBD by_IN your_PRP$ desire_NN that_IN I_PRP first_RB thought_VBD of_IN such_JJ a_DT composition_NN ._. So_RB many_JJ years_NNS have_VBP since_IN past_NN ,_, that_IN you_PRP may_MD have_VB ,_, perhaps_RB ,_, forgotten_VBN this_DT circumstance_NN :_: but_CC your_PRP$ desires_NNS are_VBP to_TO me_PRP in_IN the_DT nature_NN of_IN commands_NNS ;_: and_CC the_DT impression_NN of_IN them_PRP is_VBZ never_RB to_TO be_VB erased_VBN from_IN my_PRP$ memory_NN ._. Again_RB ,_, Sir_NNP ,_, without_IN your_PRP$ assistance_NN this_DT history_NN had_VBD never_RB been_VBN completed_VBN ._. Be_VB not_RB startled_VBN at_IN the_DT assertion_NN ._. I_PRP do_VBP not_RB intend_VB to_TO draw_VB on_IN you_PRP the_DT suspicion_NN of_IN being_VBG a_DT romance_NN writer_NN I_PRP mean_VBP no_DT more_JJR than_IN that_IN ...

Thus, instead of words...

It was by your desire that I first thought of such a composition. So many years have since past, that you may have, perhaps, forgotten this circumstance: but your desires are to me in the nature of commands; and the impression of them is never to be erased from my memory. Again, Sir, without your assistance this history had never been completed. Be not startled at the assertion. I do not intend to draw on you the suspicion of being a romance writer. I mean no more than that I partly owe to you my existence during great part of the time which I have employed in composing it: another matter which it may be necessary to remind you of; since there are certain actions of which you are apt to be extremely forgetful; but of these I hope I shall always have a better memory than yourself. Lastly, It is owing to you that the history appears what it now is. If there be in this work, as some have been pleased to say, a stronger picture of a truly benevolent mind than is to be found in any other, who that knows you, and a particular ...

... we can use tags as a 'normal' running text

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . VB RB VBN IN DT NN . PRP VBP RB VB TO VB IN PRP DT NN IN VBG DT NN NN . PRP VBP DT JJR IN IN PRP RB VBP TO PRP PRP$ NN IN JJ NN IN DT NN WDT PRP VBP VBN IN VBG PRP : DT NN WDT PRP MD VB JJ TO VB PRP IN : IN EX VBP JJ NNS IN WDT PRP VBP JJ TO VB RB JJ : CC IN DT PRP VBP PRP MD RB VB DT JJR NN IN PRP . RB , PRP VBZ VBG TO PRP IN DT NN VBZ WP PRP RB VBZ . IN EX VB IN DT NN , IN DT VBP VBN VBN TO VB , DT JJR NN IN DT RB JJ NN IN VBZ TO VB VBN IN DT JJ , WP WDT VBZ PRP , CC DT JJ NN IN NNS , MD VB NN IN NN NN VBN VBN . IN EX ...

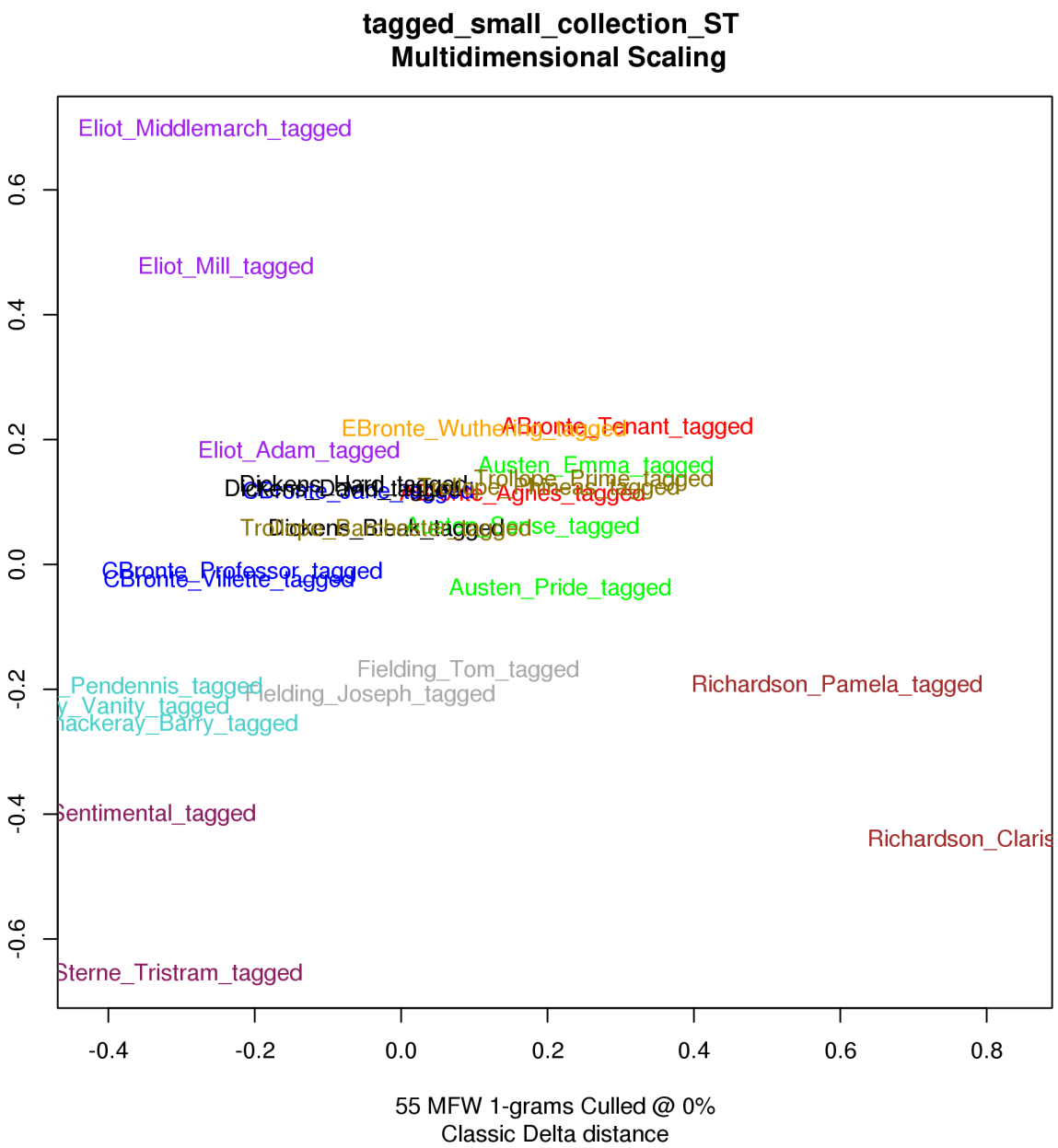
Another example: a Latin text...

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una, pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. ...

... and its grammatical representation.

N-nom ESSE-IND PRON V-PTC-nom PREP N-acc ADJ-NUM REL ADJ-NUM V-IND N-nom PRON ADJ ADJ-NUM REL DET N-nom ADJ POSS N-nom V-IND SENT DIMOS PRON N-nom N-abl N-abl PREP PRON V-IND SENT N-acc PREP ADJ-abl N-nom N-acc PREP N-abl N-nom CC N-nom V-IND SENT DIMOS PRON ADJ-SUP ESSE-IND N-nom ADV REL PREP N-abl CC N-abl N-gen ADV V-IND ADJ PREP DET N-nom ADV V-IND CC DET REL PREP V-GED N-acc V-IND V-IND ADJ ESSE-IND N-abl REL PREP N-acc V-IND V-IND ADV N-acc V-IND SENT REL PREP N-abl N-nom ADV ADJ N-acc N-abl V-IND REL ADV ADJ-abl N-abl PREP N-abl V-IND PREP CC POSS N-abl DET V-IND CC DET PREP DET N-abl N-acc V-IND SENT DET ADJ-NUM N-nom REL N-acc V-INF V-PTC-nom ESSE-IND N-nom V-IND PREP N-abl N-abl V-IND N-nom N-abl N-abl N-abl N-gen V-IND CC PREP N-abl CC N-abl N-acc N-acc V-IND PREP N-acc SENT N-nom PREP ...

POS-tags analysis: disappointing



n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN VBD_IN_PRP$

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN NN_IN_PRP

n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

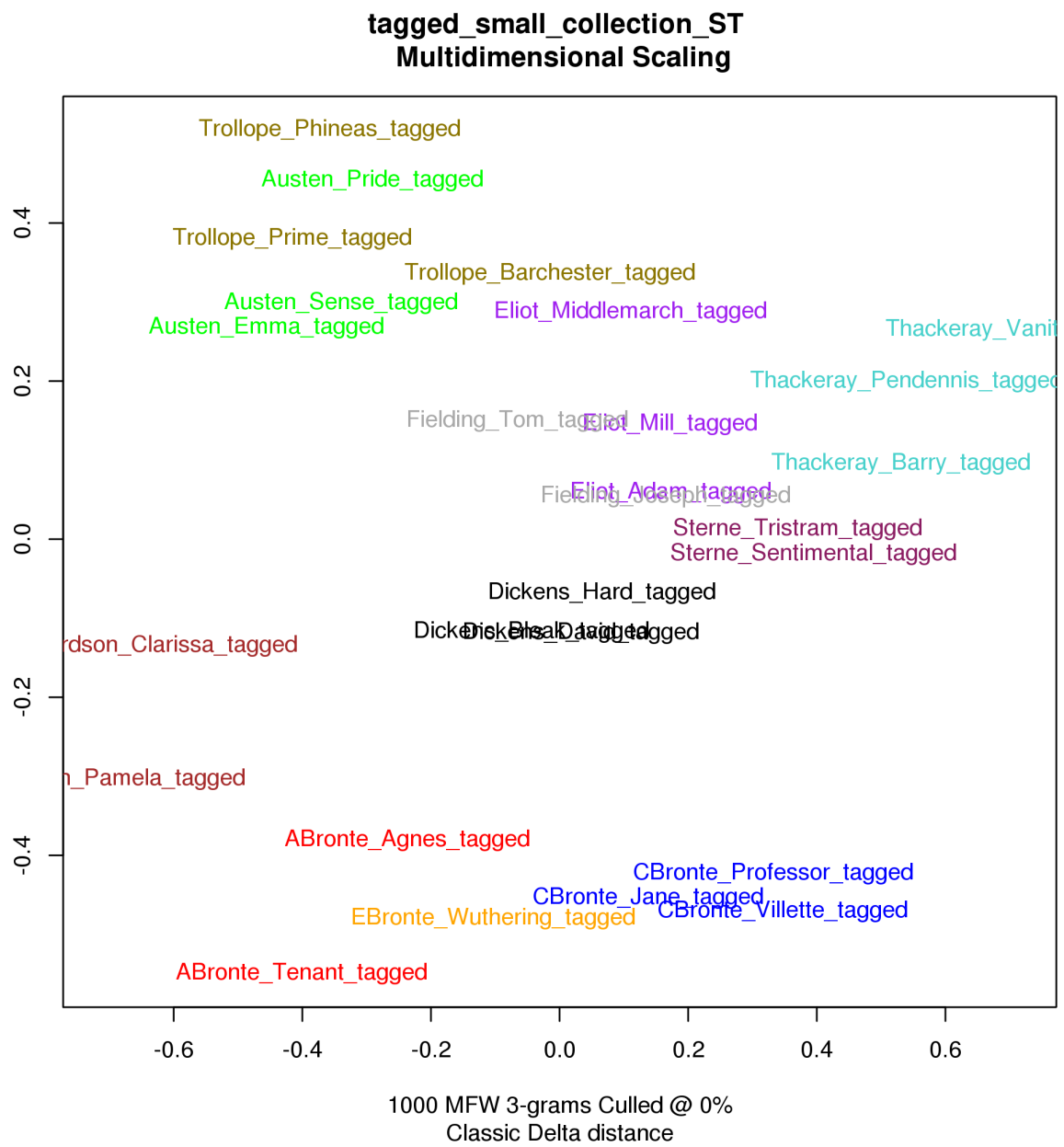
PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN NN_IN_PRP IN_PRP_RB

n-grams of POS tags

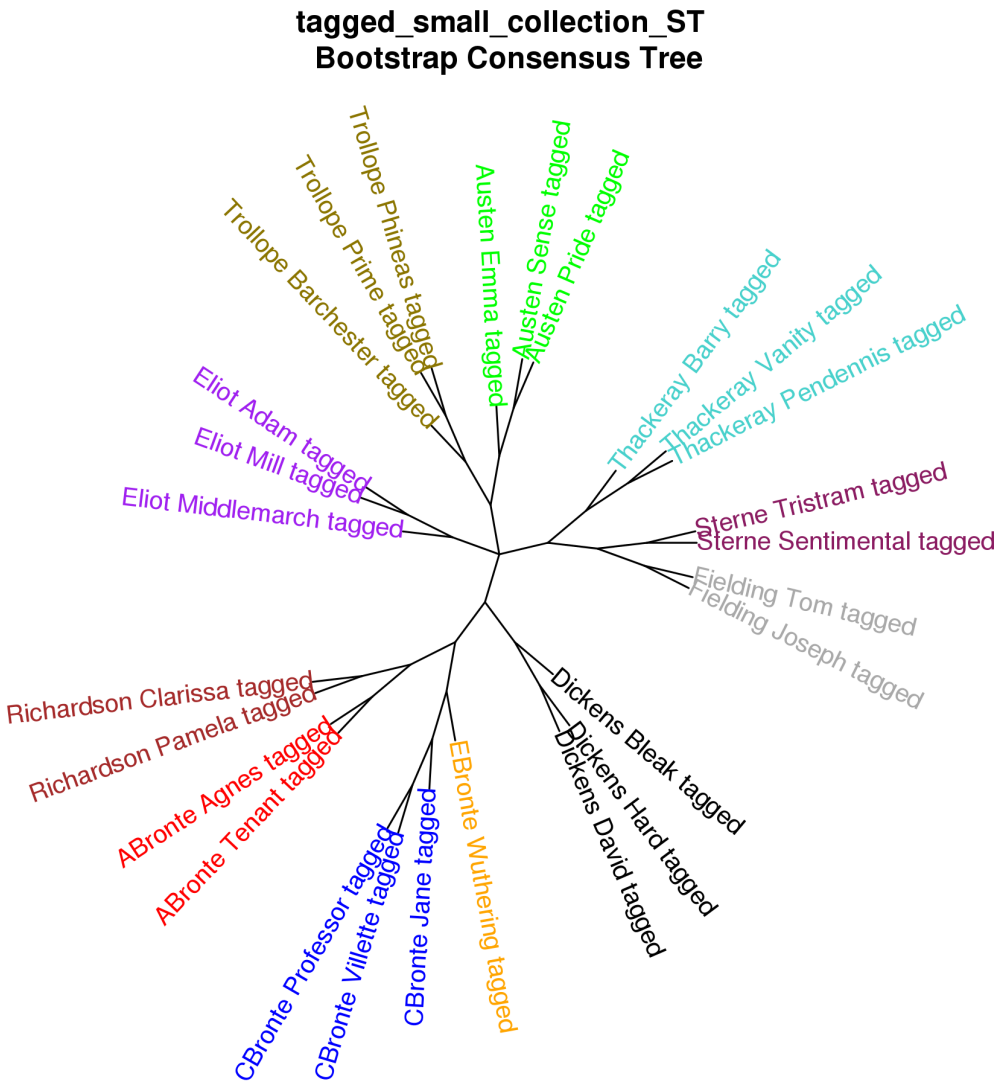
PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN NN_IN_PRP IN_PRP_RB ...

3-grams of POS tags: surprisingly accurate



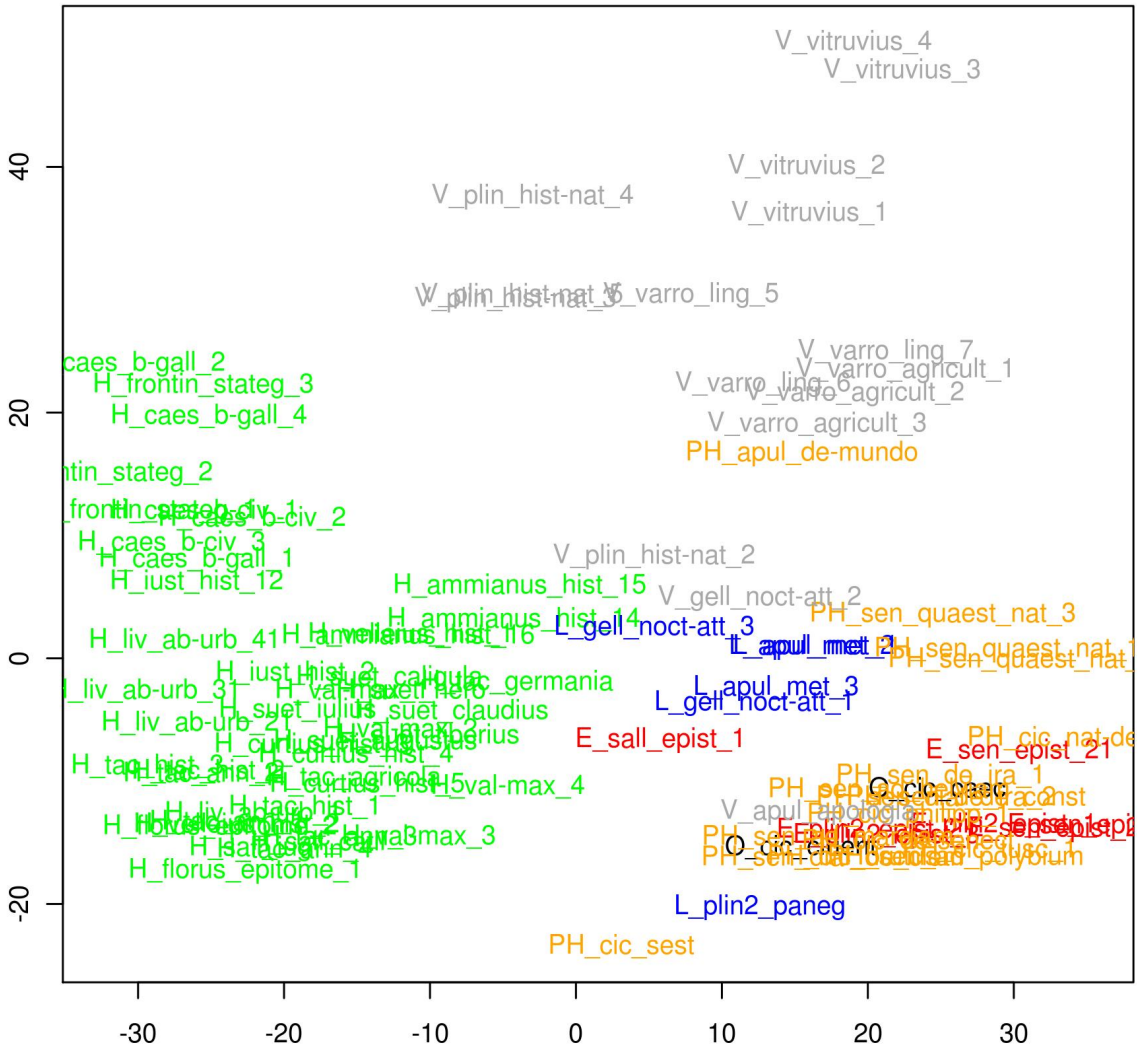
3-grams of POS tags = no words, just labels!



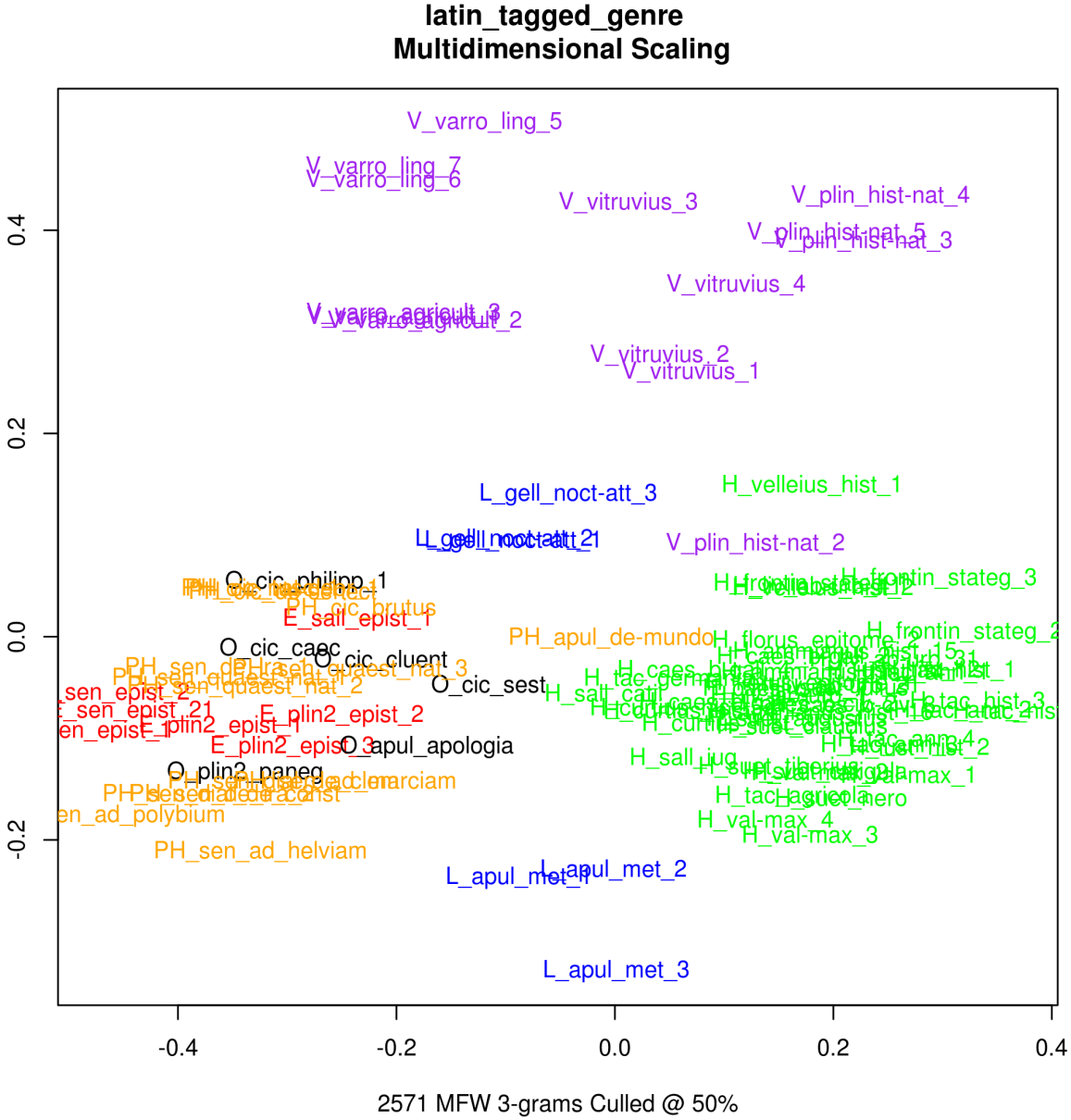
Experiment on genre recognition: example

- did Latin philosophers develop their own style?
- do POS-tags depend on genre?
- is there a difference in performance between MFWs and POS tags?

Latin genre: the most frequent words



Latin genre: 3-grams of POS tags



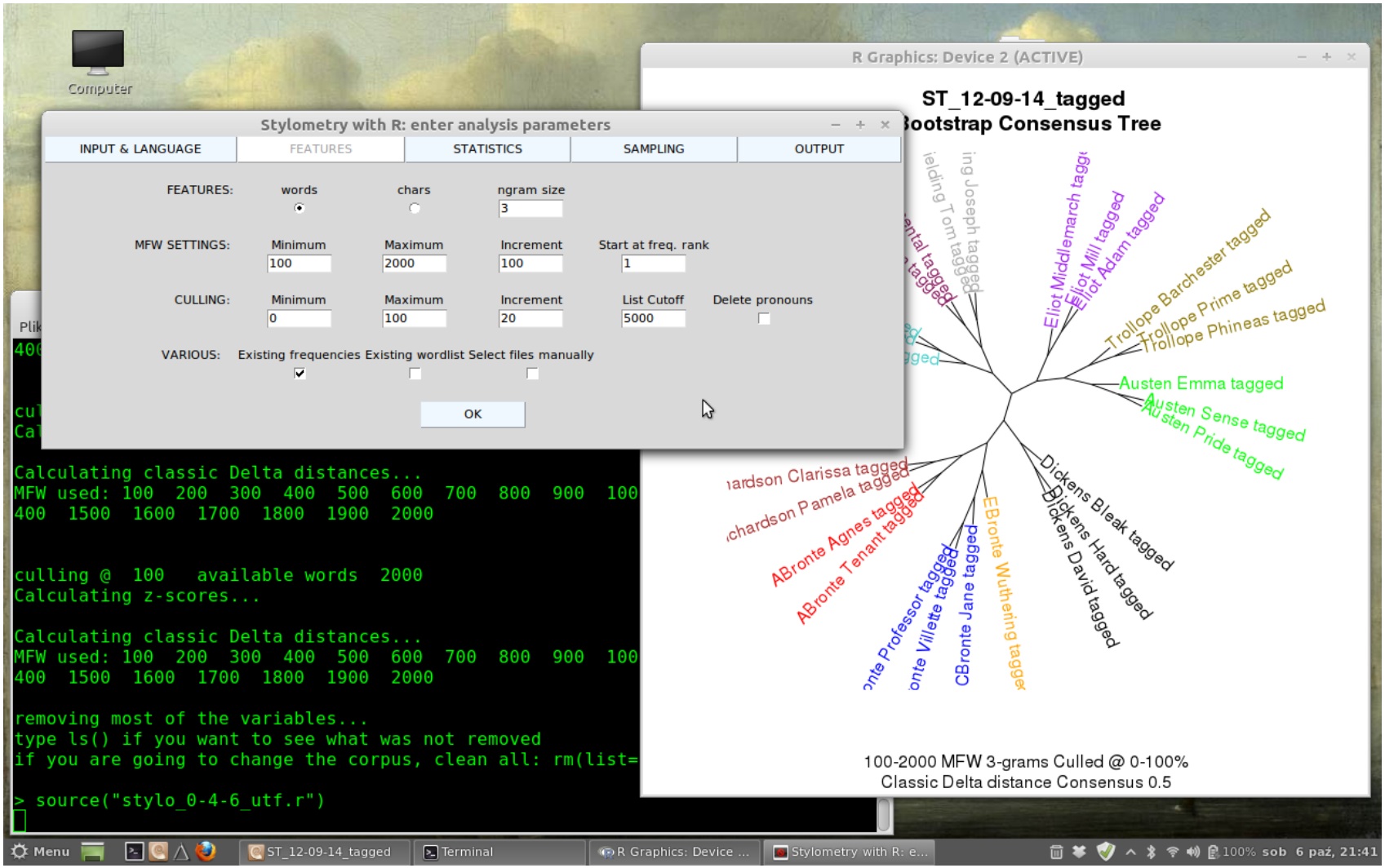
How did they know which tags were appropriate?!

n-acc v-ind sent
n-acc v-ind pun
n-abl v-ind sent
v-ind sent adv
v-ind pun cs
n-abl cc n-abl
adv v-ind pun
n-acc cc n-acc
n-abl v-ind pun
v-ind sent cc
v-ind pun adv
prep n-abl n-gen
v-ind sent pun
v-ind pun rel
prep adj-abl n-abl
...

Conclusions

- Stylometry can retrieve much information from content (i.e. 'normal') words, ...
- ... but bare function words more precisely distinguish authorial uniqueness.
- Best performance, however, is achieved when the words are thrown away!
- Thus, grammatical labels (tags) are worth to be thoroughly tested in future experiments,...
- ... even if it looks pretty heretic.

Run your own experiment!



Thank you!