Stylometry with R

mike.kestemont@gmail.com
www.mike-kestemont.org
University of Antwerp

NEH Summer Institute, Pittsburgh
July 2017

Installing R

- ***R*** (http://www.r-project.org/)
- Open-source statistical software
- Runs on all major platforms
- Install instructions: http://cran.freestatistics.org/

Sublime 3

- For viewing files today
- If you don’t have a good text editor
- (esp. if you are on Windows)
- Install Sublime 3
- Free download
- Install: http://www.sublimetext.com/2

Stylo

- “Stylometry with R”
- https://sites.google.com/site/computationalstylistics/
- Free package for easy stylometric analysis in R
- Graphical user interface (no coding!)

There's no I in team

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

Maciej

Jan

Install Stylo

- Install from within R

- Launch R: double-click icon (e.g. in `Applications`)

- To download and install, type in the console:

  -
    ```
    install.packages("stylo")
    ```

- Every time you restart R, import Stylo:

  -
    ```
    library(stylo)
    ```

Download course material

- Download course materials from:
  - tinyurl.com/y73tc2es
- Unzip the folder (pitt17)
- Place it e.g. on your Desktop

Medieval French Genres

- Jean Bodel (French poet, late 12th C.)
- Famous quote *Chanson de Saisnes*:

> *Ne sont que 3 matières à nul homme atandant,*
> *De France et de Bretaigne, et de Rome la grant.*

- Distinguishes 3 *matières* or “genres”:
  1. *Matière de France* (chansons de geste; Charlemagne)
  2. *Matière de Bretaigne* (romans arturiens; King Arthur)
  3. *Matière de Rome* (romans antiques; e.g. Troie)
- Question: can we distinguish these using stylometry?

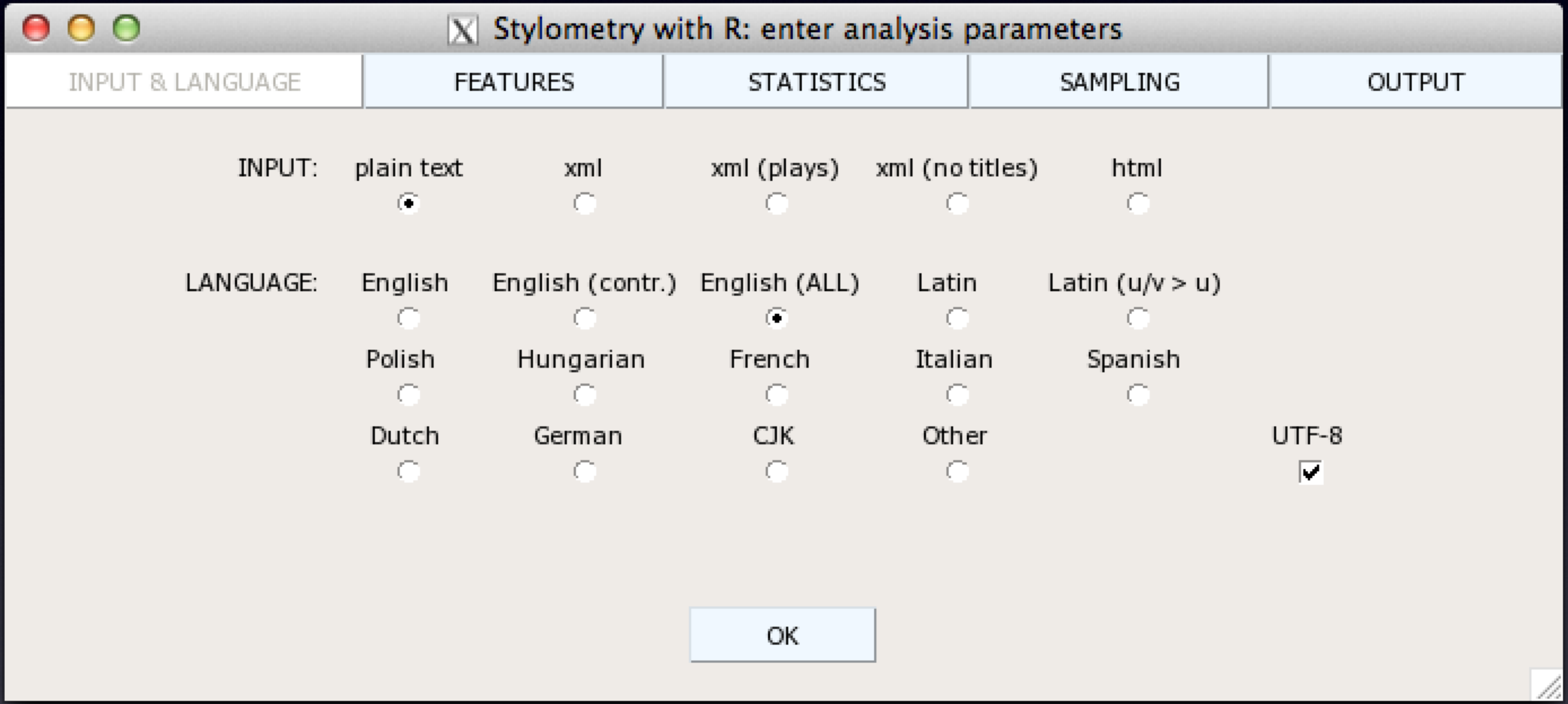
Clustering in Stylo

- Let’s do a clustering experiment on our genres
- Create a folder `corpus` under `pitt17/data/genres/`
- Copy all `bre_*` and `fra_*` texts to this folder

Run stylo

- Stylo needs to know where our data is. Type in R:
  - `setwd(“~/Desktop/pitt17/data/genres/“)`
  - (You can use tab to navigate!)
  - It has to see `corpus` (and not be inside it!)
- Make sure stylo is loaded:
  - `library(stylo)`
- Run command:
  - `stylo()`
- The GUI should load…

Stylo GUI

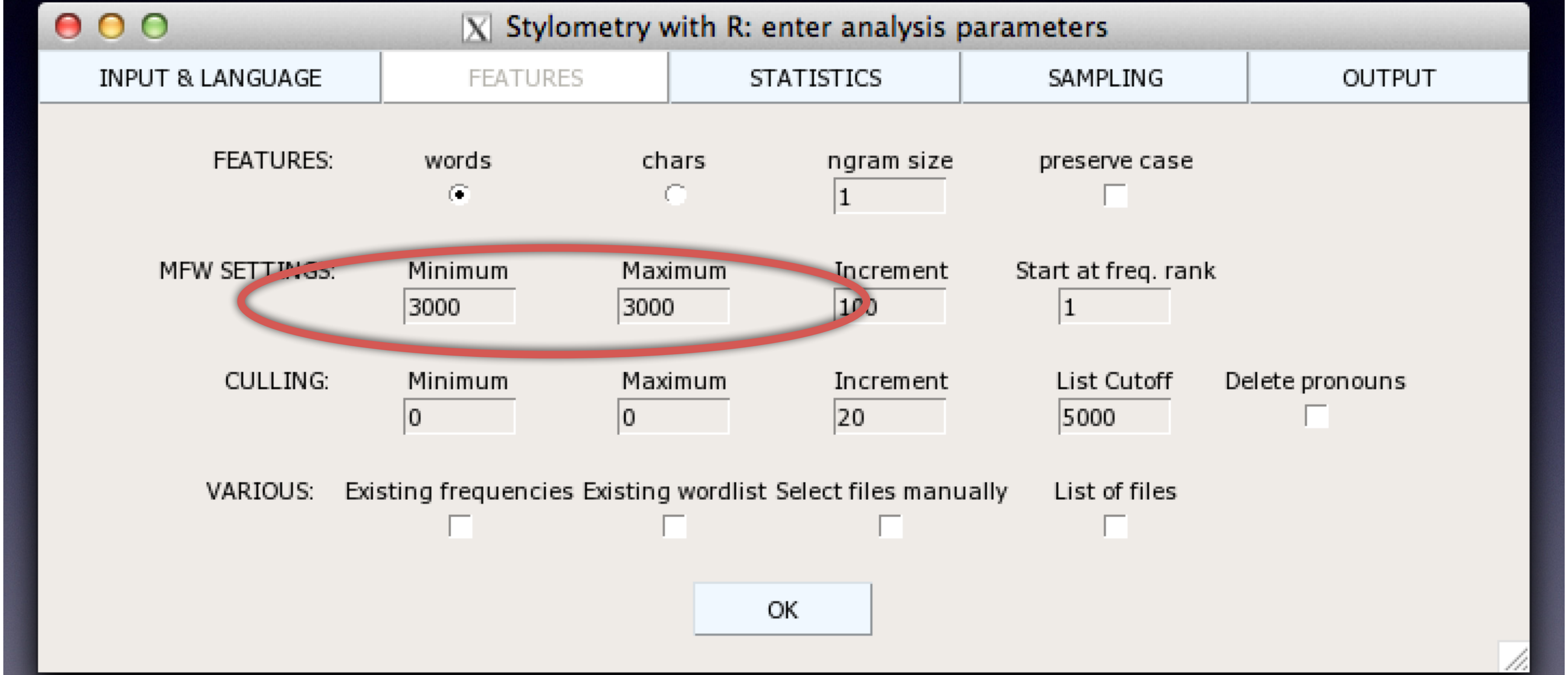


INPUT: ● plain text ○ xml ○ xml (plays) ○ xml (no titles) ○ html

LANGUAGE:
○ English ○ English (contr.) ● English (ALL) ○ Latin ○ Latin (u/v > u)
○ Polish ○ Hungarian ○ French ○ Italian ○ Spanish
○ Dutch ○ German ○ CJK ○ Other

UTF-8 ☑

Adjust parameters and hit OK



Stylometry with R: enter analysis parameters

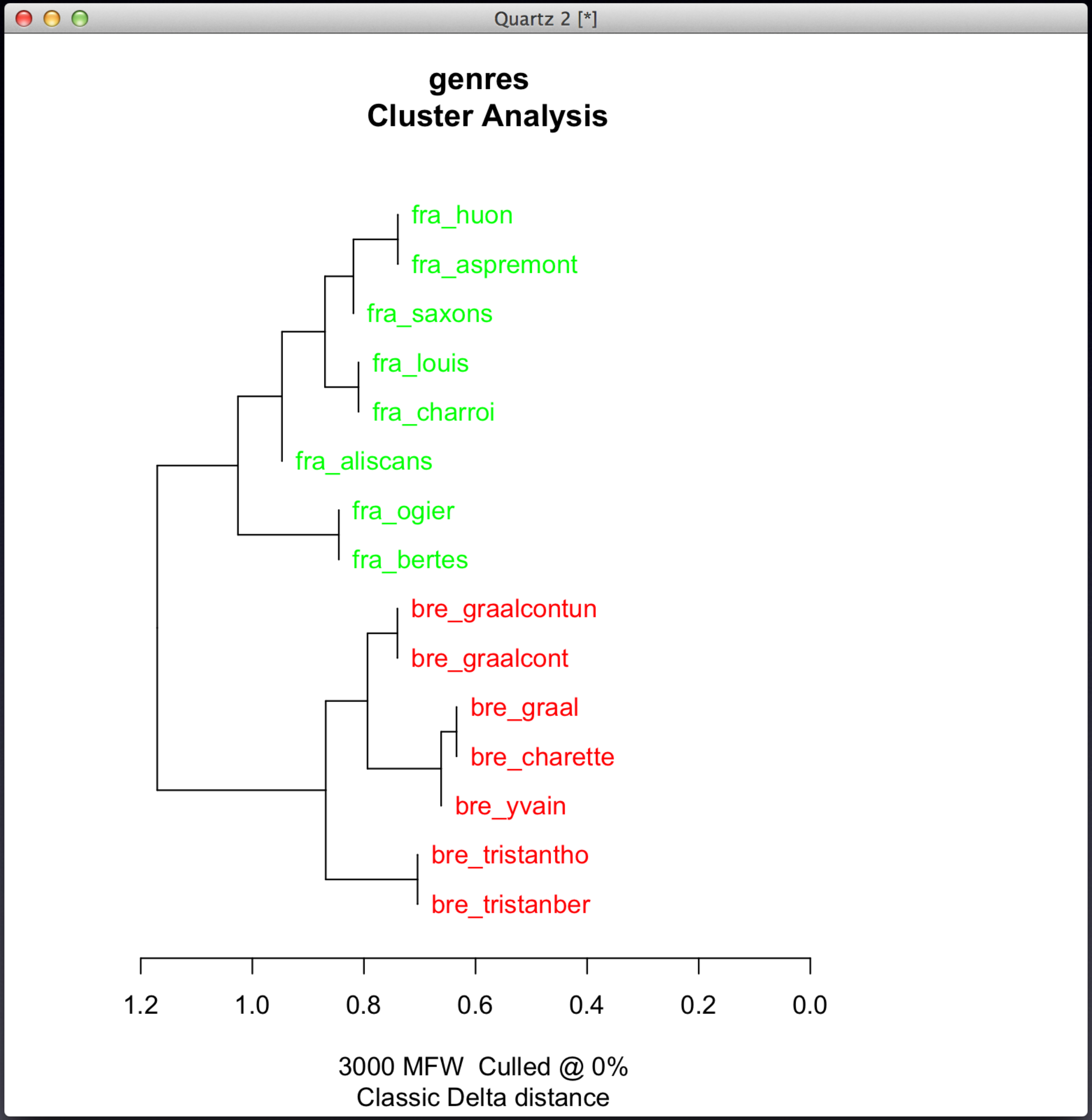
FEATURES: words ● chars ○ ngram size 1 preserve case □

MFW SETTINGS: Minimum 3000 Maximum 3000 Increment 100 Start at freq. rank 1

CULLING: Minimum 0 Maximum 0 Increment 20 List Cutoff 5000 Delete pronouns □

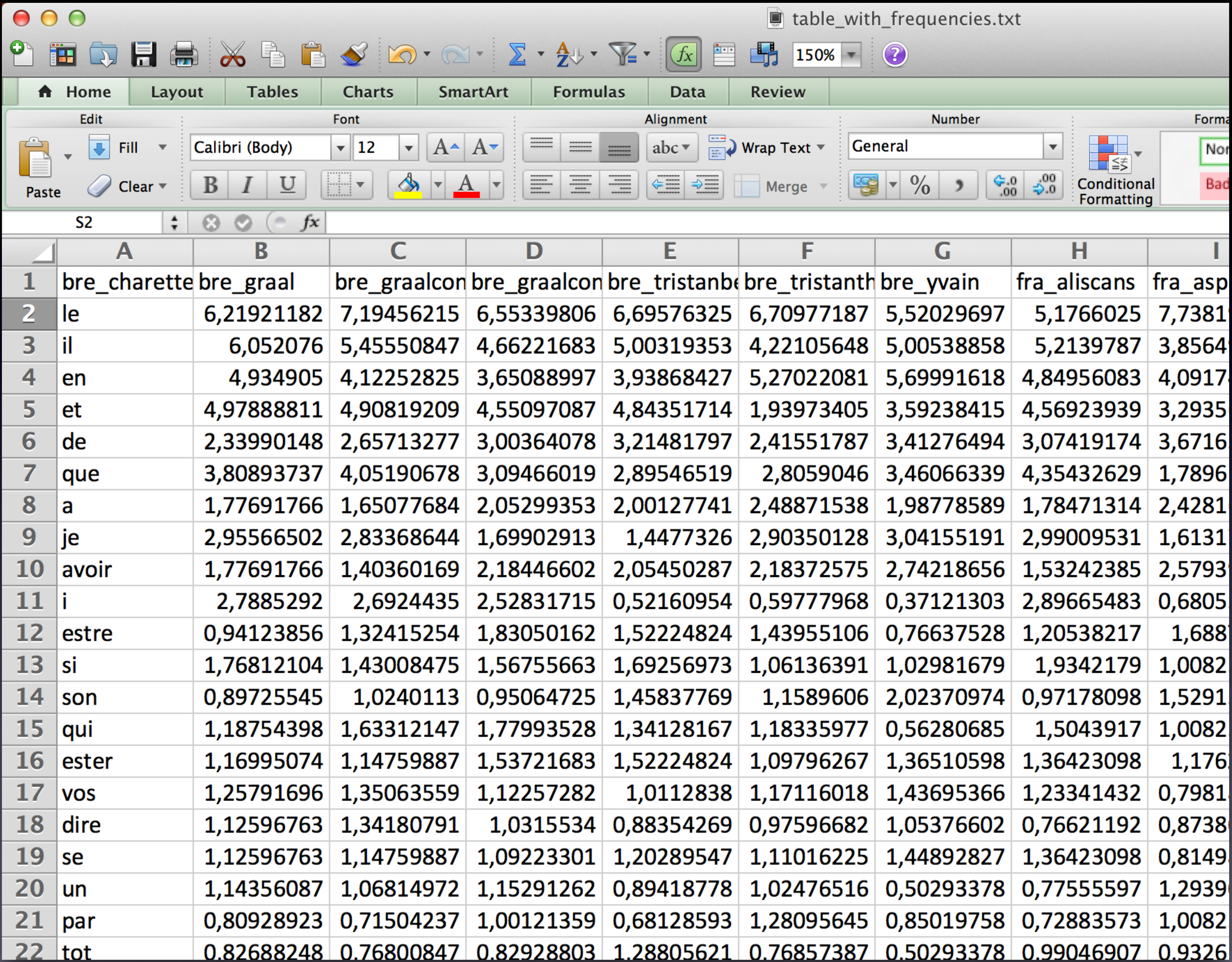
VARIOUS: Existing frequencies □ Existing wordlist □ Select files manually □ List of files □

And you should get a tree…



OK… What happened?

- We represent texts as “bags of words”
- Create a large frequency table:
  - each column = text
  - each row = word
  - each cell = relative frequency
- Check out `table_with_frequencies.txt`

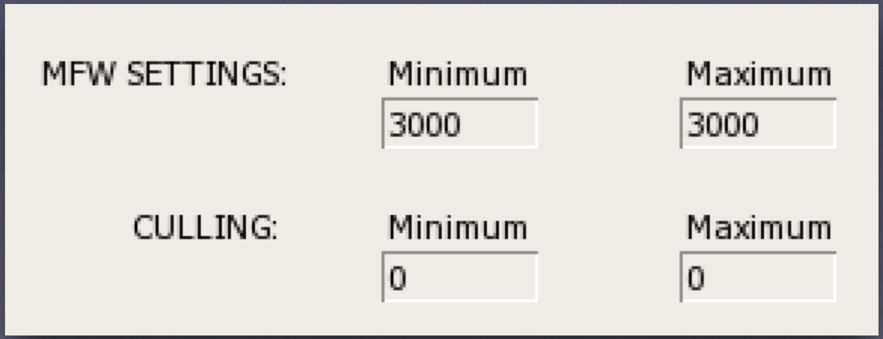


Bag of words?

- We **ignore** word order, position of word in document, syntax, …
- Only use word counts
- Relative frequencies

Only use 3,000 words

- Most Frequent Words: `MFW`
- Better for statistics
- Check out `wordlist.txt`
- What kind of words are most frequent?



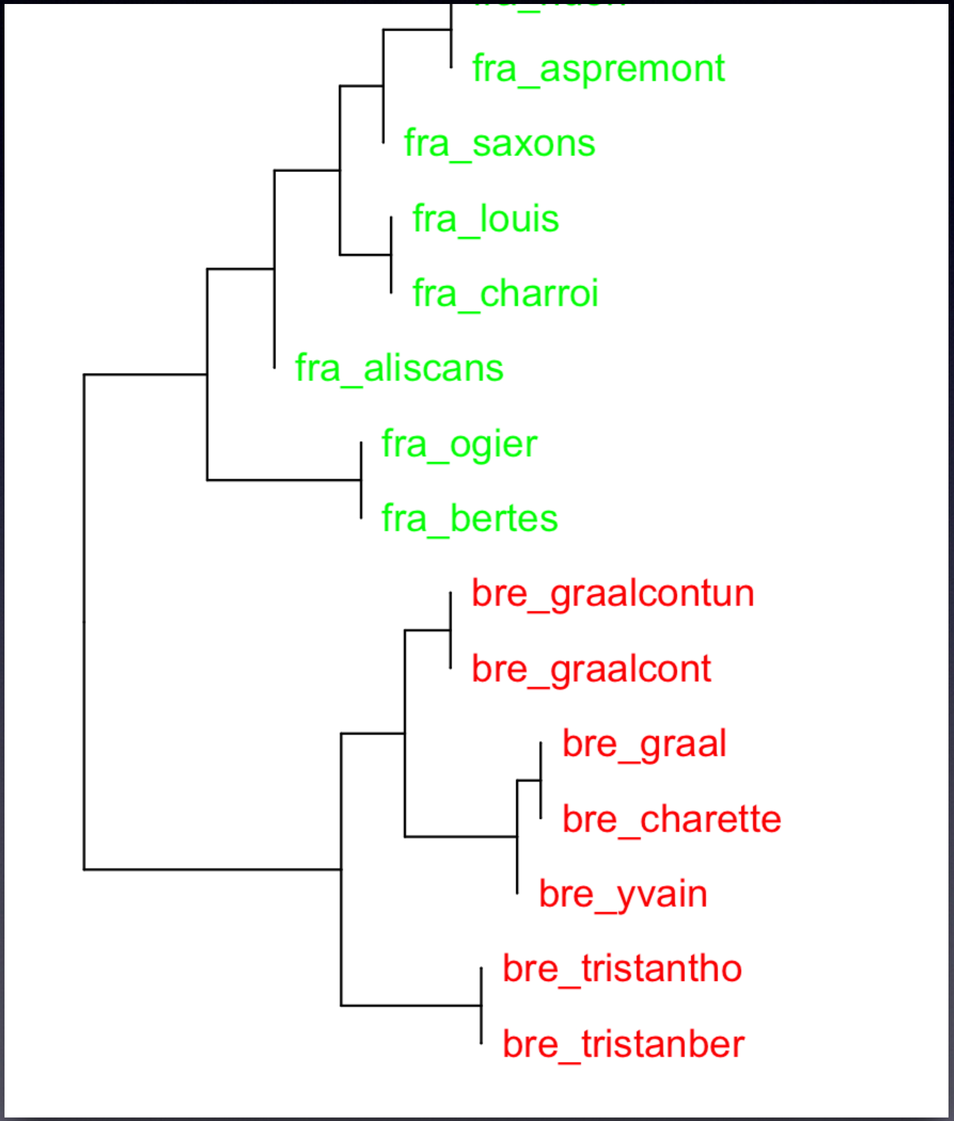
	Minimum	Maximum
MFW SETTINGS:	3000	3000
CULLING:	0	0

Distance matrix

Dist()	Text1	Text2	Text3
Text1	0.0	Dist(Text1, Text2)	Dist(Text1, Text3)
Text2	Dist(Text2, Text1)	0.0	Dist(Text2, Text3)
Text3	Dist(Text3, Text1)	Dist(Text3, Text2)	0.0

Build tree

- Now we build a tree bottom-up
- First, join 2 texts that are most similar
- Combine them in a new node
- Work you way up the three
- Until all texts are joined
- Horizontal axis reflects (dis)similarity



Do it yourself (1)

1. Try out different parameters:

   - Vary the number of `MFW` (under `features` tab): 30, 50, 1000, 5000, … (Always update `Minimum` and `Maximum` simultaneously!)

   - Vary the `distance metric` (under `statistics` tab)

   - Do you get different results? “Better” results?

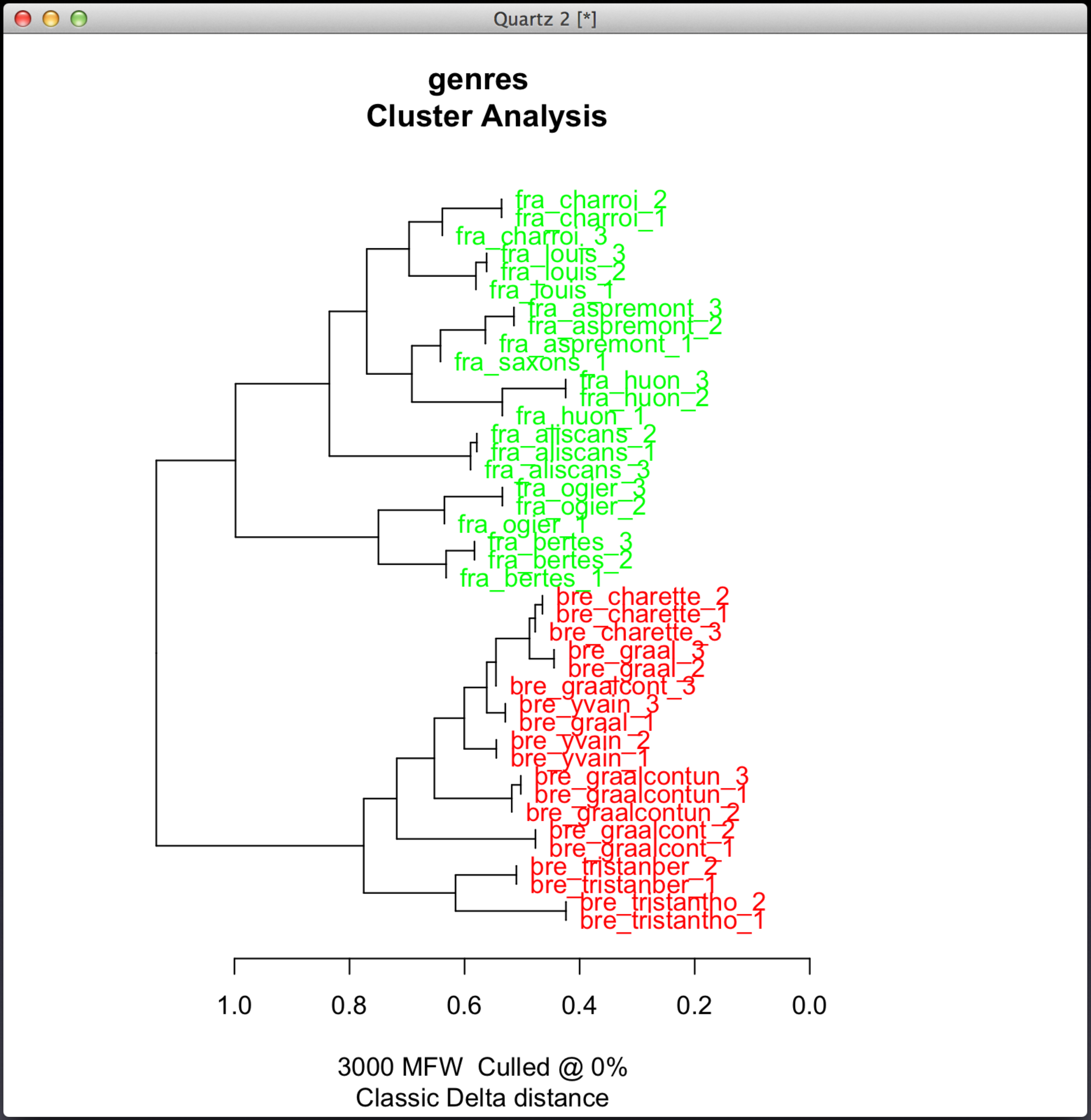
- *Graal*, *Yvain* and *Charette* always cluster together. Can you think of an explanation why?

Do it yourself (2)

1. Under the `sampling` tab, select `Normal sampling` and insert 3,000 under `Sample size`.

2. Run the analysis again. There are much data points now: can you guess what happened?

3. Set the `Sample size` at an absurd size: e.g. 20,000. Do you get an error? Why?


Quartz 2 [*]
genres
Cluster Analysis
fra_charroi_2
fra_charroi_1
fra_charroi_3
fra_louis_3
fra_louis_2
fra_louis_1
fra_aspremont_3
fra_aspremont_2
fra_aspremont_1
fra_saxons_1
fra_huon_3
fra_huon_2
fra_huon_1
fra_aliscans_2
fra_aliscans_1
fra_aliscans_3
fra_ogier_3
fra_ogier_2
fra_ogier_1
fra_bertes_3
fra_bertes_2
fra_bertes_1
bre_charette_2
bre_charette_1
bre_charette_3
bre_graal_3
bre_graal_2
bre_graalcont_3
bre_yvain_3
bre_graal_1
bre_yvain_2
bre_yvain_1
bre_graalcontun_3
bre_graalcontun_1
bre_graalcontun_2
bre_graalcont_2
bre_graalcont_1
bre_tristanber_2
bre_tristanber_1
bre_tristantho_2
bre_tristantho_1
1.0
0.8
0.6
0.4
0.2
0.0
3000 MFW Culled @ 0%
Classic Delta distance

Unstability

- Cluster Analyses can be unstable (cf. 30 > 31 `MFW`)

- Very different results for small change in parameters

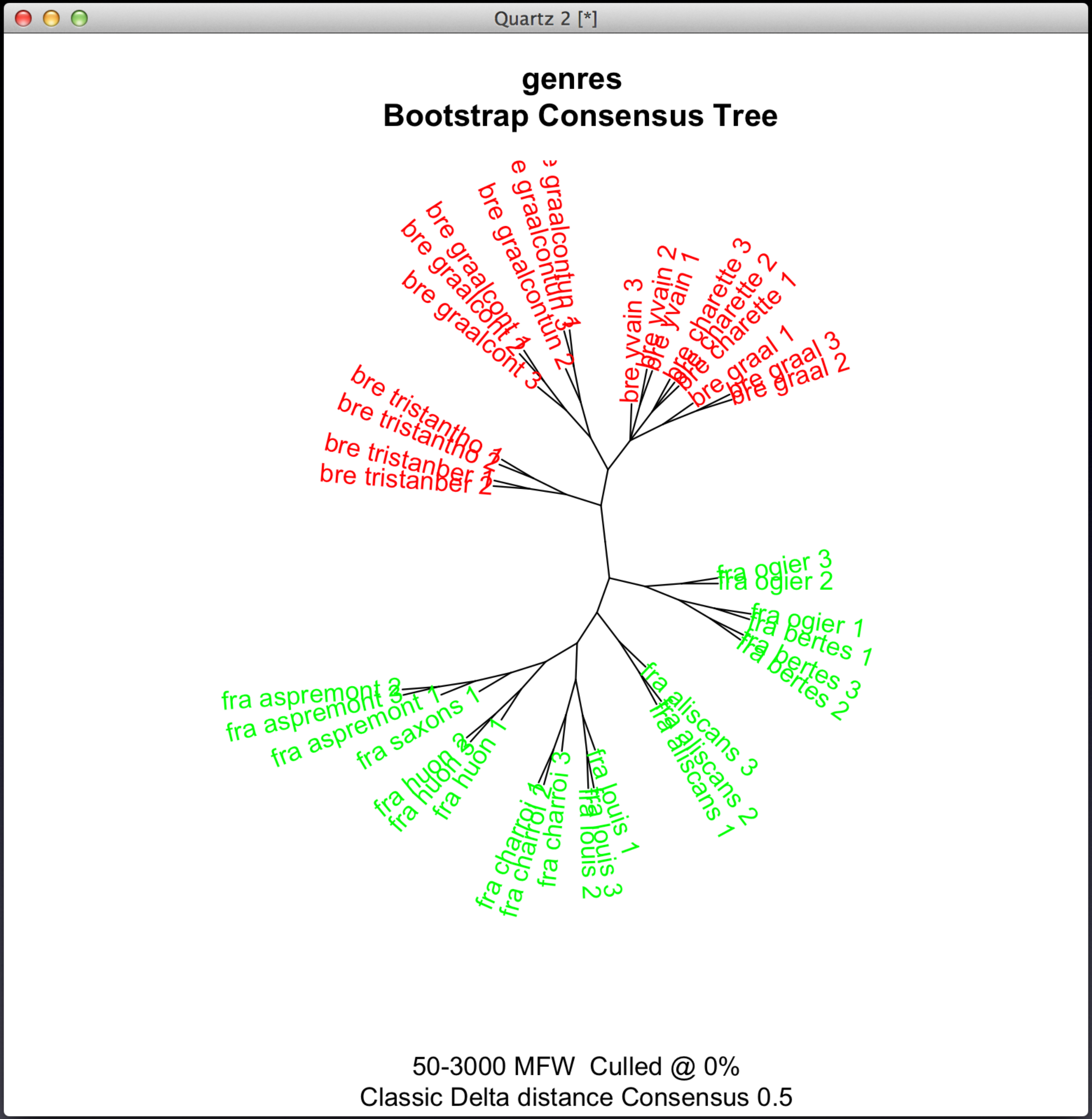
- Rerun experiment with for `MFW: Minimum=50, Maximum=3000, Increment=50`

- We now iteratively run cluster analyses for different frequency bands: `50-100 MFW, 100-150 MFW, 150-200, …, 2900-2950 MFW, 2950-3000 MFW.`

- Do you see the tree change in each picture?

Bootstrap Consensus Trees

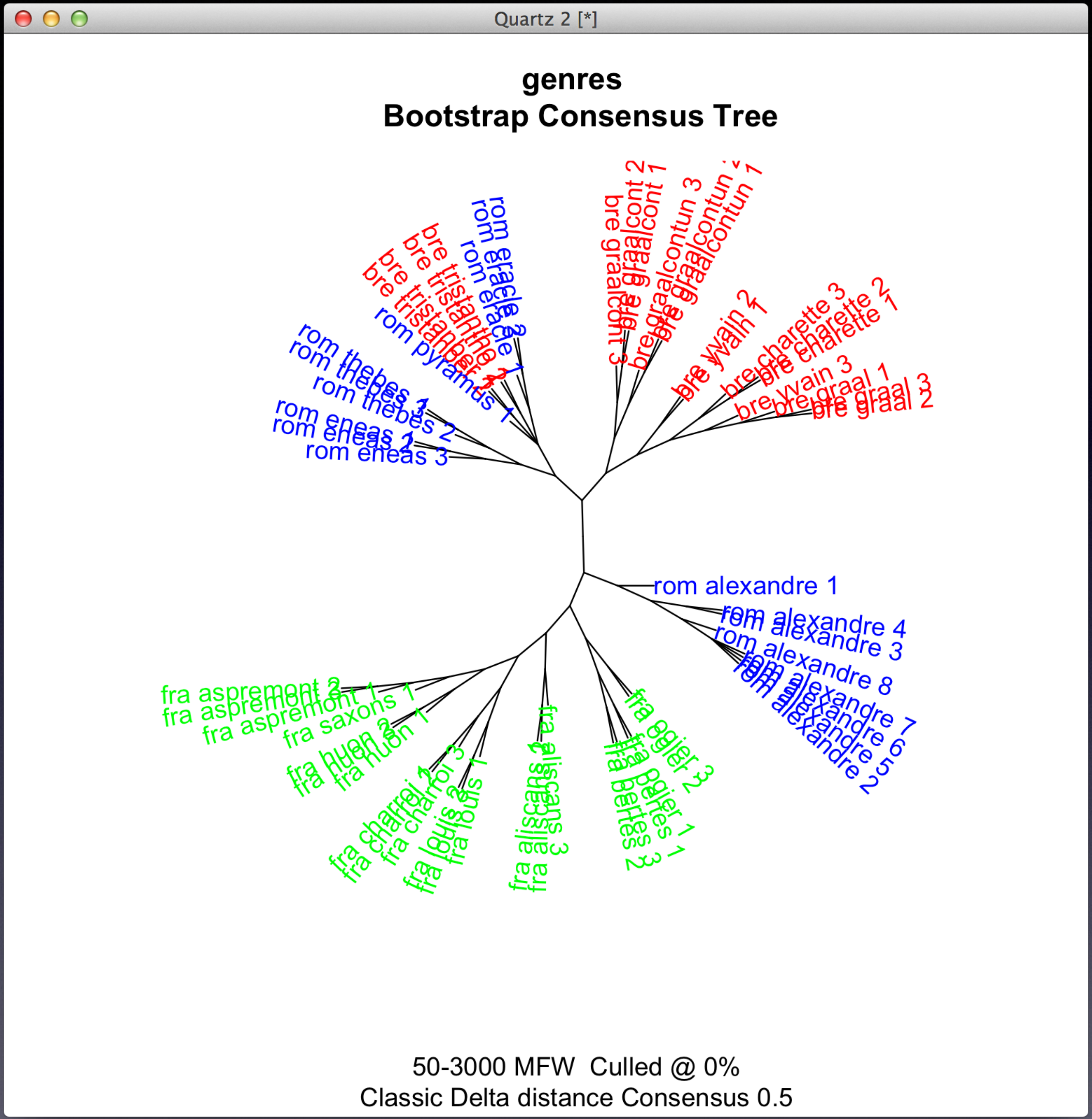
- Bootstrap Consensus Trees (BCT)
- Gives “summary” of different cluster analyses
- Only visualises nodes on which there is a consensus among the trees (50% majority vote)
- Rerun analysis, but select `Consensus Tree` (under `statistics`), but leave `Consensus strength` to `0.5`


Quartz 2 [*]
genres
Bootstrap Consensus Tree
bre graalcontun 3
bre graalcontun 1
bre graalcontun 2
bre graalcont 1
bre graalcont 2
bre graalcont 3
bre yvain 3
bre yvain 2
bre yvain 1
bre charette 3
bre charette 2
bre charette 1
bre graal 1
bre graal 3
bre graal 2
bre tristantho 1
bre tristantho 2
bre tristanber 1
bre tristanber 2
fra ogier 3
fra ogier 2
fra ogier 1
fra bertes 1
fra bertes 3
fra bertes 2
fra aliscans 3
fra aliscans 2
fra aliscans 1
fra louis 1
fra louis 3
fra louis 2
fra charroi 3
fra charroi 2
fra charroi 1
fra huon 1
fra huon 3
fra huon 2
fra saxons 1
fra aspremont 1
fra aspremont 3
fra aspremont 2
50-3000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

Do it yourself (+)

- We have seen that the cluster analyses easily distinguish Jean Bodel's *matière de Bretaigne* and *matière de France* without supervision. But what about the *matière de Rome*? Add the `rom_*` texts under data to the `corpus` folder.

- Rerun various cluster analyses on this expanded data set and experiment with the BCT. Experiment with different MFWs and sample sizes. What is the result? Do you get pretty clusters? How do you interpret this? Which two Arthurian texts behave strangely?



genres
Bootstrap Consensus Tree

50-3000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

Text selection

- Sometimes you don't want to analyse all texts under `corpus`

- Under `features`, tick `Select files manually`

- You will get a dialogue window:
    - (De)select individual texts using `Control+Click`
    - Select a range of texts using `Shift+Click`

- Try to run an analysis using only the `bre_*` and `rom_*` texts

Do it yourself (1)

- I downloaded the entire oeuvre by Dante Alighieri (1265-1321) from danteonline.it

- (I don’t know anything about Dante, and I don’t speak Italian)

- Still, analyse his oeuvre: “Distant” Reading!

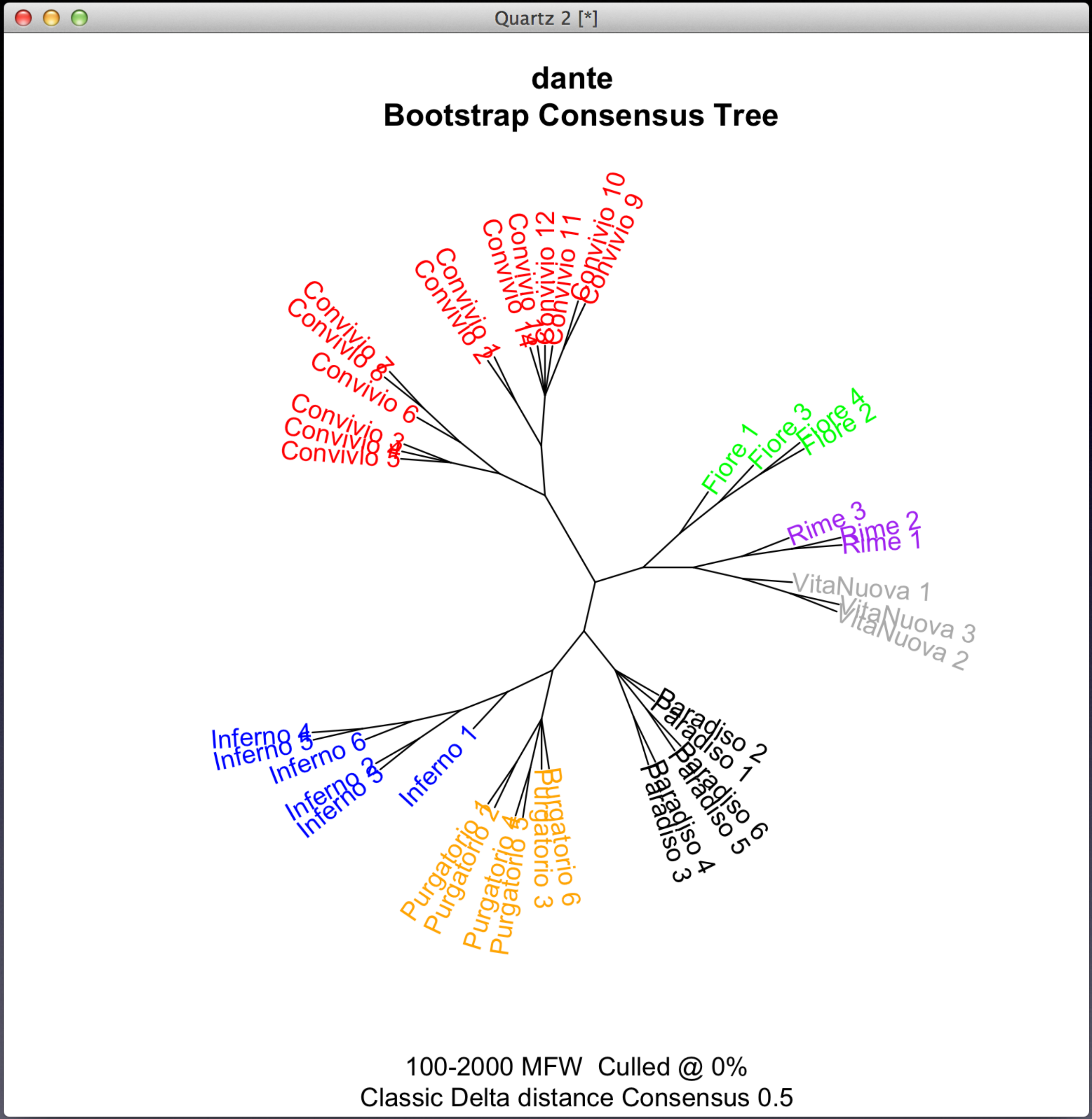
- Type `setwd(“pitt17/data/dante”)` in R to navigate to the correct directory

Do it yourself (2)

- Run various (normal) cluster analyses on Dante's work: try different `MFWs`. (Don't use sampling yet: `No sampling`) Do you see a clear clustering of texts?

- Analyze these two clusters using `oppose()`. Don't forget to create the folders necessary for this: divide the texts in a `primary` and `secondary set`. Result? Silly me! Can you too find out why these two clusters are there?

- Add cluster labels followed by "_" in the file names under corpus to sort our the colouring of the cluster plots. Each file should get a title = `clustername_title.txt`



dante
Bootstrap Consensus Tree

100-2000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

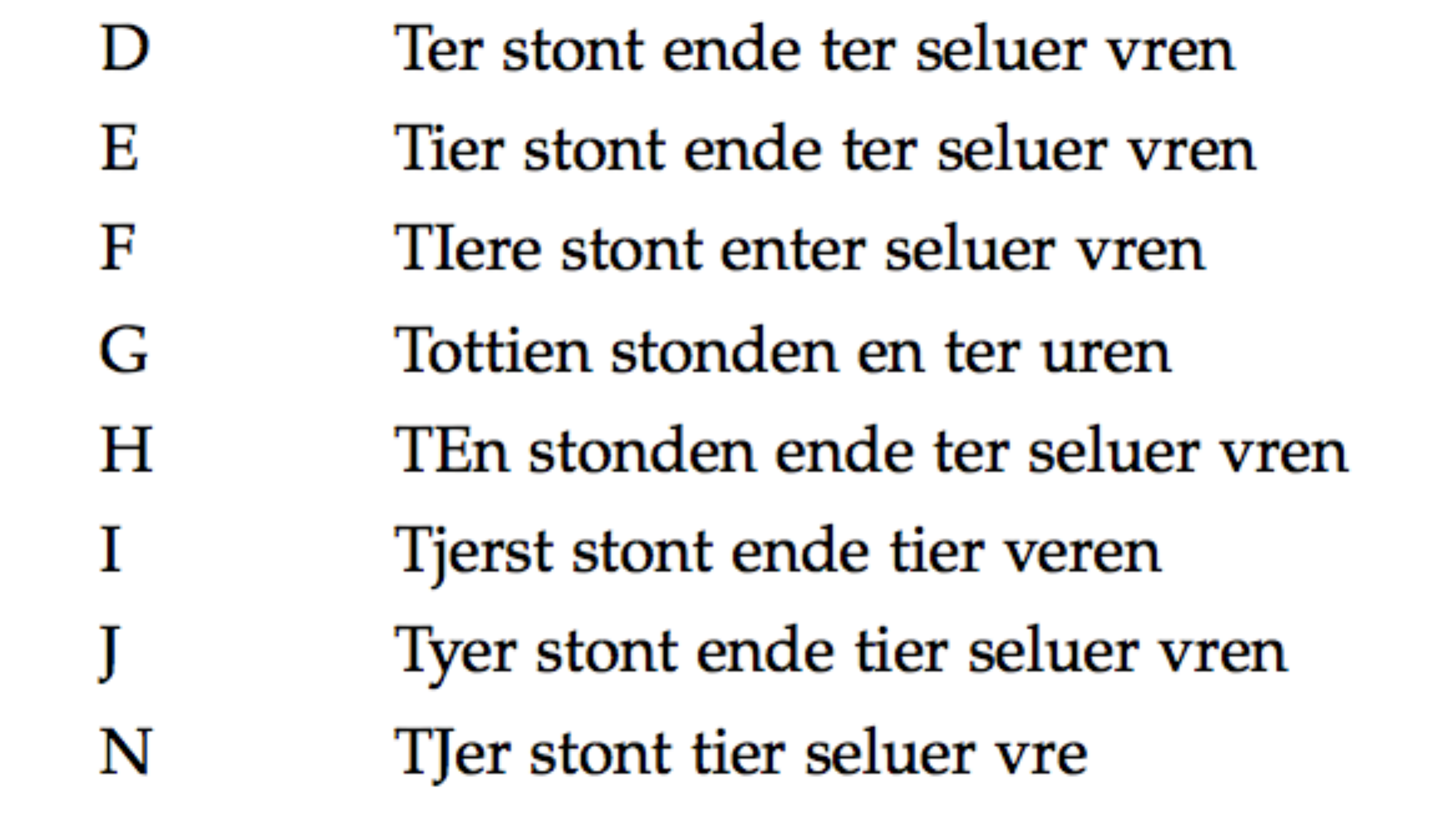
Do it yourself (3)

- Now analyse <u>only</u> the Italian works using `stylo()`.
- Now run Bootstrap Analysis Trees for various `MFWs` (adjust `Minimum, Maximum` and `Increment`).
- Try out different sample sizes (e.g. 5,000). You can leave out *DettoDAmore*, which is too short. Do you see clusters here? Can you explain them using the internet?
- Which two parts of the *Commedia* are closest to each other?
- Use `oppose()` to find out which words are typical of `Paradiso` (in comparison to the other parts).

Spelling variation

- No printing press: manual copying
- Scribes, copyists
- No standard language, spelling
- Regional, personal preferences
- Especially vernacular texts
- Each copy unique

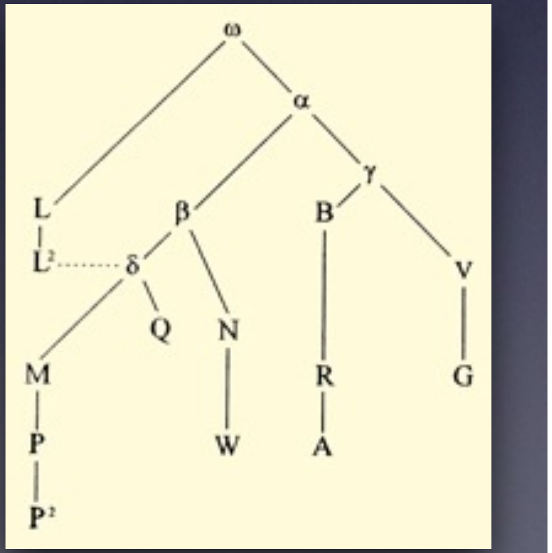
Recognizable?



D	Ter stont ende ter seluer vren
E	Tier stont ende ter seluer vren
F	TIere stont enter seluer vren
G	Tottien stonden en ter uren
H	TEn stonden ende ter seluer vren
I	Tjerst stont ende tier veren
J	Tyer stont ende tier seluer vren
N	TJer stont tier seluer vre

Huge issue

- Issue for computational text analysis
- Lemmatize, part-of-speech tag
- Often seen as problem…
- E.g. stemmatology: reconstruction
- But also interesting!
- Study scribal behaviour



Angus McIntosh

- Middle English philology
- *Linguistic Atlas of Late Medieval English*
- Scribal language
- Interested in modelling scribal behaviour

Hypothesis

- Each scribe has unique ‘profile’
- Combination of:
  - Graphetic profile (*handwriting*)
  - Linguistic profile (*language*)
- Today focus on language:
  - *alt* vs. *olt* (dialect)
  - *tijt* vs. *tyt* (spelling)

3. Chaucer

- Scribal profile in 4 MSS
- Chaucer, *Canterbury Tales*
- Well-studied scribes
- Parallel copies of 1 tale
- *The Man of Law*
- Data courtesy of J. Thaisen

Parallel content:
focus on linguistic differences

hateful harmN condiciounN of povert

with thrist with colɗ with hungR so counfoundiɗ

ohatefułł harme conɗicyouN of pouert

with thurste wiħ colde wiħ hungR so coNfoundeɗ

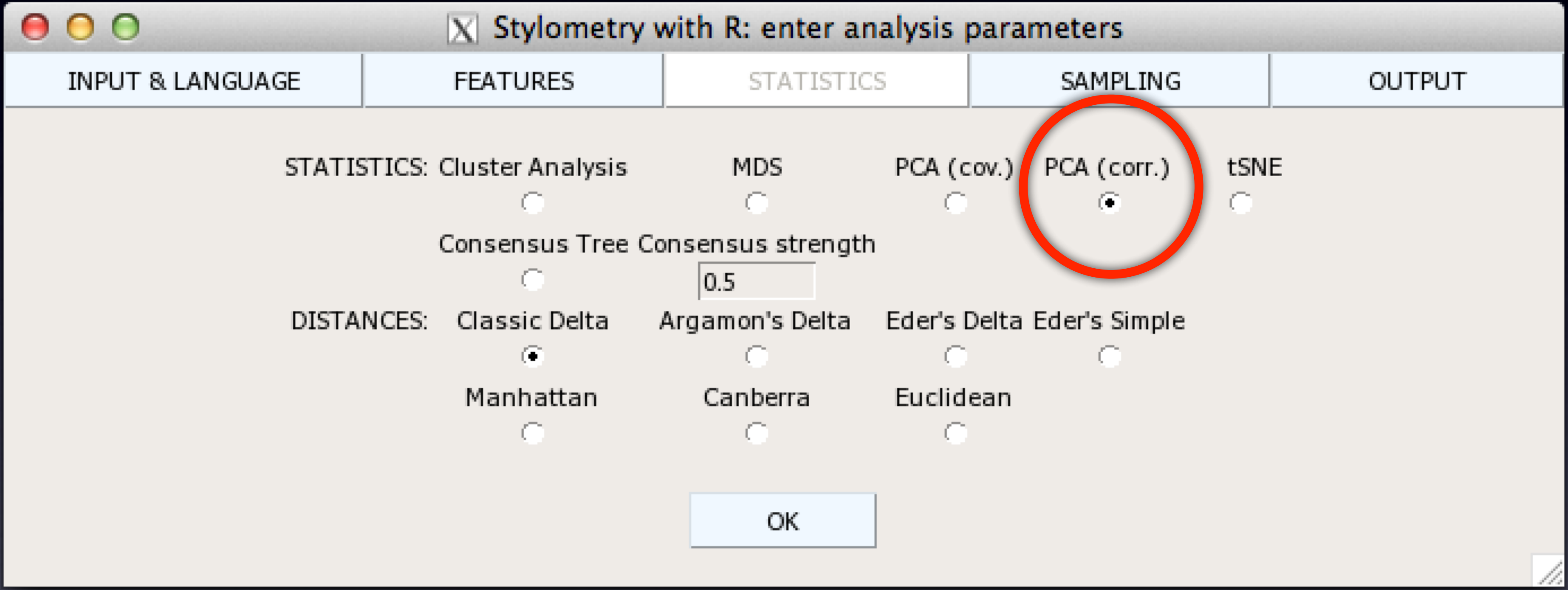
O hate full harme condiciouN of pouerte

wt thrust wt colde and honger so confounded

[Polish Ministry of Science and Higher Education, project grant no. N104 045 32/4256]

Principal Components Analysis

- (My favourite)
- enter `setwd(“~/Desktop/pitt17/data/chaucer")`
- Check out `corpus` folder
- launch `stylo()`:
    - Use `MFW=500`
    - Set method to `PCA (corr.)` under `statistics`
    - `Normal sampling; size=500 (Sampling)`



STATISTICS: Cluster Analysis ○ MDS ○ PCA (cov.) ○ PCA (corr.) ● tSNE ○

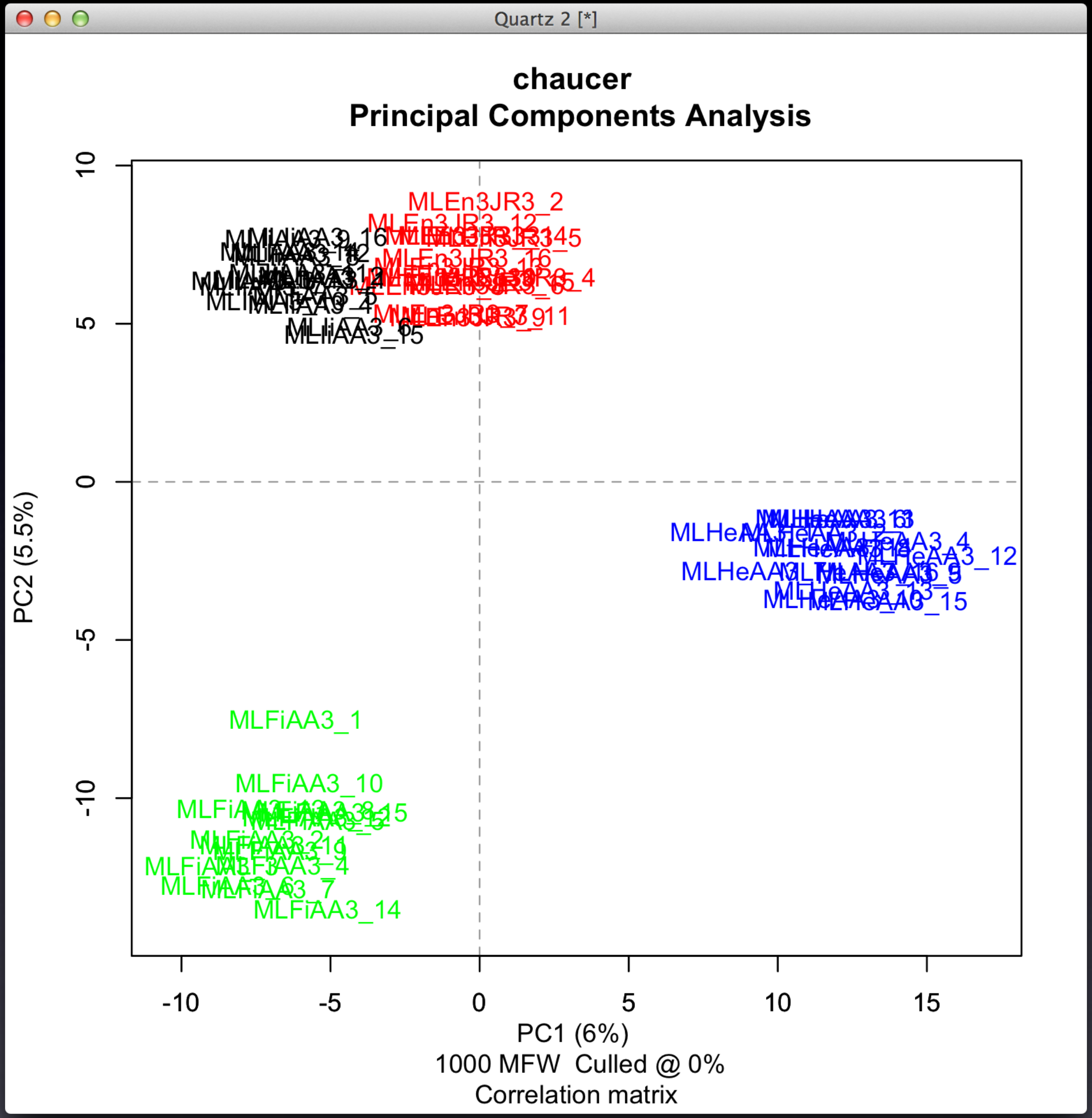
Consensus Tree ○ Consensus strength 0.5

DISTANCES: Classic Delta ● Argamon's Delta ○ Eder's Delta ○ Eder's Simple ○

Manhattan ○ Canberra ○ Euclidean ○

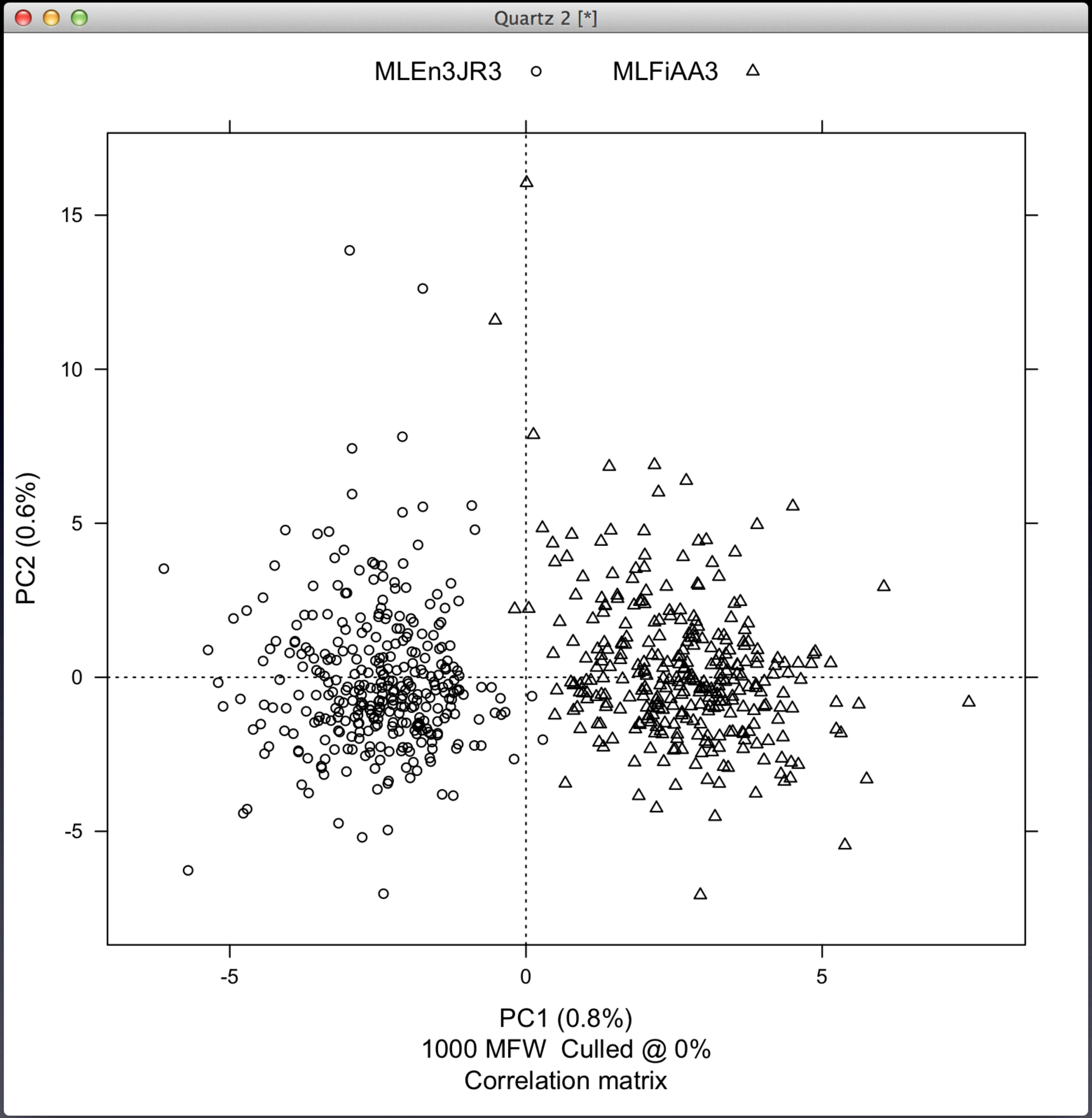
What do you see?

(Does this make sense?)



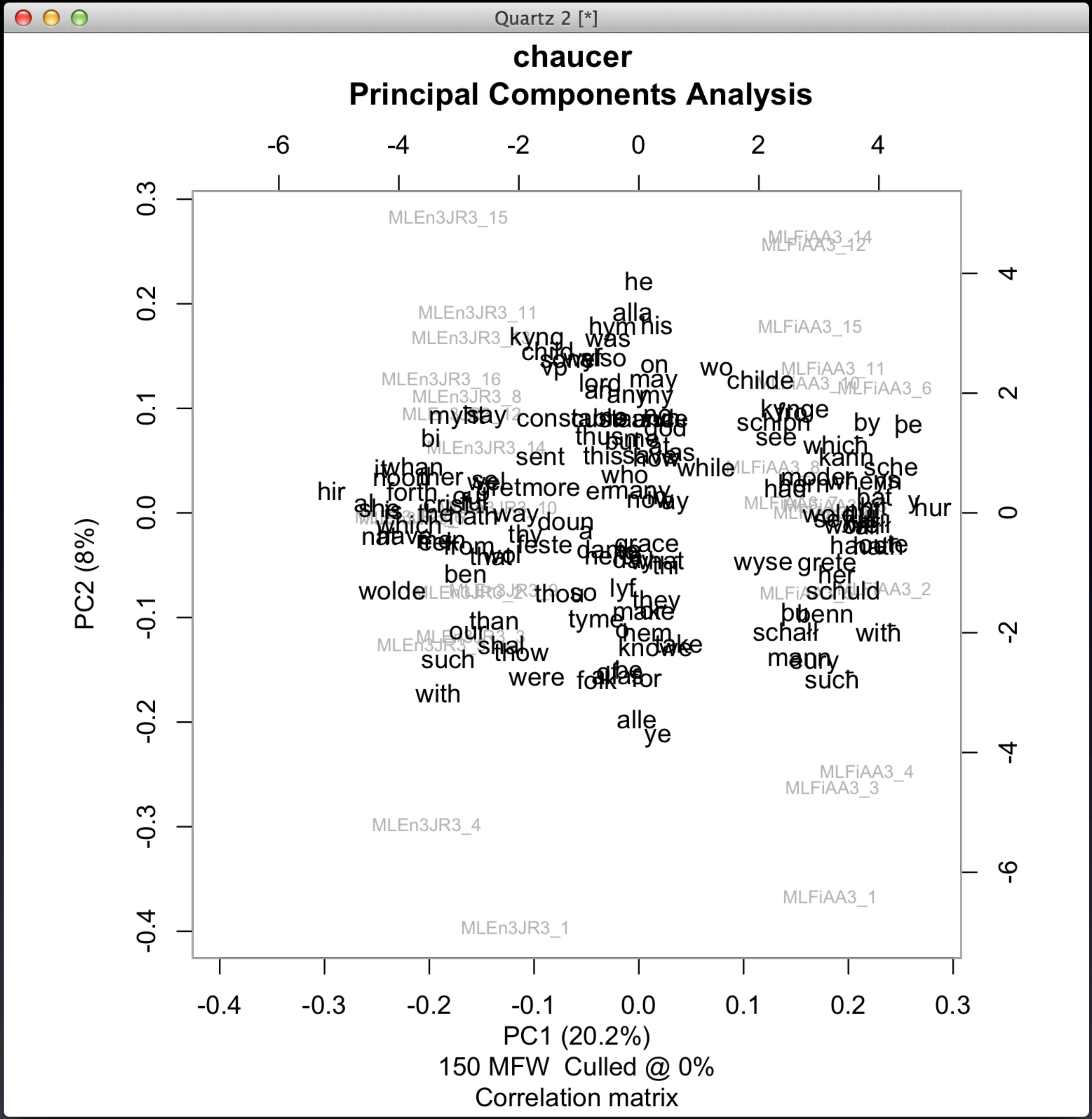
Do it yourself

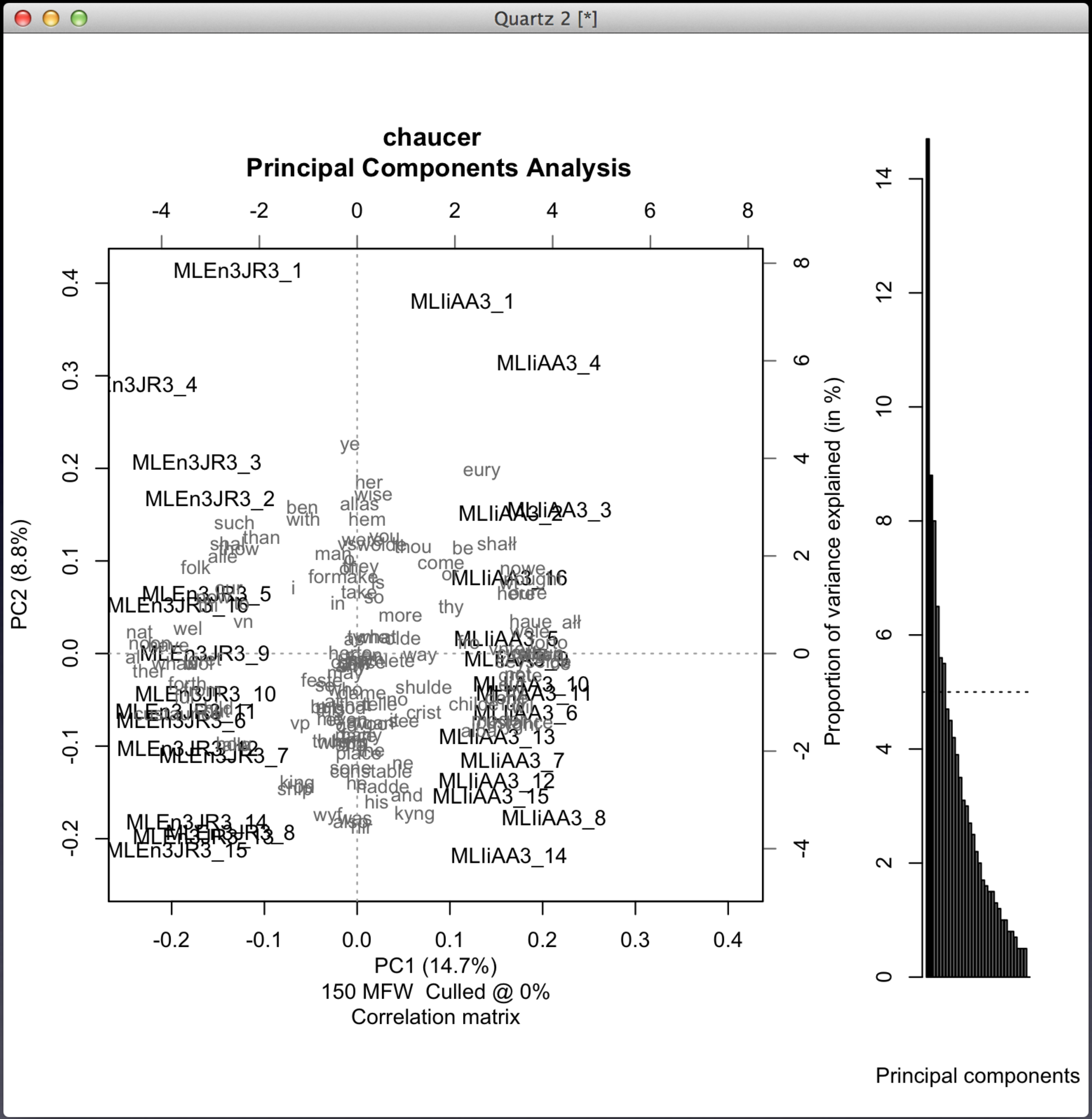
- Select a number of different manuscript *pairs* and *triples* tick `Select files manually` and use `Control+Click`. Can you describe what you see? Where are the samples positioned?
- Use 2 manuscripts. Set `PCA flavour=Symbols` and steadily decrease the sample size (500, 300, …, 50,). How small can samples get before the plot gets fuzzy? What does this tell us?


Quartz 2 [*]
MLEn3JR3 ○
MLFiAA3 △
15
10
5
0
-5
-5
0
5
PC2 (0.6%)
PC1 (0.8%)
1000 MFW Culled @ 0%
Correlation matrix

Loadings

- Extremely helpful feature of PCA
- Tells on which specific word differences the PCA is based
- Use 2 manuscripts. Set `PCA flavour= Loadings.` The loadings will be plotted in dark; the samples in lightgrey. (If difficult to read, lower the `MFW=100`)
- What is there results? Inspect the original files: do the loadings make sense?


Quartz 2 [*]
chaucer
Principal Components Analysis
-6
-4
-2
0
2
4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
PC2 (8%)
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
PC1 (20.2%)
150 MFW Culled @ 0%
Correlation matrix
MLEn3JR3_15
MLFiAA3_14
MLFiAA3_12
MLEn3JR3_11
MLFiAA3_15
MLFiAA3_11
MLEn3JR3_16
MLFiAA3_6
MLEn3JR3_8
MLEn3JR3_14
MLFiAA3_2
MLFiAA3_4
MLFiAA3_3
MLEn3JR3_4
MLFiAA3_1
MLEn3JR3_1
he
alla
hym
his
was
kyng
also
on
wo
childe
lord
may
my
kynge
by
be
schon
see
which
kann
sche
bi
sent
thus
who
while
hir
forth
more
hae
y
hur
which
way
doun
a
thy
leste
grace
wyse
grete
her
schuld
wolde
ben
lyf
they
tyme
make
than
our
them
knowe
schal
with
such
shal
thow
manny
were
folk
for
such
with
alle
ye

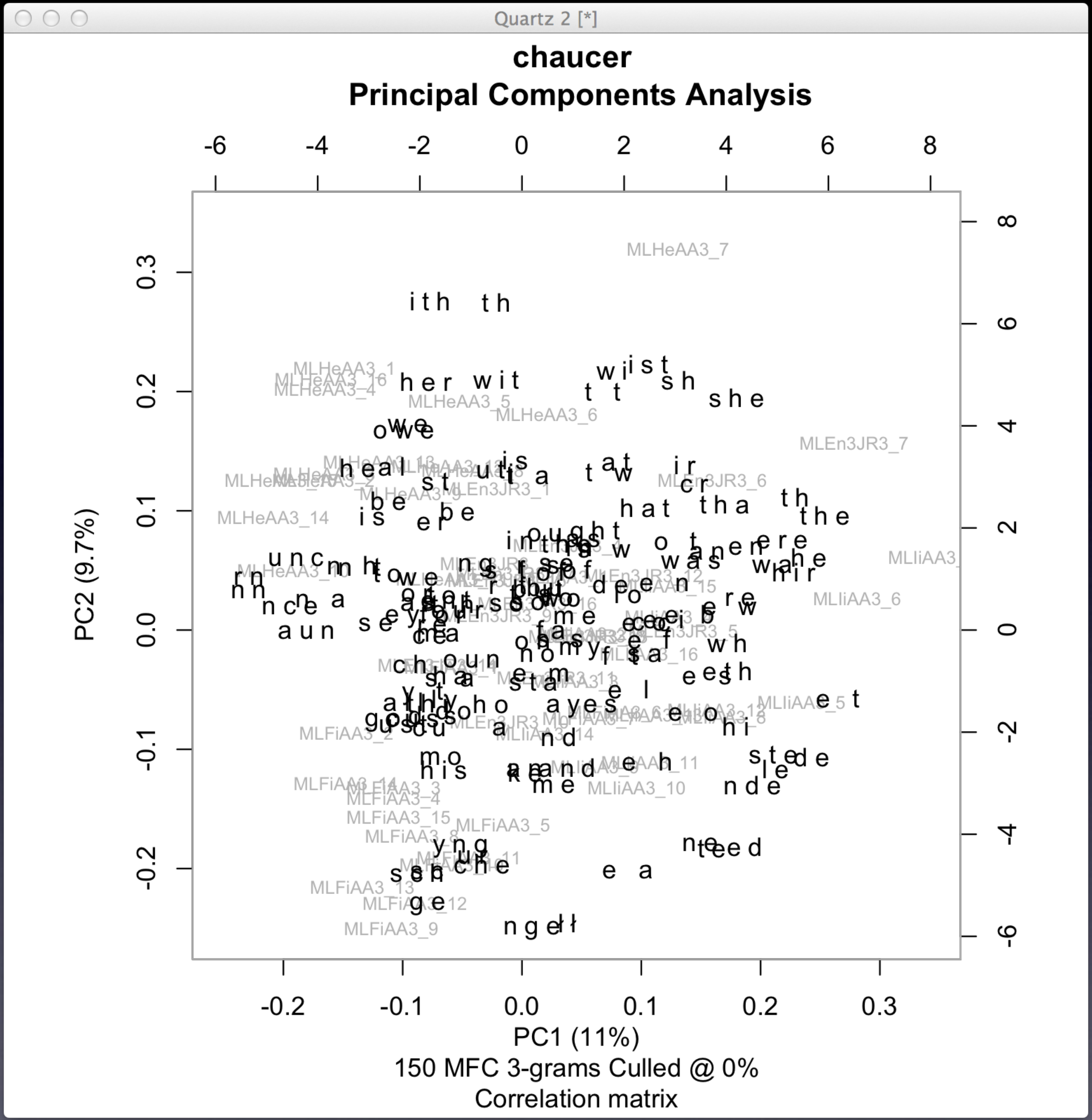


PCA flavour=

Technical

Character n-grams

- Words are not always used in stylometry
- Also character n-grams
- Under `features` tab:
  - `features = chars`
  - `ngram size = 3`
- Make sure to set `PCA flavour=Loadings`
- Can you guess what character n-grams are?


Quartz 2 [*]
chaucer
Principal Components Analysis
PC1 (11%)
PC2 (9.7%)
150 MFC 3-grams Culled @ 0%
Correlation matrix
-6
-4
-2
0
2
4
6
8
-0.2
-0.1
0.0
0.1
0.2
0.3
MLHeAA3_7
MLHeAA3_1
MLHeAA3_16
MLHeAA3_4
MLHeAA3_5
MLHeAA3_6
MLHeAA3_14
MLEn3JR3_7
MLEn3JR3_6
MLEn3JR3_1
MLIiAA3_6
MLIiAA3_5
MLIiAA3_10
MLFiAA3_2
MLFiAA3_4
MLFiAA3_15
MLFiAA3_5
MLFiAA3_8
MLFiAA3_13
MLFiAA3_12
MLFiAA3_9
ith
th
her
wit
wist
sh
she
owe
hat
tha
the
was
ught
aun
nce
und
nde
nteed
ngell

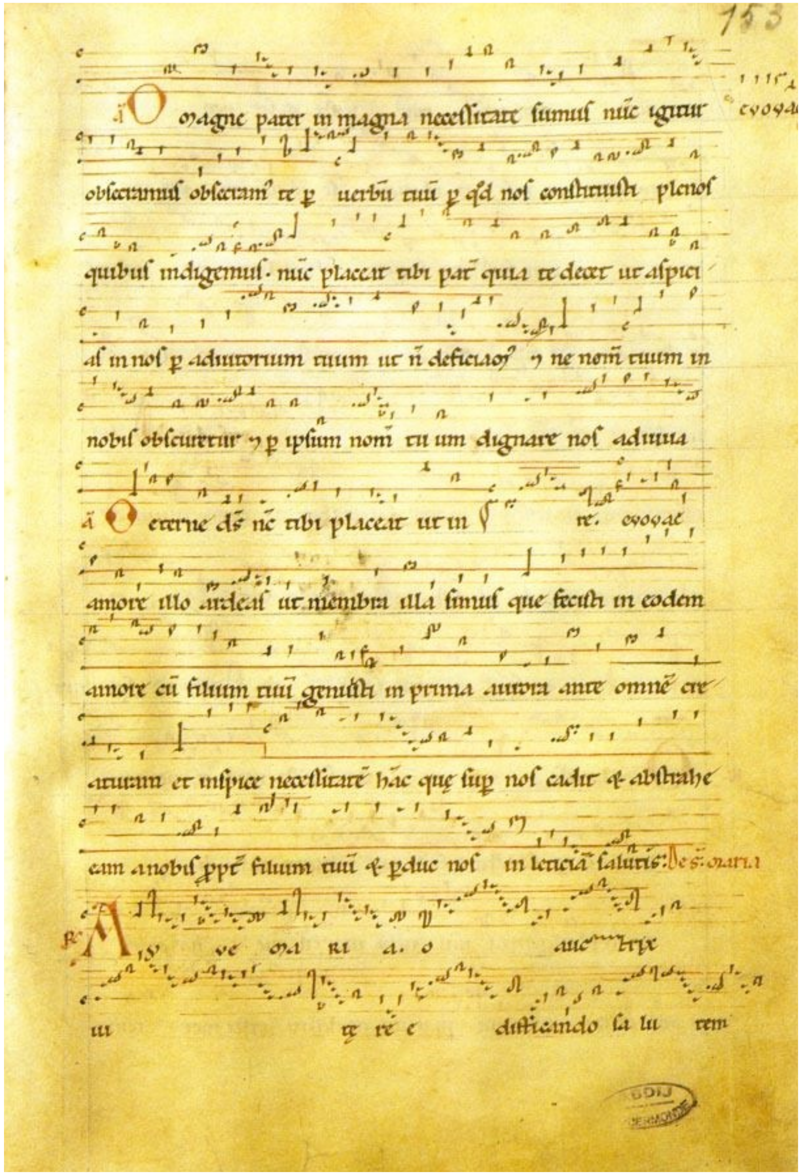
Hildegard of Bingen

- Influential women writer
- 1098–1179
- Germany
- Divine visions
- “Sybil of the Rhine”

[Wiesbaden, Landesbibliothek, 1, fol. 1r.]

Varied oeuvre

- Visions
- Music
- Scientific texts
- Recipes
- Medical treatises
- Letters (pope, emperor, ...)


153
O magne pater in magna necessitate sumus nūc igitur
euouae
obsecramus obsecram te p uerbū tuū p qd nos constituisti plenos
quibus indigemus. nūc placeat tibi pat quia te decet ut aspici
as in nos p adiutorium tuum ut ñ deficiamus et ne nomē tuum in
nobis obscuretur et p ipsum nomē tu um dignare nos adiuua
re. euouae
O eterne ds nūc tibi placeat ut in
amore illo ardeas ut membra illa simus que fecisti in eodem
amore cū filium tuū genuisti in prima aurora ante omnē cre
aturam et inspice necessitatē hanc que sup nos cadit et abstrahe
eam a nobis ppt filium tuū et pduc nos in leticia salutis.
De s. maria
A ve ma ri a. o auc trix
ui te re e dificando sa lu tem

[Dendermonde, St.-Pieters & Paulusabdij, Ms. Cod. 9]

Early 2012

- Sara Moens
- Jeroen Deploige
- Dept. History, UGhent
- Editing two texts
- Collaborate?

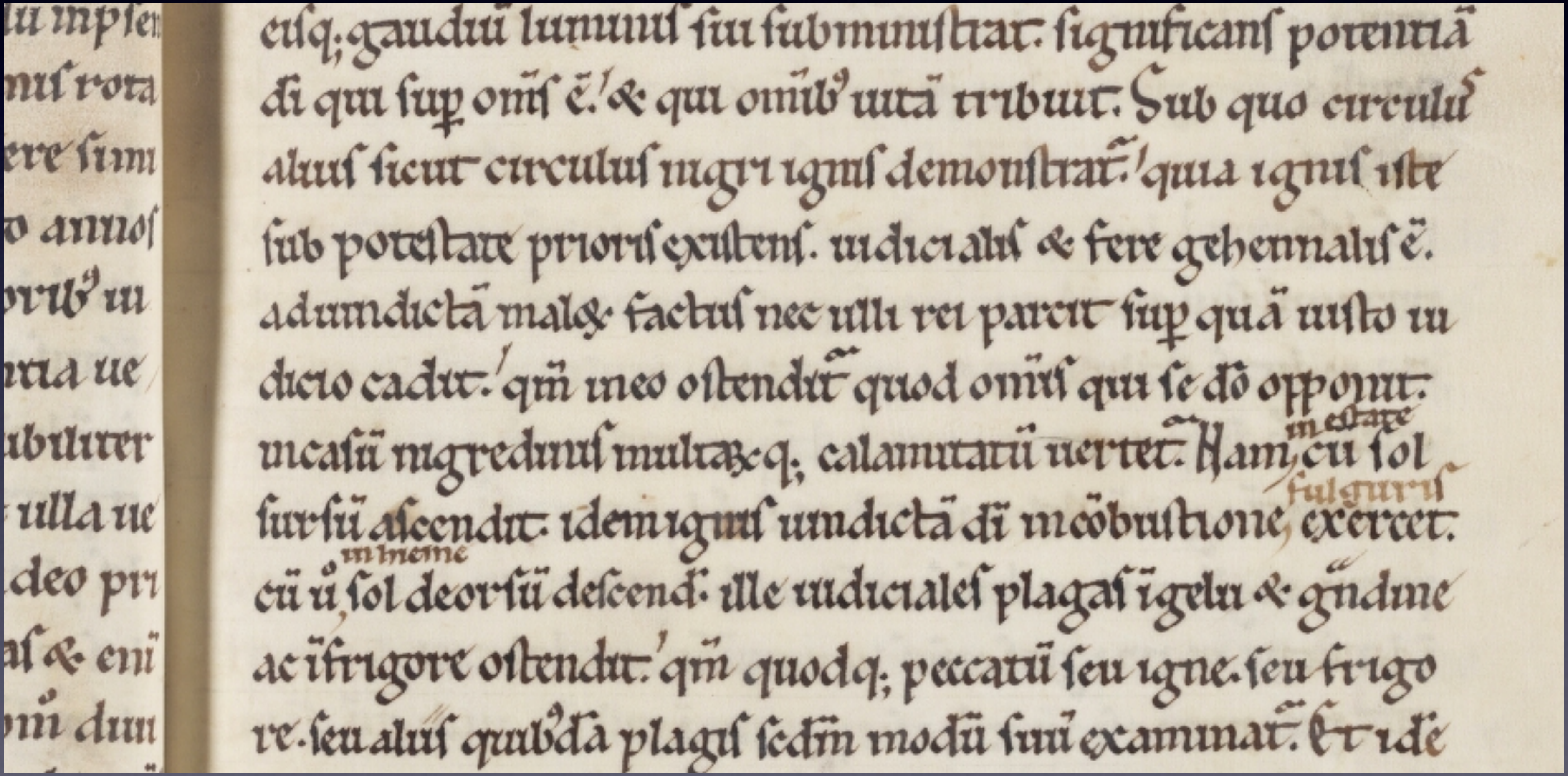
Secretaries

- Wrote in Latin...
- But was bad at it!
- No formal training as woman
- Assisted by male secretaries
- Gender issues...
- Dictated

[Hildegard and her 1st secretary Volmar]

Correction grammatical mistakes

(Only form, not content!)



eiusq; gaudiū luminis sui subministrat. significans potentiã
dī qui sup omīs ē. & qui omnib' uitã tribuit. Sub quo circulus
albus sicut circulus nigri ignis demonstrat͂. quia ignis iste
sub potestate prioris existens. iudicialis & fere gehennalis ē.
ad uindictã malę factus nec ulli rei parcit sup quã iusto iu
dicio cadit. qm̄ in eo ostendit͂ quod omīs qui se dō opponit.
in casū ingrediunis multarūq; calamitatū uertet͂. Nam in estate cū sol
sursū ascendit. idem ignis uindictã dī in cōbustione fulguris exercet.
cū u͛ in hieme sol deorsū descend͂. ille iudiciales plagas ī gelu & grandine
ac ī frigore ostendit. qm̄ quodq; peccatū seu igne. seu frigo
re. seu aliis quib'dã plagis sc̄dm̄ modū sui examinat͂. Et ide

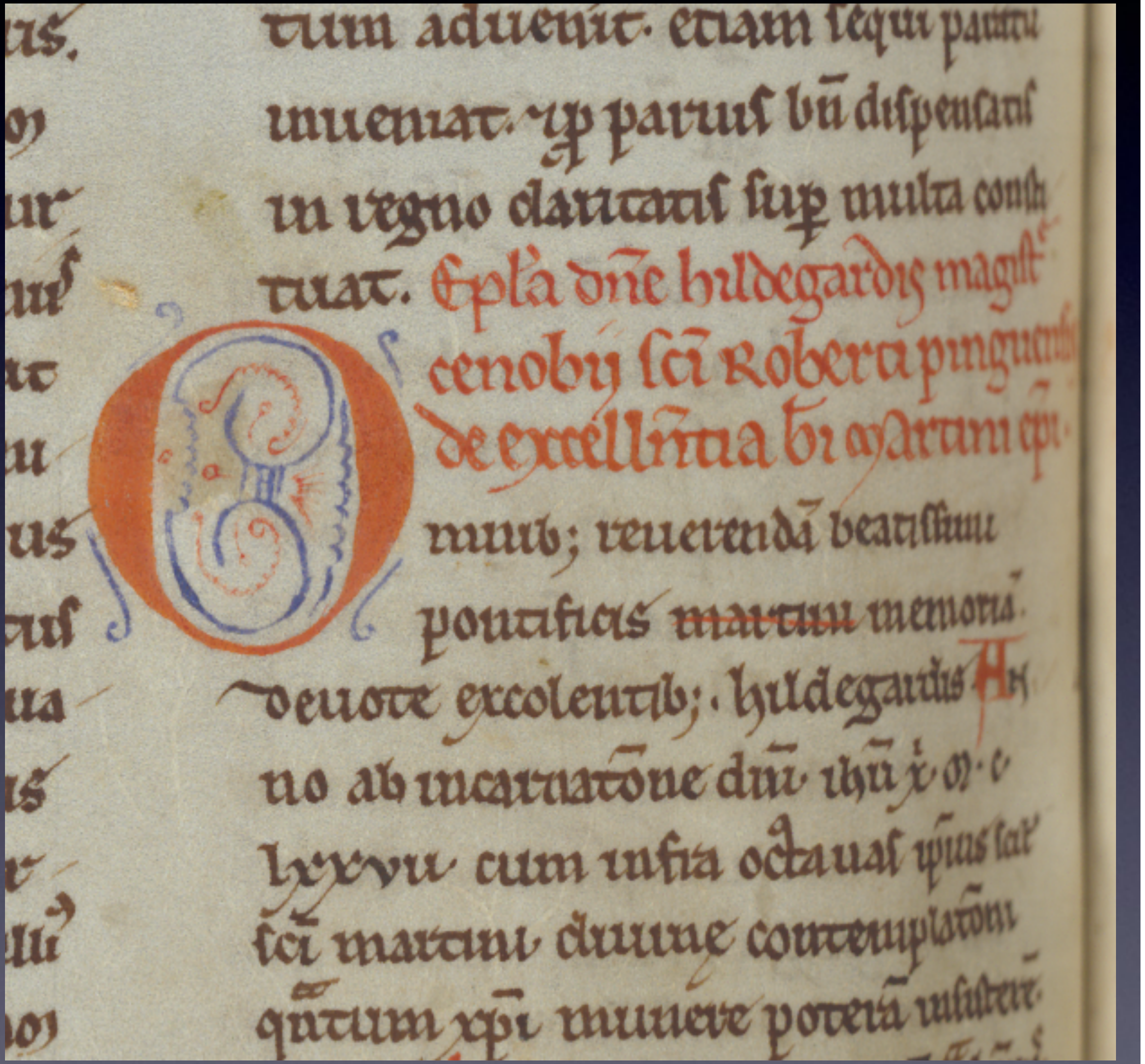
[Ghent, University Library, MS 241]

Two shorter texts...

Visio ad Guibertum missa &

Visio de sancto Martino

- “Attributed” to Hildegard
- *Opera omnia...*
- But style not typical of her
- Doubts authorship?
- Last secretary....

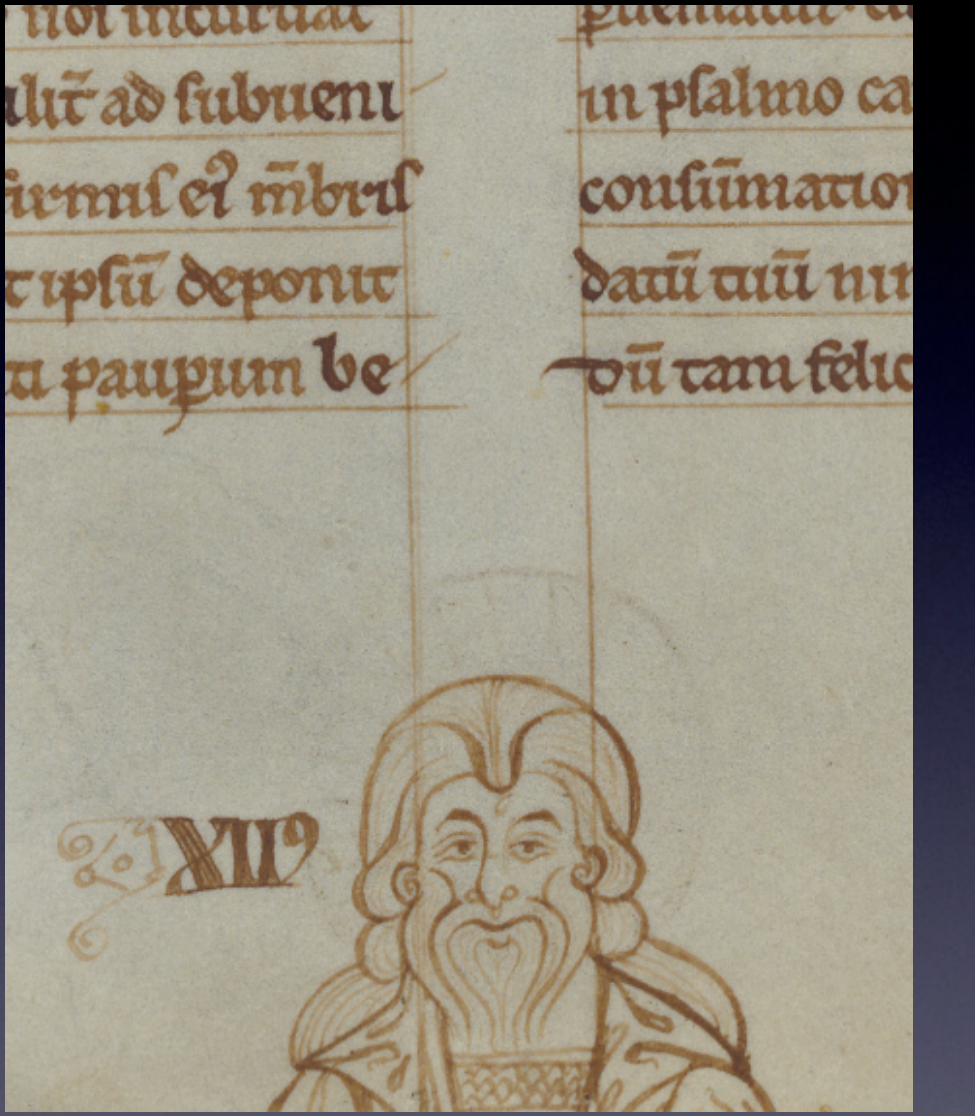


tium aduenit. etiam sequi paruu
inueniat. ꝗ paruus bñ dispensatis
in regno claritatis sup multa consti
tuat. Epła dñe hildegardis magist
cenobii scī Roberti pinguen
de excellentia bī Martini epī.
Omnib; reuerenda beatissimi
pontificis ~~martini~~ memoria
deuote excolentib;. hildegardis An
no ab incarnatione dñi ihū x. m. c
lxxvii cum infra octauas ipius scī
scī martini diuine contemplationis
gustum xpī munere potenti infuderit

[MS Brussels, Royal Library, 5527-34, fol. 141v.]

Guibert of Gembloux

- Monk from Brabant
- Hildegard’s last secretary
- Fascination St Martin
- Very elaborate style
- “Pushy”



[MS Brussels, Royal Library, 5527-34, fol. 141v.]

Stylometry?

When you correct [this text], keep to this rule: that [...] you apply your skill only to make corrections where the order or the rules of correct Latin are violated. Or if you prefer – and this is something I have conceded in this letter beyond my normal practice – you need not hesitate to clothe the whole sequence of the vision in a more becoming garment of speech, preserving the true sense

[*Visio de St. Martino*, trans. Newman, 1987, p. 23]

Corpus

- *Corpus Christianorum* (Brepols)
- Complete materials
- *Epistolaria*
- Hildegard, Guibert
- Bernard of Clairvaux (1090-1153)
- 3x +100k tokens

Do it yourself

- Check out folder `pitt17/data/hildegard`:
  - `B_ep.txt` = Letters from Bernard of Clairvaux
  - `B_Mart.txt` = Sermon about St. Martin by Bernard
  - `D_Mart.txt` = Dubious
  - `D_Missa.txt` = Dubious
  - `G_ep.txt` = Letters by Guibert
  - `H_epG.txt` = Letters by Hildegard, with Guibert
  - `H_epNG.txt` = Letters by Hildegard, before Guibert
- All texts lemmatised

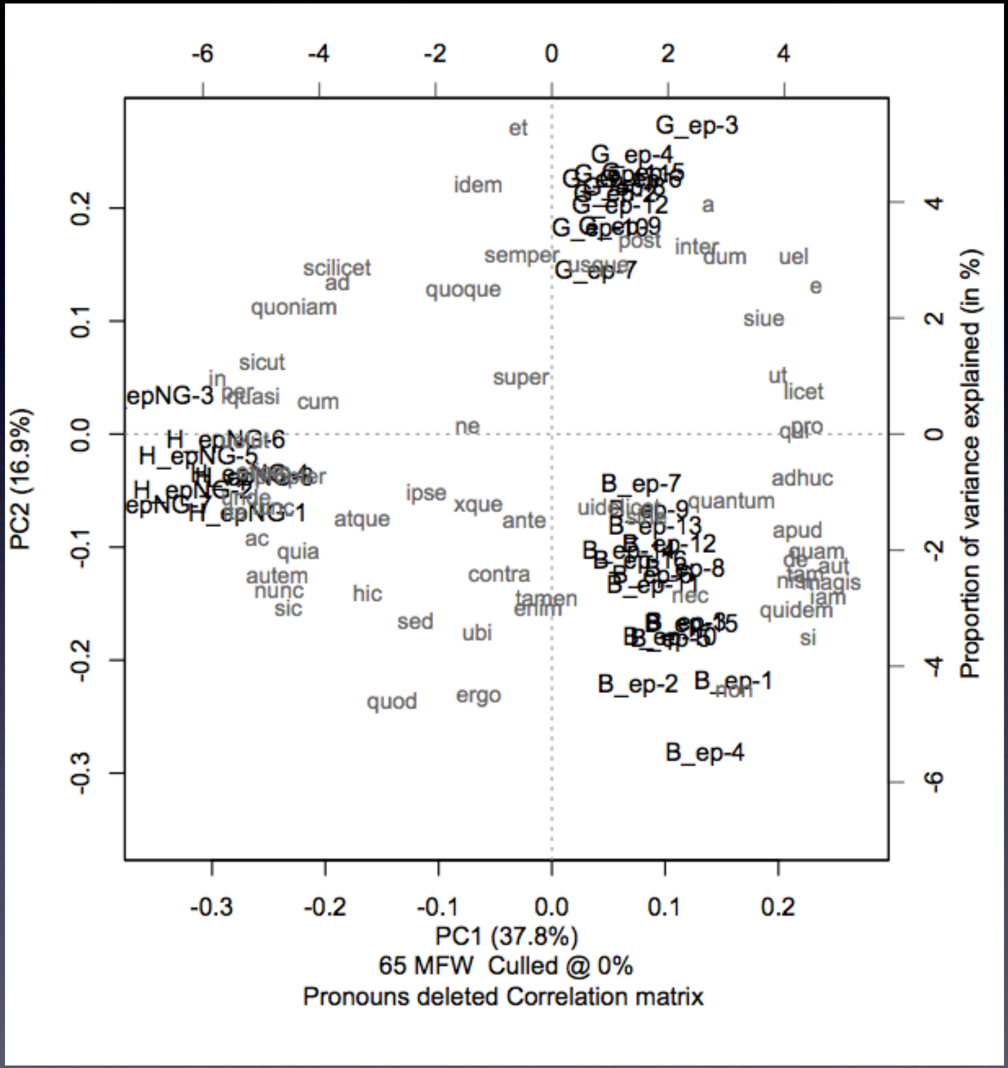
Wordlist

- Restrictive wordlist `wordlist_master.txt`
- Non-function words removed via hashtag (#)
- *Copy* `wordlist_master.txt` and rename copy to `wordlist.txt`
- Restrict analysis: tick `Use existing wordlist`
- Stylo will look for `wordlist.txt and use only these words`

Run PCA

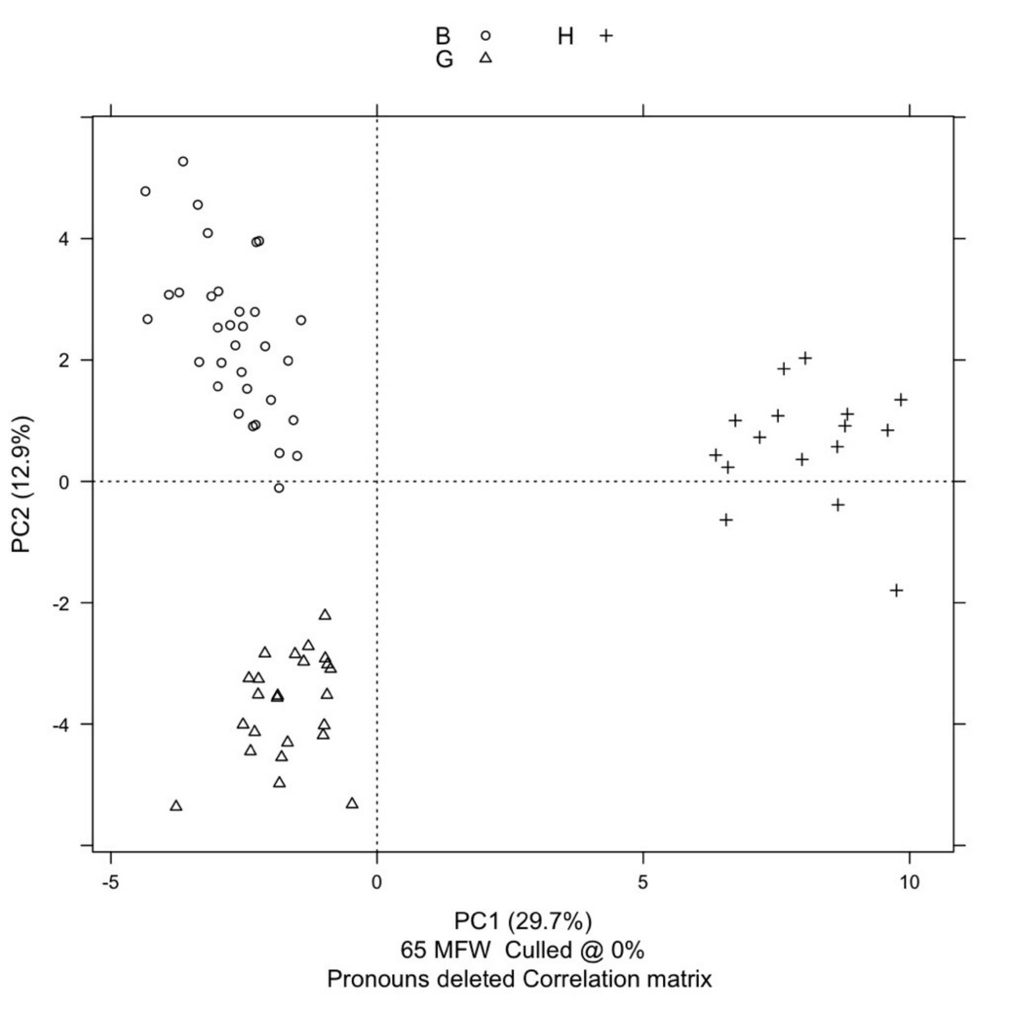
- `Sample size = 10,000`
- `MFW = 65`
- Select PCA
- `PCA Flavour = Technical`
- Select `B_ep.txt, G_ep.txt, H_epNG.txt`
- `Existing wordlist + Select Texts Manually`
- Same plot?

Test PCA

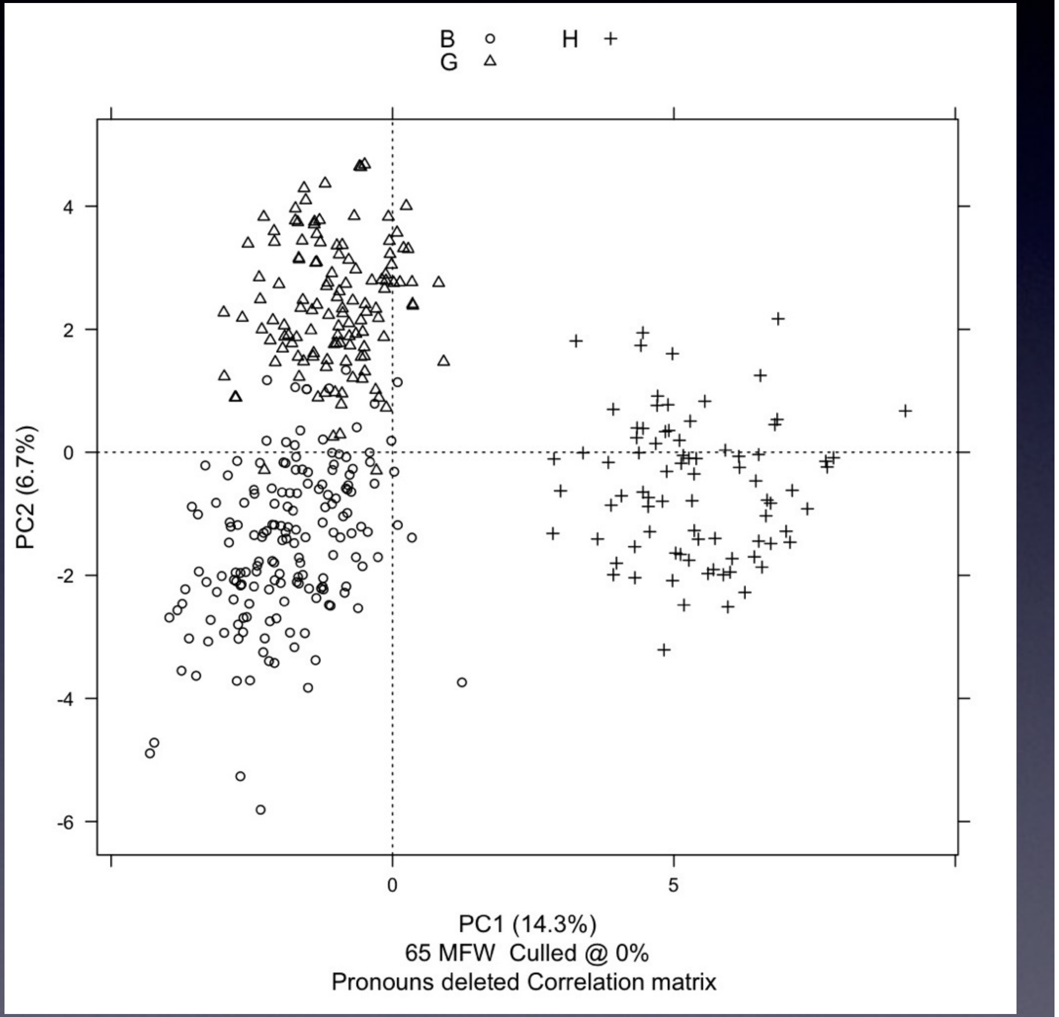


[ss=10,000; 65 MFW; content words ‘culled’]

Play with `Sample size`



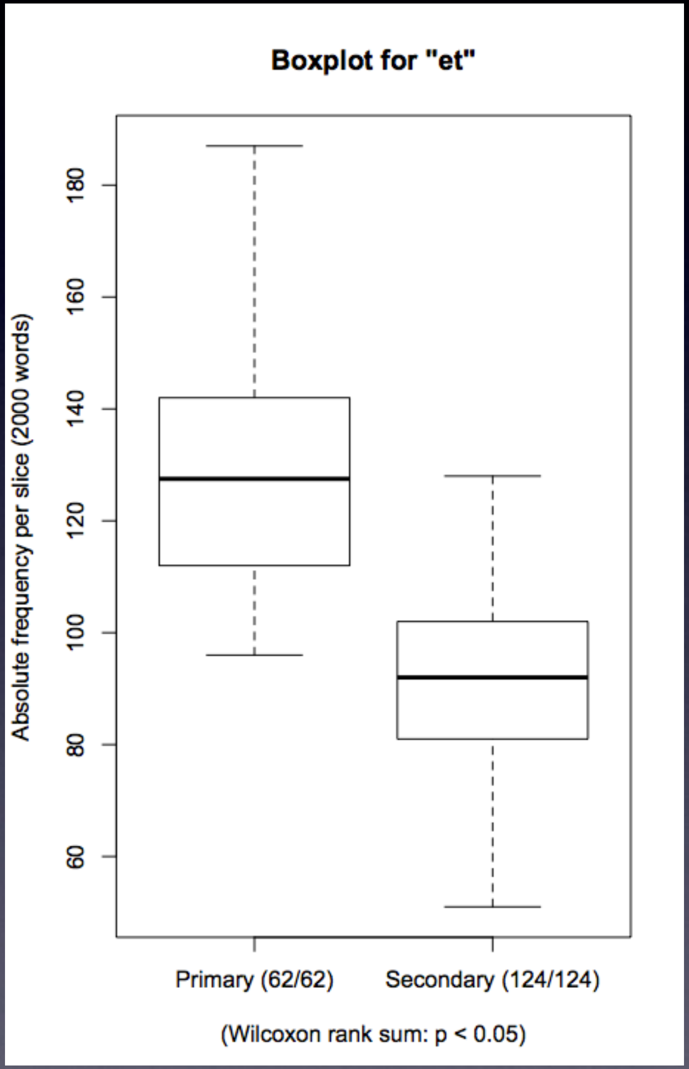
[ss=5,000]



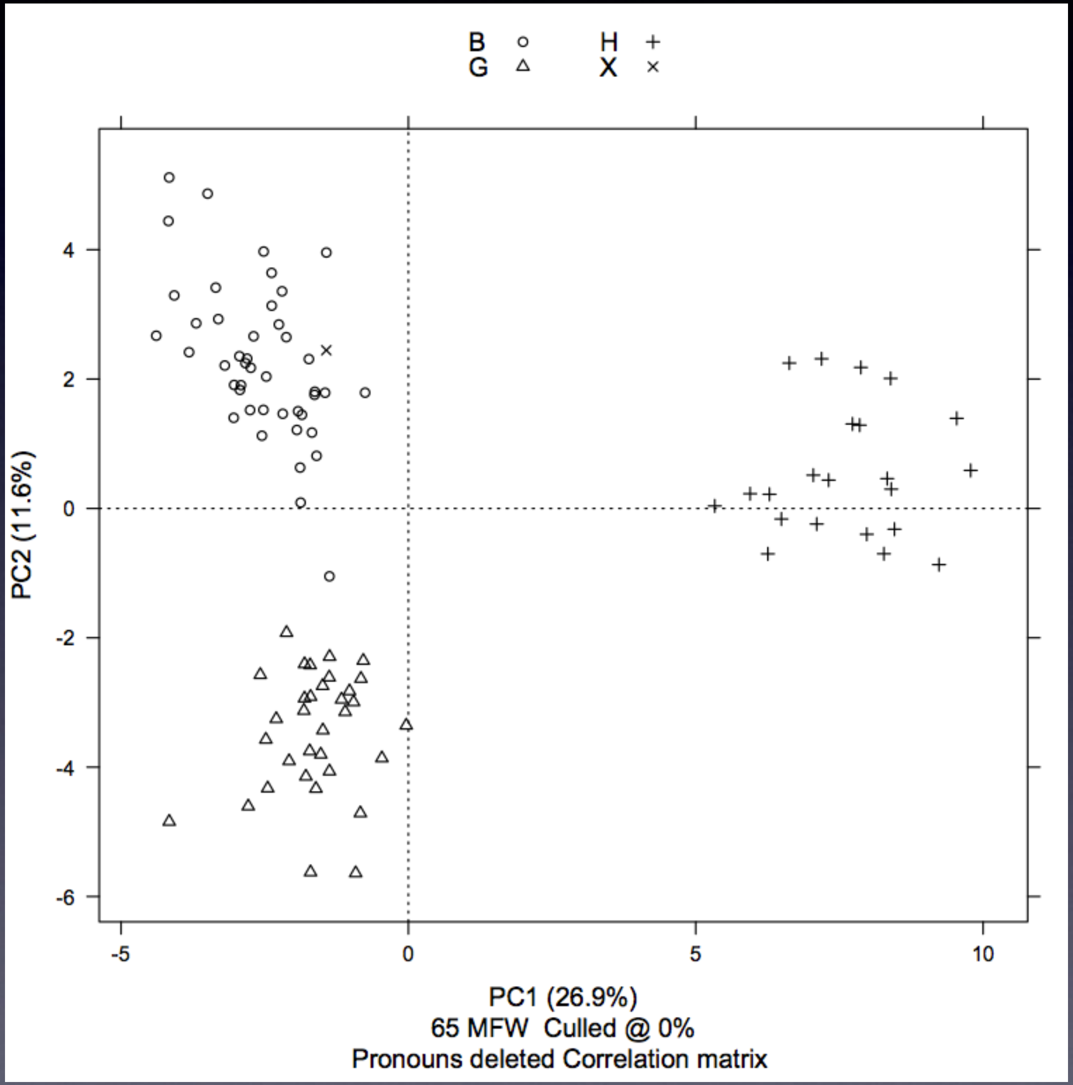
[ss=1,000]

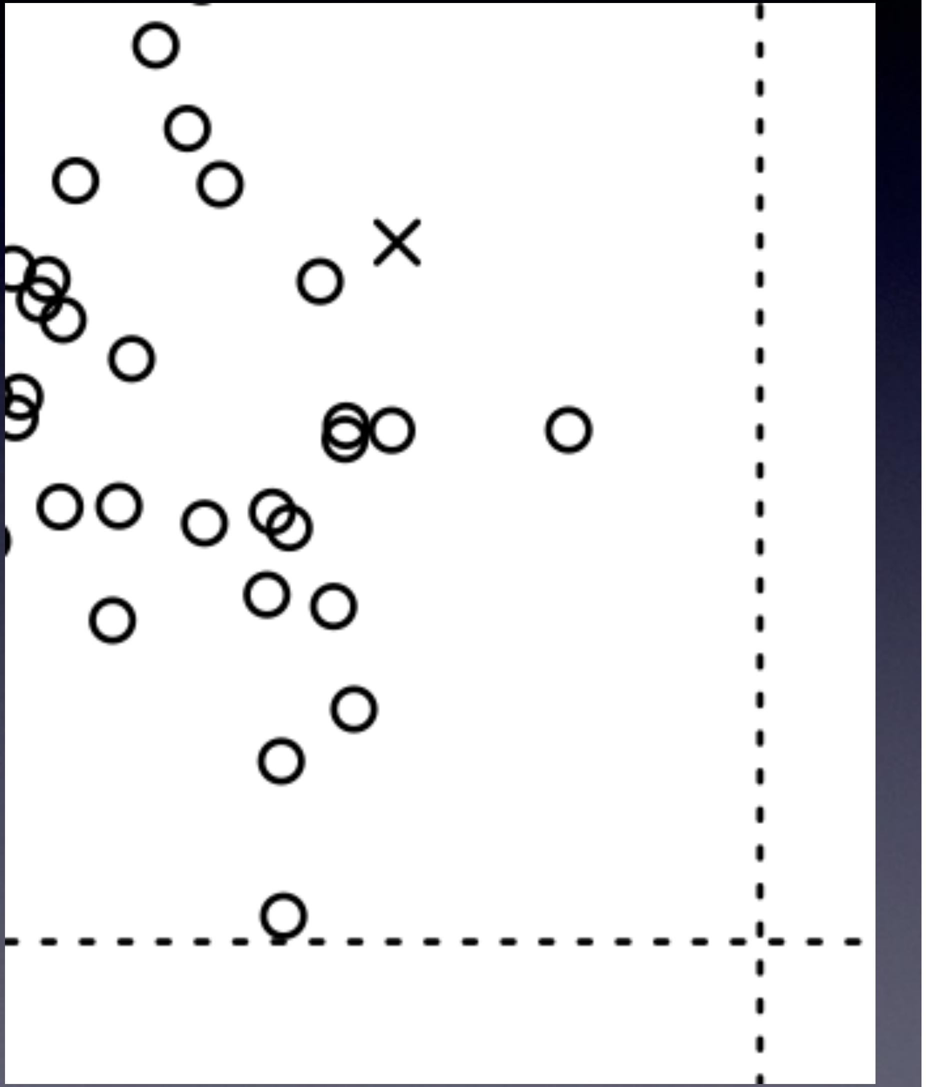
Boxplot

- Plot differences in `MFW`
- Use `oppose()`:
  - `primary_set = G_ep.txt`
  - `secondary_set = B_ep.txt + H_ep.txt`
- Also try:
  - *in* for Hildegard
  - *non* for Bernard



“Anonymous” text?



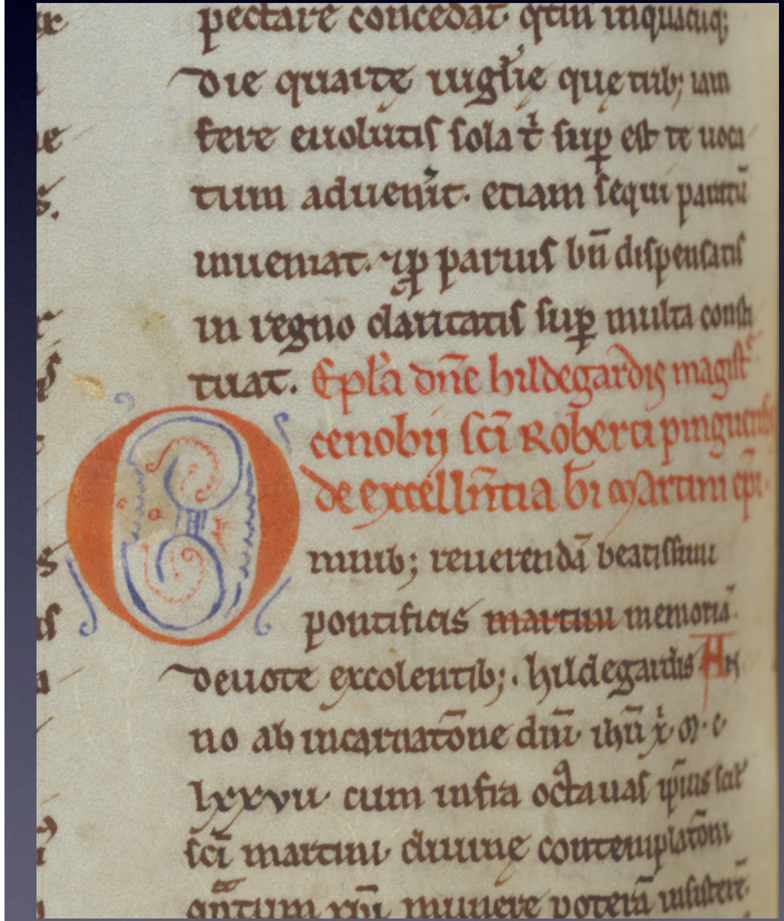


[ss=3,706; Bernard’s *Sermo in festo sancti Martini* as “anonymous test case”;]

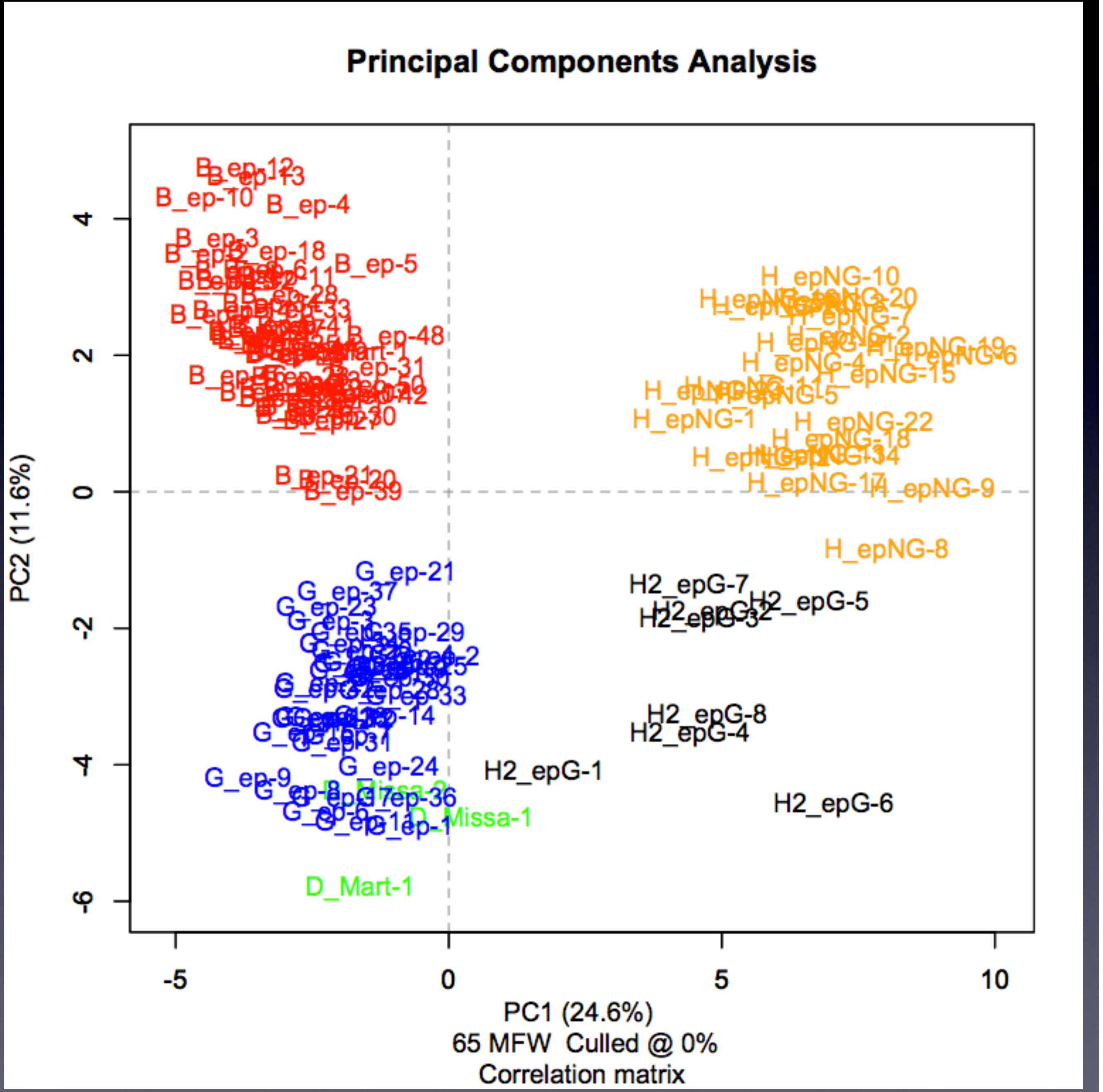
Add text

- Bernard’s *Sermo in festo sancti Martini*
- “Anonymous” test case
- Add `B_Mart.txt`
- Set `Sample size` to 3,500 (length of `B_Mart.txt`)
- Attribution?

Bigger picture



pectare concedat· quin inquacumq;
die quarte uigilie que trib; iam
fere euolutis sola t̄ sup est de noc-
tium aduenit· etiam sequi paucū
inueniat· prop paruus bū dispensatis
in regno claritatis sup multa consti-
tuat. Epl̄a dn̄e hildegardis magist
cenobij sc̄i Roberti pinguensis
de excellentia bi martini ep̄i·
Omnib; reuerendā beatissimi
pontificis ~~martini~~ memoriā
deuote excolentib;· hildegardis An-
no ab incarnatione dn̄i ihū x̄ m· c·
lxxvii· cum infra octauas ipsius sc̄i
sc̄i martini diuine contemplationi
ānitum x̄pi munere potentiā insistere·

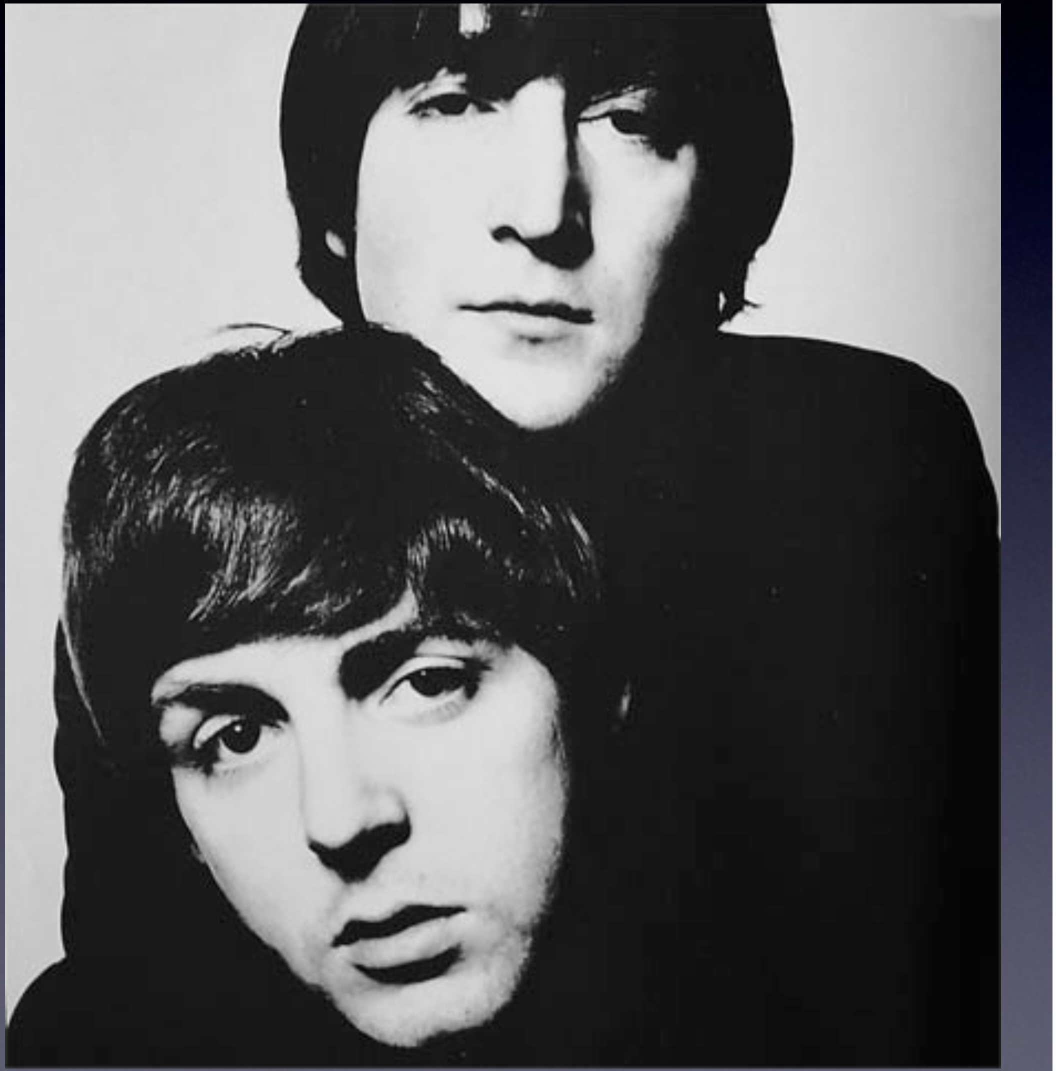


Concluding experiment

- Use all texts
  - `Sample size = 3,000`
- Don’t forget: `Existing wordlist`
  - `PCA Flavour = Classic`
- Stable? Try out different settings!

Synergy Hypothesis

- Pennebaker (e.g. 2011)
- *The Secret Life of Pronouns*
- Federalist papers and Beatles songs
- Collaborative writing style?
- “unlike either of one of the styles that the collaborating authors would produce on their own”
- Practical ànd theoretical relevance



“Hollywood version”? Online documentary

vimeo.com/70881172

References

- Argamon S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations, Literary and Linguistic Computing, 23(2): 131–47.
- Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing, 17(3): 267–87.
- Chrupała, G. et al. (2008). Learning Morphology with Morfette. In Proceedings of LREC 2008. Marrakech, Morocco: pp. 2362–67.
- Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: A Suite of Tools. In Digital Humanities 2013. Conference Abstracts. University of Nebraska-Lincoln, pp. 487–89.
- Ferrante, J. (1998). Scribe quae vides et audis. Hildegard, Her Language, and Her Secretaries. In Townsend, D. et al. (eds), The Tongue of the Fathers. Gender and Ideology in Twelfth-Century Latin. Philadelphia: University of Pennsylvania Press, pp. 102–35.
- Kestemont, M., Daelemans, W. and De Pauw, G. (2010). Weigh your Words – Memory-Based Lemmatization for Middle Dutch, Literary and Linguistic Computing, 25(3): 287–301.
- Newman, B. (1987). Sister of Wisdom. St. Hildegard's Theology of the Feminine. LA: University of California Press.
- Passarotti, M. and Dell'Orletta, F. (2010). Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin. In Calzolari, N. et al. (eds), Proceedings of LREC 2010. Valetta, Malta, pp. 1694-71.
- Pennebaker, J. (2011). The Secret Life of Pronouns. What our Words Say About Us. NY: Bloomsbury.
- Petrie, K., Pennebaker, J. and Sivertsen, B. (2008). Things We Said Today: A Linguistic Analysis of the Beatles, Psychology of Aesthetics, Creativity, and the Arts, 2(4), 197–202.