Data Analysis with R Philolab: stylometry applied to philology

Plan de Innovación y Mejora Docente

ID2018/152

2018-2019

Claudia García-Minguillán
University of Salamanca
2018-2019

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS
1218 ~ 2018

Stylometry applied to Philology

1. What is Stylometry?
2. Why? - Resource for Humanists
3. How to use it?
4. Research Project sample

Claudia García-Minguillán
University of Salamanca 2018-2019

1. What is Stylometry?

WHAT?

- Analyze to authorial writing style

HOW?

- R: programming language for statistical computing
- Word Frequency - statistics
- Attractive interface: graphical visualization
- DATA possible exportation

...and...WHY STYLO?

Claudia García-Minguillán
University of Salamanca 2018-2019

UNIVERSITY OF OXFORD

digital.humanities @ oxford

2. Why?- Resource for Humanists

- Digital Humanities
- Traditional research topics
- Scientism to your humanist job!
- Spread your Creativity

Claudia García-Minguillán
University of Salamanca 2018-2019

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS 1218 - 2018

New Test. Stud. 54, pp. 139–159. Printed in the United Kingdom © 2008 Cambridge University Press
DOI:10.1017/S0028688508000088

Recent Advances in Computational Linguistics and their Application to Biblical Studies*

J. JOSÉ ALVIAR
Faculty of Theology, University of Navarra, 31080 Pamplona, Spain

Soc. Sci. **2015**, *4*, 758–799; doi:10.3390/socsci4030758

OPEN ACCESS

social sciences

ISSN 2076-0760
www.mdpi.com/journal/socsci

Article

Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to "Shakespeare Authorship Question"

Refat Aljumily

School of English Literature, Language and Linguistics, University of Newcastle, Newcastle upon Tyne, Tyne and Wear NE1 7RU, UK; E-Mail: refat.aljumily@ncl.ac.uk; Tel.: +44-191-208-6233

Academic Editor: Martin J. Bull

Received: 1 July 2015 / Accepted: 6 September 2015 / Published: 17 September 2015

Did Shakespeare write his plays? - TED talk by Natalya St. Clair and Aaron Williams

2. Resource for Humanists

Claudia García-Minguillán
University of Salamanca 2018-2019

The Classical Quarterly 66.2 645–672 © The Classical Association (2017)
doi:10.1017/S0009838816000768

Justin Stover & Mike Kestemont (2017)

FIGURE 17: Bootstrap Consensus Tree (3000-word samples)

REASSESSING THE APULEIAN CORPUS: A COMPUTATIONAL APPROACH TO AUTHENTICITY

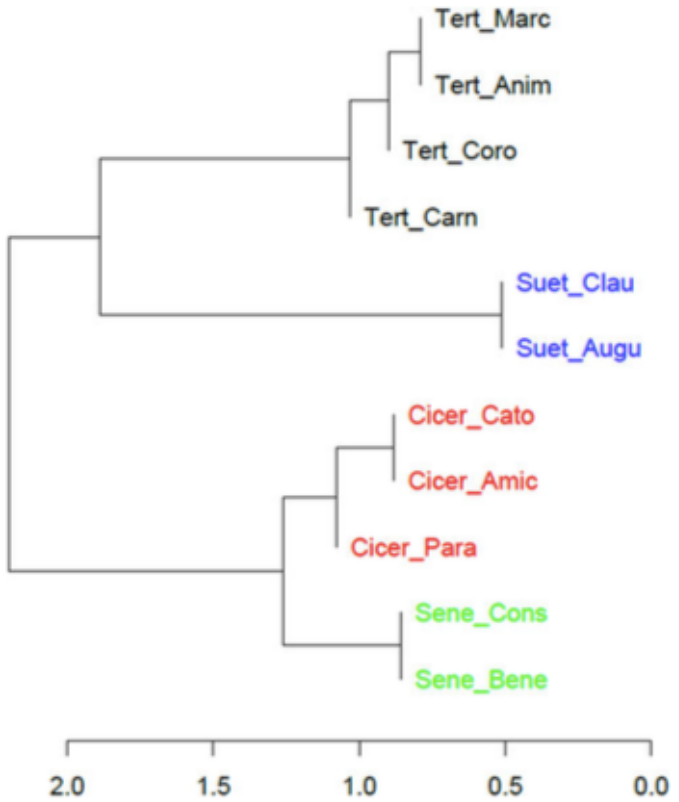


FIGURE 1: Cluster Analysis Dendrogram (no sampling), 100 MFW

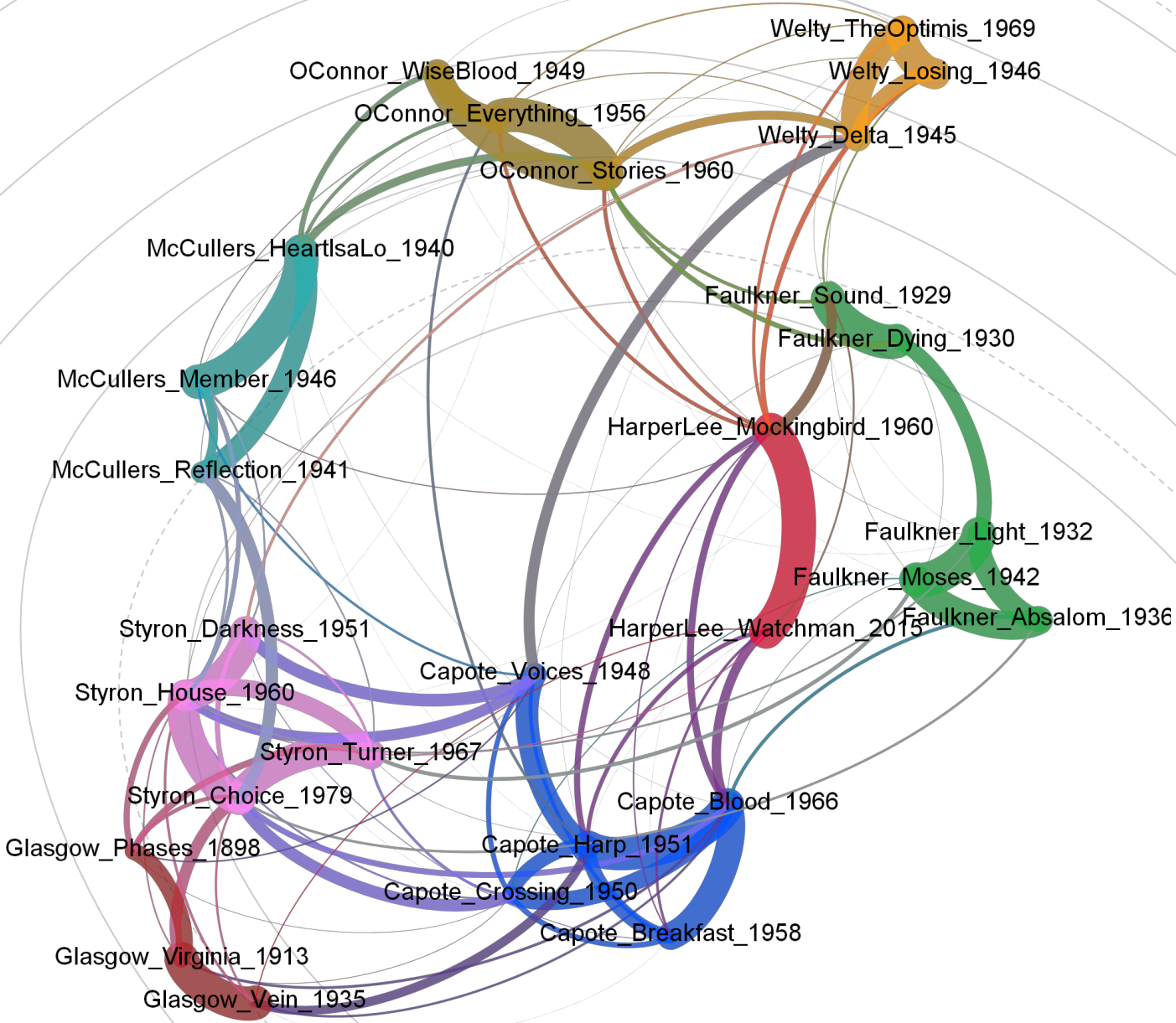
Claudia García-Minguillán
University of Salamanca 2018-2019

3. How to Stylo?

PhiloLab: Laboratory of Philology

- Authorship verification
- Genres verification: **BIG** dendrograms
- TOP Project!:
  'Chemical Linguistics: between linguistics and organic chemistry'
- Functions:
  - stylo()
  - rolling.classify ()
  - rolling.delta()
  - stylo.network()
  - oppose()
  - rolling.delta()

Claudia García-Minguillán
University of Salamanca
2018-2019

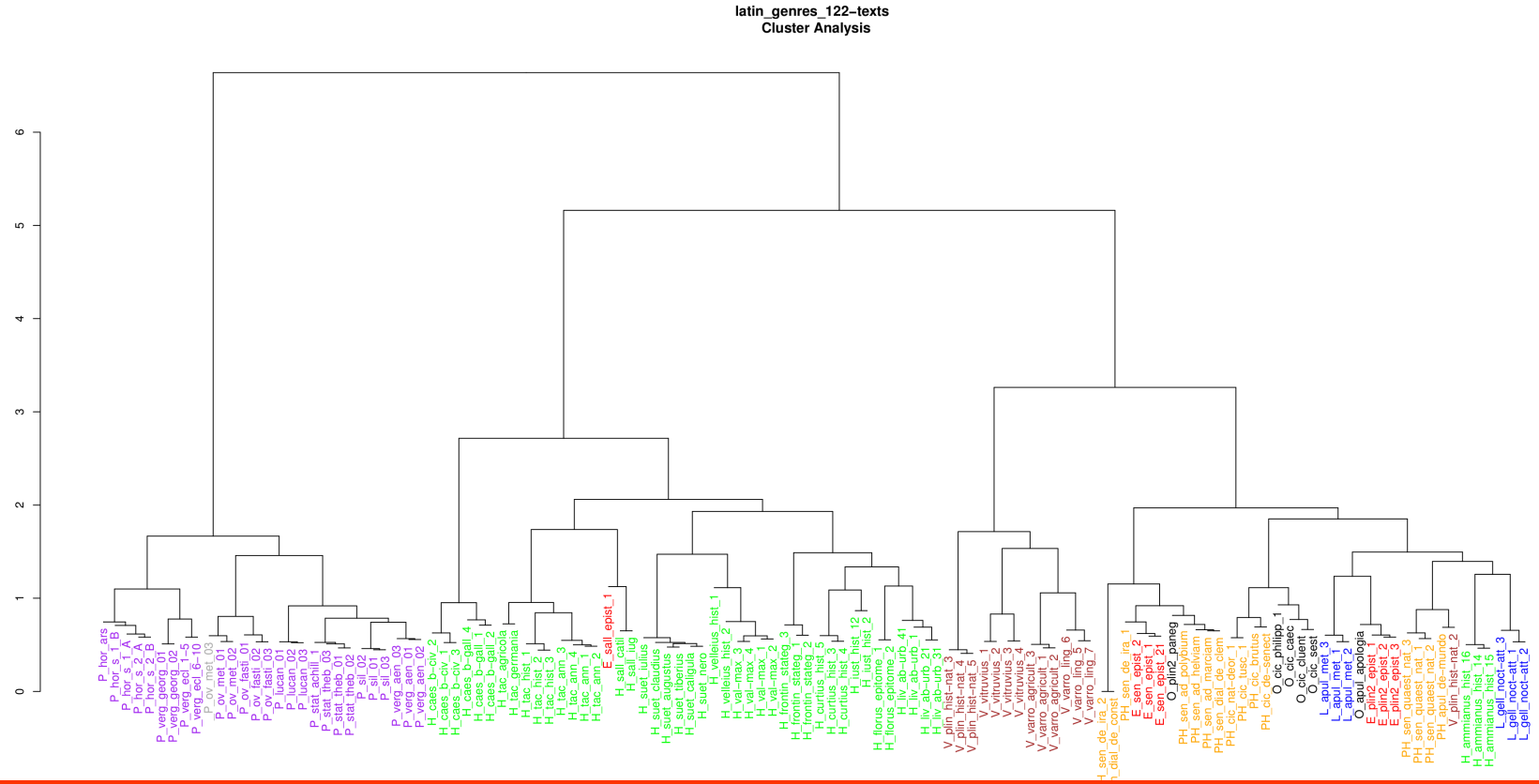


Authorship verification ~~'Go Set A Watchman while we Kill the Mockingbird in Cold Blood'~~

by

Maciej Eder and Jan Rybicki

Claudia García-Minguillán
University of Salamanca
2018-2019

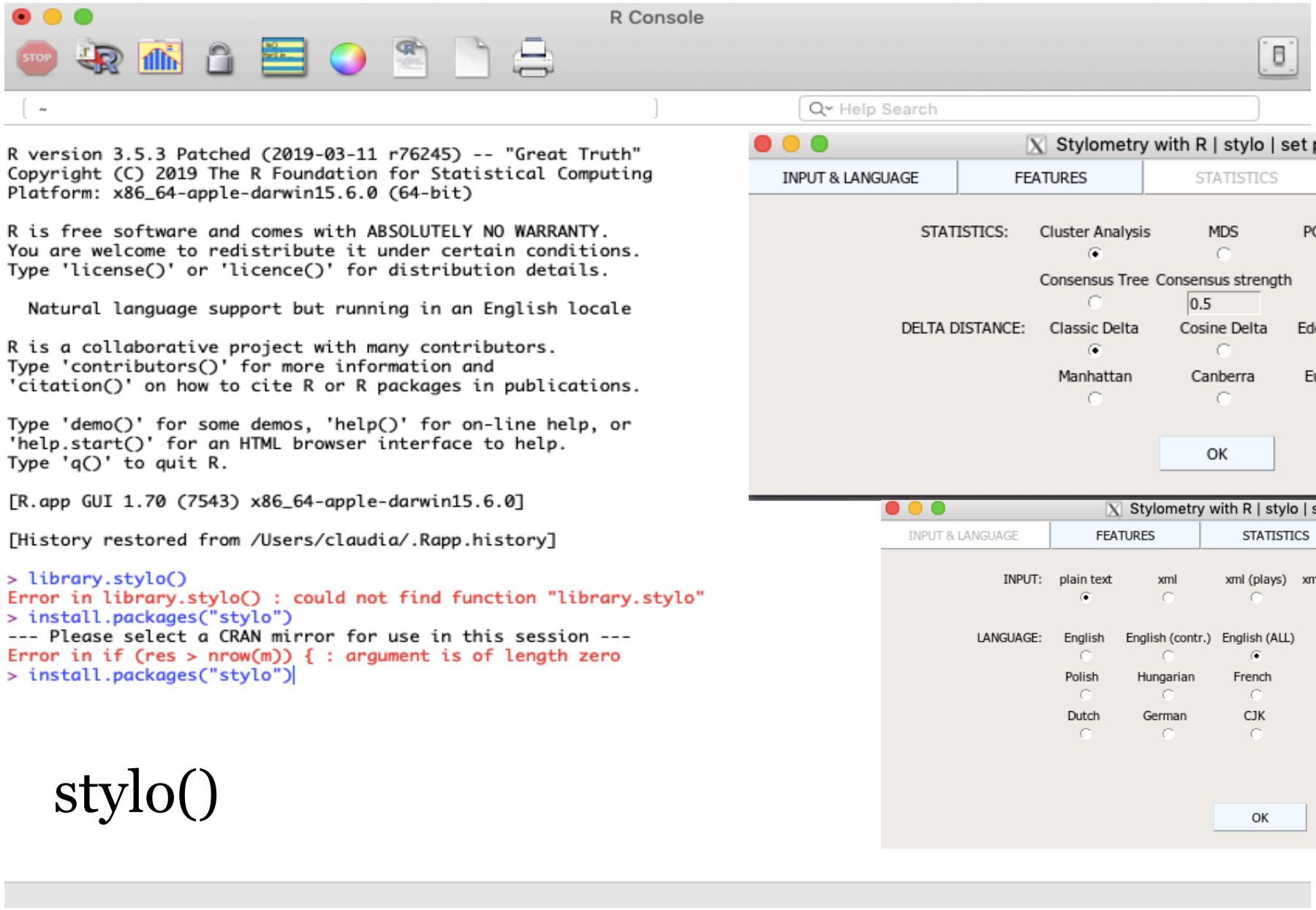


testing BIG dendrograms
by Computational Stylistics Group

Claudia García-Minguillán
University of Salamanca

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

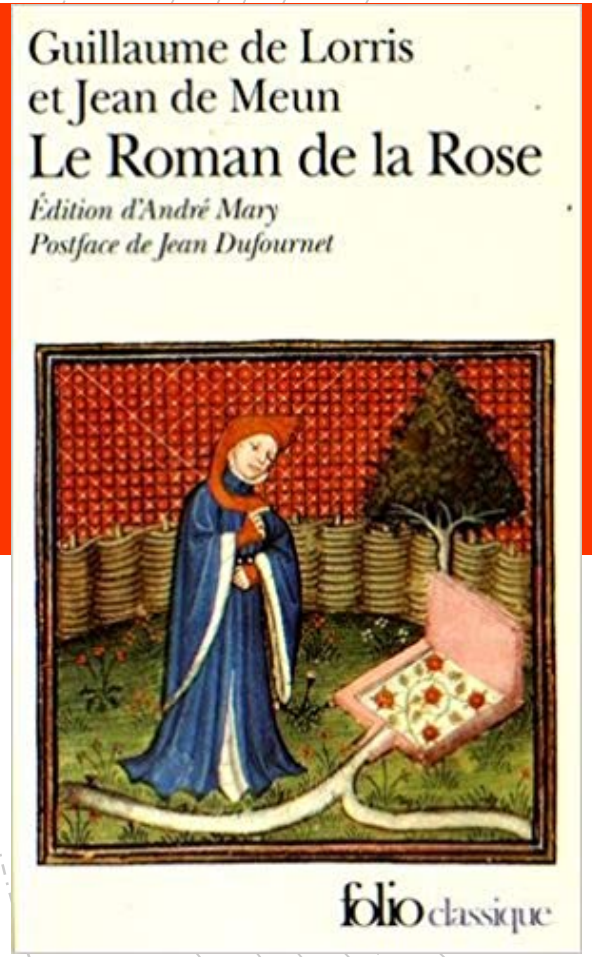
800 AÑOS
1218 - 2018



stylo()

Claudia García-Minguillán
University of Salamanca
2018-2019

rolling.classify ()



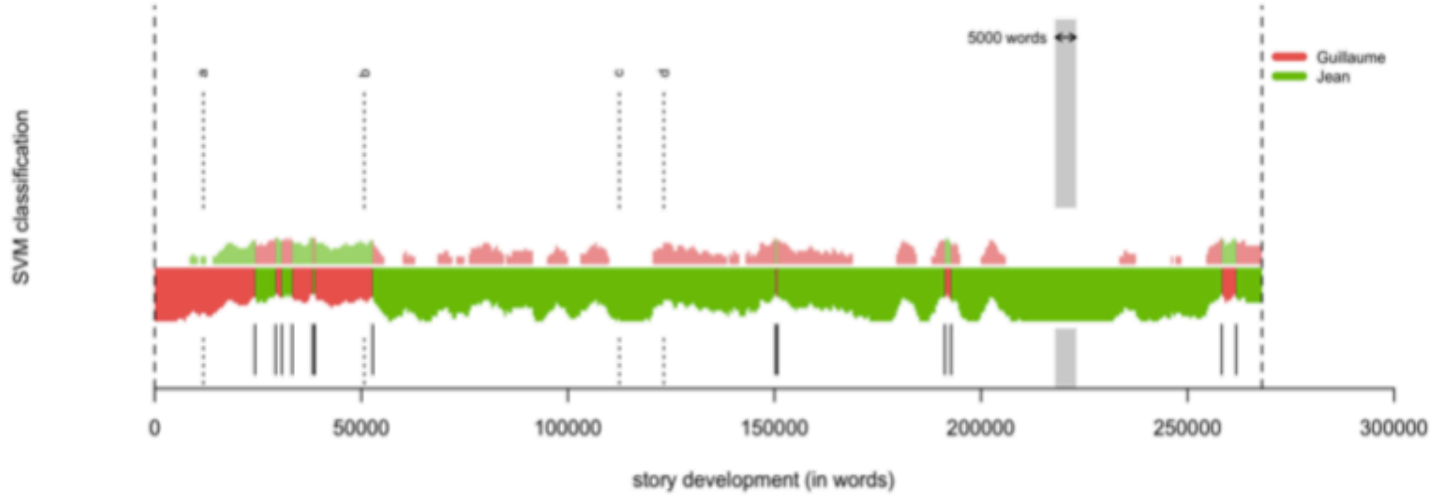


Figure 4: The Rolling Stylometry visualization. The medieval French allegoric story *Roman de la Rose* assessed using Rolling SVM and 100 MFWs; window size: 5,000 words, sample overlap: 4,500 words. Sections attributed to Guillaume de Lorris are marked red, those attributed to Jean de Meun are green. The level of certainty of the classification is indicated by the thickness of the bottom stripe. The commonly-accepted division into two authorial parts is marked with a vertical dashed line 'b'.

Source ☞
Computational Stylistics Group

Claudia García-Minguillán
University of Salamanca
2018-2019

```
'help.start()' for an HTML browse
Type 'q()' to quit R.

[R.app GUI 1.70 (7543) x86_64-app

[History restored from /Users/cla

> help()
starting httpd help server ... do
> stylo()
Error in stylo() : could not find
>  install.packages("tcltk2")
--- Please select a CRAN mirror f
trying URL 'https://cran.rediris.
Content type 'application/x-gzip'
==================================
downloaded 946 KB


The downloaded binary packages ar
    /var/folders/ty/h9hk3p355c1gx
>     install.packages("ape")
trying URL 'https://cran.rediris.
Content type 'application/x-gzip'
==================================
downloaded 2.5 MB


The downloaded binary packages ar
    /var/folders/ty/h9hk3p355c1gx
>     install.packages("class")
trying URL 'https://cran.rediris.
Content type 'application/x-gzip'
==================================
downloaded 92 KB


The downloaded binary packages are in
    /var/folders/ty/h9hk3p355c1gxfb0m4y4mmlc0000gn/T//RtmphAXumk/downloaded_packages
>     install.packages("e1071")
trying URL 'https://cran.rediris.es/bin/macosx/el-capitan/contrib/3.5/e1071_1.7-1.tgz'
Content type 'application/x-gzip' length 896841 bytes (875 KB)
==================================================
downloaded 875 KB


The downloaded binary packages are in
    /var/folders/ty/h9hk3p355c1gxfb0m4y4mmlc0000gn/T//RtmphAXumk/downloaded_packages
>     install.packages("pamr")
trying URL 'https://cran.rediris.es/bin/macosx/el-capitan/contrib/3.5/pamr_1.56.tgz'
Content type 'application/x-gzip' length 817907 bytes (798 KB)
==================================================
downloaded 798 KB


The downloaded binary packages are in
    /var/folders/ty/h9hk3p355c1gxfb0m4y4mmlc0000gn/T//RtmphAXumk/downloaded_packages
>     install.packages("tsne")
```

```
    /var/folders/ty/h9hk3p355c1gxfb0m4y4mmlc0000gn/T//RtmpD6MIHp/downloaded_packages
>     install.packages("tsne")
trying URL 'https://cran.rediris.es/bin/macosx/el-capitan/contrib/3.5/tsne_0.1-3.tgz'
Content type 'application/x-gzip' length 21437 bytes (20 KB)
==================================================
downloaded 20 KB

The downloaded binary packages are in
    /var/folders/ty/h9hk3p355c1gxfb0m4y4mmlc0000gn/T//RtmpD6MIHp/downloaded_packages
> stylo()
Error in stylo() : could not find function "stylo"
> stylo()
Error in stylo() : could not find function "stylo"
> library.stylo()
Error in library.stylo() : could not find function "library.stylo"
>   library(stylo)

### stylo version: 0.6.9 ###

If you plan to cite this software (please do!), use the following reference:
    Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R:
    a package for computational text analysis. R Journal 8(1): 107-121.
    <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>

To get full BibTeX entry, type: citation("stylo")
Warning message:
In system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE) :
  running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/library/tcltk/libs//tcltk.so'' had status 1
sty
> stylo()
using current directory...

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
Hey! The working directory should contain the subdirectory "corpus"
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!


Error in stylo() : Corpus prepared incorrectly
>
```

help( )

Claudia García-Minguillán
University of Salamanca
2018-2019

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS
1218 ~ 2018

4. Research Project Sample

'The many faces of an author's style'

...Forthcoming publication

Claudia García-Minguillán
University of Salamanca
2018-2019

How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling

Author of the *Harry Potter* books has a distinct linguistic signature

Hypothesis

Claudia García-Minguillán
University of Salamanca
2018-2019

Claudia García-Minguillán
Universidad de Salamanca 2018-2019

Fernando_Pessoa ()

- Heteronimia: Fernando Pessoa, Ricardo Reis, Bernardo Soares, Alvaro de Campos, Alberto Caeiro
- Aims:
- 1) Hypothesis
- 2) Testing author's style
- 3) Testing Stylo

Claudia García-Minguillán
Universidad de Salamanca 2018-2019

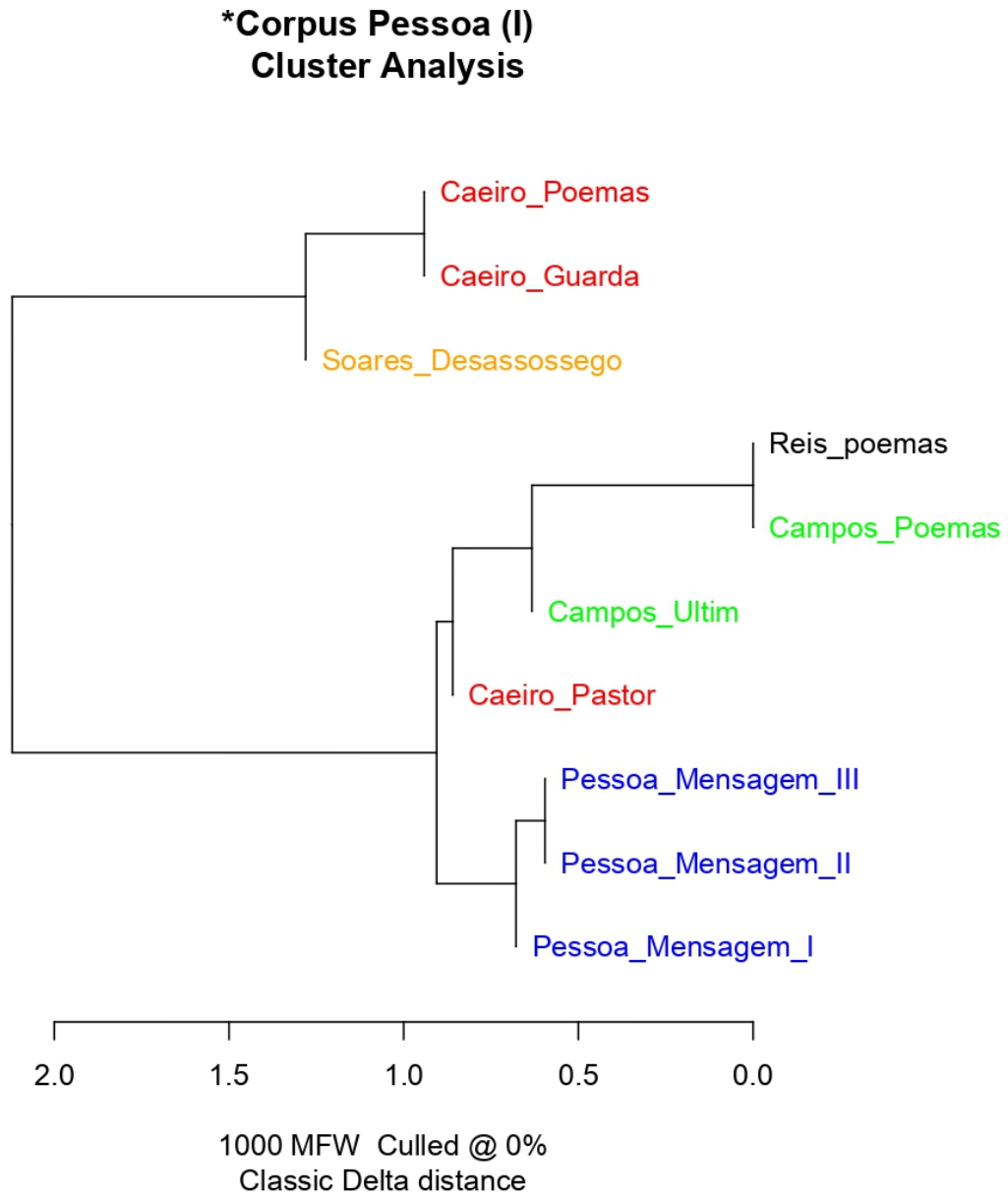
Claudia García-Minguillán
University of Salamanca
2018-2019

Mind your corpus

1. Open Access Pessoa's work	Gutenberg Projects OCR Corpora– think about sharing it at GitHub!
2. **Read** and prepare it	Collatio – 'Crítica Textual' Txt. Name your corpus
3. Remember the steps!	Parametres
4. Good practices	Be good and honest

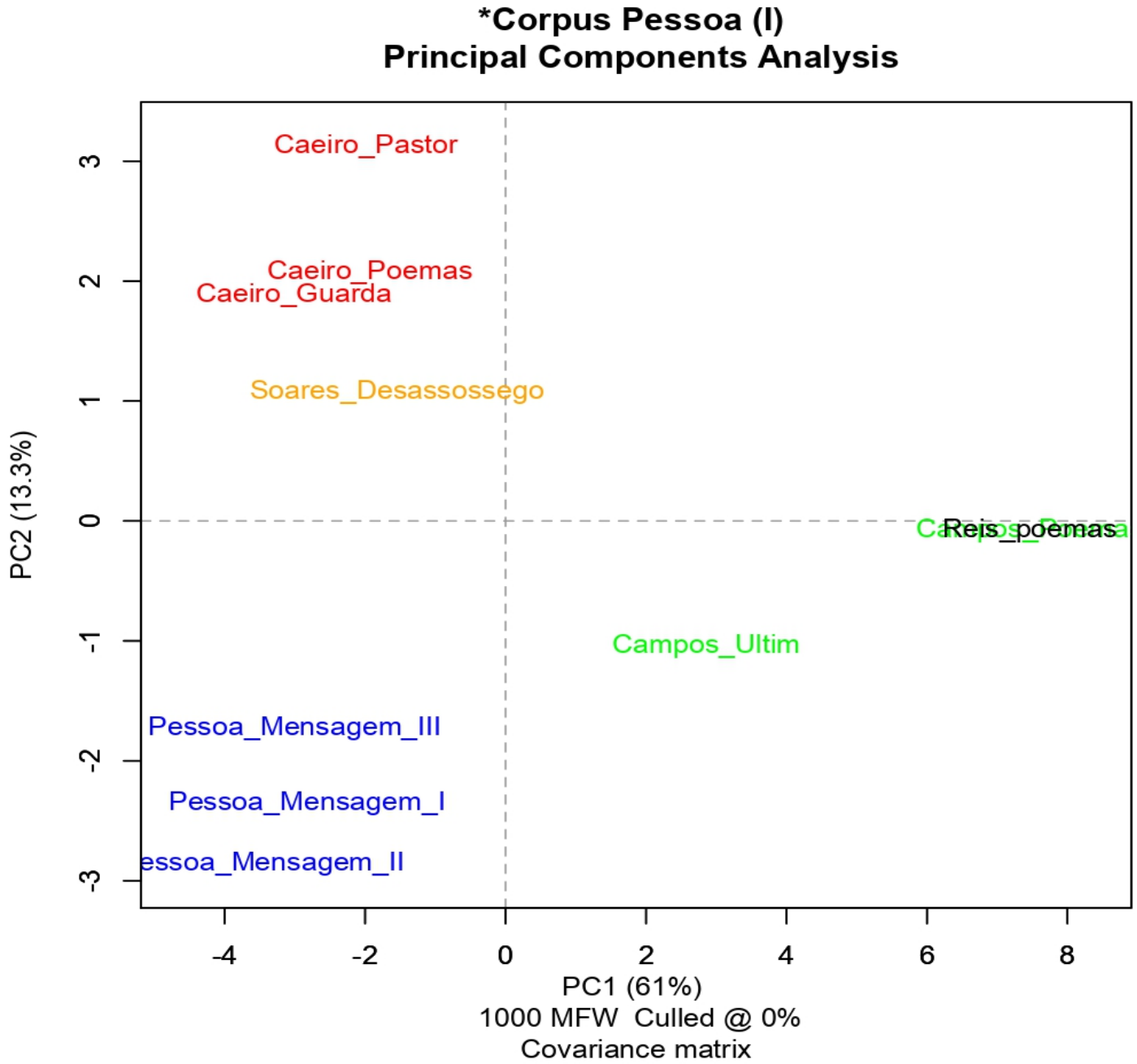

Caeiro_Guarda.txt
Caeiro_Pastor.txt
Caeiro_Poemas.txt
Campos_Poemas.txt
Campos_Ultim.txt
Pessoa_Mensagem_I.txt
Pessoa_Mensagem_II.txt
Pessoa_Mensagem_III.txt
Reis_poemas.txt
Soares_Desassossego.txt


Antologia Poética - Fernando Pessoa.txt
Cancioneiro - Fernando Pessoa.txt
Do Livro do Desassossego - Bernardo Soares.txt
Ficções do interlúdio, para além do outro oc...o de Coelho Pacheco - Fernando Pessoa.txt
Mensagem - Fernando Pessoa.txt
O Banqueiro Anarquista - Fernando Pessoa.txt
O Eu profundo e os outros Eus - Fernando Pessoa.txt
O Guardador de Rebanhos - Alberto Caeiro.txt
O Marinheiro - Fernando Pessoa.txt
O pastor amoroso - Alberto Caeiro.txt
Obra Completa - Fernando Pessoa.txt
Poemas de Álvaro de Campos - Fernando Pessoa.txt
Poemas de Fernando Pessoa.txt
Poemas de Ricardo Reis - Fernando Pessoa.txt
Poesias Inéditas - Fernando Pessoa.txt
Primeiro Fausto - Fernando Pessoa.txt



- Caeiro_Guarda.txt
- Caeiro_Pastor.txt
- Caeiro_Poemas.txt
- Campos_Poemas.txt
- Campos_Ultim.txt
- Pessoa_Mensagem_I.txt
- Pessoa_Mensagem_II.txt
- Pessoa_Mensagem_III.txt
- Reis_poemas.txt
- Soares_Desassossego.txt

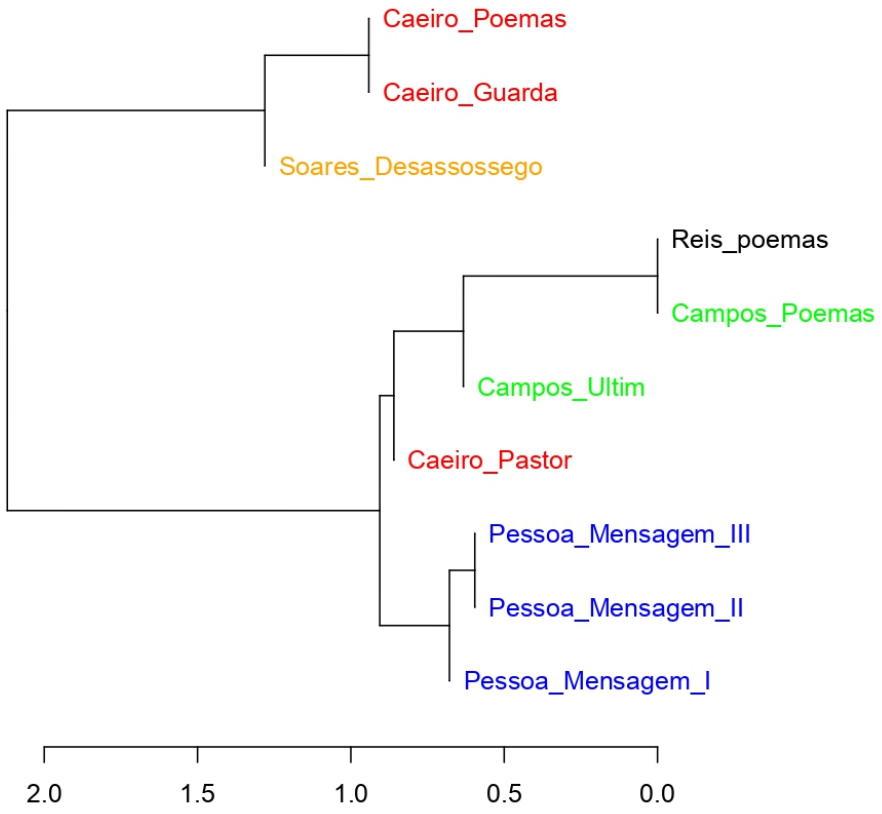
Claudia García-Minguillán
University of Salamanca
2018-2019



Claudia García-Minguillán
University of Salamanca
2018-2019

Claudia García-Minguillán
University of Salamanca
2018-2019

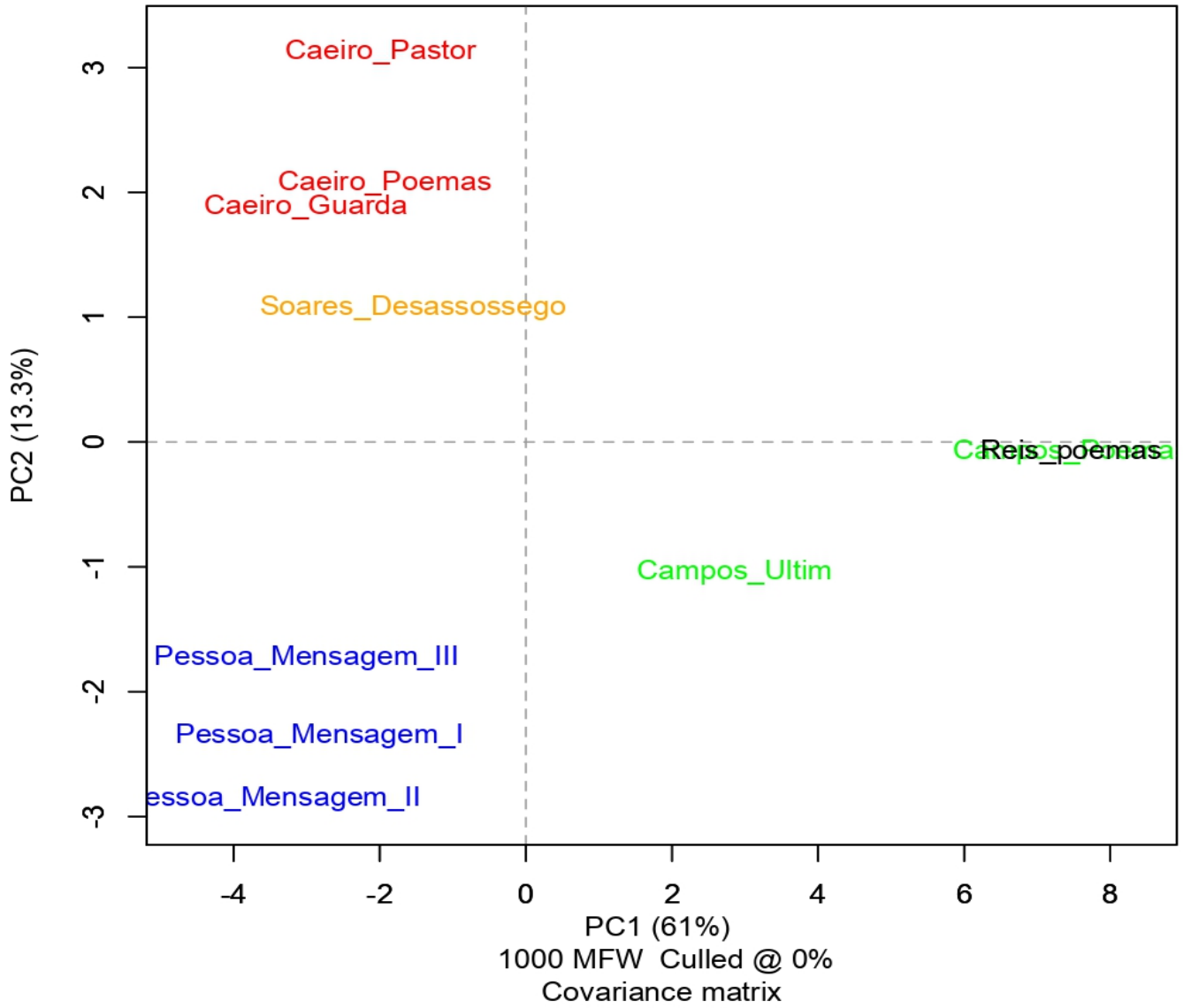
***Corpus Pessoa (I)**
Cluster Analysis



1000 MFW Culled @ 0%
Classic Delta distance

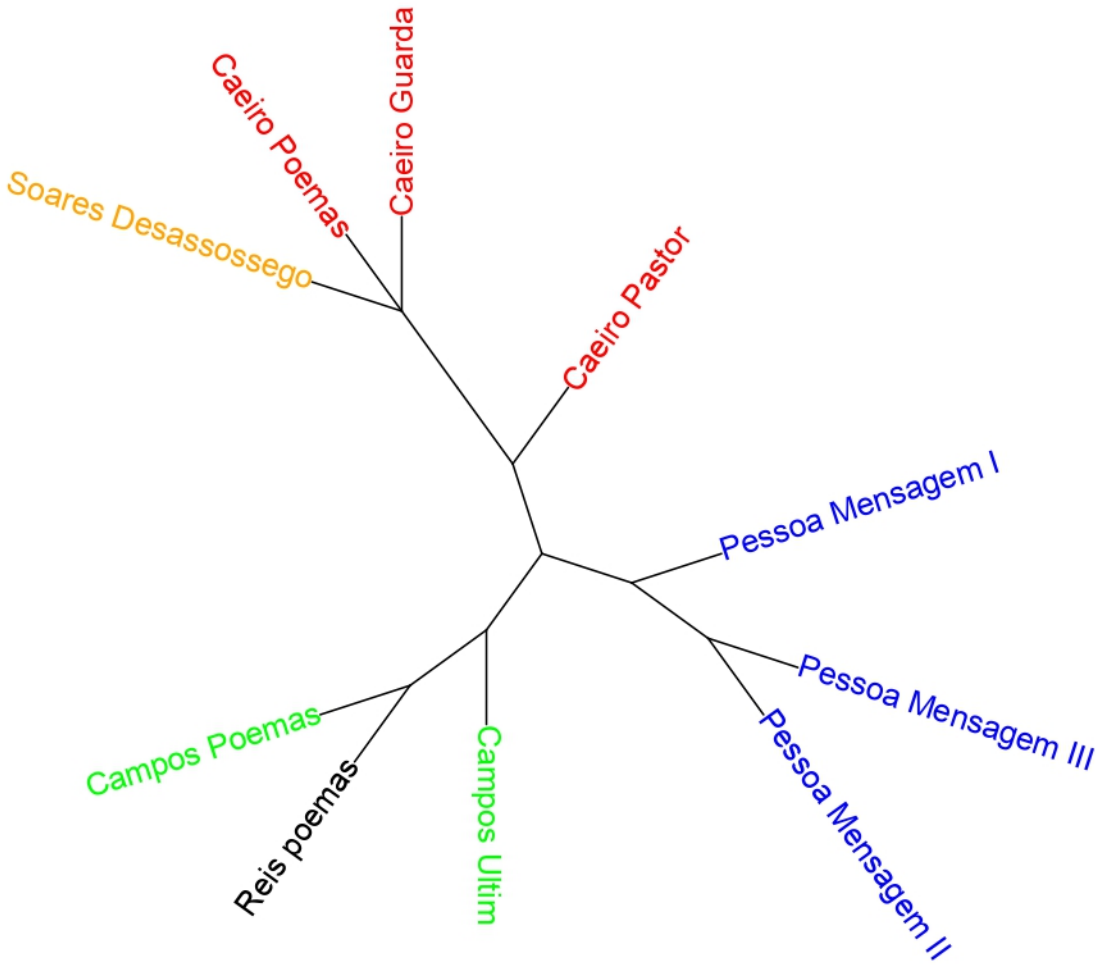
Claudia García-Minguillán
University of Salamanca
2018-2019

***Corpus Pessoa (I)**
Principal Components Analysis



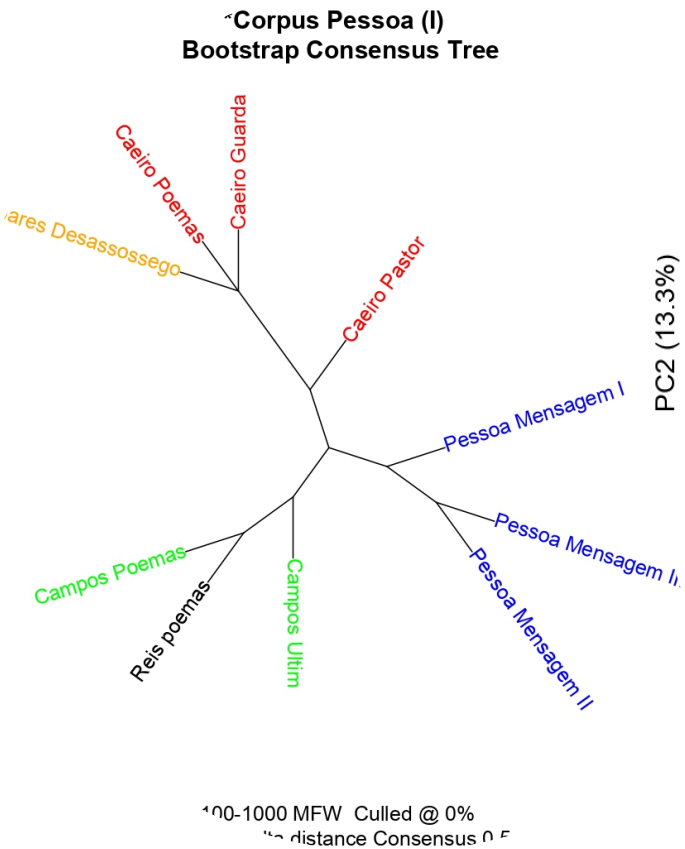
1000 MFW Culled @ 0%
Covariance matrix

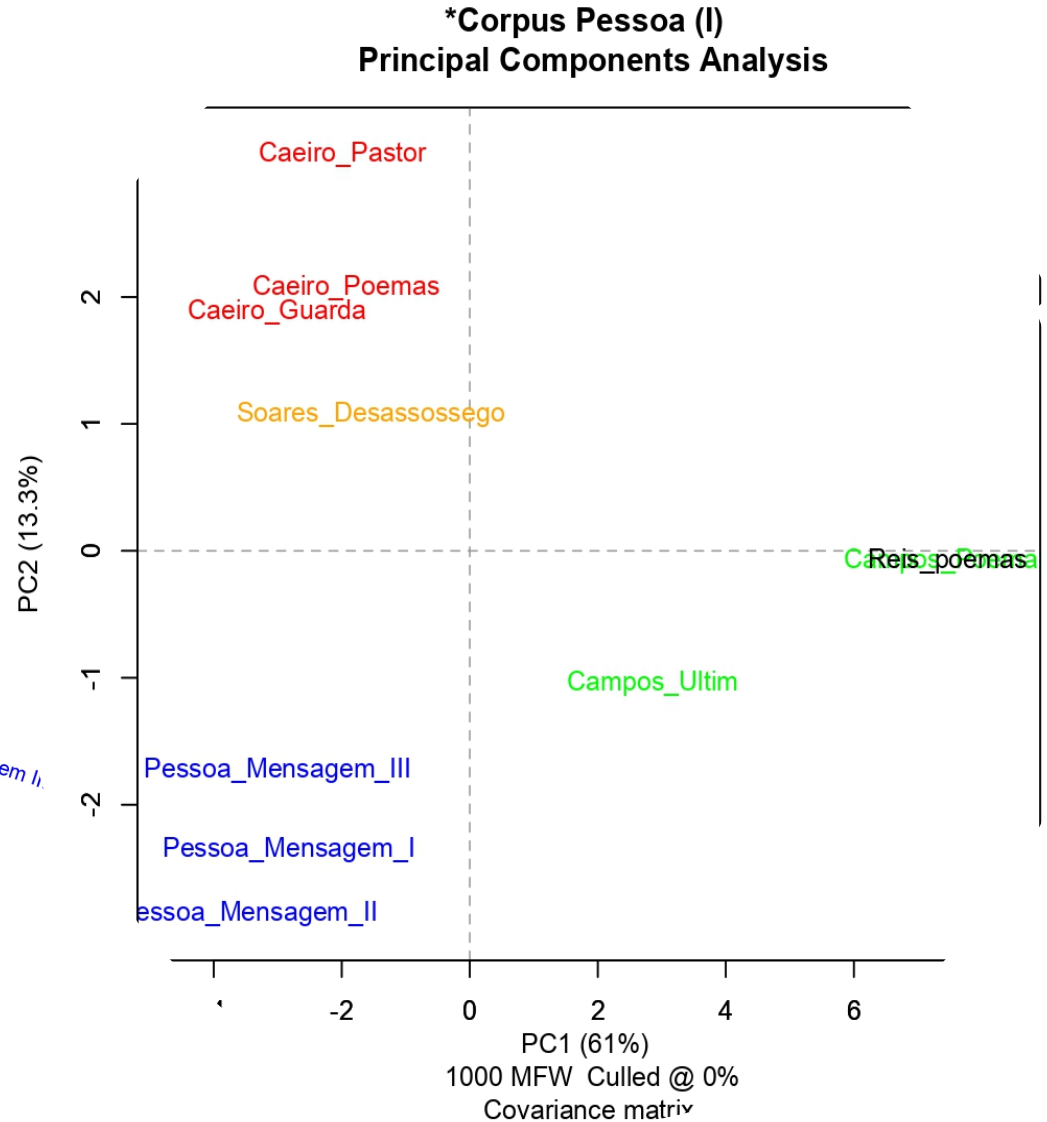
***Corpus Pessoa (I)**
Bootstrap Consensus Tree

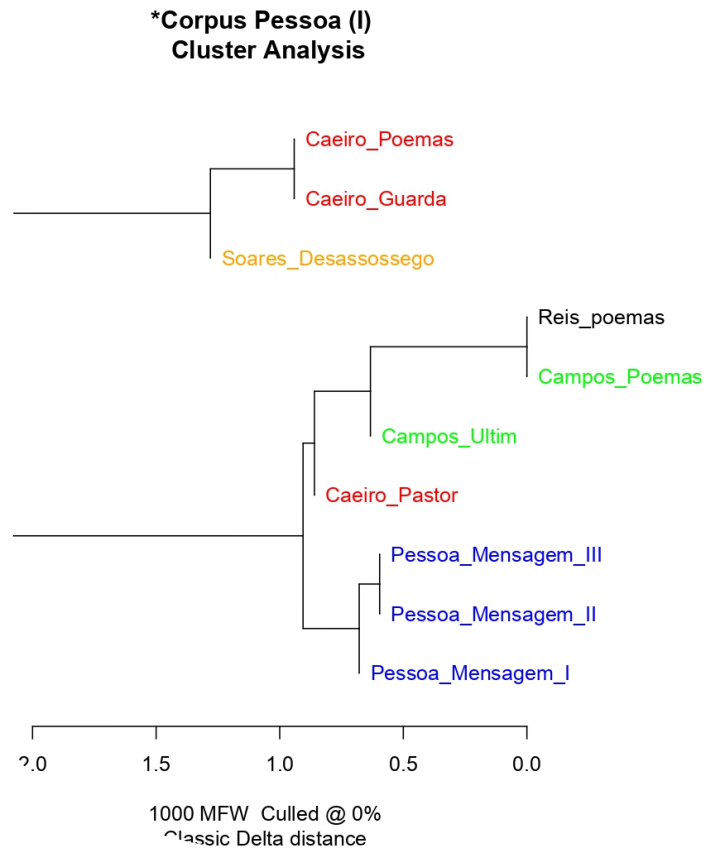


100-1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

Claudia García-Minguillán
University of Salamanca
2018-2019



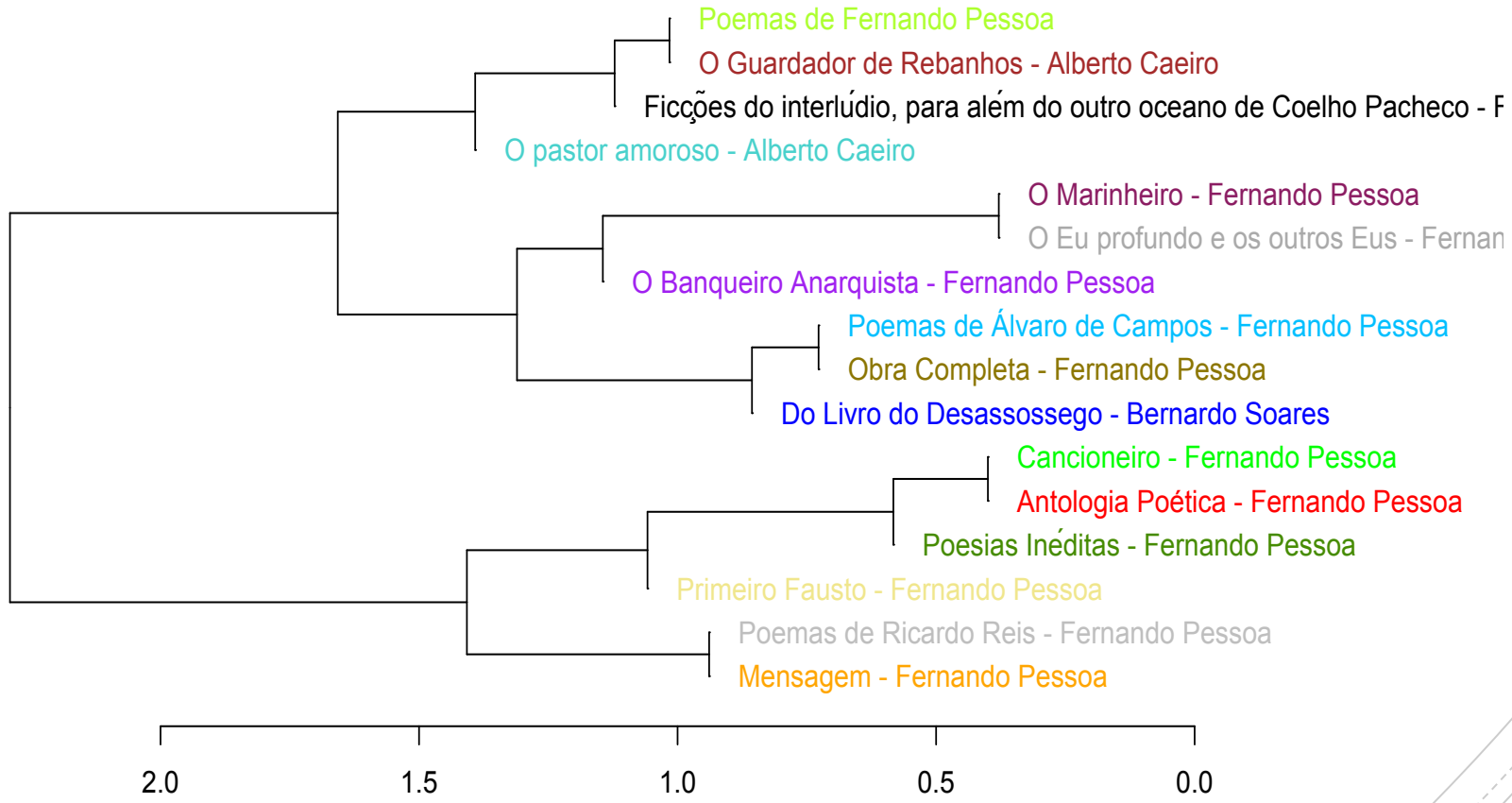




Claudia García-Minguillán
University of Salamanca
2018-2019

- Antologia Poética_Pessoa.txt
- Caeiro_Guarda.txt
- Caeiro_Pastor.txt
- Caeiro_Poemas.txt
- Campos_Poemas.txt
- Campos_Ultim.txt
- Cancioneiro_Pessoa.txt
- Ficções_Pessoa.txt
- O Banqueiro Anarquista_Pessoa.txt
- O Eu profundo e os outros Eus_Pessoa.txt
- O Marinheiro_Pessoa.txt
- Pessoa_Mensagem_I.txt
- Pessoa_Mensagem_II.txt
- Pessoa_Mensagem_III.txt
- Poesias Inéditas_Pessoa.txt
- Primeiro_Fausto_Pessoa.txt
- Reis_poemas.txt
- Soares_Desas.txt

Pessoa Corpus
Cluster Analysis

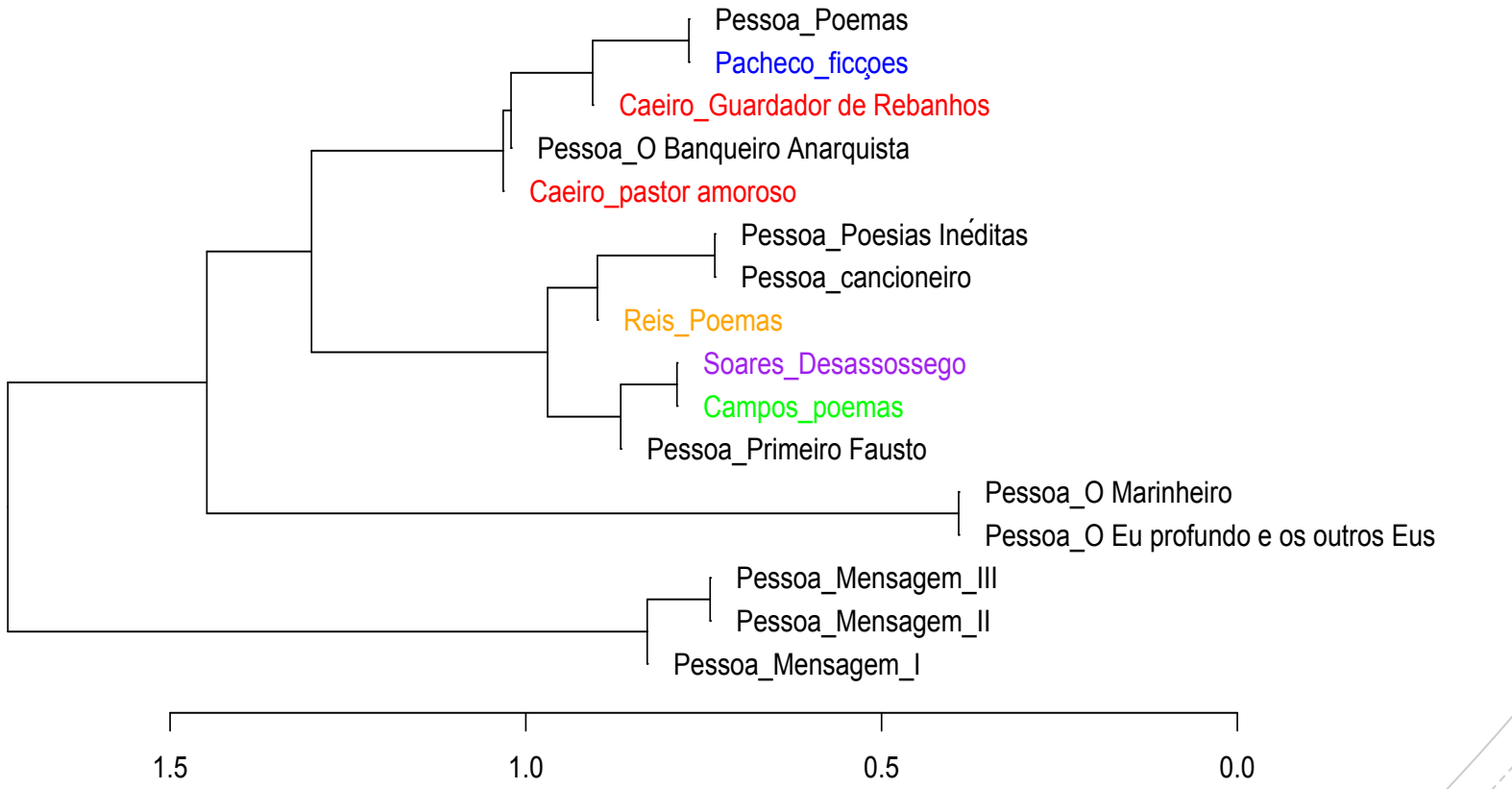


100 MFW Culled @ 0%
Classic Delta distance

Claudia García-Minguillán
University of Salamanca
2018-2019

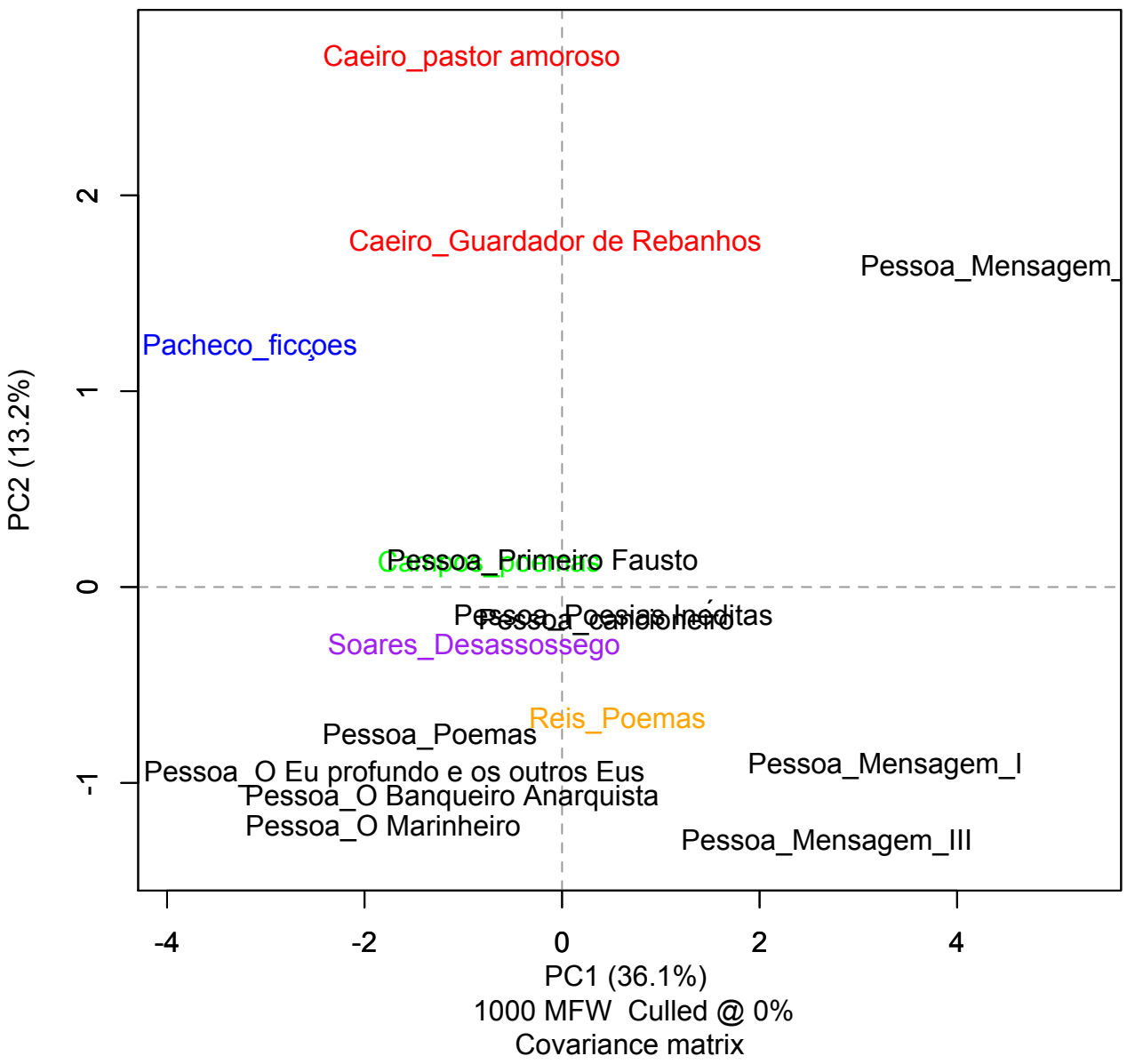
- Caeiro_Guardador de Rebanhos.txt
- Caeiro_pastor amoroso.txt
- Campos_poemas.txt
- Pacheco_ficçoes.txt
- Pessoa_cancioneiro.txt
- Pessoa_Mensagem_I.txt
- Pessoa_Mensagem_II.txt
- Pessoa_Mensagem_III.txt
- Pessoa_O Banqueiro Anarquista.txt
- Pessoa_O Eu profundo e os outros Eus.txt
- Pessoa_O Marinheiro.txt
- Pessoa_Poemas.txt
- Pessoa_Poesias Inéditas.txt
- Pessoa_Primeiro Fausto.txt
- Reis_Poemas.txt
- Soares_Desassossego.txt

***Corpus_Pessoa(II)**
Cluster Analysis



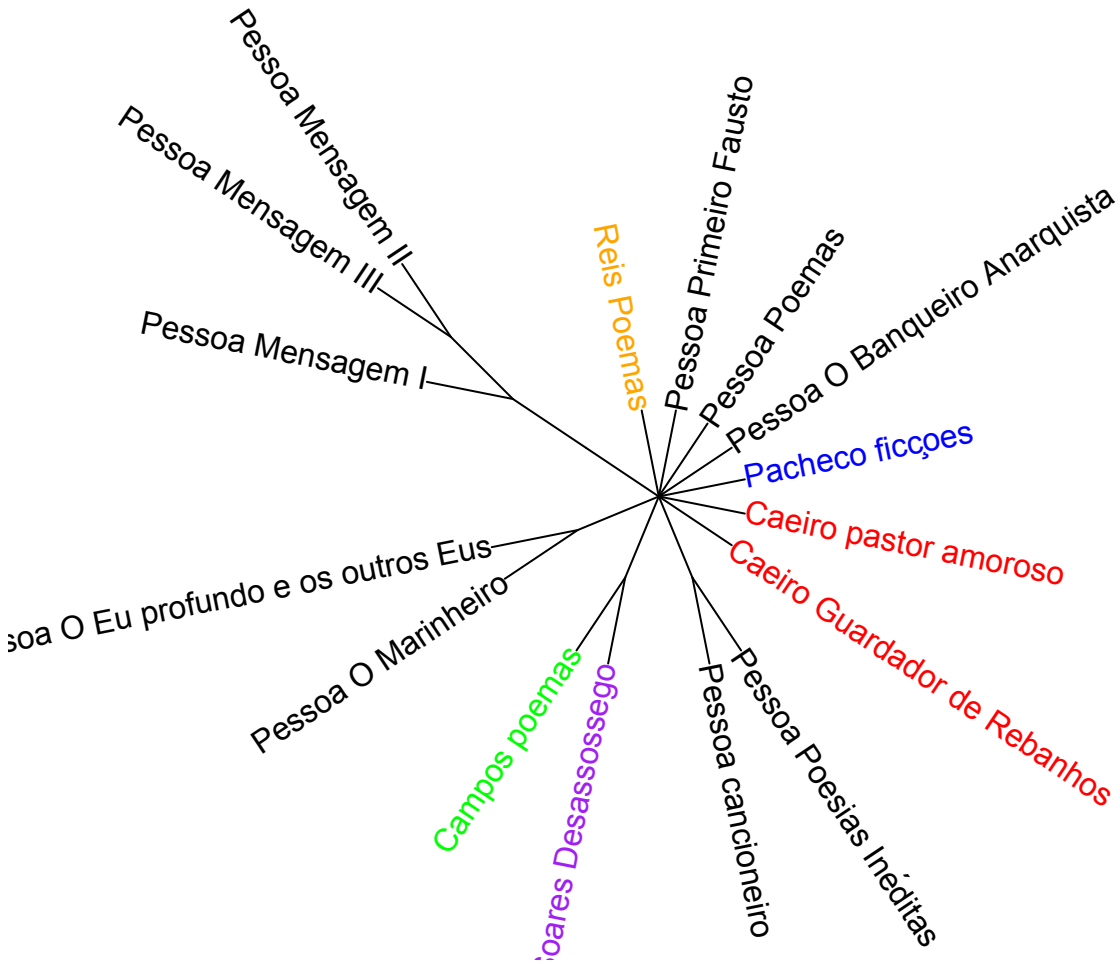
1000 MFW Culled @ 0%
Classic Delta distance

Claudia García-Minguillán
University of Salamanca
2018-2019



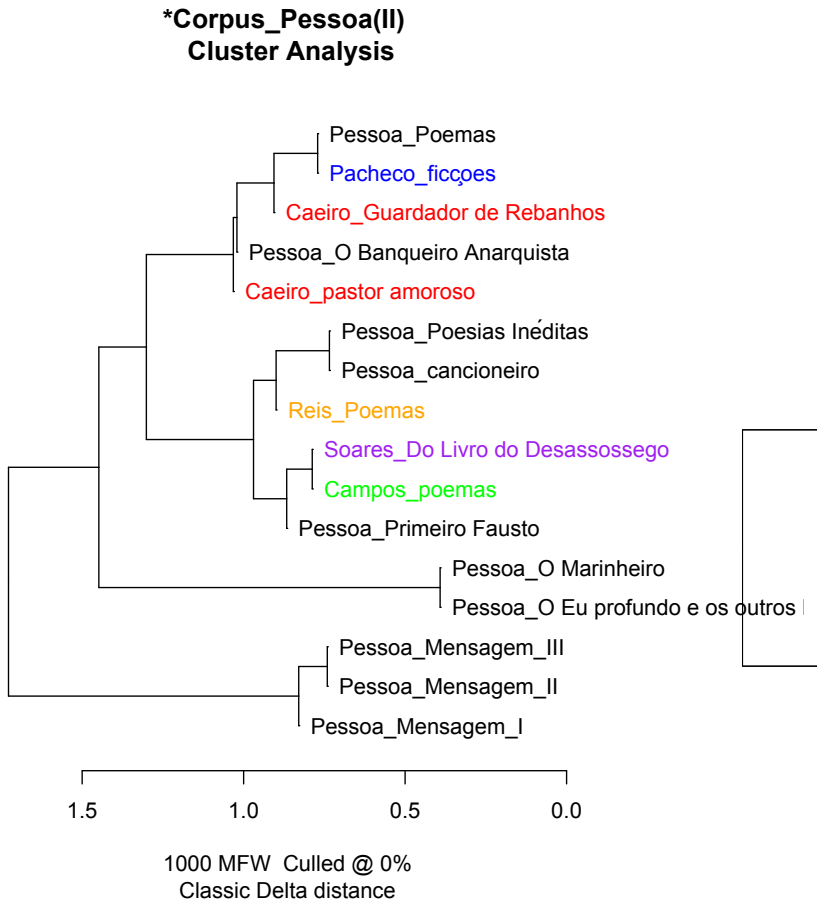
Claudia García-Minguillán
University of Salamanca
2018-2019

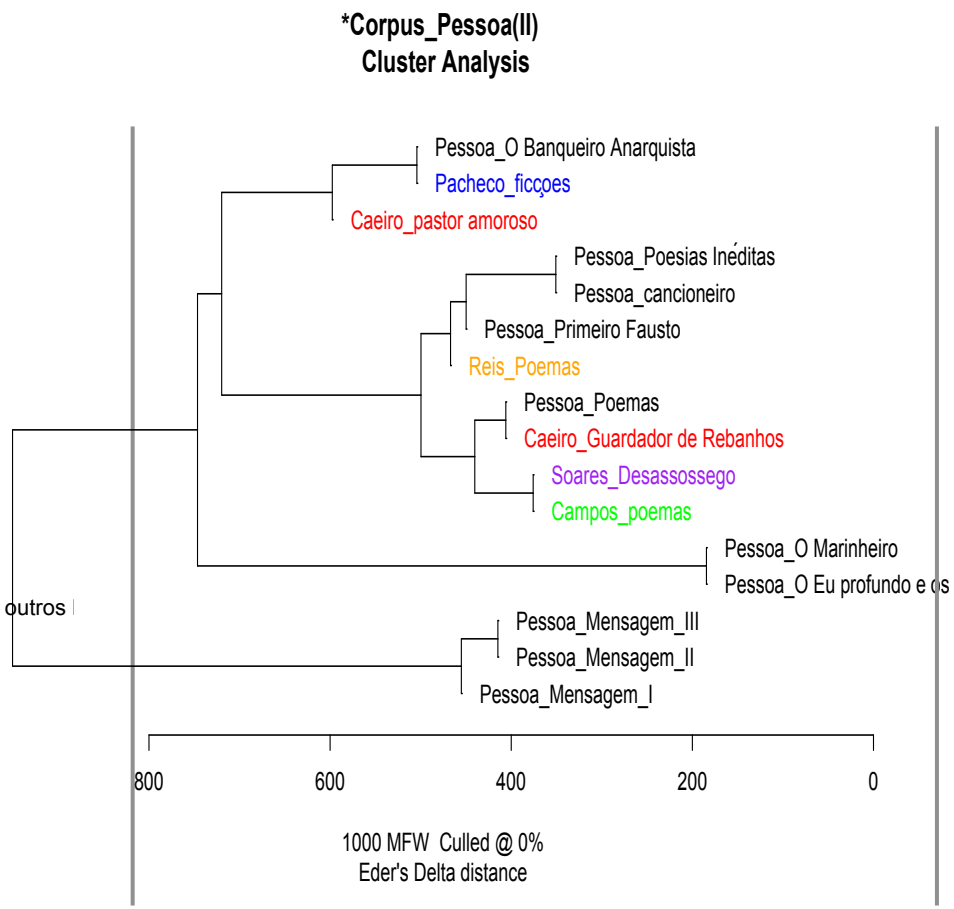
*Corpus_Pessoa(II)
Bootstrap Consensus Tree

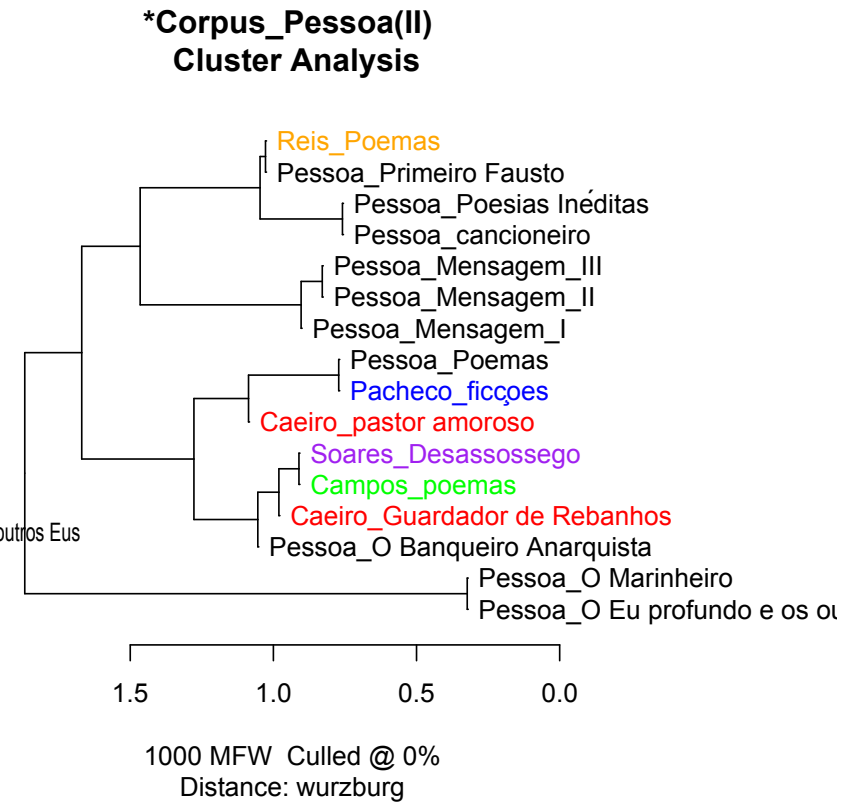


100-1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

Claudia García-Minguillán
University of Salamanca
2018-2019







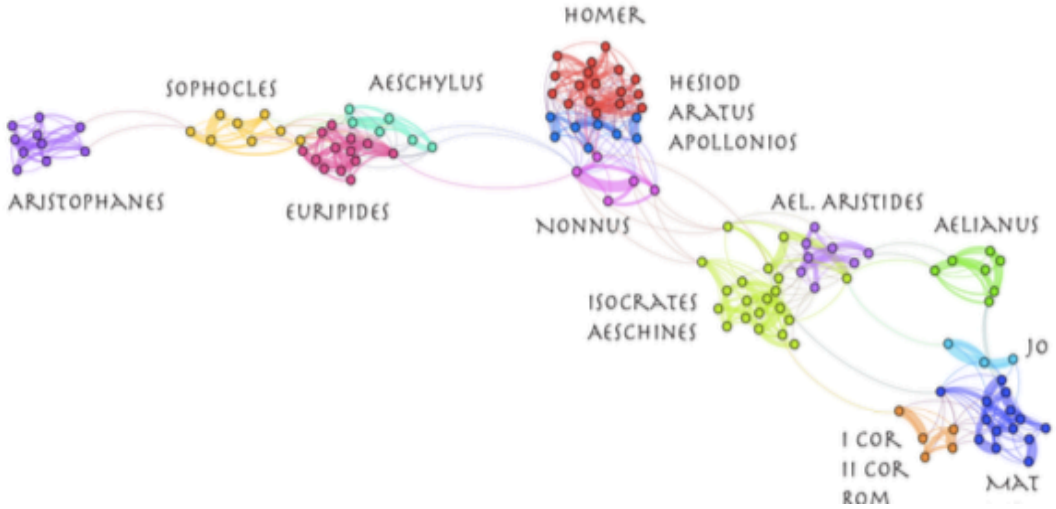
Claudia García-Minguillán
University of Salamanca
2018-2019

3. Enjoy visualization

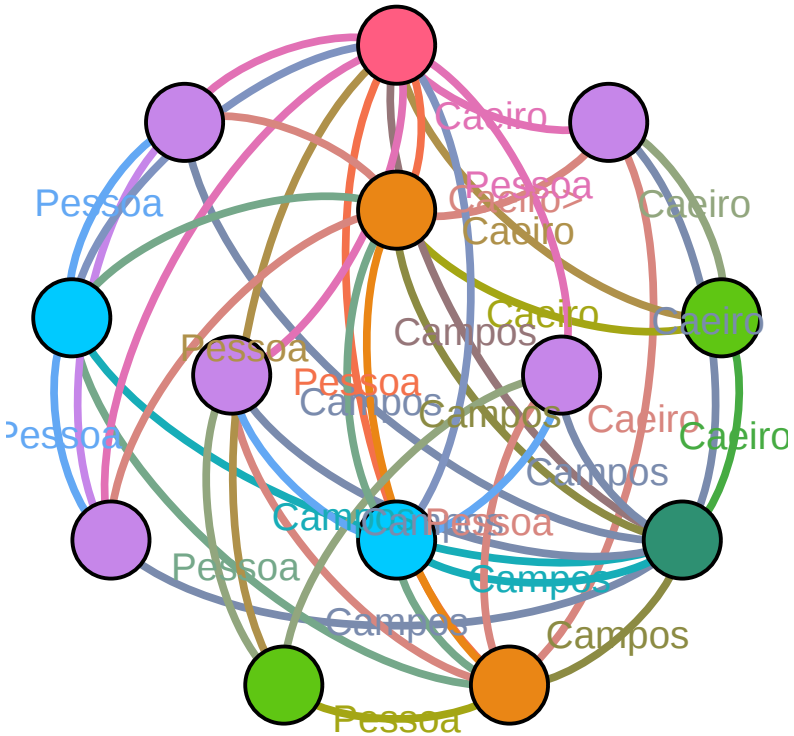
Claudia García-Minguillán
University of Salamanca
2018-2019

Gephi

visualization software



Claudia García-Minguillán
University of Salamanca
2018-2019



Claudia García-Minguillán
University of Salamanca
2018-2019

References

- Eder, M. ; Rybicki, J. & Kestemont, M. (2016): 'Stylometry with R: A Package for Computational Text Analysis', *The R Journal*, vol. 8
- Best material for learning ☞ Computational Stylistics Group
- 'Culture & Technology' European Summer University in Digital Humanities, University of Leipzig
- Digital Humanities Oxford
- 'How a Computer Pogram Helped Show J.K. Rowling write A Cuckoo's Calling', by Patrick Juola, 08/20/2013
- Máster Patrimonio Textual y Humanidades Digitales

THANK YOU FOR YOUR ATTENTION! ANY QUESTIONS?

Contact

cgmt@usal.es
Claudia Garcia-Minguillan Academia.edu
Claudia Garcia-Minguillan Researchgate

Claudia García-Minguillán
University of Salamanca
2018-2019