UVa

Estilometridieval o cómo hacer números con textos medievales

José Manuel Fradejas Rueda
Universidad de **Valladolid**

IV Jornadas del español
Todo lo que un filólogo puede hacer con un ordenador
9 de noviembre de 2017

FFI2016-75014-P AEI-FEDER, EU

Twitter, Facebook, Wikipedia ... y NLP

J. M. Fradejas Rueda

Fuente ¿Twitter?

III Jornadas de Lengua Española
Las vidas de las palabras

Grado en Español: Lengua y Literatura
Universidad de Valladolid
28-29 de marzo de 2017

Likes & Dislikes

Likes and Dislikes

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Análisis de sentimientos

¿Podemos medir / valorar el sentimiento que transmite una obra literaria?

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Análisis de sentimientos

No me refiero a

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

UVa

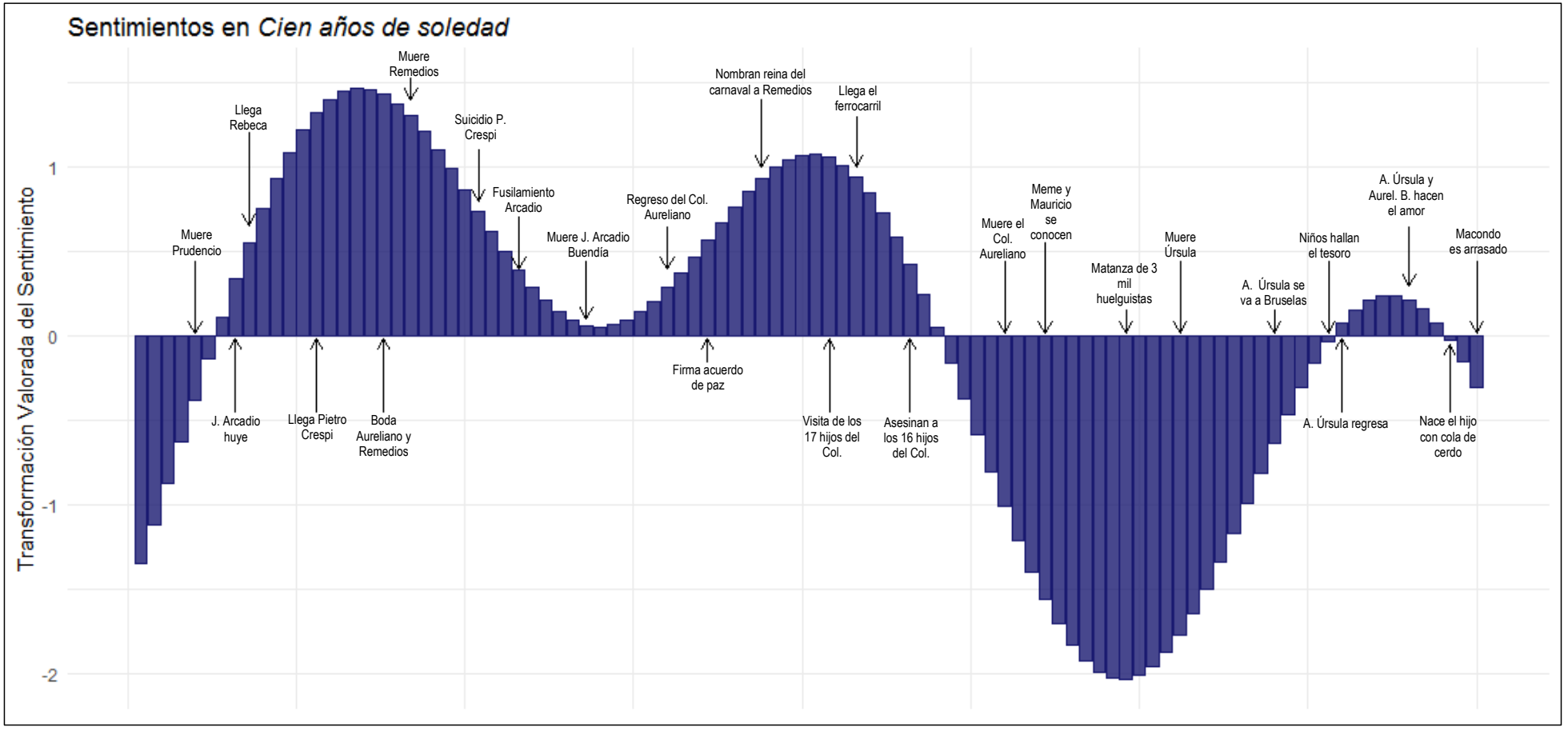
Universidad de Valladolid

Análisis de Sentimientos de *Cien años de soledad* (1967) y *El amor en los tiempos del cólera* (1985) de Gabriel García Márquez

Autor: Danny Fernando Murillo Lanza
Tutor: Dr. José Manuel Fradejas Rueda

Valladolid, 24 de julio de 2017

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

¿Quién es Elena Ferrante?

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

UVa

Análisis estadístico del estilo literario

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

Trata de

- Identificar **semejanzas** y **diferencias** entre textos y
- **Agrupar** los textos de acuerdo con su características lingüísticas

para

- **Detectar señales** en los textos (autoría, géneros, género (sexo), origen, estilo, etc.)

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

La cuestión chespiriana

Edward de Vere

Francis Bacon

Christopher Marlowe

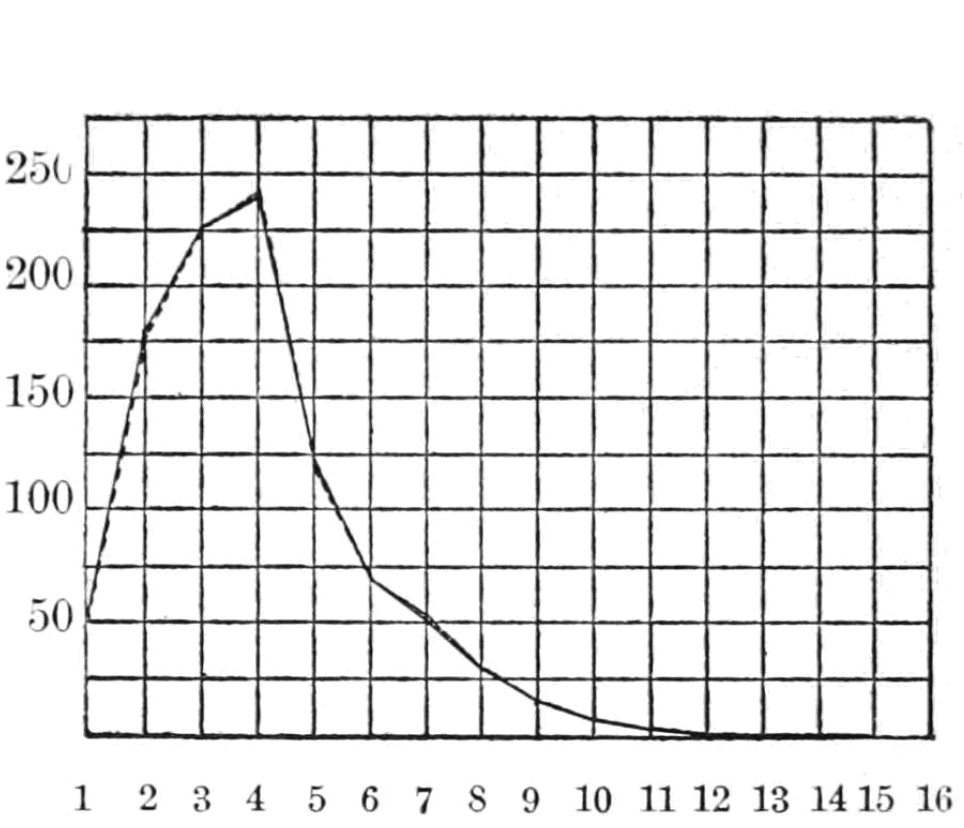
William Stanley

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

T. C. Mendenhall (1901)

"A Mechanical solution of a literary problem"



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

George Kingsley Zipf

American linguist, philologist and statistics professor for Harvard University.

Moreno-Sánchez I, Font-Clos F, Corral Á (2016) Large-Scale Analysis of Zipf's Law in English Texts. PLoS ONE 11(1): e0147073. doi:10.1371/journal.pone.0147073 (publicado 22.01.16)

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

Federalist Papers

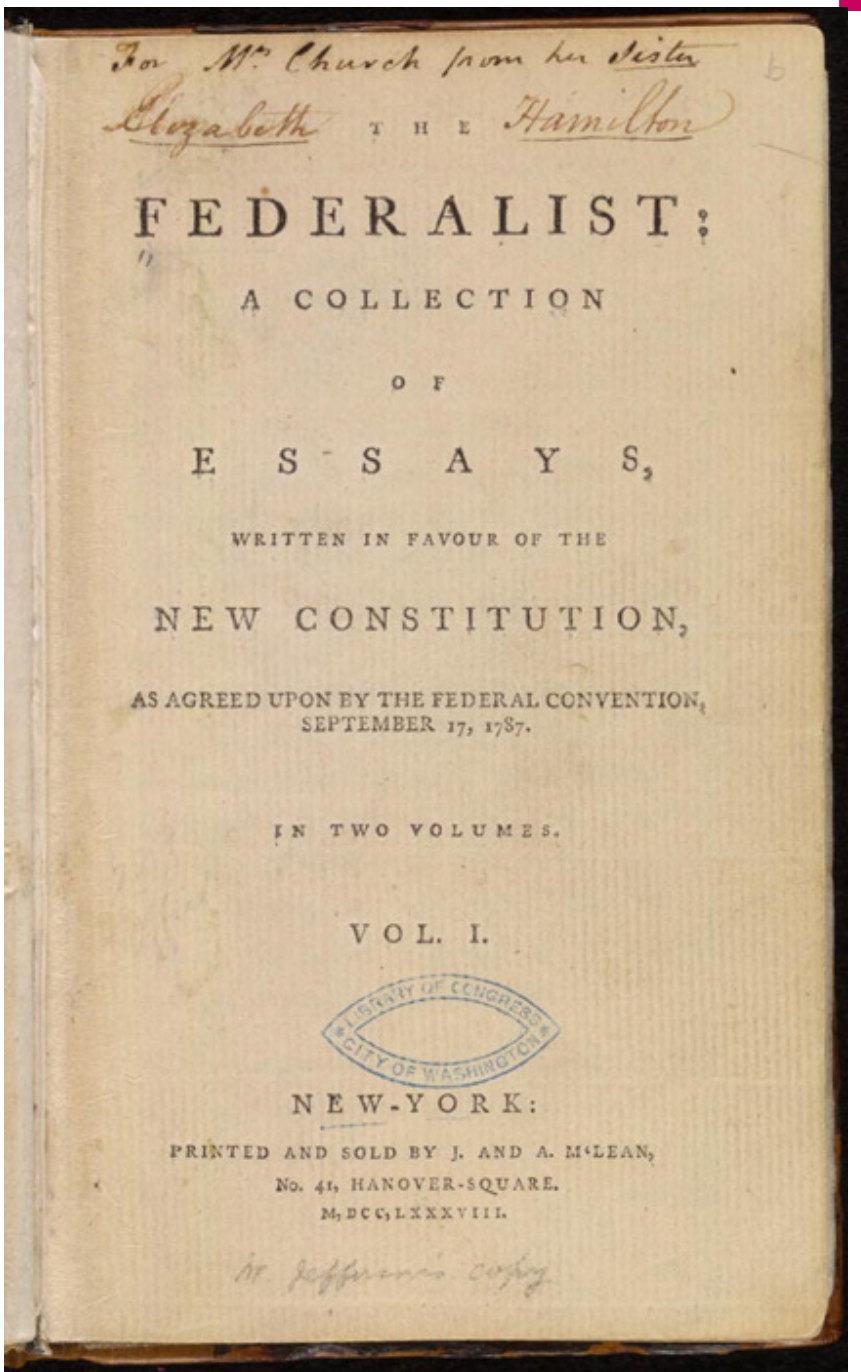
85 artículos

Publius

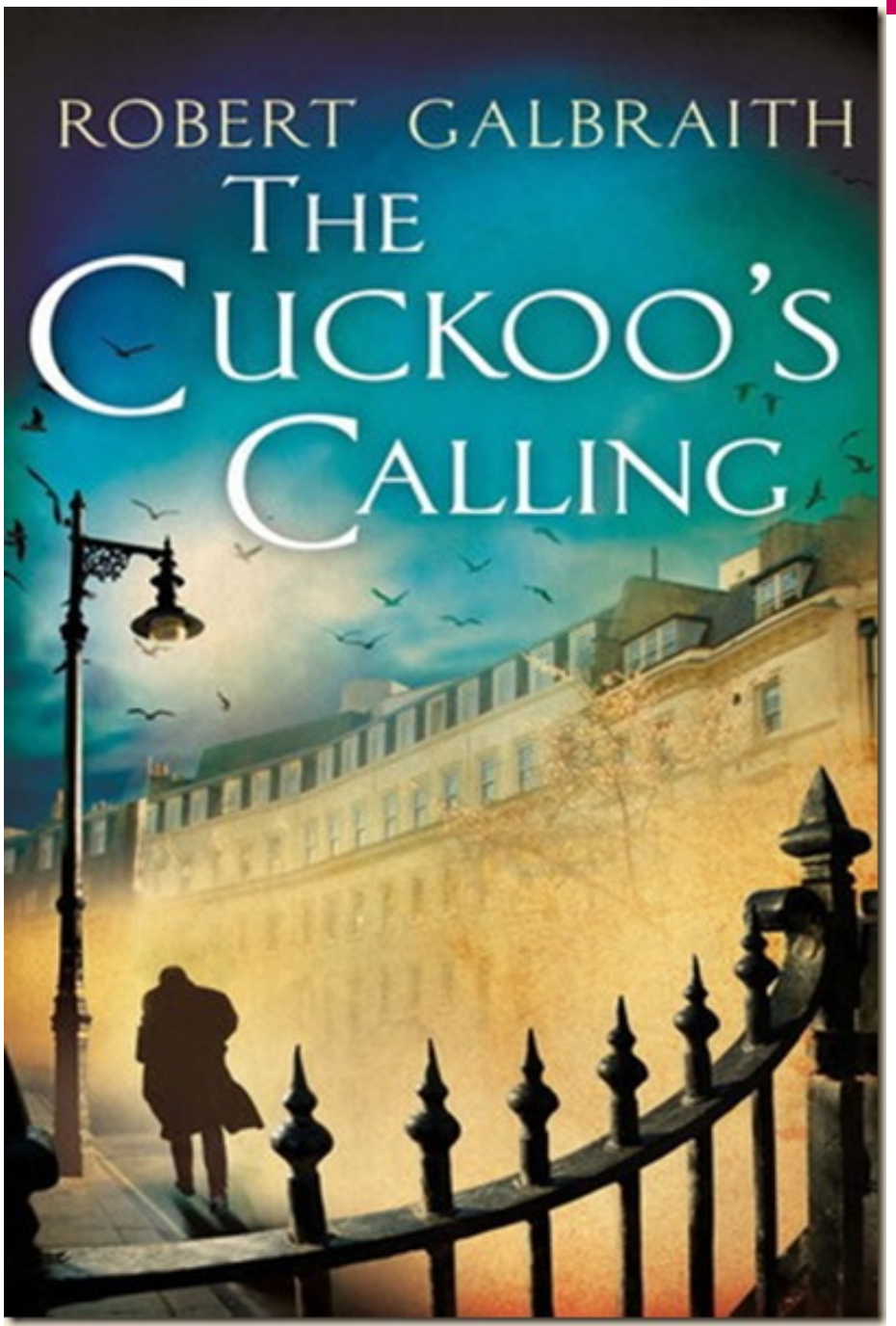
Alexander Hamilton
James Madison
John Jay

Mosteller & Wallace 1964

http://avalon.law.yale.edu/subject_menus/fed.asp


For Mrs Church from her Sister
Elizabeth
THE FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
LIBRARY OF CONGRESS
CITY OF WASHINGTON
NEW-YORK:
PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M, DCC, LXXXVIII.
Mr Jefferson's copy

Estilometría

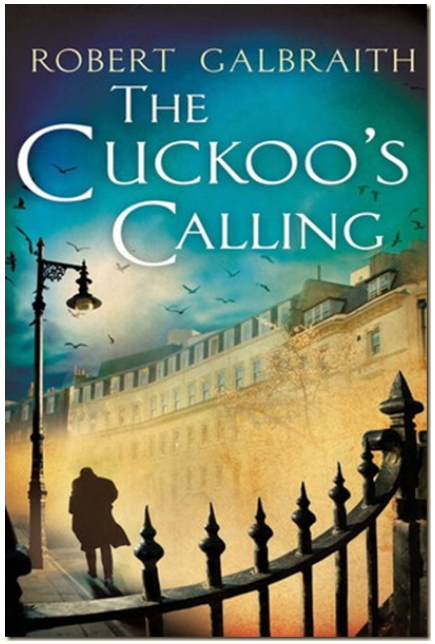


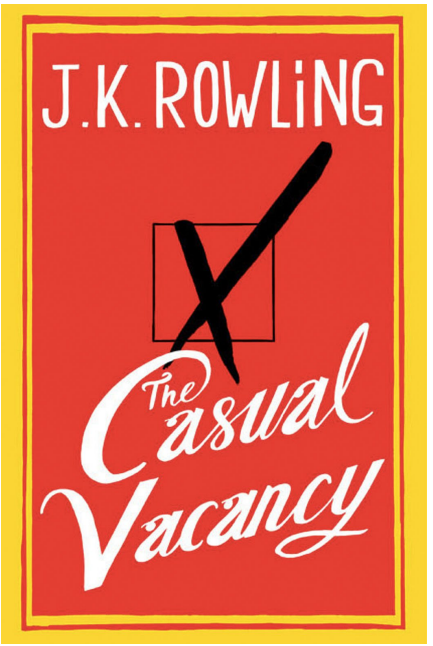
2013

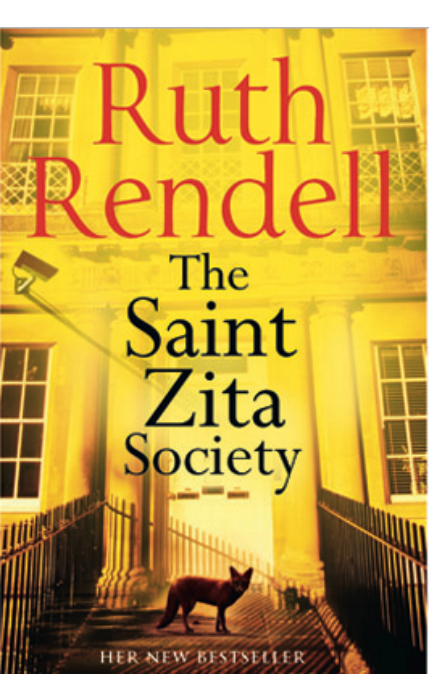
7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

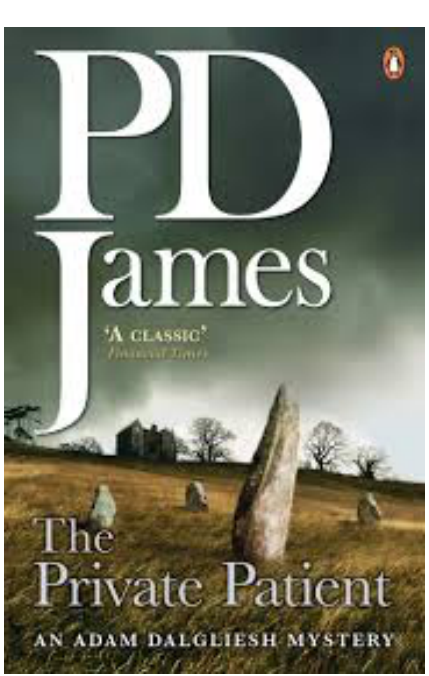
Estilometría

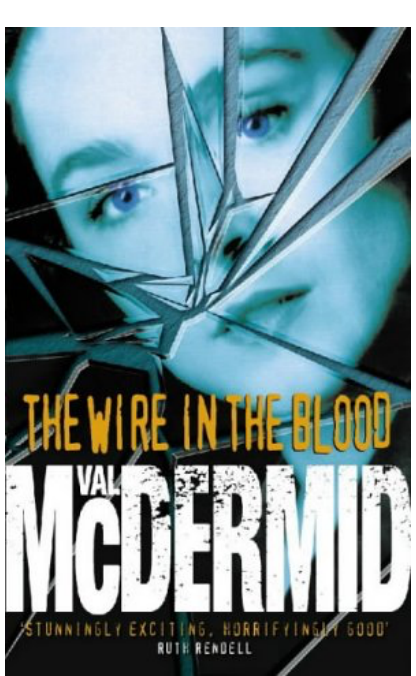
Patrick Juola

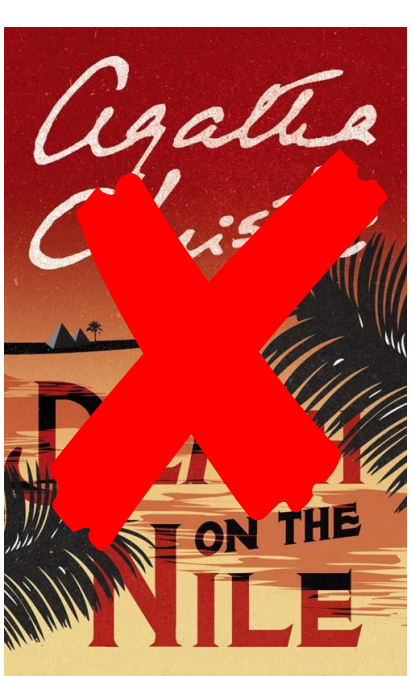












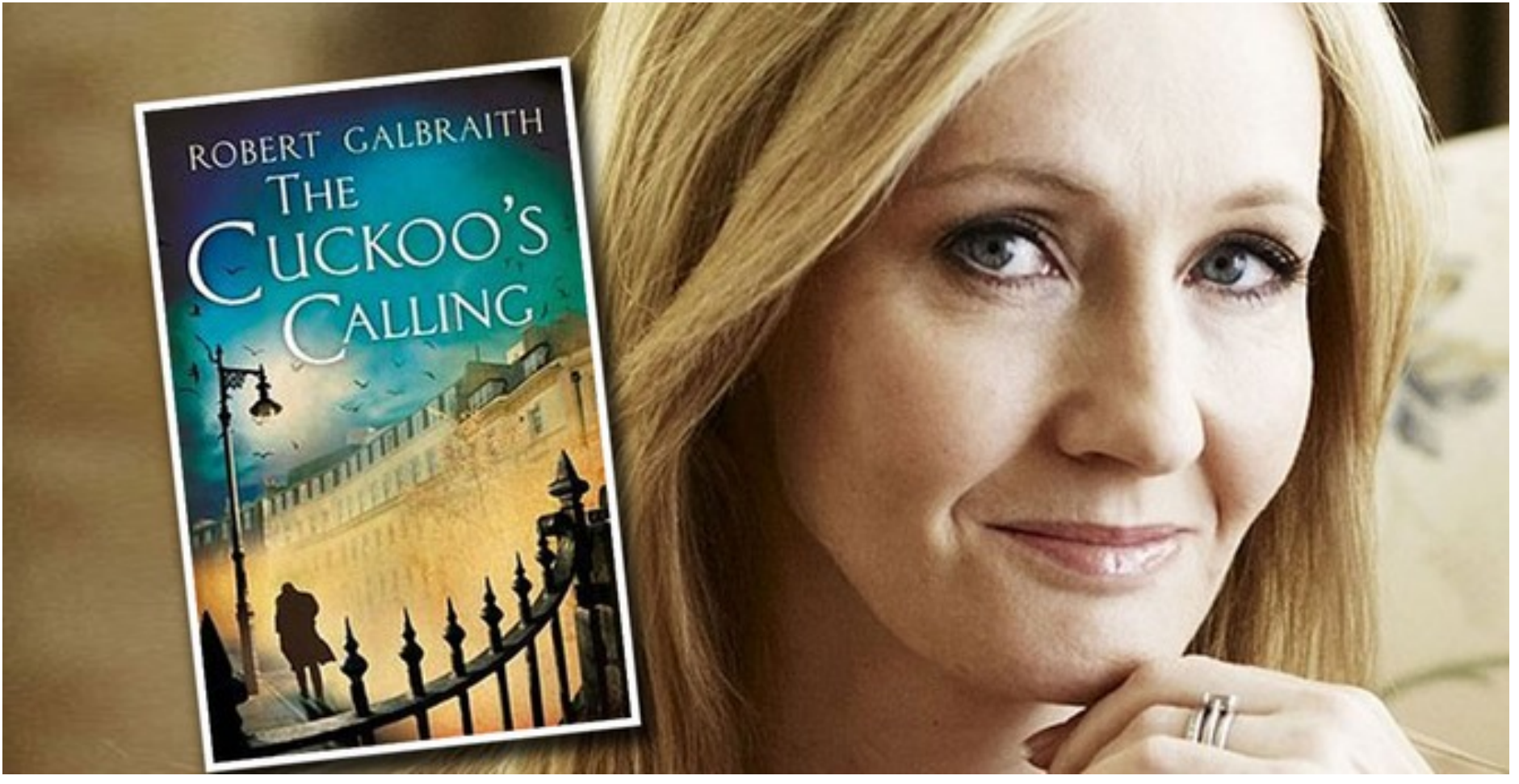
7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría

- Distribución de la longitud de las palabras
- Uso de las 100 palabras más comunes
- Distribución de 4-grams
- Distribución de bigramas

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Estilometría



"The only person consistently suggested by every analysis was Rowlin, who showed up as the winner or the runner-up in each instance"

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Minería de textos

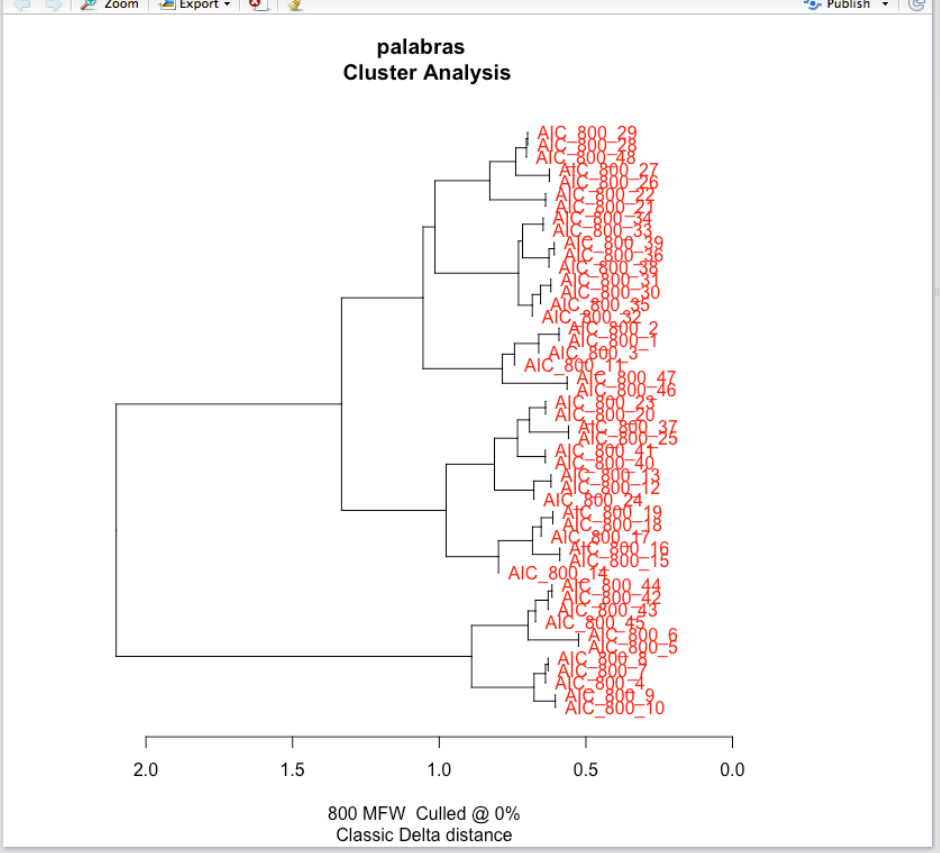
7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Minería de textos

UVa

RStudio

```
as.integer(a)
as.integer(as.roman(a))
a <- "xxi"
as.integer(as.roman(a))
library(stylo)
setwd("~/Desktop/LCA-LEXOMICA/LCA/palabras")
stylo()
```




```
loading AIC_800_41.txt  ...
loading AIC_800_42.txt  ...
loading AIC_800_43.txt  ...
loading AIC_800_44.txt  ...
loading AIC_800_45.txt  ...
loading AIC_800_46.txt  ...
loading AIC_800_47.txt  ...
loading AIC_800_48.txt  ...
loading AIC_800_5.txt   ...
loading AIC_800_6.txt   ...
loading AIC_800_7.txt   ...
loading AIC_800_8.txt   ...
loading AIC_800_9.txt   ...
slicing input text into tokens...
turning words into features, e.g. char n-grams (if applicable)...

Total nr. of samples in the corpus:  48
.....................
.....................
The corpus consists of 38686 tokens

processing  48  text samples
....
combining frequencies into a table...


culling @  0     available features (words)  4234
Calculating z-scores...

Calculating classic Delta distances...
MFW used: 800

Function call:
stylo()

Depending on your chosen options, some results should have been written
into a few files; you should be able to find them in your current
(working) directory. Usually, these include a list of words/features
used to build a table of frequencies, the table itself, a file containing
recent configuration, etc.

Advanced users: you can pipe the results to a variable, e.g.:
        my.discovery = stylo()
this will create a class "my.discovery" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(my.discovery)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")

>
```

CARACTERES
Estudios culturales y críticos de la esfera digital

mundo online | arte cultura lingüística literatura pensamiento

Revista de investigación para el análisis crítico de la cultura, el pensamiento y la sociedad digitales.

Presente en: ESCI (Thomson Reuters - WoK), ERIH Plus, MLAIB, Latindex, DIALNET, CIRC, ISOC, REDIB, DOAJ, ...

EDITORIAL DELIRIO

El análisis estilométrico aplicado a la literatura española: las novelas policiacas e históricas

Stylometric Analysis Applied to Spanish Literature: Historical and Crime Fiction

José Manuel Fradejas Rueda (Universidad de Valladolid)

Histórica		Policiaca
69	textos	49
44 252 212	caracteres	22 732 809
7 659 249	tokens	4 010 728
122 038	tipos	91 789
195 805	párrafos	149 570
1.593342	Token-Tipo Ratio	2.288587
111 003.6	Media de tokens / novela	81 851.59
12 900.86	Media tipos / novela	10 645.39
2 837.754	Media párrafos / novela	3 052.449
641 336.4	Media caracteres / novela	463 934.9
39.11672	Media palabras/párrafo	26.81506
226.0014	Media caracteres / párrafo	151.9878
5.777618	Media caracteres / palabra	5.668001

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

RStudio

codifica-TEI.R

```
texto.todo <- gsub(' \\]', '\\[', texto.todo, perl = T) # Se asegura de que no hay espacio en blanco antes
texto.todo <- gsub("<p>\\[[f|F]ol\\. (\\d+)([r|v])\\]</p>\n<p>(.+)</p>\n<p>\\[[C|c]ol\\. 1\\]</p>", '<pb n=
texto.todo <- gsub(" (\\d+)</fw>", '</fw>\n<fw type="foliacion">\\1</fw>', texto.todo, perl=T)  # Marca las
texto.todo <- gsub("<p>{CW\\. (.+)}</p>", '<fw type="signatura">\\1</fw>', texto.todo, perl = T)  # Marca l
texto.todo <- gsub("<p>\\[[c|C]ol\\. (\\w+)\\]</p>", '<cb n="\\1"/>', texto.todo, perl = T) # Marca las col
texto.todo <- gsub('</p>\\.\n', '.</p>\n', texto.todo, perl = T) # Se asegura que el punto no queda al fina
texto.todo <- gsub('</hi>\\.\n', '.</hi>\n', texto.todo, perl = T) # Se asegura que el punto no queda al fi
texto.todo <- gsub("mph>", 'x>', texto.todo, perl = T) # Convierte emph en ex para las abreviaturas
```

CODIFICA TEI — R Script

Console ~/Desktop/~STYLO/_historica/

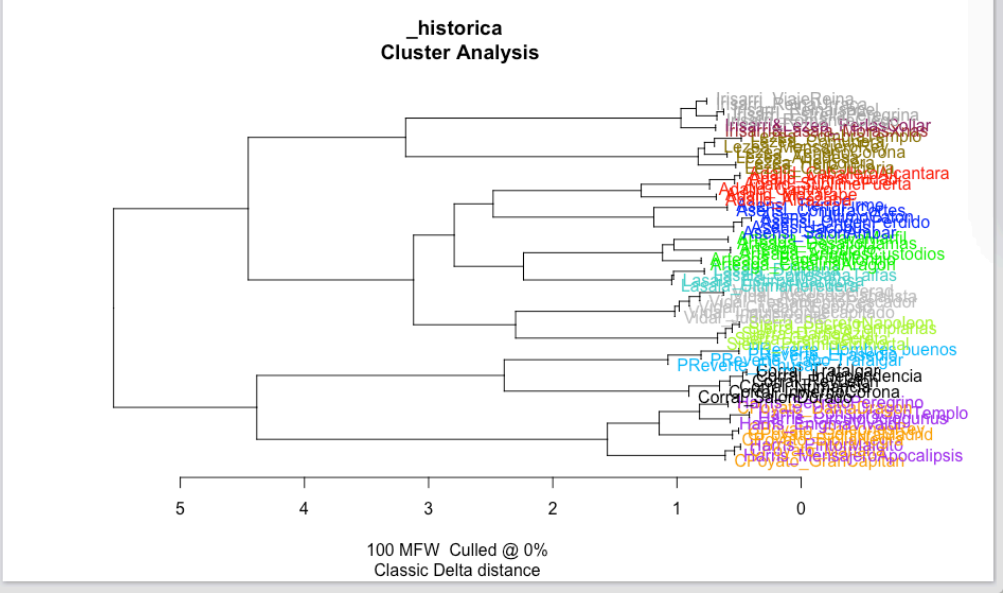
```
loading Lezea_Herbolera.txt     ...
loading Lezea_MensajeroRey.txt  ...
loading Lezea_SombraTemplo.txt  ...
loading Lezea_VenenoCorona.txt  ...
loading PReverte_Cabo Trafalgar .txt    ...
loading PReverte_El asedio .txt ...
loading PReverte_El husar .txt  ...
loading PReverte_Hombres buenos .txt    ...
loading Sierra_CenaSecreta.txt  ...
loading Sierra_DamaAzul.txt     ...
loading Sierra_PiramideInmortal.txt     ...
loading Sierra_PuertaTemplarias.txt     ...
loading Sierra_SecretoNapoleon.txt      ...
loading Vidal_AprendizCabalista.txt     ...
loading Vidal_CiudadRLeproso.txt        ...
loading Vidal_InquisidorDecapitado.txt  ...
loading Vidal_JudioErrante.txt  ...
loading Vidal_MedicoSeferad.txt ...
loading Vidal_TestamentoPescador.txt    ...
slicing input text into tokens...
turning words into features, e.g. char n-grams (if applicable)...

Total nr. of samples in the corpus:  69
.....................
.....................
...............
The corpus consists of 7655431 tokens

processing  69  text samples
......
combining frequencies into a table...


culling @  0     available features (words)  5000
Calculating z-scores...

Calculating classic Delta distances...
MFW used: 100

Function call:
```

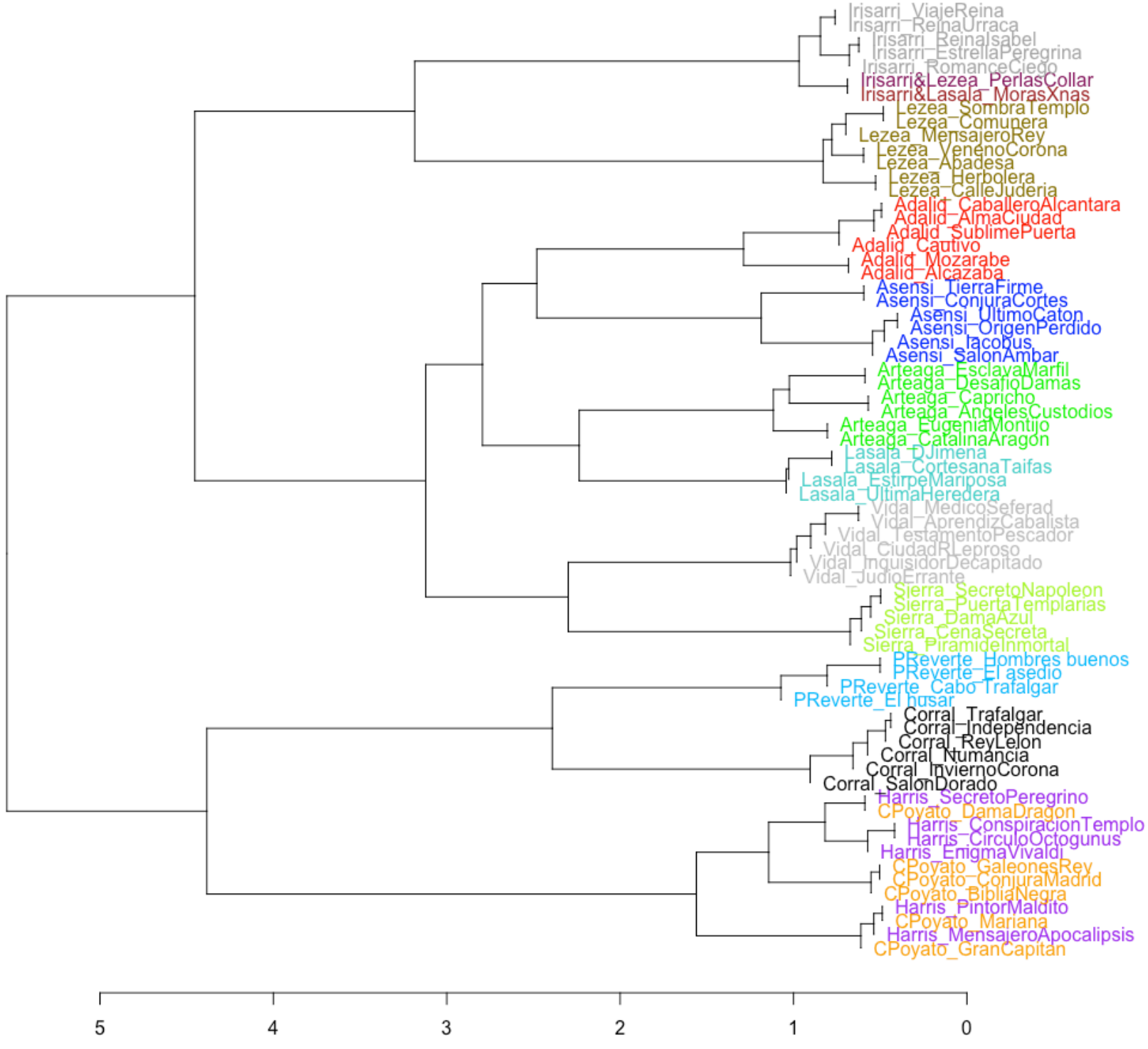
Environment | History

```
m <- addMarkers (m, lng=-2.083158, lat =39.546571, popup = "Alarcón")
m
m <- addMarkers (m, lng=-1.726366, lat =38.919603, popup = "Chinchilla")
m
m <- addMarkers (m, lng=-0.860909, lat =38.631905, popup = "Villena")
m
m <- addMarkers (m, lng=-1.1300278, lat =37.986111, popup = "Murcia")
m
install.packages("tau")
install.packages(c("tidyverse", "revest"))
library(stylo)
library(tidytext)
sentiments
library(syuzhet)
setwd("~/Desktop/~STYLO/_historica")
library(stylo)
stylo()
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish





7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

RStudio

codifica-TEI.R

```
texto.todo <- gsub('\n</p>', '</p>', texto.todo, perl = T) # Agrupa </p> aislados en una línea
texto.todo <- gsub("-</p>\n<p>", '-\n<lb break="no" rend="guion"/>', texto.todo, perl = T) # Marca fin de l
texto.todo <- gsub("-@</p>\n<p>", '\n<lb rend="guion"/>', texto.todo, perl = T) # Marca fin de línea cortad
texto.todo <- gsub("#</p>\n<p>", '-\n<lb break="no"/>', texto.todo, perl = T)  # Marca fin de línea cortado
texto.todo <- gsub("</p>\n<p>", '\n<lb/>', texto.todo, perl = T) # Marca fin de línea
texto.todo <- gsub('-</p>\n<cb n="2"/>\n<p>', '-\n<cb n="2"/>\n<lb break="no" rend="guion"/>', texto.todo,
texto.todo <- gsub('#</p>\n<cb n="2"/>\n<p>', '-\n<cb n="2"/>\n<lb break="no"/>', texto.todo, perl = T) # M
texto.todo <- gsub('</p>\n<cb n="2"/>\n<p>', '\n<cb n="2"/>\n<lb/>', texto.todo, perl = T) # Marca fin de l
texto.todo <- gsub('-</p>\n<pb n="(\\w+)"/>\n<fw type="encabezado">(.*)</fw>\n<fw type="foliacion">(\\d+)</
```

1:1 CODIFICA TEI — R Script

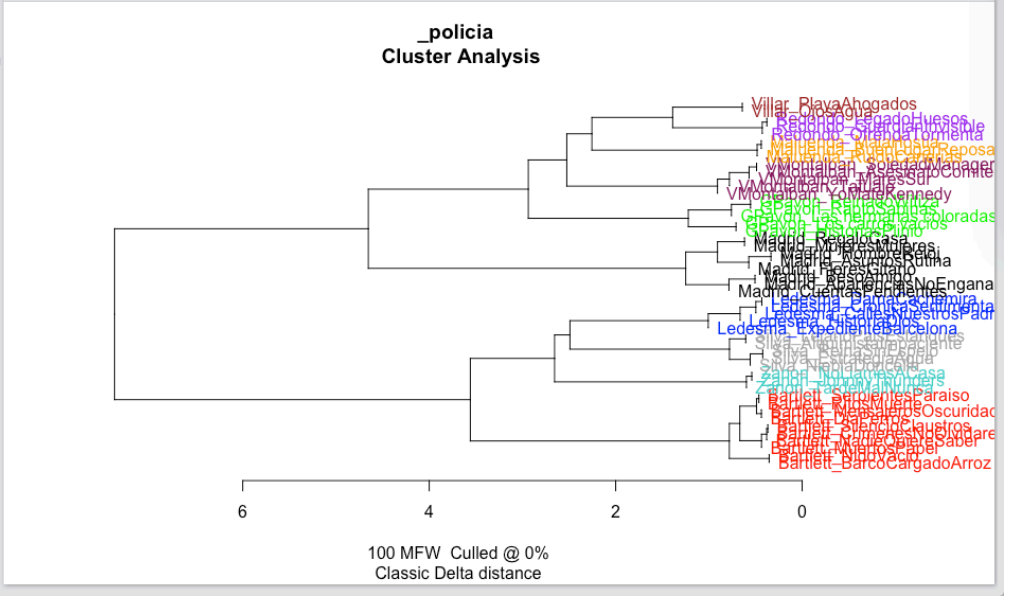
Console ~/Desktop/~STYLO/_policia/

```
loading Maluenda_MaldHostia.txt ...
loading Maluenda_RuidoCanerias.txt      ...
loading Redondo_GuardianInvisible.txt   ...
loading Redondo_LegadoHuesos.txt        ...
loading Redondo_OfrendaTormenta.txt     ...
loading Silva_AlquimistaImpaciente.txt  ...
loading Silva_EstrategiaAgua.txt        ...
loading Silva_LejanoPaisEstanques.txt   ...
loading Silva_NieblaDoncella.txt        ...
loading Silva_ReinaSinEspejo.txt        ...
loading Villar_OjosAgua.txt     ...
loading Villar_PlayaAhogados.txt        ...
loading VMontalban_AsesinatoComiteCentral.txt   ...
loading VMontalban_MaresSur.txt ...
loading VMontalban_SoledadManager.txt   ...
loading VMontalban_Tatuaje.txt  ...
loading VMontalban_YoMateKennedy.txt    ...
loading Zanon_JohnnyThunders.txt        ...
loading Zanon_NoLlamesACasa.txt ...
loading Zanon_TardeMalNunca.txt ...
slicing input text into tokens...
turning words into features, e.g. char n-grams (if applicable)...

Total nr. of samples in the corpus:  49
.....................
.....................
The corpus consists of 4008394 tokens

processing  49  text samples
....
combining frequencies into a table...


culling @  0     available features (words)  5000
Calculating z-scores...

Calculating classic Delta distances...
MFW used: 100

Function call:
stylo()
```

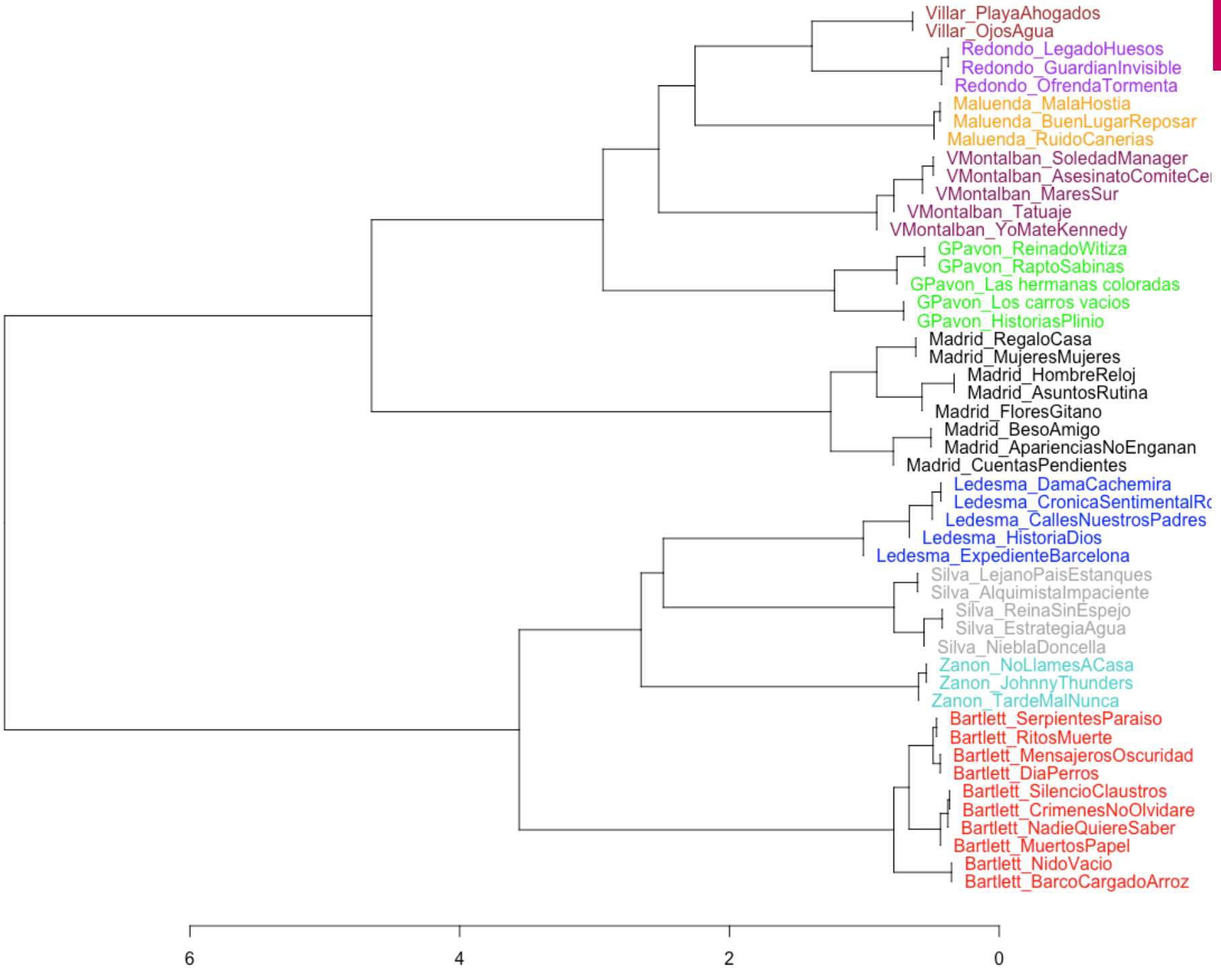
Environment | History

```
m <- addMarkers (m, lng=-1.726366, lat =38.919603, popup = "Chinchilla")
m
m <- addMarkers (m, lng=-0.860909, lat =38.631905, popup = "Villena")
m
m <- addMarkers (m, lng=-1.1300278, lat =37.986111, popup = "Murcia")
m
install.packages("tau")
install.packages(c("tidyverse", "revest"))
library(stylo)
library(tidytext)
sentiments
library(syuzhet)
setwd("~/Desktop/~STYLO/_historica")
library(stylo)
stylo()
setwd("~/Desktop/~STYLO/_policia")
stylo()
```

Files | Plots | Packages | Help | Viewer

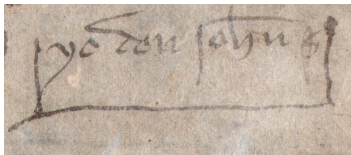


7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU



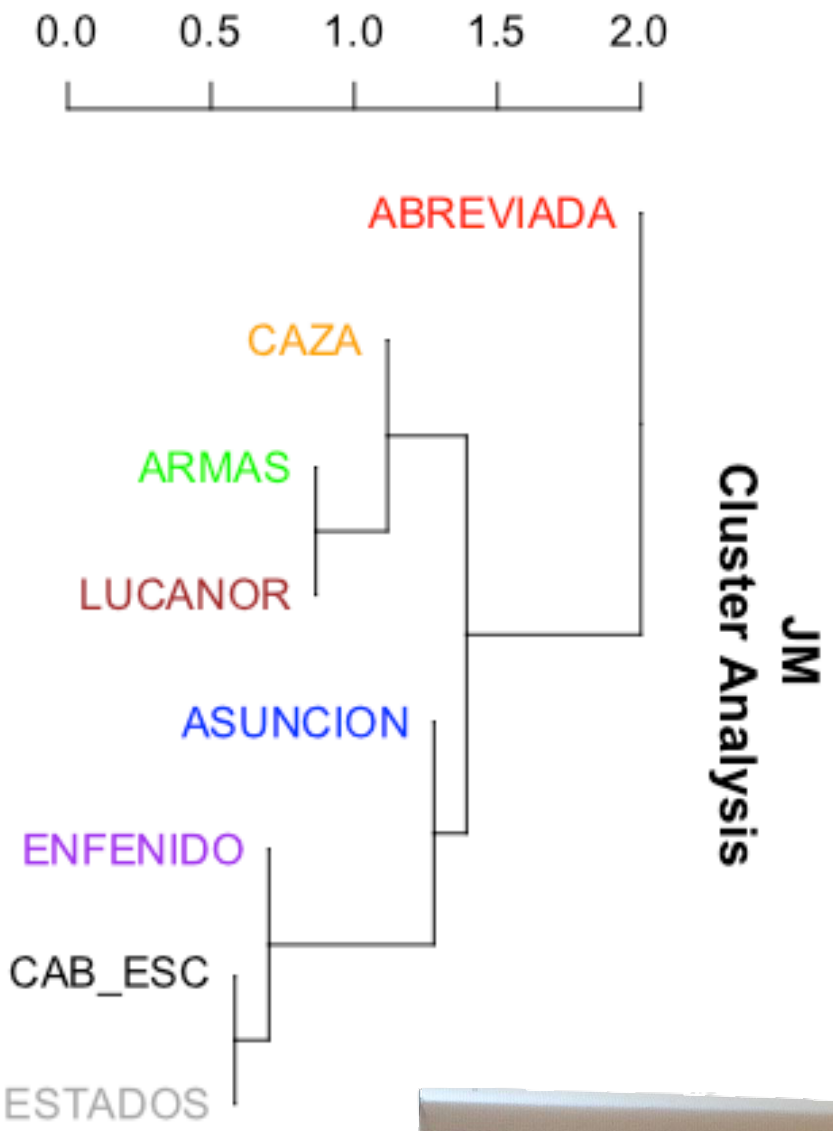
7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

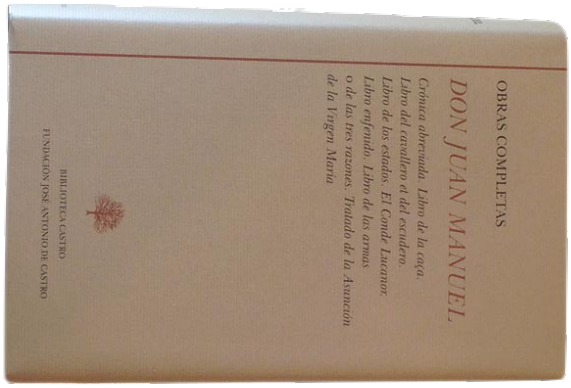
1. **Crónica abreviada**
2. **Libro de la caza**
3. **Libro del caballero y del escudero**
4. Libro de las armas
5. Libro de los estados
6. Libro enfenido
7. Conde Lucanor
8. Tratado de la Asunción



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Juan Manuel, *Obras completas*



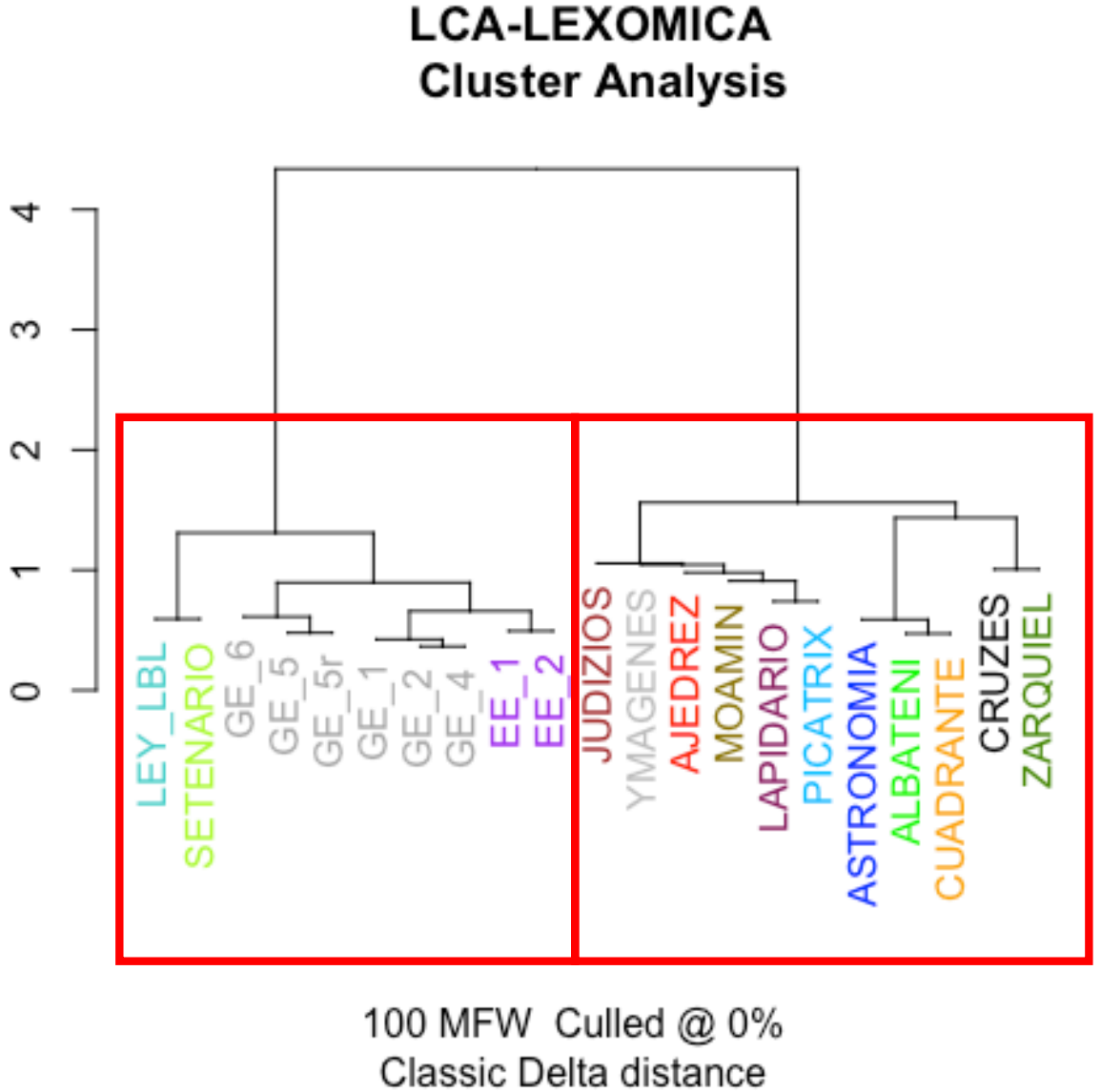


Scriptorium alfonsí

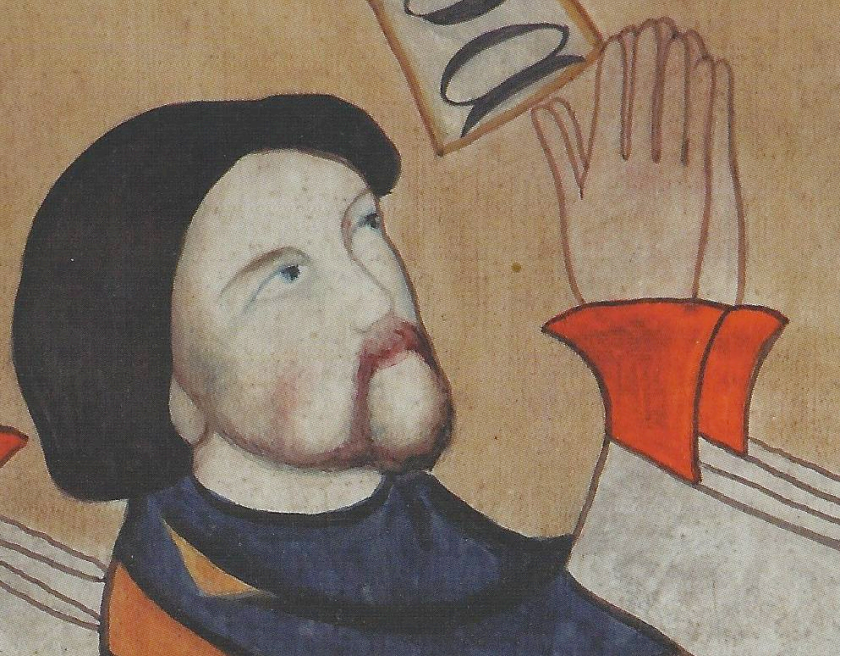
21 textos
3.234.321 palabras

Scriptorium alfonsí

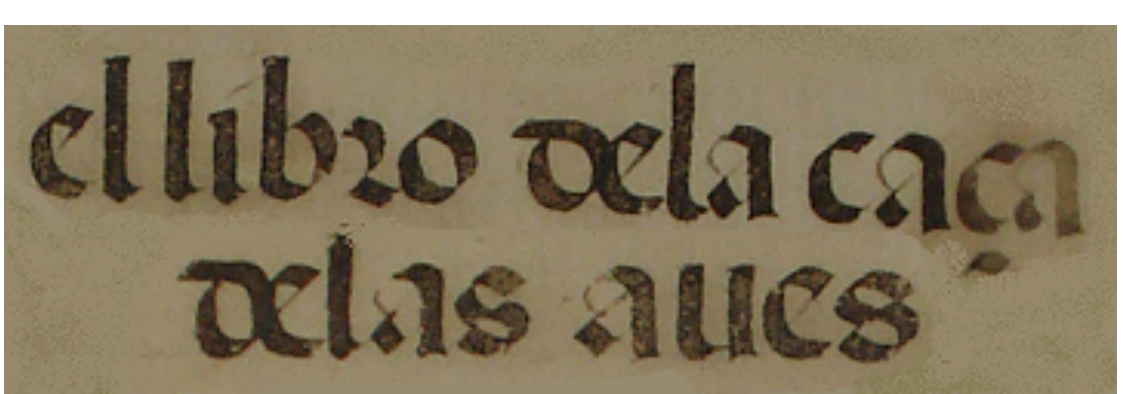
21 textos
3.234.321 palabras



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

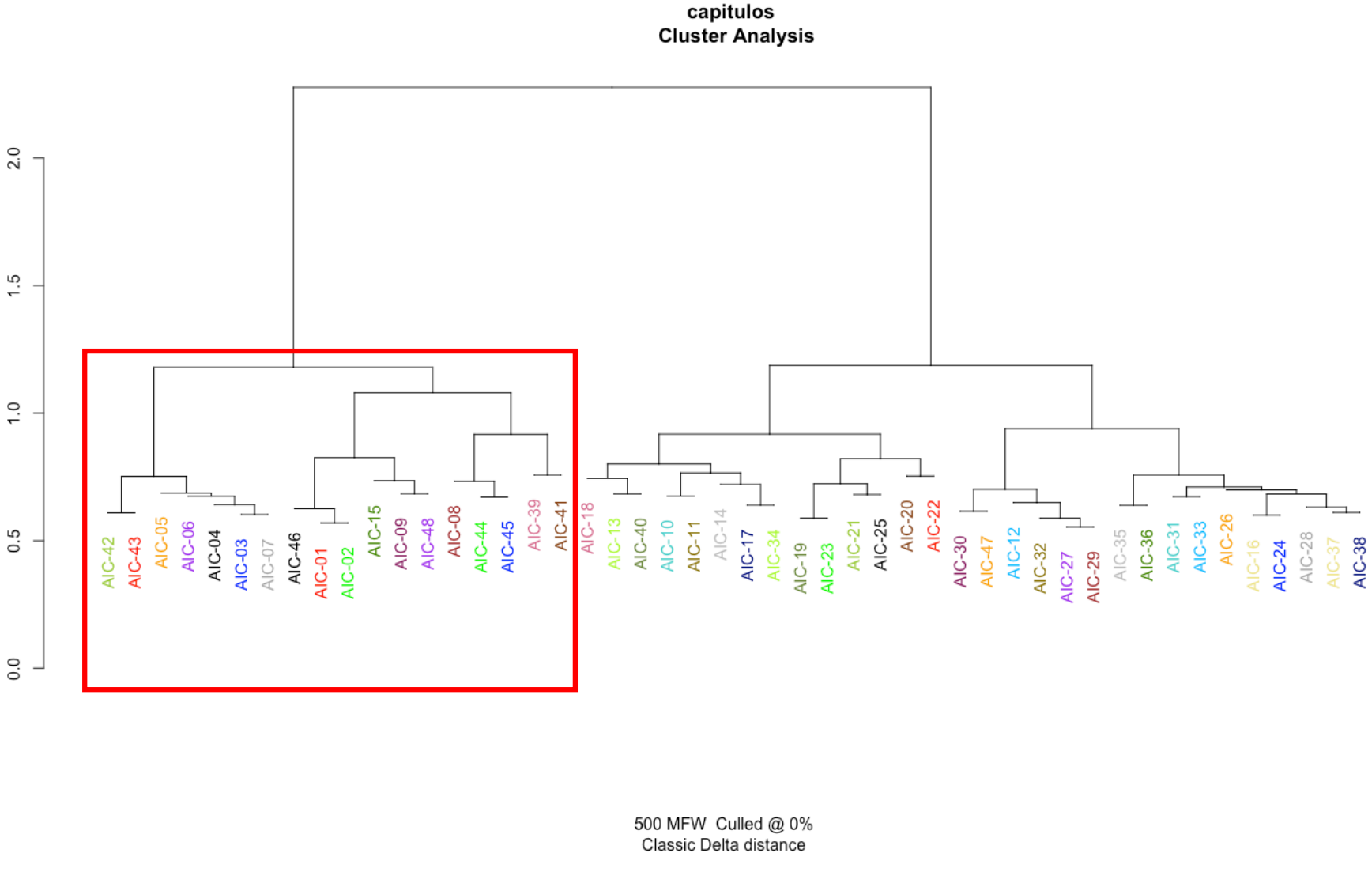


López de Ayala



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Libro de la caza de las aves
Pero López de Ayala

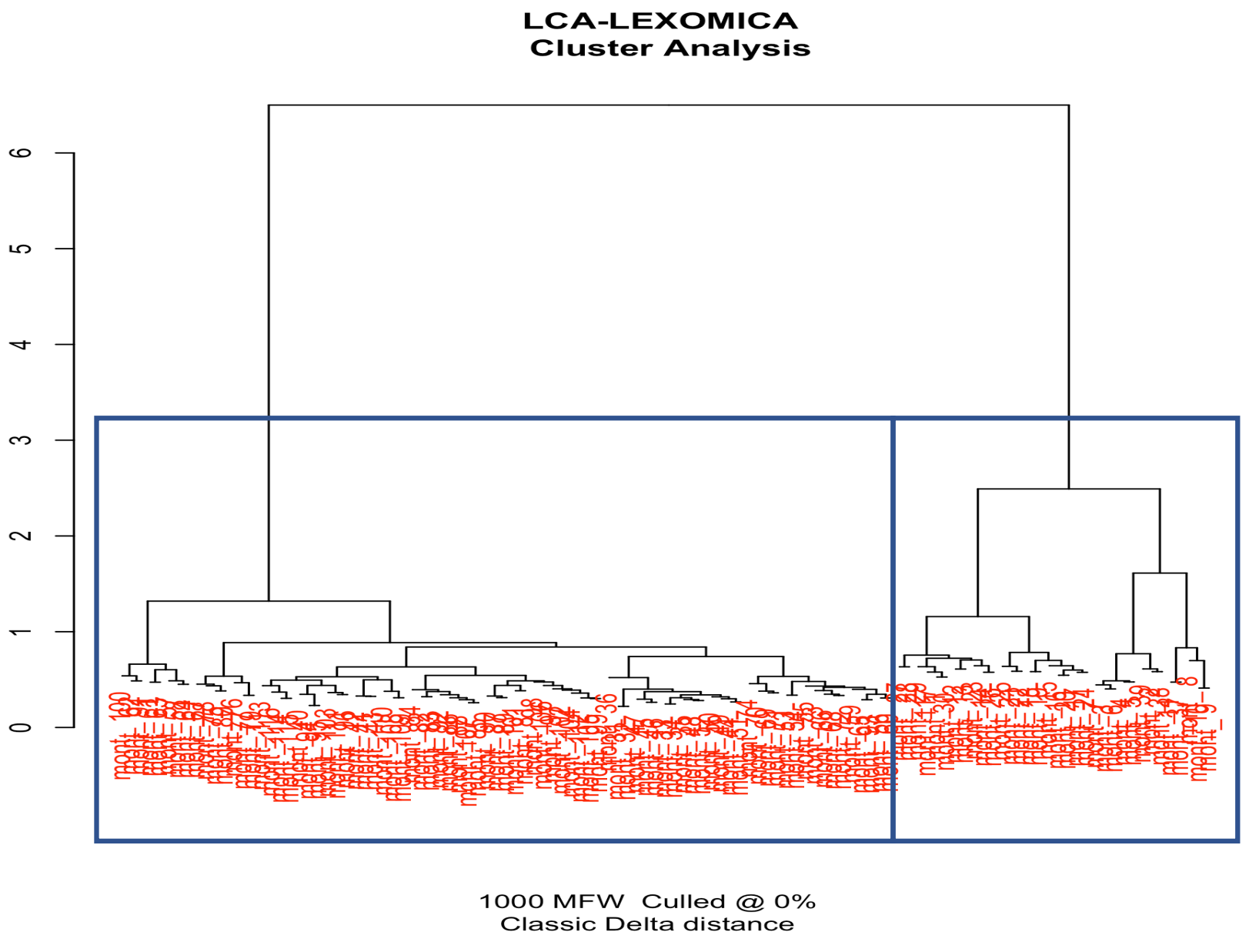


7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Libro de la montería

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Libro de la montería



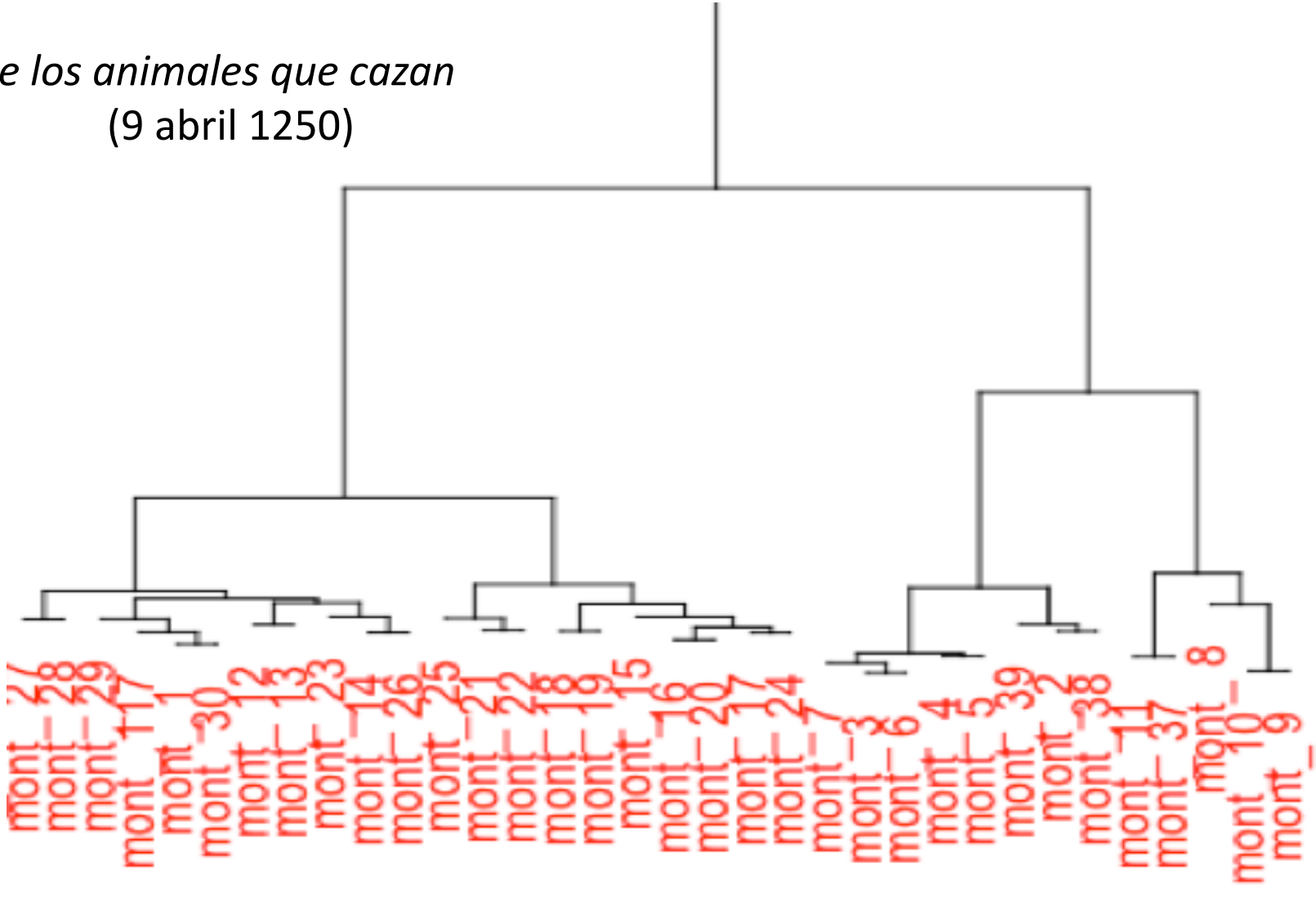
7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

Libro de la montería (detalle)

1000 MFW Delta Classic

Lo procedente del

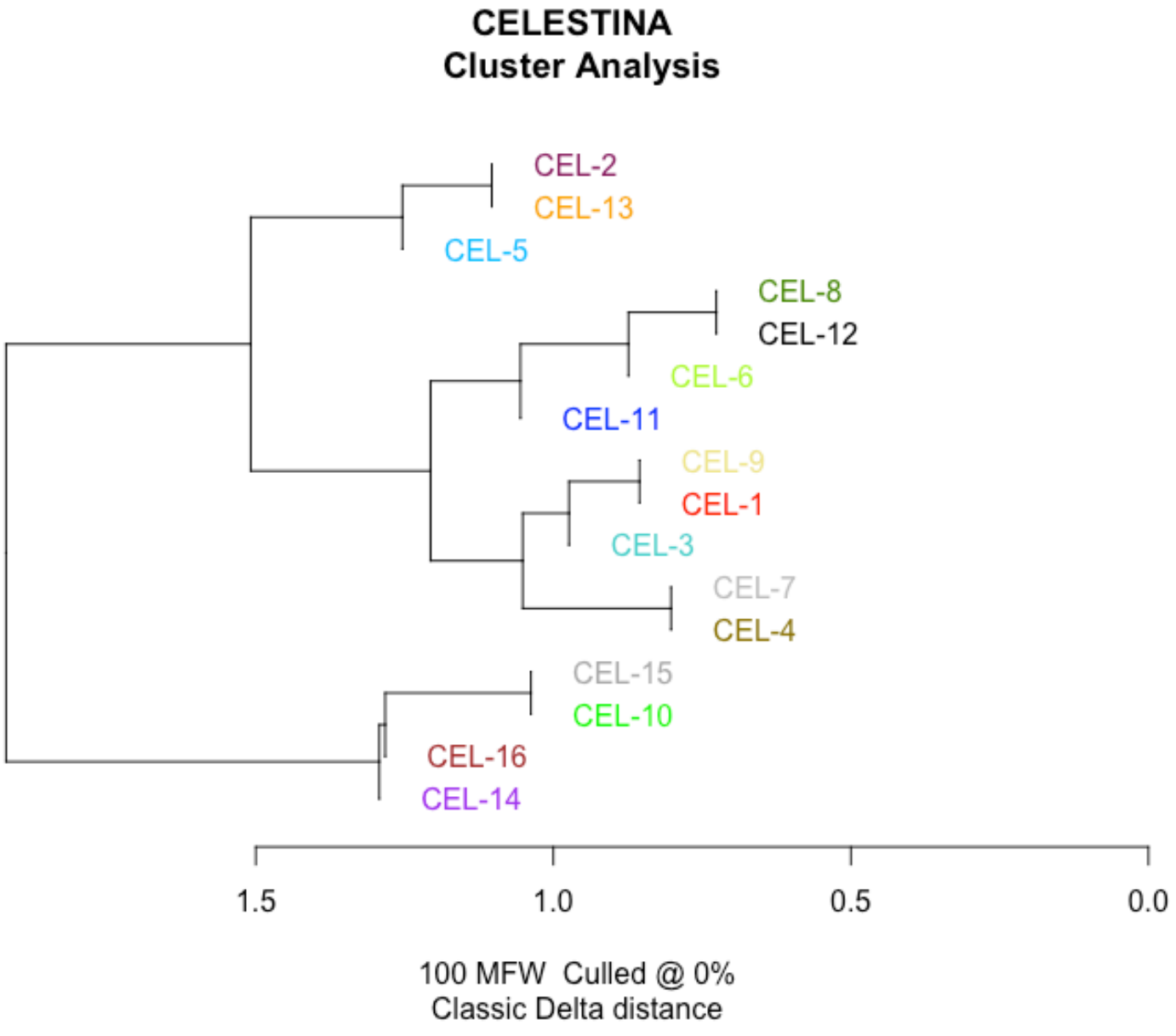
Libro de los animales que cazan
(9 abril 1250)



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

¶Argumẽto del primer auto desta comedia.

Entrãdo Calisto vna huerta empos ỏ vn falcon suyo fallo y a Melibea de cuyo amor preso comẽçole de hablar: dela qual rigorosamẽte dẽspedido: fue para su casa muy sanguistiado. hablo con vn criado suyo llamado sempronio. el qual despues de muchas razones le endereço a vna vieja llamada celestina: en cuya casa tenia el mesmo criado vna enamorada llamada elicia: la qual vinẽdo sempnio a casa ỏ celestina cõel negocio de su amo tenia a otro consigo llamado crito: al qual escondierõ. Entre tanto q̃ sempronio esta negociãdo con celestina: calisto esta razonando cõ otro criado suyo por nõbre parmeno: el qual razonamiẽto dura fasta q̃ llega Sempronio y celestina a casa de calisto. Parmeno fue conoscido de celestina: la qual mucho le dize delos fechos y co-



7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

NEW APPROACHES TO TEXTUAL DISCUSSIONS IN HERRERA'S POETRY

Laura Hernández Lorenzo
FPU researcher and teaching fellowship
lhernandez1@us.es University of Seville

UVa

7PartidasDigital – Universidad de Valladolid – MINECO / FEDER FFI1016-75014-P AEI-FEDER-EU

THE AUTHOR AND HIS WORKS

Fernando de Herrera, 'The Divine'

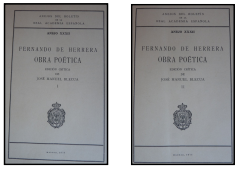
Algunas Obras, known as *H*

Fernando de Herrera (1534-1597)

- One of the most influential poets of Spanish Golden Age
- His poetic works published in life are know as *H*.
- Textual and authorship problems with the poetic works published after his death by the painter Pacheco and known as ***P*** (Cuevas 1985)

Versos de Fernando de Herrera, known as *P*

DIGITALISATION



Herrera, Fernando (1975). *Obra poética*. Volumes I and II. Annotated edition by José Manuel Blecua. Madrid: Real Academia Española.

OCR

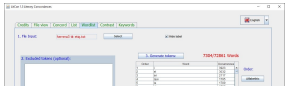
herrera2.txt = *H*
herrera3.txt = *P*
CORPUS

Digitalisation process:
1. OCR through Blecua's annotated edition
2. Generation of .docx and .txt files
3. Revision and correction of OCR and line break mistakes: â, à, á
4. Resulting format: UTF-8

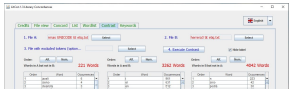
PROCESSING WITH LITCON

A Java software tool specifically developed to work with poetry

View of *H* corpus

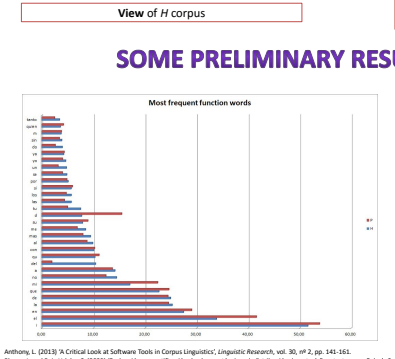


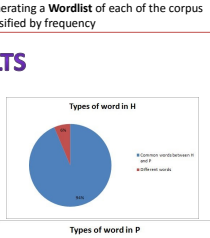
Generating a **Wordlist** of each of the corpus classified by frequency

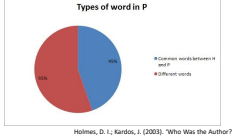


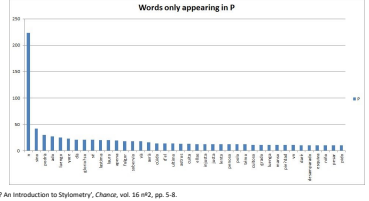
Contrast function: *H* and *P* are processed and the software show us common words between them and words which appear only in one of them, classified by frequency

SOME PRELIMINARY RESULTS









Anthony, L. (2013) 'A Critical Look at Software Tools in Corpus Linguistics', *Linguistic Research*, vol. 30, nº 2, pp. 141-161.
Blecua, J. and Ruiz Urbón, C. (2009) 'Evaluación y cuantificación de algunas técnicas de "atribución de autoría" en textos españoles', *Castilla. Estudios de Literatura*, pp. 27-49.
Blecua, J. M. (ed.) and Herrera, F. (1975) *Obra poética. Volumes I and II*, Madrid: Real Academia Española.
Cuevas, C. (ed.) and Herrera, F. de (1985): *Poesía castellana original completa*, Madrid: Cátedra.
Eder, M. (2010) 'Does size matter? Authorship attribution, small samples, big problem', *Proceedings of Digital Humanities*, pp. 132-135.
Holmes, D. I., Kardos, J. (2003). 'Who Was the Author? An Introduction to Stylometry', *Chance*, vol. 16 nº2, pp. 5-8.
Jockers, M. L. (2013) *Macroanalysis: digital methods and literary history*, Urbana: University of Illinois Press.
Kassal, A. D. (1966) *Vocabulario de la obra poética de Herrera*, Madrid: RAE.
Macri, O. (1972). *Fernando de Herrera*, Madrid: Gredos.
McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*, Cambridge: Cambridge University Press.
McEnery, T. and Wilson, A. (2005) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
Moretti, F. (2007). *La literatura vista desde lejos*, Barcelona: Marbot Ediciones.

josemanuel.fradejas@uva.es

@JMFradeRue

7partidas.hypotheses.org

github.com/7partidasdigital

FFI2016-75014-P AEI-FEDER, EU