STYLOMETRIC ANALYSIS OF GREEK, USING VOCABULARY, GRAMMAR, AND SYNTAX

Eleni Bozia, Ph.D., Dr. Phil.
Assistant Professor
Department of Classics
University of Florida

SunoikisisDC Digital Classics
Summer Semester 2019
May 9th, 2019

- The Historian's Macroscope: Big Digital History
- Introduction to Visualization
- Practical Issues in Visualization

TEXT ANALYSIS & VISUALIZATION

- http://www.online-utility.org/text/analyzer.jsp
- https://www.jasondavies.com/wordcloud/#
- www.wordsift.com
- http://textexture.com/index.php
- http://voyant-tools.org/tool/Bubbles/

FROM THE ALBUM, AMERICAN TEEN (2017)

Number of characters (including spaces) :	22872
Number of characters (without spaces) :	17212
Number of words :	4316
Lexical Density :	15.0139
Number of sentences :	27
Number of syllables :	5917

Some top phrases containing 8 words (without punctuation marks)	Occurrences
me alive keep me alive keep me alive	15
keep me alive keep me alive keep me	15
promise that you'll keep my love with ya	12
alive keep me alive keep me alive keep	12
dumb young young dumb and broke young dumb	9
young dumb young young dumb and broke young	9
keep your number saved 'cause i hope one	8
i'll keep your number saved 'cause i hope	8
the one for me you weren't the one	8
weren't the one for me you weren't the	8
your number saved 'cause i hope one day	8
for me you weren't the one for me	7
you weren't the one for me you weren't	7
that you'll keep my love with ya promise	7
one for me you weren't the one for	7
do do do do do do do do	7
young young dumb and broke young dumb young	6
burning burning burning dadadadada turning turning turning turning	6
my love with ya promise that you'll keep	6
with ya promise that you'll keep my love	6
got my heart now why won't you stop	6
love with ya promise that you'll keep my	6
burning burning dadadadada turning turning turning turning dadadadada	6
promise promise promise promise promise promise promise promise	6
me you weren't the one for me you	6
keep my love with ya promise that you'll	6
all the stupid shit that young kids do	6

1001 'GREATEST' MOVIES OF ALL TIME

Number of characters (including spaces) :	17450
Number of characters (without spaces) :	13479
Number of words :	2841
Lexical Density :	53.6079
Number of sentences :	48
Number of syllables :	4494

Some top phrases containing 7 words (without punctuation marks)	Occurrences
harry potter and the deathly hallows part	2

Some top phrases containing 6 words (without punctuation marks)	Occurrences
the lord of the rings the	3
the girl with the dragon tattoo	2
potter and the deathly hallows part	2
of the planet of the apes	2
harry potter and the deathly hallows	2

Some top phrases containing 5 words (without punctuation marks)	Occurrences
once upon a time in	3
the lord of the rings	3
lord of the rings the	3
the girl with the dragon	2
girl with the dragon tattoo	2
potter and the deathly hallows	2
harry potter and the deathly	2
how to train your dragon	2
of the planet of the	2
and the deathly hallows part	2
the planet of the apes	2

Some top phrases containing 3 words (without punctuation marks)	Occurrences
harry potter and	4
potter and the	4
planet of the	3
the rings the	3
the man who	3
a time in	3
once upon a	3
lord of the	3
of the rings	3
star wars episode	3
of the apes	3
the lord of	3
upon a time	3
of the planet	2
girl with the	2
to train your	2
dawn of the	2
indiana jones and	2
the hole the	2
jones and the	2
of the dead	2
with the dragon	2
night of the	2
the planet of	2
the girl with	2
the dark knight	2
the wind the	2
life and death	2
and the deathly	2
the curse of	2
and the beast	2
return of the	2
deathly hallows part	2
kill bill vol	2
the wrath of	2
train your dragon	2
how to train	2
the dragon tattoo	2
the deathly hallows	2
beauty and the	2
curse of the	2

Some top phrases containing 4 words (without punctuation marks)	Occurrences
harry potter and the	4
of the rings the	3
upon a time in	3
the lord of the	3
once upon a time	3
lord of the rings	3
planet of the apes	3
the planet of the	2
of the planet of	2
the girl with the	2
girl with the dragon	2
beauty and the beast	2
potter and the deathly	2
indiana jones and the	2
how to train your	2
with the dragon tattoo	2
the deathly hallows part	2
and the deathly hallows	2
the curse of the	2
to train your dragon	2

Most frequent words in the corpus:
man (24);
day (10);
life (10);
night (10);
dead (8)

https://voyant-tools.org/?corpus=02e7c0ca49b7be6160b3e876cb978097&view=TextualArc

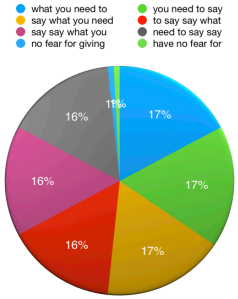
Genres Analysis

action (125) adventure (142) animation (57) biography (86)
comedy (223) crime (205) drama (706) family (70)
fantasy (84) fi (81) film (26) history (58) horror (52) musical (66) mystery (99)
romance (201) sci (81) thriller (227) war (85) western (32)

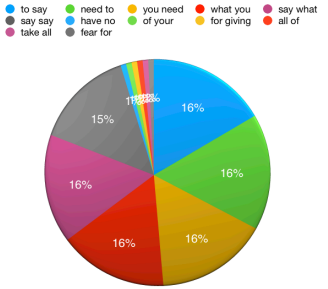
Order	**Unfiltered word count**	**Occurrences**	**Percentage**
1.	drama	706	26.6818
2.	thriller	227	8.5790
3.	comedy	223	8.4278
4.	crime	205	7.7475
5.	romance	201	7.5964
6.	adventure	142	5.3666
7.	action	125	4.7241
8.	mystery	99	3.7415
9.	biography	86	3.2502
10.	war	85	3.2124
11.	fantasy	84	3.1746
12.	sci_fi	81	3.0612
13.	family	70	2.6455
14.	history	58	2.1920
15.	animation	57	2.1542
16.	horror	52	1.9652
17.	musical	36	1.3605
18.	western	32	1.2094
19.	music	30	1.1338
20.	film_noir	26	0.9826
21.	sport	21	0.7937

Say What you need to say:

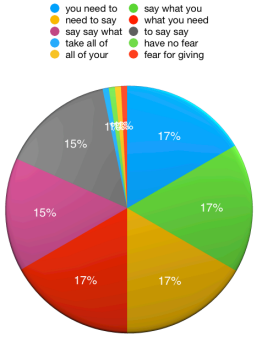
Top phrases containing 4 words	Occurrences
what you need to	41
you need to say	41
say what you need	41
to say say what	37
say say what you	37
need to say say	37
no fear for giving	2
have no fear for	2



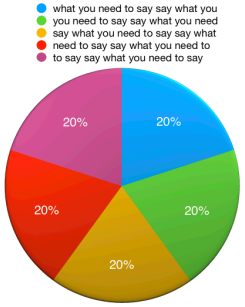
Top phrases containing 2 words	Occurrences
to say	42
need to	41
you need	41
what you	41
say what	41
say say	37
have no	2
of your	2
for giving	2
all of	2
take all	2
fear for	2



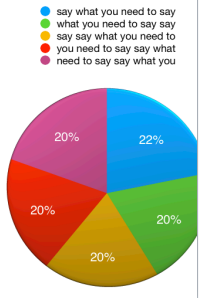
Top phrases containing 3 words	Occurrences
you need to	41
say what you	41
need to say	41
what you need	41
say say what	37
to say say	37
take all of	2
have no fear	2
all of your	2
fear for giving	2



Top phrases containing 8 words	Occurrences
what you need to say say what you	37
you need to say say what you need	37
say what you need to say say what	37
need to say say what you need to	37
to say say what you need to say	37

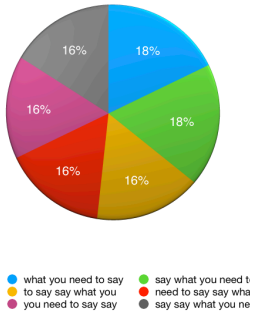


Top phrases containing 6 words	Occurrences
say what you need to say	41
what you need to say say	37
say say what you need to	37
you need to say say what	37
need to say say what you	37



Top phrases containing 7 words	Occurrences
need to say say what you need	37
say say what you need to say	37
say what you need to say say	37
what you need to say say what	37
you need to say say what you	37

Top phrases containing 5 words	Occurrences
what you need to say	41
say what you need to	41
to say say what you	37
need to say say what	37
you need to say say	37
say say what you need	37



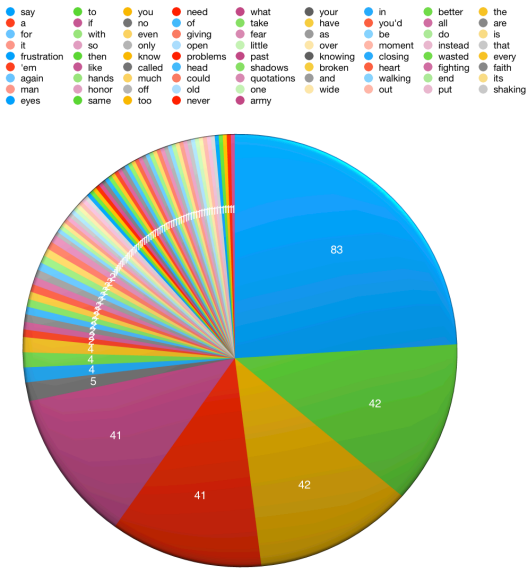
Word Count: 349

Data:

Word count:	349
say	83
to	42
you	42
need	41
what	41
your	5
in	4
better	4
the	4
a	2
if	2
no	2
of	2
take	2
have	2
you'd	2
all	2
are	2
for	2
with	2
even	2
giving	2
fear	2
as	1
be	1

do	1
is	1
it	1
so	1
only	1
open	1
little	1
over	1
moment	1
instead	1
that	1
frustration	1
then	1
know	1
problems	1
past	1
knowing	1
closing	1
wasted	1
every	1
'em	1
like	1
called	1
head	1
shadows	1
broken	1

heart	1
fighting	1
faith	1
again	1
hands	1
much	1
could	1
quotations	1
and	1
walking	1
end	1
its	1
man	1
honor	1
off	1
old	1
one	1
wide	1
out	1
put	1
shaking	1
eyes	1
same	1
too	1
never	1
army	1



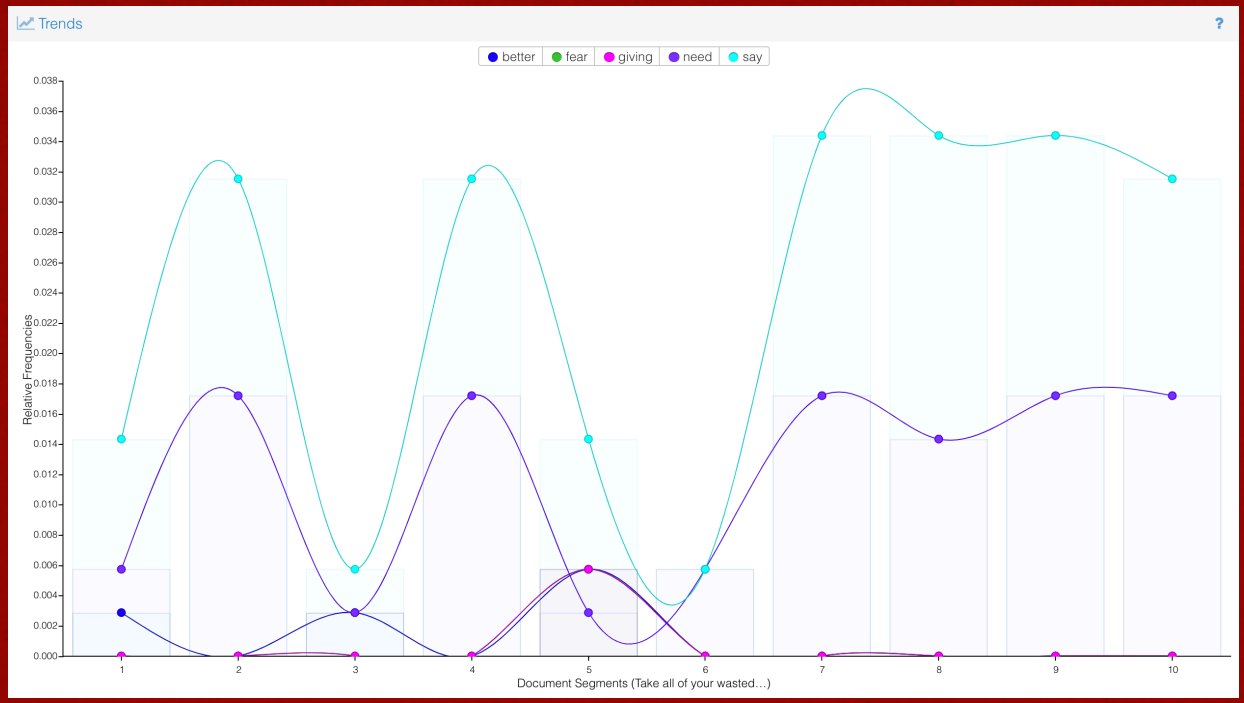
Say What You Need To Say

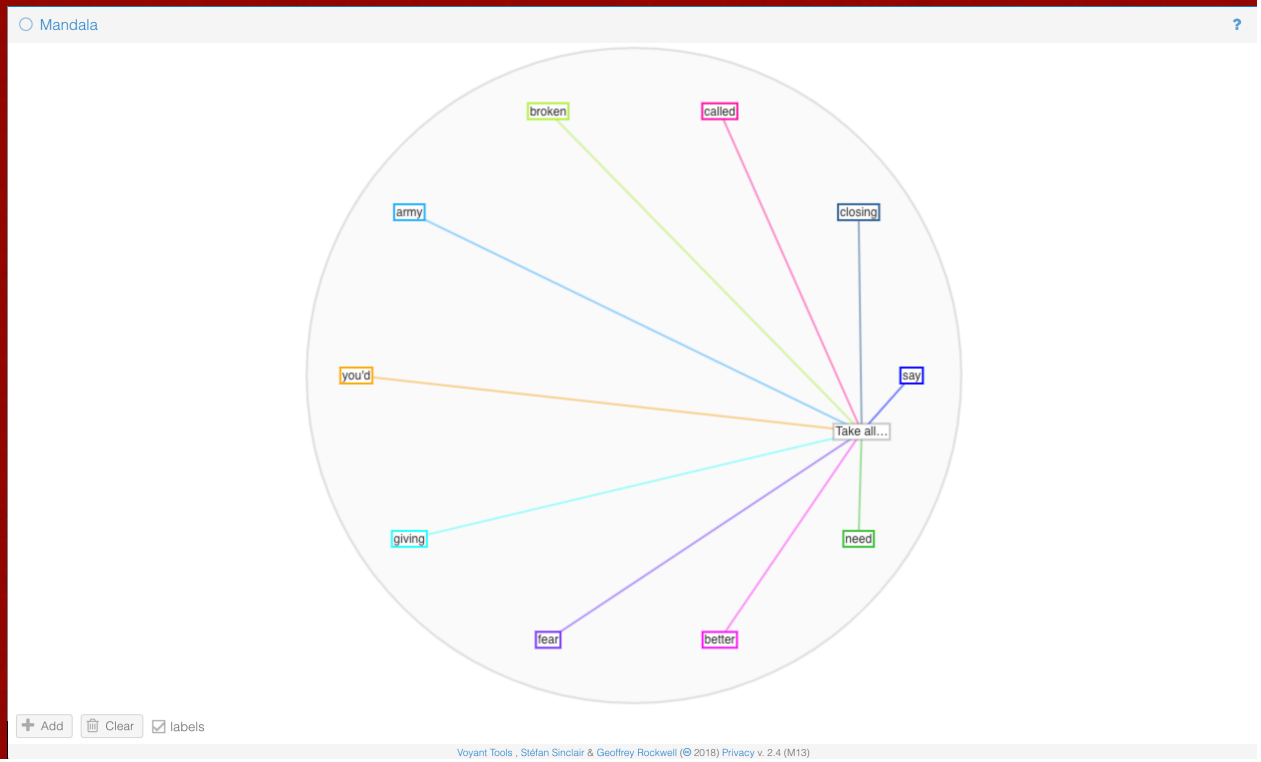
2001 By: John Mayer

VOYANT

see through your text

Say What You Need To Say:





STYLOMETRY

HISTORY

- Lorenzo da Valla in 1439

The Donation of Constantine (Latin: Donatio Constantini) is a forged Roman imperial decree by which the 4th century emperor Constantine the Great supposedly transferred authority over Rome and the western part of the Roman Empire to the Pope. Composed probably in the 8th century, it was used, especially in the 13th century, in support of claims of political authority by the papacy.

The basics of stylometry were set out by Polish philosopher Wincenty Lutos□awski in Principes de stylométrie (1890)

HISTORY

- Authorship of Ronald Reagan's Radio Addresses
- "Double Falsehood"

- The Style of Numbers behind a Number of Styles
- Making Hit Music into Science
- Forensic Linguistics
- Forensic Analysis of Instant Messaging
- Deception in Instant Messaging

METHODS

- Writer Invariant

Property of a text that is invariant of its authors

Word Lengths

Sentence Length

Average Word Length

Noun, Verb or adverb usage frequency

Vocabulary Richness

Frequency of Function Words

METHODS

- Neural Networks (70% of precision)
- Genetic Algorithms
- Rare Pairs (based on collocation)

TOOLS

- JAVA Graphical Authorship Attribution program
- The Signature Stylometric System
- Stylene (a stylometric system for Dutch)
- Stylo
- Deceiving Authorship Detection: Tools to maintain anonymity and current trends in adversarial stylometry

STYLO

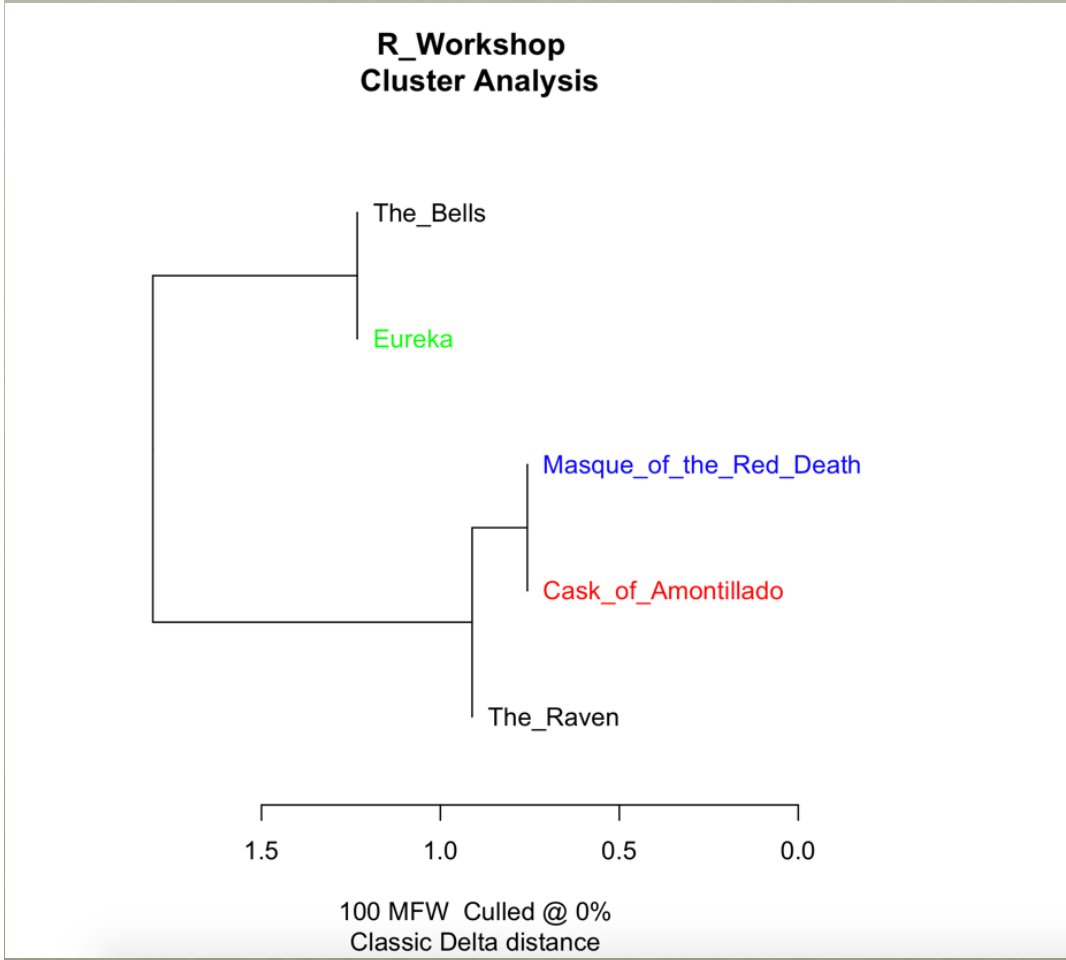
Instructions: How to set up R and stylo

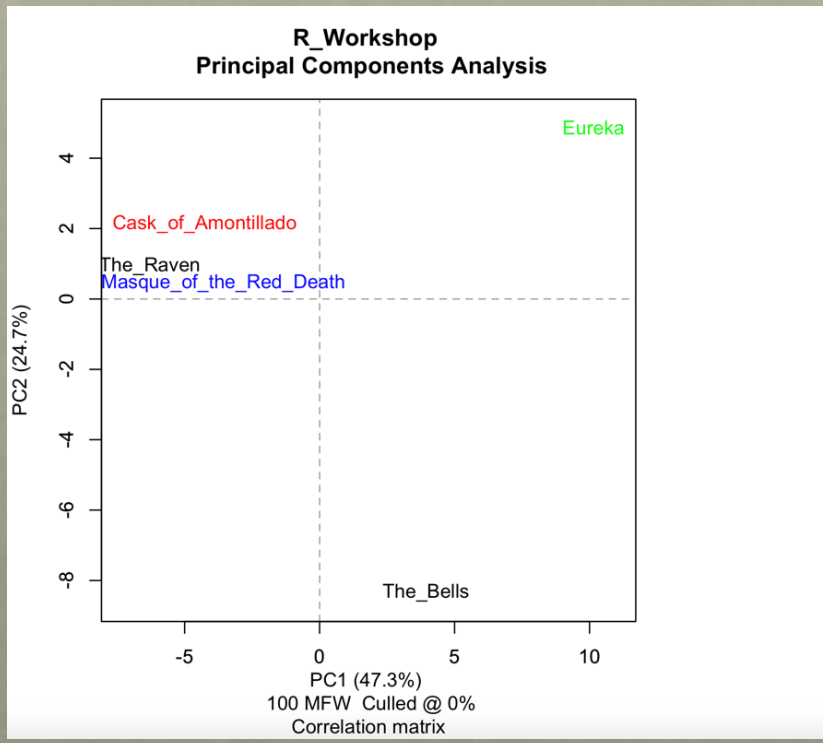
- R is a programming environment for computations, analysis, and statistics of data. It provides functions to parse data, store them as variables or arrays, and perform calculations.
- Download R for your operating system
- https://cran.cnr.berkeley.edu/
- If you use Mac, you need also the following
- http://www.xquartz.org/
- Open R and type
- install.packages("stylo")
- Activate the package (whenever you start a fresh R session):

- library(stylo)
- A manual to stylometry using "stylo" is available here.
- Setting working Directory
- 1) through the menu
- In *Windows*: go to the *File* menu, select *Change Working Directory*, and select the appropriate folder/directory
- In *Macs*: go to the *Misc* menu, select *Change Working Directory*, and select the appropriate folder/directory

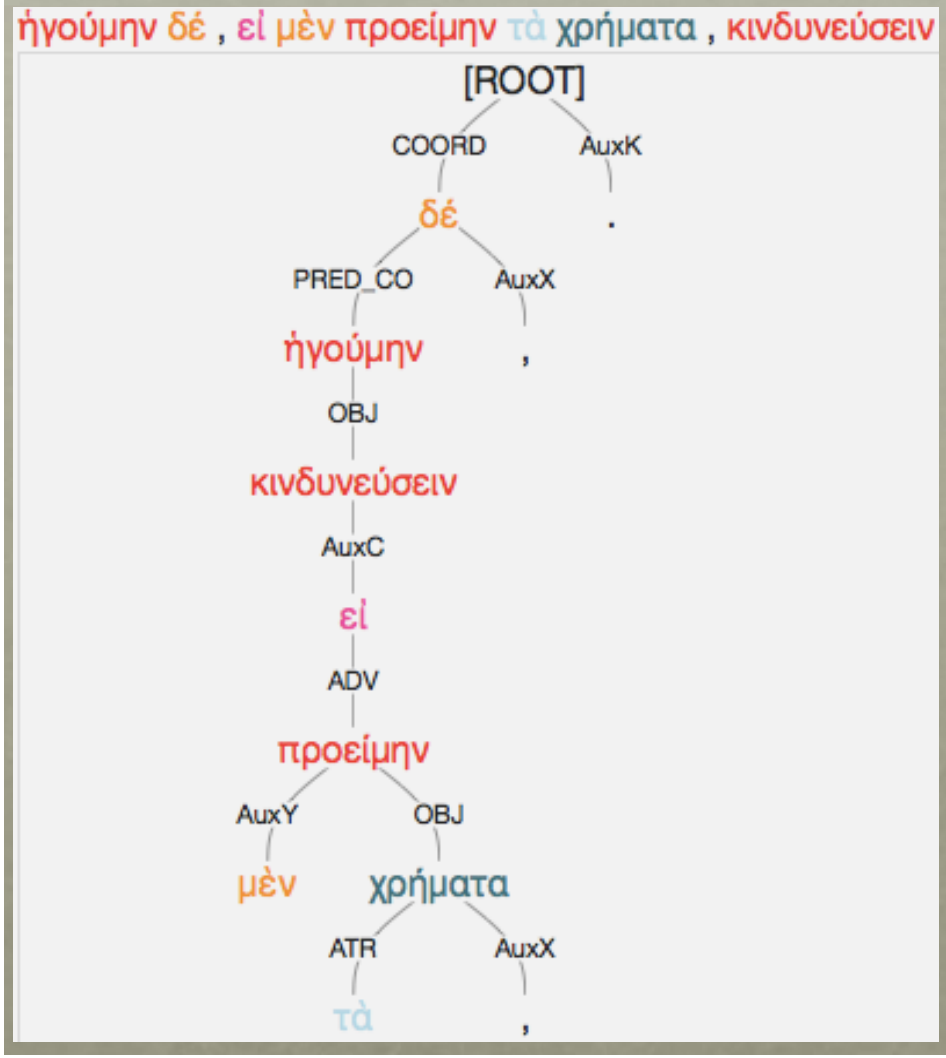
- Inside your Working Directory create a folder “corpus”
- In the “corpus” folder create at least 2 text files with the documents you want to analyze.
- In Windows, open Notepad
- In Mac, open text edit
- Go to text edit preferences and make sure “plain text” is selected under new document tab.
- Create a new document and paste some content.

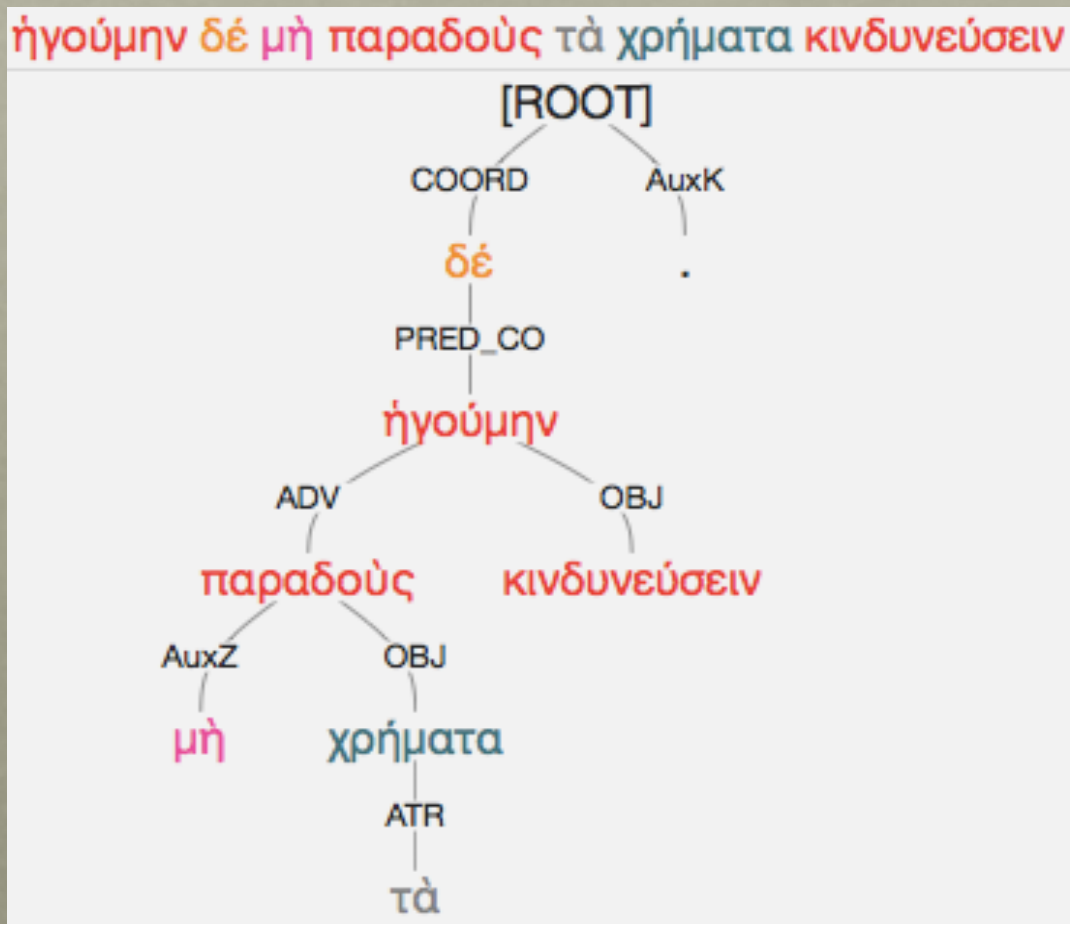
EDGAR ALLAN POE



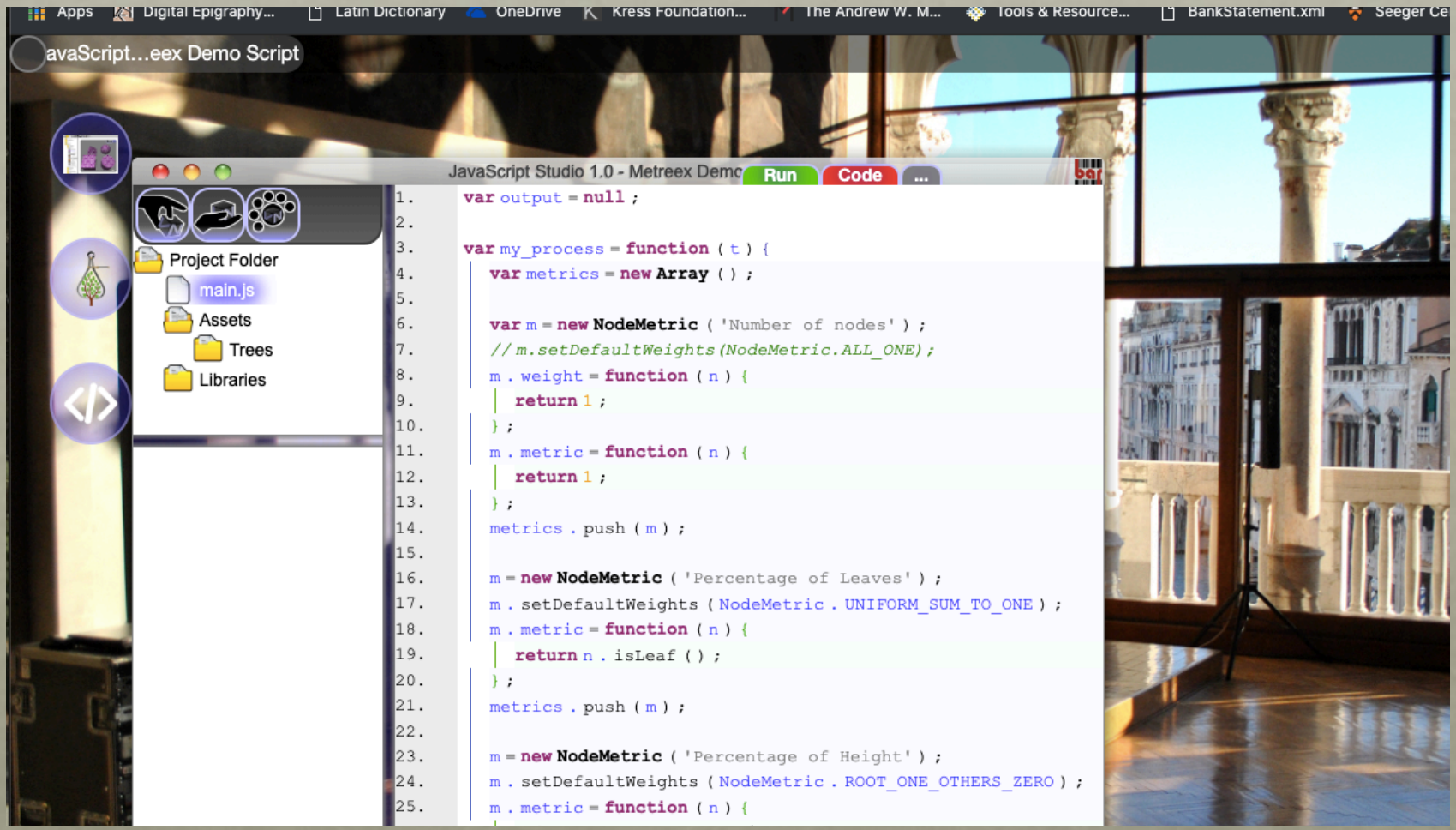


WWW.PERSEIDS.ORG





CUSTOMIZED METRICS

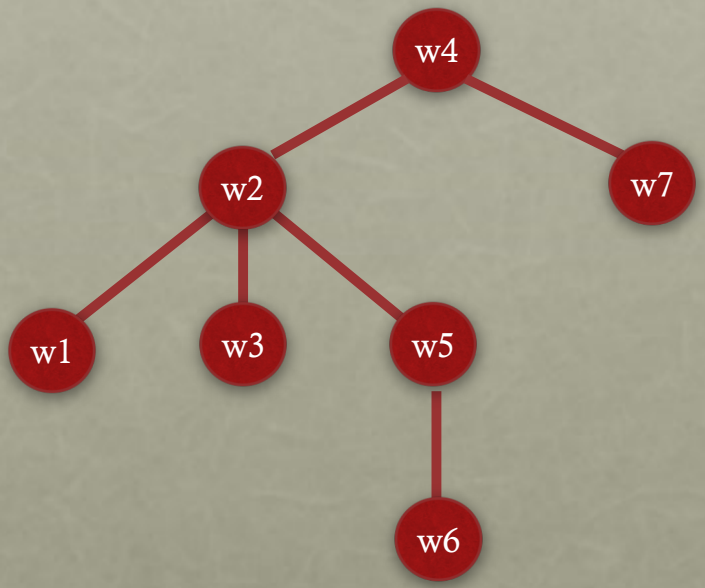


www.metreex.org

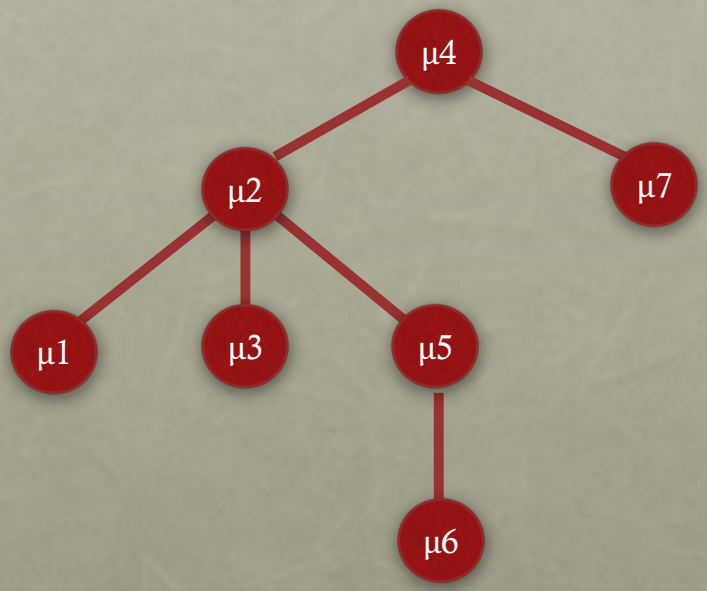
NODE-BASED METRICS

- To calculate a sentence metric we need: w and μ

Weights



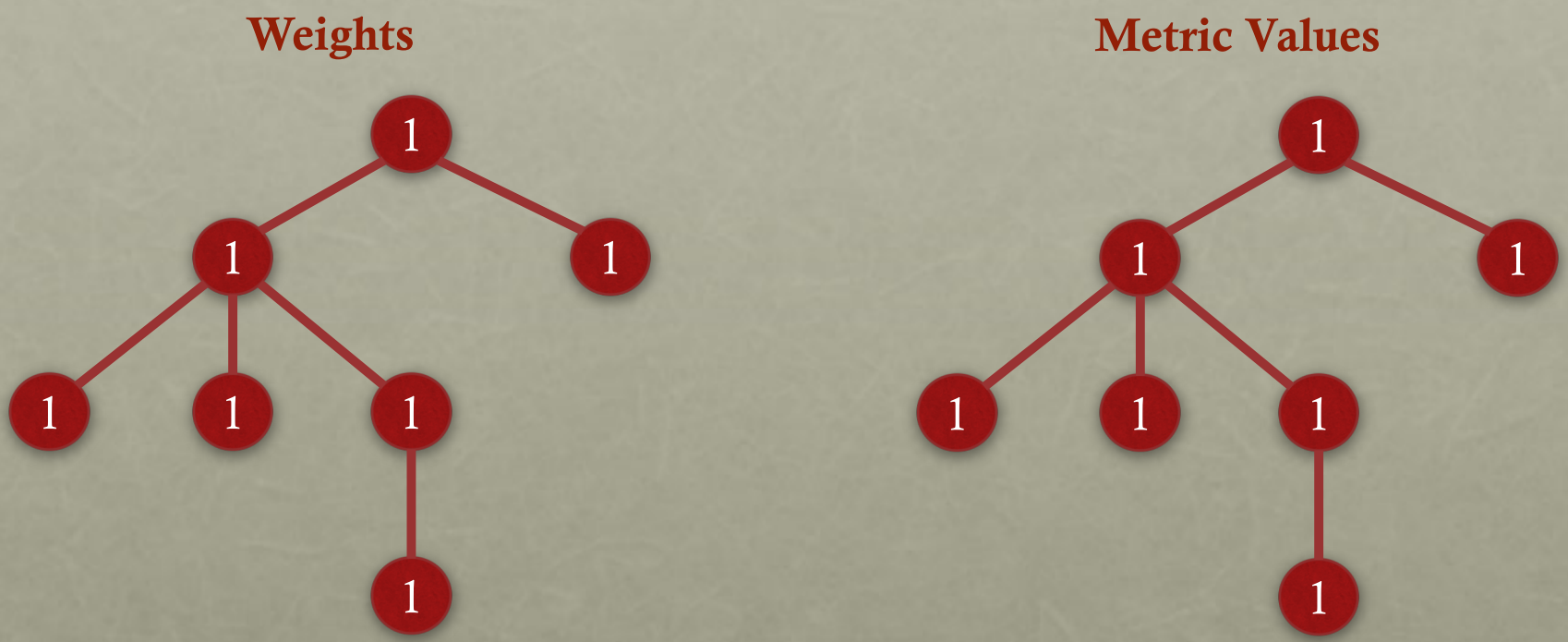
Metric Values



- Result: w1xμ1+w2xμ2+w3xμ3+w4xμ4+w5xμ5+w6xμ6+w7xμ7

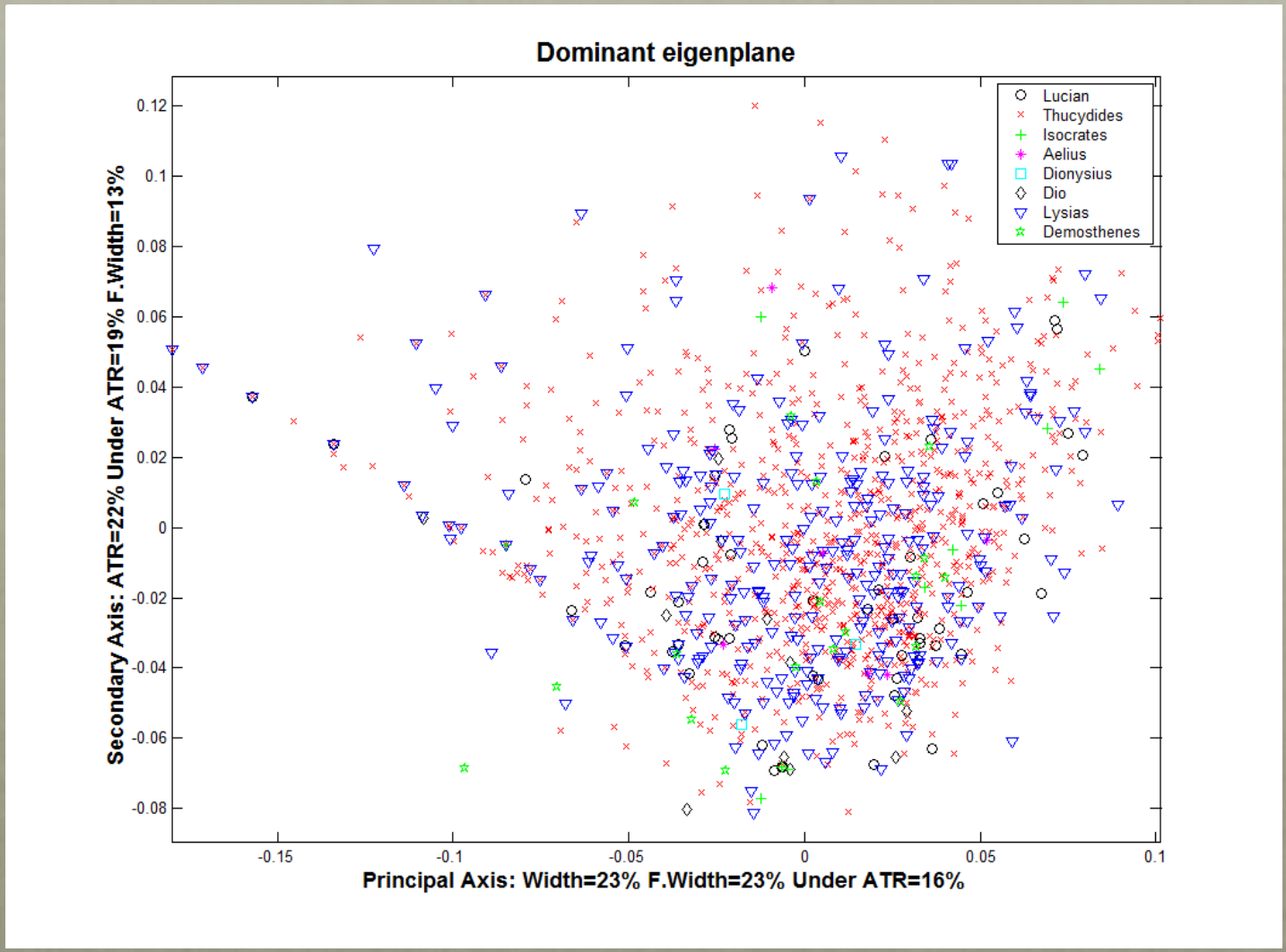
NODE-BASED METRICS

- Metric example: *Number of words*

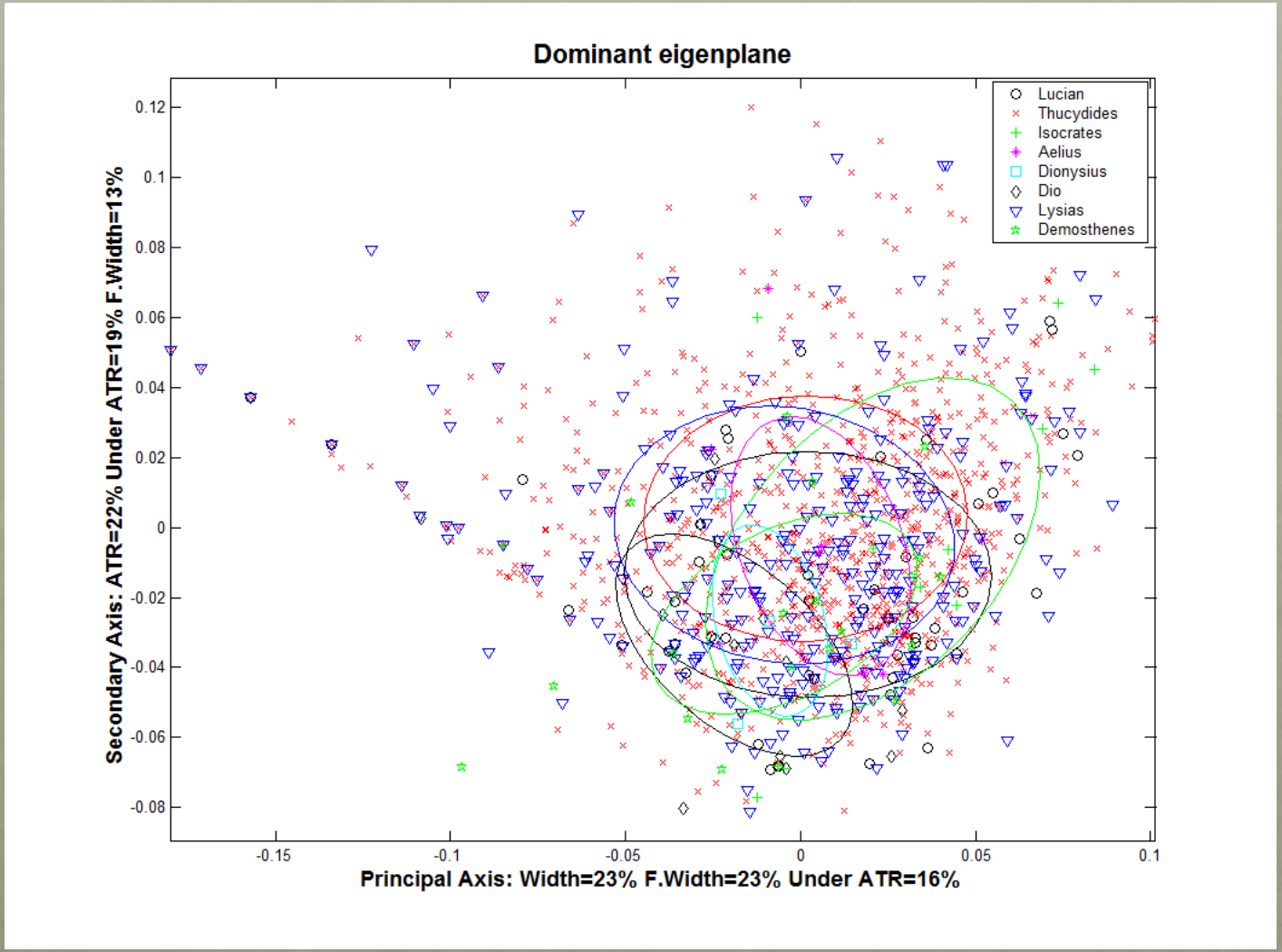


- Result: 1x1+1x1+1x1+1x1+1x1+1x1+1x1=7

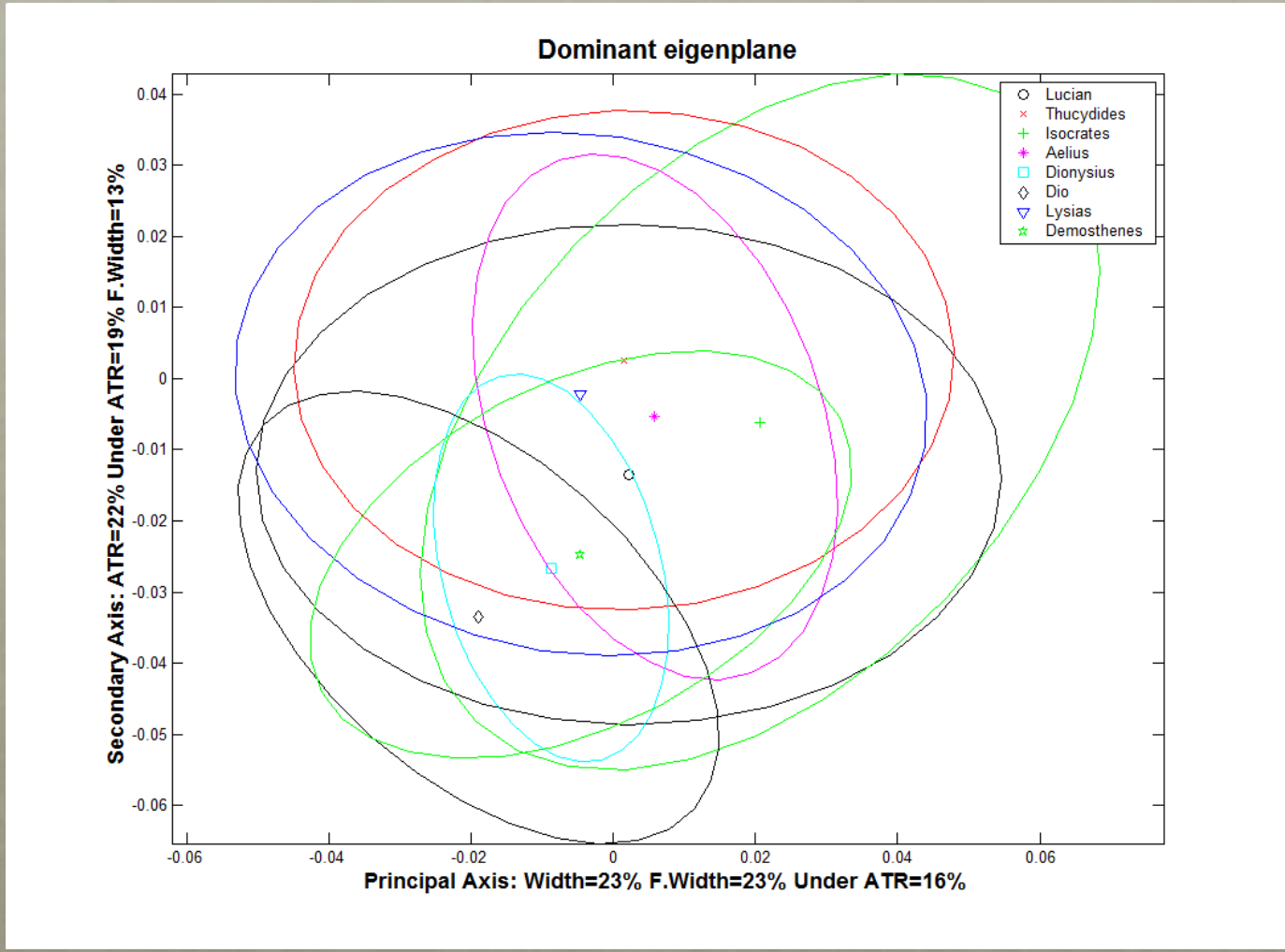
ASPECTS OF CLASSICAL & IMPERIAL ORATORY



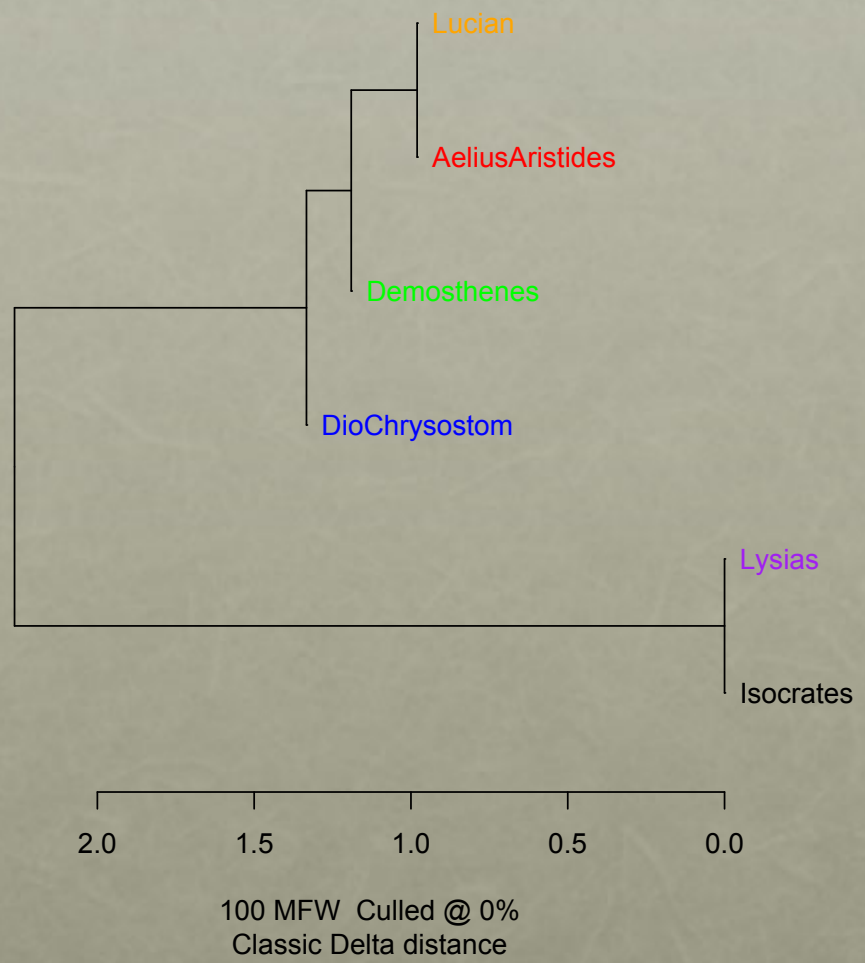
ASPECTS OF CLASSICAL & IMPERIAL ORATORY



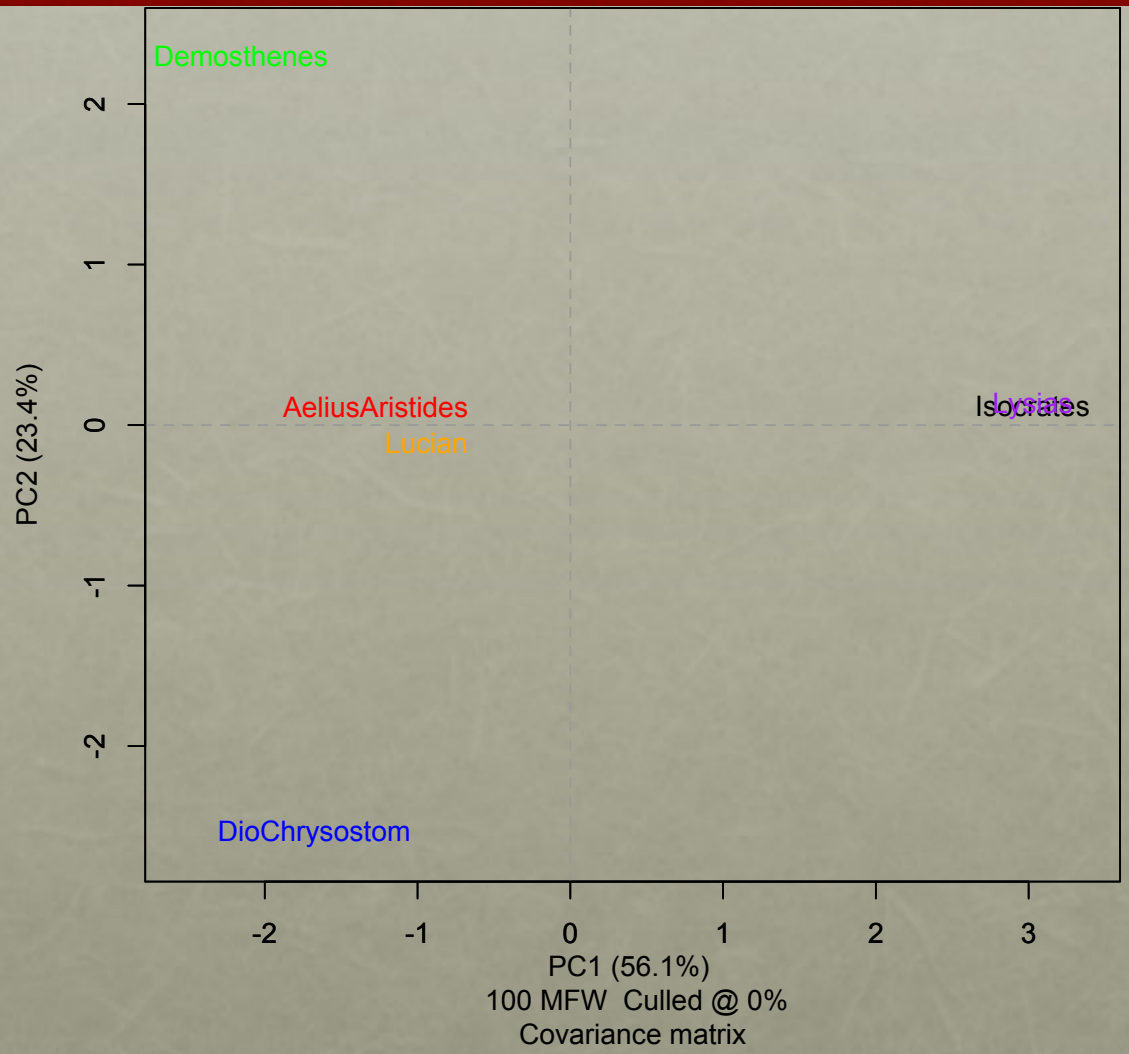
ASPECTS OF CLASSICAL & IMPERIAL ORATORY



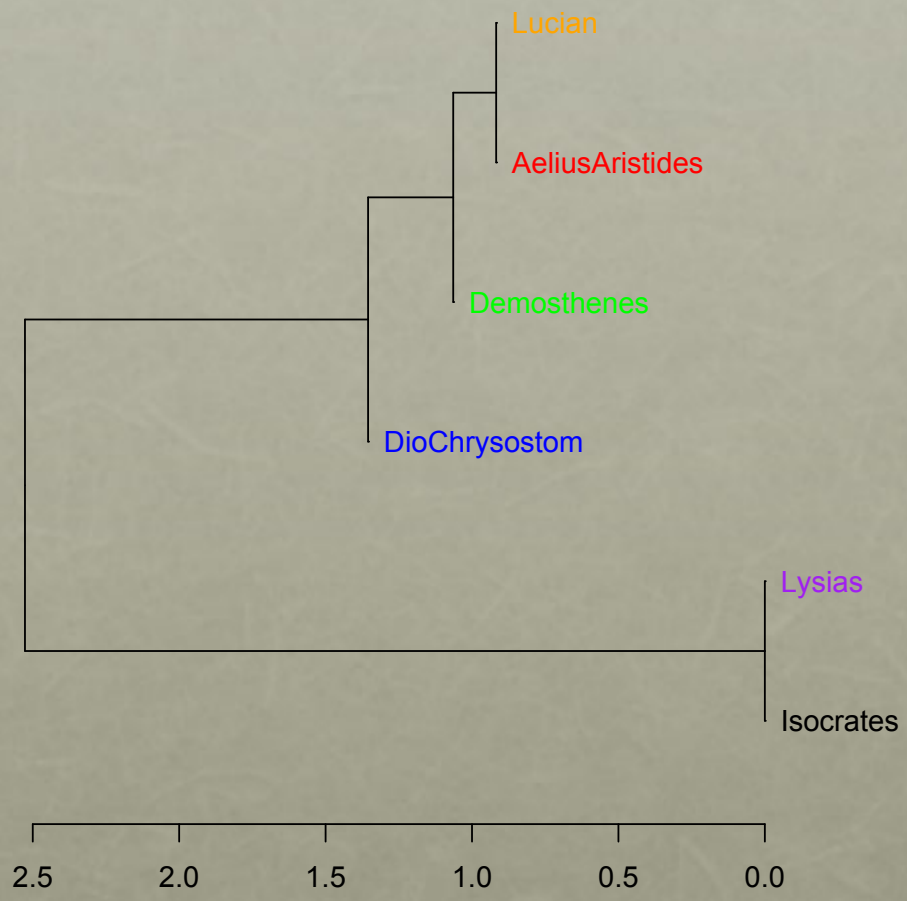
Desktop
Cluster Analysis



Desktop
Principal Components Analysis



Desktop
Cluster Analysis



THANK YOU!