

# TEXTUAL ANALYSIS— TREEBANK ANNOTATION, SYNTACTICAL & LINGUISTIC METRICS

Eleni Bozia

Lehrstuhl für Digital Humanities  
Institut für Informatik  
Universität Leipzig

Department of Classics & Digital Worlds Institute  
University of Florida

A LANGUAGE IS NOT JUST WORDS. IT'S A CULTURE, A TRADITION, A UNIFICATION OF A COMMUNITY, A WHOLE HISTORY THAT CREATES WHAT A COMMUNITY IS. IT'S ALL EMBODIED IN A LANGUAGE.

*Noam Chomsky*

# FIRST OR SECOND OR...ATTICISM

- δεικνῦσι προπερισπωμένως Ἀττικοί, δεικνύεσιν Ἕλληνες.  
Δεικνύασι δὲ οἱ δεύτεροι Ἀττικοί (Moeris, 194.29).
- χολάδας οἱ πρῶτοι Ἀττικοί, χόλικας θηλυκῶς οἱ μέσοι.  
“χολικας ἐφθάς.” τὰς χόλικας ἀρσενικῶς Ἕλληνες (Moeris, 213.2)
- Κακοδαιμονεῖν· οὕτως οἱ νόθως ἀττικίζοντες, Ἀθηναῖοι γὰρ διὰ τοῦ α κακοδαιμονᾶν λέγουσιν.” (Phrynicus the Atticist, *Eclogae* 54)

# DIONYSIUS HALICARNASSUS

- We ought to acknowledge a great debt of gratitude to the age in which we live, my most accomplished Ammaeus, for an improvement in certain fields of serious study, and especially for the considerable revival in the practice of civil oratory. In the epoch preceding our own, the old philosophic Rhetoric was so grossly abused and maltreated that it fell into a decline. From the death of Alexander of Macedon it began to lose its spirit and gradually wither away, and in our generation had reached a state of almost total extinction... I think that the cause and origin of this great revolution has been the conquest of the world by Rome, who has thus made every city focus its entire attention upon her. (Dion. Hal. *Orat. Vett.* 1-3. Trans. Usher)

He is completely **pure in his vocabulary**, and is the perfect **model of the Attic dialect** (Dionysius of Halicarnassus, *Lysias* 2)

Purity of language, correct dialect, the presentation of ideas by means of **standard, not figurative expressions**; **Clarity, brevity, concision**, terseness, vivid representation...But there is nothing sublime or imposing about the style of Lysias. It certainly does not excite us or move us to wonder, nor does it portray pungency , intensity or fear; (Dionysius of Halicarnassus, *Lysias* 13)

- His style has the following characteristics: it is as pure as that of Lysias; Not a word is used at random; And the language conforms closely to the most ordinary and familiar usage. Like its predecessor, it avoids the banality of archaic and obscure words, **but uses figurative language somewhat more than Lysias**, achieving a happy balance in this respect. In the matter of lucidity and vividness it is similar to that of Lysias; (Dionysius of Halicarnassus, *Isocrates* 2)

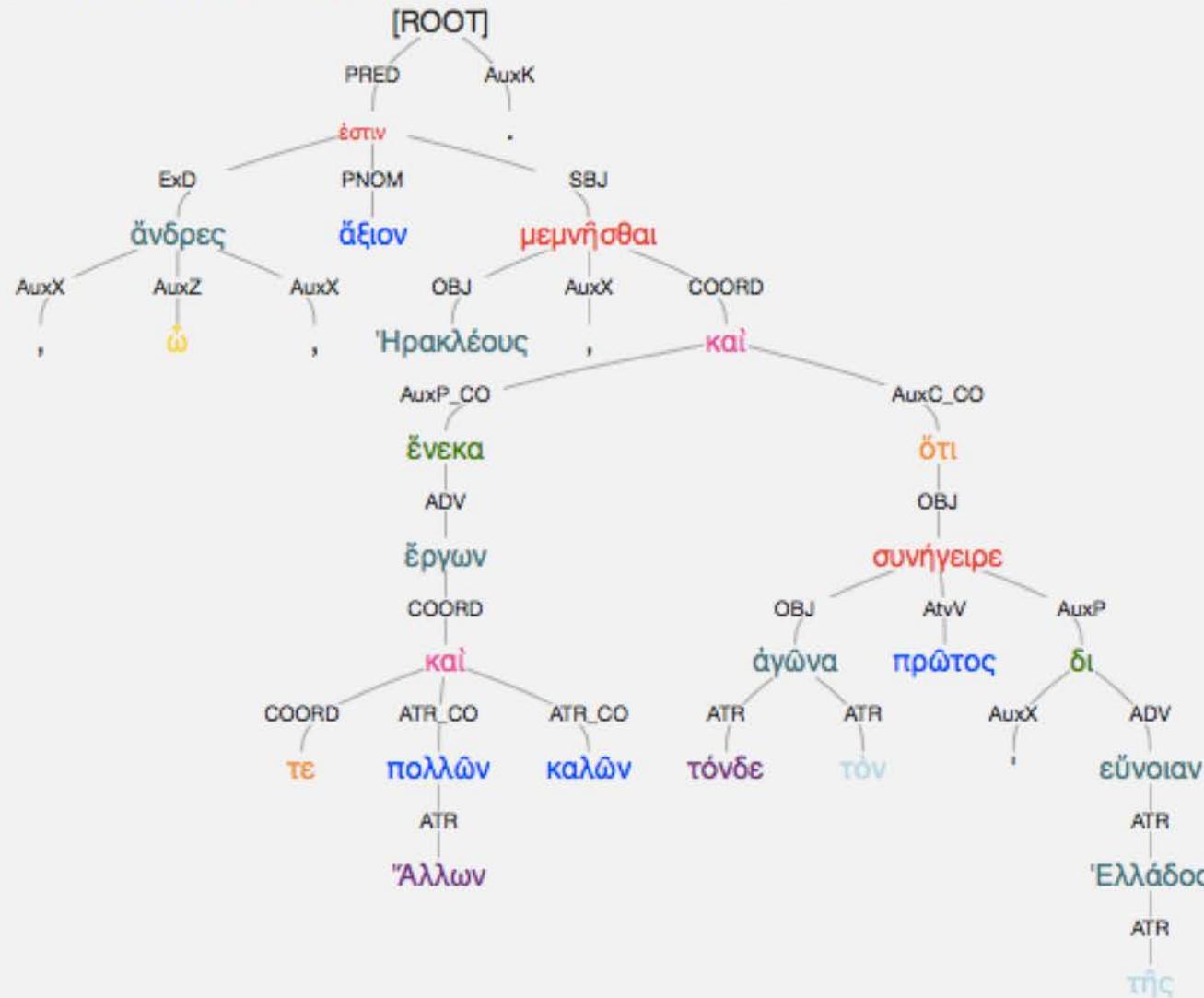
- In the first place, before rounding off the first idea (or clause if it should be called), a second idea is introduced; then a third is subjoined before the second is complete, and material belonging to the second is tacked on after the third has been completed. (Dionysius, *Demosthenes* 9)

- Coming to make you my morning salutation, which should have taken the orthodox form of Rejoice, I bade you, in a very choice fit of absent-mindedness, **Be healthy**—a good enough wish in its way, but a little untimely and unconnected with that early hour. (Luc. *Laps.* 1. All translations of Lucian are by Fowler.)

- To be honest, however, their praise caused me considerable annoyance, and when they had gone and I was left alone, I reflected as follows: “So this is the only attraction in my writings, that they are **unconventional** and keep off the beaten track. While good vocabulary, **conformity to the ancient canon**, penetration of intellect, power of perception, **Attic grace**, good construction, general competence, perhaps have no place in my work.  
(Lucian, *Zeuxis* 2)

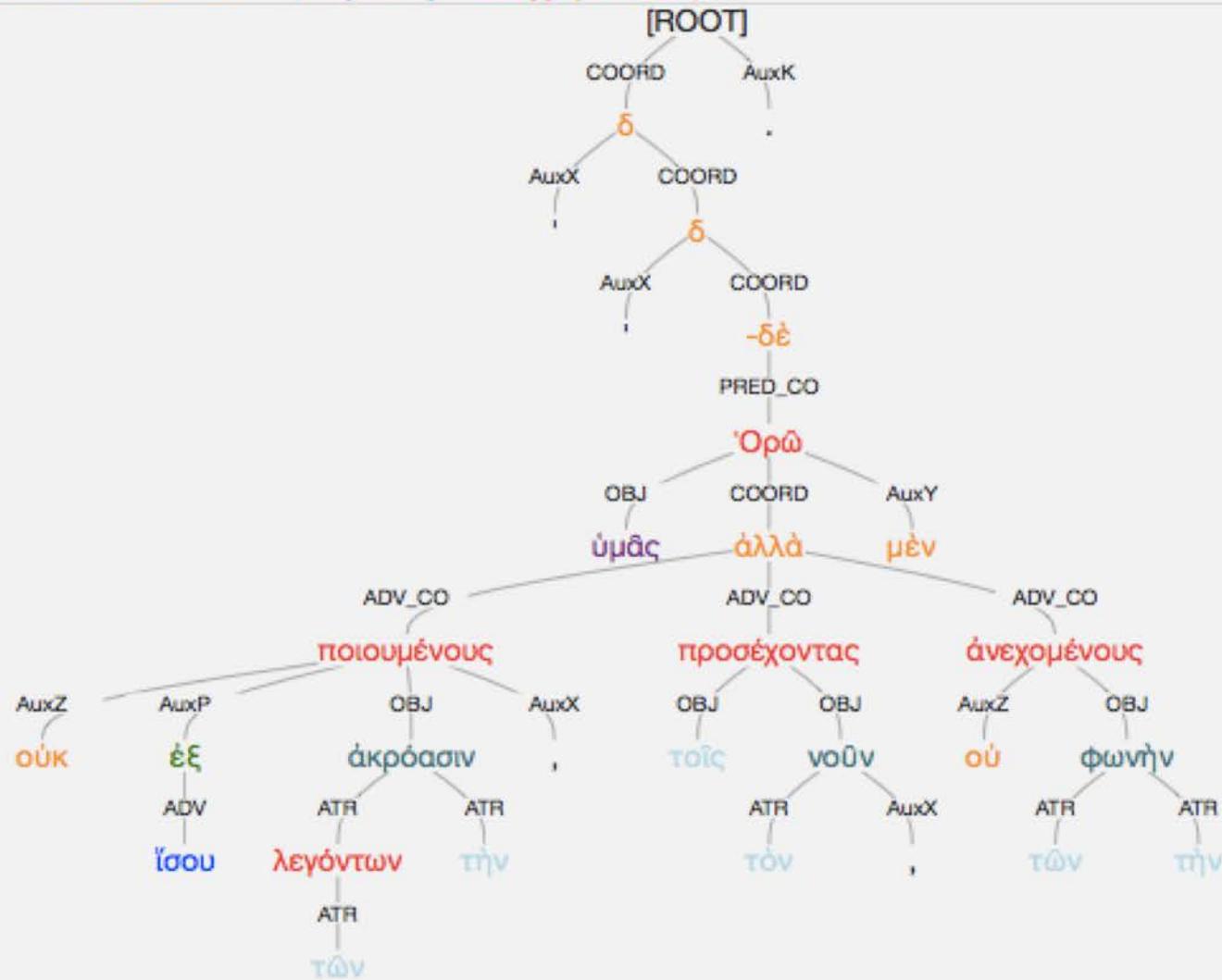
# LYSIAS, OLYMPIAKUS 1

"Άλλων τε πολλῶν καὶ καλῶν ἔργων ἔνεκα , ω̄ ἄνδρες , ἄξιον ἐστιν Ἡρακλέους μεμνῆσθαι , καὶ ὅτι τόνδε τὸν ἀγώνα πρῶτος συνήγειρε δι ' εὔνοιαν τῆς Ἑλλάδος .

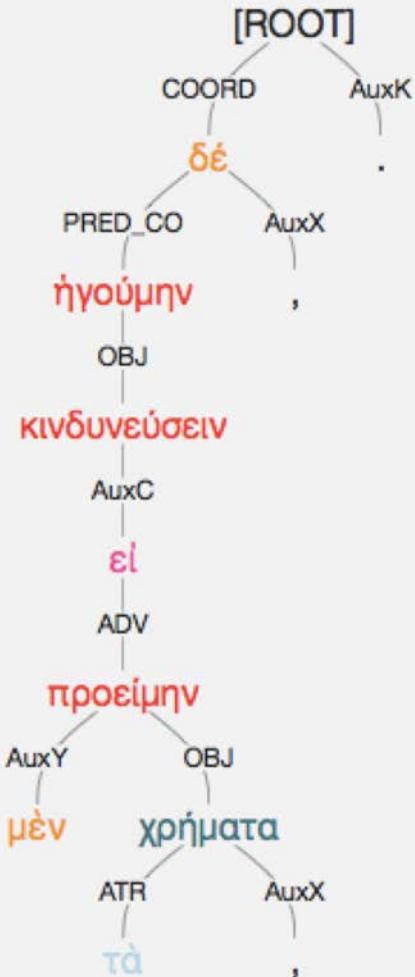


# ISOCRATES, *PANEGYRICUS* 1

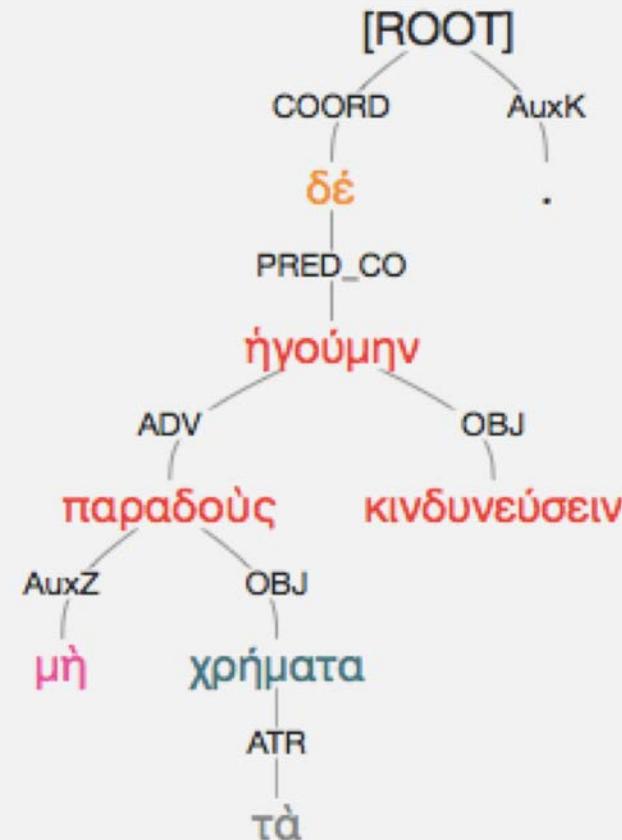
‘Ορῶ δ ' ὑμᾶς οὐκ ἔξ ἴσου τῶν λεγόντων τὴν ἀκρόασιν ποιουμένους , ἀλλὰ τοῖς μὲν προσέχοντας τὸν νοῦν , τῶν δ ' οὐ -δὲ τὴν φωνὴν ἀνεχομένους .



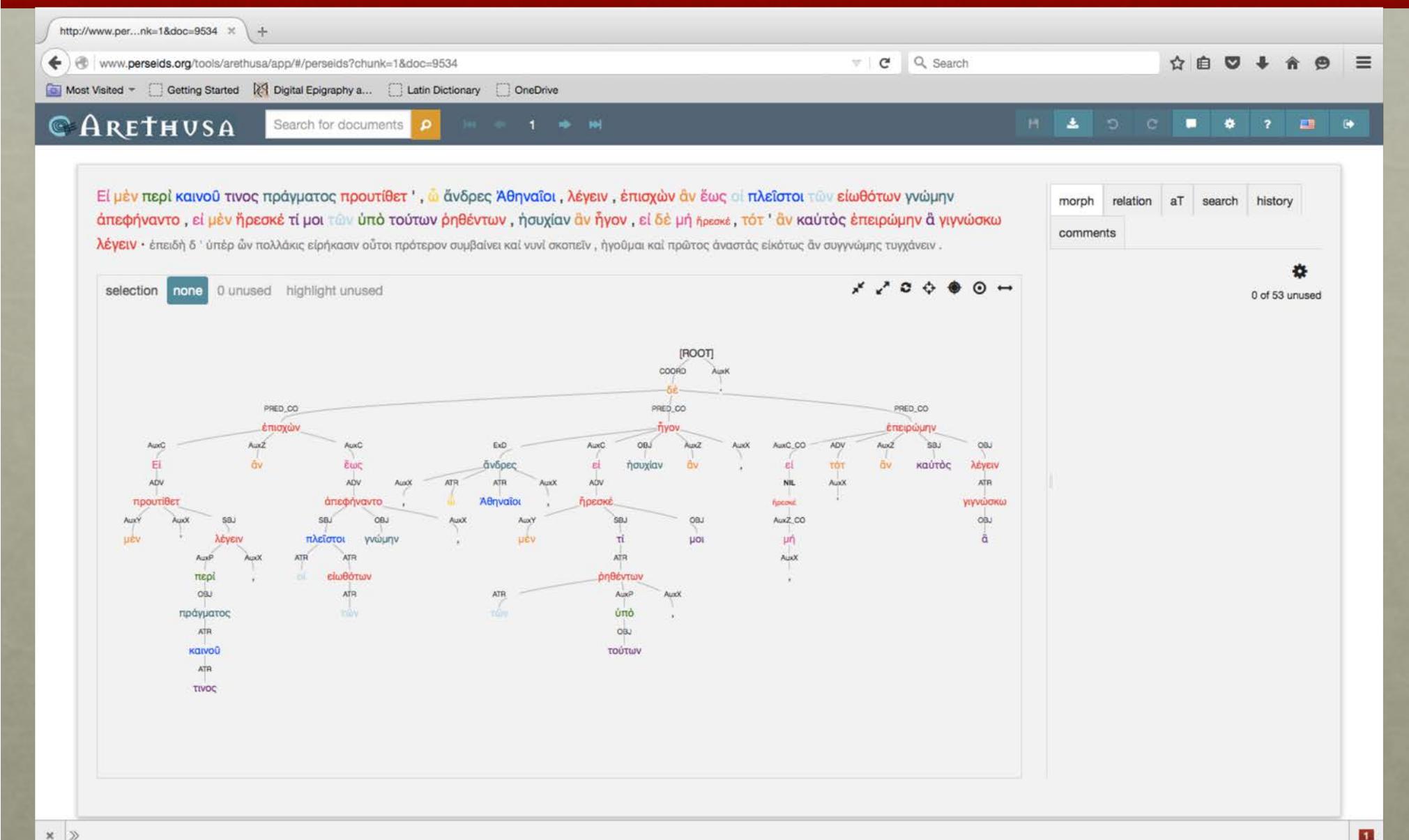
ήγούμην δέ , εἰ μὲν προείμην τὰ χρήματα , κινδυνεύσειν



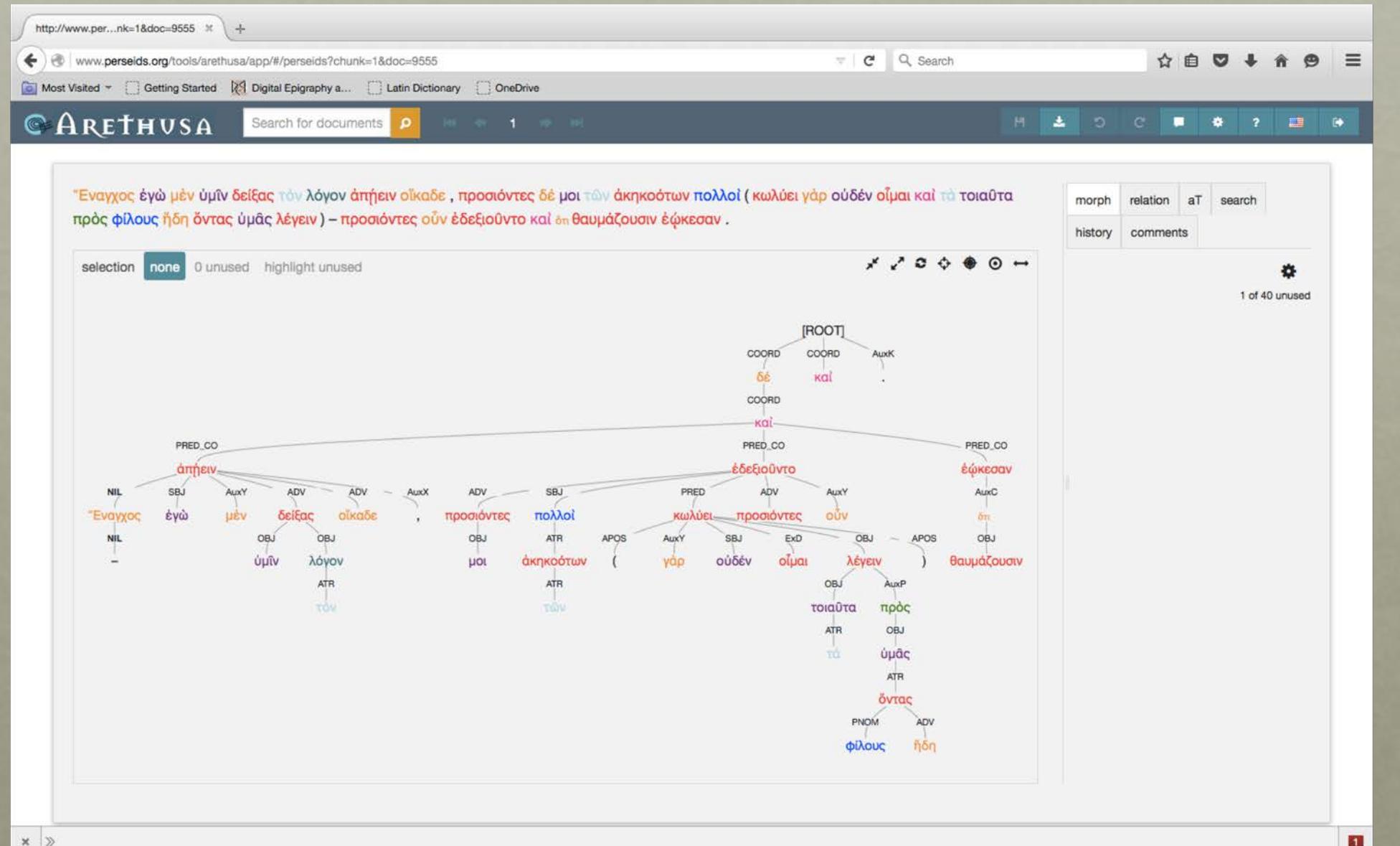
ήγούμην δέ μὴ παραδοὺς τὰ χρήματα κινδυνεύσειν



# DEMOSTHENES, *PHILIPPIC* 1.1



# LUCIAN, ZEUXIS 1



# DIO, ORATIO 42.1

http://www.per...nk=1&doc=9592

www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=9592

Most Visited Getting Started Digital Epigraphy a... Latin Dictionary OneDrive

ARETHUSA Search for documents

"Οπως μὲν ὑμεῖς διανοεῖσθε περὶ ἐμοῦ καὶ τῆς ἐμῆς εἰ̄ -τε σοφίας εἰ̄ -τε ἀμαθίας οὐ δύναμαι ξυμβαλεῖν , πότερον ὄντως ἐπιθυμεῖτε τῶν λόγων , ὡς ἀκουσόμενοι τι θαυμαστὸν καὶ τοιοῦτον ὅποιον οὐκ ἂν ἔτέρου τῶν νῦν , ἢ τι ούναντιον , ὡς ἐμὲ ἔξελέγοντες καὶ ἀποδείξοντες οὐθὲν μέγα οὐ̄ -δε σπουδαῖον ἐπιστάμενον .

morph relation aT search history  
comments

selection none 0 unused highlight unused

0 of 54 unused

[ROOT]

δύναμαι

οὐ δύναμαι ξυμβαλεῖν

πότερον

ούναντιον

οὐθὲν μέγα σπουδαῖον

ἐπιθυμεῖτε

διανοεῖσθε

μένεις

ἐμοῦ

τῆς ἐμῆς

εἰ̄ -τε σοφίας

εἰ̄ -τε ἀμαθίας

ἀκουσόμενοι

λόγων

ώς

τι

ἔξελέγοντες

ἀποδείξοντες

οὐθὲν μέγα σπουδαῖον

οὐ

οὐδεῖς

τις

θαυμαστὸν

τοιοῦτον

ὅποιον

ἄν

οὐκ

ἔτέρου

τις

τις

# AELIUS ARISTIDES, *ROME*

http://www.perseids.org/tools/arethusa/app#/perseids?chunk=1&doc=9651

www.perseids.org/tools/arethusa/app#/perseids?chunk=1&doc=9651

Most Visited Getting Started Digital Epigraphy Latin Dictionary OneDrive

ARETHUSA Search for documents

άλλι', ὁ ἄνδρες, οἱ τῆς μεγάλης ἔνοικοι πόλεως, εἴ τι μέτεστιν ὑμῖν προνοίας μὴ ἐμὲ φεύσασθαι τὴν εὔχην, συνάρασθε τῷ τολμήματι, ἵνα τῶν ἐγκωμίων πρώτον ἀρχόμενοι καὶ τούτῳ ἔχωμεν λέγειν ὅτι εὐθύς μὲν τοιούτοις ἀνδράσιν ἐντυχεῖν ἦν, ὑφ' ὧν τις, καὶ ἂν ἀμουσος ἢ τὸ πρίν, κατ' Εὐριπίδην, ἐμμελῆς τε καὶ δεξιός εὐθύς γίγνεται καὶ δύναται λέγειν καὶ περὶ τῶν μειζόνων ἢ καθ' αὐτὸν.

selection none 0 unused highlight unused

1 of 78 unused

The image displays a detailed dependency parse tree for a Greek sentence from Aelius Aristides. The tree is rooted at 'ROOT' and branches into numerous nodes representing individual words and their grammatical relationships. Key nodes include 'άλλι', 'ὁ', 'ἄνδρες', 'οἱ', 'τῆς', 'μεγάλης', 'ἔνοικοι', 'πόλεως', 'εἴ', 'τι', 'μέτεστιν', 'ὑμῖν', 'προνοίας', 'μὴ', 'ἐμὲ', 'φεύσασθαι', 'τὴν', 'εὔχην', 'συνάρασθε', 'τῷ', 'τολμήματι', 'ἵνα', 'τῶν', 'ἐγκωμίων', 'πρώτον', 'ἀρχόμενοι', 'καὶ', 'τούτῳ', 'ἔχωμεν', 'λέγειν', 'ὅτι', 'εὐθύς', 'μὲν', 'τοιούτοις', 'ἀνδράσιν', 'ἐντυχεῖν', 'ἦν', 'ὑφ', 'ὧν', 'τις', 'καὶ', 'ἀν', 'ἀμουσος', 'ἢ', 'τὸ', 'πρίν', 'κατ', 'Εὐριπίδην', 'ἐμμελῆς', 'τε', 'καὶ', 'δεξιός', 'εὐθύς', 'γίγνεται', 'καὶ', 'δύναται', 'λέγειν', 'καὶ', 'περὶ', 'τῶν', 'μειζόνων', 'ἢ', 'καθ', 'αὐτὸν'. The tree illustrates the intricate dependencies between words like 'εἴ τι' (if what), 'μέτεστιν' (is changed), and 'δεξιός' (right-handed).

# COMPLEXITY

- Aldt1.5
  - Word sequences
- Native xml structure
  - Xml is a tree structure
- Find the Children Of...
  - $O(n)$  linear complexity
  - $O(1)$  constant complexity

# COMPLEXITY

```
<sentence id="1" document_id="" subdoc="" span="">
<word id="1" form="οὐδὲν" lemma="οὐδείς" postag="p-s---na-" relation="OBJ" head="4"/>
<word id="2" form="ἄν" lemma="ἄντι" postag="d-----" relation="ADV" head="4"/>
<word id="3" form="τις" lemma="τις" postag="p-s---mn-" relation="SBJ" head="4"/>
<word id="4" form="εἴποι" lemma="εἶπον" postag="v3saoa---" relation="PRED" head="0"/>
<word id="5" form="τῆς" lemma="ό" postag="l-s---fq-" relation="ATR" head="6"/>
<word id="6" form="ἐπινοίας" lemma="ἐπίνοια" postag="n-s---fq-" relation="OBJ" head="7"/>
<word id="7" form="νεαρώτερον" lemma="νεαρός" postag="a-s---nac" relation="ATR" head="1"/>
<word id="8" form"." lemma="punc1" postag="u-----" relation="AuxK" head="0"/>
</sentence>
```

The new structure that I introduced is shown below.

```
<sentence id="1" document_id="" subdoc="" span="">
<word id="4" form="εἴποι" lemma="εἶπον" postag="v3saoa---" relation="PRED">
    <word id="1" form="οὐδὲν" lemma="οὐδείς" postag="p-s---na-" relation="OBJ">
        <word id="7" form="νεαρώτερον" lemma="νεαρός" postag="a-s---nac" relation="ATR">
            <word id="6" form="ἐπινοίας" lemma="ἐπίνοια" postag="n-s---fq-" relation="OBJ">
                <word id="5" form="τῆς" lemma="ό" postag="l-s---fq-" relation="ATR"/>
            </word>
        </word>
    </word>
    <word id="2" form="ἄν" lemma="ἄντι" postag="d-----" relation="ADV"/>
    <word id="3" form="τις" lemma="τις" postag="p-s---mn-" relation="SBJ"/>
</word>
```

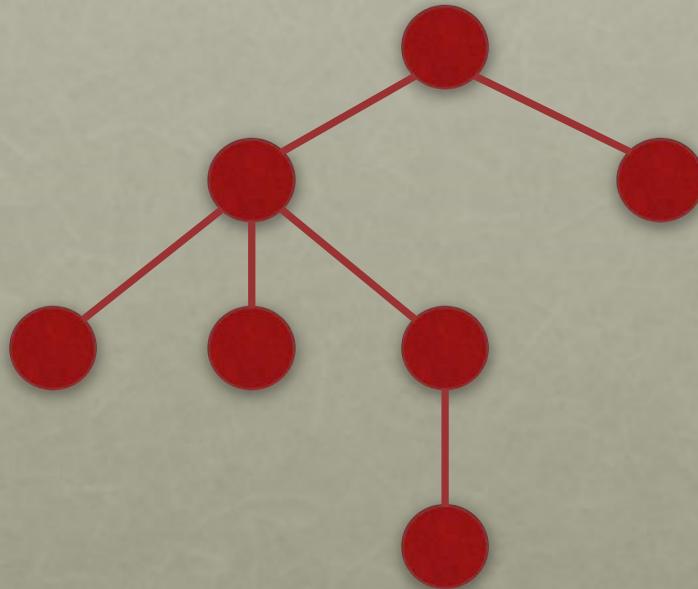
# SENTENCE AS A TREE

- $S = \{n_1, n_2, n_3, \dots, n_k\}$ 
  - where  $n_i$  is a tree node ( $n_i \in T$ )
  - and  $T$  denotes the space of tree nodes
- *Each node is a tree on its own*
- *Each node has properties such as:*
  - *children of:*  $T \rightarrow \{T^x\}$
  - *isATR:*  $T \rightarrow \{0,1\}$

$$fS = w_1 \mu n_1 + w_2 \mu n_2 + \dots + w_k \mu n_k = \sum_{i=1}^k w_i \mu n_i$$

# NODE-BASED METRICS

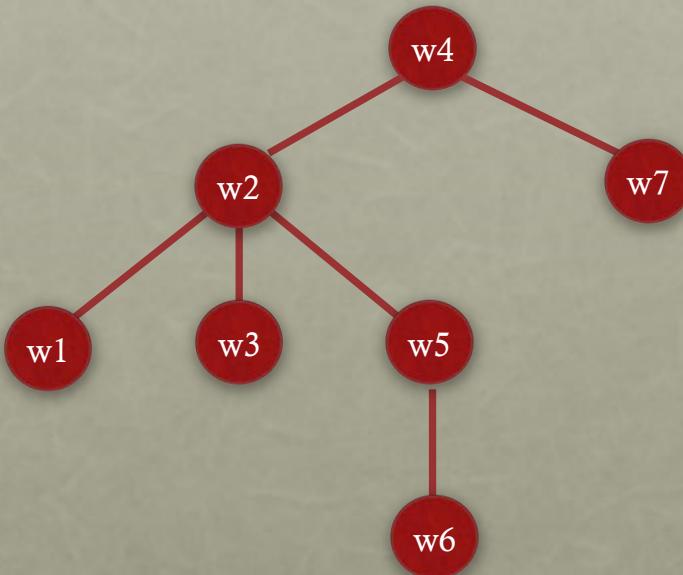
- Let's consider this example syntactically annotated sentence:



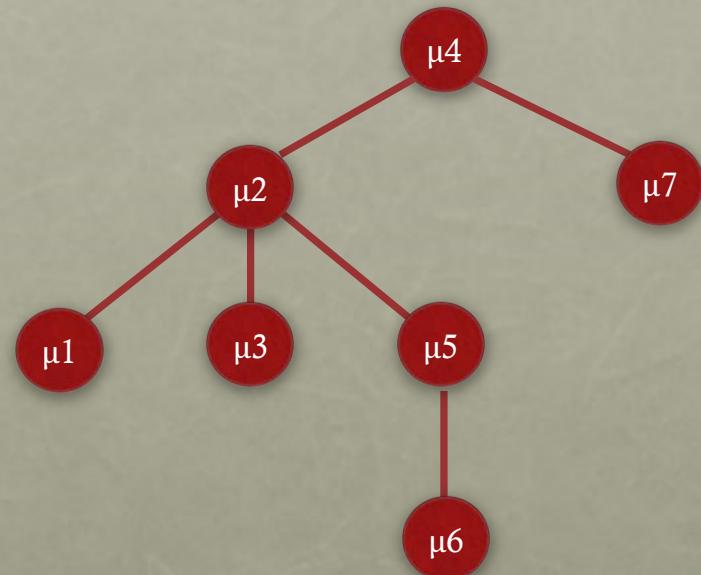
# NODE-BASED METRICS

- To calculate a sentence metric we need: w and  $\mu$

Weights



Metric Values

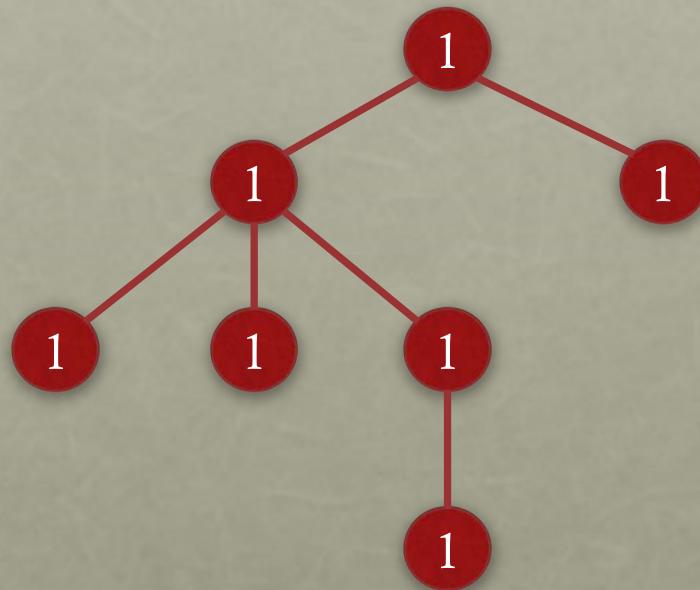


- Result:  $w_1\mu_1 + w_2\mu_2 + w_3\mu_3 + w_4\mu_4 + w_5\mu_5 + w_6\mu_6 + w_7\mu_7$

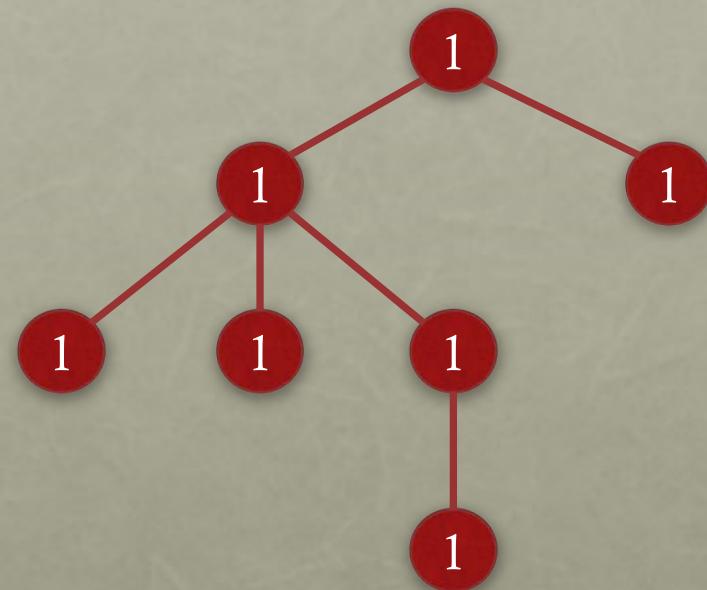
# NODE-BASED METRICS

- Metric example: *Number of words*

# Weights

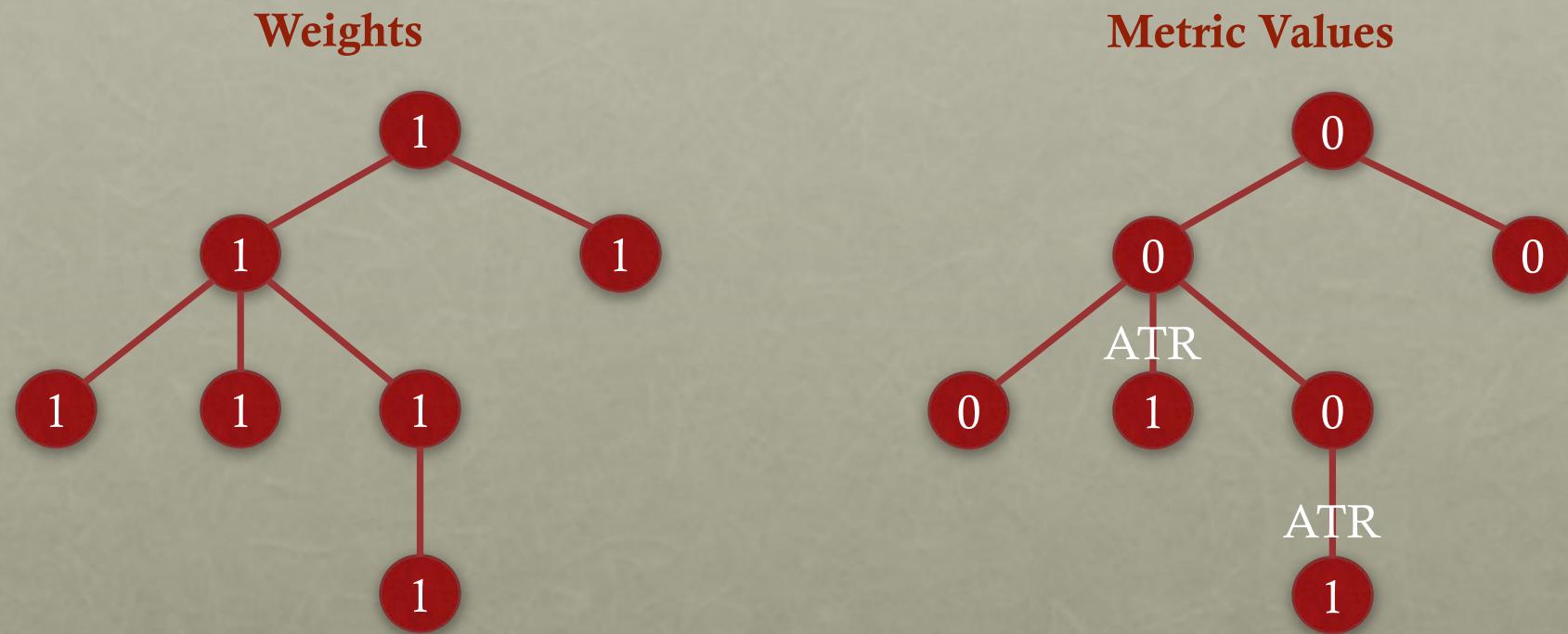


# Metric Values



# NODE-BASED METRICS

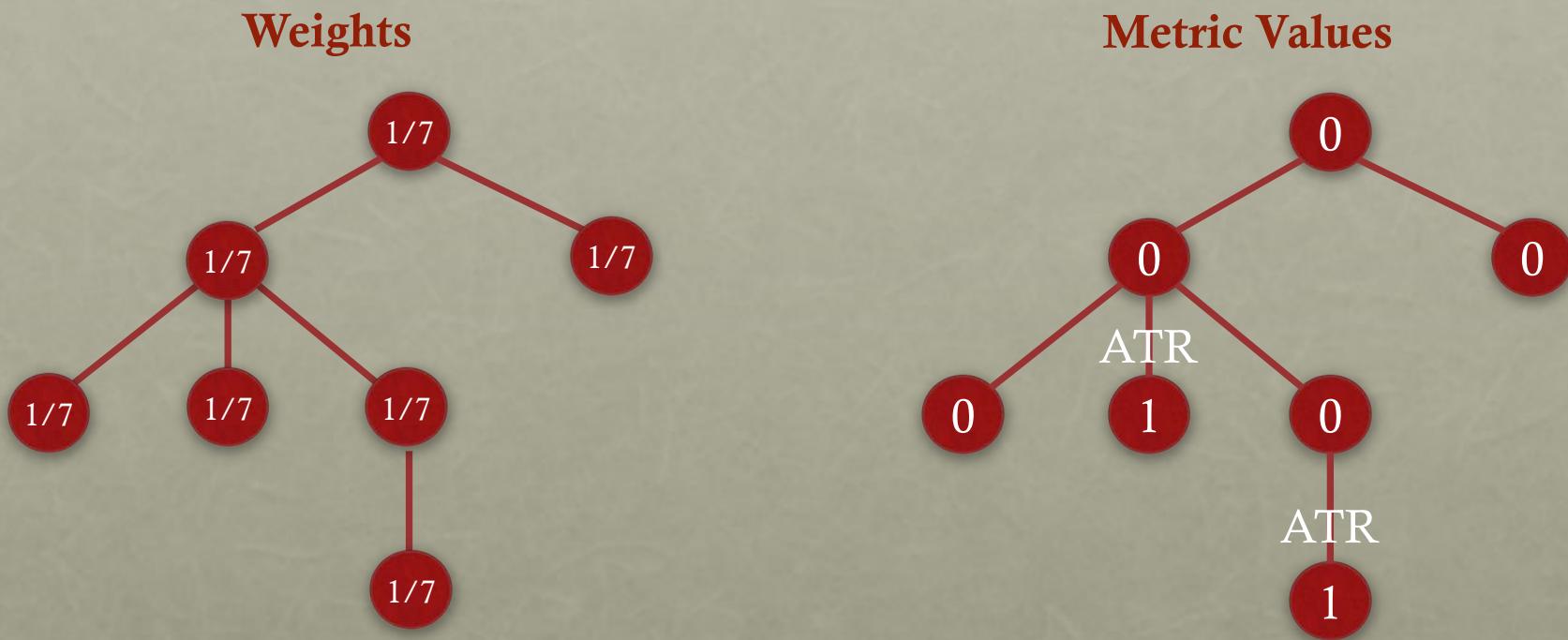
- Metric example: *Number of attributives*



- Result:  $1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 0 + 1 \times 0 = 2$

# NODE-BASED METRICS

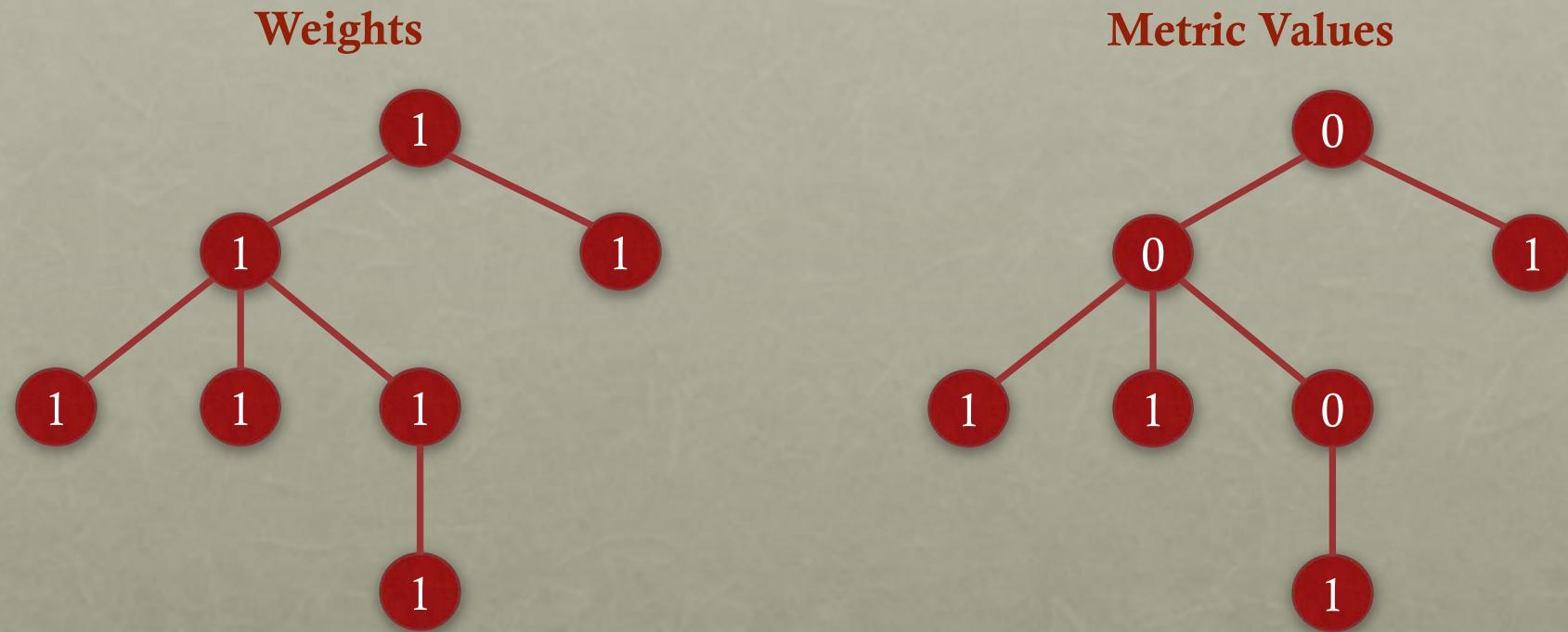
- Metric example: *Percentage of attributives*



- Result:  $1/7 \times 0 + 1/7 \times 0 + 1/7 \times 1 + 1/7 \times 0 + 1/7 \times 1 + 1/7 \times 0 + 1/7 \times 0 = 2/7$

# NODE-BASED METRICS

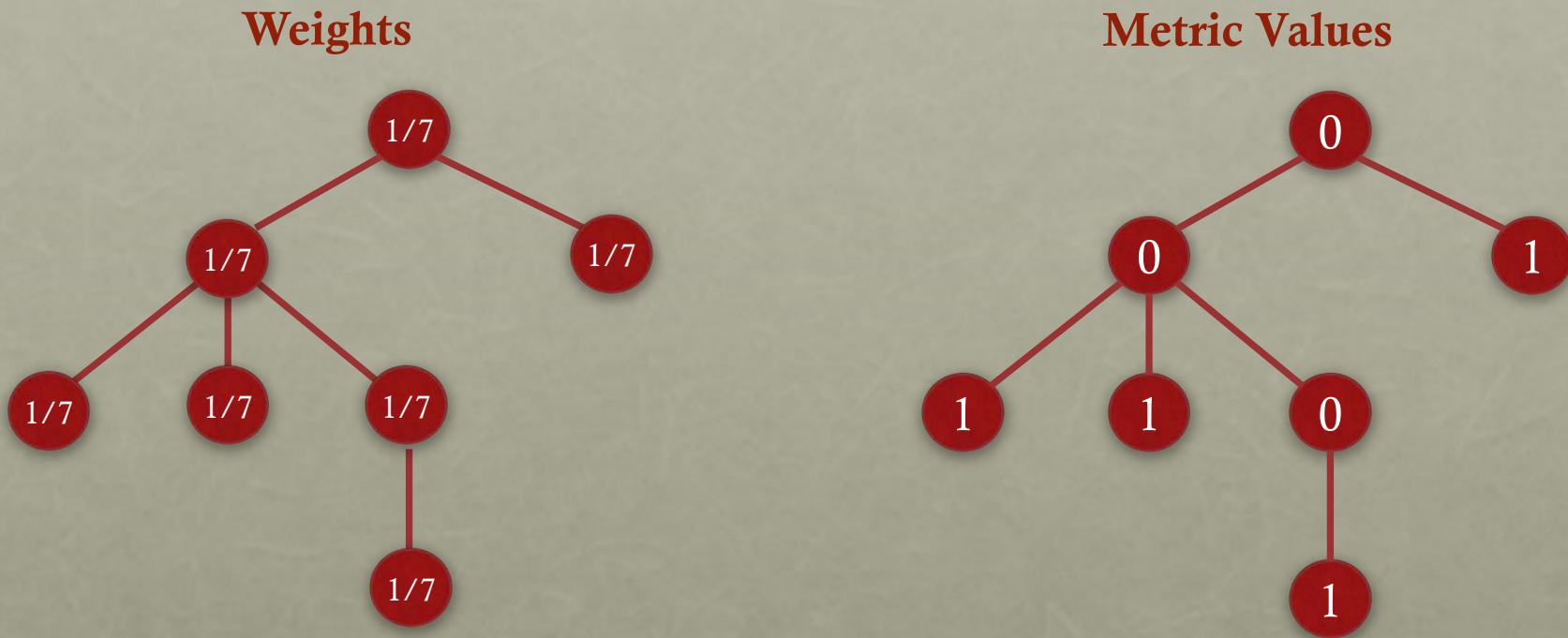
- Metric example: *Number of leaves*



- Result:  $1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 = 4$

# NODE-BASED METRICS

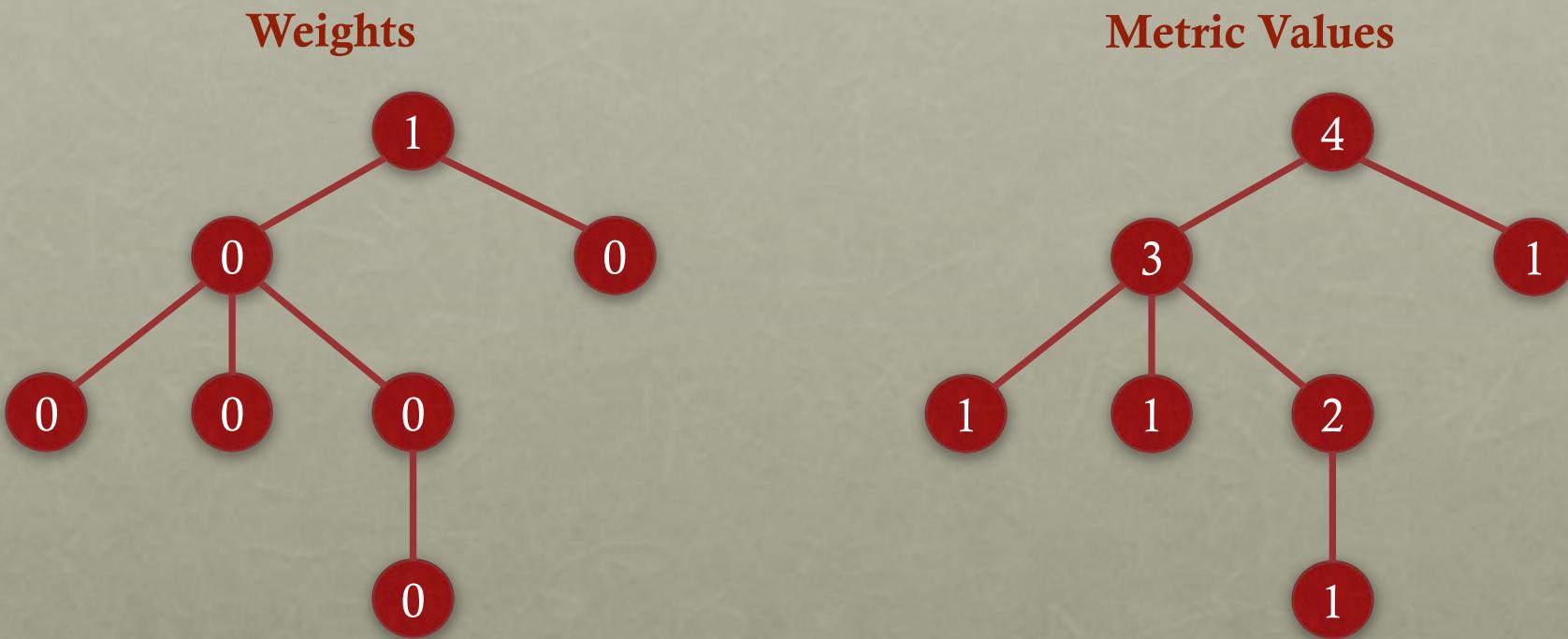
- Metric example: *Percentage of leaves*



- Result:  $1/7 \times 0 + 1/7 \times 1 + 1/7 \times 1 + 1/7 \times 0 + 1/7 \times 1 + 1/7 \times 1 + 1/7 \times 0 = 4/7$

# RECURSIVE NODE-BASED METRICS

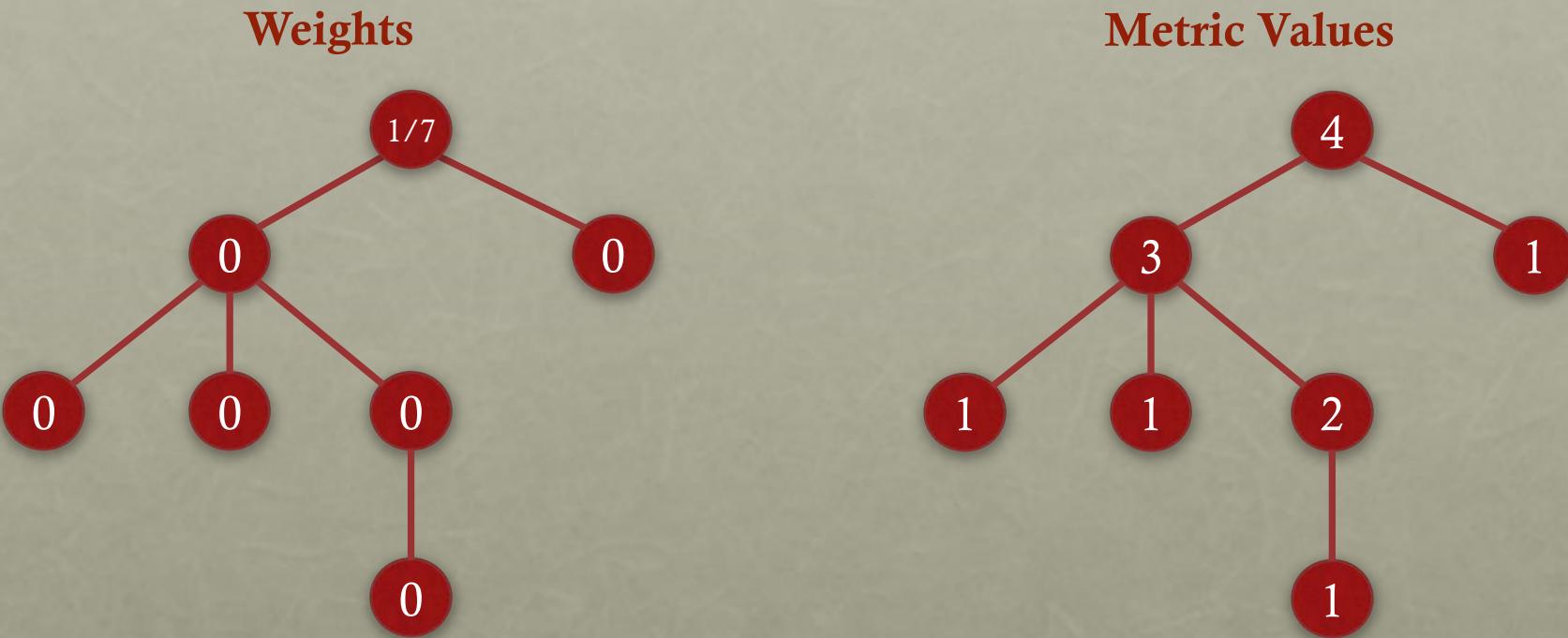
- Metric example: *Height of the tree*



- Result:  $1 \times 4 + 0 \times 3 + 0 \times 1 + 0 \times 1 + 0 \times 1 + 0 \times 2 + 0 \times 1 = 4$

# RECURSIVE NODE-BASED METRICS

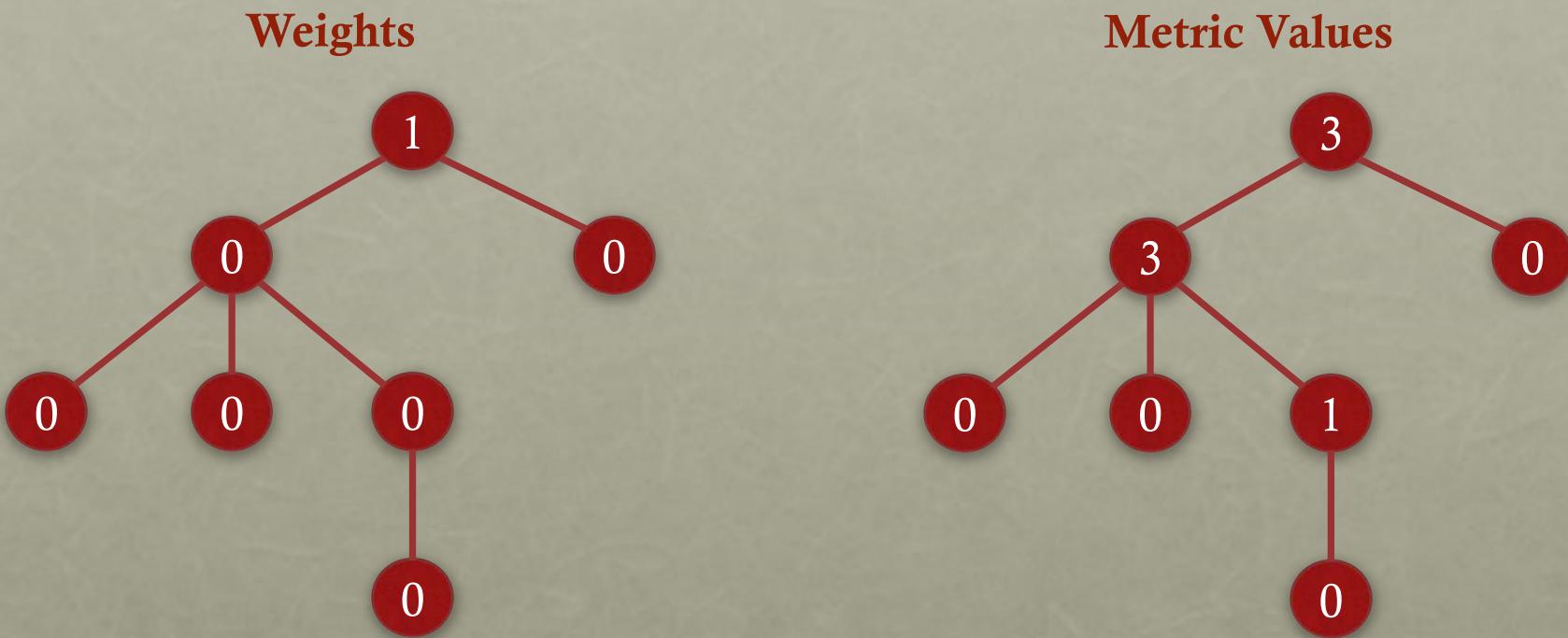
- Metric example: *Height of the tree as percentage of the words*



- Result:  $1/7 \times 4 + 0 \times 3 + 0 \times 1 + 0 \times 1 + 0 \times 1 + 0 \times 2 + 0 \times 1 = 4/7$

# RECURSIVE NODE-BASED METRICS

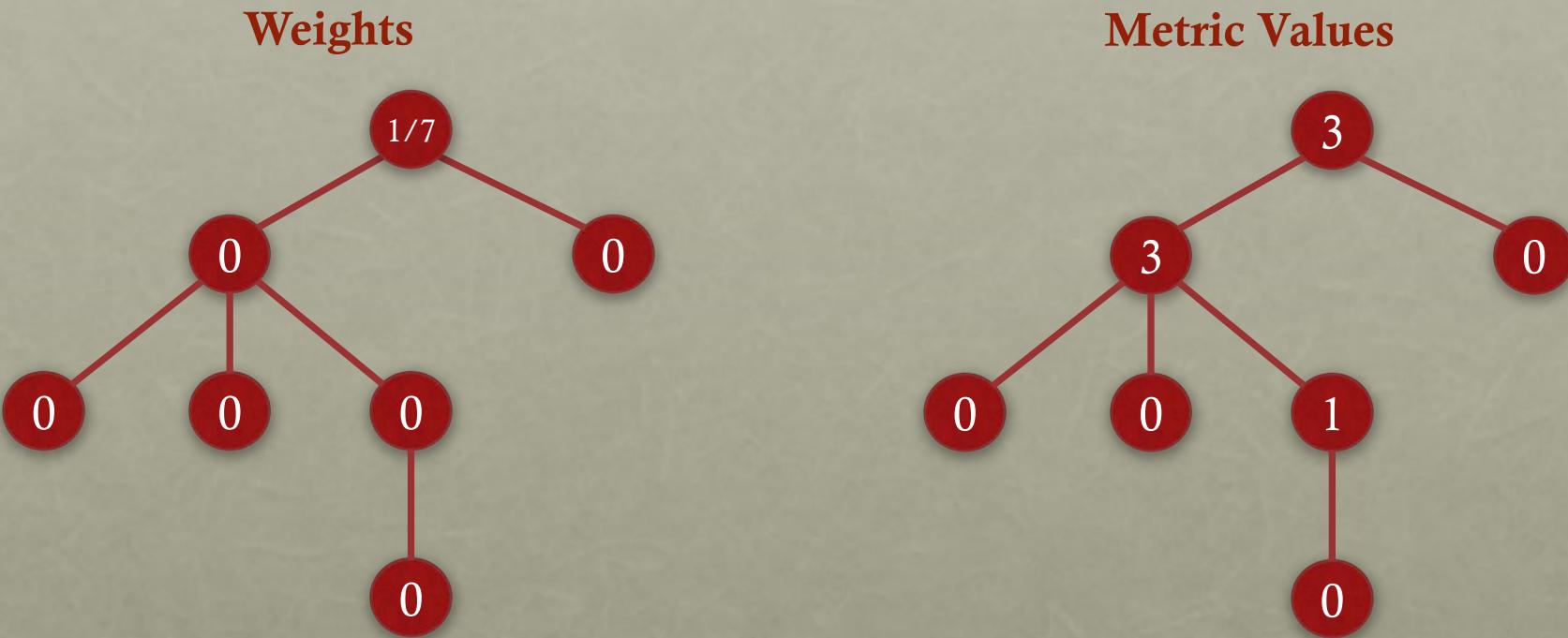
- Metric example: *Size of largest family*



- Result:  $1 \times 3 + 0 \times 3 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 1 + 0 \times 0 = 3$

# RECURSIVE NODE-BASED METRICS

- Metric example: *Size of largest family as percentage of the words*



- Result:  $1/7 \times 3 + 0 \times 3 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 1 + 0 \times 0 = 3/7$

# JAVASCRIPT

```
m=new NodeMetric('Percentage of Verb Attributives');

m.weight=function(n)
{
    return 1/n.getRoot().getNumOfWords();
};

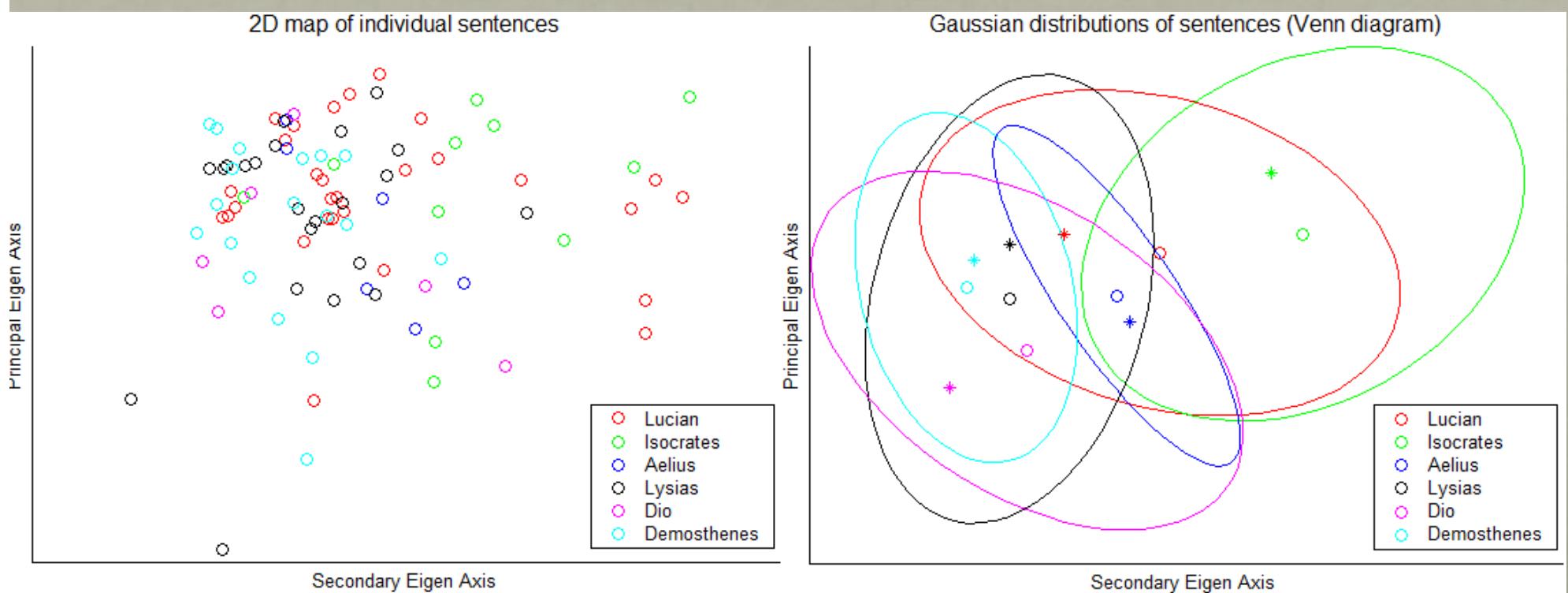
m.metric=function(n)
{
    if(n.getRelation()=='ATR' && n.getPosTag()[0]=='v')
        return 1;
    else return 0;
};
```

# METRICS

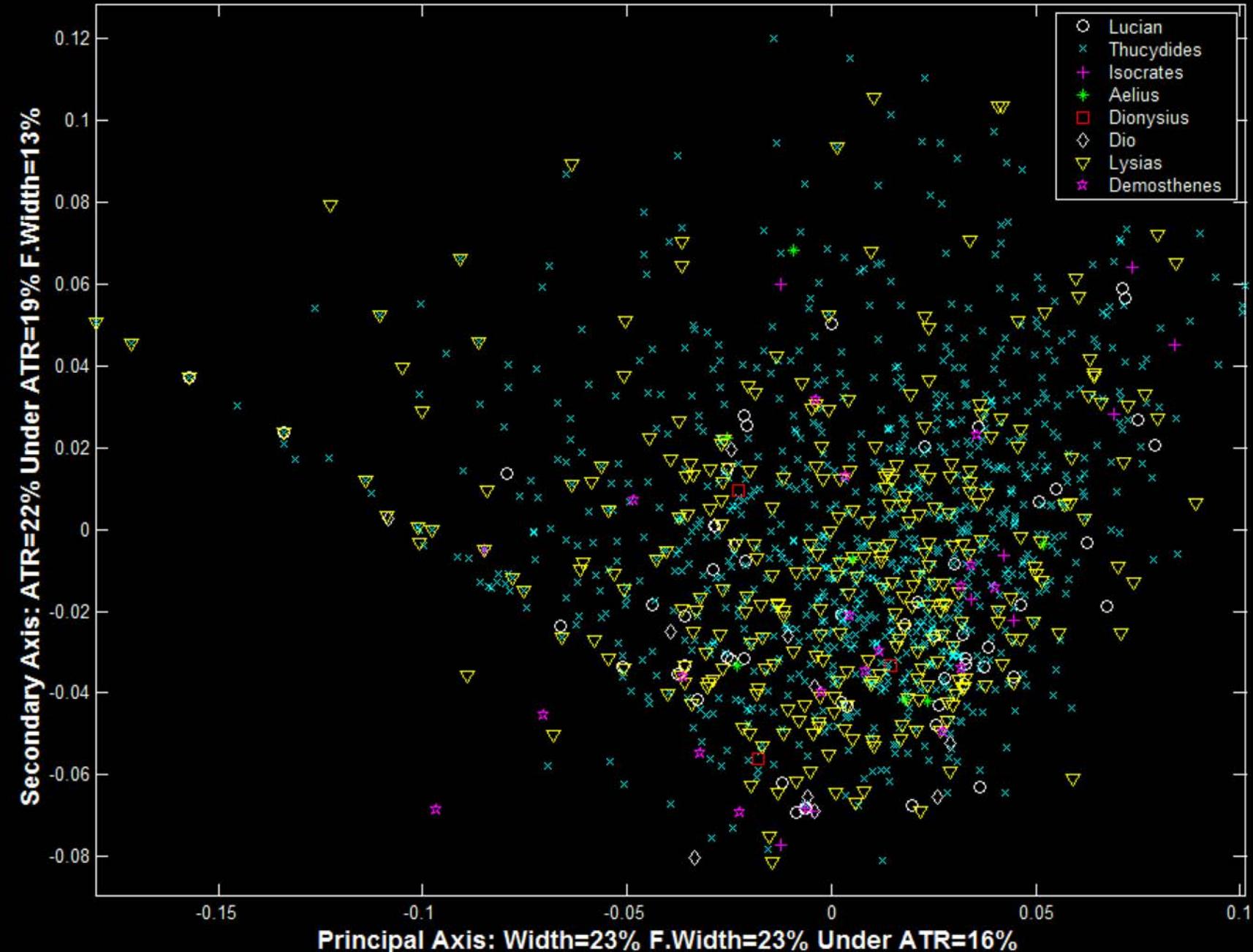
- Height of the tree as percentage of the nodes
- Width of the tree as percentage of the nodes
- Number of leaves as percentage of the nodes
- Number of attributives as percentage of the nodes
- Largest family size as percentage of the nodes
- Verb Attributives as percentage of the nodes
- Percentage of δέ coordinates
- Number of nodes under ATR as percentage of the nodes
- Adjective Attributives as percentage of the nodes

# DATA

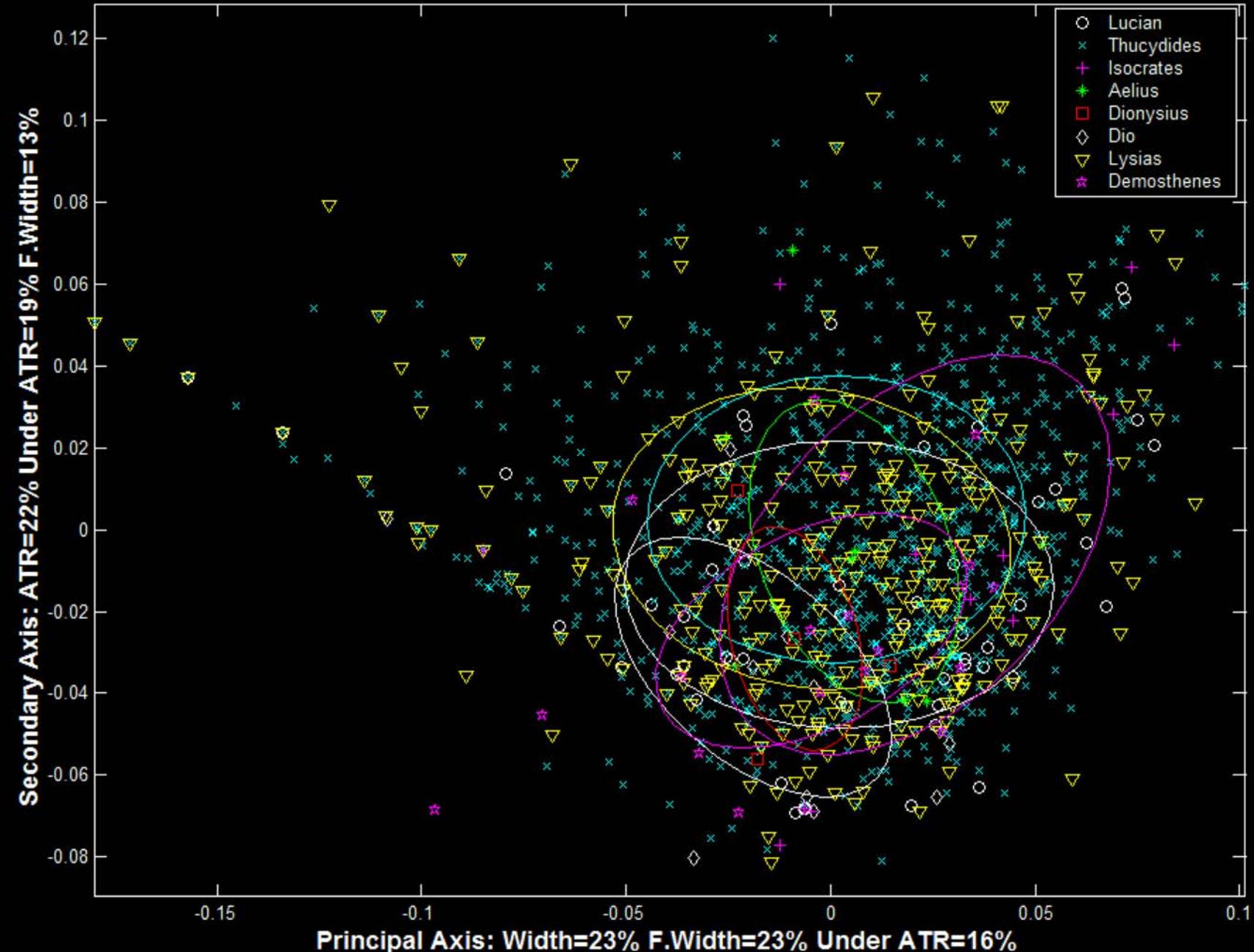
0.4000	0.3700	0.2000	0.1300	0.2700	0	0.0300	0	0.4000
0.7500	0.2500	0.7500	0.7500	0.2500	0	0	0	0.2500
0.5000	0.3300	0.2800	0.2200	0.0600	0	0	0.1100	0.0600
0.4400	0.1600	0.2400	0.0800	0.1800	0	0.0800	0.0400	0.3500
0.4700	0.2700	0.3300	0.2700	0.0700	0	0	0	0.0700
0.3500	0.3200	0.1800	0.0900	0.1800	0.0300	0.0300	0.0900	0.3400
0.5600	0.4400	0.3300	0.3300	0.1100	0	0	0	0.1100
0.4700	0.2400	0.3100	0.1100	0.3800	0.0500	0.0500	0.1300	0.6500
0.4700	0.1600	0.2400	0.1600	0.2100	0	0	0.0300	0.2100
0.5000	0.1500	0.2500	0.1000	0.1900	0.0600	0.0400	0.0200	0.3300
0.5000	0.4100	0.2700	0.1400	0	0	0	0.0500	0
0.5800	0.2500	0.4200	0.3300	0.2500	0	0	0	0.2500
0.4700	0.1500	0.2600	0.1300	0.2100	0	0.0200	0.0900	0.2800
0.4600	0.3800	0.3100	0.3100	0.3800	0	0	0	0.6200
0.5000	0.2500	0.2500	0.1400	0.2500	0.0400	0	0.0400	0.3900
0.4400	0.3300	0.4400	0.3300	0.2200	0	0.1100	0	0.3300
0.5000	0.2100	0.3600	0.2900	0.2900	0	0	0	0.4300
0.5800	0.2500	0.4200	0.4200	0.1700	0	0	0	0.1700
0.6000	0.4000	0.6000	0.6000	0	0	0	0.2000	0
0.4400	0.2200	0.2600	0.1500	0.1100	0	0.0400	0	0.1500
0.5800	0.4200	0.4200	0.4200	0	0	0	0.0800	0
0.5500	0.1900	0.2600	0.2600	0.2900	0.0300	0.0600	0.0300	0.6100
0.4800	0.2300	0.1900	0.1900	0.1900	0	0	0	0.1900
0.5300	0.1600	0.3400	0.1600	0.1600	0	0.0300	0.0500	0.1600
0.5000	0.1500	0.1800	0.0600	0.1300	0.0100	0	0.0200	0.2900
0.4800	0.2000	0.4400	0.2000	0.0400	0	0	0.0400	0.0400
0.5500	0.3600	0.4500	0.4500	0.2700	0.0900	0.0900	0	0.2700
0.5500	0.3600	0.3600	0.3600	0.1800	0	0.0900	0	0.1800
0.4300	0.3300	0.2400	0.1400	0.2900	0.1000	0	0	0.6700
0.4900	0.2300	0.1300	0.0900	0.1600	0	0.0100	0	0.1800
0.4200	0.2300	0.1600	0.0800	0.2500	0.0300	0.0200	0.0500	0.6100
0.4300	0.2000	0.2500	0.1200	0.2200	0	0.0200	0.0400	0.3100
0.3500	0.3000	0.2600	0.1300	0.2600	0.0500	0	0	0.3600
0.4400	0.3300	0.4400	0.4400	0.4400	0.2500	0	0	0.7500
0.5300	0.1800	0.2600	0.1100	0.2100	0.0300	0.0300	0	0.2400
0.5700	0.3600	0.3600	0.3600	0.1400	0	0.0700	0	0.1400
0.4400	0.2800	0.3600	0.1200	0.2400	0.0400	0	0.0400	0.3300
0.4700	0.1400	0.2300	0.0600	0.1100	0.0100	0.0500	0.0100	0.3500
0.4700	0.3000	0.1700	0.1700	0.1300	0.0300	0.0300	0	0.5000
0.5300	0.2100	0.4200	0.3200	0.1100	0	0	0.0500	0.1100
0.5500	0.2300	0.2000	0.1100	0.1400	0.0200	0.0200	0.0200	0.5900
0.3600	0.4500	0.2700	0.2700	0.3600	0.1000	0.1000	0	0.6000
0.5000	0.2500	0.2900	0.1400	0.0400	0	0	0.0700	0.0400



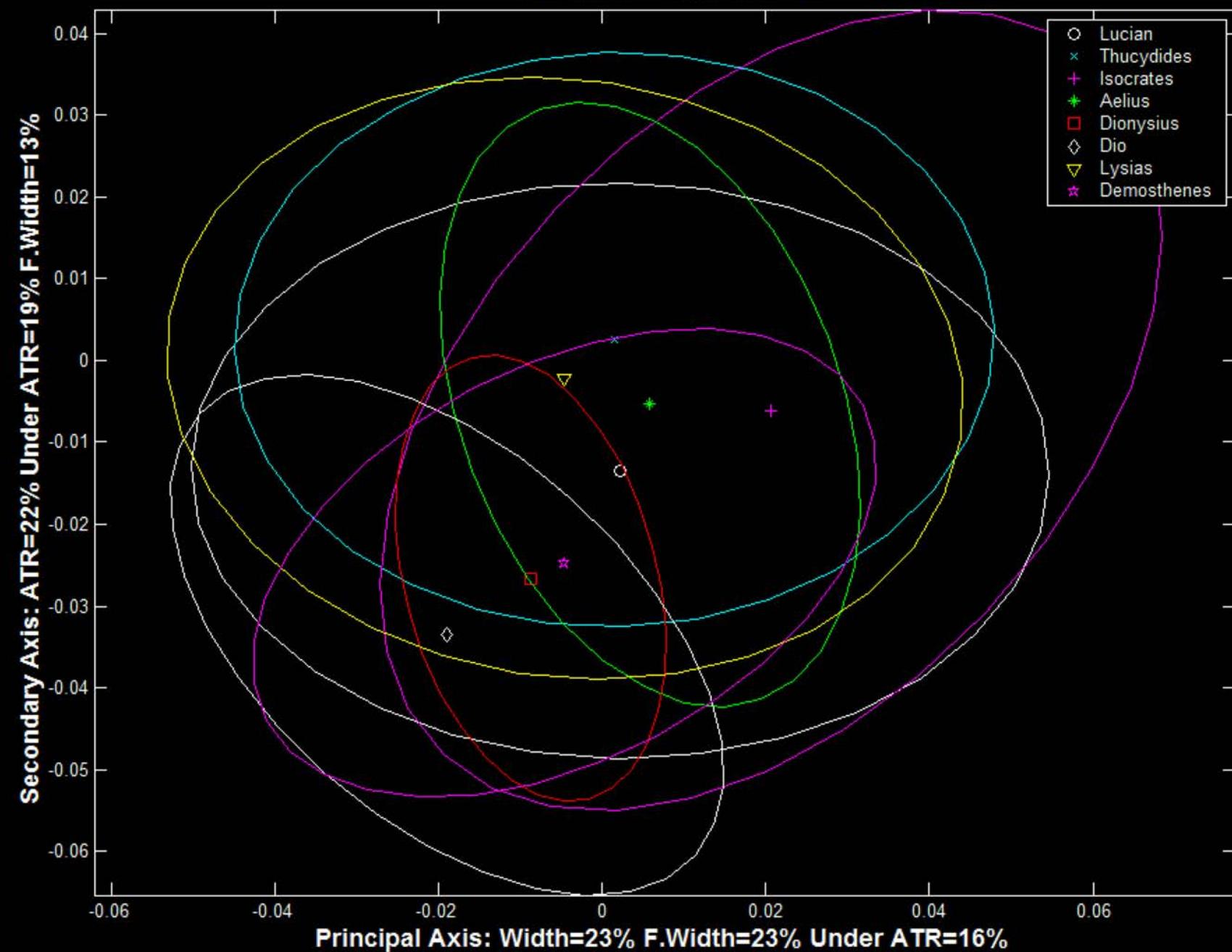
## Dominant eigenplane



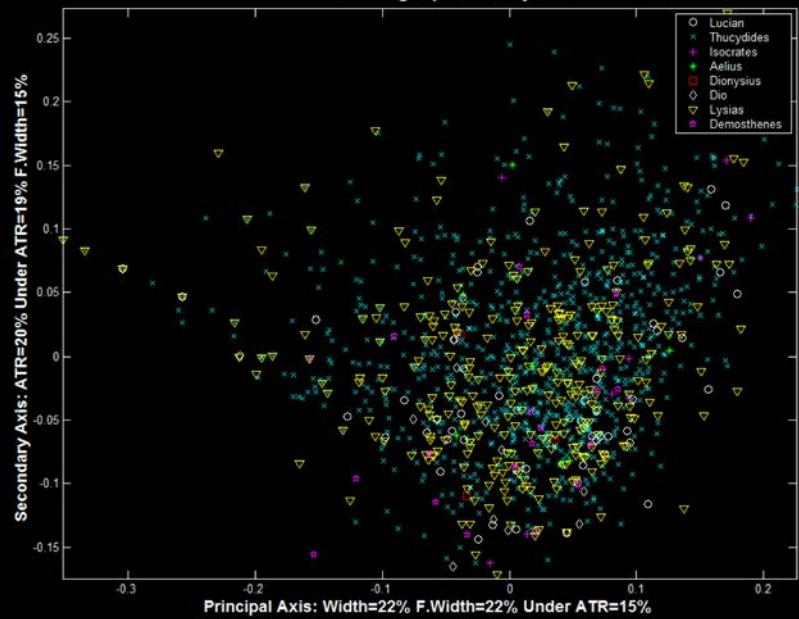
## Dominant eigenplane



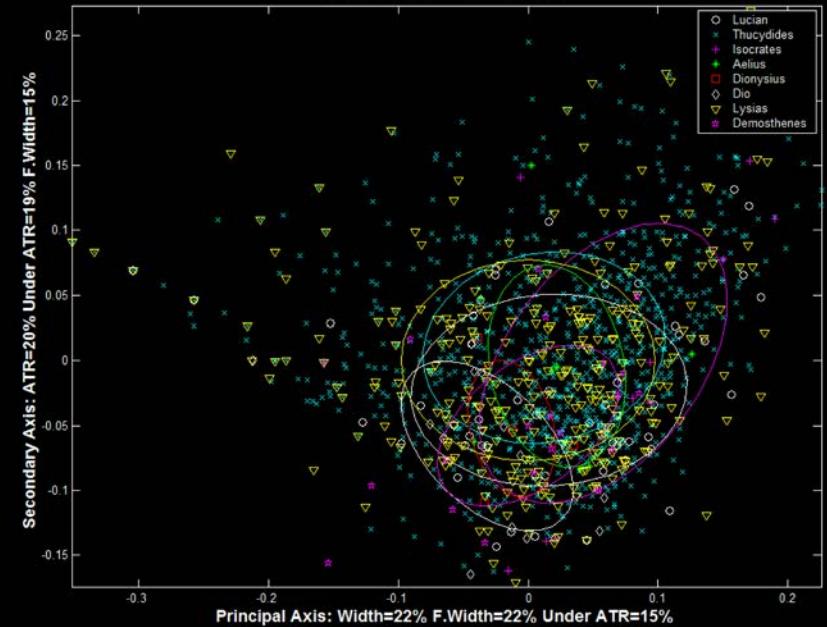
## Dominant eigenplane



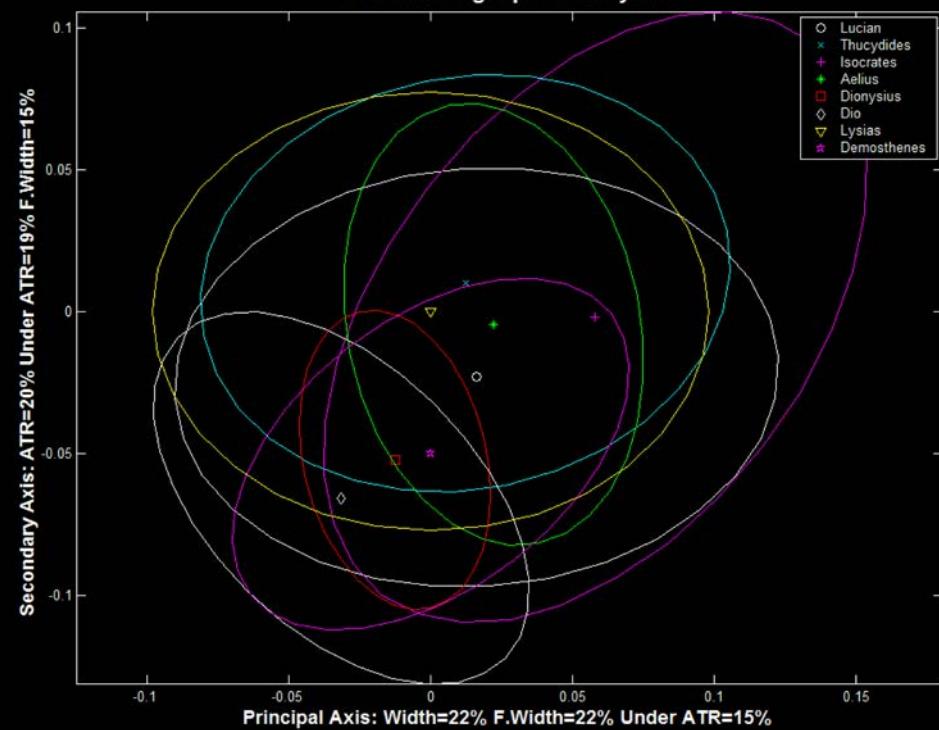
Dominant eigenplane of Lysias



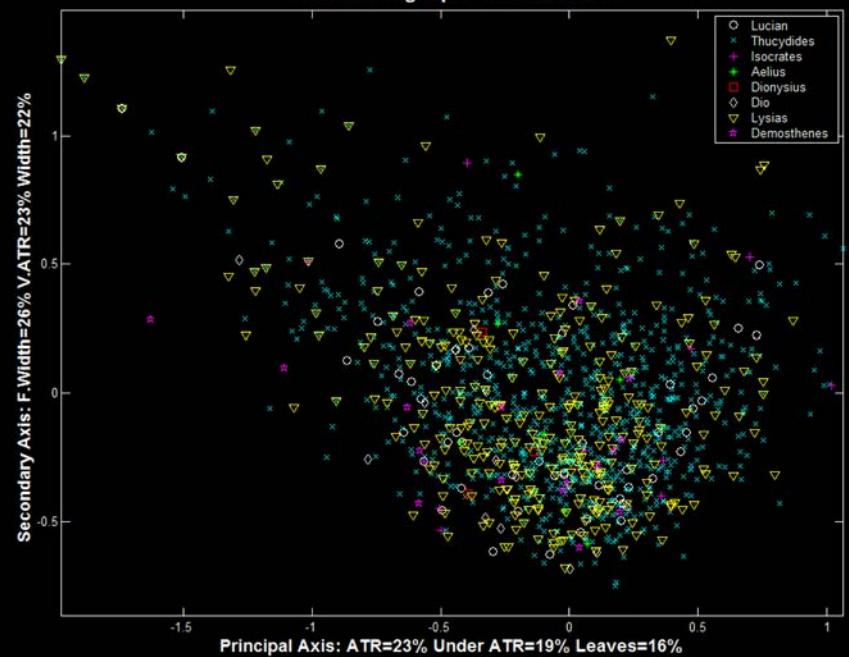
Dominant eigenplane of Lysias



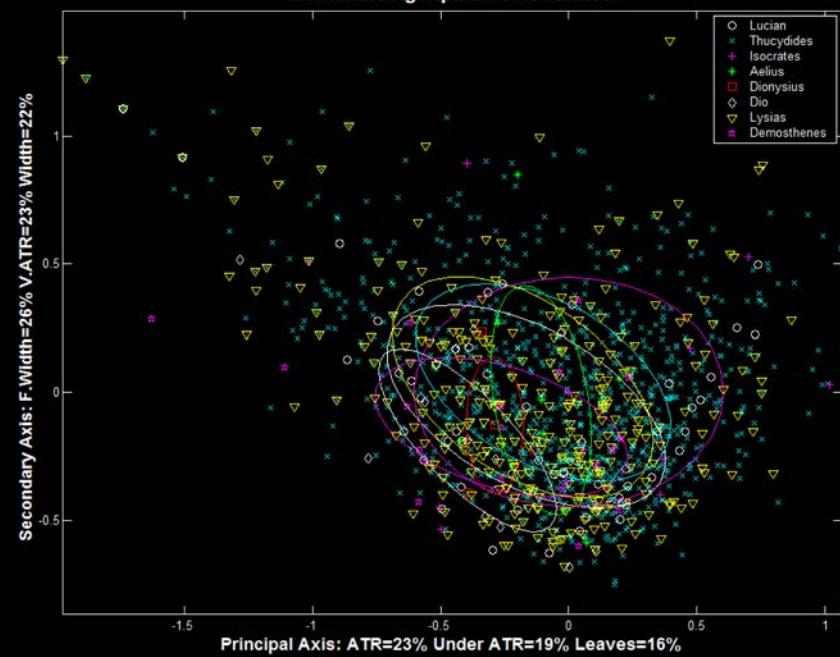
Dominant eigenplane of Lysias



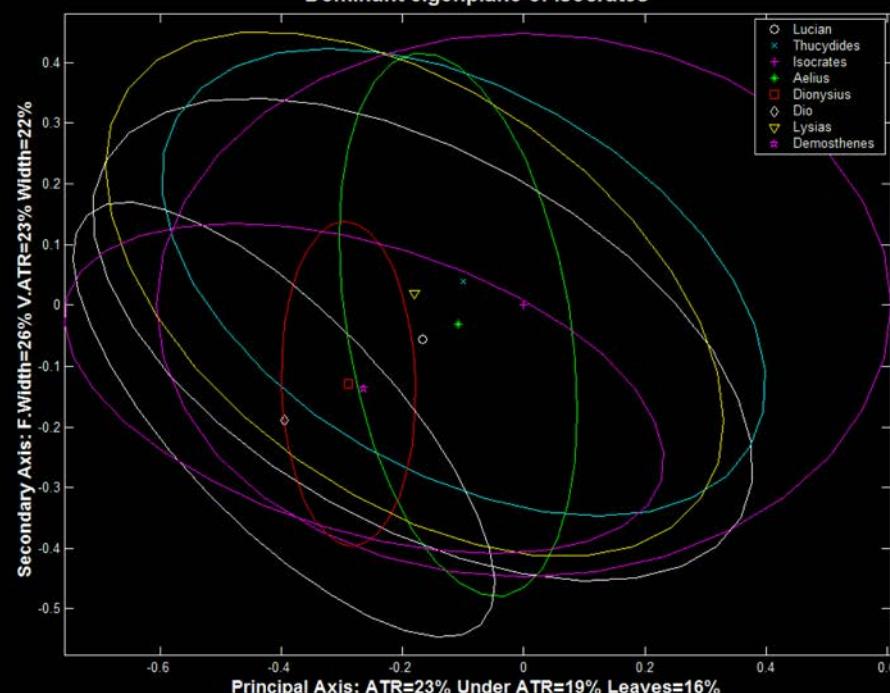
Dominant eigenplane of Isocrates

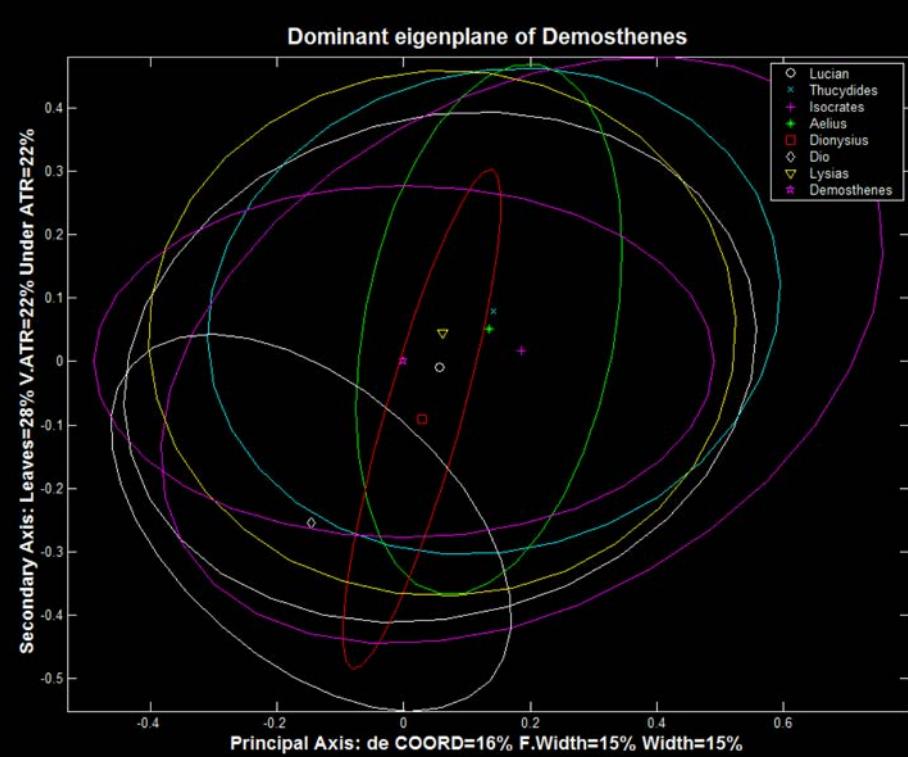
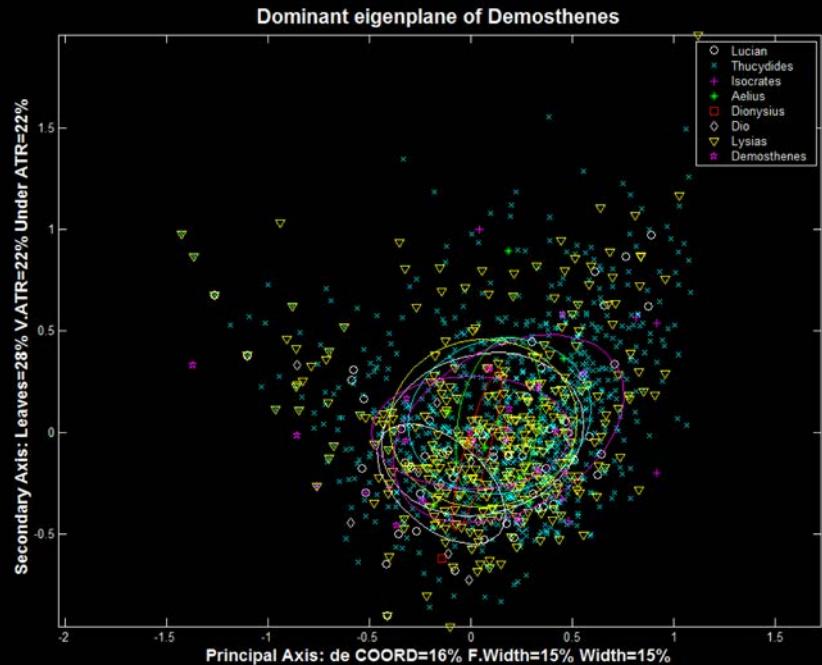
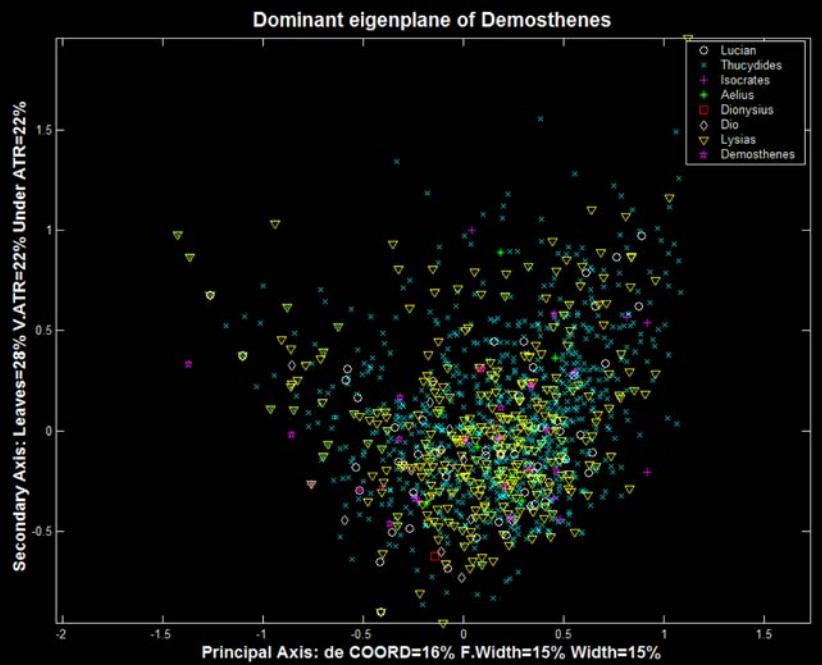


Dominant eigenplane of Isocrates

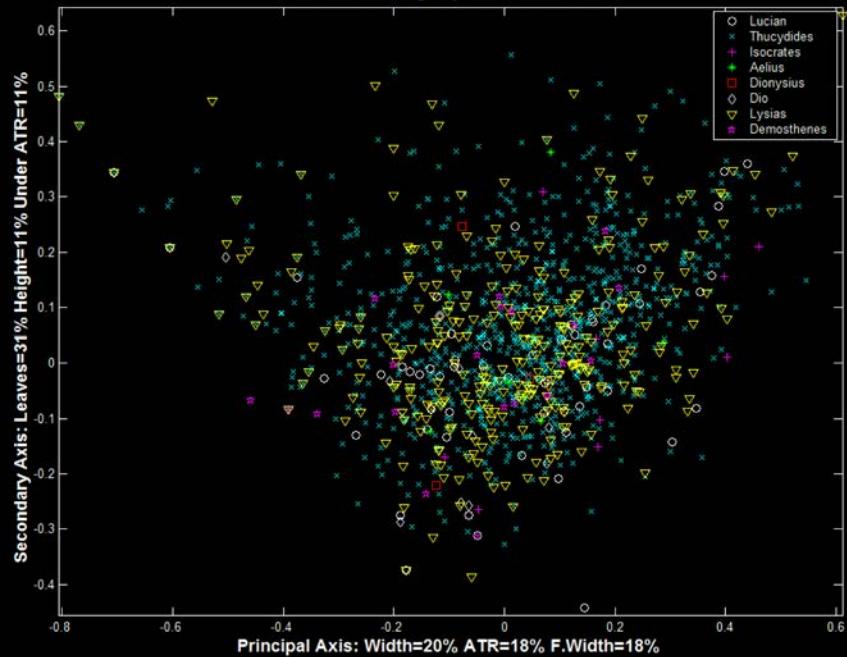


Dominant eigenplane of Isocrates

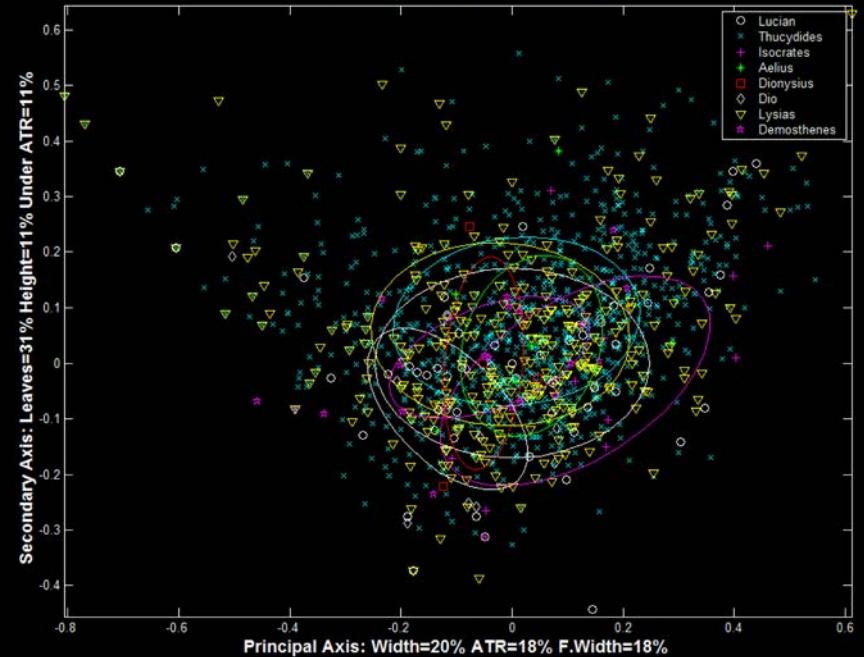




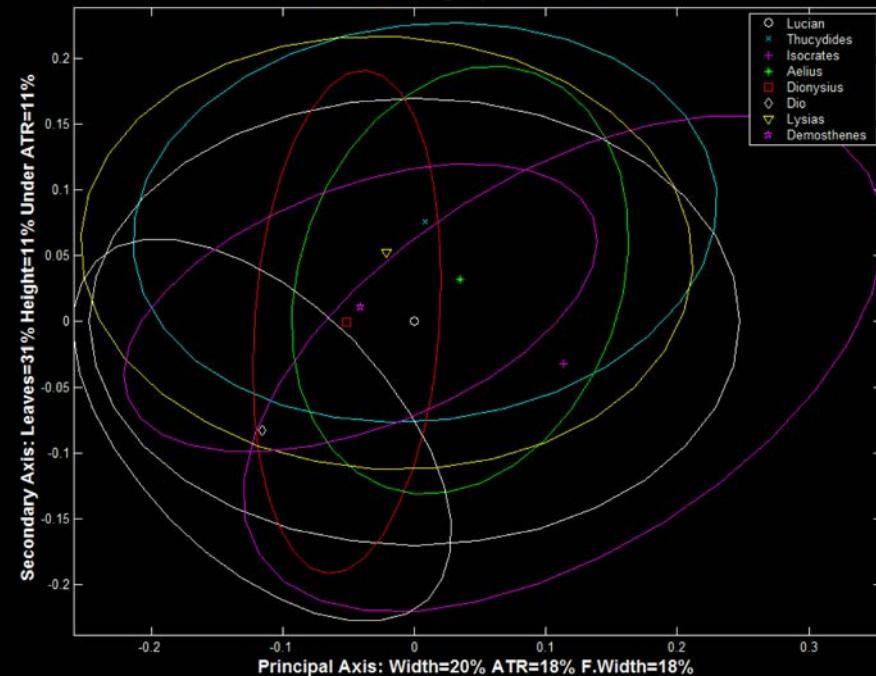
Dominant eigenplane of Lucian

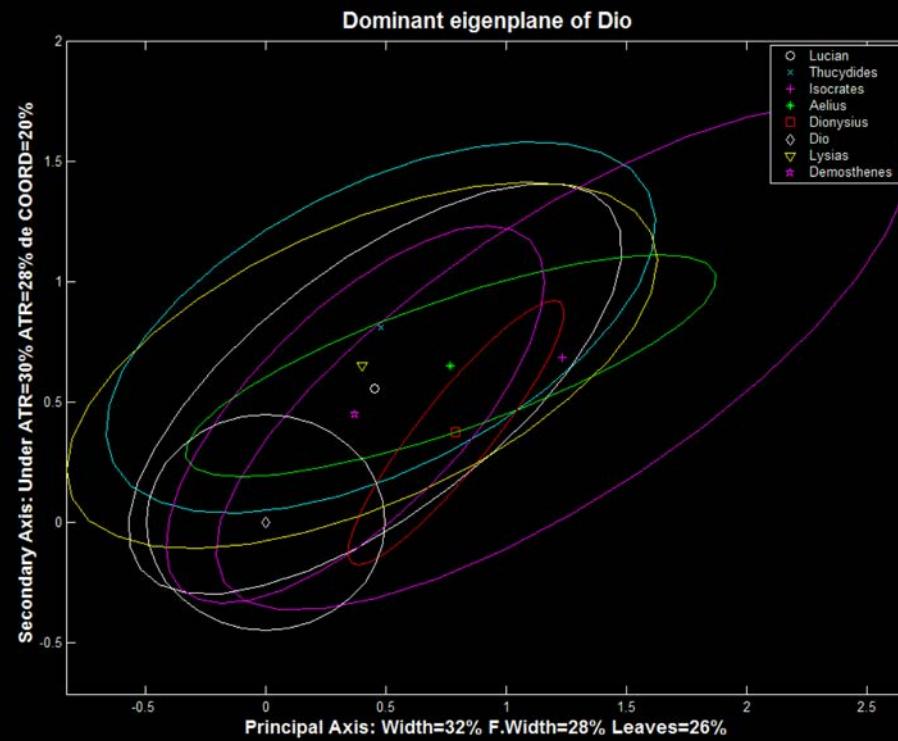
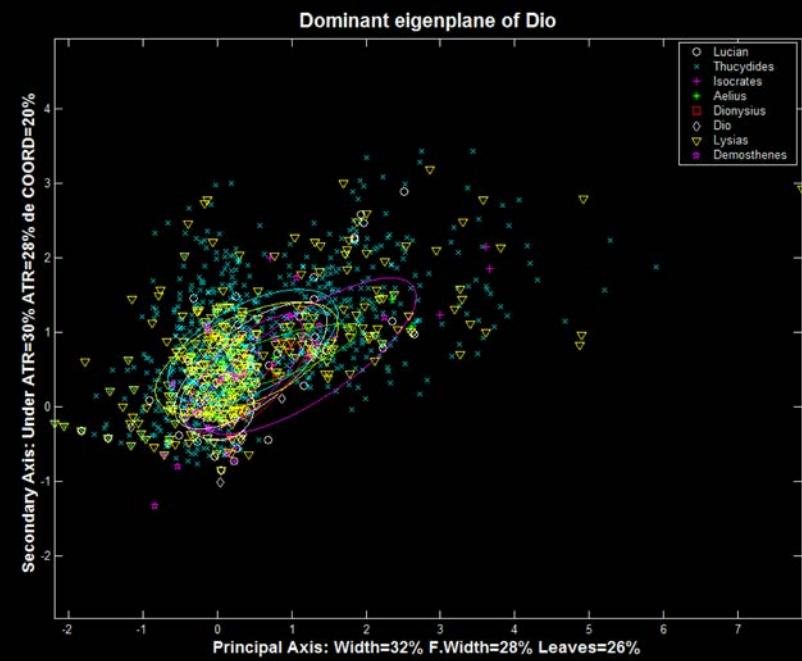
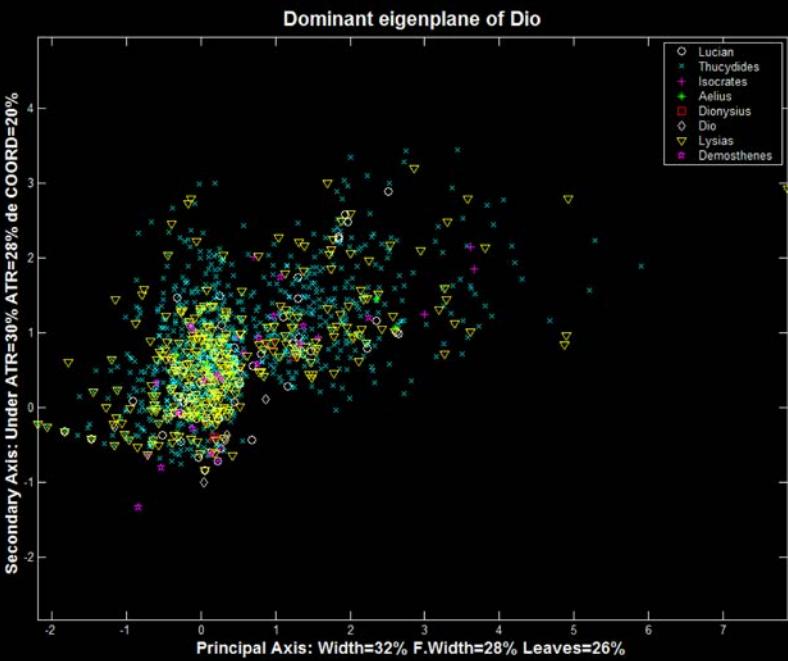


Dominant eigenplane of Lucian

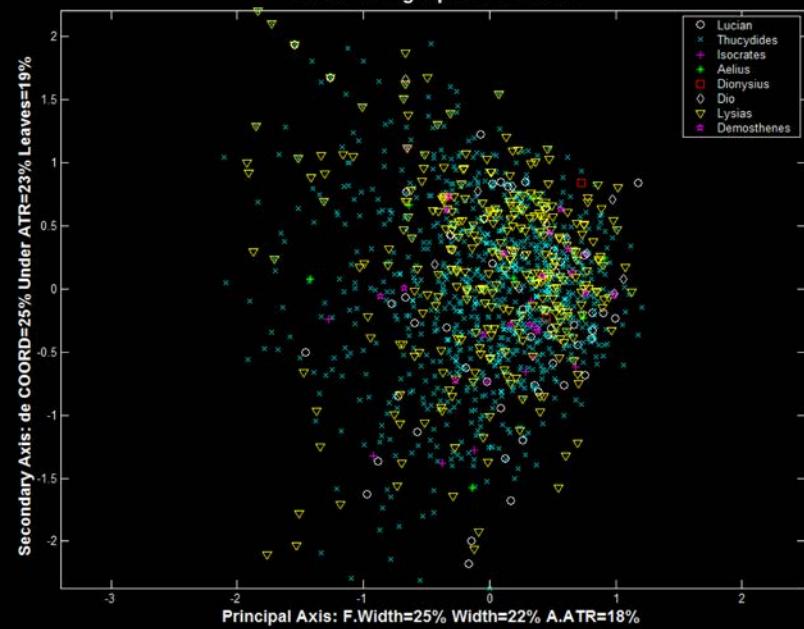


Dominant eigenplane of Lucian

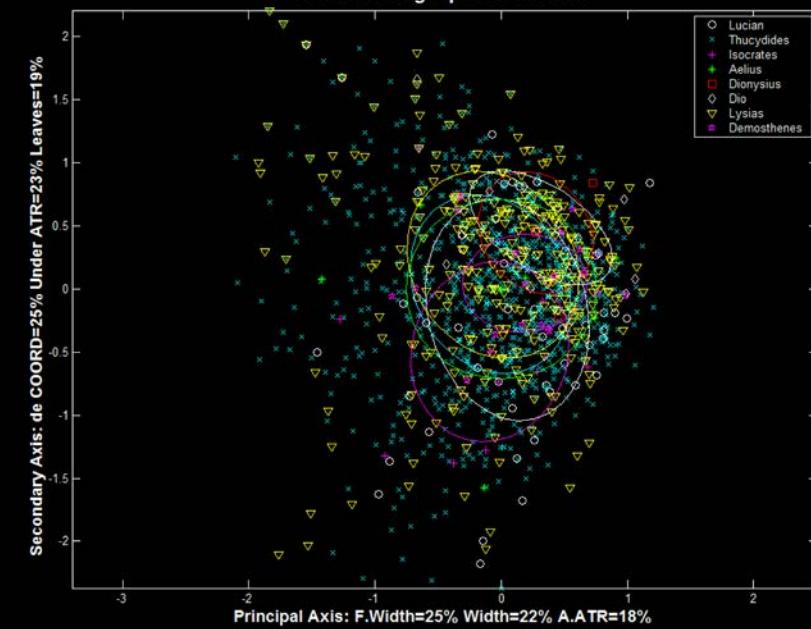




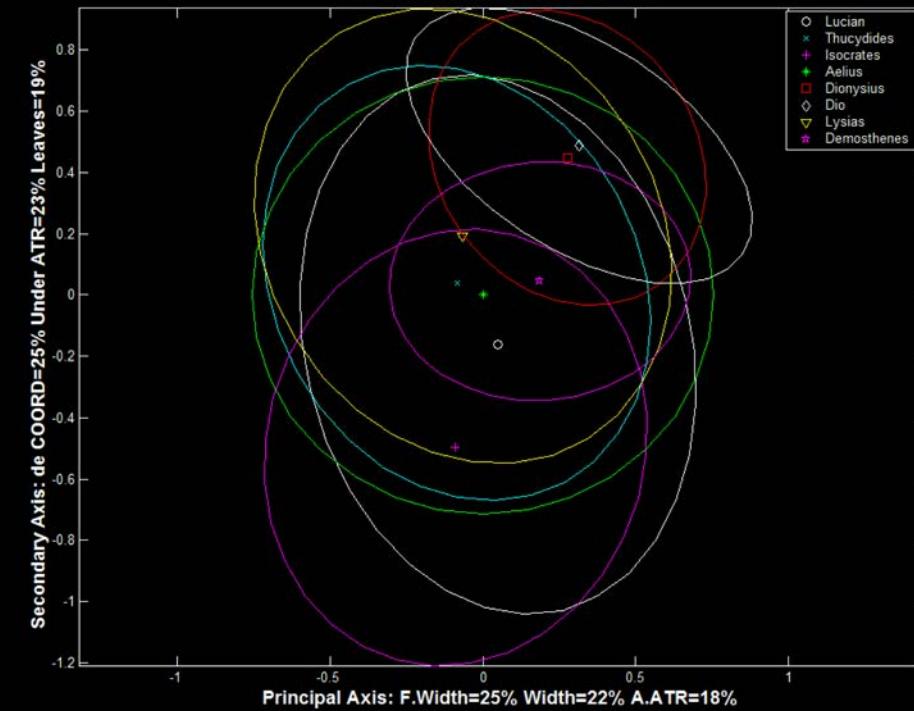
Dominant eigenplane of Aelius



Dominant eigenplane of Aelius



Dominant eigenplane of Aelius



# STYLOMETRIC ANALYSIS

- 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 Lutosławski in *Principes de stylométrie* (1890)

# CASE STUDIES

- 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
- Authorship of Ronald Reagan's Radio Addresses
- “Double Falsehood”

# STYLO

- Writer Invariant

Property of a text that is invariant of its authors

Word Lengths

Sentence Length

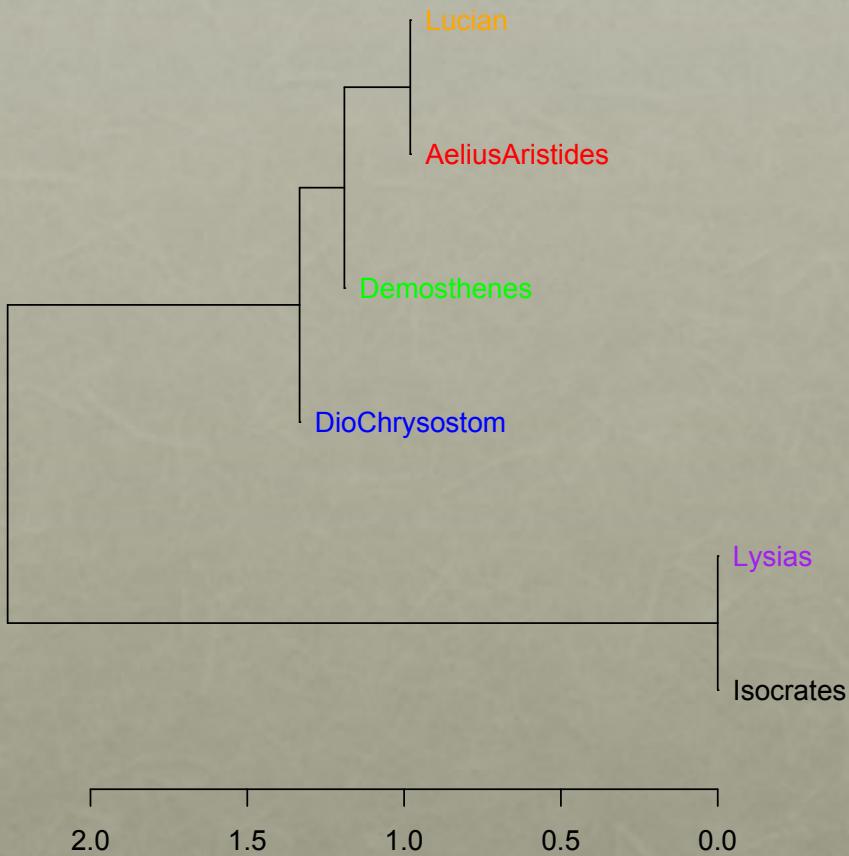
Average Word Length

Noun, Verb or adverb usage frequency

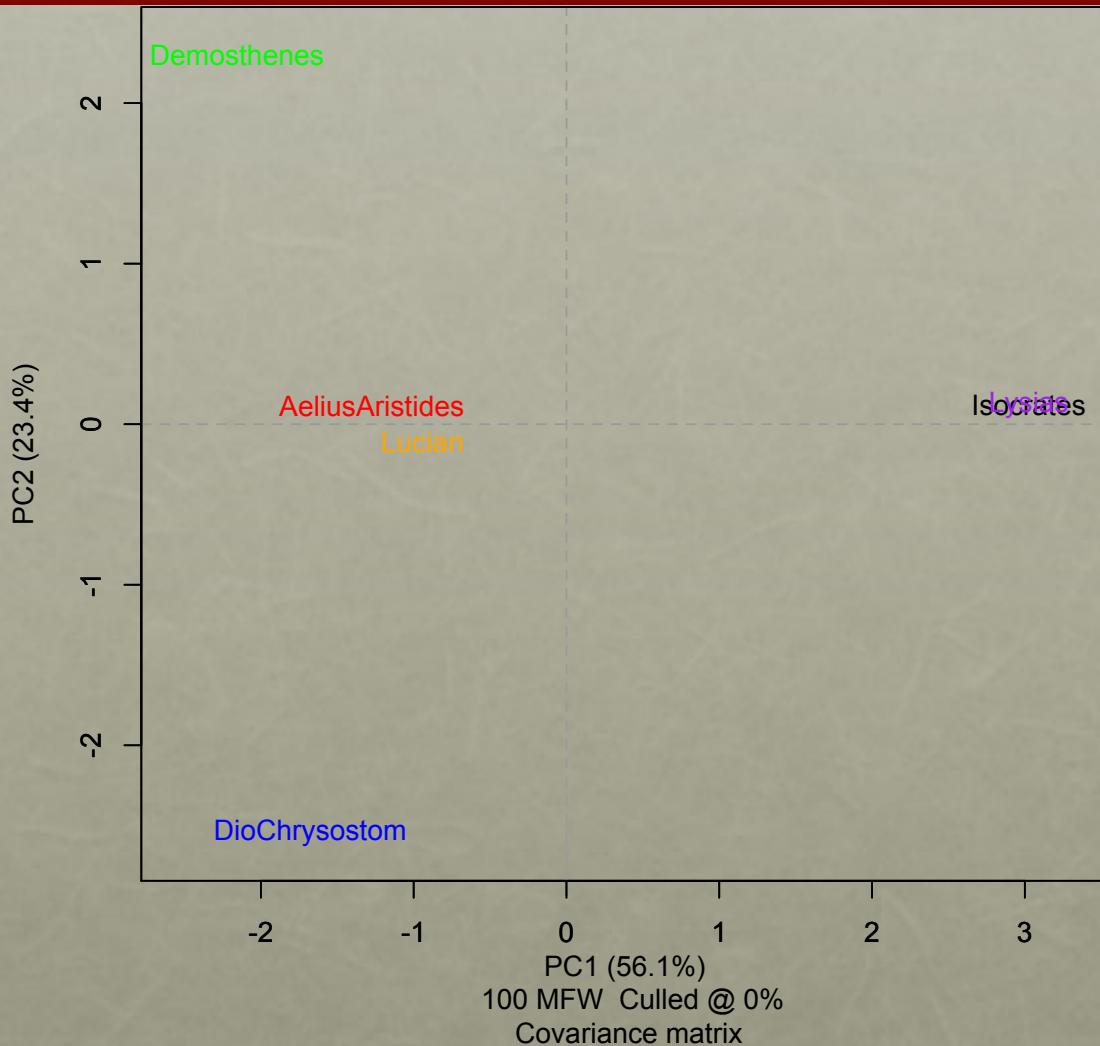
Vocabulary Richness

Frequency of Function Words

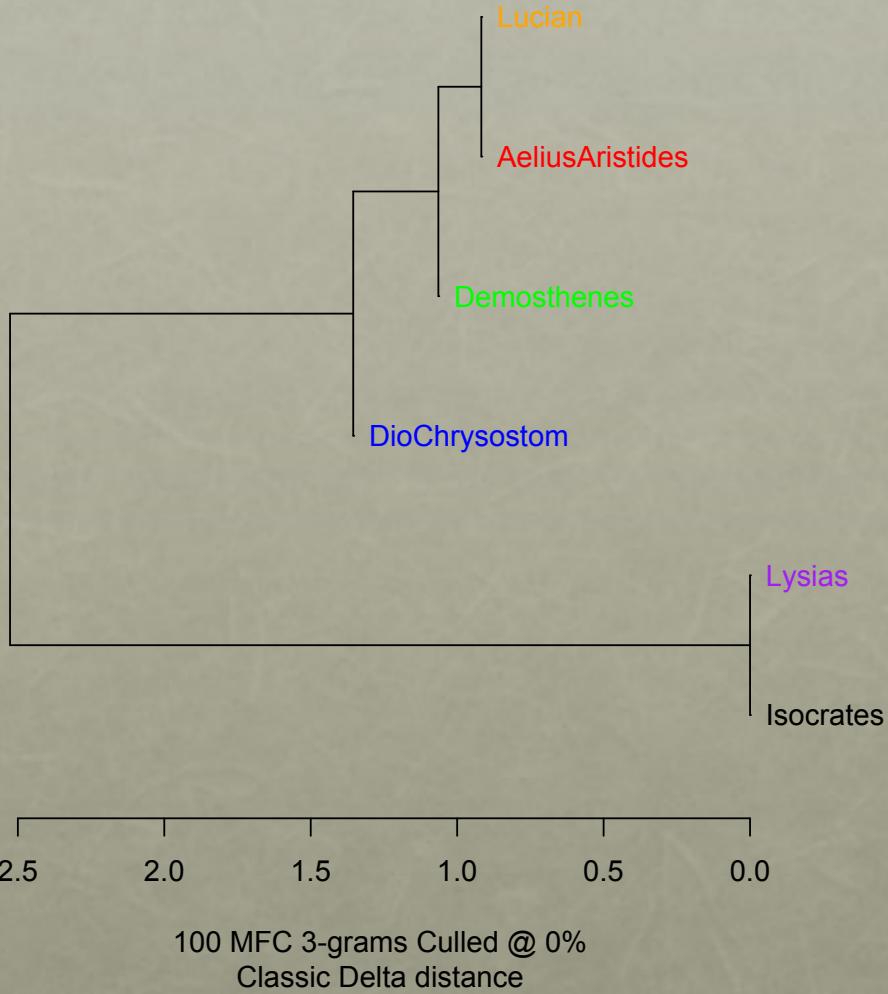
## Desktop Cluster Analysis



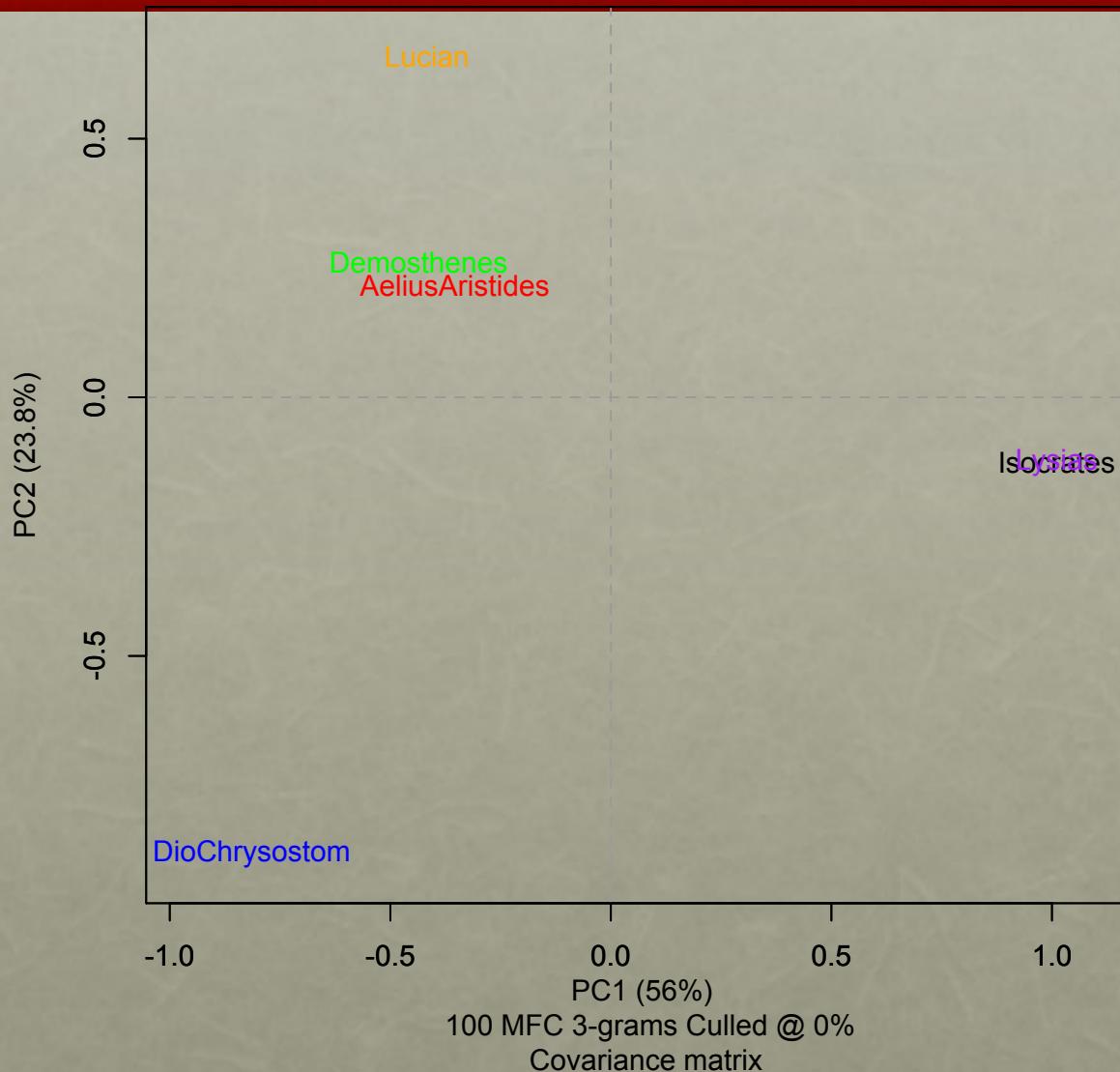
Desktop  
Principal Components Analysis



## Desktop Cluster Analysis



## Desktop Principal Components Analysis



# THANK YOU

