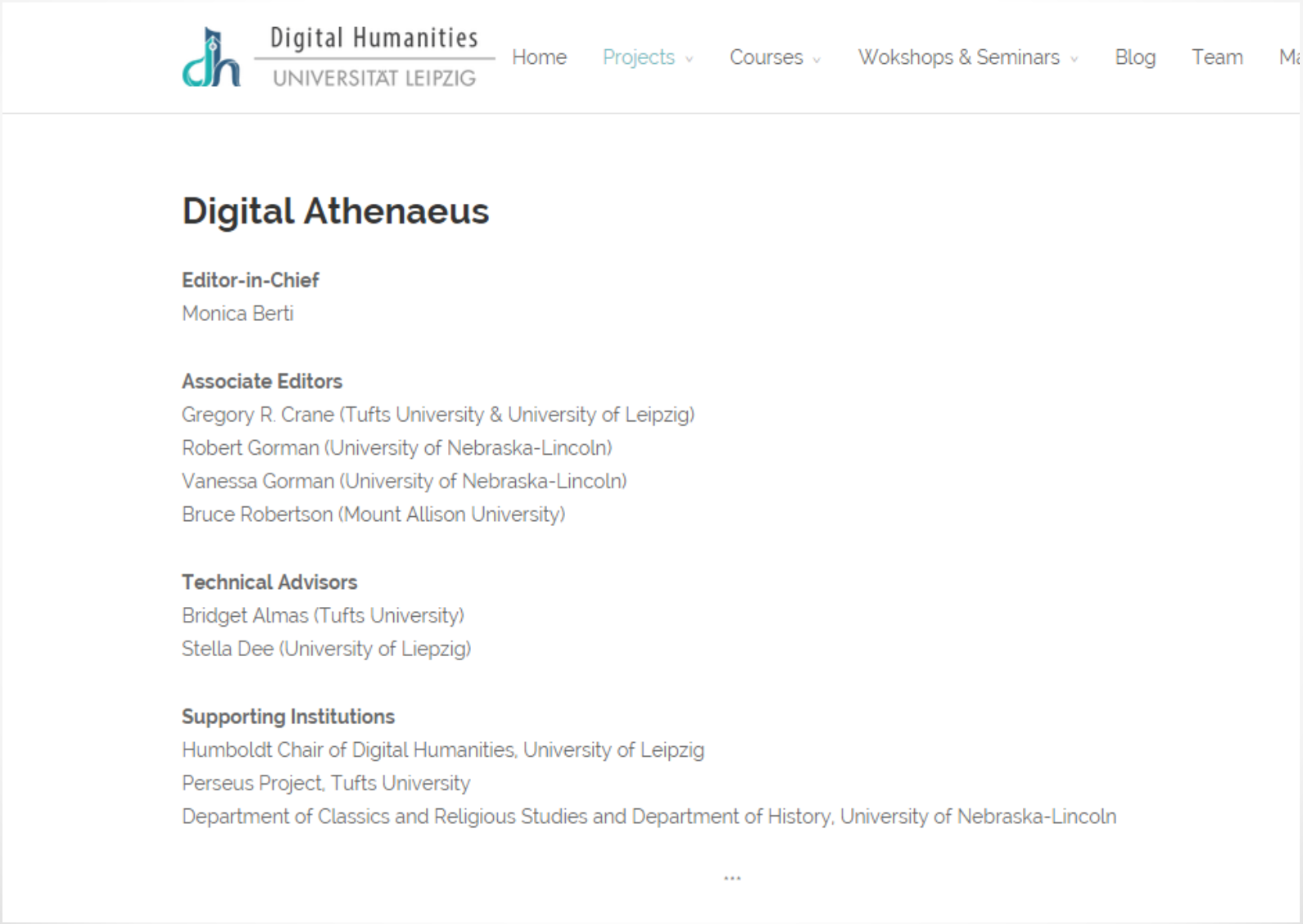
Greek Historiography Through Dependency Syntax Treebanking

Digital Classicist New England
March 25, 2015, Tufts University

Robert J. Gorman, Dept. of Classics
Vanessa B. Gorman, Dept. of History
University of Nebraska-Lincoln

http://www.dh.uni-leipzig.de/wo/projects/digital-athenaeus/



How accurate are the quotes, paraphrases, excerpts, and epitomes attributed to earlier authors?

The Layers of Athenaeus (c. 200 CE)

- Narrator (Athenaeus himself)
- The 24 Deipnosophists
- 2500+ quotes or paraphrases to 800+ writers
- All hopelessly intertangled

Corrupting Luxury in Ancient Greek Literature

By

Robert J. Gorman

and

Vanessa B. Gorman

The University of Michigan Press, Ann Arbor

Derive Syntactic "Thumbprints"

- Create a database of syntactically analyzed Greek prose
- Teach the computer to distinguish known authors (proof of concept)
- Compare directly-transmitted to epitomized prose by the same author

Epitomizers and Excerptors

- **Polybius** (2nd c. BCE) has 5 of 40 books preserved through direct transmission
  - Others mainly preserved in the excerptors working for Emperor Constantine VII Porphyrogenitus (10th c. CE)
- **Diodorus Siculus** (1st c. BCE) has 15 of 40 books preserved through direct transmission
  - Others mainly in Photius (9th century CE) and the same Constantine excerptors

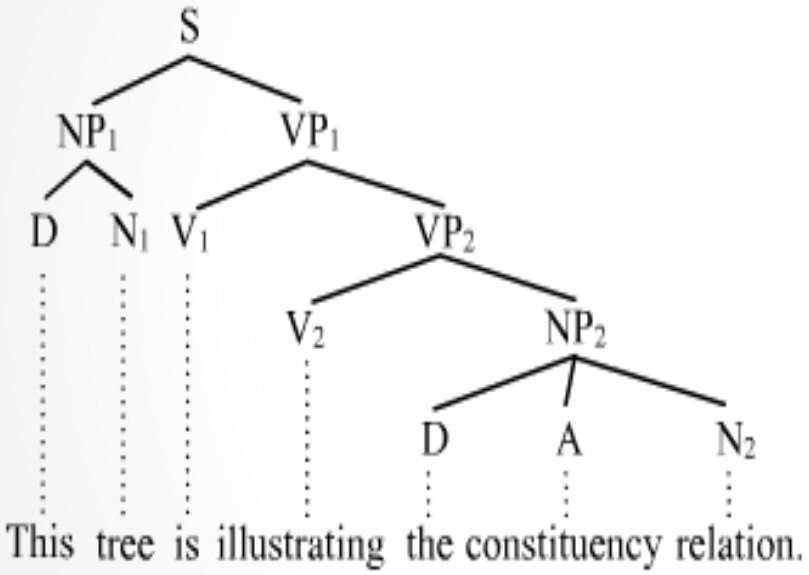
Fragments of Lost Authors

- Compare to fragments of the same author that are preserved elsewhere
- Compare to context in Athenaeus and Photius
- Does it resemble:
  - The other fragments of the same author?
  - The context in Athenaeus?

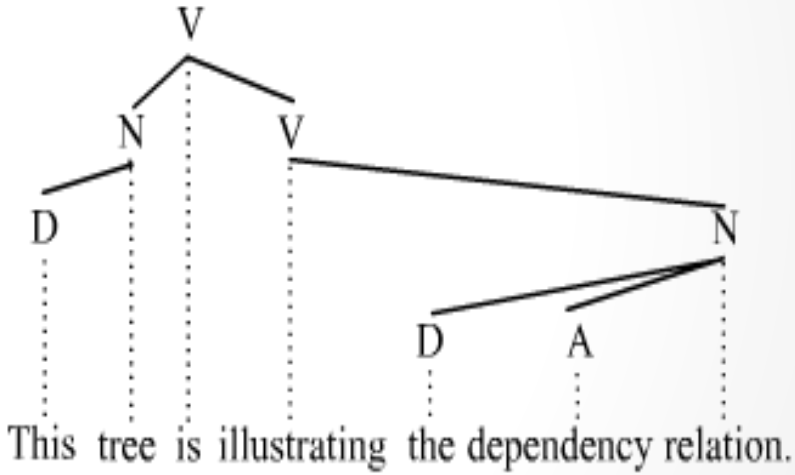
Dependency Syntax Treebanking

- Corpus Linguistics
- Annotation: create a database of syntactically-analyzed prose
- Abstraction: translate into a computer searchable dataset
- Analysis: develop algorithms to query that dataset

Dependency vs. Constituency Grammar

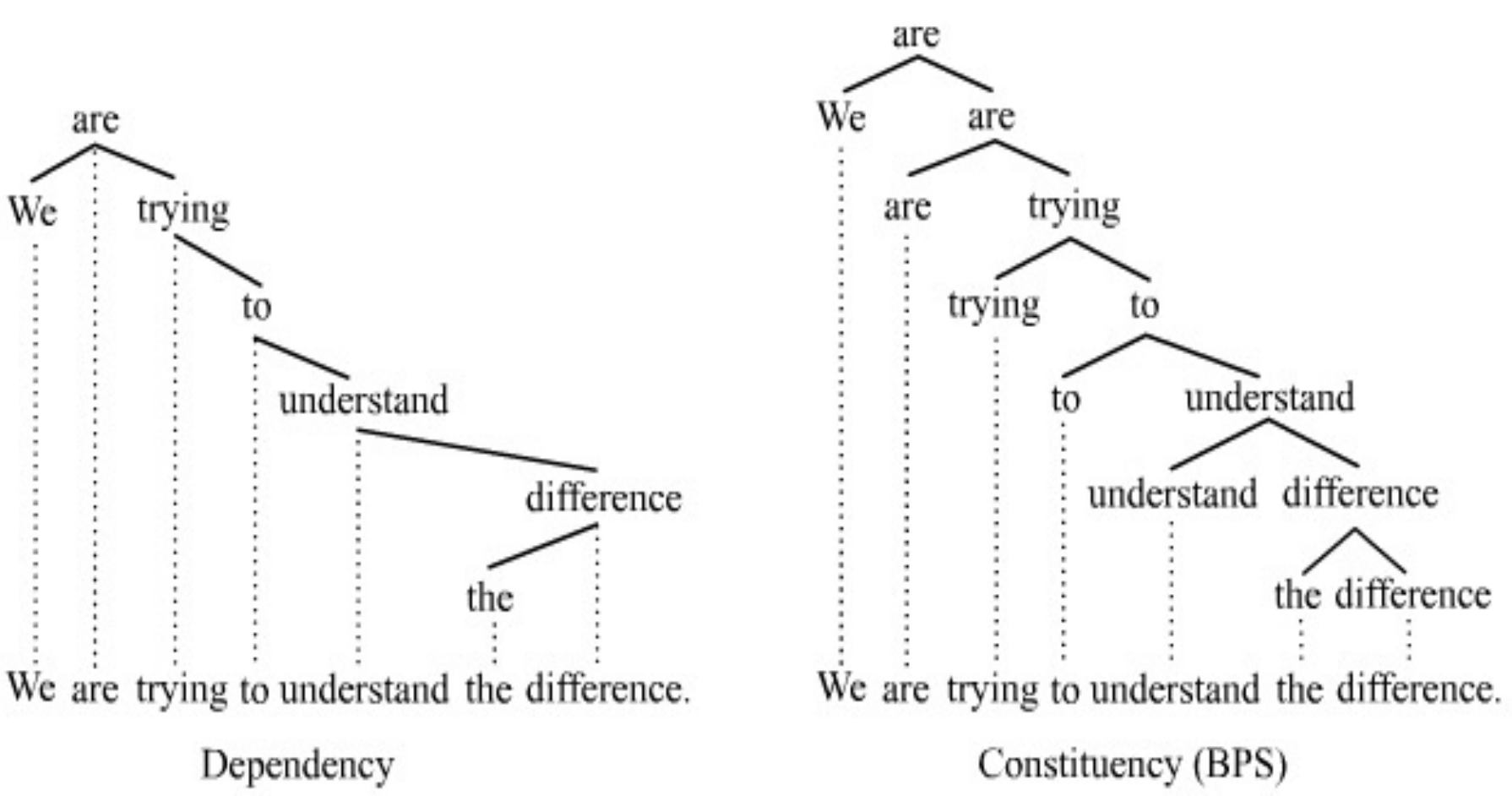


Constituency relation (PSG)



Dependency relation

Dependency vs. Constituency Grammar



http://nlp.perseus.tufts.edu/syntax/treebank/greek.html

Aeschylus	Agamemnon	Francesco Mambrini (Scuola Normale Superiore di Pisa)	9,806	XML
	Eumenides	FM	6,380	XML
	Libation Bearers	FM	6,566	XML
	Persians	FM	6,270	XML
	Prometheus Bound	FM	7,058	XML
	Seven Against Thebes	FM	6,232	XML
	Suppliants	FM	5,949	XML
Athenaeus	The Deipnosophists (Book 12)	Vanessa Gorman (University of Nebraska)	19,961	XML
Hesiod	Shield of Heracles	*standard method*	3,834	XML
	Theogony	*standard method*	8,106	XML
	Works and Days	*standard method*	6,941	XML
Homer	Iliad	*standard method*	128,102	XML
	Odyssey	*standard method*	104,467	XML
Plato	Euthyphro	Giuseppe G. A. Celano (Università degli Studi di Pavia)	6,097	XML
Sophocles	Ajax	Dan Libatique (New York University)	9,474	XML
	Antigone	Alejandro Abritta (Unvetted)	8,758	XML
	Antigone	Francesco Mambrini (Scuola Normale Superiore di Pisa)	8,751	XML
	Electra	FM	10,489	XML
	Oedipus Tyrannus	FM	11,185	XML
	Trachinae	FM	8,822	XML
Total			374,490	XML

My Dataset

AUTHOR	WORK	TOKEN COUNT	STATUS
Athenaeus	*Books 12-13*	*45,584 tokens*	*submitted*
Lysias	*Orations 1, 14, 15*	*7,650 tokens*	*submitted*
Polybius	*Book 1*	*28,288 tokens*	*submitted*
Herodotus	*Book 1*	*32,879 tokens*	*editing*
Plutarch	*Lycurgus*	*10,567 tokens*	*submitted*
Antiphon	*Oration 1*	*2,015 tokens*	*editing*
Diodorus Siculus	*Book 11*	*6,247 tokens [11.1-20 only]*	*in progress*
Thucydides	*Book 1*	*13,720 tokens [1.1-80 only]*	*in progress*
TOTAL [2/20/2015]		*146,950 tokens*	

Lysias

On the Murder of Eratosthenes

- Lys. 1 – Vanessa Gorman, Editor

Against Alcibiades 1

- Lys. 14 – Vanessa Gorman, Editor

Polybius

Histories

- Plb. 1.1-1.9 – Vanessa Gorman, Editor
- Plb. 1.10-1.19 – Vanessa Gorman, Editor
- Plb. 1.20-1.29 – Vanessa Gorman, Editor
- Plb. 1.30-1.39 – Vanessa Gorman, Editor
- Plb. 1.40-1.49, – Vanessa Gorman, Editor
- Plb. 1.50-1.59 – Vanessa Gorman, Editor
- Plb. 1.60-1.69 – Vanessa Gorman, Editor
- Plb. 1.70-1.79 – Vanessa Gorman, Editor
- Plb. 1.80-1.88 – Vanessa Gorman, Editor

παρεσκευάζετο γὰρ πολλῇ δυνάμει πλεῖν ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖ
ν
τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

"He was preparing to sail to Greece with a great force and to fight with the Greeks against the Persians."

(Diodorus 11.26.4 [sent. 58])

<word id="23" form="εὐνοίαις" lemma="" postag="" relation="nil" head=""/>
<word id="24" form="ἰδίους" lemma="" postag="" relation="nil" head=""/>
<word id="25" form="·" lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
</sentence>
<sentence id="58" document_id="urn:cts:greekLit:tlg0060.tlg001.perseus-grc3"
subdoc="11.26.4"
span="">
<word id="1" form="παρεσκευάζετο" lemma="" postag="" relation="nil" head=""/>
<word id="2" form="γὰρ" lemma="" postag="" relation="nil" head=""/>
<word id="3" form="πολλῇ" lemma="" postag="" relation="nil" head=""/>
<word id="4" form="δυνάμει" lemma="" postag="" relation="nil" head=""/>
<word id="5" form="πλεῖν" lemma="" postag="" relation="nil" head=""/>
<word id="6" form="ἐπὶ" lemma="" postag="" relation="nil" head=""/>
<word id="7" form="τὴν" lemma="" postag="" relation="nil" head=""/>
<word id="8" form="Ἑλλάδα" lemma="" postag="" relation="nil" head=""/>
<word id="9" form="καὶ" lemma="" postag="" relation="nil" head=""/>
<word id="10" form="συμμαχεῖν" lemma="" postag="" relation="nil" head=""/>
<word id="11" form="τοῖς" lemma="" postag="" relation="nil" head=""/>
<word id="12" form="Ἕλλησι" lemma="" postag="" relation="nil" head=""/>
<word id="13" form="κατὰ" lemma="" postag="" relation="nil" head=""/>
<word id="14" form="τῶν" lemma="" postag="" relation="nil" head=""/>
<word id="15" form="Περσῶν" lemma="" postag="" relation="nil" head=""/>
<word id="16" form="." lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
</sentence>
<sentence id="59" document_id="urn:cts:greekLit:tlg0060.tlg001.perseus-grc3"
subdoc="11.26.5"
span="">
<word id="1" form="ἤδη" lemma="" postag="" relation="nil" head=""/>
<word id="2" form="δ'" lemma="" postag="" relation="nil" head=""/>
<word id="3" form="αὐτοῦ" lemma="" postag="" relation="nil" head=""/>
<word id="4" form="μέλλοντος" lemma="" postag="" relation="nil" head=""/>
<word id="5" form="ποιεῖσθαι" lemma="" postag="" relation="nil" head=""/>

Overview

Edit | Edit XML | Download Copy | Change Title | History | Comment

Editing Treebank of Diod. perseus-grc3 11.21-11.30 from publication Greek Treebank Collection/2015223/1

Select a sentence to Edit

Treebank of Diod. perseus-grc3 11.21-11.30 *Next>>*

1 ὁ δὲ Γέλων καὶ αὐτὸς ἡτοιμακὼς ἦν τὴν δύναμιν , πυθόμενος δὲ τὴν τῶν Ἱμεραίων ἀθυμίαν ἀνέζευξεν ἐκ τῶν Συρακουσῶν κατὰ σπουδήν , ἔχων πεζοὺς μὲν οὐκ ἐλάττους τῶν πεντακισμυ…

2 διανύσας δὲ ταχέως τὴν ὁδὸν καὶ πλησιάσας τῇ πόλει τῶν Ἱμεραίων , ἐποίησε θαρρεῖν τοὺς πρότερον καταπεπληγμένους τὰς τῶν Καρχηδονίων δυνάμεις .

3 αὐτὸς μὲν γὰρ στρατοπεδείαν οἰκείαν βαλόμενος τῶν περὶ τὴν πόλιν τόπων , ταύτην μὲν ἀχύρωσε τάφρῳ βαθείᾳ καὶ χαρακώματι περιλαβών , τοὺς δ' ἱππεῖς ἅπαντας ἐξαπέστειλεν ἐπὶ τοὺς…

4 οὗτοι δὲ παραδόξως ἐπιφανέντες διεσπαρμένοις ἀτάκτως κατὰ τὴν χώραν , τοσούτους ἀνῆγον αἰχμαλώτους ὅσους ἕκαστος ἄγειν ἠδύνατο .

5 εἰσαχθέντων δὲ αἰχμαλώτων εἰς τὴν πόλιν πλειόνων ἢ μυρίων , ὁ μὲν Γέλων μεγάλης ἀποδοχῆς ἐτύγχανεν , οἱ δὲ κατὰ τὴν Ἱμέραν κατεφρόνησαν τῶν πολεμίων .

6 ἀκόλουθα δὲ τούτοις πράττων ὁ μὲν Γέλων ἁπάσας τὰς πύλας , ἃς διὰ φόβον πρότερον ἐνῳκοδόμησαν οἱ περὶ Θήρωνα , ταύτας τ- οὐναντίον διὰ τὴν καταφρόνησιν ἐξῳκοδόμησε , καὶ ἄλλ…

7 καθόλου δὲ Γέλων στρατηγίᾳ καὶ συνέσει διαφέρων εὐθὺς ἐζήτει , δι' οὗ τρόπου καταστρατηγήσας τοὺς βαρβάρους ἀκινδύνως αὐτῶν ἄρδην ἀνελεῖ τὴν δύναμιν .

8 συνεβάλετο δὲ αὐτῷ καὶ τὸ αὐτόματον πρὸς τὴν ἐπίνοιαν μεγάλα , τοιαύτης γενομένης περιστάσεως .

9 κρίναντος αὐτοῦ τὰς τῶν πολεμίων ναῦς ἐμπρῆσαι , καὶ τοῦ Ἀμίλκα διατρίβοντος μὲν κατὰ τὴν ναυτικὴν στρατοπεδείαν , παρασκευαζομένου δὲ θύειν τῷ Ποσειδῶνι μεγαλοπρεπῶς , ἦκον…

10 οὔσης δὲ τῆς ἡμέρας ταύτης καθ' ἣν ἔμελλε συντελεῖν τὴν θυσίαν Ἀμίλκας , κατὰ ταύτην Γέλων ἀπέστειλεν ἰδίους ἱππεῖς , οἷς ἦν προστεταγμένον περιελθεῖν τοὺς πλησίον τόπους καὶ πρ…

11 ἐξέπεμψε δὲ καὶ σκοποὺς εἰς τοὺς ὑπερκειμένους λόφους , οἷς προσέταξεν , ὅταν ἴδωσι τοὺς ἱππεῖς γενομένους ἐντὸς τοῦ τείχους , ἆραι τὸ σύσσημον .

παρεσκευάζετο γὰρ πολλῇ δυνάμει πλεῖν ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖν τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

selection none 15 unused highlight unused

morph relation aT search history comments

παρεσκευάζετο γὰρ πολλῇ δυνάμει πλεῖν ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖν τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

selection none 15 unused highlight unused

Color Legends

morph (active)

Part of Speech

long	postag
---	_
adjective	a
adposition	r
adverb	d
article	l
conjunction	c
interjection	i
irregular	x
noun	n
pronoun	p

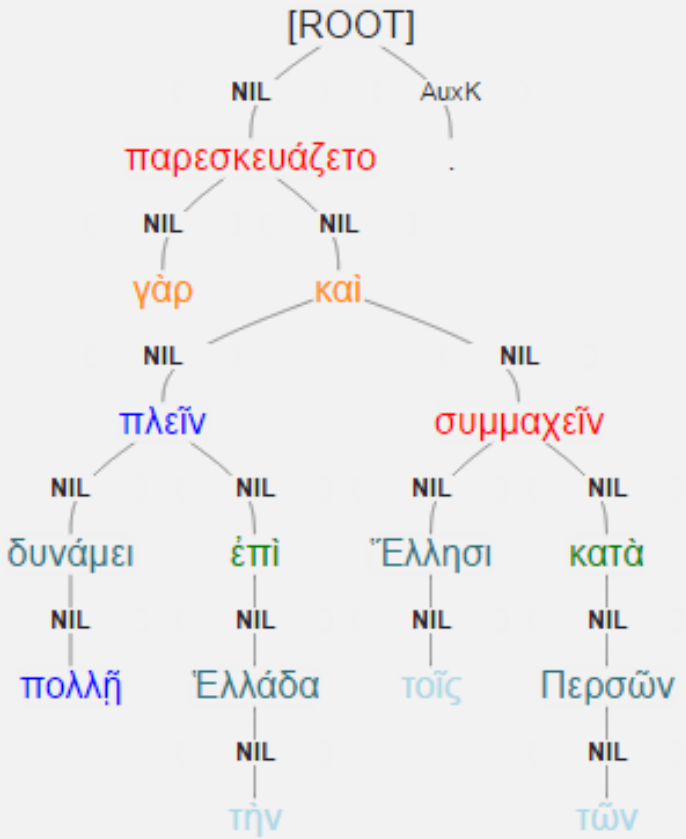
Messages

Color coding

morph | relation | aT | search | history | comments

παρεσκευάζετο γὰρ πολλῇ δυνάμει πλεῖν ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖν τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

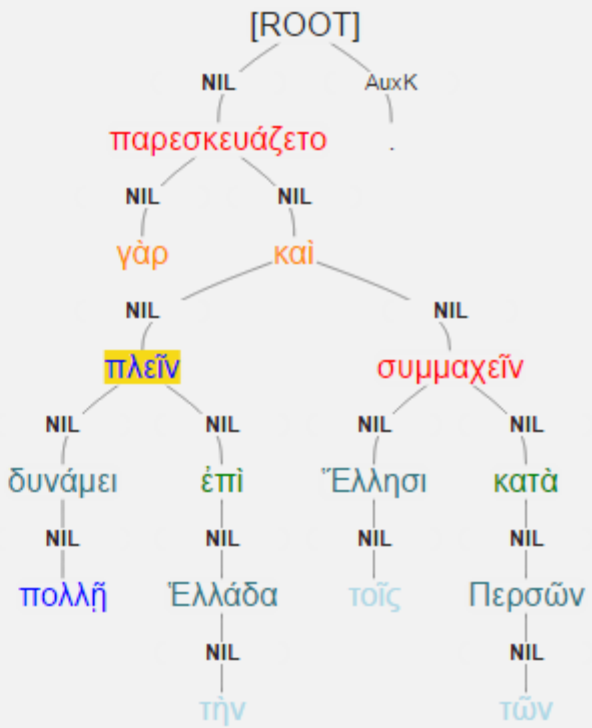
selection none 0 unused highlight unused



ments.. Diod. 11.26.4 58

ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖν τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

ed



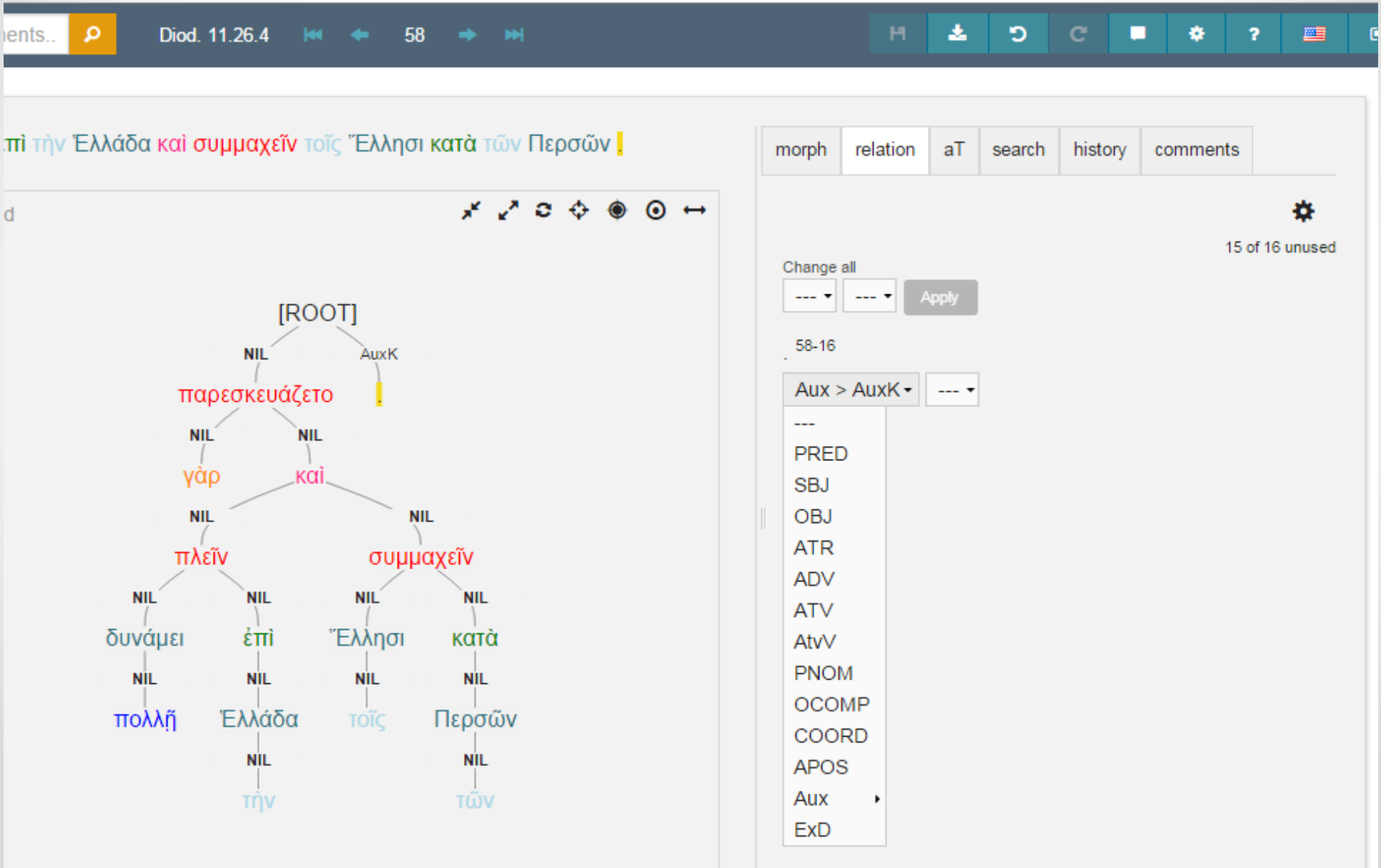
morph | relation | aT | search | history | comments

πλεῖν 58-5

☑	πλείων	a-s---nn-	bsp/morpheus
	adj.sg.neut.nom		
☐	πλείων	a-s---nv-	bsp/morpheus
	adj.sg.neut.voc		
☐	πλείων	a-s---na-	bsp/morpheus
	adj.sg.neut.acc		
☐	πλέω	v--pna---	bsp/morpheus
	verb.pr.inf.act		

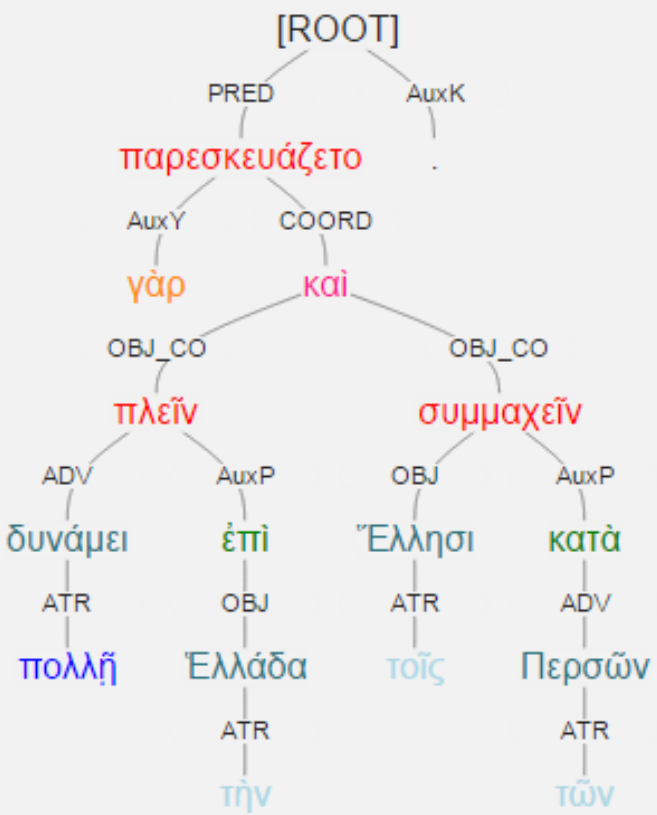
Create new form

Prague tagset



παρεσκευάζετο γὰρ πολλῇ δυνάμει πλεῖν ἐπὶ τὴν Ἑλλάδα καὶ συμμαχεῖν τοῖς Ἕλλησι κατὰ τῶν Περσῶν .

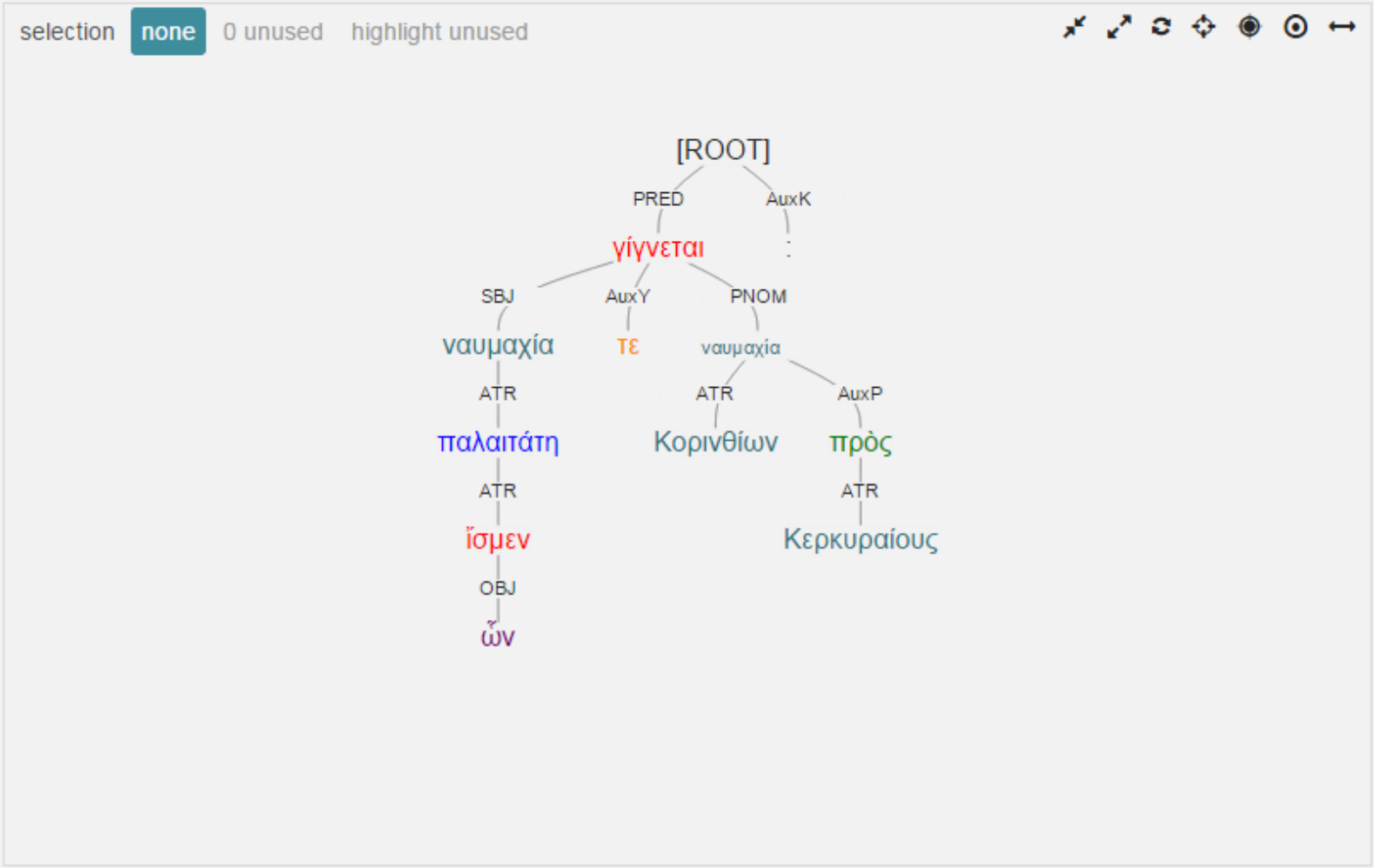
selection none 0 unused highlight unused



```
</sentence>
<sentence id="58" document_id="urn:cts:greekLit:tlg0060.tlg001.perseus-grc3"
        subdoc="11.26.4"
        span="">
  <word id="1" form="παρεσκευάζετο" lemma="παρασκευάζω" postag="v3siie---"
      relation="PRED"
      head="0"/>
  <word id="2" form="γὰρ" lemma="γάρ" postag="d--------" relation="AuxY" head="1"/>
  <word id="3" form="πολλῇ" lemma="πολύς" postag="a-s---fd-" relation="ATR" head="4"/>
  <word id="4" form="δυνάμει" lemma="δύναμις" postag="n-s---fd-" relation="ADV"
      head="5"/>
  <word id="5" form="πλεῖν" lemma="πλέω" postag="v--pna---" relation="OBJ_CO" head="9"/>
  <word id="6" form="ἐπὶ" lemma="ἐπί" postag="r--------" relation="AuxP" head="5"/>
  <word id="7" form="τὴν" lemma="ὁ" postag="l-s---fa-" relation="ATR" head="8"/>
  <word id="8" form="Ἑλλάδα" lemma="Ἑλλάς" postag="n-s---fa-" relation="OBJ" head="6"/>
  <word id="9" form="καὶ" lemma="καί" postag="c--------" relation="COORD" head="1"/>
  <word id="10" form="συμμαχεῖν" lemma="συμμαχέω" postag="v--pna---" relation="OBJ_CO"
      head="9"/>
  <word id="11" form="τοῖς" lemma="ὁ" postag="l-p---md-" relation="ATR" head="12"/>
  <word id="12" form="Ἕλλησι" lemma="Ἕλλην" postag="n-p---md-" relation="OBJ" head="10"/>
  <word id="13" form="κατὰ" lemma="κατά" postag="r--------" relation="AuxP" head="10"/>
  <word id="14" form="τῶν" lemma="ὁ" postag="l-p---mg-" relation="ATR" head="15"/>
  <word id="15" form="Περσῶν" lemma="Πέρσης" postag="n-p---mg-" relation="ADV"
      head="13"/>
  <word id="16" form="." lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
</sentence>
<sentence id="59" document_id="urn:cts:greekLit:tlg0060.tlg001.perseus-grc3"
        subdoc="11.26.5"
        span="">
  <word id="1" form="ἤδη" lemma="ἤδη" postag="d--------" relation="ADV" head="4"/>
  <word id="2" form="δ'" lemma="δέ" postag="d--------" relation="AuxY" head="9"/>
  <word id="3" form="αὐτοῦ" lemma="αὐτός" postag="p-s---mg-" relation="SBJ" head="4"/>
```

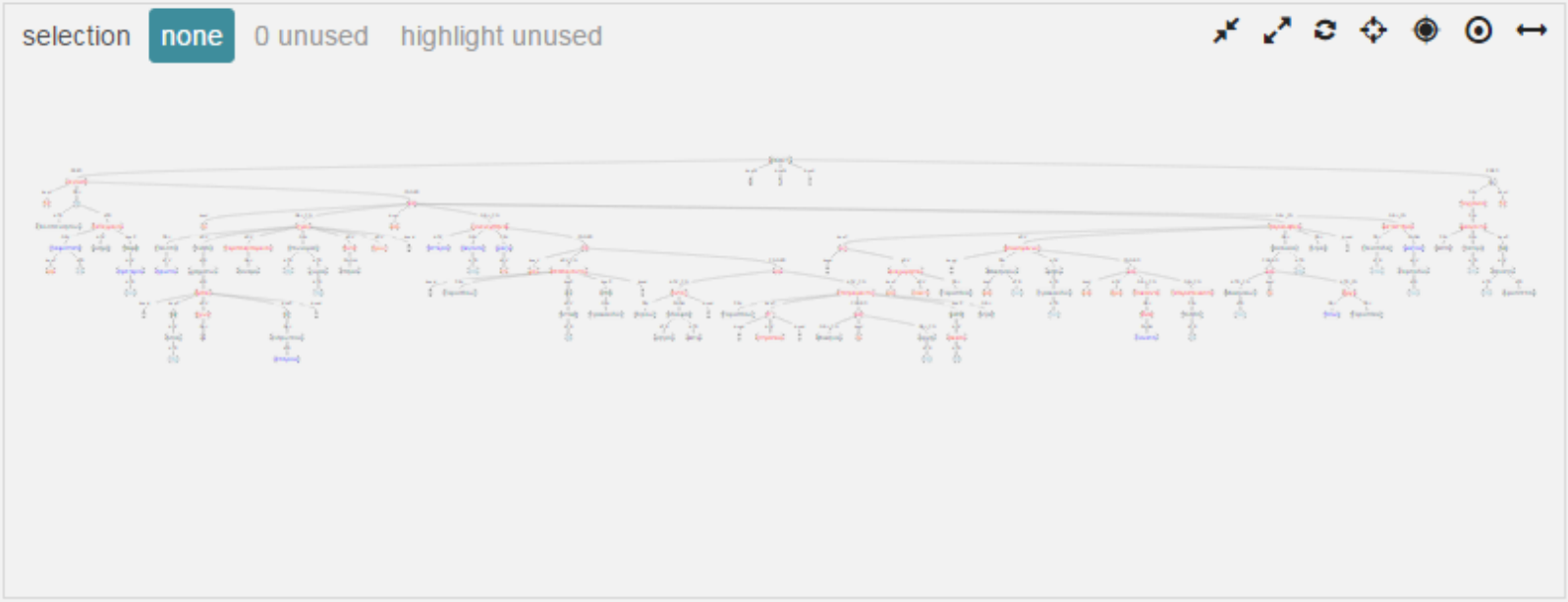
Thuc. 1.13.4 [elision]

ναυμαχία τε παλαιτάτη ὧν ἴσμεν γίγνεται Κορινθίων πρὸς Κερκυραίους : ναυμαχία



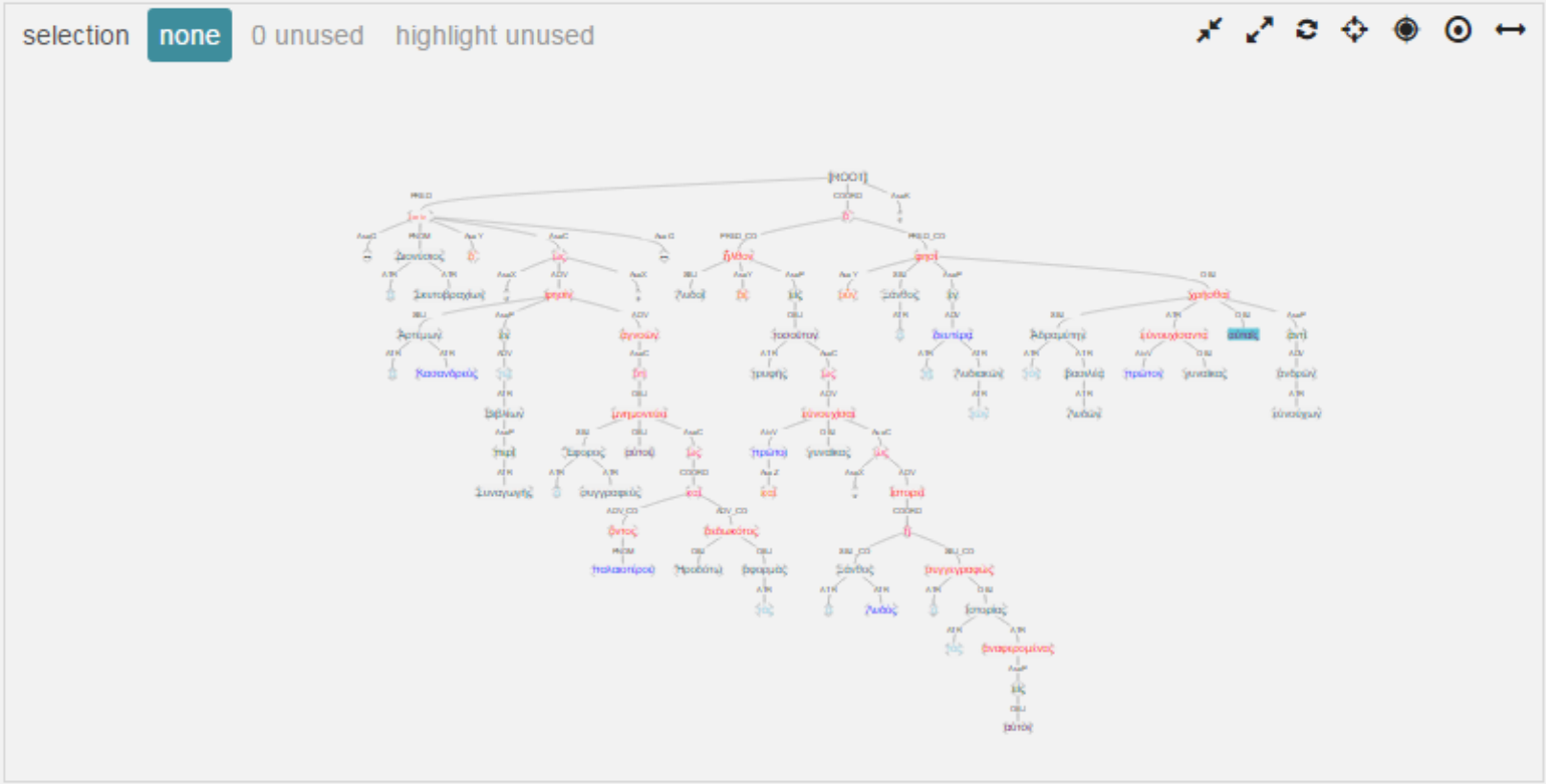
A flat tree: Thuc. 1.9.2 [135 words]

λέγουσι δὲ καὶ οἱ τὰ σαφέστατα Πελοποννησίων μνήμῃ παρὰ τῶν πρότερον δεδεγμένοι Πέλοπά τε πρῶτον πλήθει χρημάτων , ἃ ἦλθεν ἐκ τῆς Ἀσίας ἔχων ἐς ἀνθρώπους ἀπόρους , δύναμιν περιποιησάμενον τὴν ἐπωνυμίαν τῆς χώρας ἔπηλυν ὄντα ὅμως σχεῖν , καὶ ὕστερον τοῖς ἐκγόνοις ἔτι μείζω ξυνενεχθῆναι , Εὐρυσθέως μὲν ἐν τῇ Ἀττικῇ ὑπὸ Ἡρακλειδῶν ἀποθανόντος , Ἀτρέως δὲ μητρὸς ἀδελφοῦ ὄντος αὐτῷ , καὶ ἐπιτρέψαντος Εὐρυσθέως , ὅτ᾽ ἐστράτευε , Μυκήνας τε καὶ τὴν ἀρχὴν κατὰ τὸ οἰκεῖον Ἀτρεΐ (τυγχάνειν δὲ αὐτὸν φεύγοντα τὸν πατέρα διὰ τὸν Χρυσίππου θάνατον) , καὶ ὡς οὐκέτι ἀνεχώρησεν Εὐρυσθεύς , βουλομένων καὶ τῶν Μυκηναίων φόβῳ τῶν Ἡρακλειδῶν καὶ ἅμα δυνατὸν δοκοῦντα εἶναι καὶ τὸ πλῆθος τεθεραπευκότα τῶν Μυκηναίων τε καὶ ὅσων Εὐρυσθεὺς ἦρχε τὴν βασιλείαν Ἀτρέα παραλαβεῖν , καὶ τῶν Περσειδῶν τοὺς Πελοπίδας μείζους καταστῆναι . [0]



A deep tree: Athen. 12.11 [82 words]

ἐστὶν Λυδοί δὲ εἰς τοσοῦτον ἦλθον τρυφῆς ὡς καὶ πρῶτοι γυναῖκας εὐνουχίσαι , ὡς ἱστορεῖ Ξάνθος ὁ Λυδὸς ἢ ὁ τὰς εἰς αὐτὸν ἀναφερομένας ἱστορίας συγγεγραφὼς -- Διονύσιος δ' ὁ Σκυτοβραχίων , ὡς Ἀρτέμων φησὶν ὁ Κασανδρεὺς ἐν τῷ περὶ Συναγωγῆς Βιβλίων , ἀγνοῶν ὅτι Ἔφορος ὁ συγγραφεὺς μνημονεύει αὐτοῦ ὡς παλαιοτέρου ὄντος καὶ Ἡροδότῳ τὰς ἀφορμὰς δεδωκότος -- ὁ δ' οὖν Ξάνθος ἐν τῇ δευτέρᾳ τῶν Λυδιακῶν Ἀδραμύτην φησὶ τὸν Λυδῶν βασιλέα πρῶτον γυναῖκας εὐνουχίσαντα χρῆσθαι αὐταῖς ἀντὶ ἀνδρῶν εὐνούχων .



The Ancient Greek and Latin Dependency Treebanks

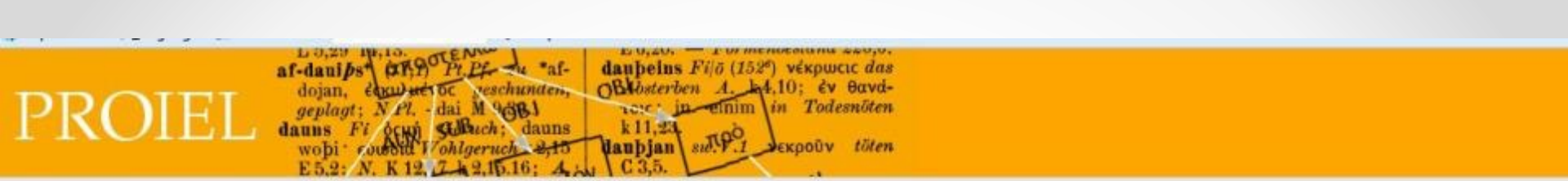
Ancient Greek Data

The Ancient Greek Dependency Treebank includes the entirety of Homer's *Iliad* and *Odyssey*; Sophocles' *Ajax*, *Antigone*, *Electra*, *Oedipus Tyrannus* and *Trachinae*; Plato's *Euthyphro*; book 12 of Athenaeus' *Deipnosophists*; and all of the works of Hesiod and Aeschylus - a total of 374,490 words.

We use two methods for building treebanks of Classical texts: a "standard" production method in which two people independently annotate each sentence and a third reconciles their differences; and a "scholarly" method where a single individual creates an annotation that stands as their published interpretation of the text. In our standard method, we attempt to remove the bias of any single individual; in a scholarly method, what we are trying to capture is exactly that unique interpretation. All of these works are freely available for download under a Creative Commons license.

We will also publish unvetted annotations via the scholarly method by individuals who complete entire texts and would like their work published. These interpretations are not included in the official Perseus Treebank but are available for download here.

Author	Work	Annotator	Word count	
Aeschylus	Agamemnon	Francesco Mambrini (Scuola Normale Superiore di Pisa)	9,806	XML
	Eumenides	FM	6,380	XML
	Libation Bearers	FM	6,566	XML
	Persians	FM	6,270	XML
	Prometheus Bound	FM	7,058	XML



You need to sign in or sign up before continuing.

Sign in

User name*

Password*

Sign in

Register new account

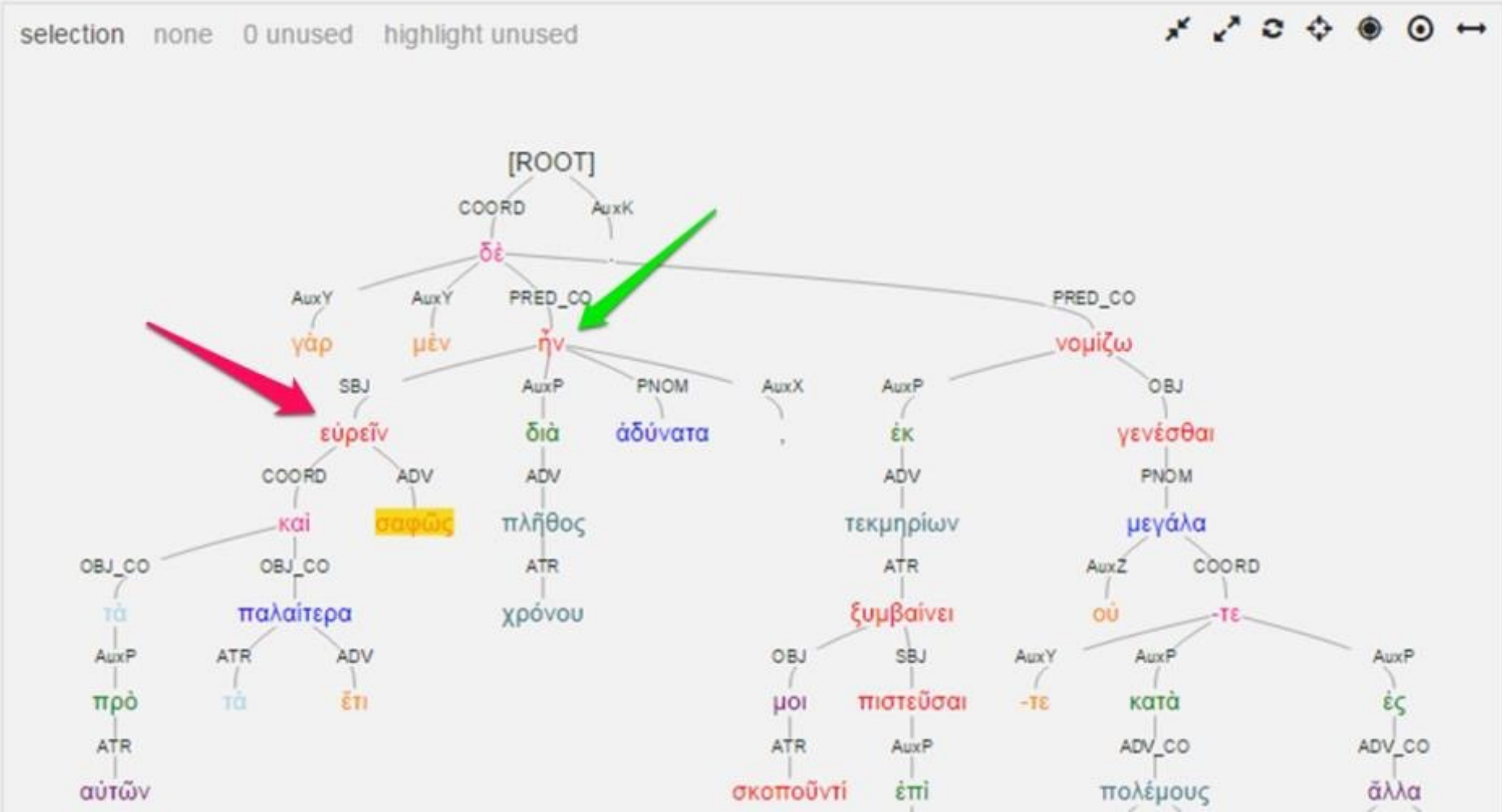
Forgot your password?

For each word in AGDT we have:

- dependency (word's parent, children)
- syntactic relation (grammatical label for dependency)
- Lemma
- Morphology
- Position in sentence

Dependency Distance

τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν , ἐκ δὲ τεκμηρίων ὧν ἐπὶ μακρότατον σκοποῦντί μοι πιστεῦσαι ξυμβαίνει οὐ μεγάλα νομίζω γενέσθαι οὔ -τε κατὰ [illegible] πολέμους οὔ -τε ἐς τὰ ἄλλα .



Dependency Degree
Linear vs. hubby structure

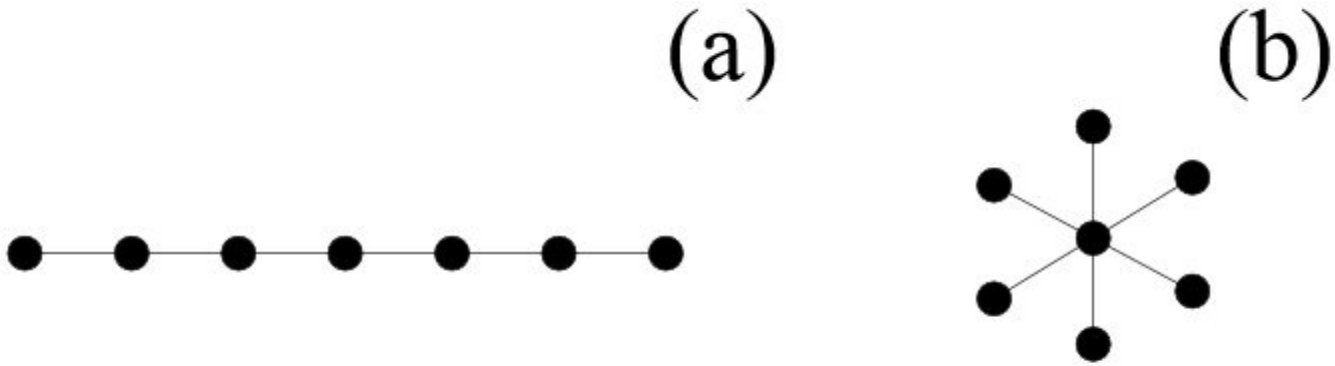
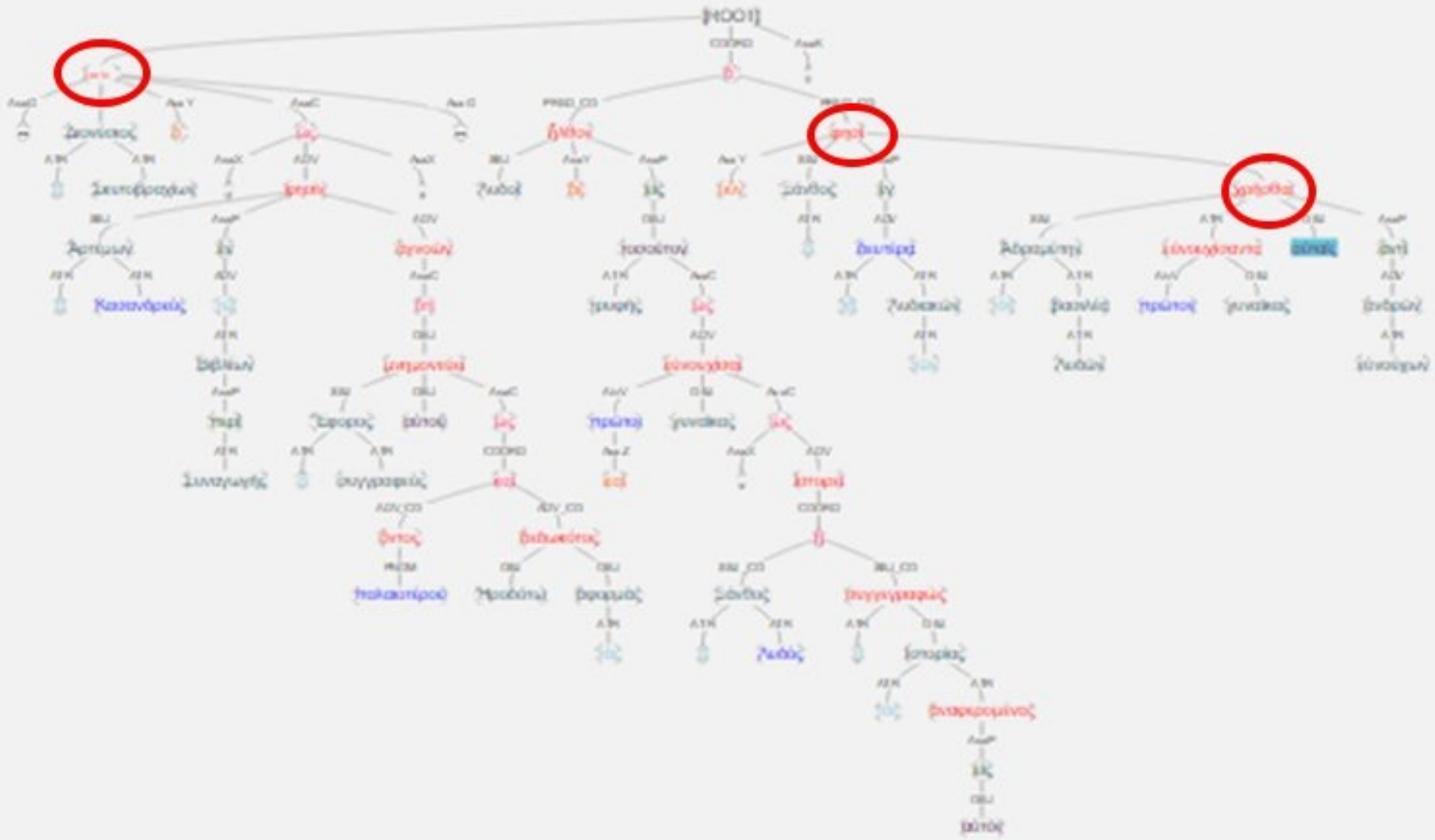
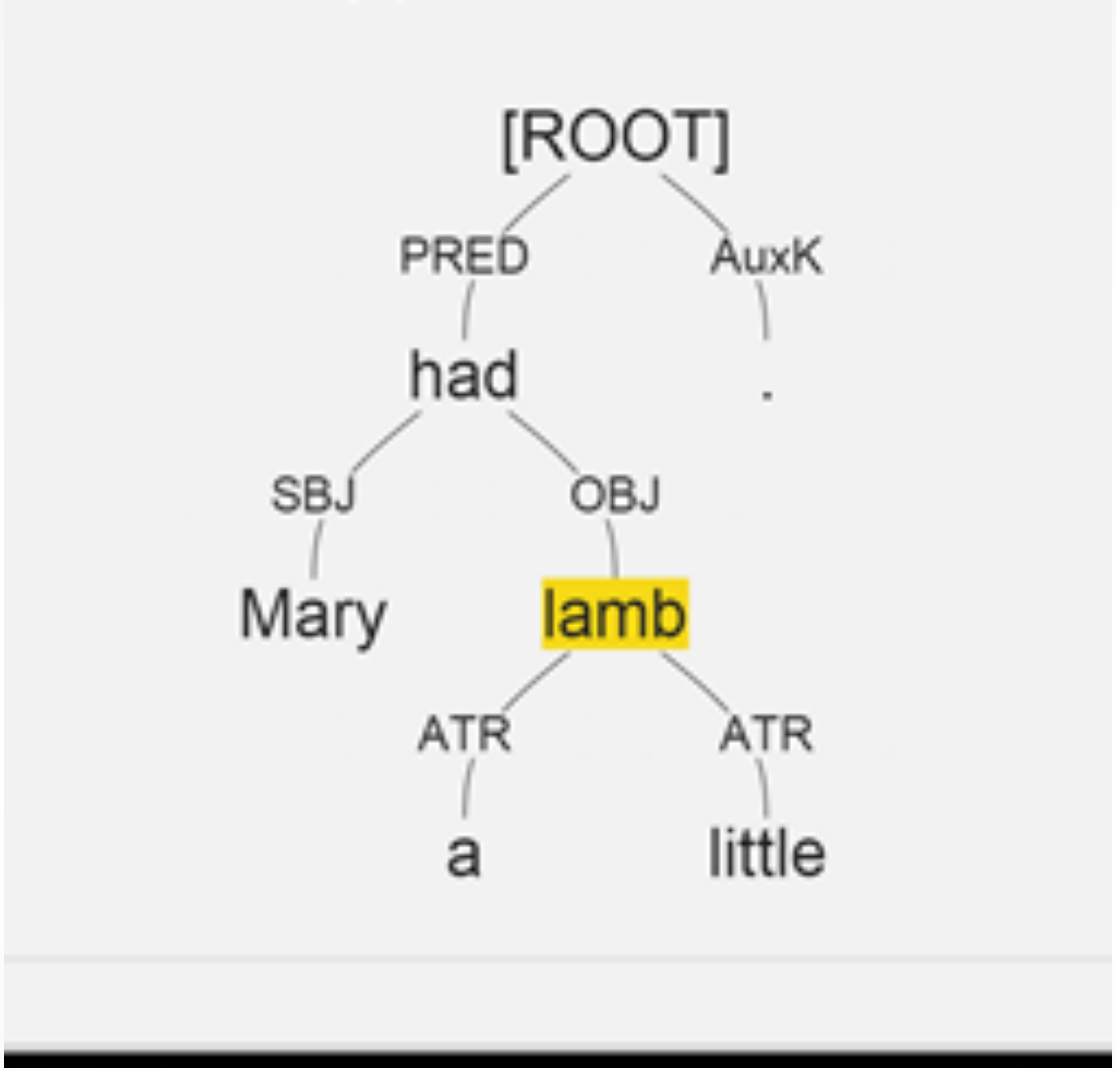
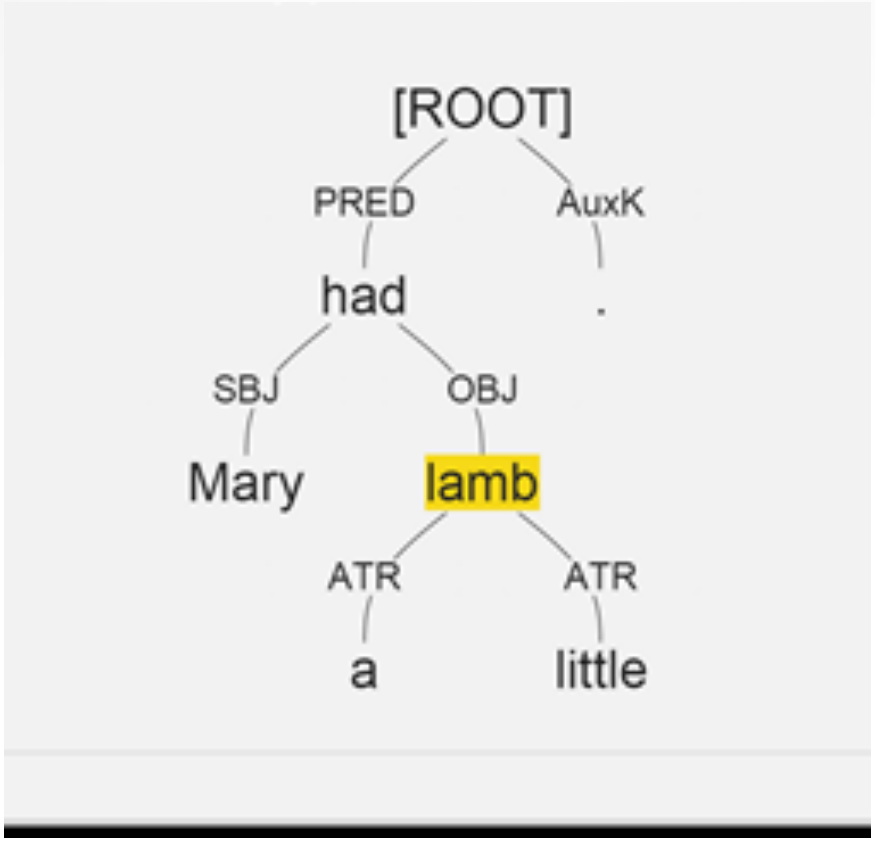


Figure 3. (a) a linear tree and (b) a star tree




[ROOT]
PRED
AuxK
had
.
SBJ
OBJ
Mary
lamb
ATR
ATR
a
little



Mary: SBJ-PRED-ROOT
had: PRED-ROOT
a: ATR-OBJ-PRED-ROOT
little: ATR-OBJ-PRED-ROOT
lamb: OBJ-PRED-ROOT

XML

```
<annotator>
  <short>Vanessa Gorman</short>
  <name>Vanessa Gorman</name>
  <address>vbgorman@gmail.com</address>
  <uri>http://data.perseus.org/sosol/users/Vanessa%20Gorman</uri>
</annotator>
<sentence id="1"
      document_id="http://perseids.org/annotsrc/urn:cts:greekLit:tlg0003.tlg001.perseus-grc1"
      subdoc="1.1.1"
      span="">
  <word id="1" form="Θουκυδίδης" lemma="Θουκυδίδης" postag="n-s---mn-" relation="SBJ"
      head="3"/>
  <word id="2" form="Ἀθηναῖος" lemma="Ἀθηναῖος" postag="n-s---mn-" relation="ATR"
      head="1"/>
  <word id="3" form="ξυνέγραψε" lemma="συγγράφω" postag="v3saia---" relation="PRED"
      head="0"/>
  <word id="4" form="τὸν" lemma="ὁ" postag="l-s---ma-" relation="ATR" head="5"/>
  <word id="5" form="πόλεμον" lemma="πόλεμος" postag="n-s---ma-" relation="OBJ_AP"
      head="10"/>
  <word id="6" form="τῶν" lemma="ὁ" postag="l-p---mg-" relation="ATR" head="7"/>
  <word id="7" form="Πελοποννησίων" lemma="Πελοποννήσιοι" postag="n-p---mg-"
      relation="ATR_CO"
      head="8"/>
  <word id="8" form="καὶ" lemma="καί" postag="c--------" relation="COORD" head="5"/>
  <word id="9" form="Ἀθηναίων" lemma="Ἀθήναιος" postag="n-p---mg-" relation="ATR_CO"
      head="8"/>
  <word id="10" form="," lemma="punc1" postag="u--------" relation="APOS" head="3"/>
  <word id="11" form="ὡς" lemma="ὡς" postag="c--------" relation="AuxC" head="10"/>
  <word id="12" form="ἐπολέμησαν" lemma="πολεμέω" postag="v3paia---" relation="OBJ_AP"
      head="11"/>
  <word id="13" form="πρὸς" lemma="πρός" postag="r--------" relation="AuxP" head="12"/>
  <word id="14" form="ἀλλήλους" lemma="ἀλλήλων" postag="p-p---ma-" relation="ADV"
      head="13"/>
  <word id="15" form="," lemma="punc1" postag="u--------" relation="AuxX" head="19"/>
  <word id="16" form="ἀρξάμενος" lemma="ἄρχω" postag="v-sapmmn-" relation="ADV_CO"
      head="19"/>
  <word id="17" form="εὐθὺς" lemma="εὐθύς2" postag="d--------" relation="ADV" head="16"/>
  <word id="18" form="καθισταμένου" lemma="καθιστάω" postag="v-sppeng-" relation="OBJ"
      head="16"/>
```

Edit summary (Briefly describe the changes you have made):

XQuery Tutorial

« W3Schools Home Next Chapter »

XQuery is to XML what SQL is to database tables.

XQuery is designed to query XML data - not just XML files, but anything that can appear as XML, including databases.

What is XQuery?

- XQuery is ***the*** language for querying XML data
- XQuery for XML is like SQL for databases
- XQuery is built on XPath expressions
- XQuery is supported by all major databases
- XQuery is a W3C Recommendation

XQuery is About Querying XML

XQuery is a language for finding and extracting elements and attributes from XML documents.

Here is an example of a question that XQuery could solve:

"Select all CD records with a price less than $10 from the CD collection stored in the XML document called cd_catalog.xml"

XQuery and XPath

XQuery 1.0 and XPath 2.0 share the same data model and support the same functions and operators. If you have already studied XPath you will have no problems with understanding XQuery.

You can read more about XPath in our XPath Tutorial.

XQuery - Examples of Use

XQuery can be used to:

sWord_Rel_withSubDoc.xquery

```
for $s in //sentence

return <sentence StndRef="Diodorus Siculus Book 11" subdoc="{$s/@subdoc}" id="{$s/@id}"  document_id="{$s/@document_id}">

{for $w in $s/word
let $wr := data($w/@relation)
let $o := $w/parent::sentence/word[@id = $w/@head]
let $or :=data($o/@relation)
let $p := $w/parent::sentence/word[@id = $o/@head]
let $pr := data($p/@relation)
let $q := $w/parent::sentence/word[@id = $p/@head]
let $qr := data($q/@relation)
let $r := $w/parent::sentence/word[@id = $q/@head]
let $rr :=data($r/@relation)
let $a := $w/parent::sentence/word[@id = $r/@head]
let $ar := data($a/@relation)
let $b := $w/parent::sentence/word[@id = $a/@head]
let $br := data($b/@relation)
let $c := $w/parent::sentence/word[@id = $b/@head]
let $cr := data($c/@relation)
let $d := $w/parent::sentence/word[@id = $c/@head]
let $dr := data($d/@relation)
let $e := $w/parent::sentence/word[@id = $d/@head]
let $er := data($e/@relation)
let $f := $w/parent::sentence/word[@id = $e/@head]
```

Output Mapping Stack

ὦ τοῦ στρατηγήσαντος ἐν Τροίᾳ ποτὲ / Ἀγαμέμνονος παῖ
"O child of Agamemnon, once leading an army at Troy"

Soph.Elec.rel.xml

```
25      </treebankSource>
26
27      <sentence id="2899145"
28            document_id="Perseus:text:9999.01.0012"
29            StndRef="Soph_Elec">
30    <sword>AuxZ*ExD*PRED#</sword>
31    <sword>ATR*ATR*ATR*ExD*PRED#</sword>
32    <sword>ATR*ATR*ExD*PRED#</sword>
33    <sword>AuxP*ATR*ATR*ExD*PRED#</sword>
34    <sword>ADV*AuxP*ATR*ATR*ExD*PRED#</sword>
35    <sword>ADV*ATR*ATR*ExD*PRED#</sword>
36    <sword>ATR*ExD*PRED#</sword>
37    <sword>ExD*PRED#</sword>
38    <sword>AuxX*ExD*PRED#</sword>
39    <sword>ADV*PRED#</sword>
40    <sword>OBJ*SBJ*PRED#</sword>
41    <sword>PRED#</sword>
42    <sword>OBJ*PRED#</sword>
43    <sword>ATR*OBJ*PRED#</sword>
44    <sword>SBJ*PRED#</sword>
45    <sword>AuxX*OBJ*SBJ*PRED#</sword>
46    <sword>ATR*PNOM*ATR*OBJ*SBJ*PRED#</sword>
47    <sword>PNOM*ATR*OBJ*SBJ*PRED#</sword>
48    <sword>ATR*OBJ*SBJ*PRED#</sword>
49    <sword>ADV*ATR*OBJ*SBJ*PRED#</sword>
50    <sword>AuxK#</sword>
```

ὦ τοῦ στρατηγήσαντος ἐν Τροίᾳ ποτὲ / Ἀγαμέμνονος παῖ
"O child of Agamemnon, once leading an army at Troy"

```
<sentence id="2899145"
      document_id="Perseus:text:9999.01.0012"
      StndRef="Soph_Elec">
<sword>AuxZ-e*ExD-n*PRED-v#</sword>
<sword>ATR-p*ATR-v*ATR-n*ExD-n*PRED-v#</sword>
<sword>ATR-v*ATR-n*ExD-n*PRED-v#</sword>
<sword>AuxP-r*ATR-v*ATR-n*ExD-n*PRED-v#</sword>
<sword>ADV-n*AuxP-r*ATR-v*ATR-n*ExD-n*PRED-v#</sword>
<sword>ADV-g*ATR-v*ATR-n*ExD-n*PRED-v#</sword>
<sword>ATR-n*ExD-n*PRED-v#</sword>
<sword>ExD-n*PRED-v#</sword>
<sword>AuxX-u*ExD-n*PRED-v#</sword>
<sword>ADV-d*PRED-v#</sword>
<sword>OBJ-a*SBJ-v*PRED-v#</sword>
<sword>PRED-v#</sword>
<sword>OBJ-p*PRED-v#</sword>
<sword>ATR-v*OBJ-p*PRED-v#</sword>
<sword>SBJ-v*PRED-v#</sword>
<sword>AuxX-u*OBJ-a*SBJ-v*PRED-v#</sword>
<sword>ATR-p*PNOM-a*ATR-v*OBJ-a*SBJ-v*PRED-v#</sword>
<sword>PNOM-a*ATR-v*OBJ-a*SBJ-v*PRED-v#</sword>
<sword>ATR-v*OBJ-a*SBJ-v*PRED-v#</sword>
<sword>ADV-d*ATR-v*OBJ-a*SBJ-v*PRED-v#</sword>
<sword>AuxK-u#</sword>
</sentence>
```

[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R 3.2.0 (Full of Ingredients) prerelease versions** will appear starting March 19. Final release is scheduled for April 16, 2015.
- **R 3.1.3 (Smooth Sidewalk) prerelease versions** will appear starting February 28. Final release is scheduled for March 9, 2015.

10Computational01
01Stylistics0101000
11Group011010110

main page

- **main page**
  - more photos
- **corpora**
- **installation hints**
- **papers and articles**
- **preprints**
- **projects**
  - Computer Methods in Textual Studies 2014
  - testing big dendrograms
  - Testing consensus networks

stylo R package

The suite of stylometric tools, so far in the form of separate scripts, has been recently ported to a regular R package. Once installed, it provides a number of functions that can be invoked from inside the R console.

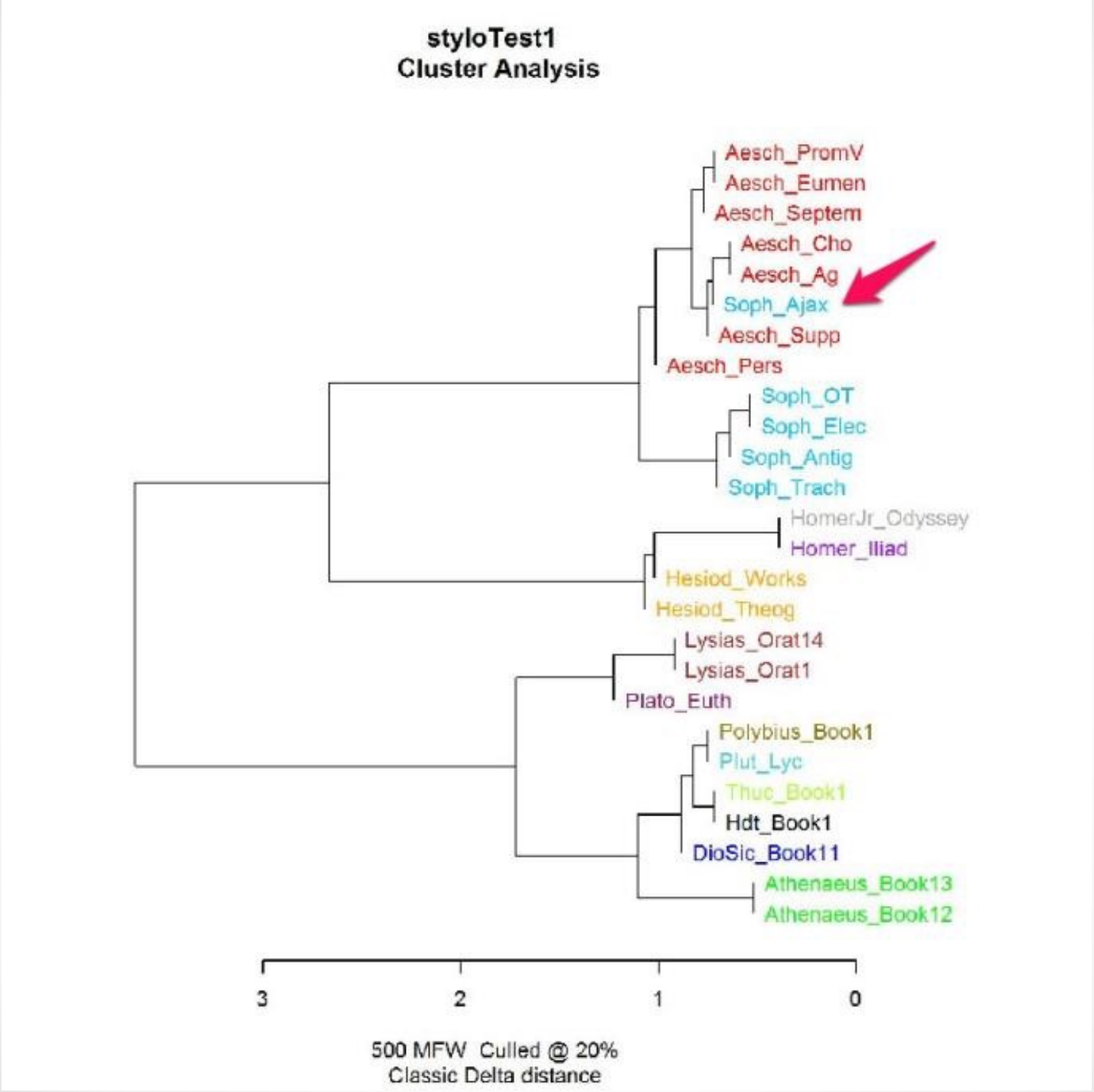
The most important functions are:

- stylo()
- classify()
- oppose()
- rolling.delta()

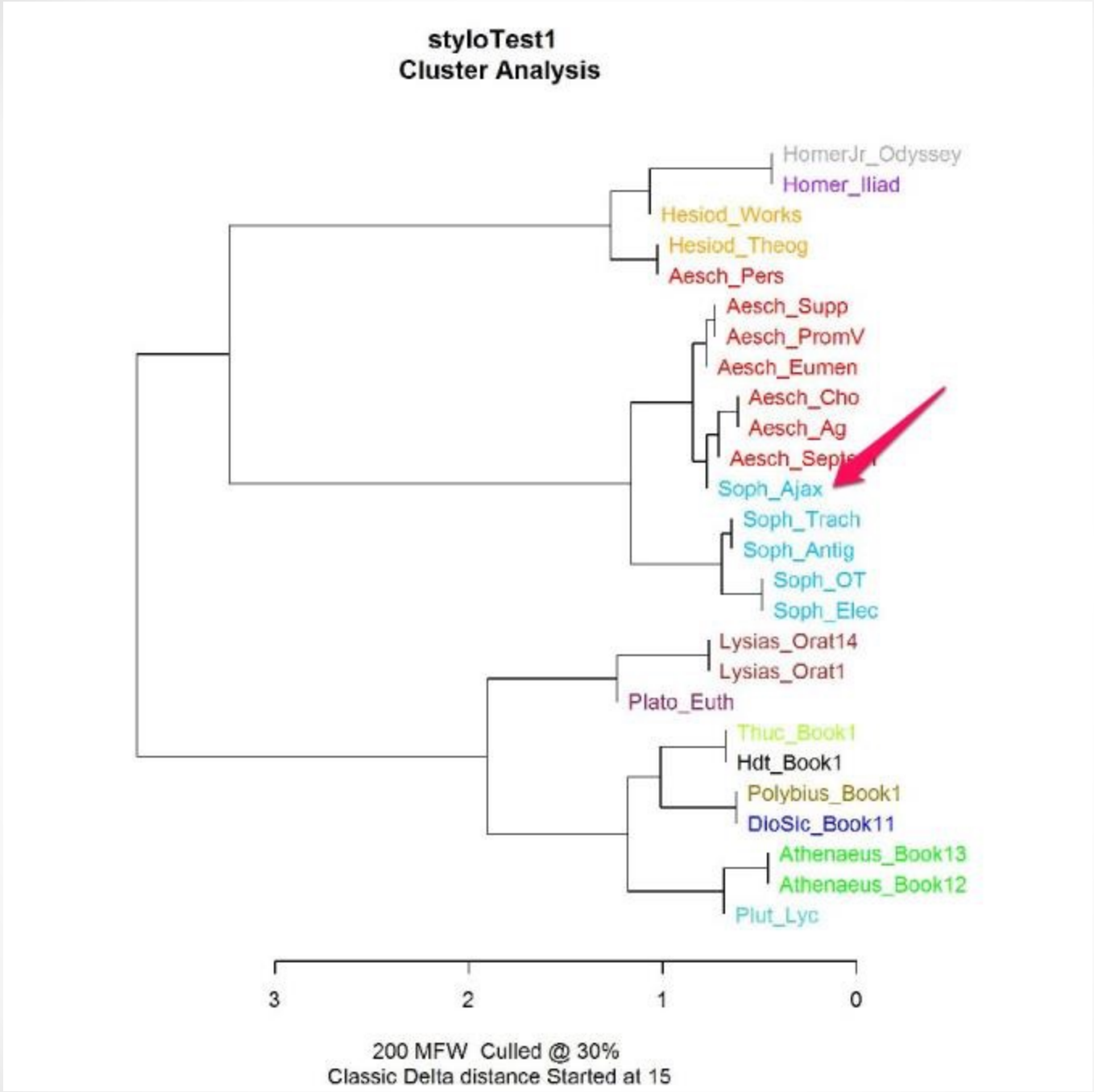
A1

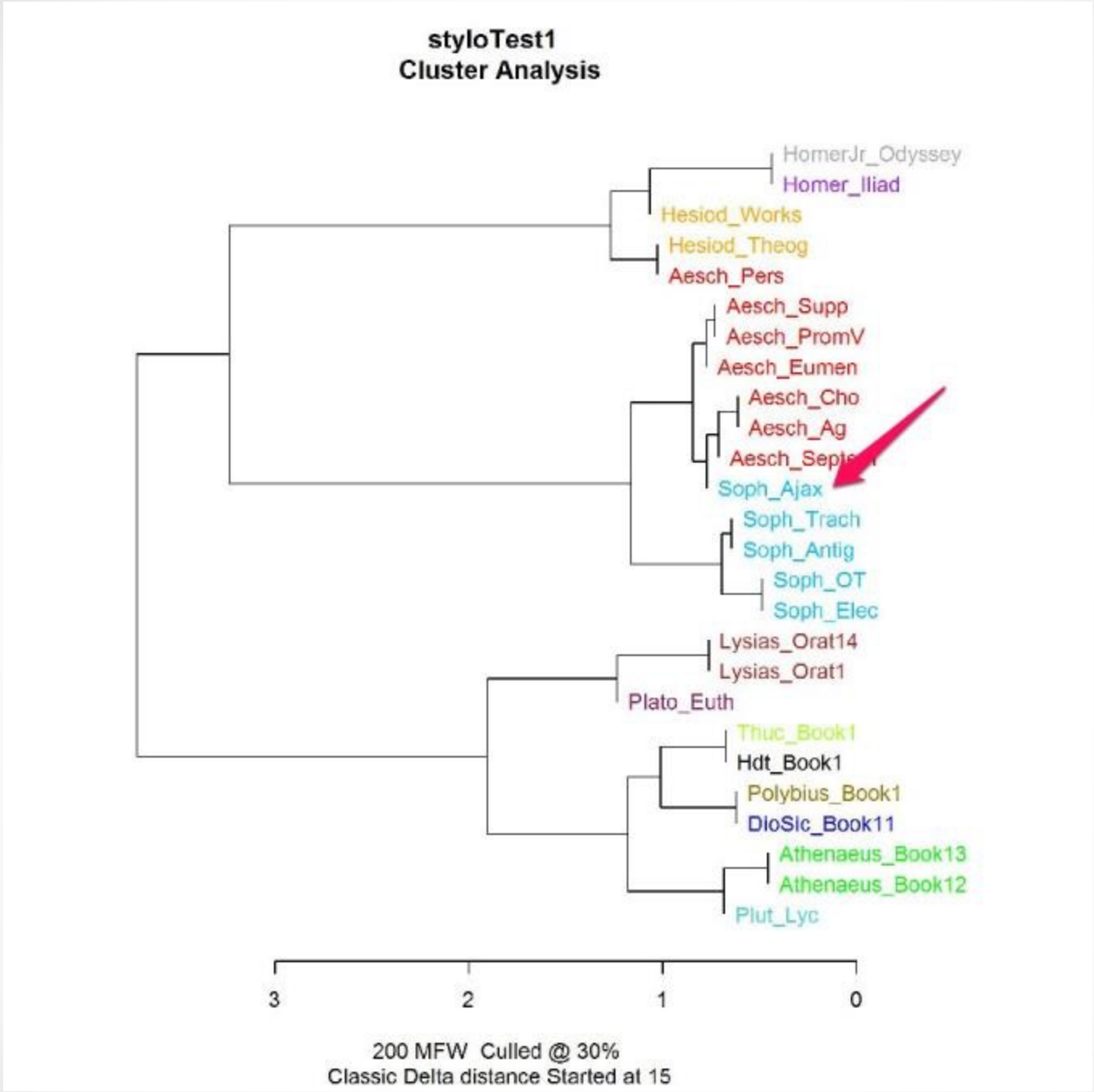
	A	B	C	D	E	F	G	H	I	J
1		Aesch_Ag	Aesch_Ch	Aesch_Eu	Aesch_Per	Aesch_Prc	Aesch_Sep	Aesch_Su	Athenaeu	Athenaeu
2	auxk-root	8.229655	8.728826	8.197492	7.6874	8.330972	8.327985	8.808203	4.953166	5.792345
3	pred-co-coord-root	2.467877	2.975736	3.181818	3.748006	3.400397	3.353659	2.58867	1.379556	1.613158
4	pred-root	4.721599	5.325805	4.937304	3.349282	5.1431	4.605263	5.58077	4.357565	5.001582
5	coord-root	2.304711	2.212727	2.257053	3.301435	2.054406	2.487163	2.067574	0.712686	0.802625
6	obj-pred-co-coord-root	1.550071	1.739661	1.61442	2.15311	1.827713	1.508344	1.412002	0.671961	0.810533
7	auxy-pred-root	2.896186	3.525103	3.197492	1.913876	3.060357	2.567394	3.630862	3.649969	3.795667
8	adv-pred-co-coord-root	1.356312	1.297116	1.285266	1.834131	1.303485	1.428113	1.008573	0.473427	0.648426
9	adv-pred-root	2.447481	2.45689	2.507837	1.770335	2.139416	2.134146	2.639099	1.939524	2.403922
10	obj-pred-root	2.967571	3.326721	3.401254	1.897927	4.122981	2.743902	3.3451	2.010792	2.653013
11	sbj-pred-co-coord-root	1.050377	1.022432	0.956113	1.594896	0.779258	1.396021	0.840477	0.488699	0.597027
12	auxy-coord-root	0.448705	0.595147	0.768025	1.212121	0.892604	0.882542	0.537906	0.743229	0.759133
13	sbj-pred-root	1.957985	1.846483	1.598746	1.339713	1.68603	1.909499	2.168432	2.48931	3.000949

	A	B	C	D	E	F	G	H	I	J	K	
1		Mean Freq	Thuc_Book1	Aesch_Ag	Aesch_Ch	Aesch_Eu	Aesch_Pe	Aesch_Prc	Aesch_Sep	Aesch_Su	Athenaeu	A
2	auxk-root	6.557	3.68075802	8.229655	8.728826	8.197492	7.6874	8.330972	8.327985	8.808203	4.953166	5
3	pred-root	3.835	2.49271137	4.721599	5.325805	4.937304	3.349282	5.1431	4.605263	5.58077	4.357565	5
4	auxy-pred-root	2.923	2.74052478	2.896186	3.525103	3.197492	1.913876	3.060357	2.567394	3.630862	3.649969	3
5	pred-co-coord-root	2.805	2.64577259	2.467877	2.975736	3.181818	3.748006	3.400397	3.353659	2.58867	1.379556	1
6	obj-pred-root	2.269	1.21720117	2.967571	3.326721	3.401254	1.897927	4.122981	2.743902	3.3451	2.010792	2
7	coord-root	2.062	1.37026239	2.304711	2.212727	2.257053	3.301435	2.054406	2.487163	2.067574	0.712686	0
8	adv-pred-root	2.058	1.44314869	2.447481	2.45689	2.507837	1.770335	2.139416	2.134146	2.639099	1.939524	2
9	sbj-pred-root	1.626	1.34110787	1.957985	1.846483	1.598746	1.339713	1.68603	1.909499	2.168432	2.48931	3
10	obj-pred-co-coord-root	1.584	1.31924198	1.550071	1.739661	1.61442	2.15311	1.827713	1.508344	1.412002	0.671961	0
11	adv-pred-co-coord-root	1.382	1.31924198	1.356312	1.297116	1.285266	1.834131	1.303485	1.428113	1.008573	0.473427	0
12	atr-obj-pred-root	1.120	0.45918367	1.65205	1.846483	1.630094	1.052632	2.238595	1.460205	2.23567	0.962126	1
13	auxy-coord-root	1.048	1.8877551	0.448705	0.595147	0.768025	1.212121	0.892604	0.882542	0.537906	0.743229	0
14	atr-sbj-pred-root	1.033	1.0058309	1.233938	1.068213	0.705329	1.307815	0.920941	1.010911	1.512859	2.199145	2
15	sbj-pred-co-coord-root	1.026	0.92565598	1.050377	1.022432	0.956113	1.594896	0.779258	1.396021	0.840477	0.488699	0
16	atr-obj-pred-co-coord-root	0.852	1.01311953	0.836223	0.717229	0.846395	1.259968	1.289317	0.866496	0.890906	0.529424	0
17	auxp-pred-root	0.755	0.65597668	0.856618	0.915611	0.736677	0.46252	0.637574	0.818357	1.092621	1.552637	


styloTest1
Cluster Analysis
Aesch_PromV
Aesch_Eumen
Aesch_Septem
Aesch_Cho
Aesch_Ag
Soph_Ajax
Aesch_Supp
Aesch_Pers
Soph_OT
Soph_Elec
Soph_Antig
Soph_Trach
HomerJr_Odyssey
Homer_Iliad
Hesiod_Works
Hesiod_Theog
Lysias_Orat14
Lysias_Orat1
Plato_Euth
Polybius_Book1
Plut_Lyc
Thuc_Book1
Hdt_Book1
DioSic_Book11
Athenaeus_Book13
Athenaeus_Book12
3
2
1
0
500 MFW Culled @ 20%
Classic Delta distance

	A	B	C	D	E	F	G	H	I	J	K	
1		Mean Freq	Thuc_Book1	Aesch_Ag	Aesch_Ch	Aesch_Eu	Aesch_Per	Aesch_Prc	Aesch_Sep	Aesch_Su	Athenaeu	A
2	auxk-root	6.557	3.68075802	8.229655	8.728826	8.197492	7.6874	8.330972	8.327985	8.808203	4.953166	5
3	pred-root	3.835	2.49271137	4.721599	5.325805	4.937304	3.349282	5.1431	4.605263	5.58077	4.357565	5
4	auxy-pred-root	2.923	2.74052478	2.896186	3.525103	3.197492	1.913876	3.060357	2.567394	3.630862	3.649969	3
5	pred-co-coord-root	2.805	2.64577259	2.467877	2.975736	3.181818	3.748006	3.400397	3.353659	2.58867	1.379556	1
6	obj-pred-root	2.269	1.21720117	2.967571	3.326721	3.401254	1.897927	4.122981	2.743902	3.3451	2.010792	2
7	coord-root	2.062	1.37026239	2.304711	2.212727	2.257053	3.301435	2.054406	2.487163	2.067574	0.712686	0
8	adv-pred-root	2.058	1.44314869	2.447481	2.45689	2.507837	1.770335	2.139416	2.134146	2.639099	1.939524	2
9	sbj-pred-root	1.626	1.34110787	1.957985	1.846483	1.598746	1.339713	1.68603	1.909499	2.168432	2.48931	3
10	obj-pred-co-coord-root	1.584	1.31924198	1.550071	1.739661	1.61442	2.15311	1.827713	1.508344	1.412002	0.671961	0
11	adv-pred-co-coord-root	1.382	1.31924198	1.356312	1.297116	1.285266	1.834131	1.303485	1.428113	1.008573	0.473427	0
12	atr-obj-pred-root	1.120	0.45918367	1.65205	1.846483	1.630094	1.052632	2.238595	1.460205	2.23567	0.962126	1
13	auxy-coord-root	1.048	1.8877551	0.448705	0.595147	0.768025	1.212121	0.892604	0.882542	0.537906	0.743229	0
14	atr-sbj-pred-root	1.033	1.0058309	1.233938	1.068213	0.705329	1.307815	0.920941	1.010911	1.512859	2.199145	2
15	sbj-pred-co-coord-root	1.026	0.92565598	1.050377	1.022432	0.956113	1.594896	0.779258	1.396021	0.840477	0.488699	0
16	atr-obj-pred-co-coord-root	0.852	1.01311953	0.836223	0.717229	0.846395	1.259968	1.289317	0.866496	0.890906	0.529424	0
17	auxp-pred-root	0.755	0.65597668	0.856618	0.915611	0.736677	0.46252	0.637574	0.818357	1.092621	1.552637	


styloTest1
Cluster Analysis
HomerJr_Odyssey
Homer_Iliad
Hesiod_Works
Hesiod_Theog
Aesch_Pers
Aesch_Supp
Aesch_PromV
Aesch_Eumen
Aesch_Cho
Aesch_Ag
Aesch_Sept
Soph_Ajax
Soph_Trach
Soph_Antig
Soph_OT
Soph_Elec
Lysias_Orat14
Lysias_Orat1
Plato_Euth
Thuc_Book1
Hdt_Book1
Polybius_Book1
DioSic_Book11
Athenaeus_Book13
Athenaeus_Book12
Plut_Lyc
3
2
1
0
200 MFW Culled @ 30%
Classic Delta distance Started at 15


styloTest1
Cluster Analysis
HomerJr_Odyssey
Homer_Iliad
Hesiod_Works
Hesiod_Theog
Aesch_Pers
Aesch_Supp
Aesch_PromV
Aesch_Eumen
Aesch_Cho
Aesch_Ag
Aesch_Sept
Soph_Ajax
Soph_Trach
Soph_Antig
Soph_OT
Soph_Elec
Lysias_Orat14
Lysias_Orat1
Plato_Euth
Thuc_Book1
Hdt_Book1
Polybius_Book1
DioSic_Book11
Athenaeus_Book13
Athenaeus_Book12
Plut_Lyc
3
2
1
0
200 MFW Culled @ 30%
Classic Delta distance Started at 15

Burrows Delta

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|$$

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

	A	B	C	D	E	F	G
1		Mean Freq	Thuc_Book1	A – μ / σ	Aesch_Ag	B – μ / σ	Δ
2	auxk-root	6.557	3.681	-1.570	8.230	0.913	2.483
3	pred-root	3.835	2.493	-1.040	4.722	0.687	1.727
4	auxy-pred-root	2.923	2.741	-0.244	2.896	-0.036	0.208
5	pred-co-coord-root	2.805	2.646	-0.121	2.468	-0.257	0.136
6	obj-pred-root	2.269	1.217	-1.222	2.968	0.812	2.034
7	coord-root	2.062	1.370	-0.595	2.305	0.208	0.803
8	adv-pred-root	2.058	1.443	-1.151	2.447	0.730	1.881
9	sbj-pred-root	1.626	1.341	-0.549	1.958	0.638	1.187
10	obj-pred-co-coord-root	1.584	1.319	-0.315	1.550	-0.041	0.274
11	adv-pred-co-coord-root	1.382	1.319	-0.070	1.356	-0.029	0.042
12	atr-obj-pred-root	1.120	0.459	-1.269	1.652	1.021	2.290
13	auxy-coord-root	1.048	1.888	1.744	0.449	-1.244	2.988
14	atr-sbj-pred-root	1.033	1.006	-0.052	1.234	0.393	0.446
15	sbj-pred-co-coord-root	1.026	0.926	-0.156	1.050	0.038	0.194
16	atr-obj-pred-co-coord-root	0.852	1.013	0.494	0.836	-0.050	0.543
17	auxp-pred-root	0.755	0.656	-0.294	0.857	0.303	0.596
18							
19	Burrows Δ (AB)						1.115

1		Aesch_Ag	Aesch_Cho	Aesch_Eumen
2	Aesch_Ag	0	0.71543335	0.763167063
3	Aesch_Cho	0.71543335	0	0.74264963
4	Aesch_Eumen	0.76316706	0.74264963	0
5	Aesch_Pers	0.94373888	1.04667227	0.92321543
6	Aesch_PromV	0.87854255	0.8363469	0.782961412
7	Aesch_Septem	0.77628055	0.88979861	0.774476877
8	Aesch_Supp	0.71903384	0.80702961	0.844401159
9	Athenaeus_Book12	0.78709041	1.05752813	0.99988182
10	Athenaeus_Book13	0.73990822	0.95643313	0.916504514
11	DioSic_Book11	1.04109611	1.25095598	1.164238931
12	Hdt_Book1	0.81175972	0.97358297	0.934143579
13	Hesiod_Theog	1.12645659	1.21475149	1.060578625
14	Hesiod_Works	1.22172962	1.30519872	1.105966348
15	Homer_Iliad	1.80017686	1.75842255	1.65654519
16	HomerJr_Odyssey	1.63761741	1.61084298	1.474928814
17	Lysias_Orat1	0.98393602	1.17724583	1.123183176
18	Lysias_Orat14	0.96168621	1.1402985	1.063068141
19	Plato_Euth	0.90944349	1.05783982	0.929843668
20	Plut_Lyc	0.77060676	0.95247095	0.894406667
21	Polybius_Book1	0.90628942	1.11990842	1.040748688
22	Soph_Ajax	0.65576243	0.83199545	0.749789301
23	Soph_Antig	0.65029711	0.79423157	0.723471679
24	Soph_Elec	0.65060508	0.74893157	0.748172668
25	Soph_OT	0.64515663	0.81648197	0.775861402
26	Soph_Trach	0.67815046	0.82706971	0.828417623
27	Thuc_Book1	0.85059646	1.06735586	0.902156032

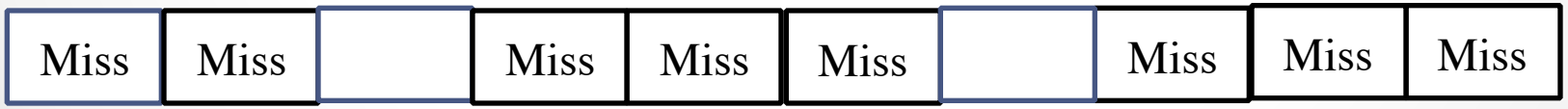
styloOppose
Craig's Zeta
Score
-1.0
-0.5
0.0
0.5
1.0
Avoided
Preferred
0
10
20
30
40
Rank of the item
adv-pnom-pred-root
adv-sbj-pred-root
atr-sbj-adv-pred-root
auxx-pred-root
-adv-co-coord-pred-root
sbj-sbj-pred-root
tr-adv-auxp-adv-pred-root
obj-sbj-pred-root
obj-obj-adv-pred-root
auxp-sbj-pred-root
atr-obj-sbj-pred-root
uxx-adv-co-coord-pred-root
auxc-adv-pred-root
coord-coord-pred-root
auxx-obj-co-coord-pred-root
atr-adv-adv-pred-root
atr-sbj-coord-pred-root
xx-pred-co-coord-coord-root
auxp-obj-adv-pred-root
atr-atr-obj-adv-pred-root
auxx-root
atr-obj-adv-coord-root
atr-obj-adv-co-coord-pred-root
v-auxp-adv-co-coord-pred-root
atr-auxp-sbj-pred-root
sbj-adv-pred-root
atr-auxp-obj-pred-root
undefined-root
sbj-ap-exd-a
atr-sbj-ap-exd-
auxy-adv-auxc-
adv-exd-adv-pred
sbj-exd-pred-co-coord-ro
obj-ap-exd-apos-pred-co
adv-co-coord-auxc-pred-
auxy-exd-pred-co-coord-
adv-exd-pred-co-coord-ro
auxx-coord-auxc-pred-co
obj-adv-co-coord-auxc-pr
auxy-adv-auxc-coord-root
coord-auxc-pred-co-coord
atv-obj-pred-co-coord-root
atr-sbj-exd-pred-co-coord-
sbj-ap-exd-apos-pred-root
auxz-exd-pred-co-coord-ro
auxz-adv-auxc-pred-co-co
auxz-sbj-pred-co-coord-roo
adv-adv-auxc-pred-co-coor
atv-sbj-pred-co-coord-root
pnom-exd-pred-co-coord-ro
atr-sbj-ap-exd-apos-pred-ro
sbj-atr-obj-pred-co-coord-ro
auxx-atr-sbj-pred-co-coord-r
auxy-adv-pred-co-coord-roo
obj-exd-pred-co-coord-root
auxx-atr-obj-pred-co-coord-r
atr-obj-adv-auxc-pred-co-coo
obj-co-coord-auxp-pred-co-co
adv-adv-co-coord-auxc-pred-
atr-coord-pred-co-coord-root
exd-coord-root
auxy-exd-pred-root
sbj-atr-sbj-pred-co-coord-root
atvv-pred-co-coord-root
auxx-coord-auxc-pred-root
atr-obj-ap-exd-apos-pred-co-c
sbj-exd-pred-root
coord-auxc-coord-root
coord-apos-pred-co-coord-root
obj-atr-sbj-pred-co-coord-root
atr-co-coord-obj-pred-co-coord-
pnom-exd-pred-root

Craig's Zeta

- Divide corpus 1 into segments of equal size (size = n)
- Segments with at least 1 example of given feature are hits.
- Each hit is worth 1 point.

- Hits/segments = preferred feature score

- Divide corpus 2 into segments of size n.
- Segments with no examples of feature are misses.
- Each miss is worth -1 point.



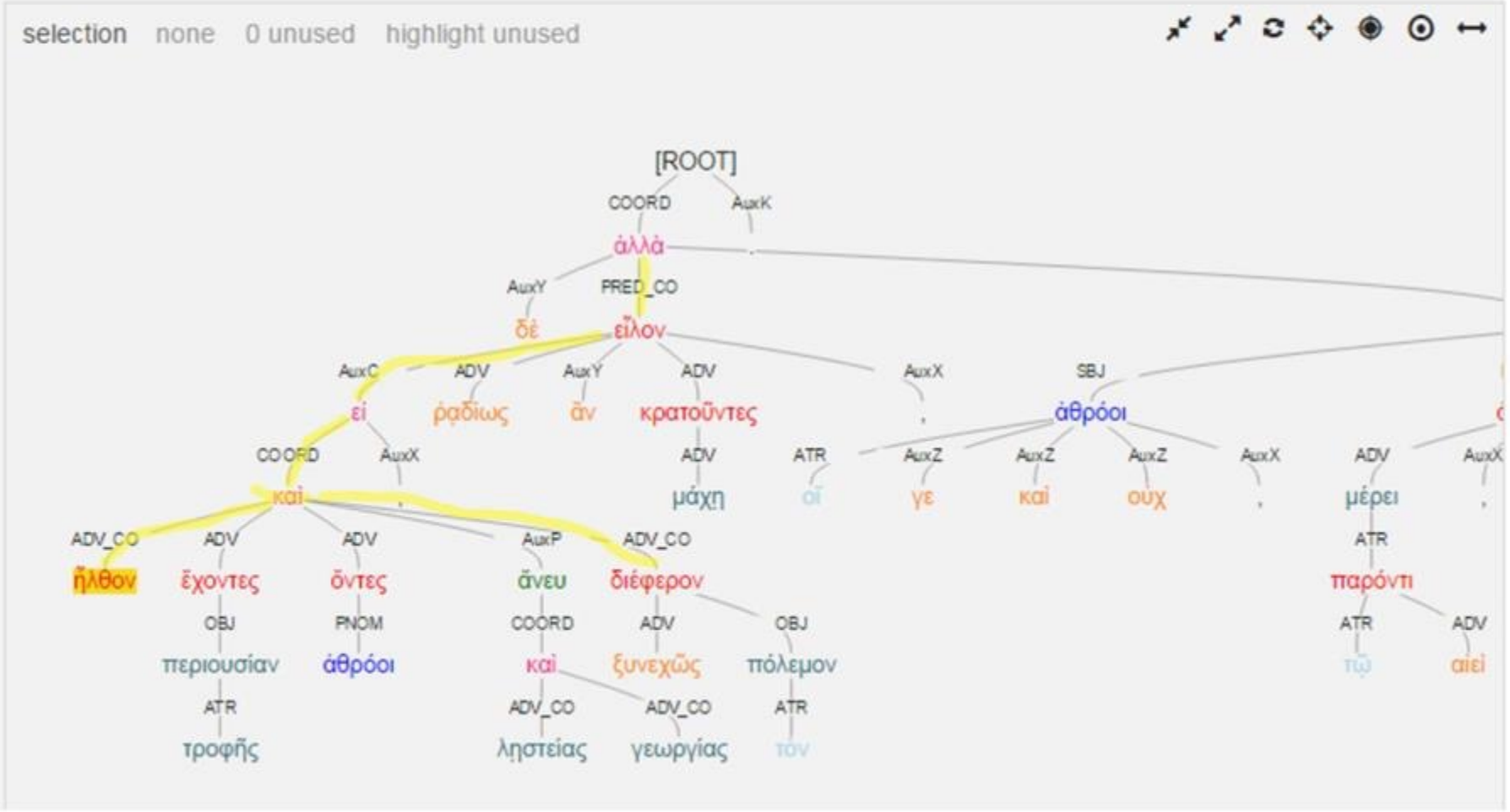
- Misses/segments = avoided feature score

styloOppose
Craig's Zeta
Score
-1.0
-0.5
0.0
0.5
1.0
Avoided
Preferred
0
10
20
30
40
Rank of the item
adv-pnom-pred-root
adv-sbj-pred-root
atr-sbj-adv-pred-root
auxx-pred-root
-adv-co-coord-pred-root
sbj-sbj-pred-root
tr-adv-auxp-adv-pred-root
obj-sbj-pred-root
obj-obj-adv-pred-root
auxp-sbj-pred-root
atr-obj-sbj-pred-root
uxx-adv-co-coord-pred-root
auxc-adv-pred-root
coord-coord-pred-root
auxx-obj-co-coord-pred-root
atr-adv-adv-pred-root
atr-sbj-coord-pred-root
xx-pred-co-coord-coord-root
auxp-obj-adv-pred-root
atr-atr-obj-adv-pred-root
auxx-root
atr-obj-adv-coord-root
atr-obj-adv-co-coord-pred-root
v-auxp-adv-co-coord-pred-root
atr-auxp-sbj-pred-root
sbj-adv-pred-root
atr-auxp-obj-pred-root
undefined-root
sbj-ap-exd-a
atr-sbj-ap-exd-
auxy-adv-auxc-
adv-exd-adv-pred
sbj-exd-pred-co-coord-ro
obj-ap-exd-apos-pred-co
adv-co-coord-auxc-pred-
auxy-exd-pred-co-coord-
adv-exd-pred-co-coord-ro
auxx-coord-auxc-pred-co
obj-adv-co-coord-auxc-pr
auxy-adv-auxc-coord-root
coord-auxc-pred-co-coord
atv-obj-pred-co-coord-root
atr-sbj-exd-pred-co-coord-
sbj-ap-exd-apos-pred-root
auxz-exd-pred-co-coord-ro
auxz-adv-auxc-pred-co-co
auxz-sbj-pred-co-coord-roo
adv-adv-auxc-pred-co-coor
atv-sbj-pred-co-coord-root
pnom-exd-pred-co-coord-ro
atr-sbj-ap-exd-apos-pred-ro
sbj-atr-obj-pred-co-coord-ro
auxx-atr-sbj-pred-co-coord-r
auxy-adv-pred-co-coord-roo
obj-exd-pred-co-coord-root
auxx-atr-obj-pred-co-coord-r
atr-obj-adv-auxc-pred-co-coo
obj-co-coord-auxp-pred-co-co
adv-adv-co-coord-auxc-pred-
atr-coord-pred-co-coord-root
exd-coord-root
auxy-exd-pred-root
sbj-atr-sbj-pred-co-coord-root
atvv-pred-co-coord-root
auxx-coord-auxc-pred-root
atr-obj-ap-exd-apos-pred-co-c
sbj-exd-pred-root
coord-auxc-coord-root
coord-apos-pred-co-coord-root
obj-atr-sbj-pred-co-coord-root
atr-co-coord-obj-pred-co-coord-
pnom-exd-pred-root

Thucydides

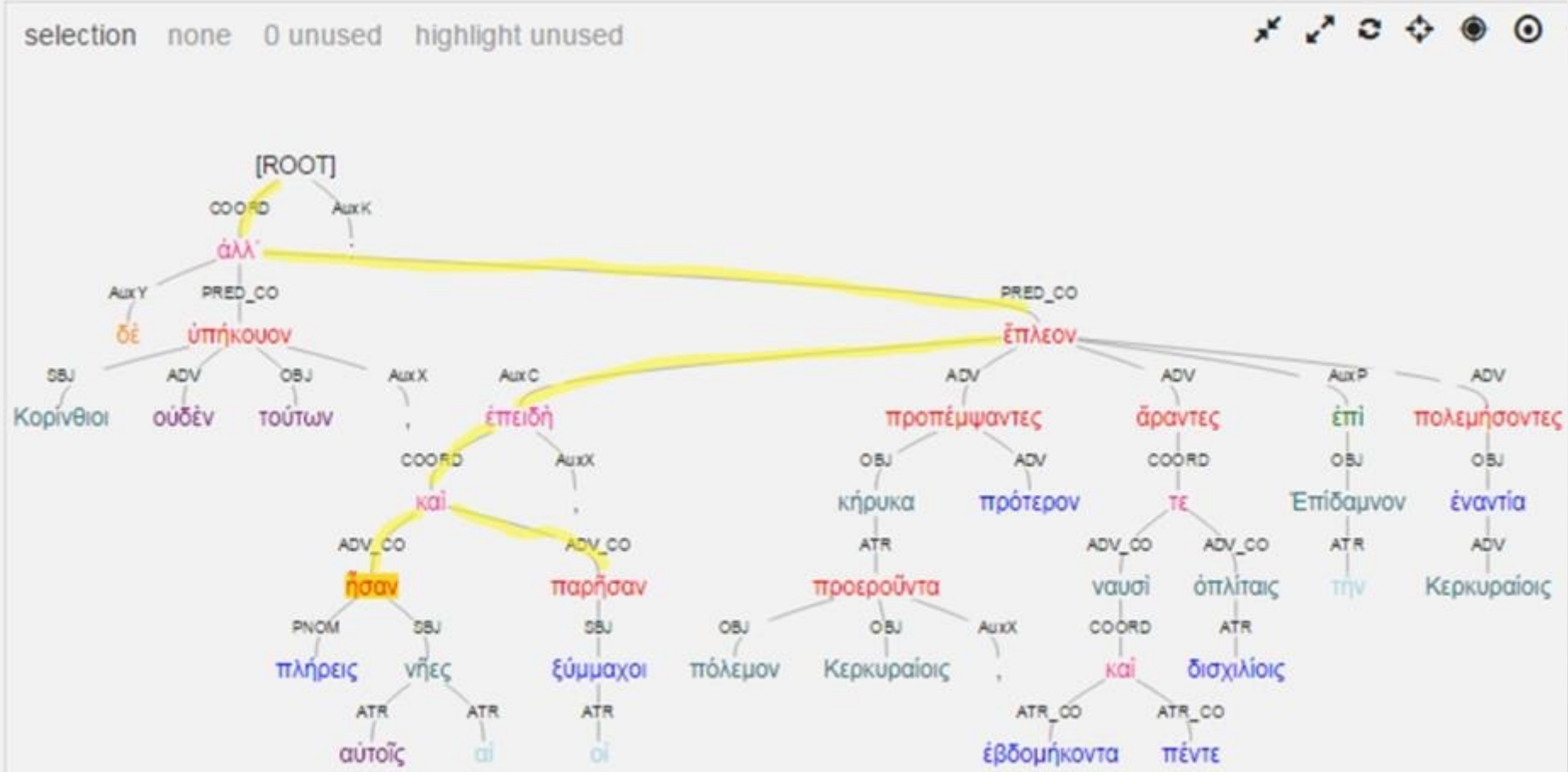
#1a Adv_co*Coord*AuxC*Pred_co*Coord*root
[Thuc. 1.11.2, sent 64, word 3]

περιουσίαν δὲ εἰ ἦλθον ἔχοντες τροφῆς καὶ ὄντες ἀθρόοι ἄνευ λῃστείας καὶ γεωργίας ξυνεχῶς τὸν πόλεμον διέφερον , ῥᾳδίως ἂν μάχῃ κρατοῦντες εἶλον , οἵ γε καὶ οὐχ ἀθρόοι , ἀλλὰ μέρει τῷ αἰεὶ παρόντι ἀντεῖχον , πολιορκίᾳ δ' ἂν προσκαθεζόμενοι ἐν ἐλάσσονί τε χρόνῳ καὶ ἀπονώτερον τὴν Τροίαν εἶλον .



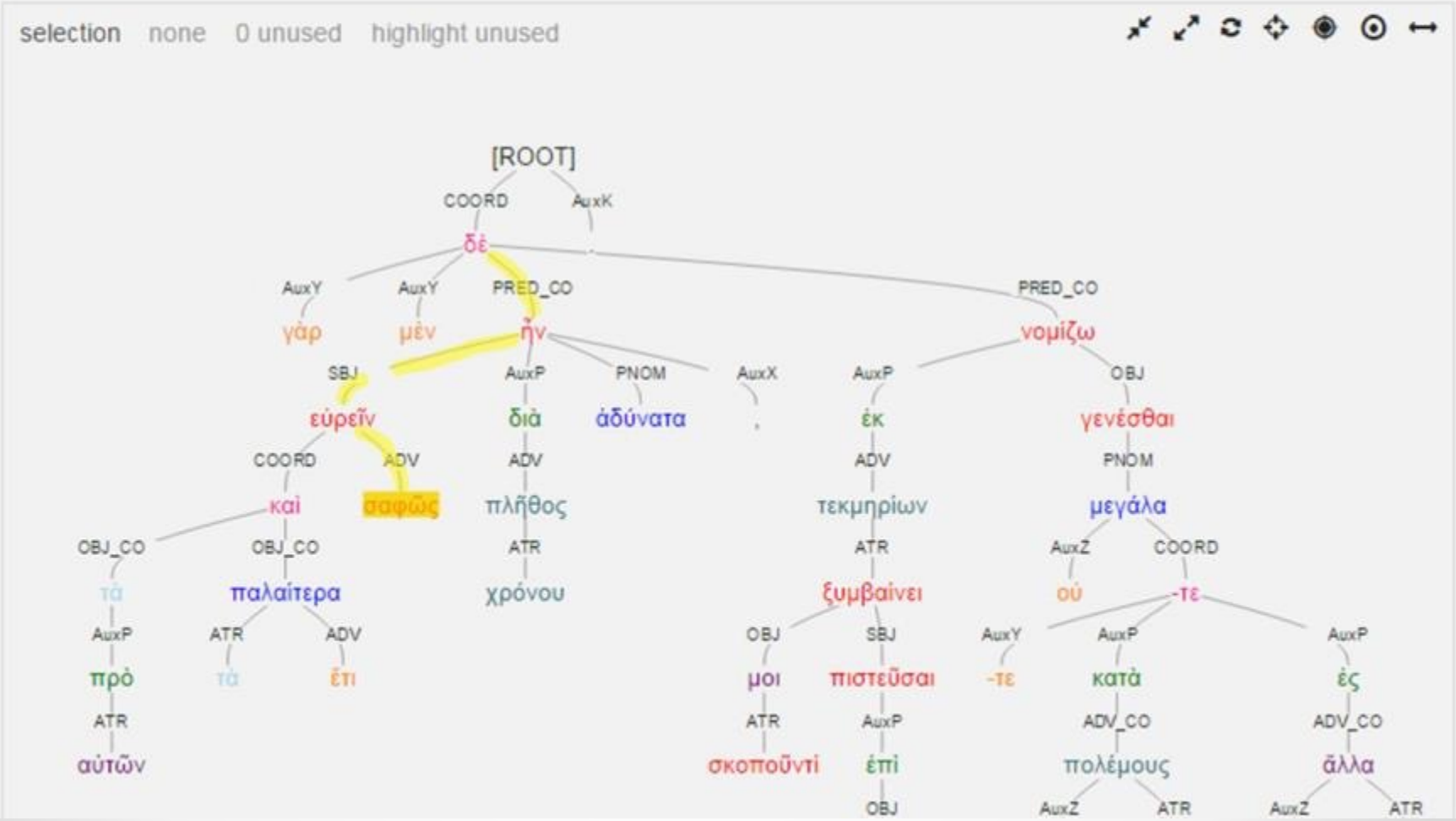
1b Adv_co*Coord*AuxC*Pred_co*Coord*root [Thuc. 1.29.1, sent 58, word 11]

Κορίνθιοι δὲ οὐδὲν τούτων ὑπήκουον , ἀλλ᾽ ἐπειδὴ πλήρεις αὐτοῖς ἦσαν αἱ νῆες καὶ οἱ ξύμμαχοι παρῆσαν , προπέμψαντες κήρυκα πρότερον πόλεμον προεροῦντα Κερκυραίοις , ἄραντες ἑβδομήκοντα ναυσὶ καὶ πέντε δισχιλίοις τε ὁπλίταις ἔπλεον ἐπὶ τὴν Ἐπίδαμνον Κερκυραίοις ἐναντία πολεμήσοντες :



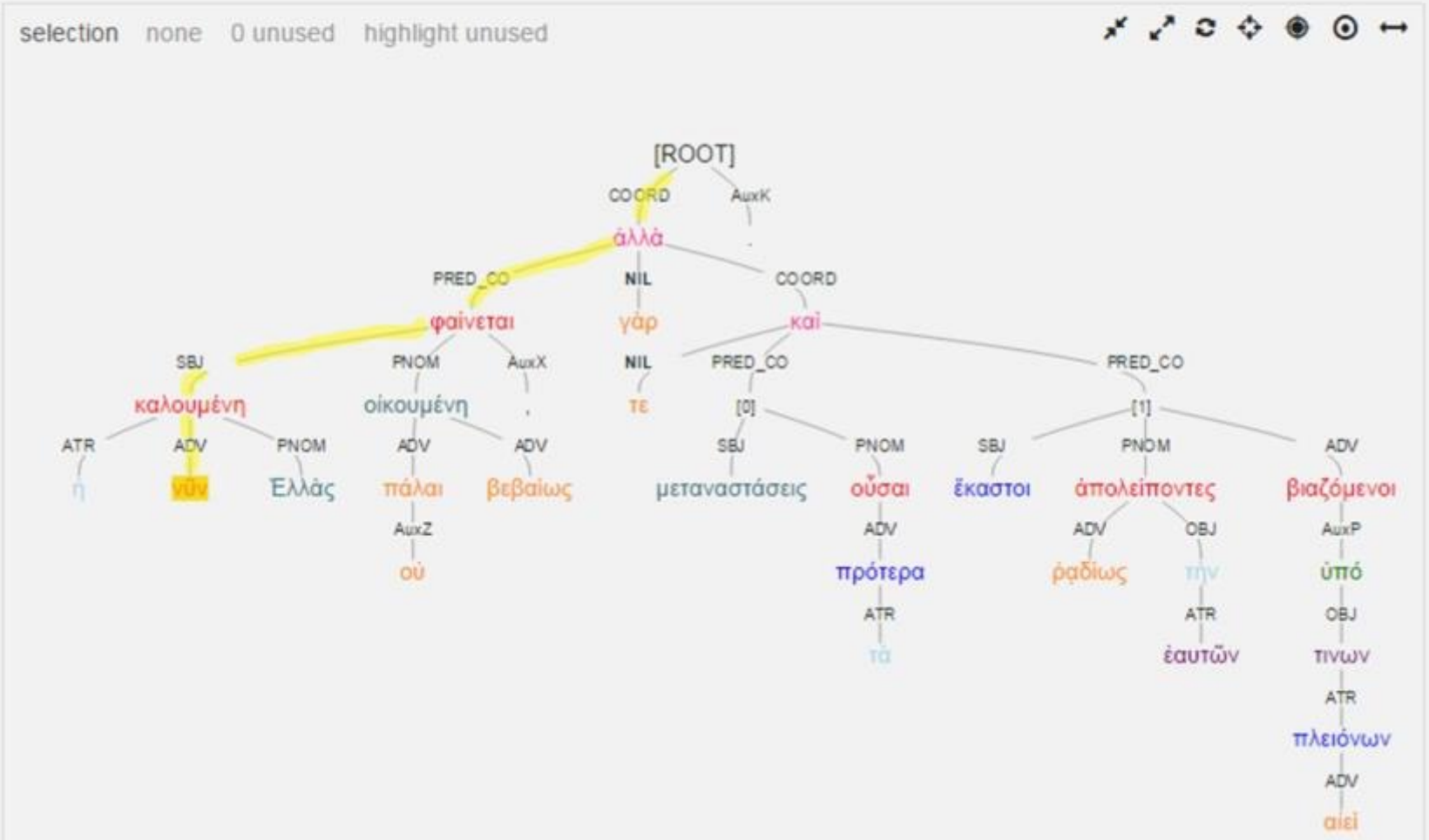
#2a Adv*Sbj*Pred_co*Coord*Root [Thuc. 1.1.3, sent 3, word 9]

τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν , ἐκ δὲ τεκμηρίων ὧν ἐπὶ μακρότατον σκοποῦντί μοι πιστεῦσαι ξυμβαίνει οὐ μεγάλα νομίζω γενέσθαι οὔ -τε κατὰ τοὺς πολέμους οὔ -τε ἐς τὰ ἄλλα .



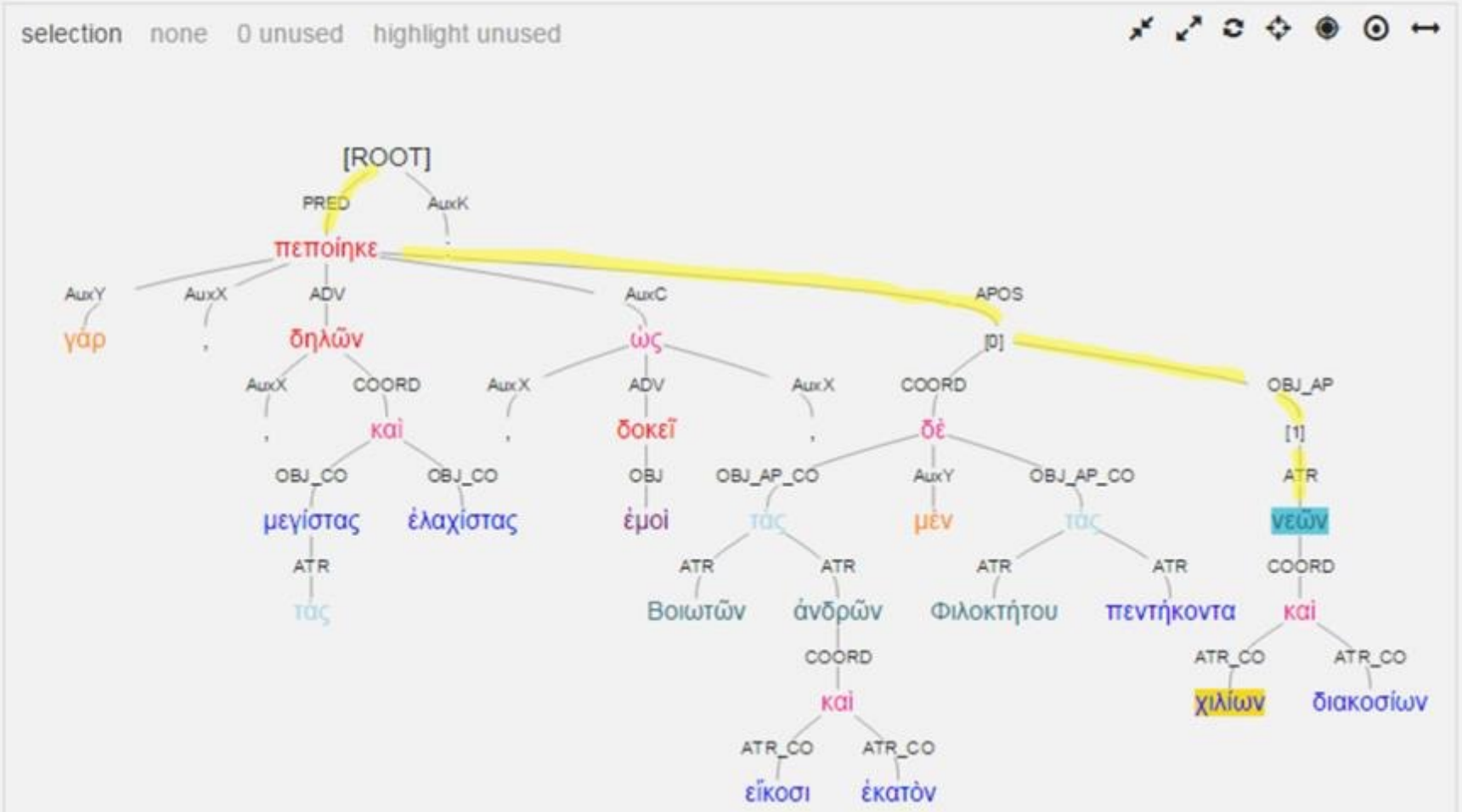
#2b Adv*Sbj*Pred_co*Coord*Root [Thuc. 1.2.1, sent 4, word 4]

φαίνεται γὰρ ἡ νῦν Ἑλλὰς καλουμένη οὐ πάλαι βεβαίως οἰκουμένη , ἀλλὰ μεταναστάσεις τε οὖσαι τὰ πρότερα καὶ ῥᾳδίως ἕκαστοι τὴν ἑαυτῶν ἀπολείποντες βιαζόμενοι ὑπό τινων αἰεὶ πλειόνων . [0] [1]



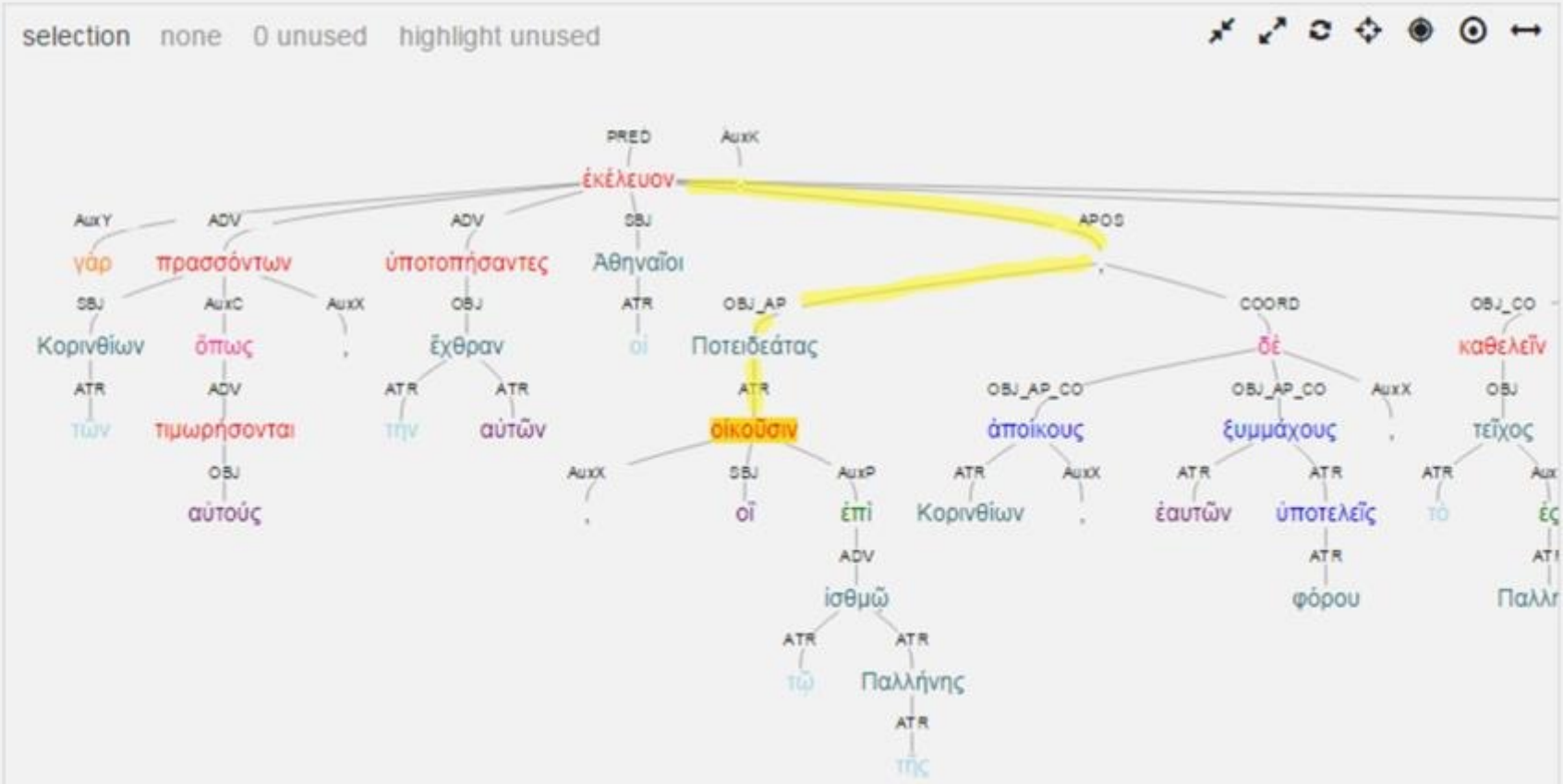
#3a Atr*Obj_ap*Apos* Pred*Root [Thuc. 1.10.4, sent 55, word 3]

πεποίηκε γὰρ χιλίων καὶ διακοσίων νεῶν τὰς μὲν Βοιωτῶν εἴκοσι καὶ ἑκατὸν ἀνδρῶν , τὰς δὲ Φιλοκτήτου πεντήκοντα , δηλῶν , ὡς ἐμοὶ δοκεῖ , τὰς μεγίστας καὶ ἐλαχίστας : [0] [1]



#3b Atr*Obj_ap*Apos* Pred*Root [Thuc. 1.56.2, sent 90, word 18

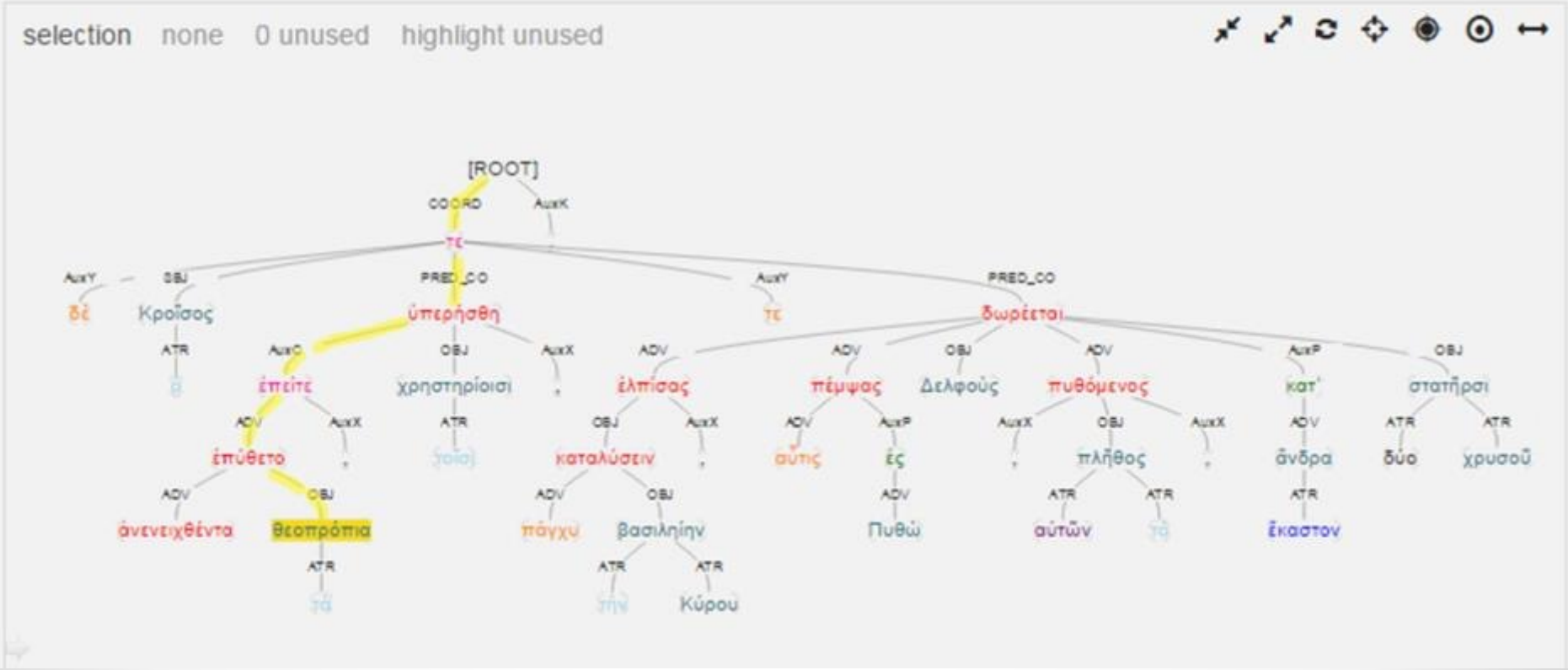
τῶν γὰρ Κορινθίων πρασσόντων ὅπως τιμωρήσονται αὐτούς , ὑποτοπήσαντες τὴν ἔχθραν αὐτῶν οἱ Ἀθηναῖοι Ποτειδεάτας , οἳ οἰκοῦσιν ἐπὶ τῷ ἰσθμῷ τῆς Παλλήνης , Κορινθίων ἀποίκους , ἑαυτῶν δὲ ξυμμάχους φόρου ὑποτελεῖς , ἐκέλευον τὸ ἐς Παλλήνην τεῖχος καθελεῖν καὶ ὁμήρους δοῦναι , τούς τε ἐπιδημιουργοὺς ἐκπέμπειν καὶ τὸ λοιπὸν μὴ δέχεσθαι οὓς κατὰ ἔτος ἕκαστον Κορίνθιοι ἔπεμπον , δείσαντες μὴ ἀποστῶσιν ὑπό τε Περδίκκου πειθόμενοι καὶ Κορινθίων , τούς τε ἄλλους τοὺς ἐπὶ Θράκης ξυναποστήσωσι ξυμμάχους .



Herodotus

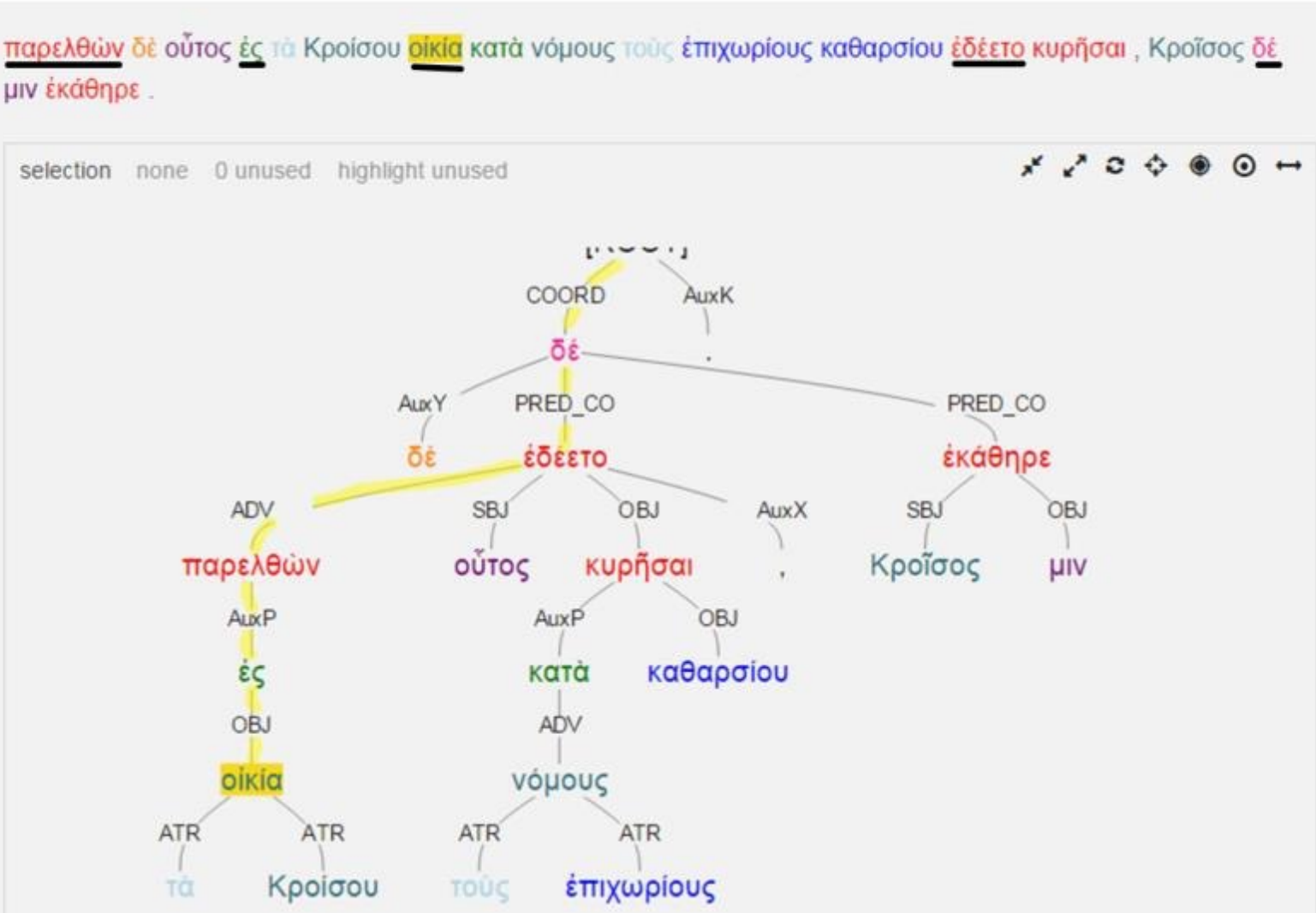
#4b obj*adv*auxc*pred_co*coord*root [Hdt. 1.54] (Sent. 66/67, word 5)

ἐπείτε δὲ ἀνενειχθέντα τὰ θεοπρόπια ἐπύθετο ὁ Κροῖσος , ὑπερήσθη τε τοῖσι χρηστηρίοισι , πάγχυ τε ἐλπίσας καταλύσειν τὴν Κύρου βασιληίην , πέμψας αὖτις ἐς Πυθὼ Δελφοὺς δωρέεται , πυθόμενος αὐτῶν τὸ πλῆθος , κατ' ἄνδρα δύο στατῆρσι ἕκαστον χρυσοῦ .



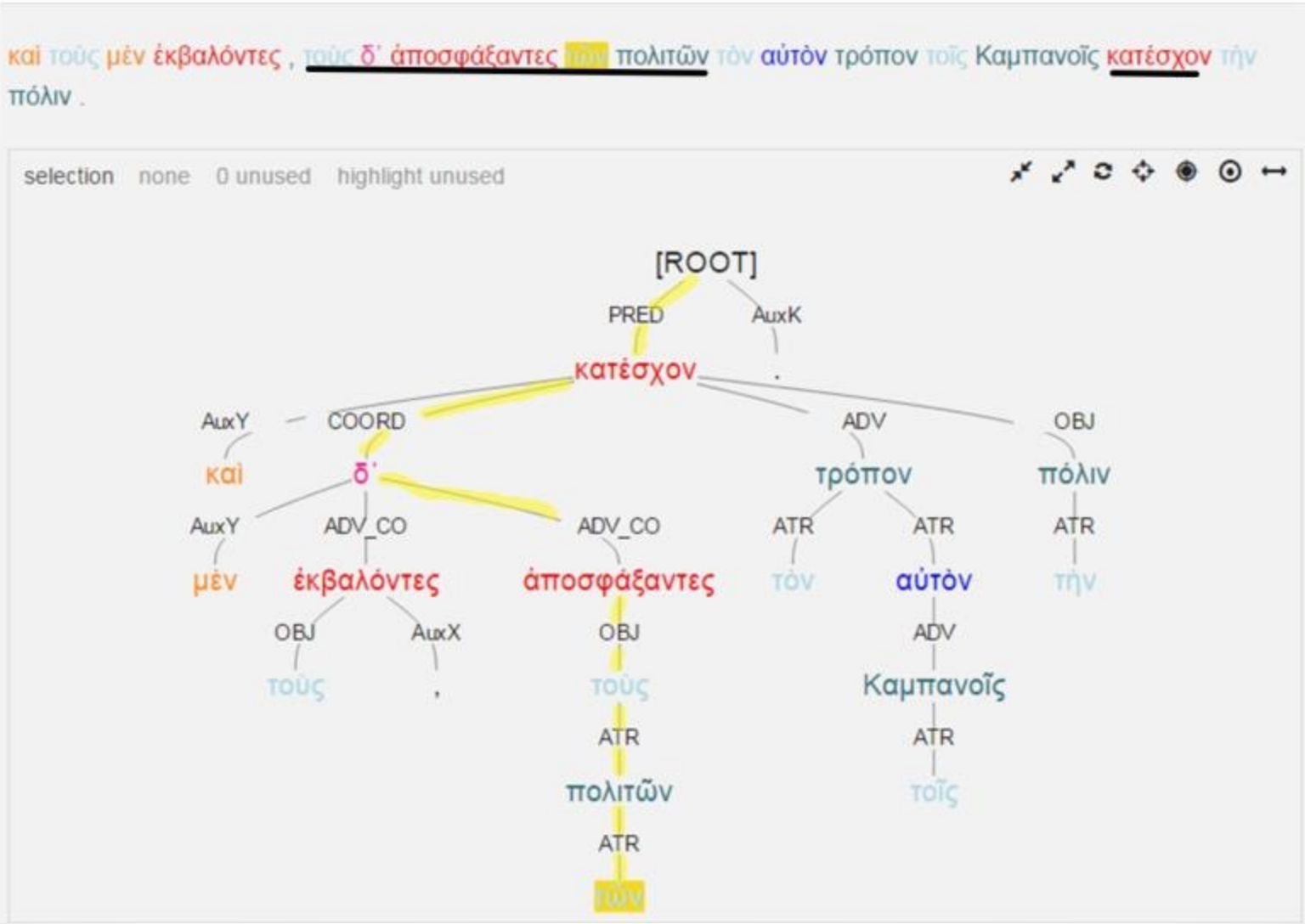
#6a Obj*auxp*adv*pred_co*coord*root [Hdt. 1.35] (Sent. 110, word 7)

παρελθὼν δὲ οὗτος ἐς τὰ Κροίσου οἰκία κατὰ νόμους τοὺς ἐπιχωρίους καθαρσίου ἐδέετο κυρῆσαι , Κροῖσος δέ μιν ἐκάθηρε .



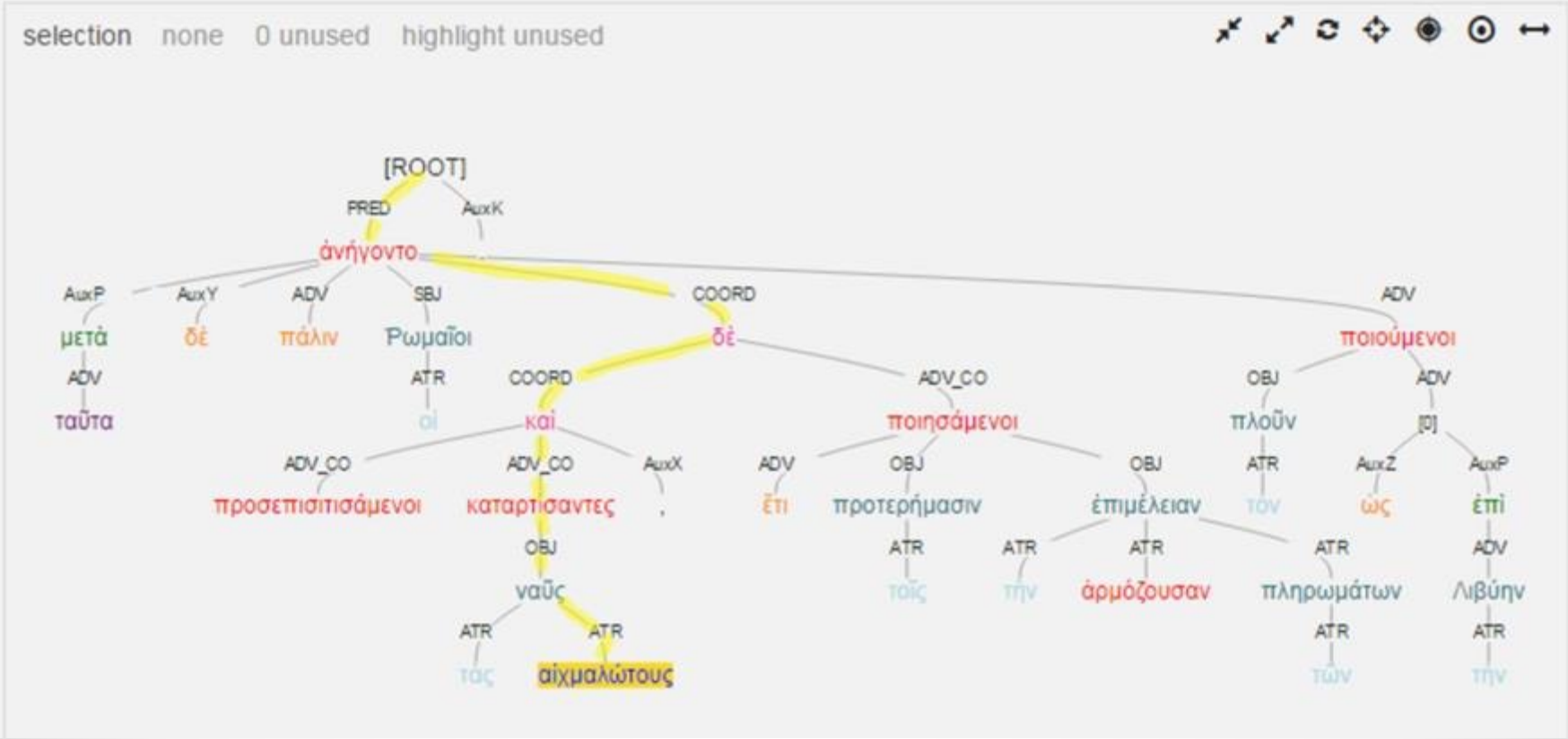
Polybius

#7a Atr*atr*obj*adv_co*coord*pred*root [Polyb. 1.7] (Sent. 62, word 9)



#9a atr*obj*adv_co*coord*coord*pred*root [Polyb. 1.29] (Sent. 129, word 10)

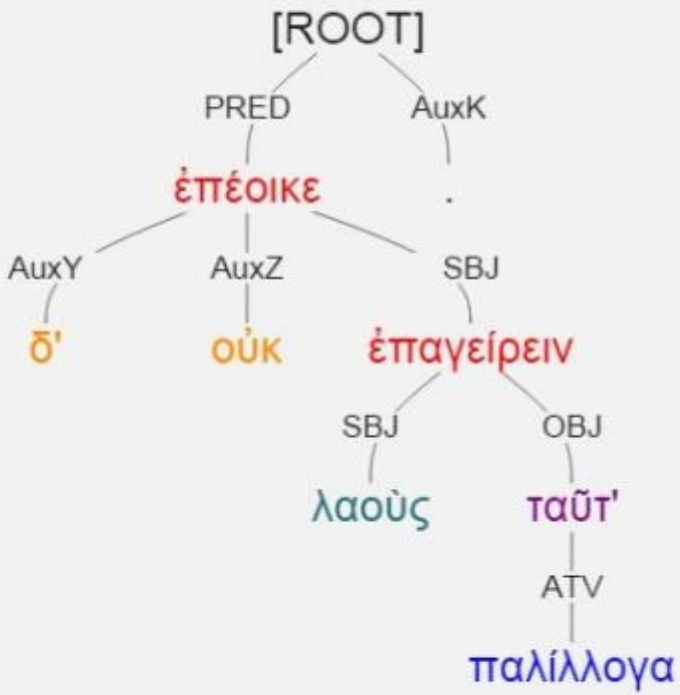
μετὰ δὲ ταῦτα πάλιν οἱ Ῥωμαῖοι προσεπισιτισάμενοι καὶ τὰς αἰχμαλώτους ναῦς καταρτίσαντες , ἔτι δὲ τὴν ἁρμόζουσαν τοῖς προτερήμασιν ἐπιμέλειαν ποιησάμενοι τῶν πληρωμάτων ἀνήγοντο ποιούμενοι τὸν πλοῦν ὡς ἐπὶ τὴν Λιβύην . [0]

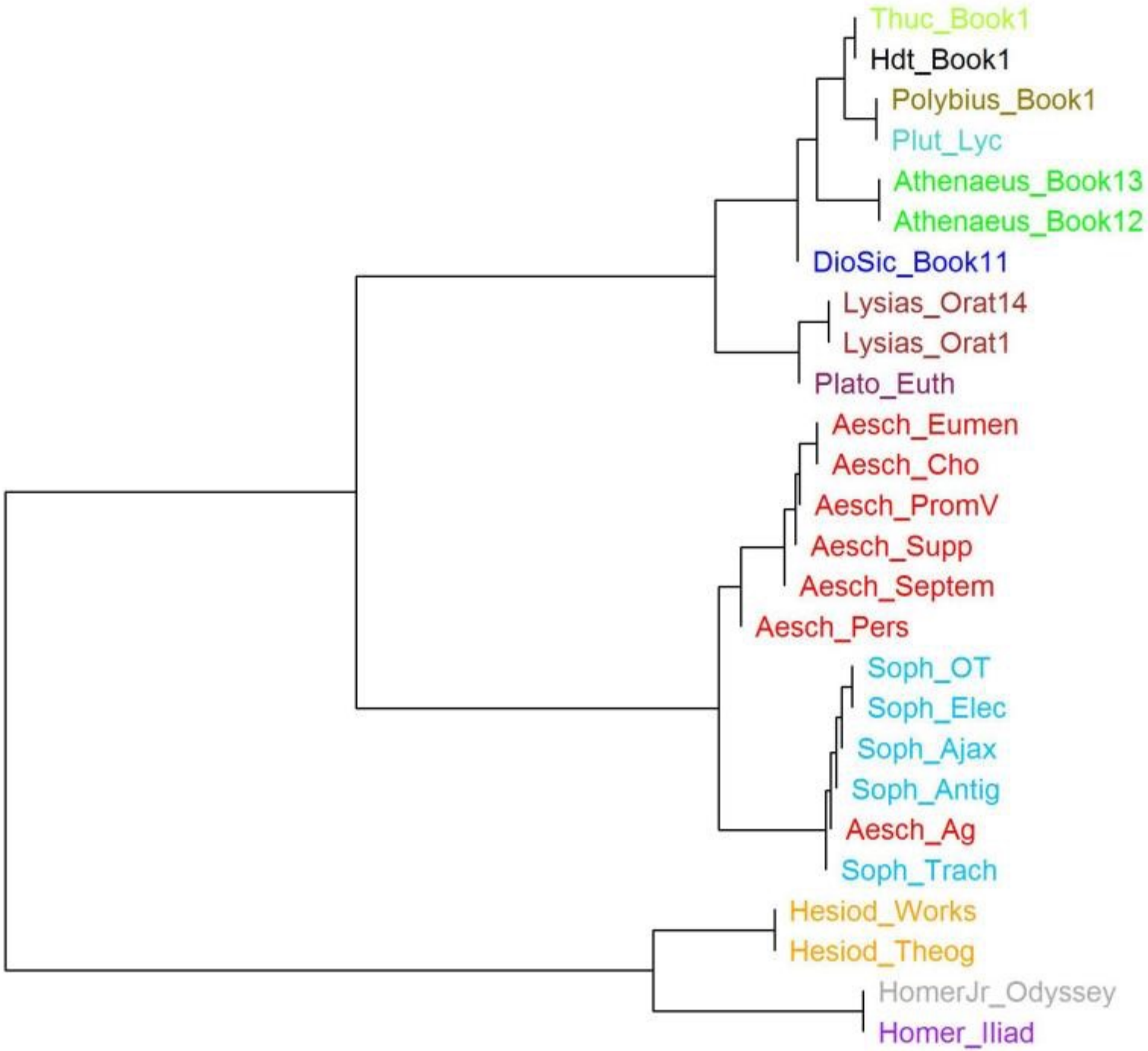


Homer

λαοὺς δ' οὐκ ἐπέοικε παλίλλογα ταῦτ' ἐπαγείρειν .

selection none 0 unused highlight unused




Thuc_Book1
Hdt_Book1
Polybius_Book1
Plut_Lyc
Athenaeus_Book13
Athenaeus_Book12
DioSic_Book11
Lysias_Orat14
Lysias_Orat1
Plato_Euth
Aesch_Eumen
Aesch_Cho
Aesch_PromV
Aesch_Supp
Aesch_Septem
Aesch_Pers
Soph_OT
Soph_Elec
Soph_Ajax
Soph_Antig
Aesch_Ag
Soph_Trach
Hesiod_Works
Hesiod_Theog
HomerJr_Odyssey
Homer_Iliad

Does size matter? Authorship attribution, short samples, big problem

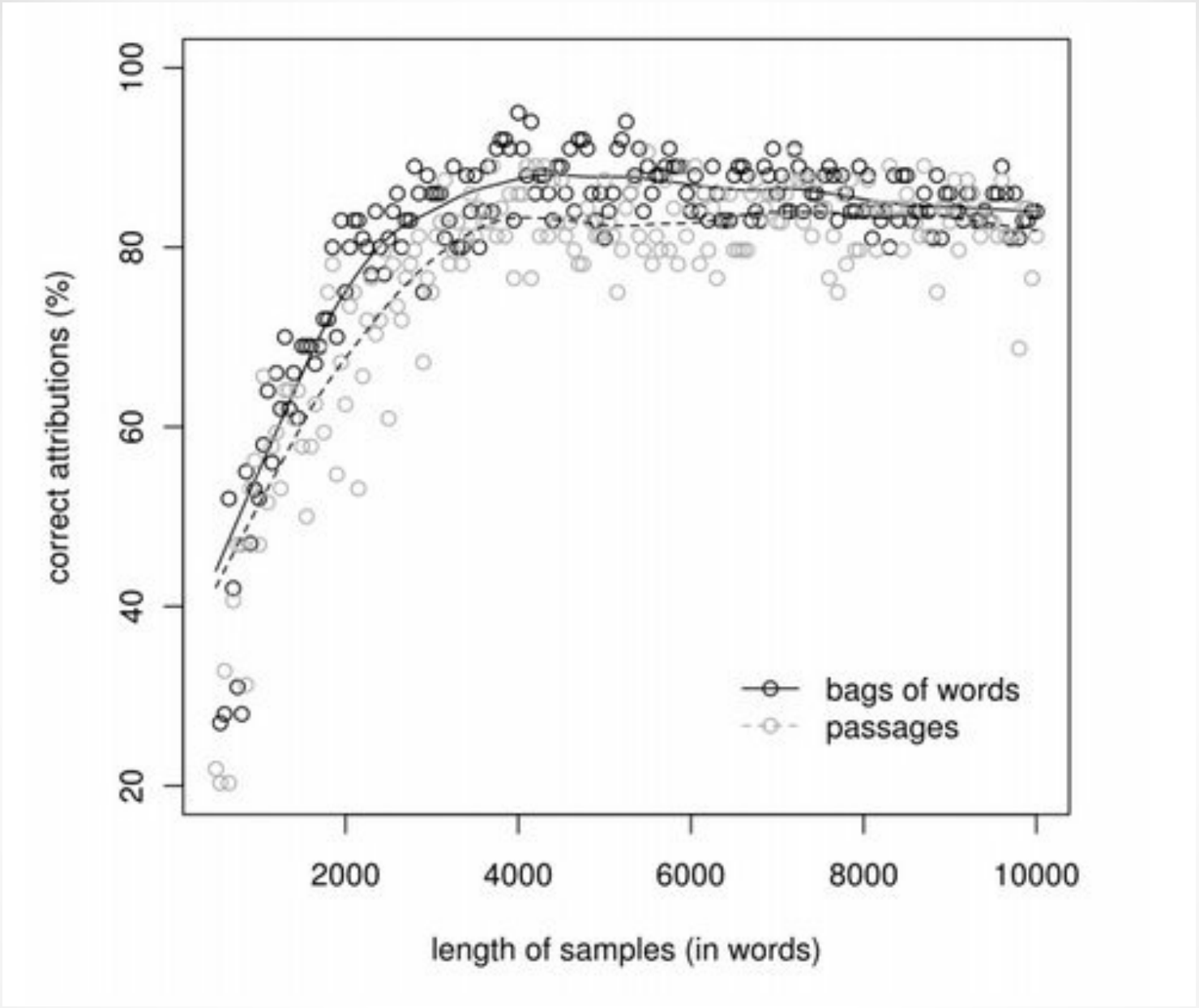
Maciej Eder

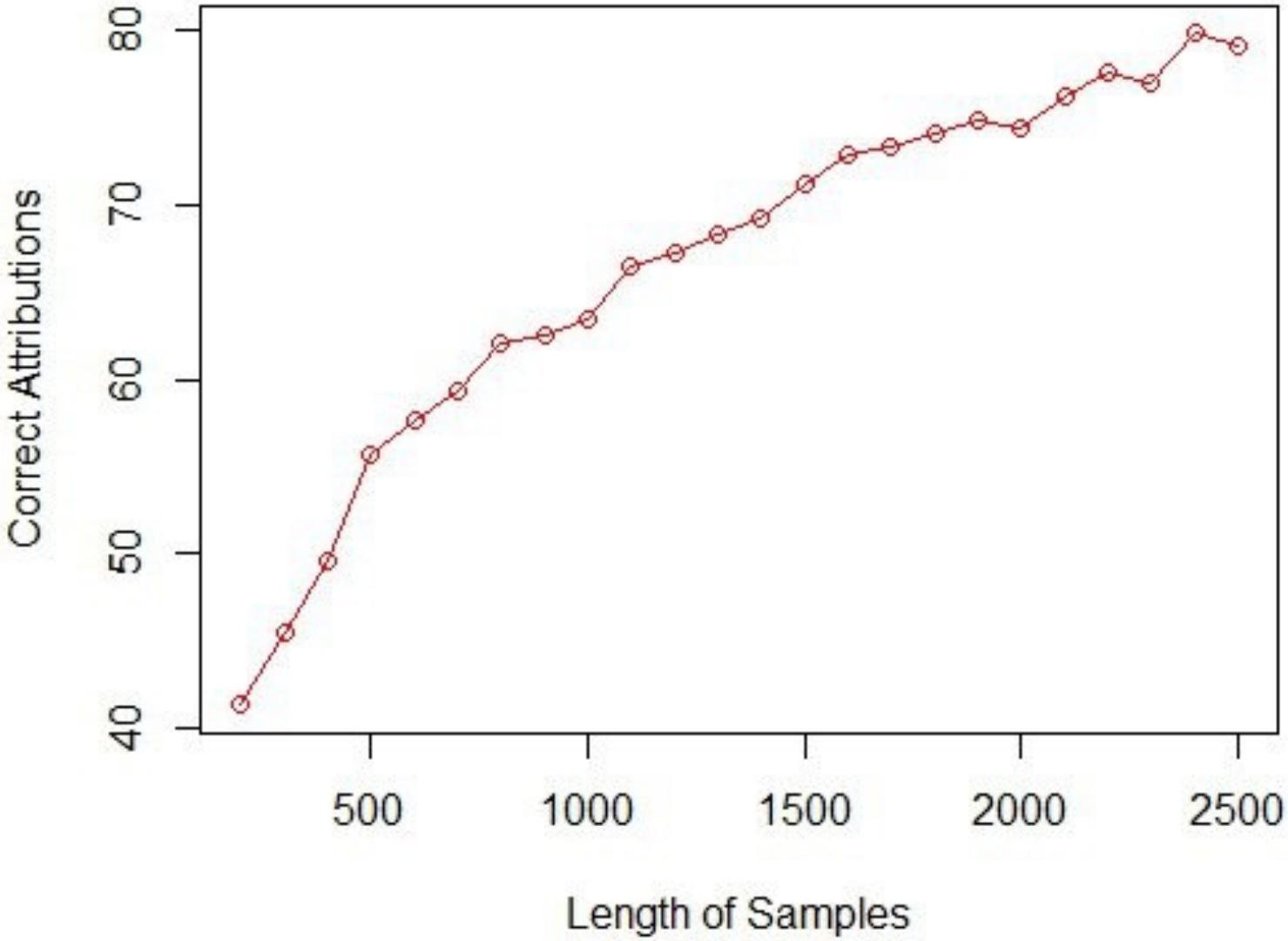
Abstract

The aim of this study is to find such a minimal size of text samples for authorship attribution that would provide stable results independent of random noise. A few controlled tests for different sample lengths, languages and genres are discussed and compared. Depending on the corpus used, the minimal sample length varied from 2,500 words (Latin prose) to 5,000 or so words (in most cases, including English, German, Polish and Hungarian novels). Another observation is connected with the method of sampling: contrary to common sense, randomly excerpted 'bags of words' turned to be much more effective than the classical solution, i.e. using original sequences of words ('passages') of desired size. Although the tests have been performed using the Delta method (Burrows, 2002) applied to the most frequent words (MFWs), some additional experiments have been conducted for SVM and k-NN applied to MFWs, character 3-grams, character 4-grams, and POS-tag 3-grams. Despite significant differences in overall attributive success between particular methods and/or style-markers, the minimal amount of textual data needed for reliable authorship attribution turned out to be method-independent.

1 Introduction

In the field of computational stylistics, and especially in authorship attribution, the reliability of short sample effect (Hoover, 2003: 439). In another instance, Rybicki discovered that his own results of remarkable similarities in the patterns of distance between idiolects in two different translations of the


correct attributions (%)
100
80
60
40
20
2000
4000
6000
8000
10000
length of samples (in words)
bags of words
passages


Correct Attributions
40
50
60
70
80
500
1000
1500
2000
2500
Length of Samples

Length	Mean Success	Min Success	Max Success	St. Dev
2500	79.074	58.583	89.908	5.9
2400	79.849	58.219	91.071	5.78
2300	76.945	61.538	89.166	5.76
2200	77.655	58.959	88.652	5.36
2100	76.261	51.194	87.919	5.68
2000	74.362	55.056	85.542	5.83
1900	74.95	58.762	86.227	5.28
1800	74.092	60.526	86.294	5.47
1700	73.337	62.211	84.803	4.76
1600	72.955	61.085	85.549	4.56
1500	71.19	57.083	81.896	4.08
1400	69.313	59.558	80	4.63
1300	68.406	55.629	78.245	4.36

Length	Mean Success	Min Success	Max Success	St. Dev
1200	67.275	54.179	77.741	4.92
1100	66.578	55.585	75.842	3.84
1000	63.5	53.535	72.651	3.66
900	62.602	52.517	70.111	3.75
800	62.106	50.863	70.425	3.56
700	59.304	49.462	65.874	3.14
600	57.736	50.584	64.083	2.82
500	55.293	48.51	60.297	2.45
400	49.649	43.726	54.22	2.14
300	45.458	41.456	50.987	1.87
200	41.326	37.618	44.614	1.51

What Next?

- Test! Test! Test!
- Cast the net as widely as possible:
  - Many flavors of sWord
    - With POS, with Dependency Distance …
    - N-grams
  - Many computational approaches

What next?

- Test! Test! Test!
- Aim directly at research question
  - Athenaeus and fragments
  - Are fragments of single author distinguishable according to transmitting source?

What's needed?

- Trees! Trees! Trees!
- Metadata
  - Digital Athenaeus
  - Digital *Fragmenta Historicorum Graecorum*
- Scalable workflow
  - Stable identification for each token

The Vision Thing

- Treebanker's Utopia
  - Real time feedback for annotators
    - Is this syntactic structure feasible?
    - Is this structure prone to inter-annotator disagreement?
- Philologist's Elysium
  - Real time feedback for close readers
  - How does this text compare to others:
    - Lexically, syntactically, semantically?
    - Pragmatically, acoustically, etc.?

- Leipzig Open Philology Project
  - Digital Athenaeus Project
- Perseus and Perseids Projects, Tufts University
  - Perseus Open Publication Series
- University of Nebraska–Lincoln
  - Dept. of History
  - Dept. of Classics and Religious Studies