

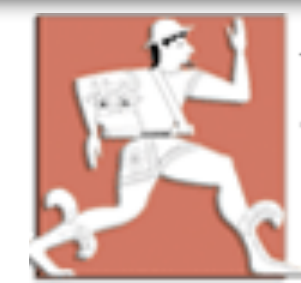
IMPLEMENTING GREEK MORPHOLOGY

Digital Humanities 2009

Helma Dik & Richard Whaling, University of Chicago

perseus.uchicago.edu

The texts in our system come from the Perseus Project. We have made them searchable by string, lemma, and morphological attributes.



PERSEUS DIGITAL LIBRARY
GREGORY R. CRANE, EDITOR-IN-CHIEF
TUFTS UNIVERSITY

```
<?xml version="1.0" encoding="utf-8"?>
<TEI.2><text><group><text n="Apol." /><body><p><milestone unit="section"
n="17a" /><milestone ed="P" unit="para" /><O(TI MEIN U(MEI=S,W)=A)/
NDRES *)AQHNAI=OI,PEPO/NQATE U(POI,TW=N E)MW=N KATHGO/
RWV, OUIK OI)=DA: E)GWV D' OUI)=N KAI AU)TOIS U(P' AU)TW=N O)LI/
GOU E)MAUTOU= E)PELAQO/MHN, OU(TW PIQANW=S E)/LEGON. KAI/
TOI A)LHQE/S GEW/S E)/POS E)PEI=N OUI)DEIN E)I)RH/KASIN. MA/LISTA
DEI AU)TW=N E(N E)QAU/MASA TW=N POLLW=N W=(N E)YEU/SANTO,
TOU=TO E)N W(= E)/LEGON V(S XRH=N U(MA=S EU)LABEL=SQAI MH-
U(P' E)MOU= E)CAPATHQH=TE
</p></body></text></group></text></TEI.2>
```

[17a]
ὅτι μὲν ὁμίεις, ὧ ἄνδρες Ἀθηναῖοι, πεπόνθατε ὑπὸ τῶν
ἐμῶν κατηγορῶν, οὐκ οἶδα· ἐγὼ δ' οὐκ αὐτὸς ὑπ'
αὐτῶν ὀλίγου ἑμαυτοῦ ἐπελαθόμην, οὕτω πιθανῶς ἔλεγον.
καίτοι ἀληθές γε ὡς ἔπος εἰπὲν οὐδὲν εἰρήκασιν.

ἐμῶς						vomit, throw up
(Show lexicon entry in LSJ Middle Liddell Autenrieth) (search)						
ἐμῶν	part sg fut part act	masc nom			no user votes	4.3% [vote]
ἐμῶν	part sg pres part act	masc nom			no user votes	12.7% [vote]
Word Frequency Statistics (more statistics)						
Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name	
40,841	146	35.748	0	0	Plato, <i>Apology</i>	
ἐμός						mine
(Show lexicon entry in LSJ Middle Liddell Slater Autenrieth) (search)						
ἐμῶν	adj pl fem gen			no user votes	25.3%	[vote]
ἐμῶν †	adj pl masc gen			no user votes	30.1%	[vote]
ἐμῶν	adj pl neut gen			no user votes	27.6%	[vote]

u(po)
<NL>N u(po,u(po/ indeclform prep</NL>
tw=n
<NL>N of fem gen pl indeclform article</NL>
<NL>N of masc/neut gen pl indeclform article</NL>
e)mw=n
<NL>P e)me/w fut part act masc nom sg attic epic doric contr ew_fut_e_stem</NL>
<NL>P e)me/w pres part act masc nom sg attic epic doric contr ew_pr_e_stem</NL>
<NL>N e)mo/s fem gen pl os_h_on</NL>
<NL>N e)mo/s masc/neut gen pl os_h_on</NL>
kathgo/rwn
<NL>N kath/goros masc/fem/neut gen pl os_on</NL>
ou)k
<NL>N ou) proclitic indeclformadverb</NL>
oi)=da
<NL>V oi)=da perf ind act 1st sg ath_primary</NL>



IL.1.1 μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος¹¹
IL.1.1 SING, goddess, the anger of Peleus' son Achilles

IL.1.2 οὐλομένην, ἣ μυρ' Ἀχαιοῖς ἔα' ἔθηκε,²¹
IL.1.2 and its devastation, which put pains thousandfold upon the Achaians,

IL.1.3 πολλὰς δ' ἰφθίμους ψυχὰς Ἄϊδι προΐαφεν⁴¹
IL.1.3 hurled in their multitudes to the house of Hades strong souls

IL.1.4 ἥρώων, αὐτοὺς δὲ ἑλώρια τεύχε κύνεσσιν
IL.1.4 of heroes, but gave their bodies to be the delicate feasting

IL.1.5 οἰωνοῖσί τε ⁵πάσι, ⁶¹Διὸς⁵¹ δ' ἔτελεετο⁷¹ βουλῇ,⁶¹
IL.1.5 of dogs, of all birds, and the will of Zeus was accomplished

IL.1.6 ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε
IL.1.6 since that time when first there stood in division of conflict

IL.1.7 Ἄτρεΐδης τε ⁸ἄναξ ἀνδρῶν⁹¹ ¹⁰καὶ ¹⁰Διὸς Ἀχιλλεύς⁹¹¹⁰
IL.1.7 Atreus' son the lord of men and brilliant Achilles.

We converted the disambiguated Homer into TreeTagger format:

μῆνιν	n--s---fa-	μῆνιν
ἄειδε	v-2spma---	ἄειδω
θεὰ	n--s---fv-	θεά
Πηληϊάδεω	ne-s---mg-	Πηληϊάδης
Ἀχιλλῆος	ne-s---mg-	Ἀχιλλεύς
οὐλομένην	a--s---fa-	οὐλόμενος
	COMM	
ἣ	pr-s---fn-	ὅς
μυρ'	a--p---na-	μυρίος

The problem with the disambiguated data that are available, Homer and the New Testament, is that they are not from the Classical period. So we did our own in-house disambiguation on classical texts.

Θουκυδίδης Ἀθηναῖος **ἐννεύραψε** τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡς **ἐπολέμησαν** πρὸς ἀλλήλους, ἀρξάμενος εὐθύς καθισταμένου καὶ ἐλπίσας **μέγαν** τε ἔσεσθαι καὶ ἀξιολογώτατον τῶν προγεγενημένων, τεκμαίρομενος ὅτι ἀκμάζοντες τε ἦσαν ἐς αὐτὸν ἀμώτεροι παρασκευῇ τῇ πάσῃ καὶ τὸ ἄλλο Ἑλληνικὸν ὄρον ἔρυστατόμενον πρὸς ἐκατέρους, τὸ μὲν εὐθύς, τὸ δὲ καὶ διανοοῦμενον.

κίνησις γὰρ αὕτη μεγίστη δὴ τοῖς Ἑλλήσιν ἐγένετο καὶ μέρει πινί τῶν βαρβάρων, ὡς δὲ εἰπὲν καὶ ἐπὶ πλείστον ἀνθρώπων.

προγεγενημένων	Token #27
προγίγνομαι	
verb perfect middle-passive participle feminine genitive plural	0.333333
verb perfect middle-passive participle masculine genitive plural	0.333333
verb perfect middle-passive participle neuter genitive plural	0.333333

On its own website, Perseus offers Greek texts and a word study tool to go along with it.

The texts transformed: First of all, converted to Unicode..

```
<milestone unit="section" n="17a"/><?τι μὲν ὁμίεις, ὧ ἄνδρες Ἀθηναῖοι,
πεπόνθατε ὑπὸ τῶν ἐμῶν κατηγορῶν, οὐκ οἶδα· ἐγὼ δ' οὐκ αὐτὸς ὑπ'
αὐτῶν ὀλίγου ἑμαυτοῦ ἐπελαθόμην, οὕτω πιθανῶς ἔλεγον. καίτοι ἀληθές γε ὡς
ἔπος εἰπὲν οὐδὲν εἰρήκασιν. μάλιστα δὲ αὐτῶν ἐν ἑθαύμασα τῶν πολλῶν ὧν
ψεύεσαντο, τοῦτο ἐν ᾧ ἔλεγον ὡς χρὴν ὑμᾶς εὐλαβεῖσθαι μὴ ὑπ' ἑμοῦ
ἐξαπατηθῇτε
<milestone unit="section" n="17b"/>
```

Then tokenized and given word id-s.

```
<w id="3708995">?τι</w> <w id="3708996">μὲν</w>
<w id="3708997">ὁμίεις</w>, <w id="3708999">ὧ</w>
<w id="3709000">ἄνδρες</w> <w id="3709001">Ἀθηναῖοι</w>,
<w id="3709003">πεπόνθατε</w> <w id="3709004">ὕπ</w>
<w id="3709005">τῶν</w> <w id="3709006">ἐμῶν</w>
<w id="3709007">κατηγορῶν</w>, <w id="3709009">οὐκ</w>
<w id="3709010">οἶδα</w> <w id="3709012">ἐγὼ</w>
<w id="3709013">δ'</w> <w id="3709014">οὐν</w>
<w id="3709015">καί</w> <w id="3709016">αὐτὸς</w>
<w id="3709017">ὕπ</w> <w id="3709018">αὐτῶν</w>
<w id="3709019">ἀληθῶς</w> <w id="3709020">ἑμαυτοῦ</w>
<w id="3709021">ἐπελαθόμην</w>, <w id="3709023">οὕτω</w>
<w id="3709024">πιθανῶς</w> <w id="3709025">ἔλεγον</w>.
<w id="3709027">καίτοι</w> <w id="3709028">ἀληθῆς</w>
<w id="3709029">γ</w> <w id="3709030">ὡς</w>
<w id="3709031">ἔπος</w> <w id="3709032">εἰπὲν</w>
<w id="3709033">οὐδέν</w> <w id="3709034">εἰρήκασιν</w>.
```

TreeTagger

Documented in Helmut Schmid (1995), Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop. <http://www.jms.unilswilfrut.ac.uk/pubs/publications/tree-tagger.pdf>

```
ὅτι 0-----?τι 0.988714
μεν 0-----μὲν 1.000000
ὁμίεις pr-p---?ς 0.100000
COMM , 1.000000
δ 0-----δ 0.997947
ἄνδρες pr-p---?ς 0.646358 pr-p---?ς 0.353559
ἀθηναῖοι pr-p---?ς 0.795737 pr-p---?ς 0.139543
κατηγορῶν 0-p---?ς 0.861688
COMM , 1.000000
τῶν 0-----τῶν 1.000000
ἐμῶν 0-----ἐμῶν 1.000000
κατηγορῶν 0-p---?ς 0.171051
COMM , 1.000000
οὐκ 0-----οὐ 0.100000
ἐγὼ v-1st-a---?ς 1.000000
δ 0-----δ 1.000000
SENT . 1.000000
ἀληθῶς pr-s---?ς 1.000000
ἑμαυτοῦ pr-s---?ς 1.000000
ἐπελαθόμην 0-p---?ς 1.000000
οὕτω 0-----οὕτω 1.000000
καίτοι 0-----καί 1.000000
ἀληθῆς 0-----ἀλ 0.981171 d-----?ς 0.018829
```

A proprietary part of speech tagger that uses decision trees, rather than a pure Markov model, to tag and lemmatize input streams. This allows it to handle infrequent tags accurately and elegantly.

We project the 1,844 extant configurations of case, gender, tense, voice, etc., onto a homogeneous string format for tagging, storage, and retrieval.

In addition to the training input of any part of speech tagger, TreeTagger also accepts a lexical database as input. We initially trained TreeTagger with approximately 100,000 words from the New Testament, supplemented with the output of Morpheus for our entire corpus.

ἐμός	mine
ἐμῶν	possessive pronoun neut. gen. pl. 1
ἐμῶν	possessive pronoun fem. gen. pl. 0
ἐμῶν	possessive pronoun masc. gen. pl. 0
ἐμέω	(Look up in LSJ, Autenrieth, Slater, Middle Liddell)
to vomit, throw up	
ἐμῶν	present active participle masc. nom. sg. 0
ἐμῶν	future active participle masc. nom. sg. 0

Look Up A Word

The big question: Will users start voting in significant numbers? We'll get back to you on that one!

Visible are light-green, for parses by a single person, brighter for agreement of more than one, and red for disagreement. Below, three parses identified by the system and space for comments.

voteid	tokenid	parseid	lexid
4	3709006	NULL	205676
5	3709006	NULL	205676
6	3709006	NULL	205677

Looking up word 3709006PrvNxt

ἐμός	mine
ἐμῶν	possessive pronoun masc. gen. pl. 0.514341
ἐμῶν	possessive pronoun neut. gen. pl. 0.480822
ἐμῶν	possessive pronoun fem. gen. pl. 0

ἐμέω	(Look up in LSJ, Autenrieth, Slater, Middle Liddell)
to vomit, throw up	
ἐμῶν	present active participle masc. nom. sg. 0
ἐμῶν	future active participle masc. nom. sg. 0

Look Up A Word

SQLite is the heart of our system — most of our development effort has gone into middleware and GUI's for getting data in and out of it.



Master Table (1)
Tables (5)
Lexicon
parses
shortdefs
tokens
votes
Views (0)
Indexes (5)
file_seq
headwordlookup
headwords
parse_tokens
vote_index
Triggers (0)

We compile the data from Perseus, Morpheus, and TreeTagger into 3 relational tables. Two further tables record user votes and short definitions. After the database has been seeded, we can stream tokens in and out of TreeTagger on demand. This allows us to update our parses as we accumulate more training data.

tokenid	content	seq	type	file
3708995	ὅτι	2	word	PlatoApologyGr.xml
3708996	μὲν	3	word	PlatoApologyGr.xml
3708997	ὁμίεις	4	word	PlatoApologyGr.xml
3708998	,	5	punct	PlatoApologyGr.xml
3708999	ὧ	6	word	PlatoApologyGr.xml
3709000	ἄνδρες	7	word	PlatoApologyGr.xml
3709001	Ἀθηναῖοι	8	word	PlatoApologyGr.xml
3709002	,	9	punct	PlatoApologyGr.xml
3709003	πεπόνθατε	10	word	PlatoApologyGr.xml
3709004	ὕπ	11	word	PlatoApologyGr.xml
3709005	τῶν	12	word	PlatoApologyGr.xml
3709006	ἐμῶν	13	word	PlatoApologyGr.xml
3709007	κατηγορῶν	14	word	PlatoApologyGr.xml
3709008	,	15	punct	PlatoApologyGr.xml
3709009	οὐκ	16	word	PlatoApologyGr.xml
3709010	οἶδα	17	word	PlatoApologyGr.xml
3709011	,	18	punct	PlatoApologyGr.xml

Token table

parseid	tokenid	lex	code	lemma	authority	file	prob
6189562	3709003	480649	NULL	NULL	NULL	NULL	1
6189563	3709004	619718	NULL	NULL	NULL	NULL	1
6189564	3709005	611103	NULL	NULL	NULL	NULL	0.515
6189565	3709005	611104	NULL	NULL	NULL	NULL	0.317349
6189566	3709005	611102	NULL	NULL	NULL	NULL	0.16684
6189567	3709006	205678	NULL	NULL	NULL	NULL	0.514341
6189568	3709006	205679	NULL	NULL	NULL	NULL	0.480822
6189569	3709007	343604	NULL	NULL	NULL	NULL	0.444227
6189570	3709007	343605	NULL	NULL	NULL	NULL	0.384722
6189571	3709007	343603	NULL	NULL	NULL	NULL	0.171051
6189572	3709009	450900	NULL	NULL	NULL	NULL	1

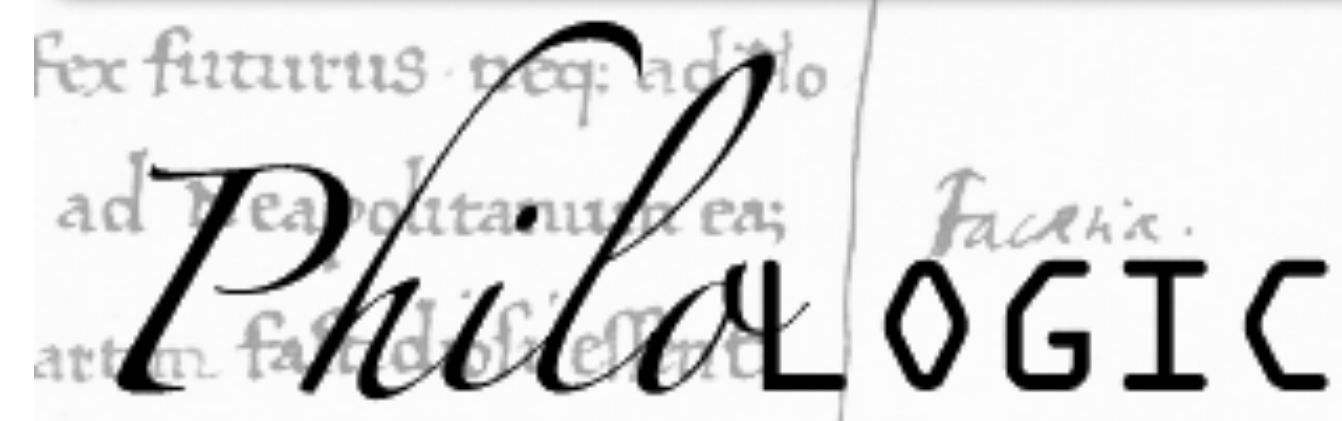
Parse table, fed by TreeTagger and approved votes, feeds TreeTagger in its turn.

id	token	code	lemma	alt_jsj
205670	ἐμῶν	ps-s---md-	ἐμός	NULL
205671	ἐμῶν	ps-s---md-	ἐμός	NULL
205672	ἐμῶν	ps-s---md-	ἐμός	NULL
205673	ἐμῶν	ps-s---md-	ἐμός	NULL
205674	ἐμῶν	v-2salm---	ἐμῶν	NULL
205675	ἐμῶν	v--3pamn---	ἐμῶν	NULL
205676	ἐμῶν	v--3pamn---	ἐμῶν	NULL
205677	ἐμῶν	ps-p---fg-	ἐμός	NULL
205678	ἐμῶν	ps-p---mg-	ἐμός	NULL
205679	ἐμῶν	ps-p---ng-	ἐμός	NULL
205680	ἐμῶν	v-3sala---	ἐμῶν	NULL
205681	ἐν	m--s---nn-	ἐίς	NULL
205682	ἐν	m--s---nn-	ἐίς	NULL
205683	ἐν	m--s---na-	ἐίς	NULL
205684	ἐν	v-3pala---	ἐίς	NULL
205685	ἐν	v--3pamn---	ἐίς	NULL
205686	ἐν	v--3pamn---	ἐίς	NULL
205687	ἐν	v--3pamn---	ἐίς	NULL
205688	ἐν	m--s---nn-	ἐίς	NULL

Lexicon

Short definitions

Finally, all roads lead to Philologic! Its powerful indexing now also keeps track of the morphological information.



A user can still search and browse the texts, see translations, and full dictionary entries.

Perseus Greek Texts

search help report a problem Pl. Ap. 17a Citation Lookup

Plato, *Apology* (XML Header) [genre: prose] [word count] [Pl. Ap.], <<Pl. Ap. 17a PL Ap. 17a (English) >>Pl. Ap. 18a

17a ὅτι μὲν ὁμίεις, ὧ ἄνδρες Ἀθηναῖοι, πεπόνθατε ὑπὸ τῶν ἐμῶν κατηγορῶν, οὐκ οἶδα· ἐγὼ δ' οὐκ αὐτὸς ὑπ' αὐτῶν ὀλίγου ἑμαυτοῦ ἐπελαθόμην, οὕτω πιθανῶς ἔλεγον. καίτοι ἀληθές γε ὡς ἔπος εἰπὲν οὐδὲν εἰρήκασιν. μάλιστα δὲ αὐτῶν ἐν ἑθαύμασα τῶν πολλῶν ὧν ψεύεσαντο, τοῦτο ἐν ᾧ ἔλεγον ὡς χρὴν ὑμᾶς εὐλαβεῖσθαι μὴ ὑπ' ἑμοῦ ἐξαπατηθῇτε

But searching by lemma.. (find me 'my')

Search for: lemma:ἐμός Search Clear

Orthography: ☒ Accented ☐ Accentless ☐ Romanized

Display: ☒ Context ☐ KWIC ☐ Similarity Search

by morphology.. (find me a possessive pronoun)

Search for: pos:ps* Search Clear

Orthography: ☒ Accented ☐ Accentless ☐ Romanized

Display: ☒ Context ☐ KWIC ☐ Similarity Search

or by a combination of criteria (find me a 3rd singular form of the verb δοκέω followed by personal pronoun in the dative)..is all there as well.

Search for: lemma:dokew;pos:v-3s* pos:pp*d- Search Clear

Orthography: ☐ Accented ☐ Accentless ☒ Romanized

Display: ☒ Context ☐ KWIC ☐ Similarity Search

Clicking on a word will bring up the parse information, and provides links to the dictionaries.

The identification of