CALLIDUS, 5-10-2017, FU BERLIN

LATIN TREEBANKING: THE STATUS OF THE ART

Giuseppe G. A. Celano, Leipzig University, DH

OVERVIEW

- Latin treebanks
- Latin treebanks and UD
- Prospects

THE TREEBANKS

- Perseus Treebank: 79 697 tk
- PROIEL Treebank: 195 158 tk
- Index Thomisticus Treebank (+ 3 classical texts): 345 140 tk

THE PERSEUS TREEBANK (LAST RELEASE, 2.1)

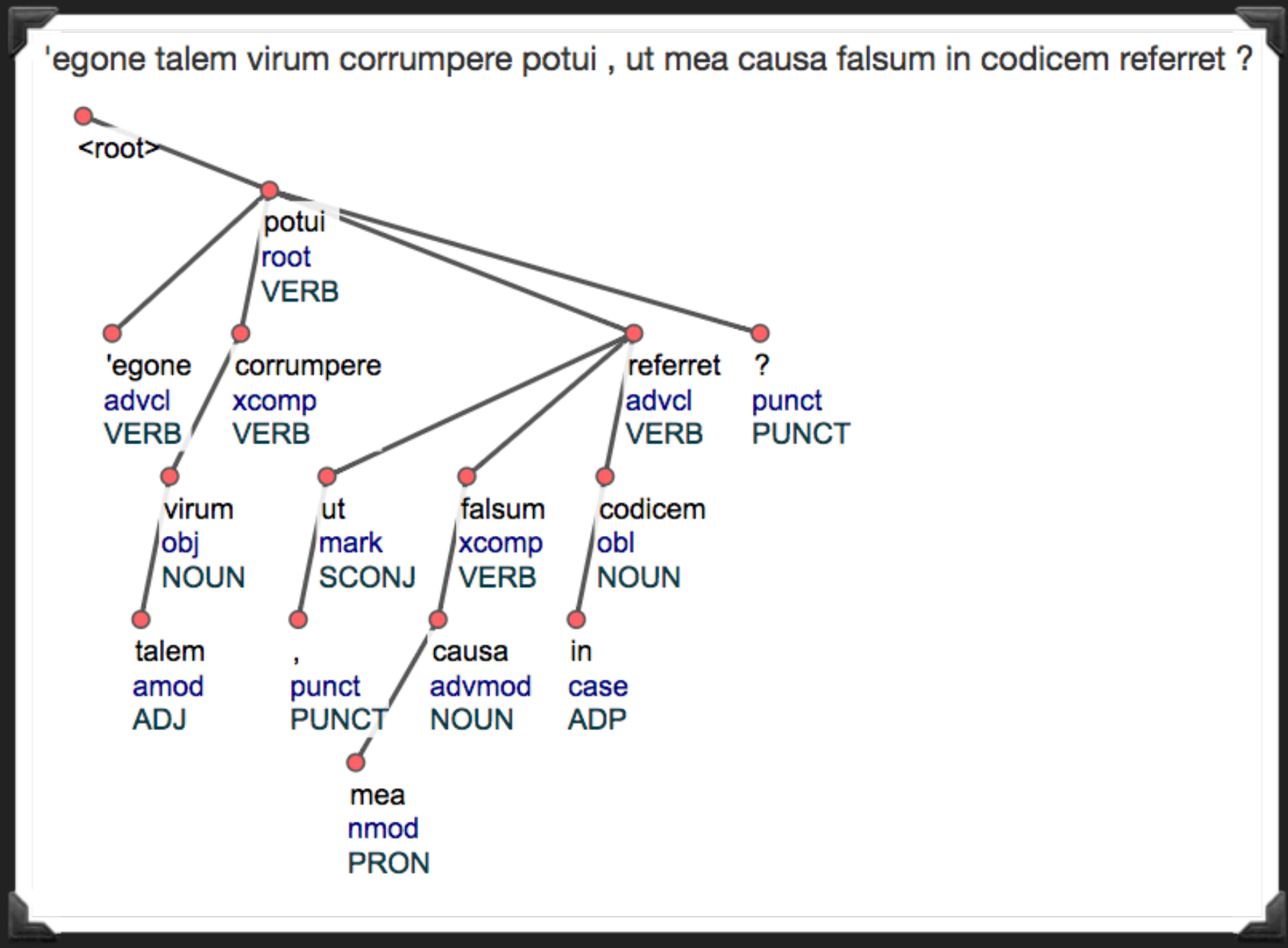
- 12 texts

composition date	text	token number
63 BC	Cicero, In Catilinam	6652
51 BC	Caesar, De Bello Gallico	1556
post 44 BC	Sallust, Bellum Catilinae	13191
ca 25 BC	Prop. Elegiae	5297
29-19 BC	Vergil, Aeneid	2839
ca 8 AD	Ov., Metamorphoses	5209
14 AD	Aug., Res Gestae	3035
15-50 AD	Ph., Fabulae	6588
ca 100 AD?	Petr., Satyricon	14177
ca 100-110 AD	Tac. Historiae	3531
117-138 AD	Suet., Vita Divi Augusti	8313
ca 400 AD	Ger. Vulgata	9309

THE PERSEUS TREEBANK: ANNOTATION

- one-annotator annotation (with the exception of the Phaedrus text)
- ‘open’ annotation model
- annotators: students and scholars
- native annotation scheme based on the PDT (labeled directed acyclic graphs
- Serialization: XML
- data available on GitHub

LABELED DIRECTED ACYCLIC GRAPHS



THE PERSEUS TREEBANK: XML SERIALIZATION

```
    <word id="9" form="sine" lemma="sino1" postag="v2spma---" relation="PRED_CO" head="8"/>
    <word id="10" form="nos" lemma="nos1" postag="p-p---ma-" relation="SBJ" head="16"/>
    <word id="11" form="cursu" lemma="cursus1" postag="n-s---mb-" relation="ADV" head="16"/>
    <word id="12" form="," lemma="comma1" postag="u--------" relation="AuxX" head="14"/>
    <word id="13" form="quo" lemma="qui1" postag="p-s---mb-" relation="ADV" head="14"/>
    <word id="14" form="sumus" lemma="sum1" postag="v1ppia---" relation="ATR" head="11"/>
    <word id="15" form="," lemma="comma1" postag="u--------" relation="AuxX" head="14"/>
    <word id="16" form="ire" lemma="eo1" postag="v--pna---" relation="OBJ" head="9"/>
    <word id="17" form="pares" lemma="par1" postag="a-p---ma-" relation="ATV" head="10"/>
    <word id="18" form="!" lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
  </sentence>
  <sentence id="73" document_id="urn:cts:latinLit:phi0620.phi001.perseus-lat1" subdoc="1.5">
    <word id="1" form="quid" lemma="quis1" postag="p-s---na-" relation="OBJ" head="3"/>
    <word id="2" form="tibi" lemma="tu1" postag="p-s---md-" relation="OBJ" head="3"/>
    <word id="3" form="vis" lemma="volo1" postag="v2spia---" relation="PRED" head="0"/>
    <word id="4" form="," lemma="comma1" postag="u--------" relation="AuxX" head="5"/>
    <word id="5" form="insane" lemma="insanus1" postag="a-s---mv-" relation="ExD" head="3"/>
    <word id="6" form="?" lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
  </sentence>
  <sentence id="74" document_id="urn:cts:latinLit:phi0620.phi001.perseus-lat1" subdoc="1.5">
    <word id="1" form="meos" lemma="meus" postag="p-p---ma-" relation="ATR" head="3"/>
    <word id="2" form="sentire" lemma="sentio1" postag="v--pna---" relation="OBJ" head="5"/>
    <word id="3" form="furores" lemma="furor2" postag="n-p---ma-" relation="OBJ" head="2"/>
```

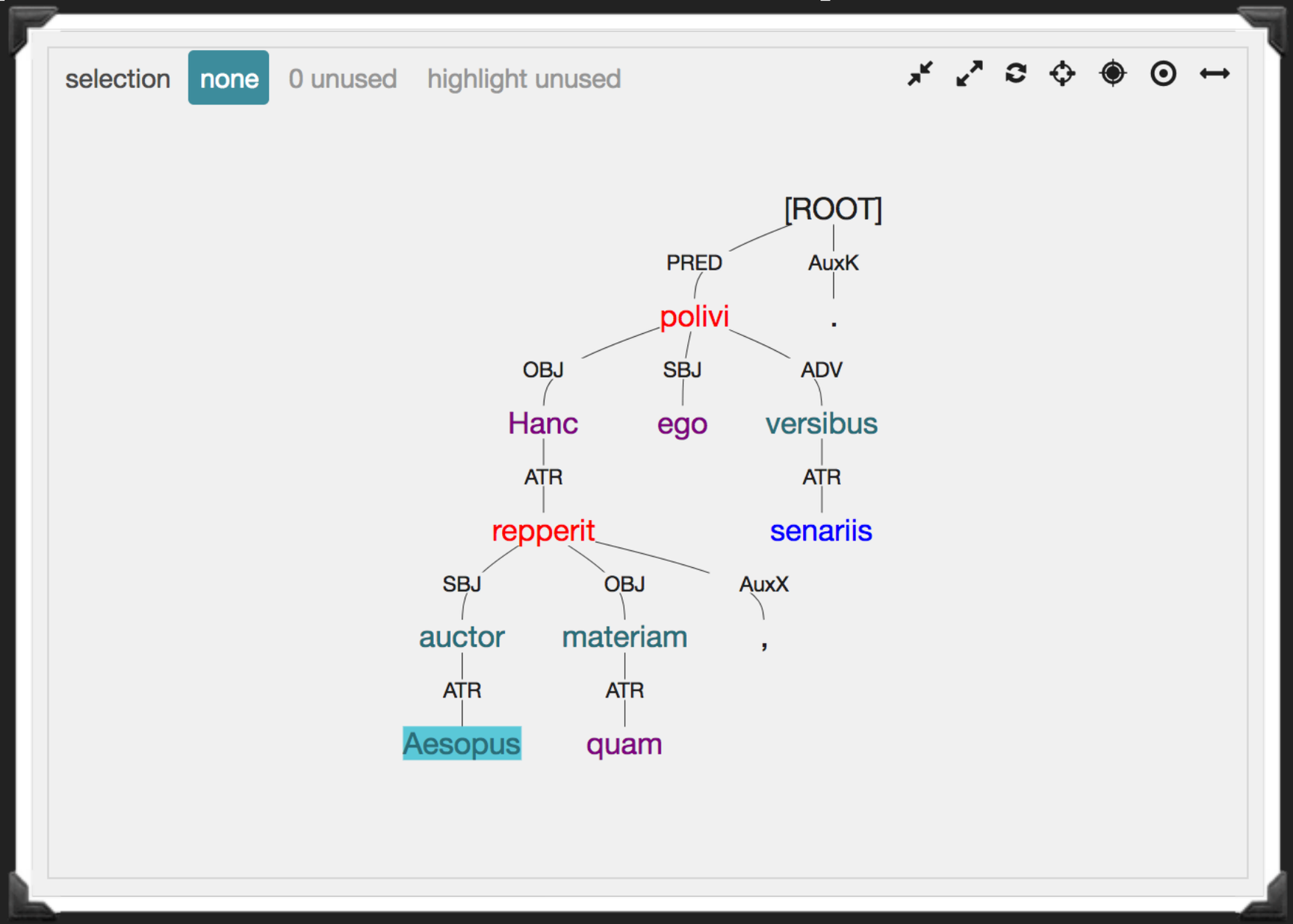
THE PERSEUS TREEBANK: OPEN ISSUES

- student annotation model
- no inter-annotator agreement calculation
- not specific enough guidelines
- complex texts
- errors and inconsistencies
- tokenization/normalization

THE PERSEUS TREEBANK: ANNOTATION ISSUES

Aesopus auctor quam materiam repperit, hanc ego polivi versibus senariis

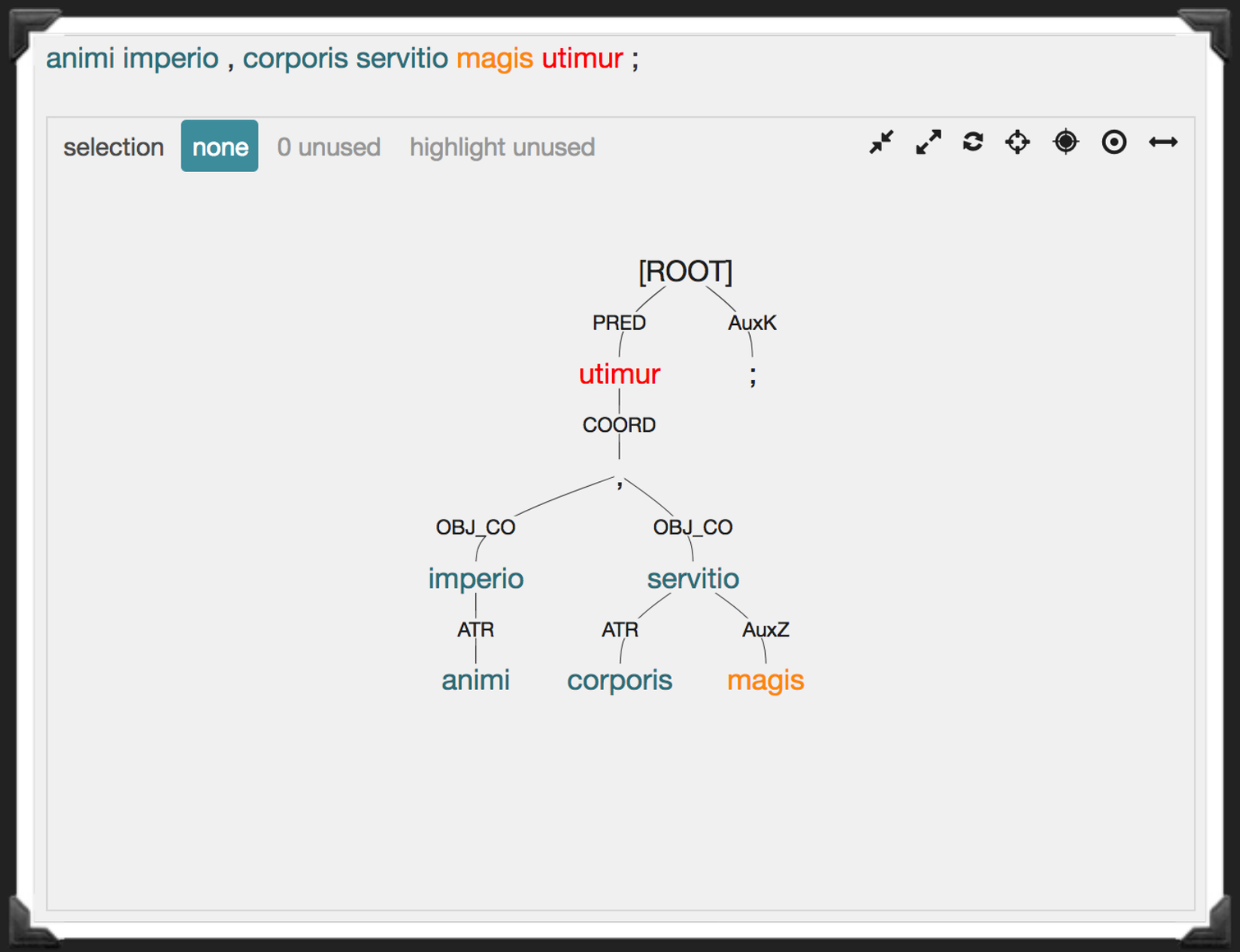
Aesopus author which material found, this I polished verse senarian



THE PERSEUS TREEBANK: ANNOTATION ISSUES

animi imperio, corporis servitio magis utimur

mind authority, body servitude rather (we)use



THE PROEIL TREEBANK (LAST RELEASE, 20170214)

- 4 texts

composition date	text	token number
51 BC	Caesar *De Bello Gallico*	28608
68-44 BC	Cicero *Ad Atticum*	41915
ca 400 AD	Gerolamus Vulgata	106279
ca 400 AD	*[Anonymous]* *Peregrinatio Etheriae*	18356

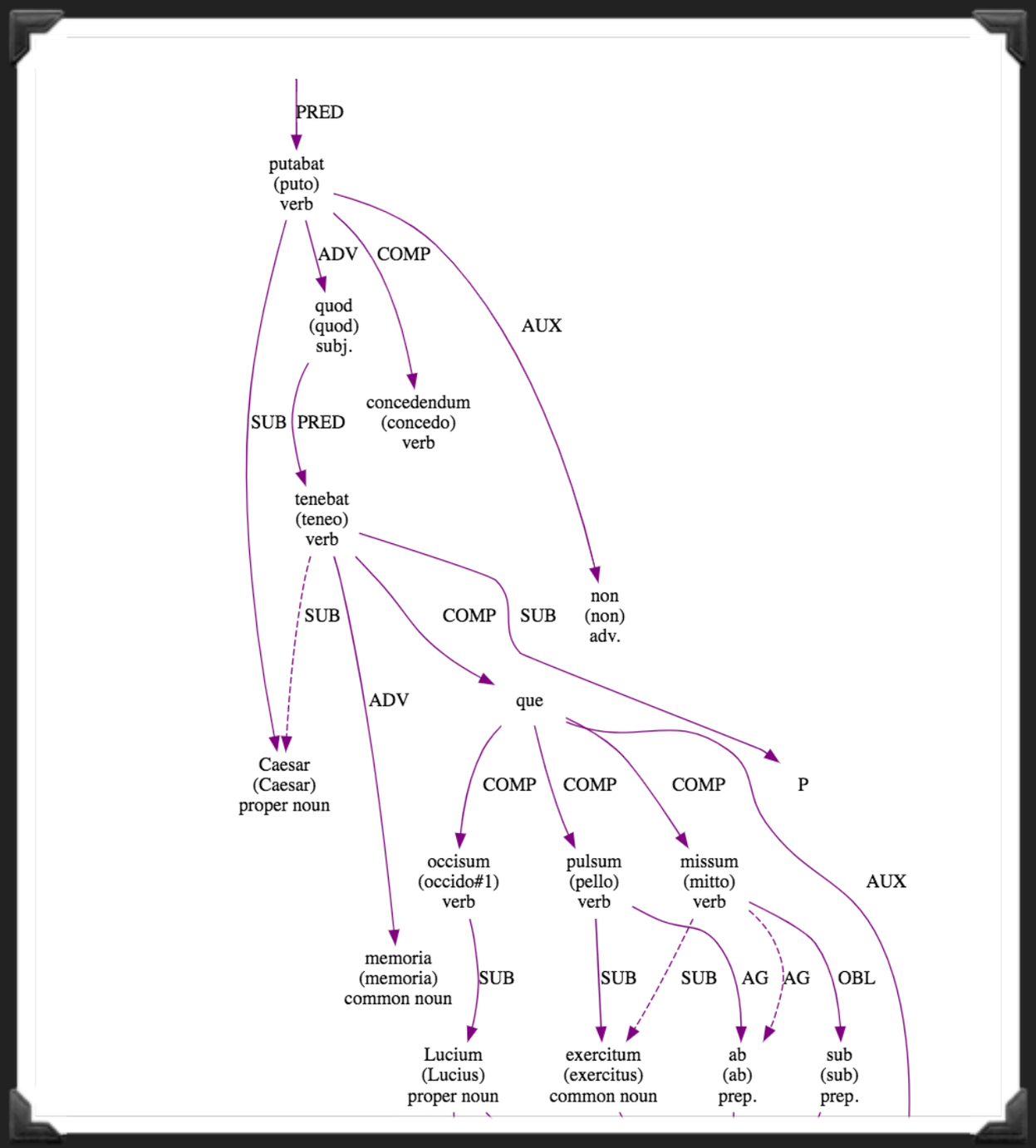
THE PROEIL TREEBANK: ANNOTATION

- one-annotator annotation + second annotator review
- ‘closed’ annotation model
- annotators: students
- native annotation scheme based on the PDT
- annotation of information structure/some coreference resolution
- Serialization: XML
- data available on GitHub

THE PROIEL TREEBANK: XML SERIALIZATION

```
      <token id="1236846" empty-token-sort="P" head-id="1084592" relation="sub" antecedent-id="10
      <token id="1084592" form="putem" citation-part="1.14.4" lemma="puto" part-of-speech="V-" mo
    </sentence>
    <sentence id="76109" status="reviewed">
      <token id="1084593" form="Romanae" citation-part="1.14.5" lemma="Romanus" part-of-speech="A
      <token id="1084594" form="autem" citation-part="1.14.5" lemma="autem" part-of-speech="Df" m
      <token id="1084595" form="se" citation-part="1.14.5" lemma="se" part-of-speech="Pk" morphol
      <token id="1084596" form="res" citation-part="1.14.5" lemma="res" part-of-speech="Nb" morph
      <token id="1084597" form="sic" citation-part="1.14.5" lemma="sic" part-of-speech="Df" morph
      <token id="1084598" form="habent" citation-part="1.14.5" lemma="habeo" part-of-speech="V-"
    </sentence>
    <sentence id="76110" status="reviewed" presentation-after=" ">
      <token id="1171879" form="senatus" citation-part="1.14.5" lemma="senatus" part-of-speech="N
      <token id="1171880" form="Ἄρειος" citation-part="1.14.5" lemma="greek expression" part-of-s
      <token id="1171881" form="πάγος" citation-part="1.14.5" lemma="greek expression" part-of-sp
        <slash target-id="1171879" relation="xsub"/>
      </token>
      <token id="1171882" empty-token-sort="V" citation-part="1.14.5" relation="pred"/>
    </sentence>
    <sentence id="82350" status="reviewed" presentation-after=" ">
      <token id="1171883" form="nihil" citation-part="1.14.5" lemma="nihil" part-of-speech="Px" m
      <token id="1171884" form="constantius" citation-part="1.14.5" lemma="constans" part-of-spee
        <slash target-id="1171883" relation="xsub"/>
      </token>
      <token id="1171885" empty-token-sort="V" citation-part="1.14.5" relation="pred"/>
    </sentence>
    <sentence id="82351" status="reviewed" presentation-after=" ">
      <token id="1171886" form="nihil" citation-part="1.14.5" lemma="nihil" part-of-speech="Px" m
      <token id="1171887" form="severius" citation-part="1.14.5" lemma="severus" part-of-speech="
        <slash target-id="1171886" relation="xsub"/>
```

A TREE IN THE PROIEL TREEBANK



THE IT TREEBANK

- 4 texts

composition date	text	token number
51 BC	Caesar *De Bello Gallico*	1488
63 BC	Cicero *In Catilinam*	6229
post 44 BC	Sallustius De Coniuratione Catilinae	15072
1265-1274 AD	ITTB	322351

THE IT TREEBANK: ANNOTATION

- one-annotator + one reviewer
- same annotation scheme as the PT
- tectogrammatical annotation for De Bello Gallico, In Catilinam, De Coniuratione Catilinae
- XML Serialization closely following the PDT
- data freely downloadable from proprietary website

RECENT LATIN TAGGERS

- Steffen Eger et al, 2016. Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. In LREC2016 (~90% all tags MarMot: ~95% lemma MarMot)

- Steffen Eger et al., 2015. Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (~85% all tags, LAPOS; ~95 lemmas)

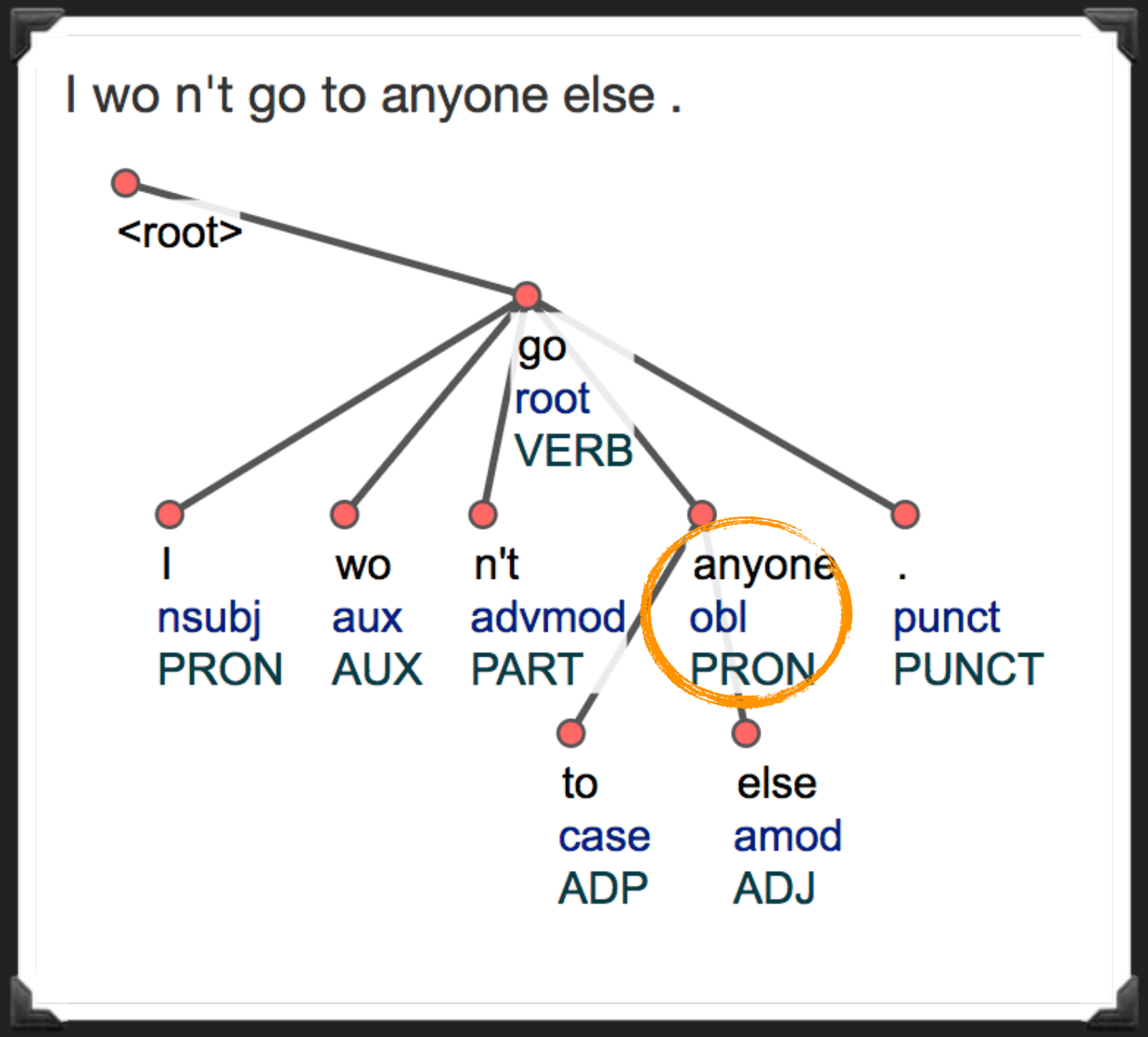
UNIVERSAL DEPENDENCIES

- one morphosyntactic annotation scheme for all languages (50 languages so far)
- language customization as sub-categories
- emphasis on form rather than meaning
- primacy of content words over function words
- asymmetrical annotation for coordination

LANGUAGE-SPECIFIC FEATURES

- 15 in in ADP S4|stRL AdpType=Prep 17 case _ _
- 33 etiam etiam ADV O4|vgr1|stRL _ 35 advmod:emph _ _

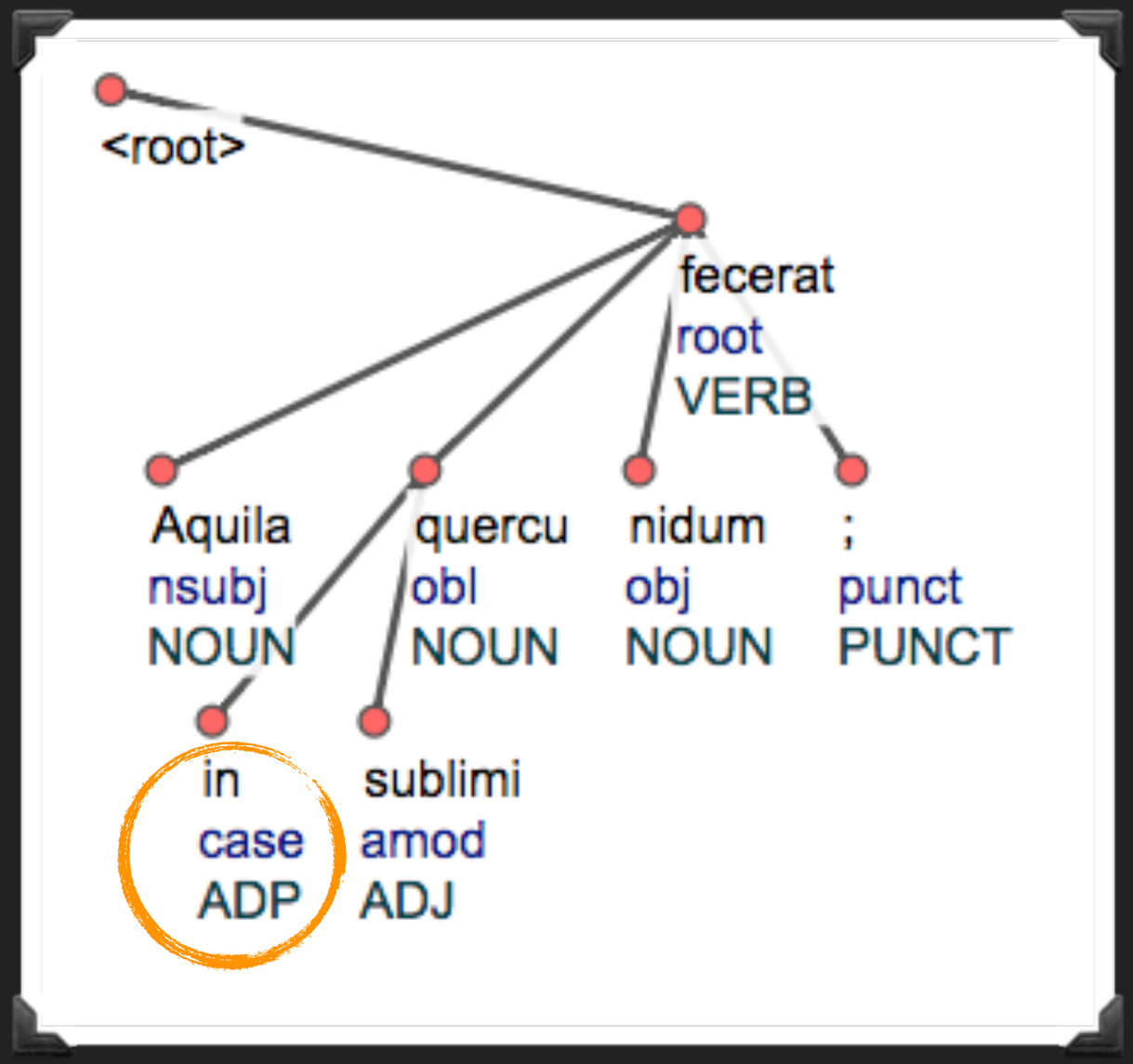
NO ARGUMENT STRUCTURE ANNOTATION



PRIORITY OF CONTENT WORDS

aquila in sublimi quercu nidum fecerat

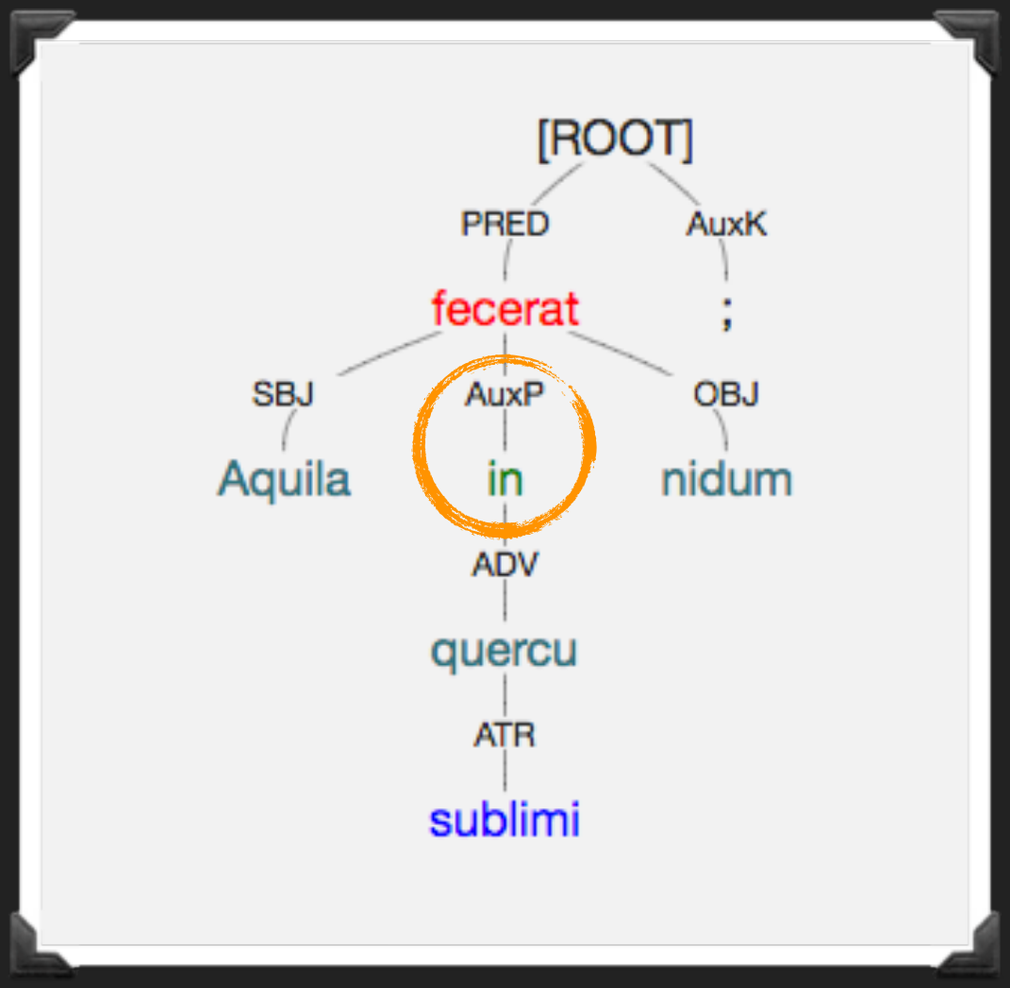
‘an eagle did a nest on a high oak’



PRIORITY OF CONTENT WORDS

aquila in sublimi quercu nidum fecerat

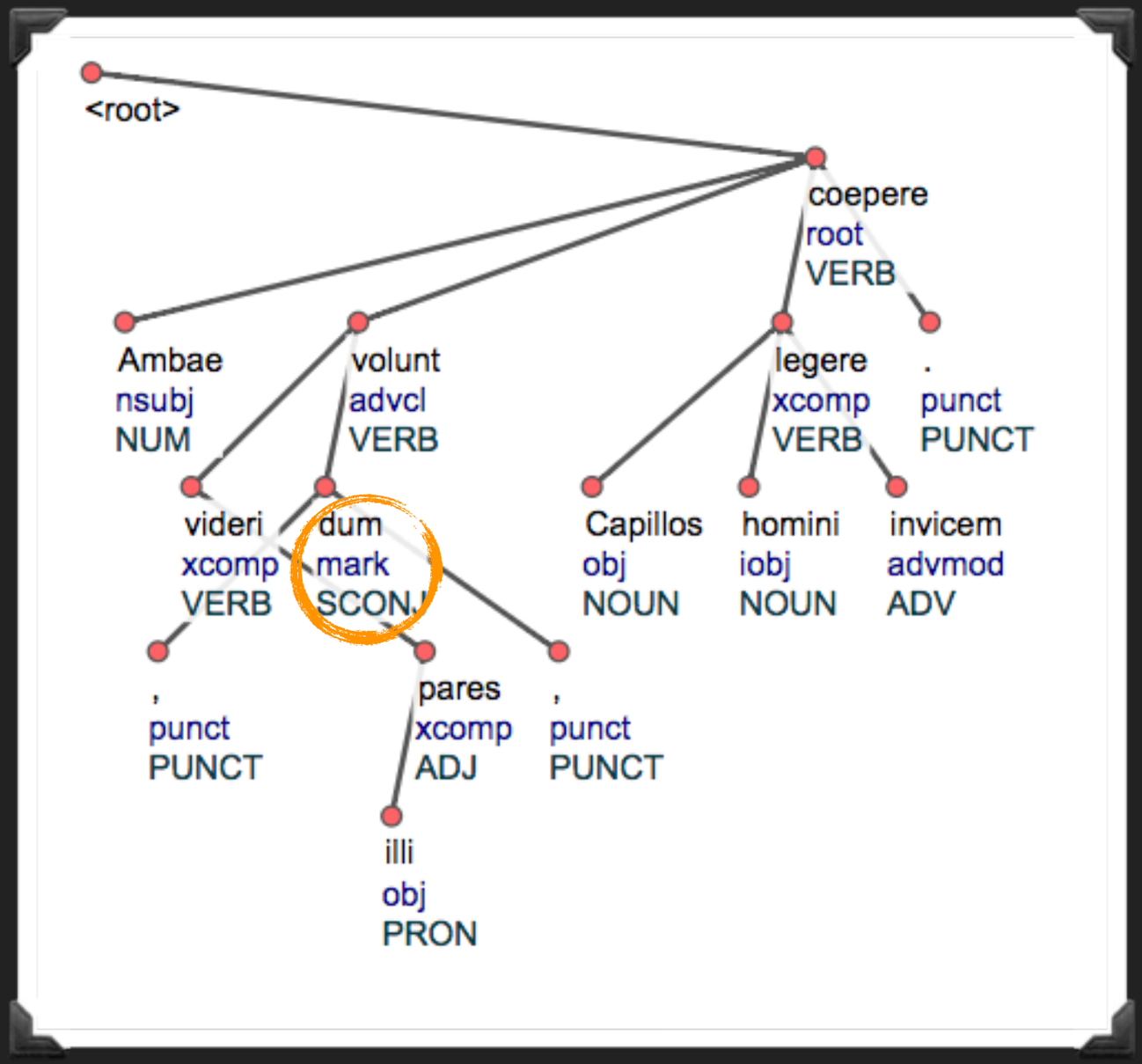
‘an eagle did a nest on a high oak’



PRIORITY OF CONTENT WORDS

ambae, videri dum volunt illi pares, capillos homini legere coepere invicem

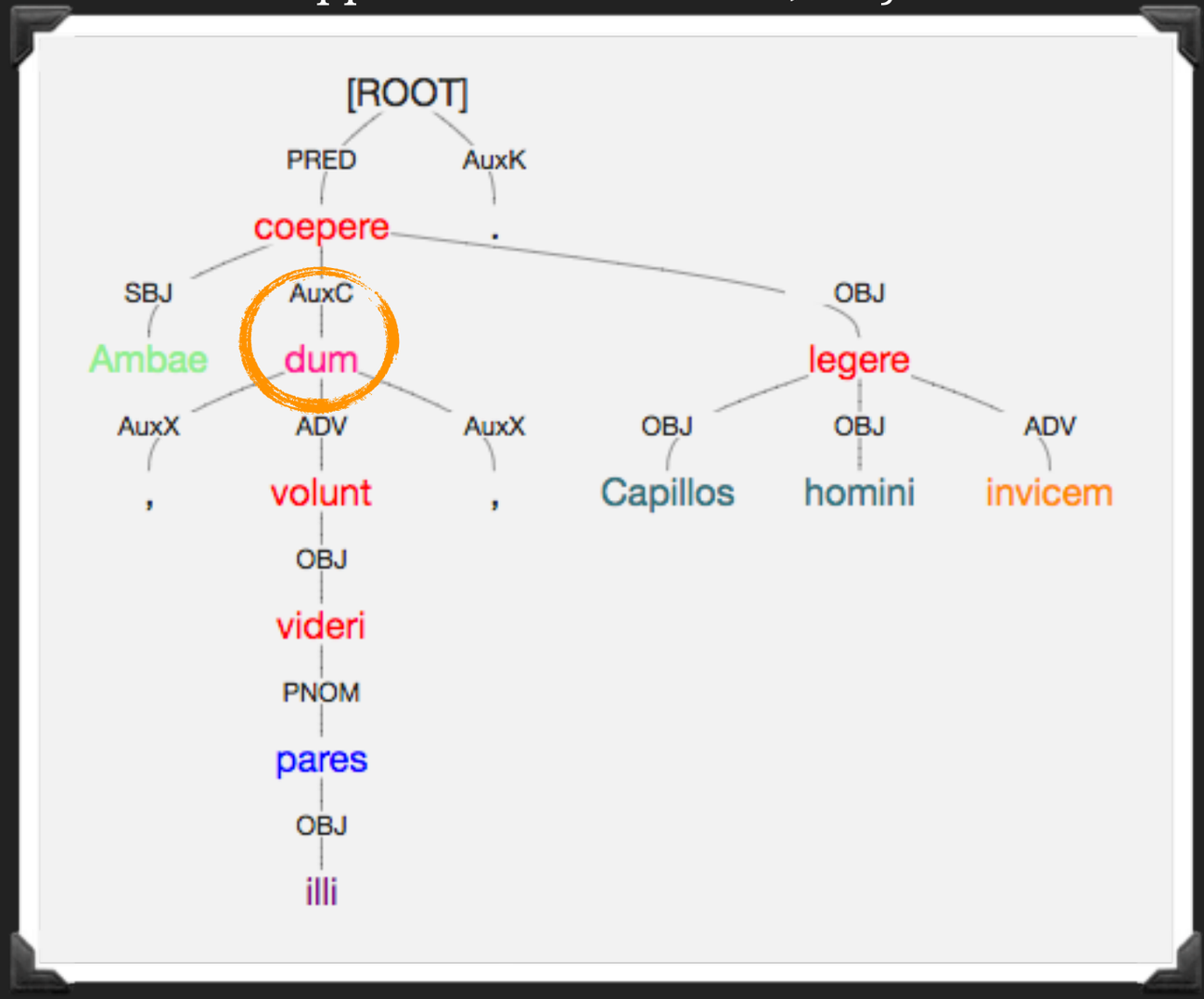
‘while they both wanted to appear as old as he was, they started to tear off his hair’



PRIORITY OF CONTENT WORDS

ambae, videri dum volunt illi pares, capillos homini legere coepere invicem

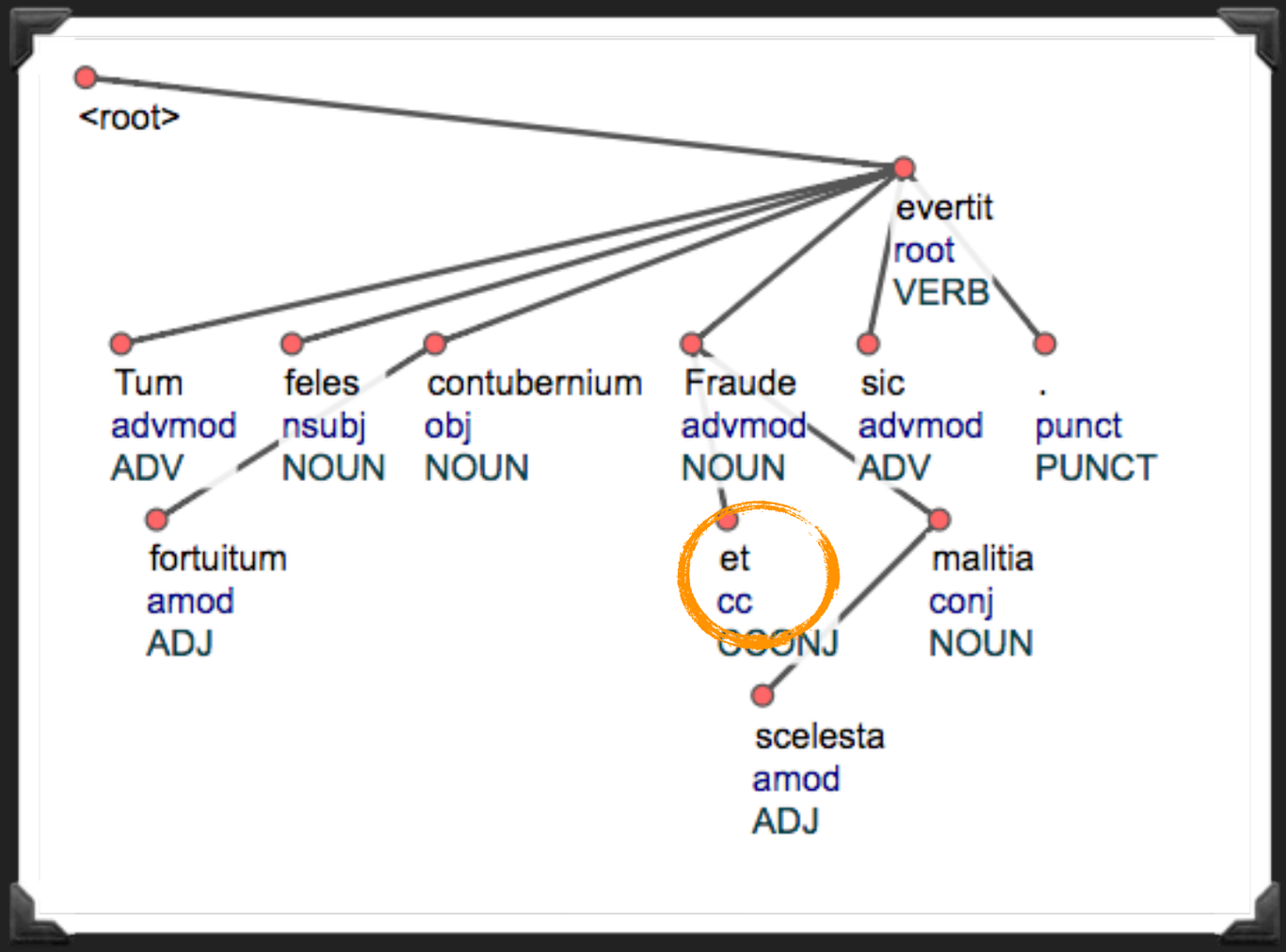
‘while they both wanted to appear as old as he was, they started to tear off his hair’



PRIORITY OF CONTENT WORDS

tum fortuitum feles contubernium fraude et scelesta sic evertit malitia

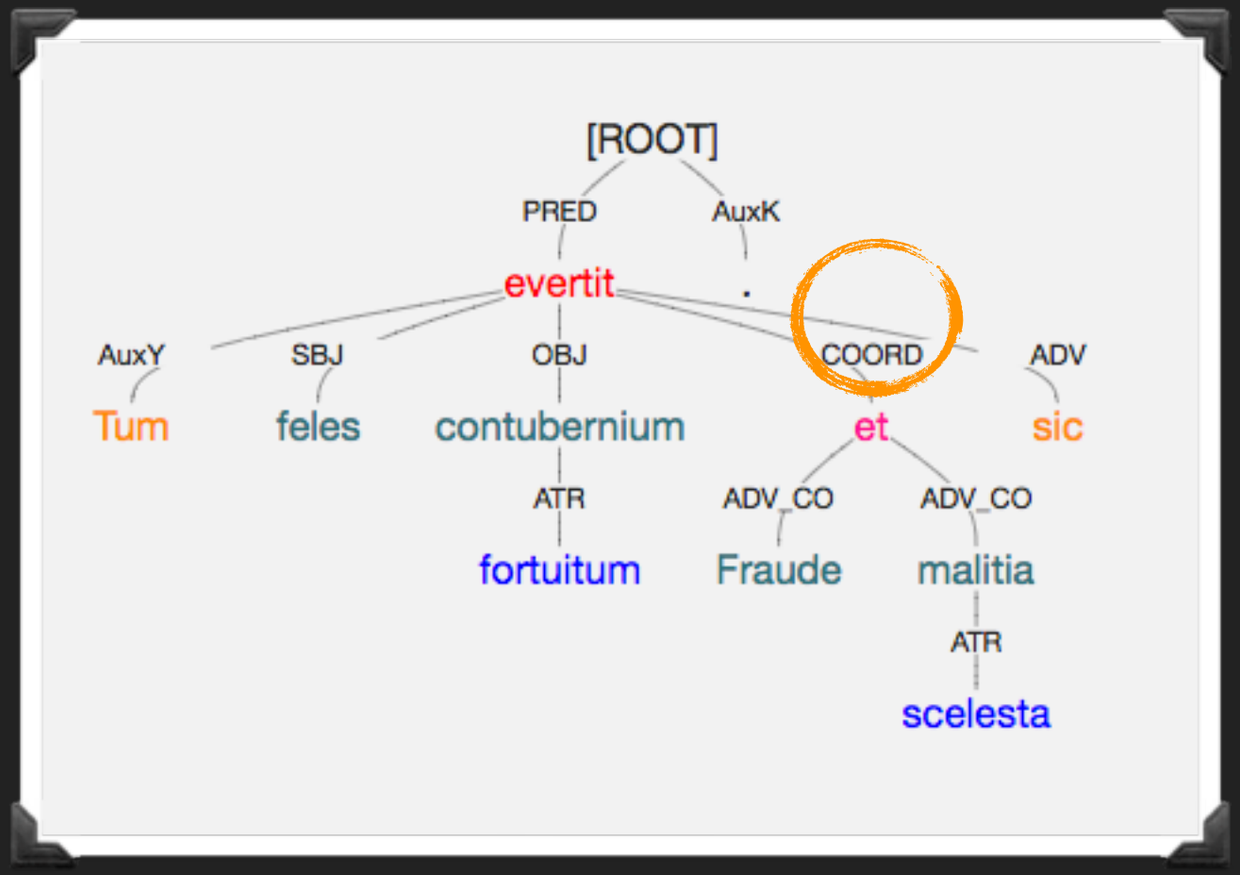
‘then the cat destroyed the cohabitation with deceit and heinous treachery’



PRIORITY OF CONTENT WORDS

tum fortuitum feles contubernium fraude et scelesta sic evertit malitia

‘then the cat destroyed the cohabitation with deceit and heinous treachery’



SHAREDTASK 2017 (HTTP://UNIVERSALDEPENDENCIES.ORG/CONLL17/)

1. Stanford (Stanford)	76.30 ± 0.12
2. C2L2 (Ithaca)	75.00 ± 0.12
3. IMS (Stuttgart)	74.42 ± 0.13
4. HIT-SCIR (Harbin)	72.11 ± 0.14
5. LATTICE (Paris)	70.93 ± 0.13
6. NAIST SATO (Nara)	70.14 ± 0.13
7. Koç University (İstanbul)	69.76 ± 0.13
8. ÚFAL – UDPipe 1.2 (Praha)	69.52 ± 0.13
9. UParse (Edinburgh)	68.87 ± 0.14
10. Orange – Deskiñ (Lannion)	68.61 ± 0.13

UDPIPE

treebank		w	s	upos	xpos	feats	all	lemma	UAS	LAS
Perseus	raw	100	98	83.4	67.6	72.5	67.6	51.2	56.5	46
Perseus	g-tok	-	-	83.4	67.6	72.5	67.6	51.2	56.6	46.1
Perseus	g-tok+m	-	-	-	-	-	-	-	67.8	61.5
PROIEL	raw	99.9	31	94.9	95	87.7	86.7	94.8	66.1	60.7
PROEIL	g-tok	-	-	95.2	95.2	88.4	87.4	95	75.3	69.4
PROIEL	g-tok+m	-	-	-	-	-	-	-	79	75
ITTB	raw	99.9	82.5	97.2	92.7	93.5	91.3	97.8	79.7	76
ITTB	g-tok		-	97.3	92.8	93.6	91.4	97.9	81.8	78.1
ITTB	g-tok+m	-	-	-	-	-	-	-	87.6	85.2

OGL TEXTS

- canonical latin repository: 5 819 581 tk
- CSEL repository: 7 080 417 tk
- full morphosyntactic annotations available at: https://github.com/gcelano/LatinUD

FUTURE WORK

- which model performs best with with which text?
- aligning the automatic annotations to spot differences/errors
- manually check a sample of the corpus
- use the lexica to improve lemmatization and check morphological annotation

THANK YOU FOR YOUR ATTENTION!