Treebanking and linguistic research with BaseX

Giuseppe G. A. Celano
University of Leipzig
at the Humboldt Chair in Digital Humanities
11 February 2016

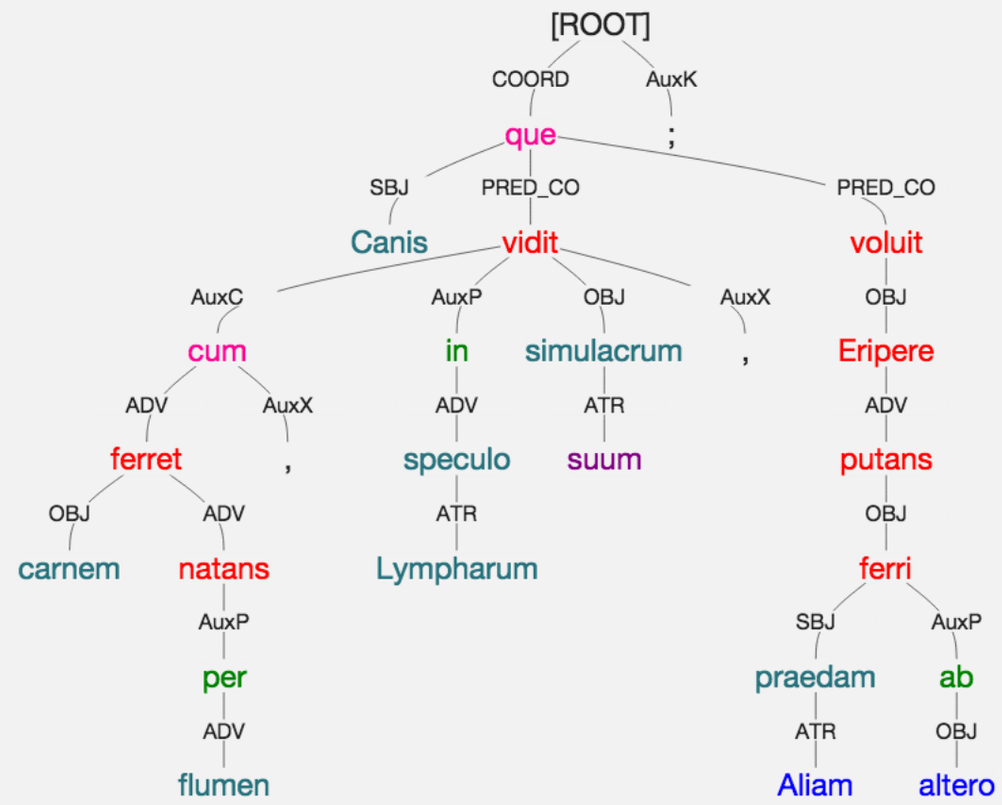
Treebank: a definition

a **treebank** is a corpus
containing linguistic trees

A graphical representation for a tree

Canis per flumen carnem cum ferret natans , Lympharum in speculo vidit simulacrum suum , Aliam que praedam ab altero ferri putans Eripere voluit ;

selection none 0 unused highlight unused



The AGLD Treebank

Ancient greek texts:

- 15 authors
- 32 (parts of) works
- 557.922 tokens

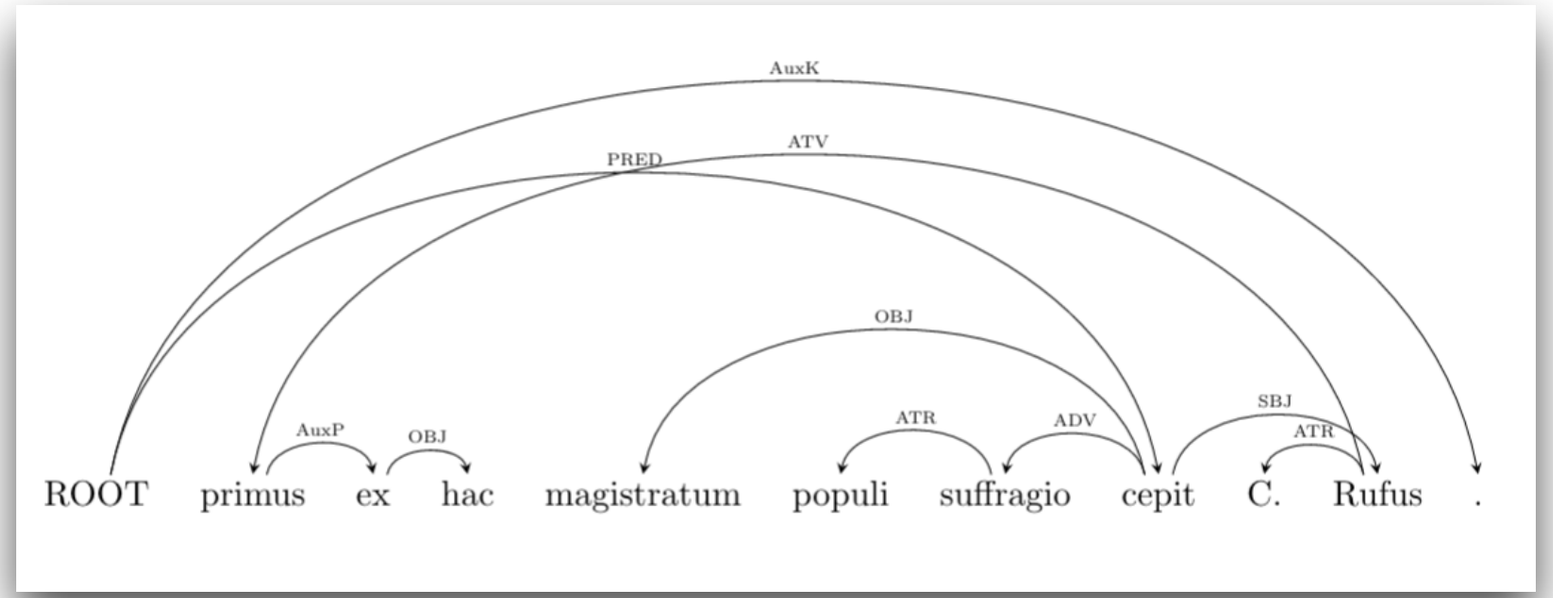
Latin texts:

- 9 authors
- 9 (parts of) works
- 64.979 tokens

Dependency tree: a formal definition

A dependency tree can be defined as a dependency graph, i.e., a labeled directed graph

(cfr. J. Nivre, 2009, *Dependency Parsing*)



The underlying representation

```
<sentence id='42' document_id='' subdoc='1:4' span=''>
  <word id='1' form='Canis' lemma='canis1' postag='n-s---mn-' relation='SBJ' ref='1:4' head='17'/>
  <word id='2' form='per' lemma='per' postag='r--------' relation='AuxP' ref='1:4' head='7'/>
  <word id='3' form='flumen' lemma='flumen' postag='n-s---na-' relation='ADV' ref='1:4' head='2'/>
  <word id='4' form='carnem' lemma='caro2' postag='n-s---fa-' relation='OBJ' ref='1:4' head='6'/>
  <word id='5' form='cum' lemma='cum' postag='c--------' relation='AuxC' ref='1:4' head='12'/>
  <word id='6' form='ferret' lemma='fero' postag='v3sisa---' relation='ADV' ref='1:4' head='5'/>
  <word id='7' form='natans' lemma='nato' postag='v-spp-mn-' relation='ADV' ref='1:4' head='6'/>
  <word id='8' form=',' lemma=',' postag='u--------' relation='AuxX' ref='1:4' head='5'/>
  <word id='9' form='Lympharum' lemma='lympha' postag='n-p---fg-' relation='ATR' ref='1:4' head='11'/>
  <word id='10' form='in' lemma='in' postag='r--------' relation='AuxP' ref='1:4' head='12'/>
  <word id='11' form='speculo' lemma='speculum' postag='n-s---nb-' relation='ADV' ref='1:4' head='10'/>
  <word id='12' form='vidit' lemma='video' postag='v3sria---' relation='PRED_CO' ref='1:4' head='17'/>
  <word id='13' form='simulacrum' lemma='simulacrum' postag='n-s---na-' relation='OBJ' ref='1:4' head='12'/>
  <word id='14' form='suum' lemma='suus' postag='p-s---na-' relation='ATR' ref='1:4' head='13'/>
  <word id='15' form=',' lemma=',' postag='u--------' relation='AuxX' ref='1:4' head='12'/>
  <word id='16' form='Aliam' lemma='alius2' postag='a-s---fap' relation='ATR' ref='1:4' head='18'/>
  <word id='17' form='que' lemma='que' postag='c--------' relation='COORD' ref='1:4' head='0'/>
  <word id='18' form='praedam' lemma='praeda' postag='n-s---fa-' relation='SBJ' ref='1:4' head='21'/>
  <word id='19' form='ab' lemma='ab' postag='r--------' relation='AuxP' ref='1:4' head='21'/>
  <word id='20' form='altero' lemma='alter' postag='a-s---mb-' relation='OBJ' ref='1:4' head='19'/>
  <word id='21' form='ferri' lemma='fero' postag='v--pnp---' relation='OBJ' ref='1:4' head='22'/>
  <word id='22' form='putans' lemma='puto' postag='v-spp-mn-' relation='ADV' ref='1:4' head='23'/>
  <word id='23' form='Eripere' lemma='eripio' postag='v--pna---' relation='OBJ' ref='1:4' head='24'/>
  <word id='24' form='voluit' lemma='volo1' postag='v3sria---' relation='PRED_CO' ref='1:4' head='17'/>
  <word id='25' form=';' lemma=';' postag='u--------' relation='AuxK' ref='1:4' head='0'/>
</sentence>
```

xml serialization

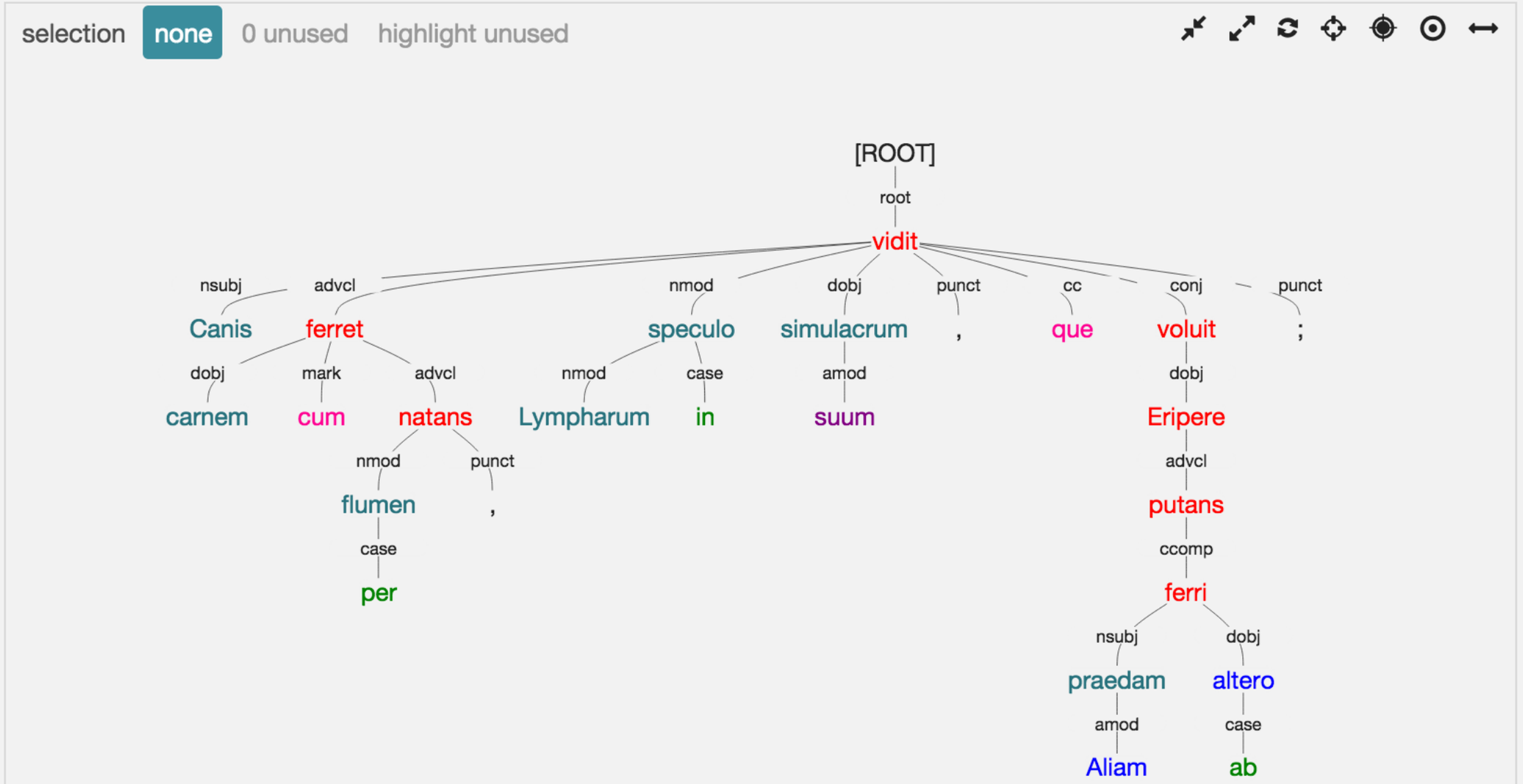
Universal Dependency

- A common annotation scheme for all languages
- dependency grammar formalism
- more than 40 languages included in the current version
- modern and ancient languages

http://universaldependencies.org/

A Universal Dependencies tree

Canis per flumen carnem cum ferret natans , Lympharum in speculo vidit simulacrum suum , Aliam que praedam ab altero ferri putans Eripere voluit ;



The underlying representation

```
1→primus→primus→_→m-s---mn-→_→9→ATV→_→_
2→ex→ex→_→r---------→_→1→AuxP→_→_
3→hac→hic→_→p-s---fb-→_→2→OBJ→_→_
4→magistratum→magistratus→_→n-s---ma-→_→7→OBJ→_→_
5→populi→populus1→_→n-s---mg-→_→6→ATR→_→_
6→suffragio→suffragium→_→n-s---nb-→_→7→ADV→_→_
7→cepit→capio1→_→v3sria---→_→0→PRED→_→_
8→C.→Caius→_→n-s---mn-→_→9→ATR→_→_
9→Rufus→Rufus2→_→n-s---mn-→_→7→SBJ→_→_
10→.→.→_→u---------→_→0→AuxK→_→_
```

a plain text serialization

BaseX

- extremely light piece of software
- well-documented
- versatile: basexgui and restxq

BaseX as an XQuery processor

- XQUF
- File module
- Fetch module
- JSON module
- Process module

XQUF and File module

file* | file2* | file3* | file4* | file5*

```
declare variable $l := collection("/Users/mycomputer/Documents/UDTransformLastLatin/1.5/dat/");
 (: /Users/mycomputer/Documents/treebank-ultimate/textwithcorrectnouns/ :)
for $g in $l
return
copy $e := $g
modify (
for $n in $e//sentence

for $g in ($n//word[@relation = "COORD"])[1]

(: dependents :)
let $h := ($g/parent::sentence/word[@head = $g/@id]
                                   [contains(@relation, "_CO")
                                    or contains(@relation, "COORD")])[1]

where $h
return
replace value of node $g/@head with $h/@id

)
return
  file:write(concat("/Users/mycomputer/Documents/UDTransformLastLatin/1.5/data/",
            file:name(base-uri($e))) ,$e)
```

XQUF and File module

file* | file2* | file3*

```
declare variable $l := collection("/Users/mycomputer/Documents/UDTransformLastLatin/1.5/afterPrepCoordConjAposPnom");

for $g in $l
return
copy $e := $g
modify (

for $n in $e//sentence
where $n/word[contains(@relation, "ExD0")]
return
delete node $n

)

return
  file:write(concat("/Users/mycomputer/Documents/UDTransformLastLatin/1.5/afterPrepCoordConjAposPnom/",
             file:name(base-uri($e))) ,$e)
```

L-processor

- A work-in-progress XQuery Library module to process (ancient) languages

Dolphin

- A tool to calculate inter-coder agreement

Thanks for your attention!