Introduction to treebanking (I)

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

10 February 2016

Overview

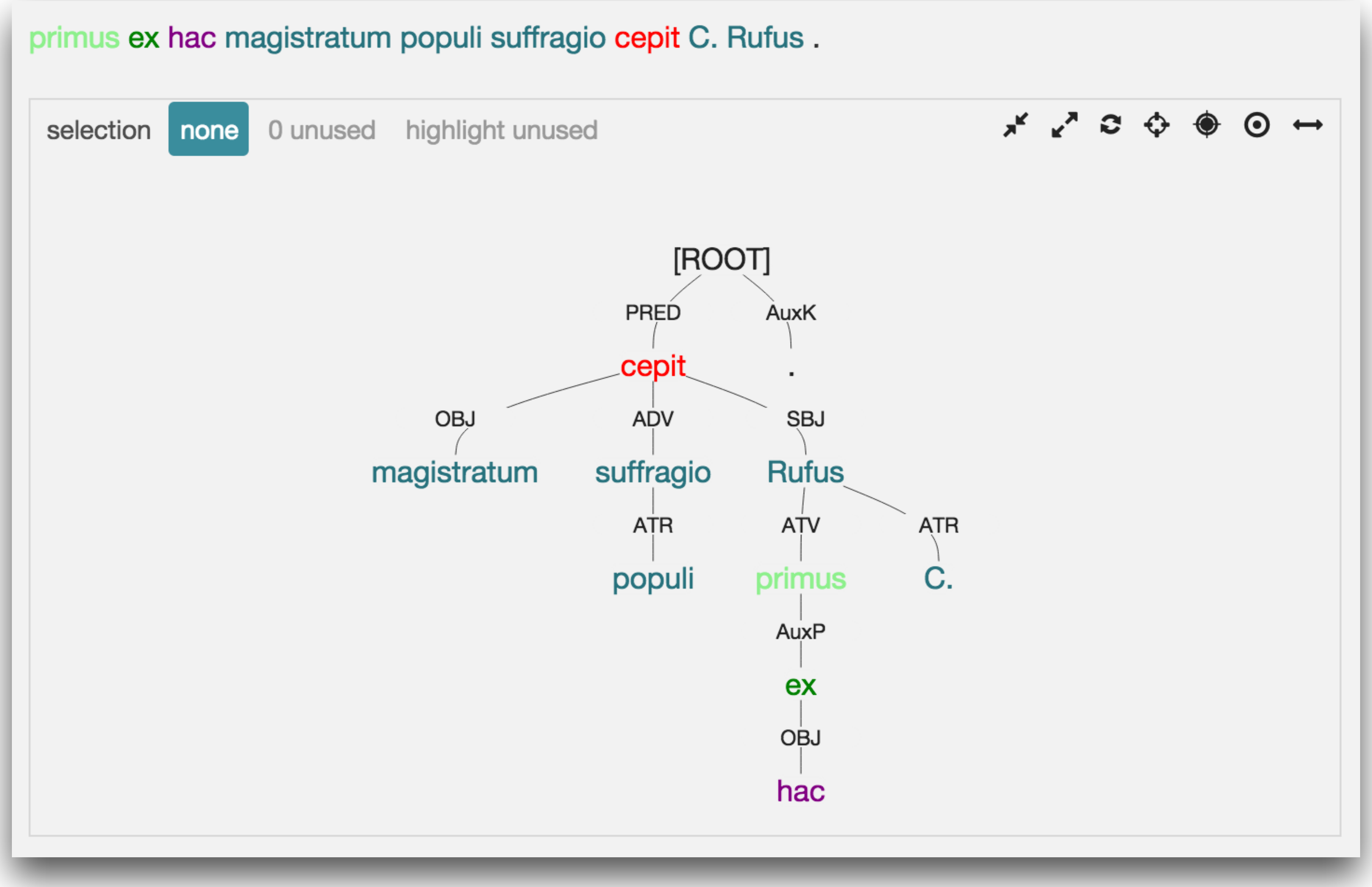
1. Treebank: the basics
2. The Ancient Greek and Latin Dependency Treebank
3. Treebanking in action

Treebanking: an informal definition

treebanking is the activity of building linguistic trees

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

A graphical representation

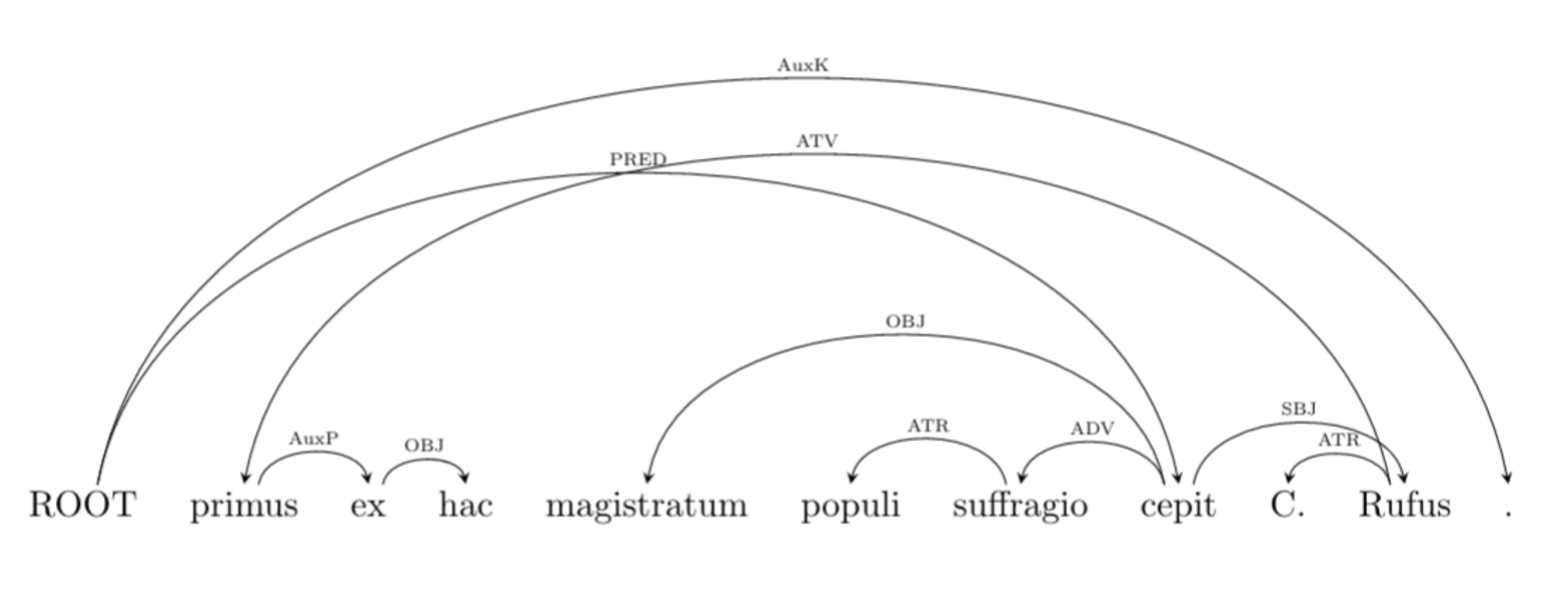


primus ex hac magistratum populi suffragio cepit C. Rufus .

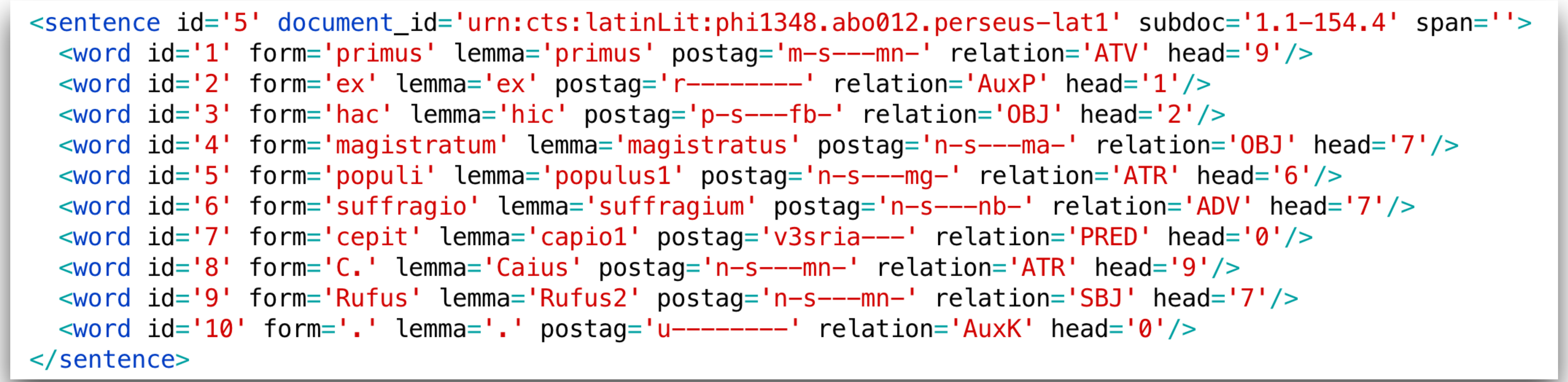
Dependency tree: a formal definition

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

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



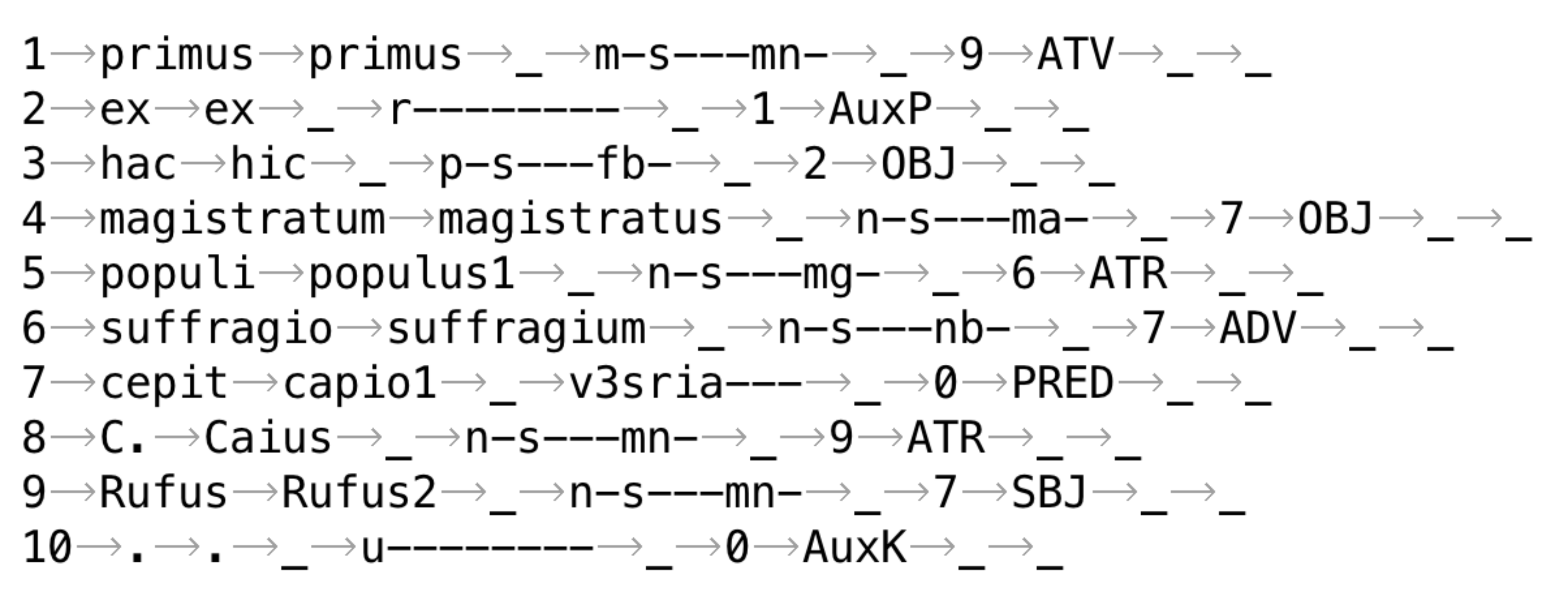
The underlying representation



```
<sentence id='5' document_id='urn:cts:latinLit:phi1348.abo012.perseus-lat1' subdoc='1.1-154.4' span=''>
  <word id='1' form='primus' lemma='primus' postag='m-s---mn-' relation='ATV' head='9'/>
  <word id='2' form='ex' lemma='ex' postag='r--------' relation='AuxP' head='1'/>
  <word id='3' form='hac' lemma='hic' postag='p-s---fb-' relation='OBJ' head='2'/>
  <word id='4' form='magistratum' lemma='magistratus' postag='n-s---ma-' relation='OBJ' head='7'/>
  <word id='5' form='populi' lemma='populus1' postag='n-s---mg-' relation='ATR' head='6'/>
  <word id='6' form='suffragio' lemma='suffragium' postag='n-s---nb-' relation='ADV' head='7'/>
  <word id='7' form='cepit' lemma='capio1' postag='v3sria---' relation='PRED' head='0'/>
  <word id='8' form='C.' lemma='Caius' postag='n-s---mn-' relation='ATR' head='9'/>
  <word id='9' form='Rufus' lemma='Rufus2' postag='n-s---mn-' relation='SBJ' head='7'/>
  <word id='10' form='.' lemma='.' postag='u--------' relation='AuxK' head='0'/>
</sentence>
```

an xml serialization

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→_→_
```

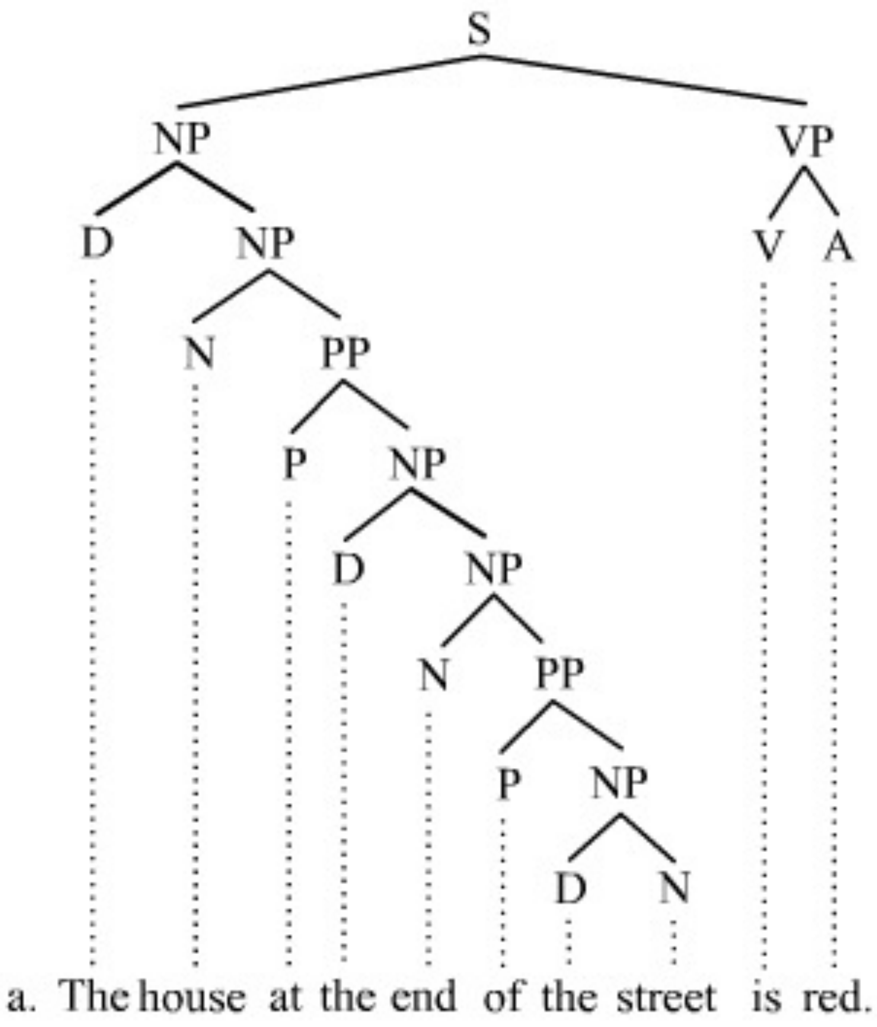
a plain text serialization

The linguistic formalism

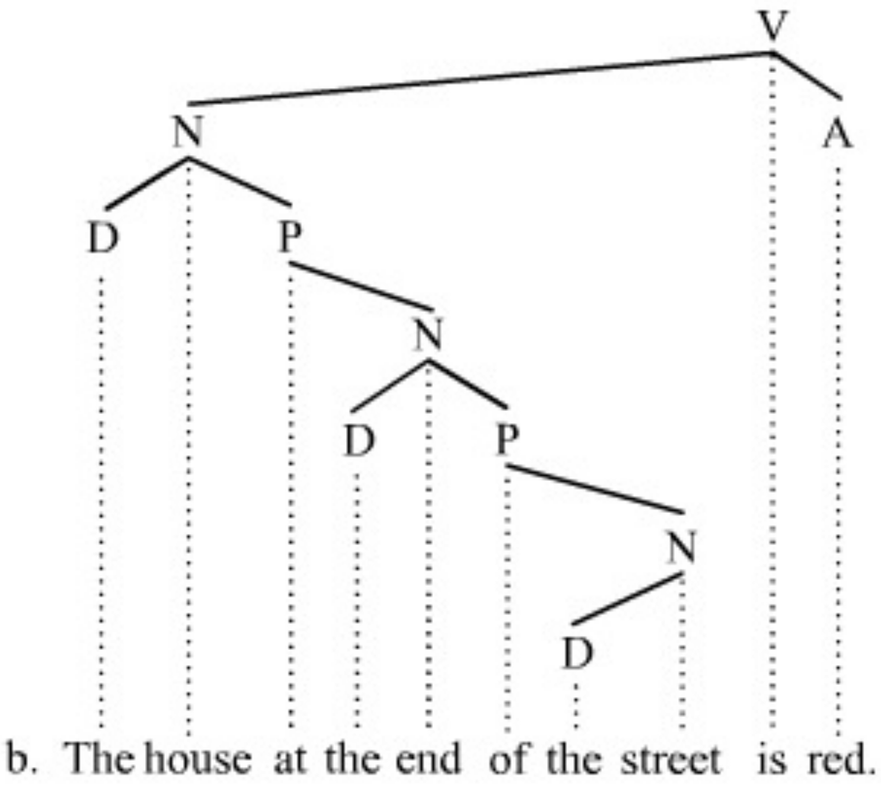
two main kinds of treebanks:

- constituency treebank
- dependency treebank

The linguistic formalism



Constituency structure



Dependency structure

(Treebank. In Wikipedia)

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/

Open Source treebanks for Ancient Greek and Latin

- Ancient Greek Dependency treebank
- PROIEL treebank
- Index Thomisticus treebank

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

The AGLD Treebank

The essential tools of the treebanker:

- The guidelines
- A tool for annotating

Learning by doing: treebanking Phaedrus

Canis per Fluvium Carnem Ferens

Amittit merito proprium qui alienum adpetit.
Canis, per fluvium carnem cum ferret, natans
lympharum in speculo vidit simulacrum suum,
aliamque praedam ab altero ferri putans
eripere voluit; verum decepta aviditas
et quem tenebat ore dimisit cibum,
nec quem petebat adeo potuit tangere.

Thanks for your attention!