



UNIVERSITÀ
CATTOLICA
del Sacro Cuore



Linked Open Treebanks

Latin treebanks in the LiLa Knowledge Base

Francesco Mambrini and Marco Passarotti

{francesco.mambrini}{marco.passarotti}@unicatt.it

SyntaxFest – TLT 2019 | Paris | August 29, 2019



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.

Introduction

- Latin treebanks

- The LiLa Knowledge Base

Populating LiLa

- Lemmas

- Treebanks

Potential use cases

Conclusions

- ▶ Latin Dependency Treebank (2006-): Classical Lat., prose and poetry, about 50k tokens;
- ▶ Index Thomisticus Treebank (2006-): Medieval Lat., only 1 author (Thomas Aquinas), about 400k tokens;
- ▶ PROIEL (2008): Late and Classical prose, transl. of NT (Jerome's *Vulgate*, 4th CE), plus other prose, about 250k;
- ▶ Late Latin Charter Treebank (2011-): 8th-9th century notary documents (charters) from Central Italy, about 250k.

- ▶ **Latin Dependency Treebank** (2006-): Classical Lat., prose and poetry, about 50k tokens;
- ▶ **Index Thomisticus Treebank** (2006-): Medieval Lat., only 1 author (Thomas Aquinas), about 400k tokens;
- ▶ PROIEL (2008): Late and Classical prose, transl. of NT (Jerome's *Vulgate*, 4th CE), plus other prose, about 250k;
- ▶ Late Latin Charter Treebank (2011-): 8th-9th century notary documents (charters) from Central Italy, about 250k.

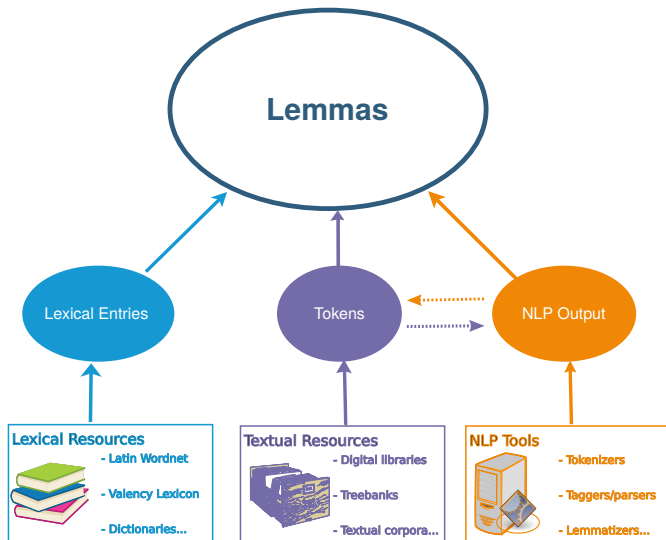
- ▶ **Latin Dependency Treebank** (2006-): Classical Lat., prose and poetry, about 50k tokens;
- ▶ **Index Thomisticus Treebank** (2006-): Medieval Lat., only 1 author (Thomas Aquinas), about 400k tokens;
- ▶ **PROIEL** (2008): Late and Classical prose, transl. of NT (Jerome's *Vulgate*, 4th CE), plus other prose, about 250k;
- ▶ Late Latin Charter Treebank (2011-): 8th-9th century notary documents (charters) from Central Italy, about 250k.

- ▶ Create a **Knowledge Base** of linguistic resources for Latin
 - ▶ corpora
 - ▶ lexicons
 - ▶ NLP tools
- ▶ Create common **vocabularies** to describe them
- ▶ Use the **LOD** paradigm

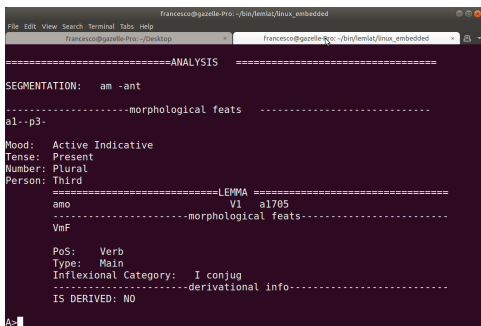


The lemma

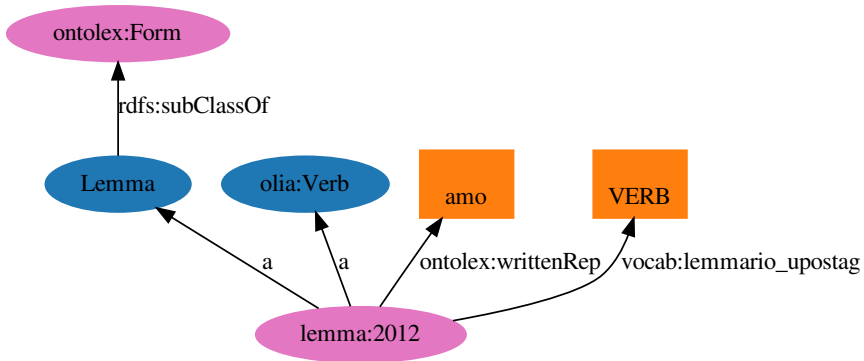
a gateway to Latin linguistic resources



- ▶ 43,432 lemmas from Georges, 1913-1918; *OLD* and Gradenwitz, 1904;
- ▶ 82,556 lemmas from Du Cange, 1883-1887;
- ▶ 26,250 lemmas from Forcellini, 1940.

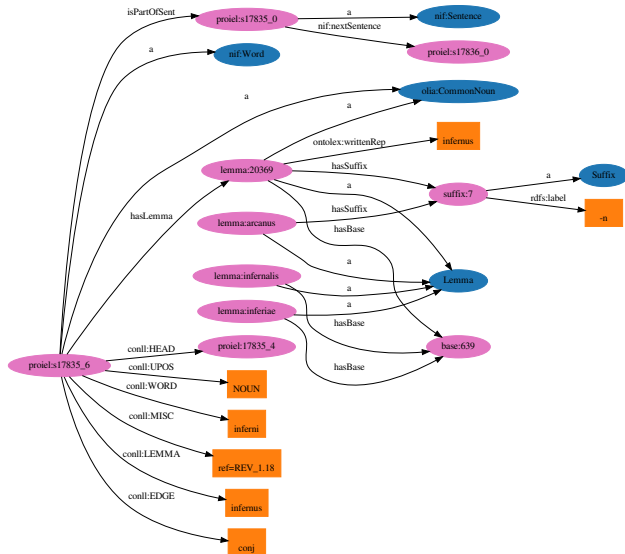


```
Francesco@gazelle-Pro: ~/bin/lemlat/linux_embedded
File Edit View Search Terminal Tabs Help
Francesco@gazelle-Pro: ~/bin/lemlat/linux_embedded
=====ANALYSIS=====
SEGMENTATION:  am -ant
-----morphological feats-----
a1-p3-
Mood:  Active Indicative
Tense: Present
Number: Plural
Person: Third
=====LEMMA=====
amo          V1  a1705
-----morphological feats-----
VmF
PoS:  Verb
Type: Main
Inflectional Category:  I conjug
-----derivational info-----
IS DERIVED: NO
A>
```

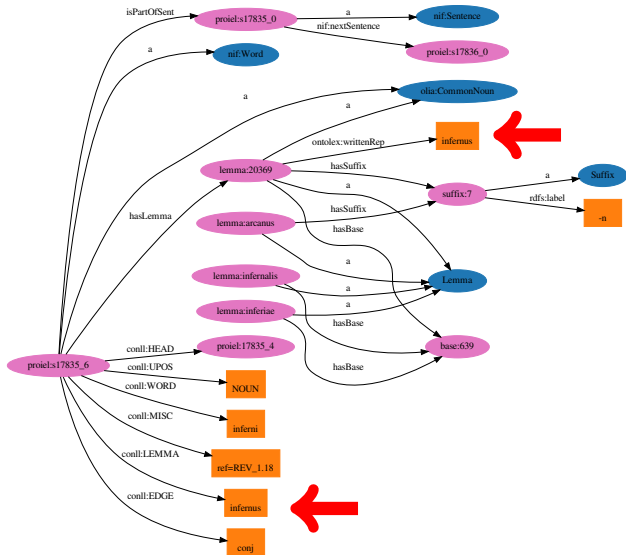



- ▶ start from a shallow **conversion** from TB format to RDF triples
- ▶ compare the **string** of the lemmatized token with the written representation(s) of a LEMLAT lemma
- ▶ **link** the token to the lemma via the `hasLemma` property

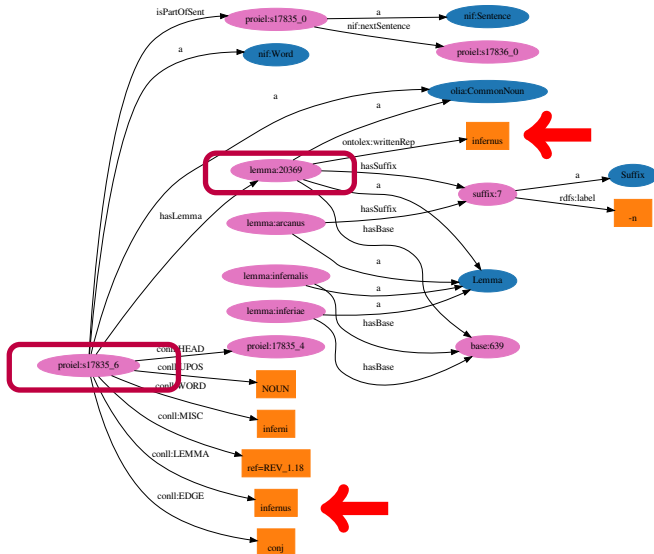
Linking corpora and lemmas



Linking corpora and lemmas



Linking corpora and lemmas



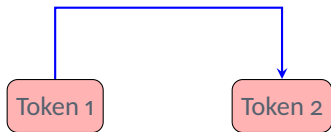
A wealth of interlinked information that can be queried!



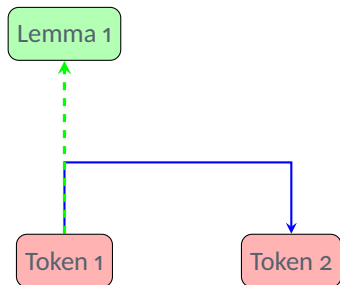
Token 1

Token 2

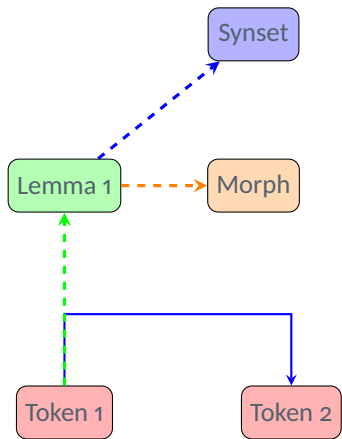
A wealth of interlinked information that can be queried!



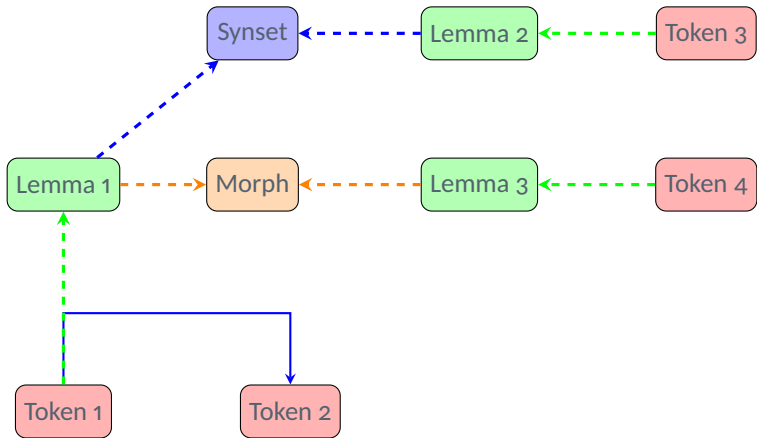
A wealth of interlinked information that can be queried!



A wealth of interlinked information that can be queried!



A wealth of interlinked information that can be queried!



Querying with SPARQL

All verbs that govern subjects formed with affix “-(t)or”



```
SELECT ?g ?headlab ?deplab WHERE {
  SERVICE <http://lila-erc.eu:3030/lemlat/sparql> {
    ?suff a lemlat_base:Suffix ;
        rdfs:label '-(t)or' .
    ?lemma lemlat_base:hasSuffix ?suff ;
        ontolex:writtenRep ?deplab . }
  GRAPH ?g {
    ?tok lemlat_base:hasLemma ?lemma ;
        conll:EDGE |'nsubj' ;
        conll:HEAD ?head .
    ?head conll:UPOS 'VERB' ;
        lemlat_base:hasLemma ?l .
  }
  SERVICE <http://lila-erc.eu:3030/lemlat/sparql> {
    ?l ontolex:writtenRep ?headlab . }
}
```

- (1) **gladiatores** audio **pugnare** mirifice
gladiators.ACC.PL hear.1SG fight.INF superbly
I hear that your gladiators fight superbly.
(Cic. Att.. 4.4a.2)

Wordcloud of results from the Index Thomisticus

“the Interpreter (of Aristotle, i.e. Averroes) says...”



- ▶ Language is complex! Morpho-syntactic description is not enough to capture all complexities
- ▶ LOD provide a way to link treebank annotation and information on other levels (semantics, derivational morphology...)
- ▶ a lexically based approach (using lemmas as hub node) is one way to do it!

- ▶ Language is complex! Morpho-syntactic description is not enough to capture all complexities
- ▶ LOD provide a way to link treebank annotation and information on other levels (semantics, derivational morphology...)
- ▶ a lexically based approach (using lemmas as hub node) is one way to do it!
- ▶ **but** (future works)...

- ▶ Language is complex! Morpho-syntactic description is not enough to capture all complexities
- ▶ LOD provide a way to link treebank annotation and information on other levels (semantics, derivational morphology...)
- ▶ a lexically based approach (using lemmas as hub node) is one way to do it!
- ▶ **but** (future works)...
 - ▶ we need to **harmonize** the tagsets (ontologies)

- ▶ Language is complex! Morpho-syntactic description is not enough to capture all complexities
- ▶ LOD provide a way to link treebank annotation and information on other levels (semantics, derivational morphology...)
- ▶ a lexically based approach (using lemmas as hub node) is one way to do it!
- ▶ **but** (future works)...
 - ▶ we need to **harmonize** the tagsets (ontologies)
 - ▶ we need to find sustainable, **scalable solutions** together with the projects that own and maintain the resources

Thanks!

Get in touch



The LiLa Team

Università Cattolica del Sacro Cuore
CIRCSE Research Centre



info@lila-erc.eu



<https://github.com/CIRCSE>



<https://lila-erc.eu>



[@ERC_LiLa](https://twitter.com/ERC_LiLa)



Largo Gemelli 1, 20123 Milan, Italy



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.