UNIVERSITÀ
CATTOLICA
del Sacro Cuore

From Treebanks to Linked Open Data

The LiLa Project

Francesco Mambrini
francesco.mambrini@unicatt.it

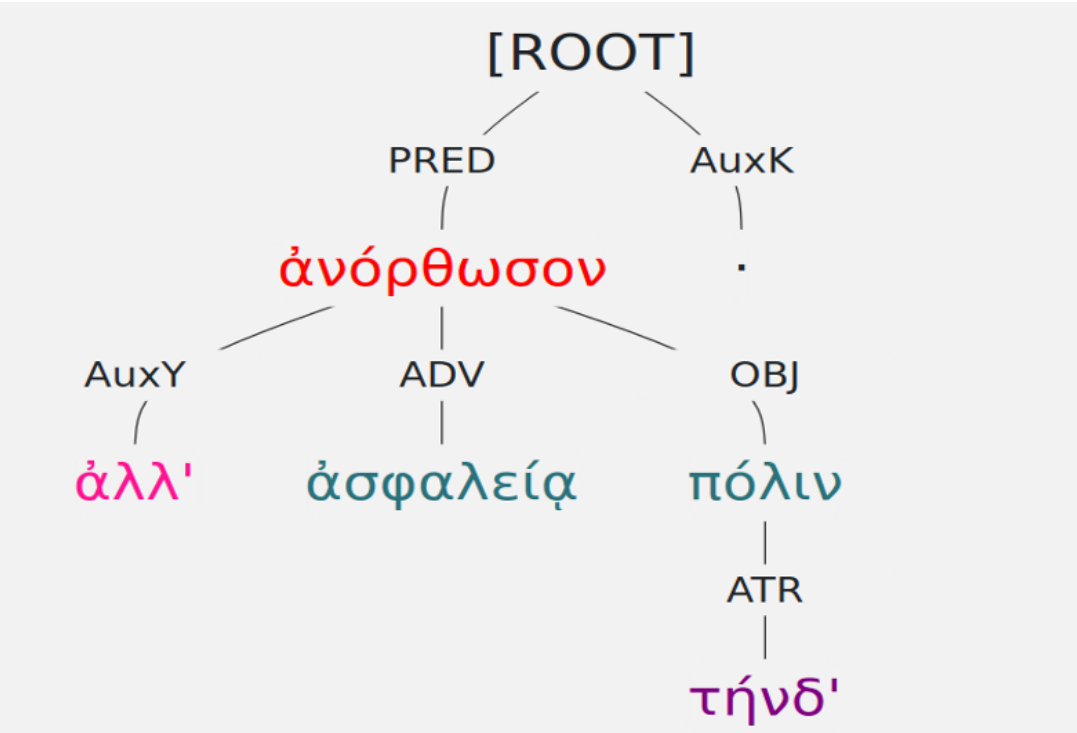
BridgeClassics | Humboldt Universiät Berlin | 7 Dec, 2021

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.

Table of Contents

. . . but what is a treebank, again?

- annotated corpus
- split into sentences
- with word-by-word annotation on:
  - morphology
  - syntax
  - perhaps more



4 treebanks of Latin

- 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;
- UDante (2021): complete annotation of the (5) Latin works by Dante Alighieri using the UD schema, about 56k

Universal Dependencies (UD)

http://universaldependencies.org/

- a shared framework
- more than 200 treebanks
- more than 100 languages
- both universal and language-specific annotation

Motivation
Resources for Latin

- Textual Resources
  - Digital Libraries (Perseus), Collections of Texts (LASLA, ALIM, Musisque Deoque, Patrologia Latina, PHI, Library of Latin Texts etc.), Treebanks
- Lexical Resources
  - Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange, TLL, OLD, TTLF, DMLBS & DB of Latin Dictionaries c/o Brepols)
- NLP Tools
  - Morphological Analysers (LEMLAT, Words, LatMor), PoS Taggers (TreeTagger, Collatinus, UDPipe), Dependency Parsers (UDPipe), CLTK

Motivation
Resources for Latin

- Textual Resources
  - Digital Libraries (Perseus), Collections of Texts (LASLA, ALIM, Musisque Deoque, Patrologia Latina, PHI, Library of Latin Texts etc.), Treebanks
- Lexical Resources
  - Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange, TLL, OLD, TTLF, DMLBS & DB of Latin Dictionaries c/o Brepols)
- NLP Tools
  - Morphological Analysers (LEMLAT, Words, LatMor), PoS Taggers (TreeTagger, Collatinus, UDPipe), Dependency Parsers (UDPipe), CLTK

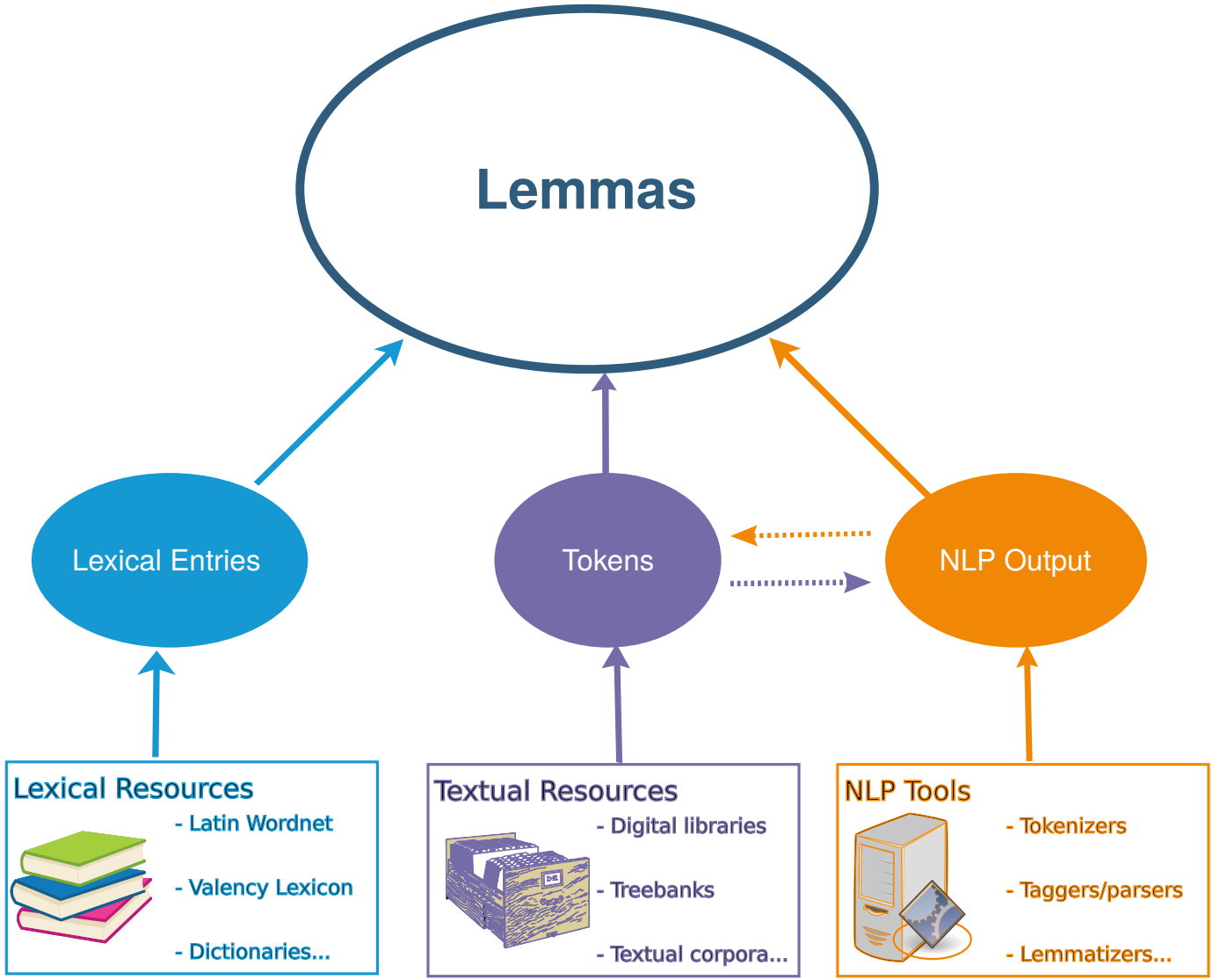
Scattered and Unconnected

Aims

- 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



LiLa: Overview

Resources connected and upcoming connections

- **Textual Resources**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] **UDante: ca. 55,000 tokens**
  - [x] *Liber Abbaci*, *Chapter VIII*: ca. 30,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexical Resources**
  - [x] Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
  - [x] Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
  - [x] LatinAffectus: ca. 4,000 entries
  - [x] Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
  - [x] Latin WordNet: ca. 1,000 manually checked entries
  - [x] Latin Vallex 2.0: Valency Lexicon
  - [x] Lewis & Short Dictionary
  - [ ] Lemma Embeddings
- **NLP tools**
  - [x] Lemma Bank: ca. 200,000 lemmas
- **TOTAL: approximately 15 million triples**

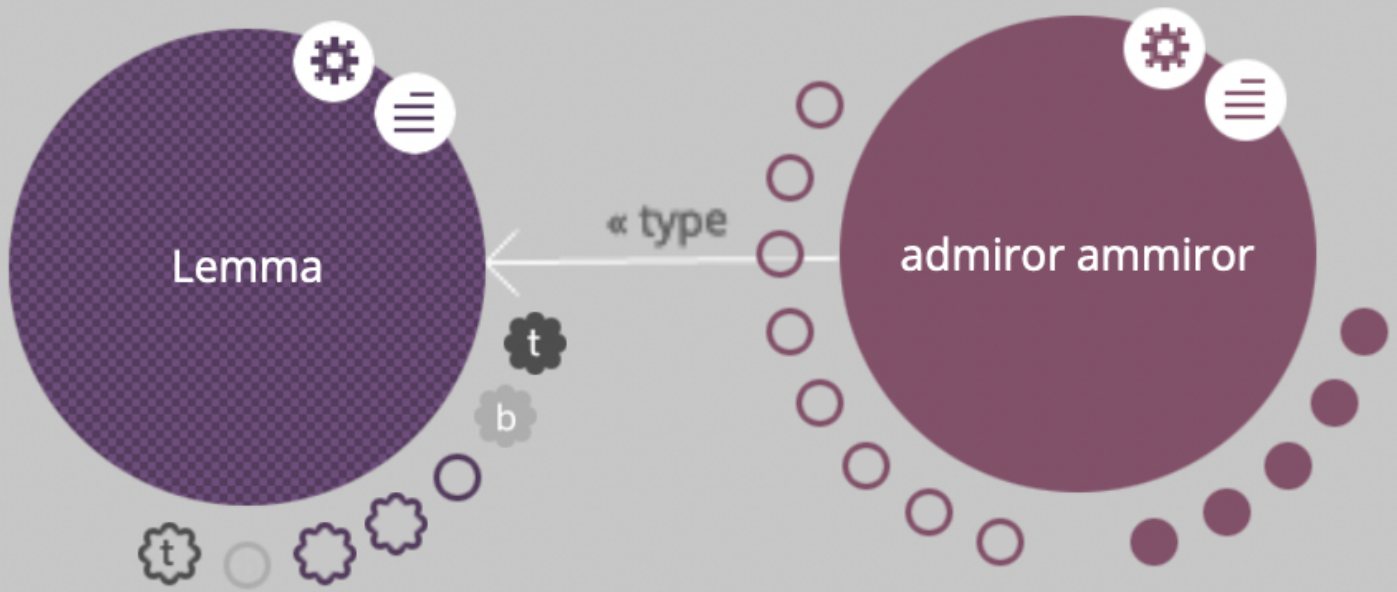
LiLa: Overview

Resources connected and upcoming connections

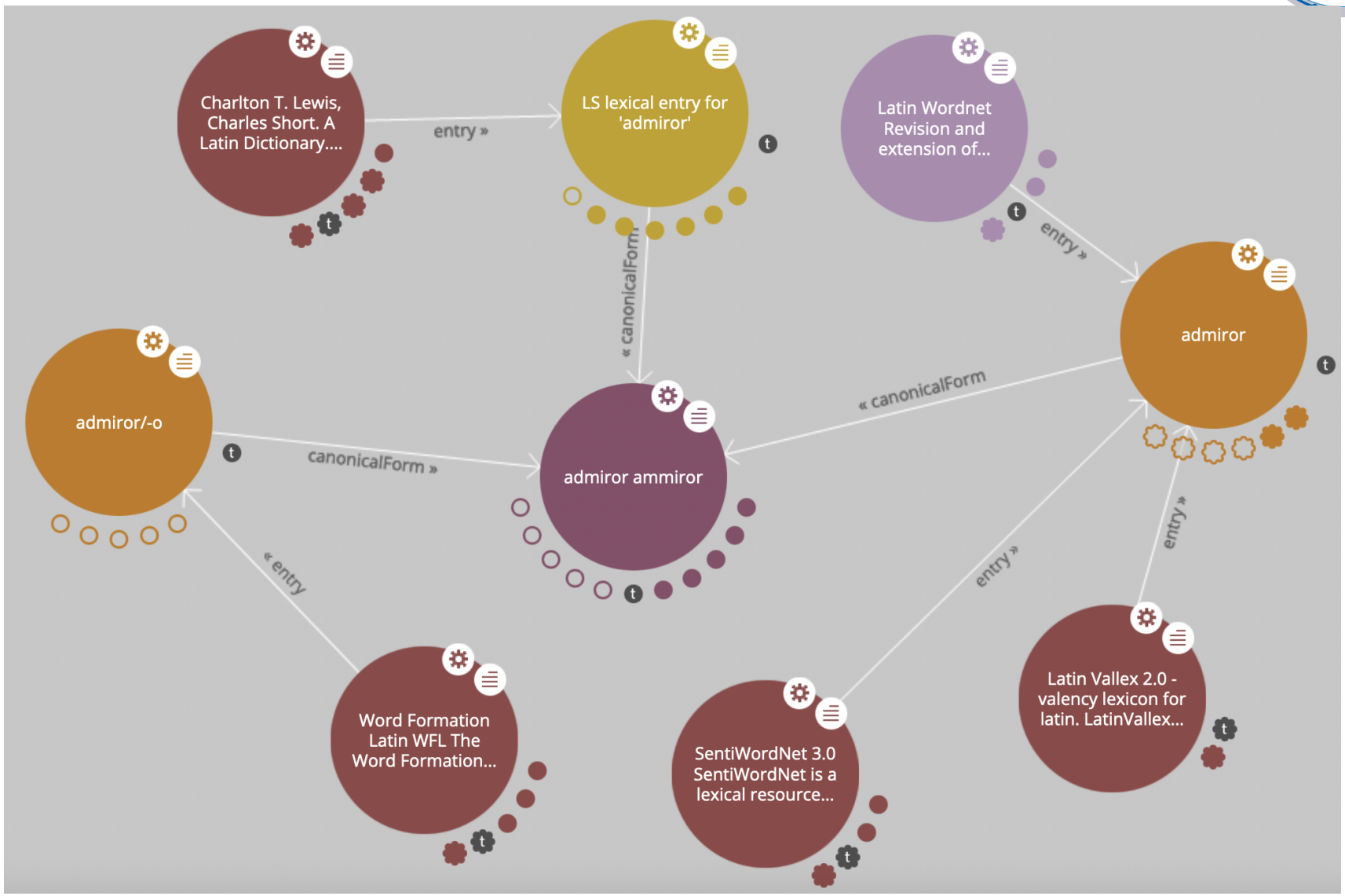
- **Textual Resources**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] **UDante: ca. 55,000 tokens**
  - [x] *Liber Abbaci*, *Chapter VIII*: ca. 30,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexical Resources**
  - [x] Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
  - [x] Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
  - [x] LatinAffectus: ca. 4,000 entries
  - [x] Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
  - [x] Latin WordNet: ca. 1,000 manually checked entries
  - [x] Latin Vallex 2.0: Valency Lexicon
  - [x] Lewis & Short Dictionary
  - [ ] Lemma Embeddings
- **NLP tools**
  - [x] Lemma Bank: ca. 200,000 lemmas
- **TOTAL: approximately 15 million triples**

Meet a lemma!

http://lila-erc.eu/data/id/lemma/87541



Lemmas and Lexical Entries

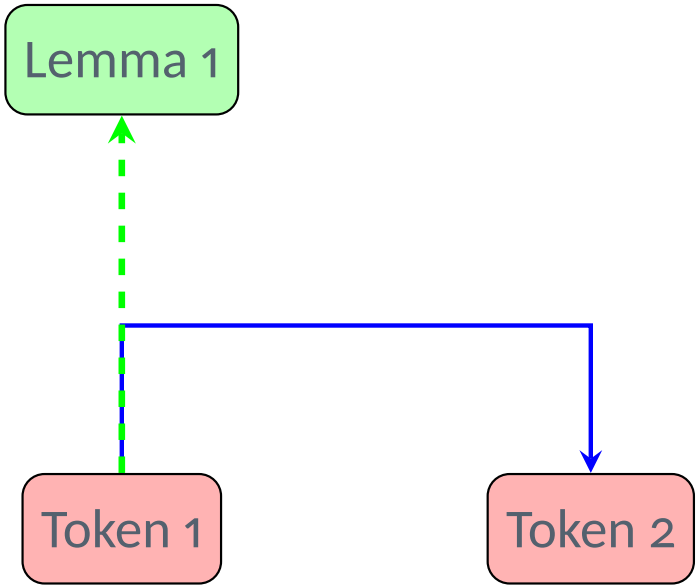


A wealth of interlinked information
that can be queried!

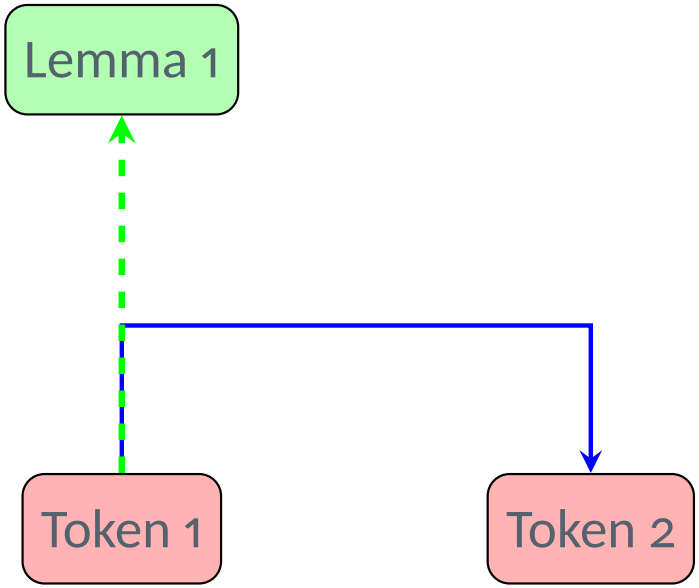
Token 1

Token 2

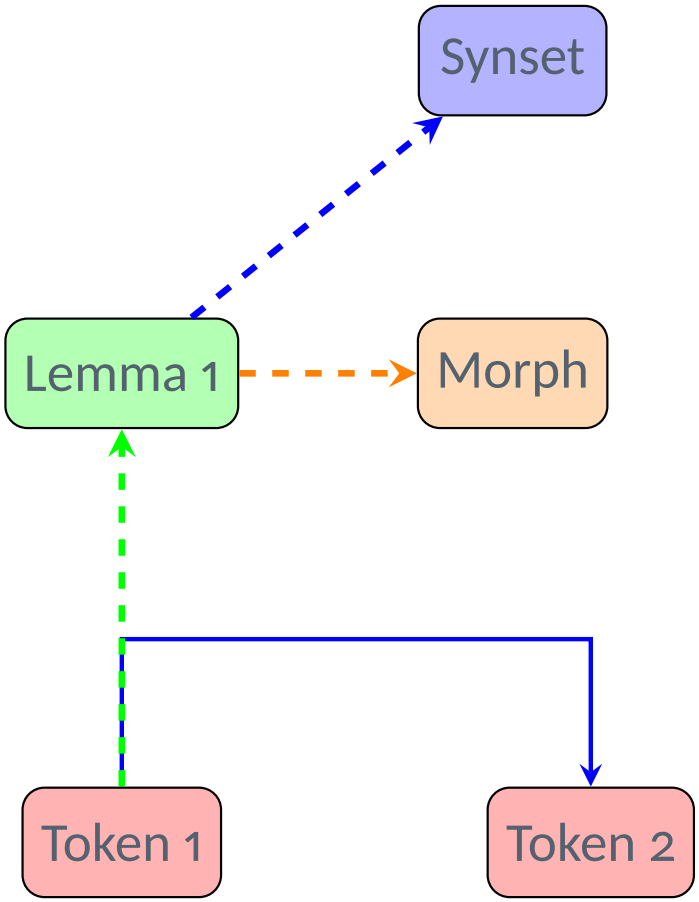
A wealth of interlinked information
that can be queried!



A wealth of interlinked information
that can be queried!

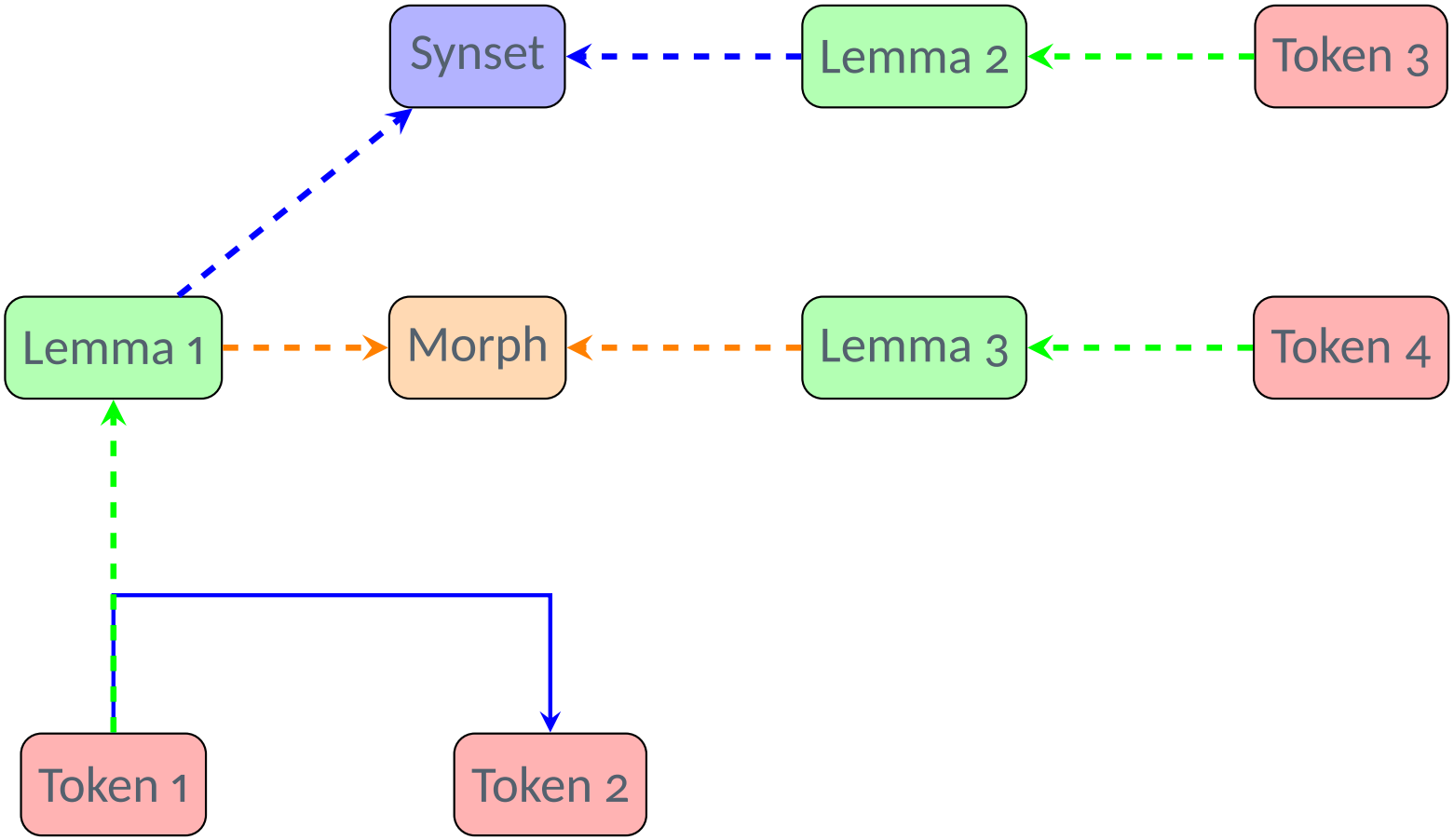


A wealth of interlinked information
that can be queried!



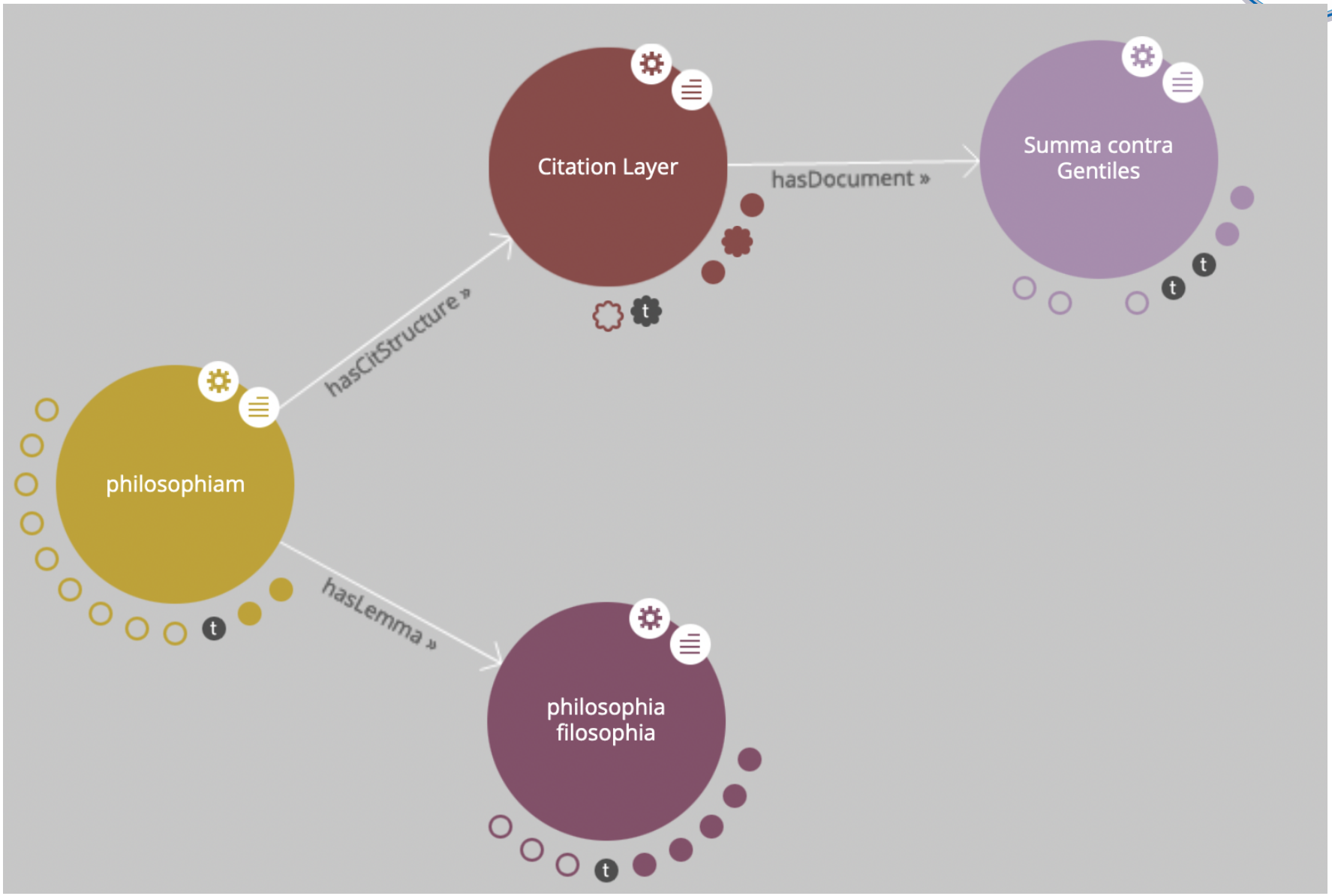
A wealth of interlinked information

that can be queried!



Linked tokens in actions

http://lila-erc.eu/data/corpora/ITTB/id/token/005.SCG*LB1.CP--++1.N.1-1.2-4W4

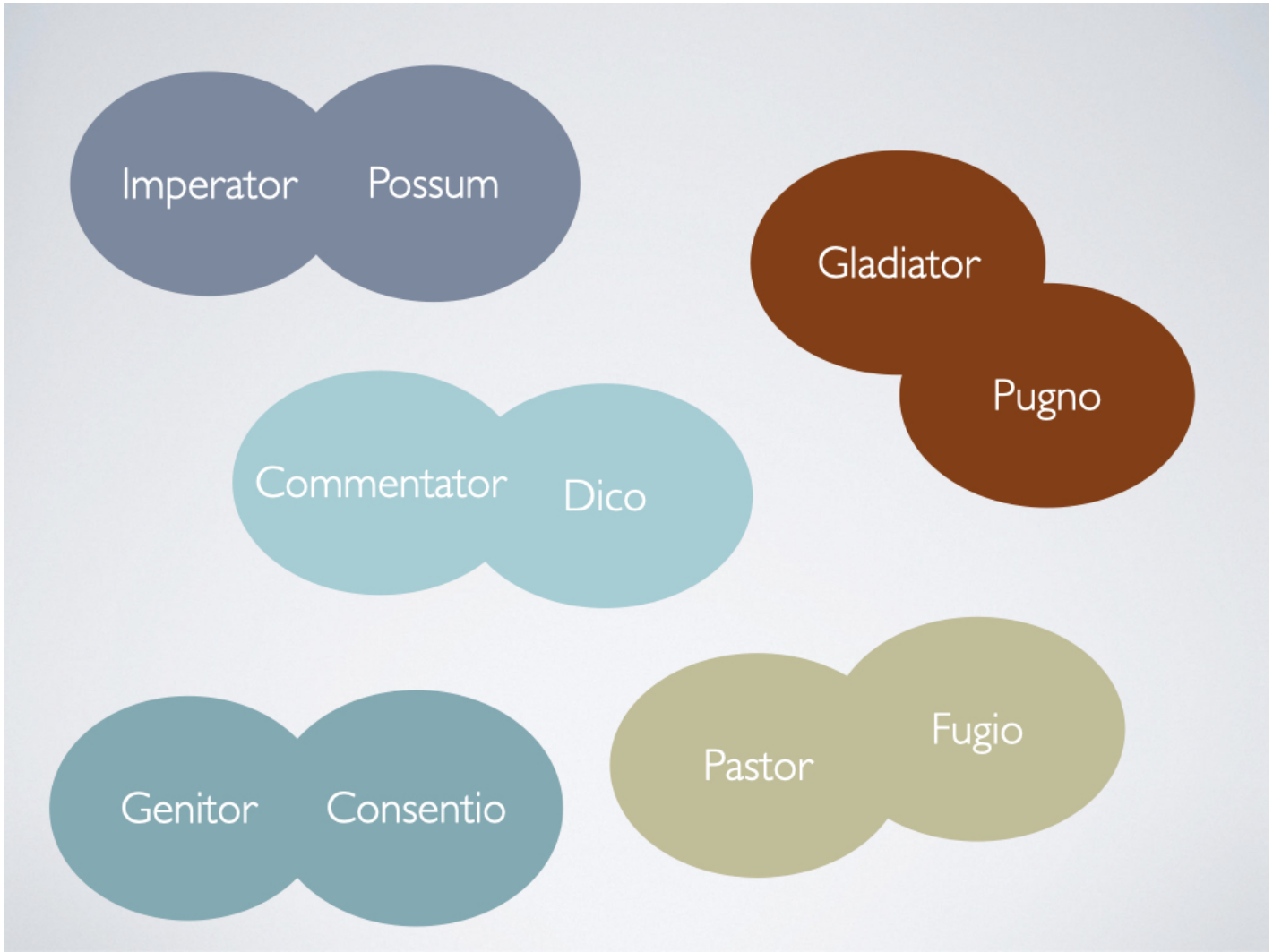


Querying with SPARQL

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 . }
}
```

Some interesting couplets



Wordcloud of results from the Index Thomisticus

"the Interpreter (of Aristotle, i.e. Averroes) says…"

Conclusions
and 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!

Conclusions
and 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)…

Conclusions
and 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)

Conclusions
and 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)
  - 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

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.