The LiLa Knowledge Base in a Nutshell

Eleonora Litta

LiLa Tutorial

Mobility & Humanities: Digital Experiences and Tools (2021)

1 December, Università degli studi di Padova

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.

Overview

Introduction and fundamentals

- LiLa: mission and architecture

LiLa now!

- Lemma Bank and Lexical Resources
- Textual Resources
- To sum up

Tutorials

- What, How and Who
- Text Linker

Research question
State of affairs

We have built and collected (for Latin and other languages):

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

ERC Consolidator Grant 2018-2023

A collection of multifarious, interoperable linguistic resources described with the same vocabulary for knowledge description (by using common data categories and ontologies)

Interlinking as a Form of Interaction

CLARIN
Common Language Resources and Technology Infrastructure

Infrastructure

Interoperability

The Linked Data Principles

FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus) - FINDABLE
- Use HTTP URIs to allow people (and machines) to look up things - ACCESSIBLE
- Use web standards to represent/query (meta)data, such as RDF and SPARQL - INTEROPERABLE
- Include links to other URIs - INTEROPERABLE
- metadata are released with a clear and accessible data usage license - REUSABLE

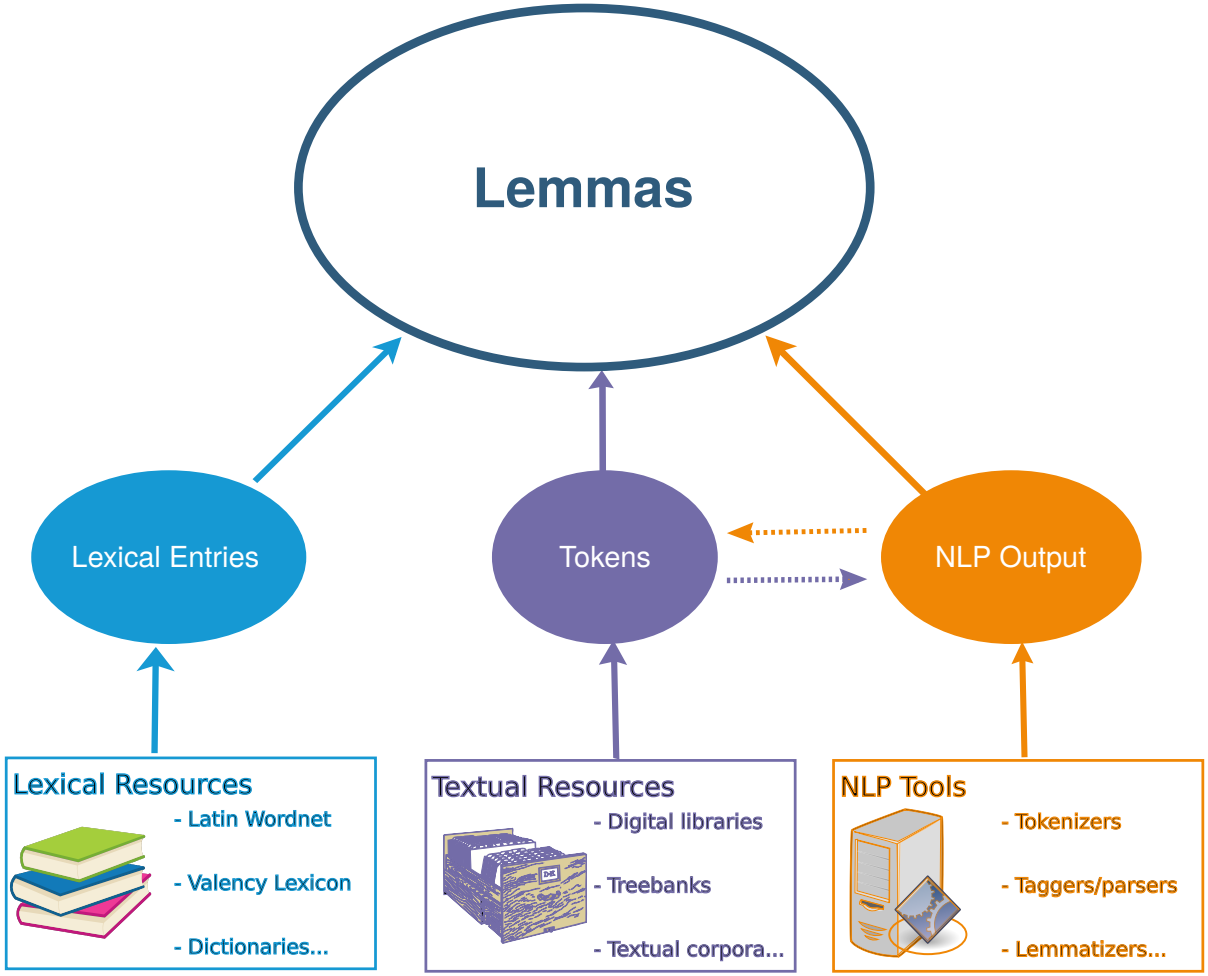
LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)
- **Object properties**: ways in which classes and individuals can be related to one another: RDF triples.
  Labels from a restricted vocabulary of knowledge description:
  `hasLemma`, `hasPoS`

Each component of the ontology is uniquely identified through a URI.

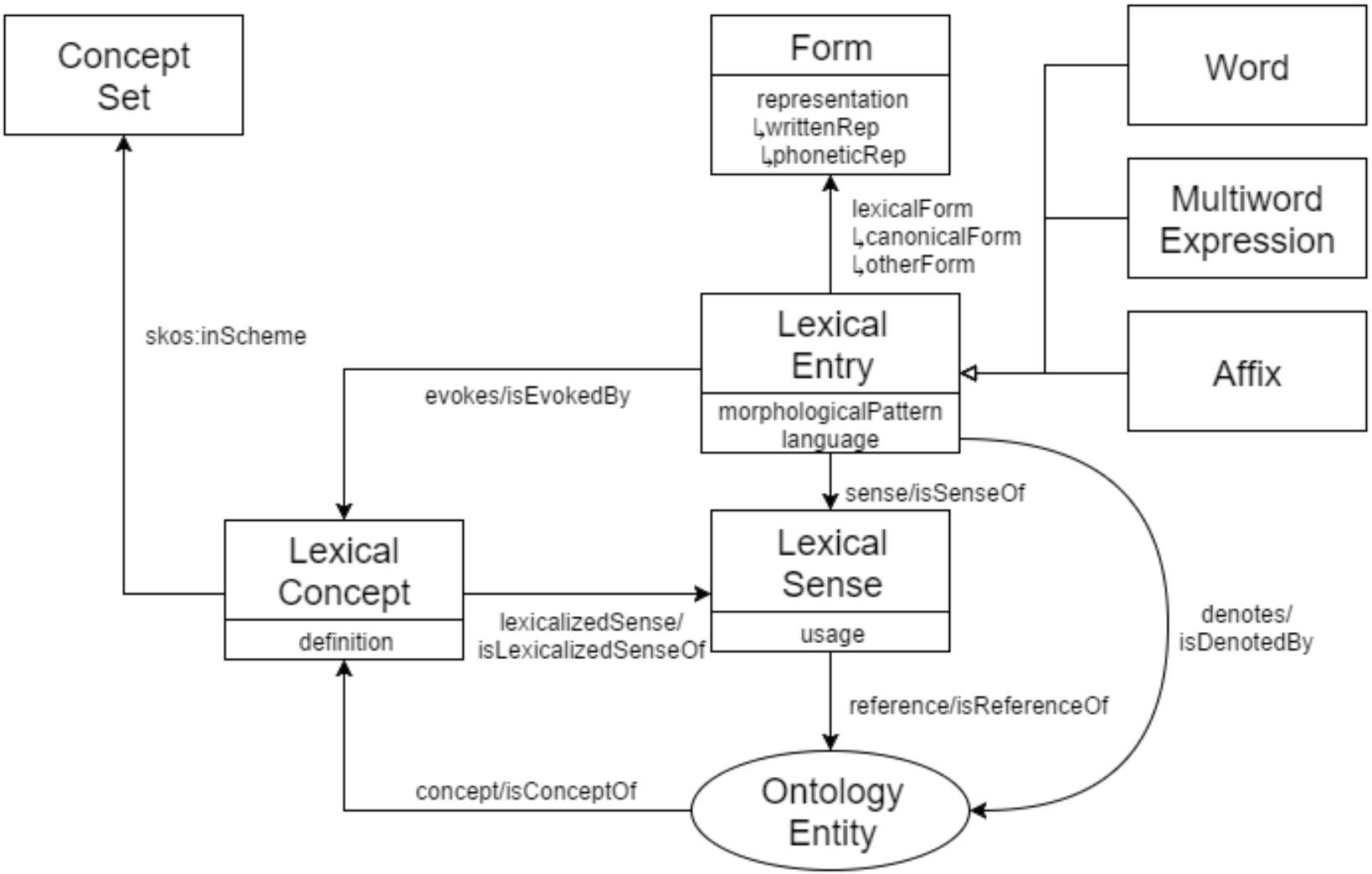
LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

A *de facto* W3C standard for publishing lexical data as LLOD



Lemma Bank and Lexical Resources

The core of LiLa. Source: LEMLAT (extended)

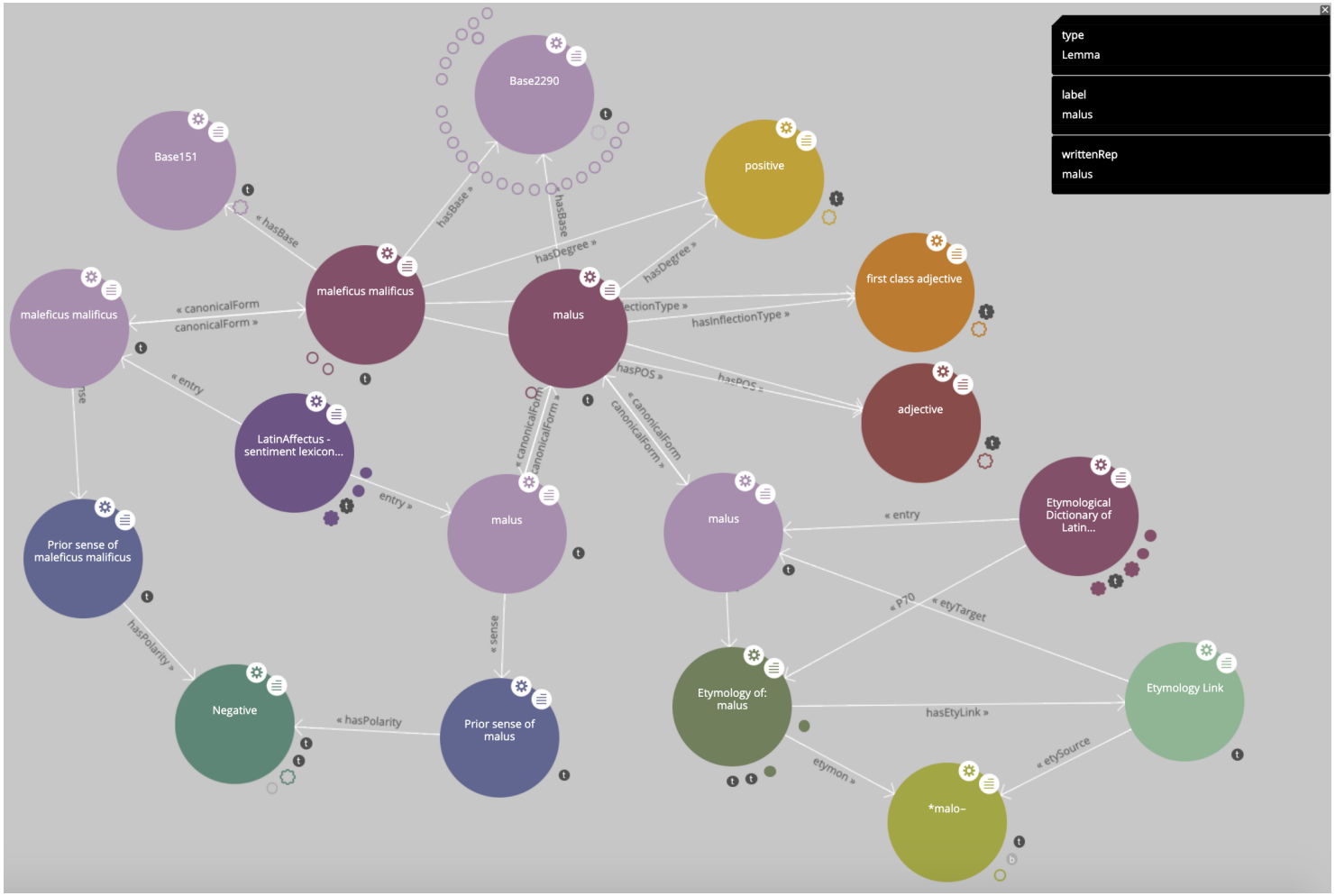
Lemma *admiror* 'to admire, to respect'

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

- Lemma Bank
- A derivational lexicon (Word Formation Latin)
- A polarity lexicon (LatinAffectus)
- An etymological dictionary (De Vaan)
- A Valency Lexicon (Latin Vallex)
- A manually checked subset of the Latin WordNet
- A bilingual Latin-English Dictionary (Lewis & Short)

LiLa: Structure. All together now!

Lemma Bank, Derivational Morphology, Etymology and Polarity

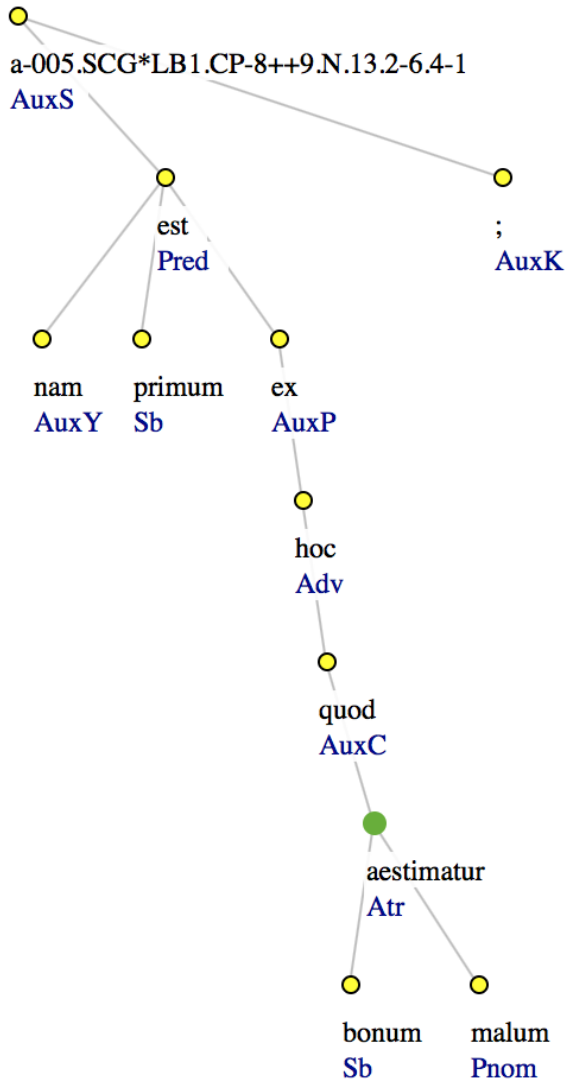


(Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (original scheme)

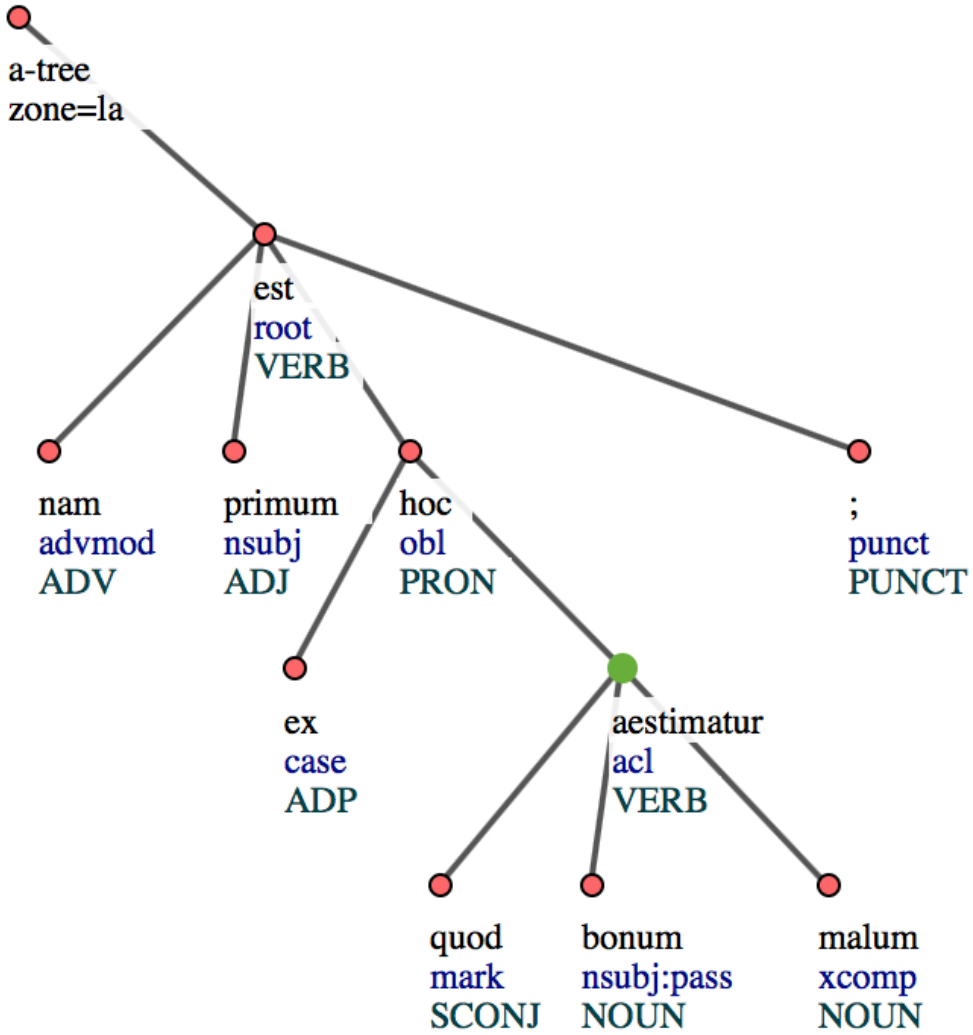
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

for the first arises because the good ***is judged*** *to be evil*; (Trans. Anton C. Pegis)



(Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (UD scheme)

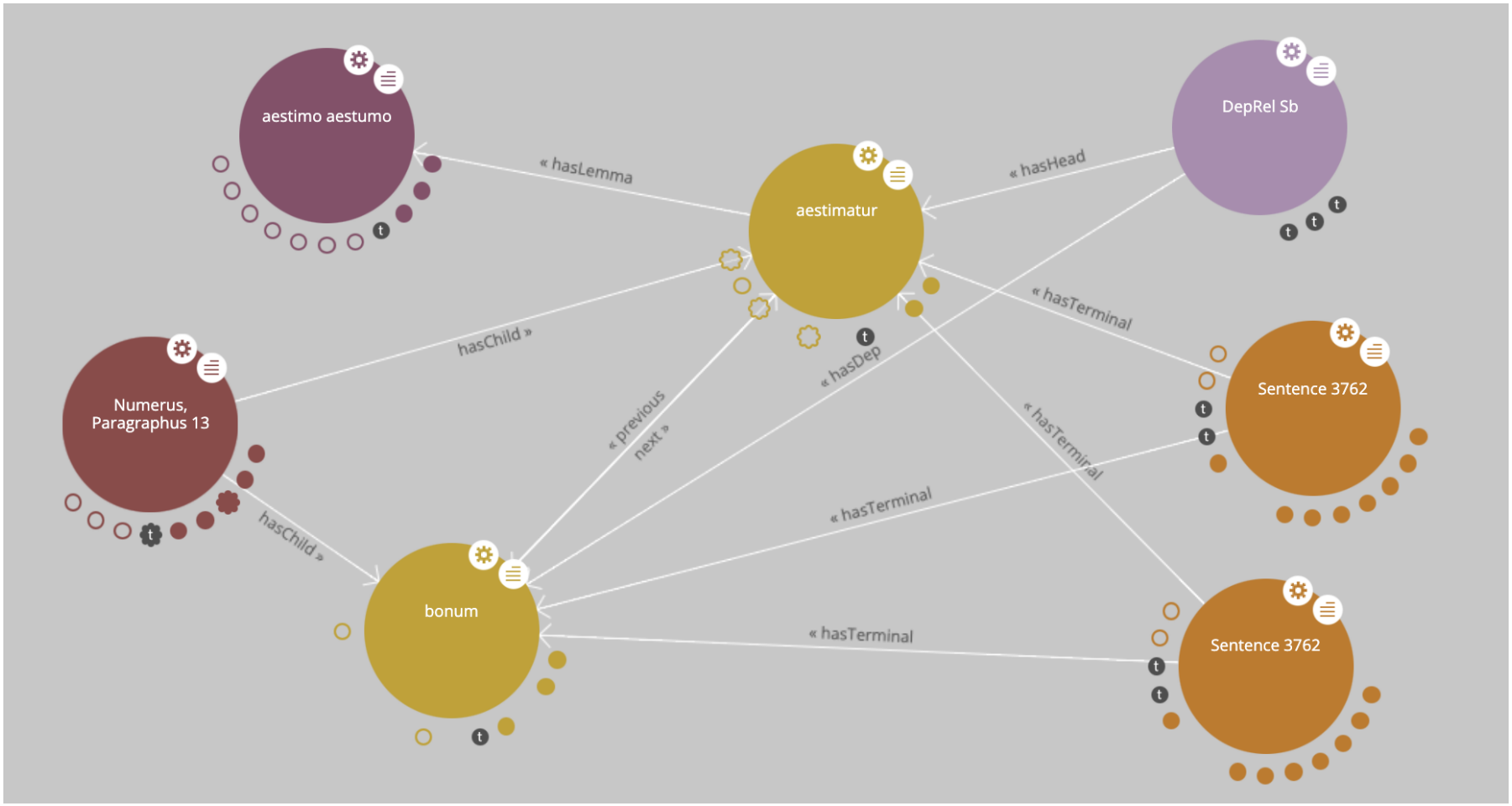


Token *aestimatur*

http://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.2-6.4-1W8

Texts, tokens, relations and lemmas

Phenomena and noumena



LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] Dante Search (700th death anniversary): ca. 46,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexica**
  - [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. 2,300 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
  - [ ] Lewis & Short Dictionary
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 13 million triples**

Query Interface, Triplestore

- Query interface
- Triplestore

Linguistic Resources. Corpora

- Index Thomisticus Treebank
- UDante
- Querolus sive Aulularia
- Liber Abbaci

Linguistic Resources. Lexica

- Word Formation Latin
- Etymological Dictionary of Latin & the Other Italic Languages
- LatinAffectus
- Index Graecorum Vocabulorum in Linguam Latinam
- Latin WordNet
- Latin Vallex 2.0

- **WHAT**: to show the workflow we employ to connect a linguistic resource for Latin to the LiLa Knowledge Base, and to demonstrate the way in which LiLa can be queried.
- **HOW**: to teach participants how to perform automatic lemmatisation and RDF-isation (i.e. format conversion) in order to link a Latin text to LiLa.
- **WHO-1**: anyone who wishes to publish Latin texts on the web.
- **WHO-2**: anyone interested in the different aspects involved in the construction of a Linguistic Linked Open Data knowledge base.

LiLa reflects the annotation granularity of the resources it connects

No data enrichment or further analysis is performed
…but we can help you to enrich your (meta)data

LiLa: Requirements

Connecting resources in the Knowledge Base

To enter the LiLa Knowledge Base, a textual/lexical resource must be:

- Lemmatised
- Part-of-Speech tagged (ideally, using the Universal Dependencies tagset)
- Online!

TextLinker

Welcome screen

Figure: LiLa's Text Linker

TextLinker
Processed output

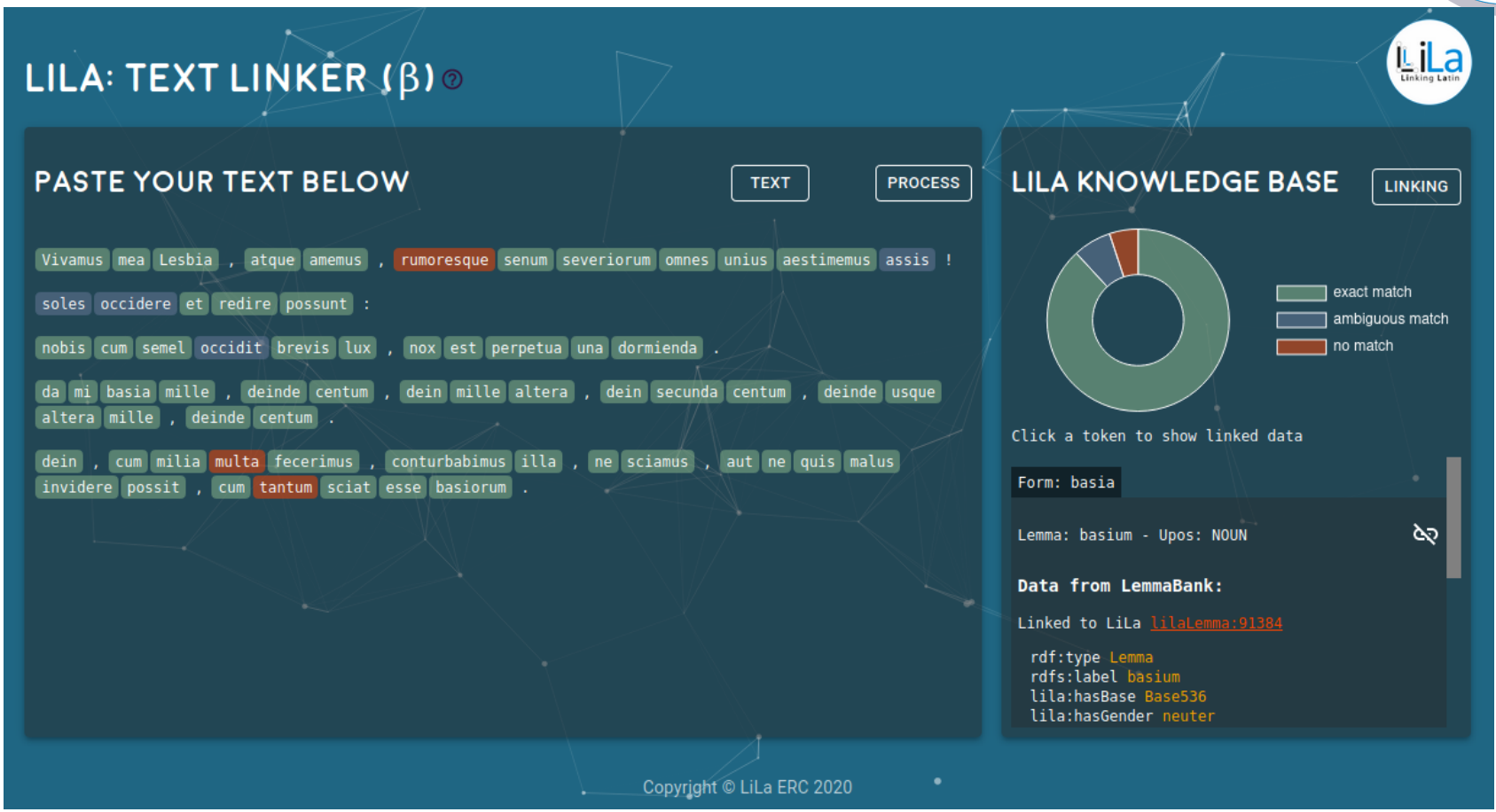


Figure: Text processed against the LiLa Knowledge Base

http://lila-erc.eu:8080/LiLaTextLinker/

Useful Links

- LiLa Lemma Bank query interface: `https://lila-erc.eu/query/`
- LiLa SPARQL interface: `https://lila-erc.eu/sparql/`
- Repository for the tutorial: `https://github.com/CIRCSE/Tutorials/tree/main/MobiLab`
- Collaborative notebook: `https://tinyurl.com/2p83buww`
- TextLinker (**!!Beta Version!!**): `http://lila-erc.eu:8080/LiLaTextLinker/`
- Repository of SPARQL queries: `https://github.com/CIRCSE/SPARQL-queries`

Thanks!
Get in touch

LiLa: Linking Latin
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.