



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore



# The LiLa Knowledge Base in a Nutshell

The LiLa Team

LiLa Tutorial

Conference *Language, Data and Knowledge* (LDK 2021)

1 September, Zaragoza, Spain



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 and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

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

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

- ▶ Textual Resources

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

- ▶ Textual Resources
- ▶ Lexical Resources

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

- ▶ Textual Resources
- ▶ Lexical Resources
- ▶ NLP Tools

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



Infrastructure



Interoperability

# The Linked Data Principles

...just to be FAIR



# The Linked Data Principles

...just to be FAIR



- ▶ Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)

# The Linked Data Principles

...just to be FAIR



- ▶ Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- ▶ Use HTTP URIs to allow people (and machines) to look up things

# The Linked Data Principles

...just to be FAIR



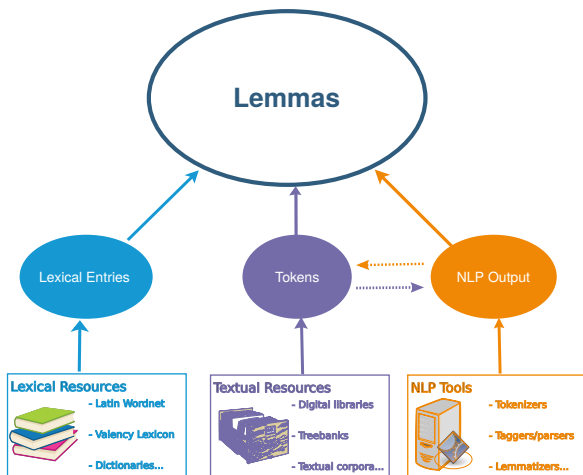
- ▶ Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- ▶ Use HTTP URIs to allow people (and machines) to look up things
- ▶ Use web standards to represent/query (meta)data, such as RDF and SPARQL

# The Linked Data Principles

...just to be FAIR



- ▶ Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- ▶ Use HTTP URIs to allow people (and machines) to look up things
- ▶ Use web standards to represent/query (meta)data, such as RDF and SPARQL
- ▶ Include links to other URIs



**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:

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

- ▶ Lemmatised

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

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

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

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

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

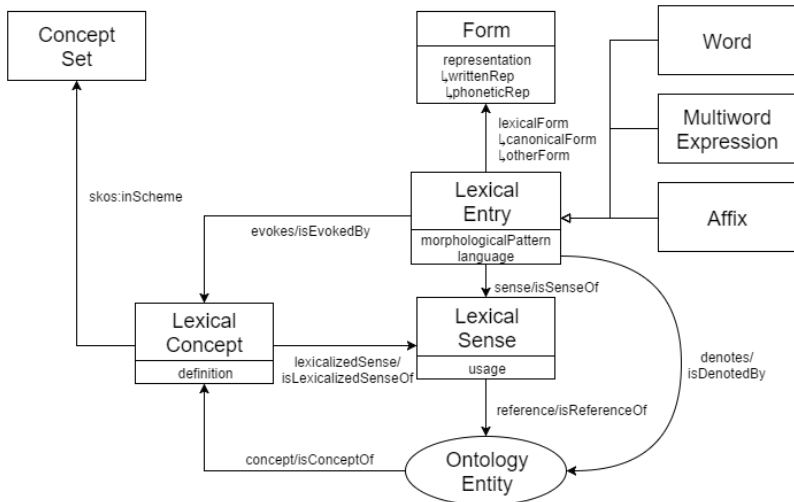
## Tutorial

What, How and Who

Programme & Communication Tools

# LiLa and Ontolex Lemon

A de facto W3C standard for publishing lexical data as LLOD



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

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

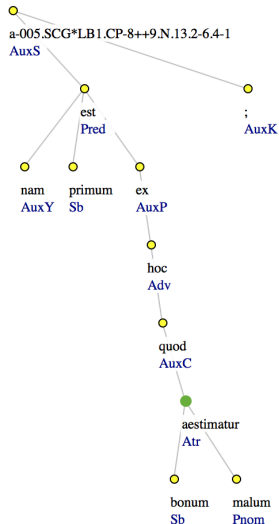


# (Annotated) Corpora in LiLa

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

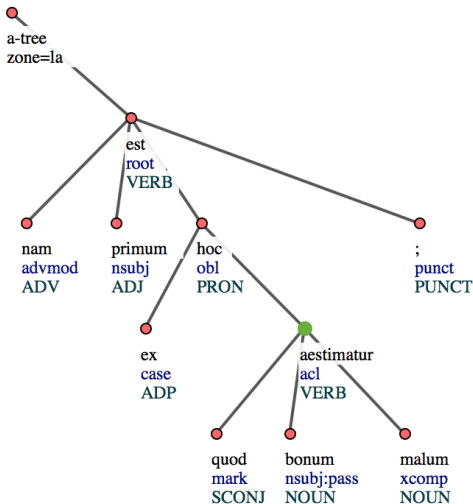
*nam primum est ex hoc  
quod bonum **aestimatur**  
malum;* (IT-TB: SCG, lib. 1,  
cap. 89, n. 13)

*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`

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools



Figure: LiLa's Text Linker

**LILA: TEXT LINKER (β)**

PASTE YOUR TEXT BELOW

TEXT PROCESS

Vivamus mea Lesbia, atque amemus, rumoresque senum severiorum omnes unius aestimemus assis!


soles occidere et redire possunt:

nobis cum semel occidit brevis lux, nox est perpetua una dormienda.

da mi basia mille, deinde centum, dein mille altera, dein secunda centum, deinde usque altera mille, deinde centum.

dein, cum milia multa fecerimus, conturbabimus illa, ne sciamus, aut ne quis malus invidere possit, cum tantum sciat esse basiorum.

**LILA KNOWLEDGE BASE LINKING**



Click a token to show linked data

Form: basia

Lemma: basium - Upod: NOUN

Data from LemmaBank:

Linked to LiLa [LiLaLemma:91394](#)

rdf:type Lemma  
rdfs:label basium  
lila:hasBase Base536  
lila:hasGender neuter

Copyright © LiLa ERC 2020

Figure: Text processed against the LiLa Knowledge Base

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

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

## ► Corpora

- ✓ Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
- ✓ Dante Search (700th death anniversary): ca. 46,000 tokens
- ✓ *Querolus sive Aulularia*: ca. 17,000 tokens
- PROIEL and LLCT treebanks
- Computational Historical Semantics, LASLA and CroALa Corpora

## ► Lexica

- ✓ Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
- ✓ Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
- ✓ LatinAffectus: ca. 2,300 entries
- ✓ Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
- ✓ Latin WordNet: ca. 1,000 manually checked entries
- ✓ Latin Vallex 2.0: Valency Lexicon
- Lewis & Short Dictionary

## ► NLP tools

- ✓ LEMLAT (lemma bank): ca. 150,000 lemmas

## ► TOTAL: approximately 13 million triples



## Query Interface, Triplestore and Linker

- ▶ Query interface; Triplestore
- ▶ Linker

## Linguistic Resources. Corpora

- ▶ Index Thomisticus Treebank
- ▶ Dante Search
- ▶ *Querolus sive Aulularia*

## 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

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

- ▶ **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.

- ▶ **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.

- ▶ **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.

- ▶ **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.

## Introduction and fundamentals

LiLa: mission and architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

## Tutorial

What, How and Who

Programme & Communication Tools

## Programme (CET)

- ▶ 9:00-10:30: Introduction, quiz
- ▶ 10:30-10:45: 🍷 break
- ▶ 10:45-13:00: Hands-on work (test on Horace, *Carmina* 1), Q&A

☕ other breaks agreed together as the activity progresses

## Tools

- ▶ Chat/Raise hand on Teams: for quick help
- ▶ Google Drive: for collaborative note-taking:  
<https://tinyurl.com/abhpnwzu>



- ▶ Repository for the tutorial:  
<https://github.com/CIRCSE/Tutorials/tree/main/LDK21>
- ▶ Collaborative notebook: <https://tinyurl.com/abhpnwzu>
- ▶ TextLinker (!!Beta Version!!):  
<http://lila-erc.eu:8080/LiLaTextLinker/>
- ▶ Repository of SPARQL queries:  
<https://github.com/CIRCSE/SPARQL-queries>
- ▶ LiLa Lemma Bank query interface: <https://lila-erc.eu/query/>
- ▶ LiLa SPARQL interface: <https://lila-erc.eu/sparql/>

## LiLa: Linking Latin

Università Cattolica del Sacro Cuore  
CIRCSE Research Centre



[info@lila-erc.eu](mailto: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.