



UNIVERSITÀ
CATTOLICA
del Sacro Cuore



Interlinking through Lemmas

The LiLa Knowledge Base
of Interlinked Linguistic Resources for Latin

Marco Passarotti, Francesco Mambrini

Ewa Wipszycka's Late Antique Seminar
March 18, 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.

Introduction and fundamentals

- LiLa: mission and architecture

- The LiLa Lemma Bank

LiLa now! lexica and corpora

- Lexical Resources

- Textual Resources

- Text Linker

- To sum up

Introduction and fundamentals

LiLa: mission and architecture

The LiLa Lemma Bank

LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

To sum up

Research question

State of affairs



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

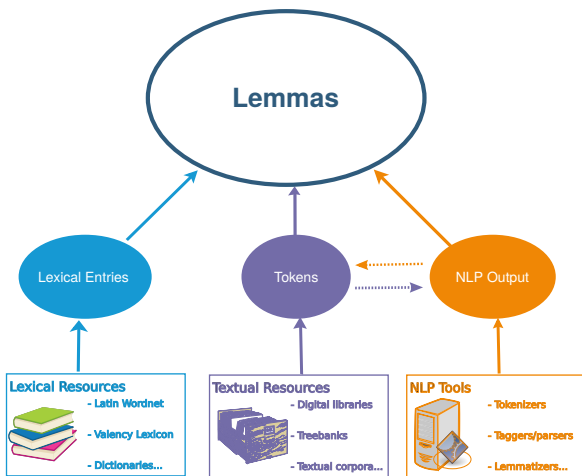
...just to be FAIR



- ▶ 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

- ▶ 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

- ▶ 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

The LiLa Lemma Bank

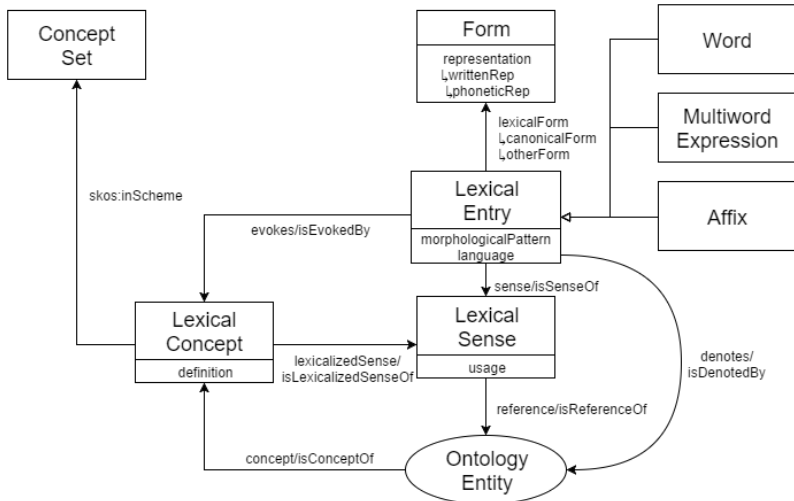
LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

To sum up



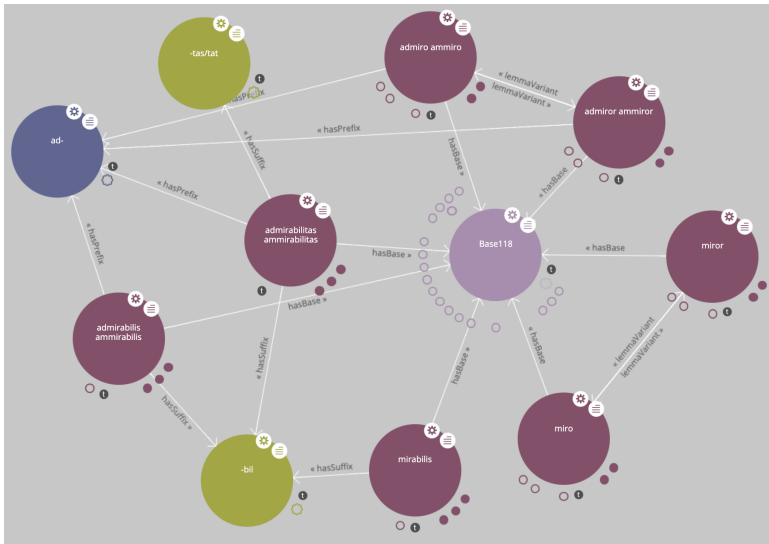
Lemma *admiror* 'to admire, to respect'

<https://lila-erc.eu/data/id/lemma/87541>

- ▶ WFL: directed tree-graphs indicating derivational path of each lemma (procedural)
- ▶ New Approach: Construction Morphology (declarative), words analysed in their internal structure
- ▶ WFL in LiLa:
 - ▶ Three classes of objects:
 1. Lemma
 2. Prefix and Suffix
 3. Base (connectors between lemmas of the same WF family)
 - ▶ Connected by three relationships:
 1. hasPrefix
 2. hasSuffix
 3. hasBase

Derivational Morphology

Source: *Word Formation Latin* (CIRCSE Research Centre)



Introduction and fundamentals

LiLa: mission and architecture

The LiLa Lemma Bank

LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

To sum up

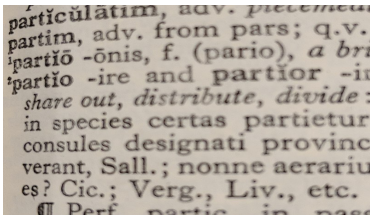


Figure: Retro-digitized dictionaries

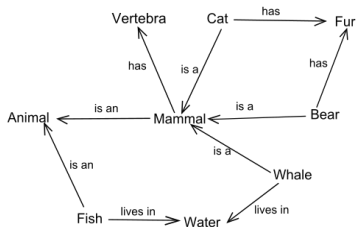
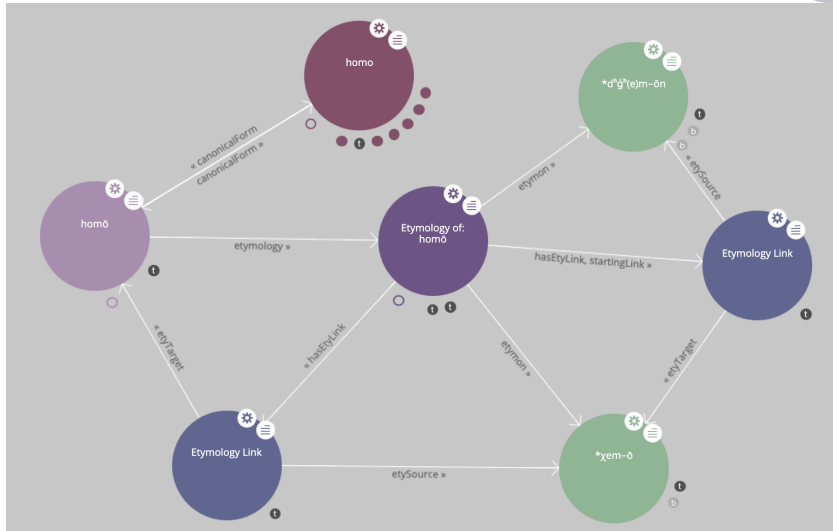


Figure: Digital-born lexicons

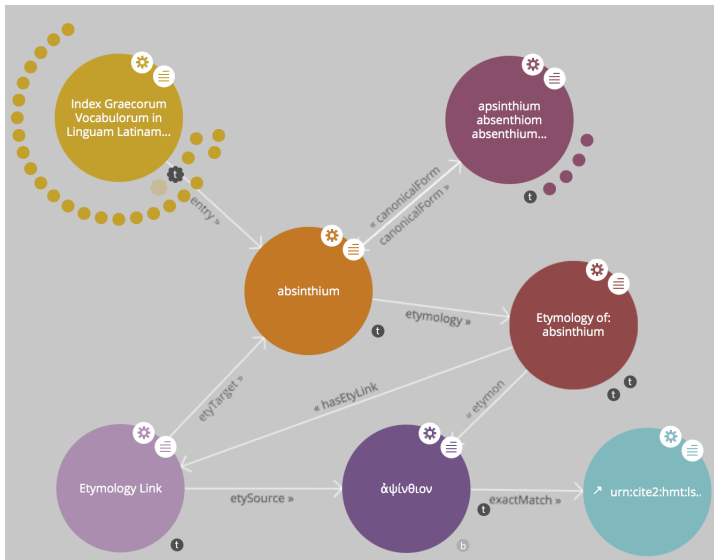
Etymology

Source: *Etymological dictionary of Latin and the other Italic Languages* (De Vaan, 2008)



Etymology

Source: *Index Graecorum Vocabulorum in Linguam Latinam* (Saalfeld, 1874)



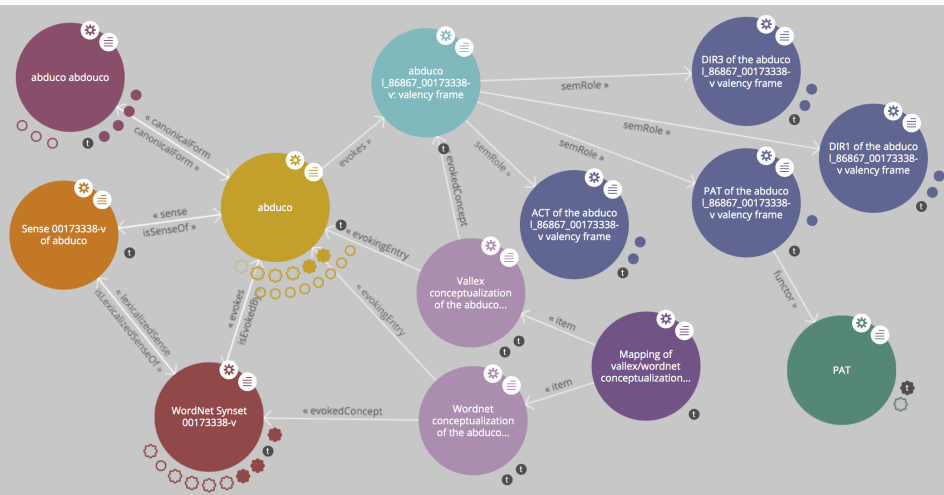
Latin WordNet

Lexical Entries, Senses and Concepts



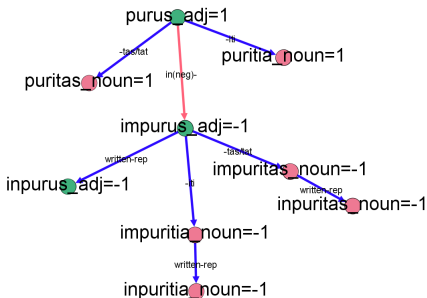
Latin WordNet and Latin VALLEX

Mapping Senses and Frames



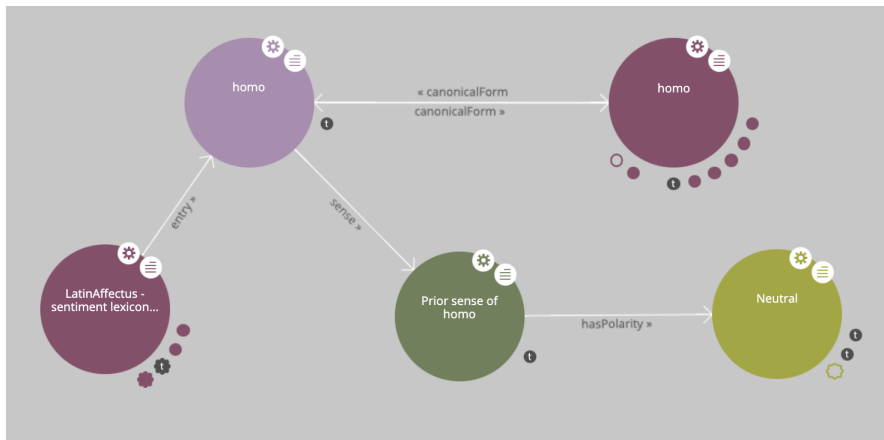
List of 2,437 nouns and adjectives associated to their **out-of-context sentiment score**: from -1 (very negative, e.g. *abominatio*) to +1 (very positive, e.g. *amor*)

- ▶ **Gold Standard**: manually created by 2 Latin language and culture experts + 1 supervisor
- ▶ **Silver Standard**: automatically created by deriving new entries from the Gold Standard



Polarity

Source: *Latin Affectus* (CIRCSE Research Centre)



Introduction and fundamentals

LiLa: mission and architecture

The LiLa Lemma Bank

LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

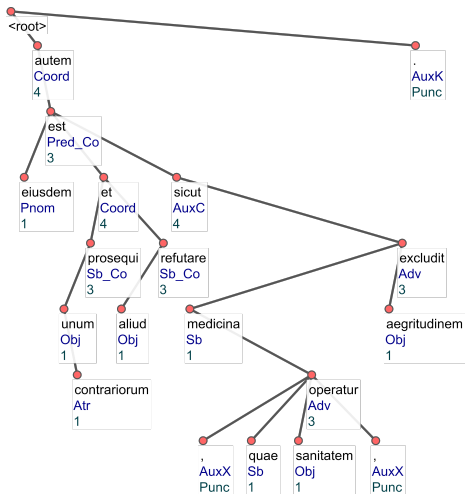
To sum up

(Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (CIRCSE Research Centre): Dependency trees

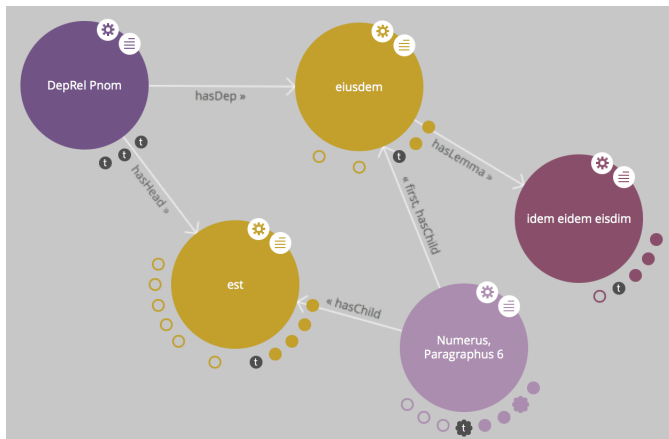
eiusdem autem est unum contrariorum prosequi et aliud refutare sicut medicina , quae sanitatem operatur , aegritudinem excludit . (IT-TB: SCG, lib. 1, cap. 1, n. 6)

Now it **belongs to the same thing** to pursue one contrary and to remove the other: thus medicine, which effects health, removes sickness. (Trans. Laurence Shapcote)



Texts, tokens, relations and lemmas

Phenomena and noumena



Introduction and fundamentals

LiLa: mission and architecture

The LiLa Lemma Bank

LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

To sum up



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. |

Copyright © LiLa ERC 2020

Figure: LiLa's Text Linker


LILA: TEXT LINKER (β)

PASTE YOUR TEXT BELOW

TEXT PROCESS

LILA KNOWLEDGE BASE LINKING

Vivamus nea Lesbia , atque amemus , rumoresque senum severiorum omnes unius aestinemus 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 sciāt esse basiorum .



- exact match
- ambiguous match
- no match

Click a token to show linked data

Form: basia

Lemma: basium - Upos: NOUN

Data from LemmaBank:

Linked to LiLa [lilaLemma:91384](#)

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

Copyright © LiLa ERC 2020

Figure: Text processed against the LiLa Knowledge Base

Introduction and fundamentals

LiLa: mission and architecture

The LiLa Lemma Bank

LiLa now! lexica and corpora

Lexical Resources

Textual Resources

Text Linker

To sum up

▶ Corpora

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

- ▶ <https://lila-erc.eu/query/>; <https://lila-erc.eu/sparql/>
- ▶ <http://lila-erc.eu:8080/LiLaTextLinker/>

Linguistic Resources. Corpora

- ▶ <https://lila-erc.eu/data/corpora/ITTB/id/corpus>
- ▶ <https://lila-erc.eu/data/corpora/DanteSearch/id/corpus>
- ▶ <https://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia>

Linguistic Resources. Lexica

- ▶ <https://lila-erc.eu/data/lexicalResources/BrilledDL/Lexicon>
- ▶ <https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon>
- ▶ <https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon>
- ▶ <http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon>
- ▶ <https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon>

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.