



UNIVERSITÀ
CATTOLICA
del Sacro Cuore



Words and Classes. Branches and Links

Interlinking (Latin) Resources
in the Linguistic Linked Open Data World
through the LiLa Knowledge Base

Marco Passarotti

Fred Jelinek Seminar Series
April 26, 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

LiLa now!

LiLa Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up

Introduction and fundamentals

LiLa: mission and architecture

LiLa now!

LiLa Lemma Bank and 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



- ▶ 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

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

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)



Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)



- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)



- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- ▶ Structural Interoperability: HTTP, URIs, RDF

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

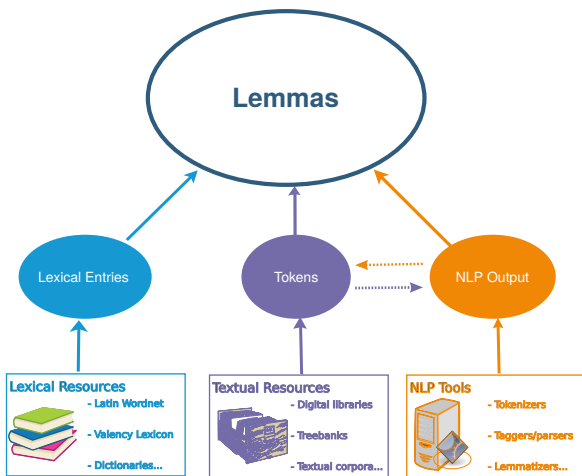


- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- ▶ Structural Interoperability: HTTP, URIs, RDF
- ▶ Conceptual Interoperability: common ontologies to understand how to use the URIs

- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- ▶ Structural Interoperability: HTTP, URIs, RDF
- ▶ Conceptual Interoperability: common ontologies to understand how to use the URIs
- ▶ Federation: to combine information from physically separated repositories

- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- ▶ Structural Interoperability: HTTP, URIs, RDF
- ▶ Conceptual Interoperability: common ontologies to understand how to use the URIs
- ▶ Federation: to combine information from physically separated repositories
- ▶ Dynamicity: to provide access to the most recent version of a resource

- ▶ Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- ▶ Structural Interoperability: HTTP, URIs, RDF
- ▶ Conceptual Interoperability: common ontologies to understand how to use the URIs
- ▶ Federation: to combine information from physically separated repositories
- ▶ Dynamicity: to provide access to the most recent version of a resource
- ▶ Ecosystem: maintained by a large and active community with common tools and practices



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

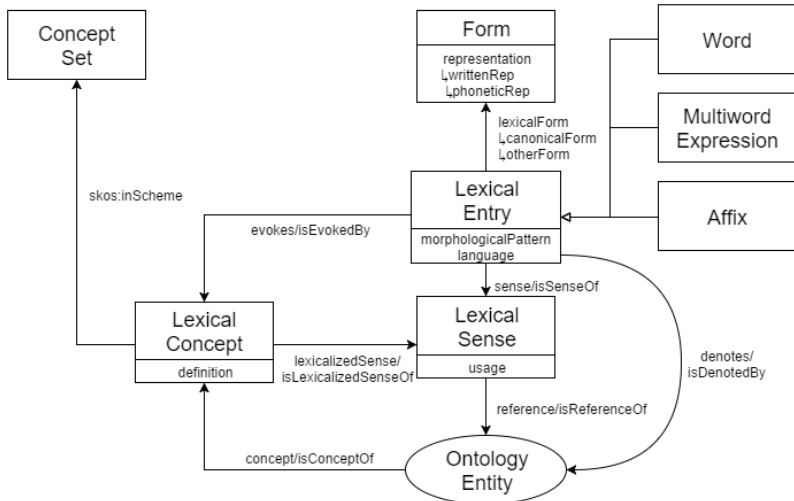
LiLa now!

LiLa Lemma Bank and Lexical Resources

Textual Resources

Text Linker

To sum up



Lemma *admiror* 'to admire, to respect'

<https://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!

LiLa Lemma Bank and Lexical Resources

Textual Resources

Text Linker

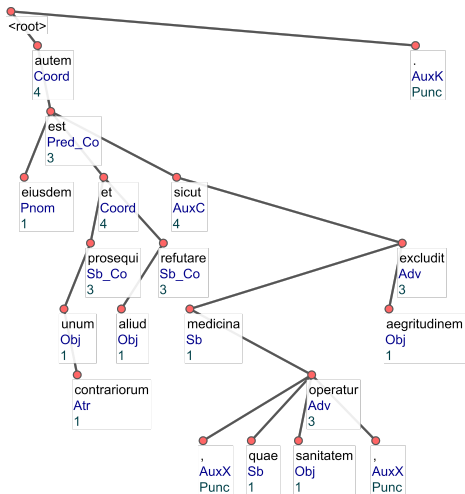
To sum up

(Annotated) Corpora in LiLa

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

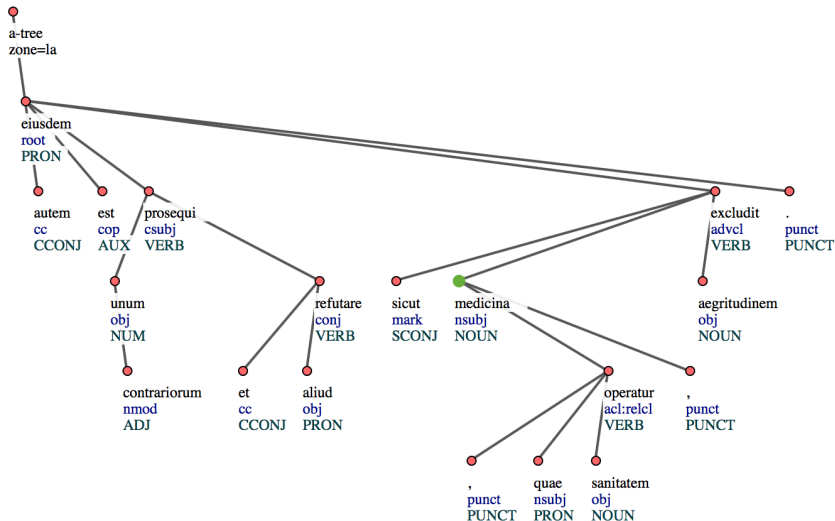
eiusdem autem est unum contrariorum prosequi et aliud refutare sicut medicina , quae sanitatem operatur , aegritudinem excludit . (IT-TB: SCG, lib. 1, cap. 1, n. 6)

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)



(Annotated) Corpora in LiLa

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



Token *medicina*

‘the healing art, medicine, surgery’

`https://lila-erc.eu/lodview/data/corpora/
ITTB/id/token/005.SCG*LB1.CP--++1.N.-6.
1-1.4-1W11`

Introduction and fundamentals

LiLa: mission and architecture

LiLa now!

LiLa Lemma Bank and Lexical Resources

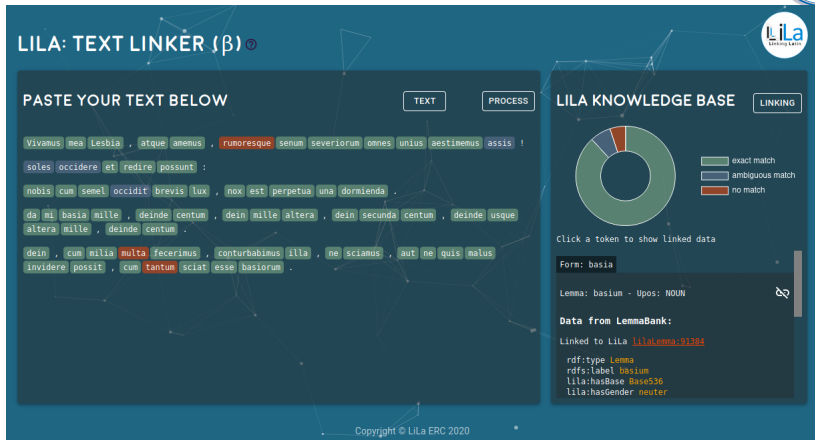
Textual Resources

Text Linker

To sum up



Figure: LiLa's Text Linker



LILA: TEXT LINKER (β)

PASTE YOUR TEXT BELOW

TEXT PROCESS

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 sciat esse basiorum .

LILA KNOWLEDGE BASE LINKING

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: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!

LiLa Lemma Bank and 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): 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

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](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.