The LiLa Knowledge Base of Interlinked Linguistic Resources for Latin

Marco Passarotti

Conference Digital Humanities and Neo-Latin Studies
April 15, 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.
Overview

Introduction and fundamentals
LiLa: mission and architecture

LiLa now!
LiLa Lemma Bank and Lexical Resources
Textual Resources
Text Linker
To sum up
Table of Contents

Introduction and fundamentals
  LiLa: mission and architecture

LiLa now!
  LiLa Lemma Bank and Lexical Resources
  Textual Resources
  Text Linker
  To sum up
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
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
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 Knowledge Base
Lexically-based architecture and (meta)data sources

Lemmas

Lexical Entries

Tokens

NLP Output

Lexical Resources
- Latin Wordnet
- Valency Lexicon
- Dictionaries...

Textual Resources
- Digital libraries
- Treebanks
- Textual corpora...

NLP Tools
- Tokenizers
- Taggers/parsers
- Lemmatizers...

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

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

Introduction and fundamentals
LiLa: mission and architecture

LiLa now!
LiLa Lemma Bank and Lexical Resources
Textual Resources
Text Linker
To sum up
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’
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
Table of Contents

Introduction and fundamentals
  LiLa: mission and architecture

LiLa now!
  LiLa Lemma Bank and Lexical Resources
  Textual Resources
  Text Linker
  To sum up
eiusdem autem est unum contrariorum prosequi et aliud refutare sicut medicina, quae sanitatem operatur, aegritudinem exclusit. (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)

Marco Passarotti
Token *medicina*
‘the healing art, medicine, surgery’
https://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCGLB1.CP--++1.N.-6.1-1.4-1W11
Table of Contents

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
Figure: Text processed against the LiLa Knowledge Base

http://lila-erc.eu:8080/LiLaTextLinker/
Table of Contents

Introduction and fundamentals
   LiLa: mission and architecture

LiLa now!
   LiLa Lemma Bank and Lexical Resources
   Textual Resources
   Text Linker
   To sum up
LiLa: Overview
Resources connected and upcoming connections

- **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

Marco Passarotti |
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
📍 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.
This work is licensed under a Creative Commons “Attribution 4.0 International” license.