Interoperabilità tra risorse linguistiche per il latino

I Linguistic Linked Open Data e la LiLa Knowledge Base

Marco Passarotti

UNIVERSITÀ CATTOLICA del Sacro Cuore

CIRCSE

Digital Humanities in pratica
26 Febbraio 2022, Collegio Ghislieri, Pavia

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

Services and Tools

To sum up

Table of Contents

Introduction and Fundamentals

- LiLa: Mission and Architecture

LiLa now!

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

Research question

State of affairs

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

Research question

State of affairs

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

- Textual Resources

Research question

State of affairs

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

- Textual Resources
- Lexical Resources

Research question

State of affairs

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

- Textual Resources
- Lexical Resources
- NLP Tools

Research question

State of affairs

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

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

LiLa Knowledge Base

Approach: Linked Data paradigm

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

LiLa

Linking Latin

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

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

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
- Federation: to combine information from physically separated repositories

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
- Federation: to combine information from physically separated repositories
- Dynamicity: to provide access to the most recent version of a resource

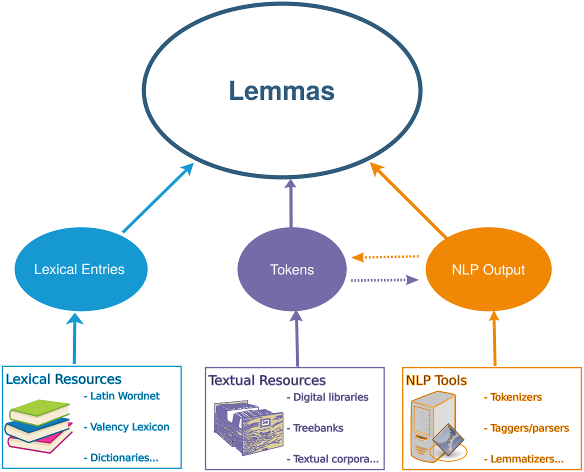
Benefits of Applying LD to Linguistic Resources

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
- 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 Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa: Structure

LiLa as mere reflection

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:

LiLa: Requirements

Connecting resources in the Knowledge Base

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

- Lemmatised

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)

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!

Table of Contents

Introduction and Fundamentals

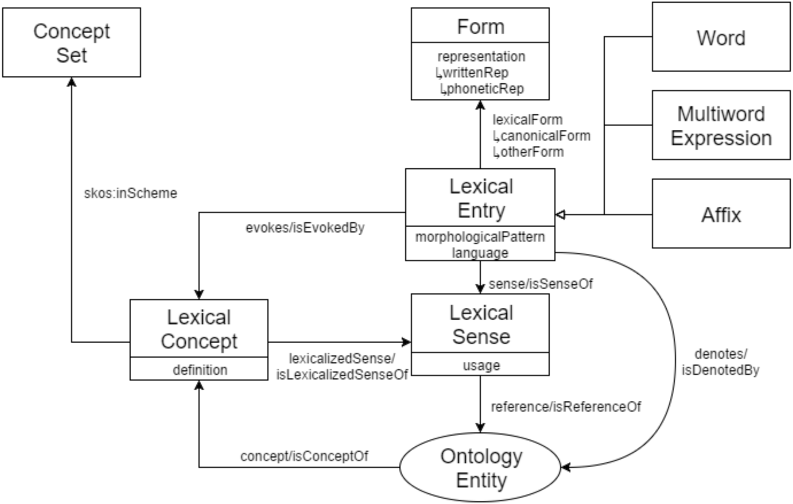
- LiLa: Mission and Architecture

LiLa now!

- Lemma Bank and Lexical Resources
- Textual Resources
- Services and Tools
- 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’

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

- Lemma Bank
- A bilingual Latin English dictionary (Lewis & Short)
- 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!

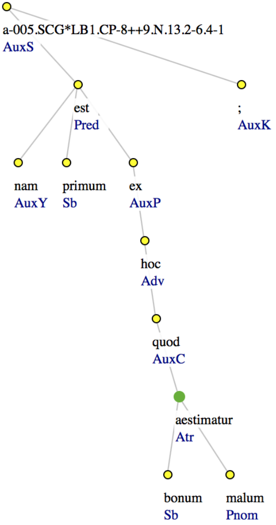
- Lemma Bank and Lexical Resources
- Textual Resources
- Services and Tools
- To sum up

(Annotated) Corpora in LiLa

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

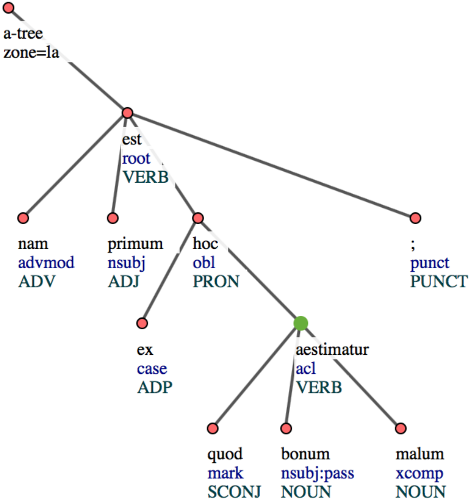
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

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)



(Annotated) Corpora in LiLa

A token from the *Index Thomisticus* Treebank in LiLa

Token *aestimatur*

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

Table of Contents

Introduction and Fundamentals

- LiLa: Mission and Architecture

LiLa now!

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

Lemma Bank Query Interface

Lemma Bank Query Interface

`https://lila-erc.eu/query/`

SPARQL Access Point

SPARQL Access Point

`https://lila-erc.eu/sparql/`

TextLinker

TextLinker

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

Table of Contents

Introduction and Fundamentals

LiLa: Mission and Architecture

LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up

LiLa: Overview

Resources connected and upcoming connections

Textual Resources

- ☑ Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 400,000 nodes
- ☑ UDante Treebank: ca. 46,000 tokens
- ☑ *Querolus sive Aulularia*: ca. 17,000 tokens
- ☑ *Liber Abbaci* (ch. VIII) by Leonardo Fibonacci: ca. 30,000 tokens
- ☑ LASLA Corpus: ca. 1.7 million tokens
- □ PROIEL and LLCT treebanks, CompHistSem, CroALa, Musisque DeoQue

Lexical Resources

- ☑ Lemma Bank: ca. 200,000 canonical forms
- ☑ Word Formation Latin: ca. 36,000 lemmas (Classical Latin)
- ☑ Etymological Dict. of Latin & the Other Italic Langs.: ca. 1,500 entries
- ☑ LatinAffectus: ca. 3,300 entries
- ☑ Index Graecorum Vocabulorum in L. Latinam Transl.: ca. 1,800 entries
- ☑ Latin WordNet: ca. 2,500 manually checked entries
- ☑ Latin Vallex 2.0: ca. 2,000 entries
- ☑ Lewis & Short Dictionary: ca. 50,000 entries
- □ Lexikon der Indogermanischen Verben (LIV). Wiktionary, BabelNet

TOTAL: approximately 33 million triples

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.