Towards connecting scholarly editions to corpora in the LiLa (Linking Latin) Knowledge Base of linguistic resources

Greta Franzini
greta.franzini@unicatt.it

Conference | Wuppertal, Germany | 17 December 2019

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.

Table of Contents

Table of Contents

Computational Linguistics *is an interdisciplinary field concerned with the processing of language by computers.* (Mitkov, 2004)

Computational Linguistics
Definition

Computational Linguistics is an interdisciplinary field concerned with the processing of language by computers. (Mitkov, 2004)

Computational Linguistics

Natural Language Processing

Computational Linguistics
Definition

Computational Linguistics *is an interdisciplinary field concerned with the processing of language by computers.* (Mitkov, 2004)

Computational Linguistics

Develops computational **methods** and **formalisms** to answer linguistics questions.

Natural Language Processing

Computational Linguistics
Definition

Computational Linguistics is an interdisciplinary field concerned with the processing of language by computers. (Mitkov, 2004)

Computational Linguistics

Develops computational **methods** and **formalisms** to answer linguistics questions.

Natural Language Processing

Solves **engineering** problems arising from the analysis of natural language text.

(adapted from Eisner, 2016)

Automatic language processing requires linguistic resources and NLP tools

Dictionary collection of words and phrases with information about them

lexicon dictionary/list of words, typically for computational purposes

thesaurus words grouped together according to similarity of meaning

Dictionary collection of words and phrases with information about them

lexicon dictionary/list of words, typically for computational purposes

thesaurus words grouped together according to similarity of meaning

Ontology inventory of objects or processes in a domain, together with a specification of some or all of the relations that hold among them, generally arranged as a hierarchy

Dictionary collection of words and phrases with information about them

- **lexicon** dictionary/list of words, typically for computational purposes
- **thesaurus** words grouped together according to similarity of meaning

Ontology inventory of objects or processes in a domain, together with a specification of some or all of the relations that hold among them, generally arranged as a hierarchy

Corpus a body of linguistic data in machine readable form, gathered according to some principled sampling method and criterion. A syntactically/semantically-annotated corpus is known as a *treebank*

Dictionary collection of words and phrases with information about them

- **lexicon** dictionary/list of words, typically for computational purposes
- **thesaurus** words grouped together according to similarity of meaning

Ontology inventory of objects or processes in a domain, together with a specification of some or all of the relations that hold among them, generally arranged as a hierarchy

Corpus a body of linguistic data in machine readable form, gathered according to some principled sampling method and criterion. A syntactically/semantically-annotated corpus is known as a *treebank*

Grammar systematic analysis of the structure of a language

Tokeniser performs tokenisation and determines the boundaries for individual tokens in text (words, numbers, punctuation)

Tokeniser performs tokenisation and determines the boundaries for individual tokens in text (words, numbers, punctuation)

Tagger assigns tags to words or expressions in a text (e.g. part of speech, named entity)

Tokeniser performs tokenisation and determines the boundaries for individual tokens in text (words, numbers, punctuation)

Tagger assigns tags to words or expressions in a text (e.g. part of speech, named entity)

Parser analyses a sentence or other string of words into its constituents, producing a parse tree of syntactic relations between them

Tokeniser	performs tokenisation and determines the boundaries for individual tokens in text (words, numbers, punctuation)
Tagger	assigns tags to words or expressions in a text (e.g. part of speech, named entity)
Parser	analyses a sentence or other string of words into its constituents, producing a parse tree of syntactic relations between them
Lemmatiser	groups the inflected forms of a word together under a base form, recovers the base form from an inflected form. Can be *morphological* (no context, ambiguity) or *morpho-syntactic* (context, no ambiguity).

Tokeniser	performs tokenisation and determines the boundaries for individual tokens in text (words, numbers, punctuation)
Tagger	assigns tags to words or expressions in a text (e.g. part of speech, named entity)
Parser	analyses a sentence or other string of words into its constituents, producing a parse tree of syntactic relations between them
Lemmatiser	groups the inflected forms of a word together under a base form, recovers the base form from an inflected form. Can be *morphological* (no context, ambiguity) or *morpho-syntactic* (context, no ambiguity).

... and more.

Corpora Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Index Thomisticus Treebank, PROIEL Latin Treebank, etc.

Corpora Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Index Thomisticus Treebank, PROIEL Latin Treebank, etc.

Lexica Vallex, IT-VaLex, Latin WordNet, Oxford Latin Dictionary, Du Cange Glossarium Mediae et Infimae Latinitatis, Thesaurus Lingua Latinae, Thesaurus Formarum Totius Latinitatis, Lexicon musicum Latinum medii aevi, etc.

Corpora Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Index Thomisticus Treebank, PROIEL Latin Treebank, etc.

Lexica Vallex, IT-VaLex, Latin WordNet, Oxford Latin Dictionary, Du Cange Glossarium Mediae et Infimae Latinitatis, Thesaurus Lingua Latinae, Thesaurus Formarum Totius Latinitatis, Lexicon musicum Latinum medii aevi, etc.

NLP Tools LEMLAT, Whitaker's Words, LatMor, TreeTagger, Collatinus, UDPipe, Chiron, etc.

Corpora Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Index Thomisticus Treebank, PROIEL Latin Treebank, etc.

Lexica Vallex, IT-VaLex, Latin WordNet, Oxford Latin Dictionary, Du Cange Glossarium Mediae et Infimae Latinitatis, Thesaurus Lingua Latinae, Thesaurus Formarum Totius Latinitatis, Lexicon musicum Latinum medii aevi, etc.

NLP Tools LEMLAT, Whitaker's Words, LatMor, TreeTagger, Collatinus, UDPipe, Chiron, etc.

Latin is the **most resourced** historical language

These resources and tools, however, are:

These resources and tools, however, are:

- Scattered and isolated

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

No interoperability!

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

No interoperability! Interoperability:

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

No interoperability! Interoperability:

- Increases productivity

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

No interoperability! Interoperability:

- Increases productivity
- Improves efficiency

These resources and tools, however, are:

- Scattered and isolated
- Developed for specific tasks
- Follow different annotation schemas and conceptual models

No interoperability! Interoperability:

- Increases productivity
- Improves efficiency
- More effective knowledge organisation

Linked Data: Semantic Web technology

Linked Data
A solution

Linked Data: Semantic Web technology
Semantic Web: set of standards and best practices to share data across the web, and to help machines make inferences and understand the meaning of this data.

Linked Data
A solution

Linked Data: Semantic Web technology
Semantic Web: set of standards and best practices to share data across the web, and to help machines make inferences and understand the meaning of this data.

Advantages:

Linked Data: Semantic Web technology
Semantic Web: set of standards and best practices to share data across the web, and to help machines make inferences and understand the meaning of this data.

Advantages:

- Connects and defines relationships between heterogeneous datasets

Linked Data
A solution

Linked Data: Semantic Web technology
Semantic Web: set of standards and best practices to share data across the web, and to help machines make inferences and understand the meaning of this data.

Advantages:

- Connects and defines relationships between heterogeneous datasets
- Aggregates distributed datasets to reduce dispersion and increase (serendipitous) knowledge discovery (i.e. discoverability of the resource)

Linked Data: Semantic Web technology
Semantic Web: set of standards and best practices to share data across the web, and to help machines make inferences and understand the meaning of this data.

Advantages:

- Connects and defines relationships between heterogeneous datasets
- Aggregates distributed datasets to reduce dispersion and increase (serendipitous) knowledge discovery (i.e. discoverability of the resource)
- Allows us to build systems that can reason across the web and answer complex questions

Linked Data
How does it work?

Linked Data technology describes data as **triples** (statements):

Linked Data
How does it work?

Linked Data technology describes data as **triples** (statements):

Linked Data technology describes data as **triples** (statements):

- `OBJECT` of one triple can be the `SUBJECT` of another triple

Linked Data
How does it work?

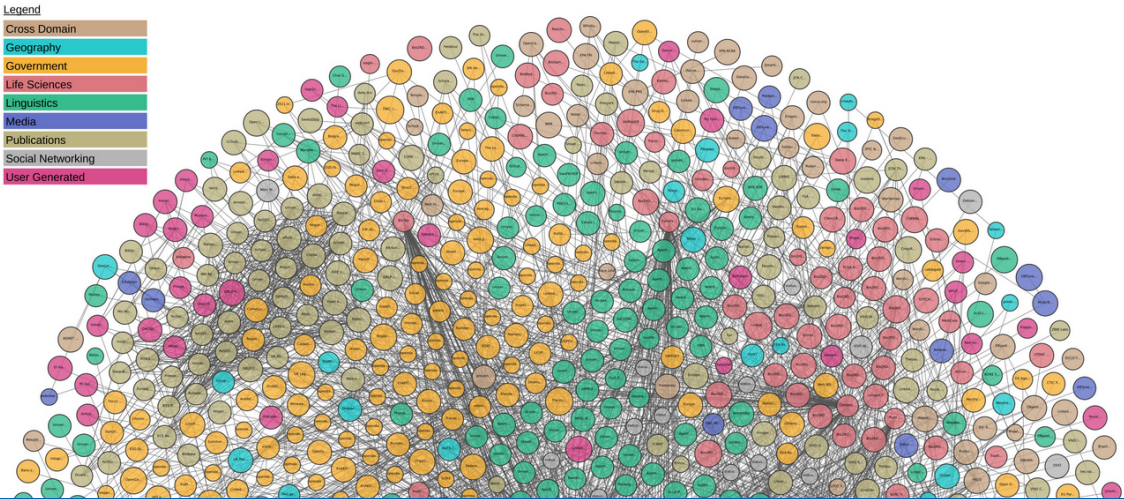
Linked Data technology describes data as **triples** (statements):

- `OBJECT` of one triple can be the `SUBJECT` of another triple
- Nodes and edges are assigned persistent **Uniform Resource Identifiers** (URIs) for unambiguous identification across the web

Linked Data technology describes data as **triples** (statements):

- `OBJECT` of one triple can be the `SUBJECT` of another triple
- Nodes and edges are assigned persistent **Uniform Resource Identifiers** (URIs) for unambiguous identification across the web
- Relationships are described by **ontologies** or vocabularies of knowledge representation

Linked Data
Many domains



Linguistic Linked Data

Linguistics

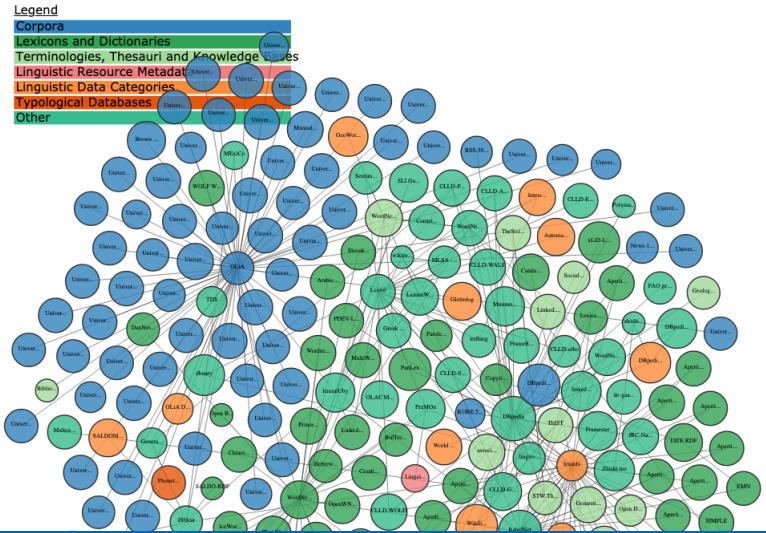


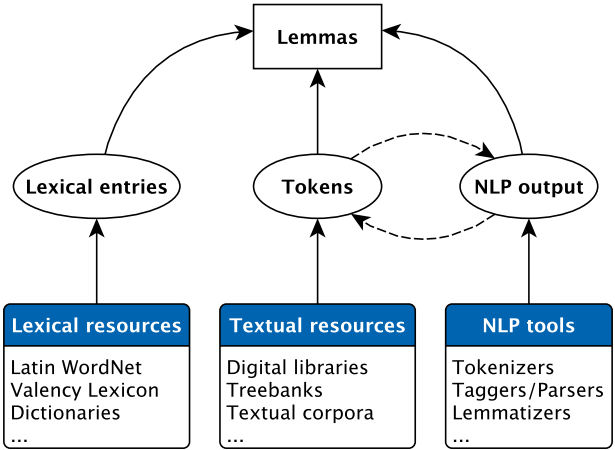
Table of Contents

- **Funding**: ERC Consolidator Grant, 2M EUR
- **Duration**: 2018-2023
- **Team**: 9 staff + student assistants
- **Website**: `https://lila-erc.eu`

- **Objective**: Knowledge Base of Linguistic Resources & Natural Language Processing Tools
- **Method**: Linked Data paradigm (FAIR principles)
- **Purpose**: Foster resource/data interoperability

LiLa: Structure

Lemmas as connectors



LiLa: Structure
Lemma bank

Lemma bank of **LEMLAT**, our morphological analyser. **Over 150,000 lemmas**, including:

Lemma bank of **LEMLAT**, our morphological analyser. **Over 150,000 lemmas**, including:

- Classical: 43,432 lemmas from Georges & Georges (1913-1918), Glare (1982), Gradenwitz (1904)

Lemma bank of **LEMLAT**, our morphological analyser. **Over 150,000 lemmas**, including:

- Classical: 43,432 lemmas from Georges & Georges (1913-1918), Glare (1982), Gradenwitz (1904)
- Medieval and Late: 82,556 lemmas from Du Cange (1883-1887)

Lemma bank of **LEMLAT**, our morphological analyser. **Over 150,000 lemmas**, including:

- Classical: 43,432 lemmas from Georges & Georges (1913-1918), Glare (1982), Gradenwitz (1904)
- Medieval and Late: 82,556 lemmas from Du Cange (1883-1887)
- Onomasticon: 26,250 lemmas from Forcellini (1940)

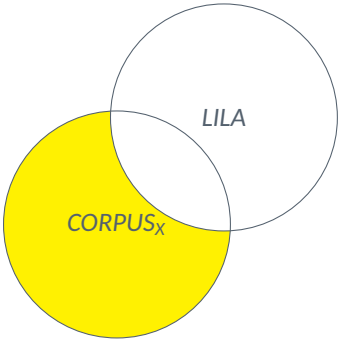
LiLa: Structure
Lemma bank

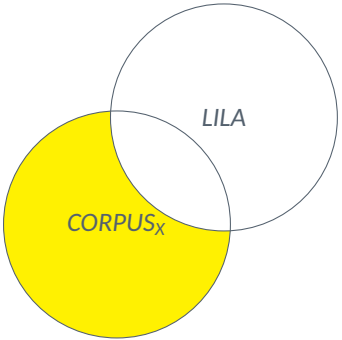
Lemma bank of **LEMLAT**, our morphological analyser. **Over 150,000 lemmas**, including:

- Classical: 43,432 lemmas from Georges & Georges (1913-1918), Glare (1982), Gradenwitz (1904)
- Medieval and Late: 82,556 lemmas from Du Cange (1883-1887)
- Onomasticon: 26,250 lemmas from Forcellini (1940)

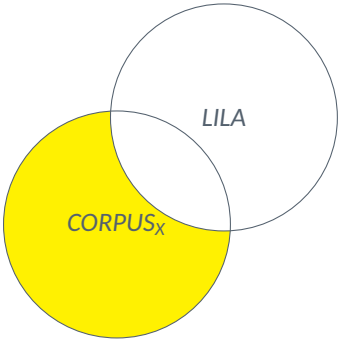
`http://www.lemlat3.eu/`

LiLa: Structure
Lemma bank



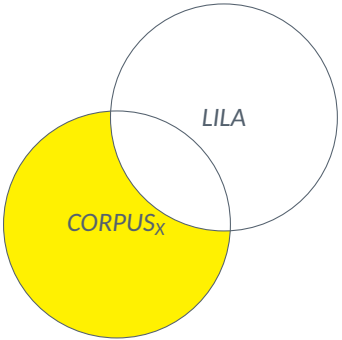


Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

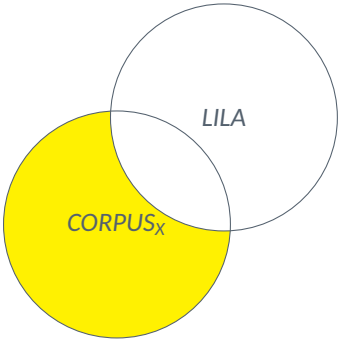
- L_C doesn't exist in L_L



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

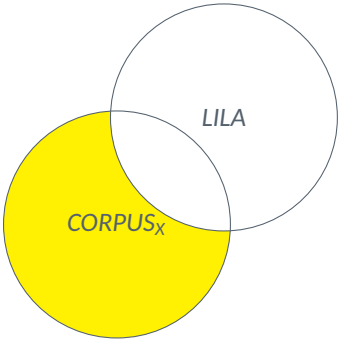
- L_C doesn't exist in L_L
- L_C is a different written representation of a L_L, e.g. *annuncio* vs. *adnuntio*

LiLa: Structure
Lemma bank



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

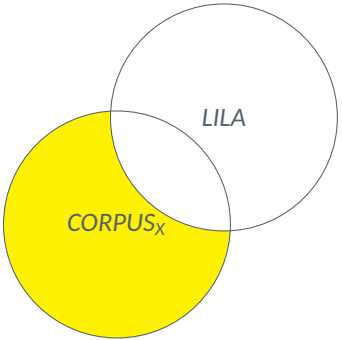
- L_C doesn't exist in L_L
- L_C is a different written representation of a L_L, e.g. *annuncio* vs. *adnuntio*
- L_C is a lemma variant of a L_L, e.g. *anthropomorphita* vs. *anthropomorphitae* (*pluralia tantum*)



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

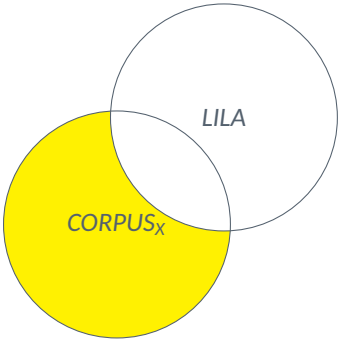
- L_C doesn't exist in L_L
- L_C is a different written representation of a L_L, e.g. *annuncio* vs. *adnuntio*
- L_C is a lemma variant of a L_L, e.g. *anthropomorphita* vs. *anthropomorphitae* (*pluralia tantum*)
- L_C is a pseudo-lemma, i.e. non Latin words

LiLa: Structure
Lemma bank



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

- L_C doesn't exist in L_L
- L_C is a different written representation of a L_L, e.g. *annuncio* vs. *adnuntio*
- L_C is a lemma variant of a L_L, e.g. *anthropomorphita* vs. *anthropomorphitae* (*pluralia tantum*)
- L_C is a pseudo-lemma, i.e. non Latin words
- lemmatisation errors, e.g. *pbiectum* instead of *obiectum*



Corpus lemmas (L_C) can't connect to LiLa lemmas (L_L) when:

- L_C doesn't exist in L_L
- L_C is a different written representation of a L_L, e.g. *annuncio* vs. *adnuntio*
- L_C is a lemma variant of a L_L, e.g. *anthropomorphita* vs. *anthropomorphitae* (*pluralia tantum*)
- L_C is a pseudo-lemma, i.e. non Latin words
- lemmatisation errors, e.g. *pbiectum* instead of *obiectum*

Manual fix

To build and **define relationships between datasets** (triples), **LiLa reuses** the following **ontologies**:

To build and **define relationships between datasets** (triples), **LiLa reuses** the following **ontologies**:

- OntoLex (Lemon): for lexical information

To build and **define relationships between datasets** (triples), **LiLa reuses** the following **ontologies**:

- OntoLex (Lemon): for lexical information
- OLiA (Ontologies of Linguistic Annotation) bundle: for part-of-speech tagging

To build and **define relationships between datasets** (triples), **LiLa reuses** the following **ontologies**:

- OntoLex (Lemon): for lexical information
- OLiA (Ontologies of Linguistic Annotation) bundle: for part-of-speech tagging
- NIF (NLP Interchange Format) and POWLA (OWL + PAULA, Potsdamer Austauschformat Linguistischer Annotationen) for corpus annotation

LiLa

LiLa: Structure
Triplestore

LiLa = database of triples =

LiLa = database of triples = **triplestore**

LiLa: Overview

Resources connected and upcoming connections

- Corpora
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles*)
- Lexica
  - ☑ Word Formation Latin (Classical Latin)
- NLP tools
  - ☑ LEMLAT (lemma bank)

LiLa: Overview

Resources connected and upcoming connections

- Corpora
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles*)
  - ☐ Dante (700th death anniversary coming up!)
- Lexica
  - ☑ Word Formation Latin (Classical Latin)
- NLP tools
  - ☑ LEMLAT (lemma bank)

LiLa: Overview

Resources connected and upcoming connections

- Corpora
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles*)
  - ☐ Dante (700th death anniversary coming up!)
- Lexica
  - ☑ Word Formation Latin (Classical Latin)
- NLP tools
  - ☑ LEMLAT (lemma bank)

LiLa: Overview

Resources connected and upcoming connections

- Corpora
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*)
  - [ ] Dante (700th death anniversary coming up!)
- Lexica
  - [x] Word Formation Latin (Classical Latin)
  - [ ] BRILL Etymological dictionary of Latin and the other Italic Languages
- NLP tools
  - [x] LEMLAT (lemma bank)

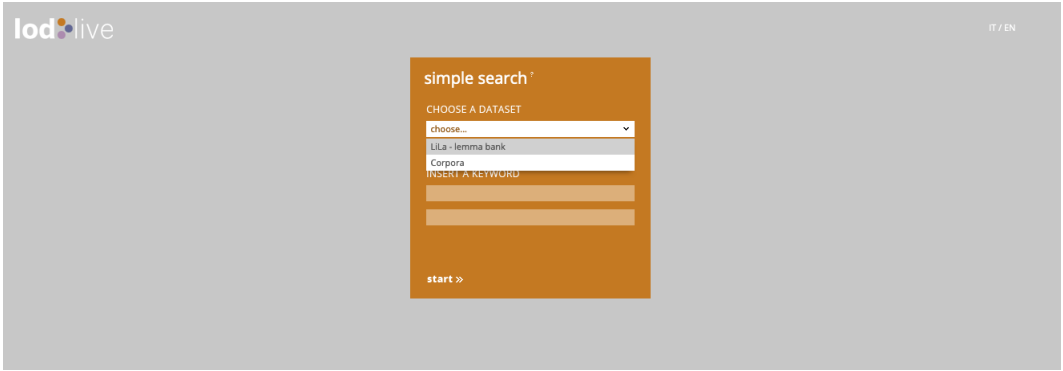
LiLa: Overview

Resources connected and upcoming connections

- Corpora
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*)
  - [ ] Dante (700th death anniversary coming up!)
- Lexica
  - [x] Word Formation Latin (Classical Latin)
  - [ ] BRILL Etymological dictionary of Latin and the other Italic Languages
  - [ ] Latin WordNet
- NLP tools
  - [x] LEMLAT (lemma bank)

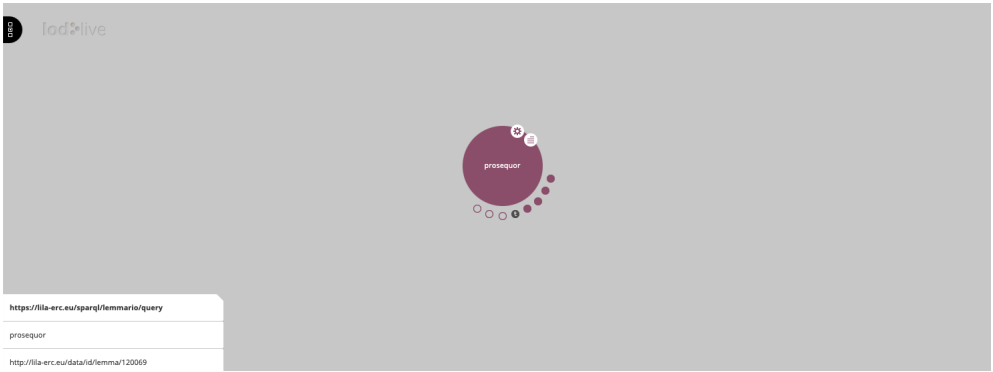
LiLa: Structure

Querying the lemma bank



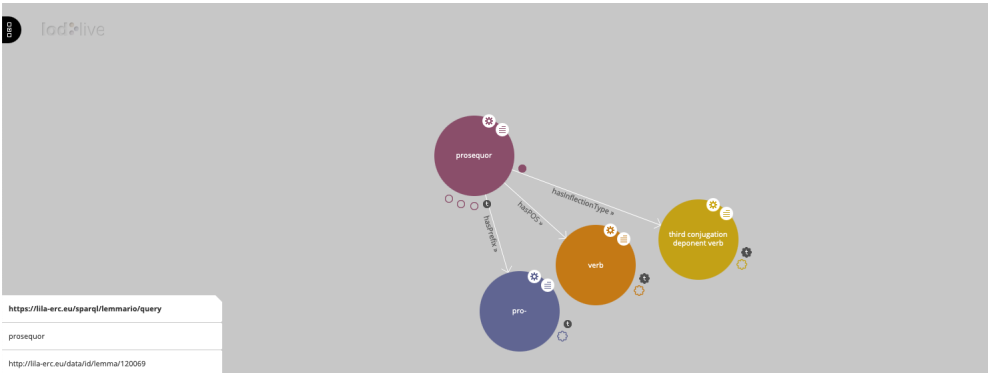
LiLa: Structure

An example: LOD view of ITTB token lemma *prosequor*



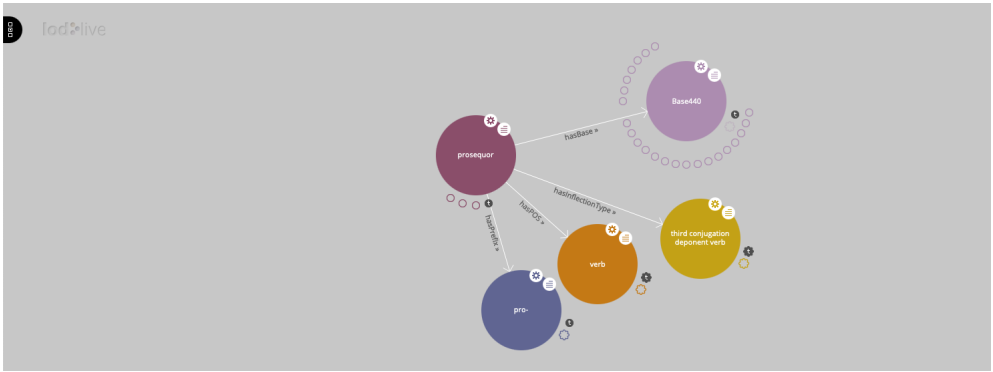
LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*



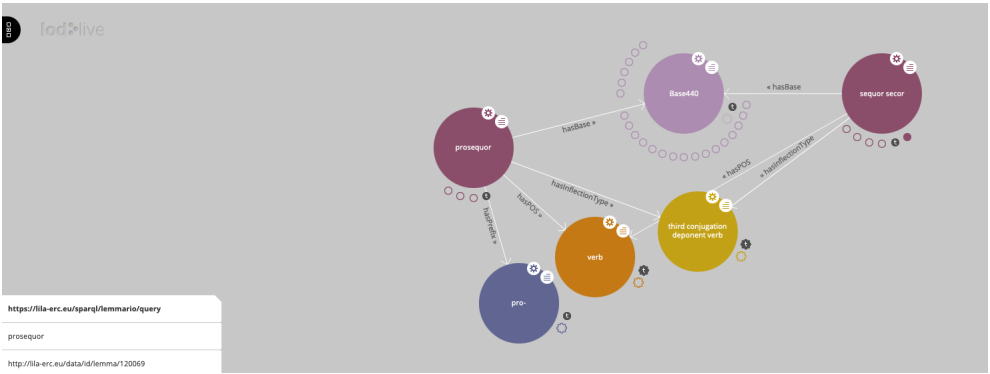
LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*



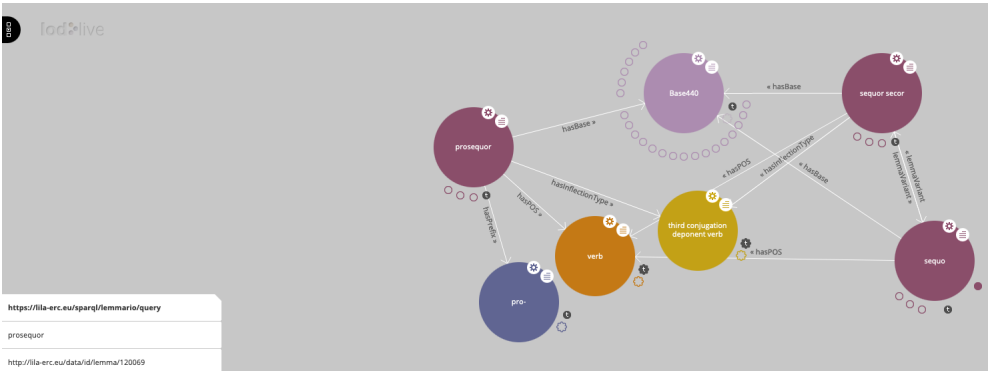
LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*



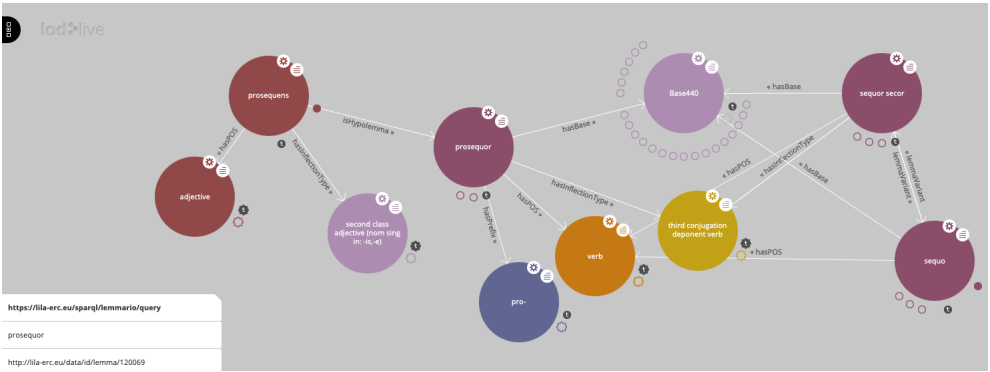
LiLa: Structure

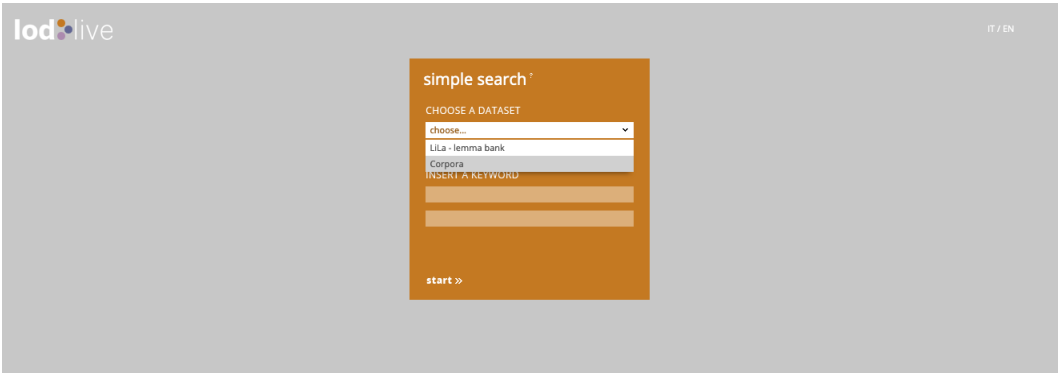
An example: LOD view of LEMLAT lemma *prosequor*



LiLa: Structure

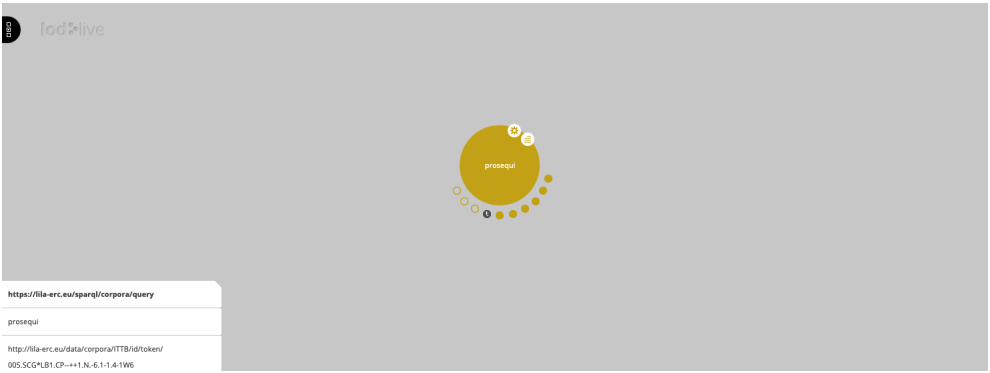
An example: LOD view of LEMLAT lemma *prosequor*




lodlive
IT / EN
simple search
CHOOSE A DATASET
choose...
LiLa - lemma bank
Corpora
INSERT A KEYWORD
start »

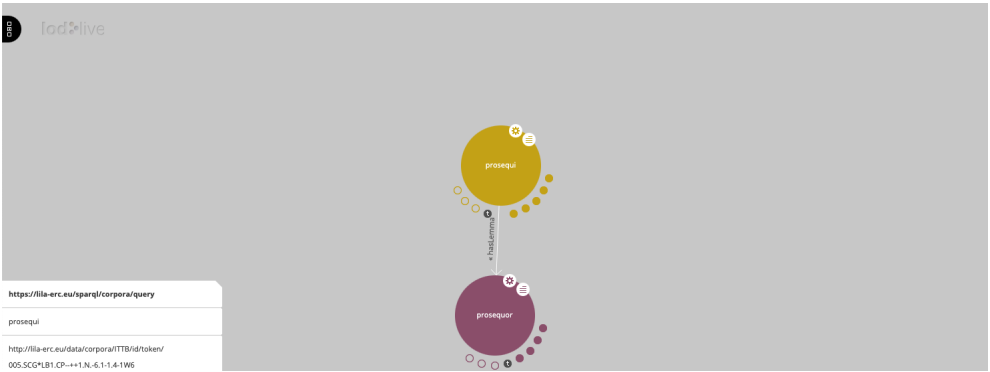
LiLa: Structure

An example: LOD view of ITTB token *prosequi*



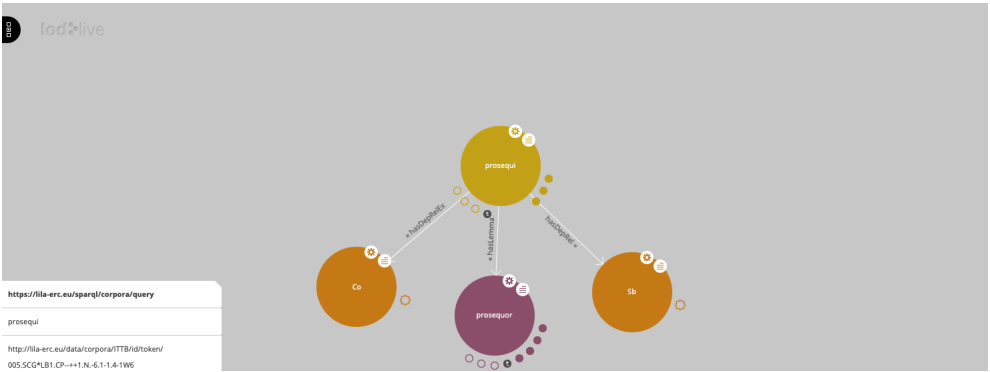
LiLa: Structure

An example: LOD view of ITTB token *prosequi*



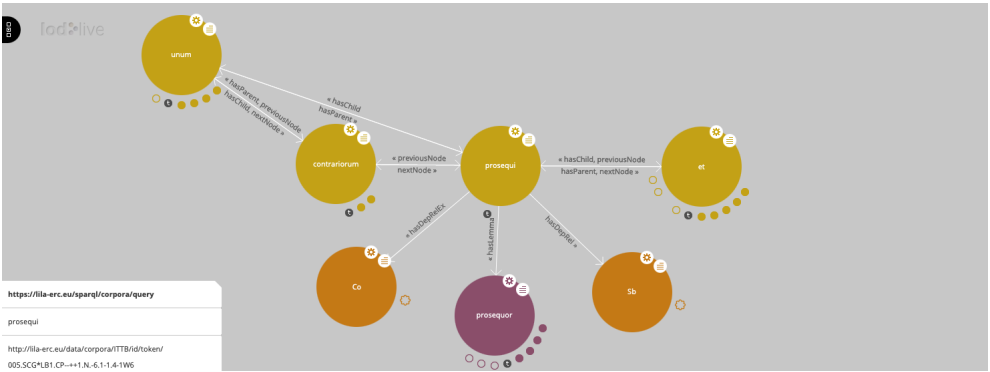
LiLa: Structure

An example: LOD view of ITTB token *prosequi*



LiLa: Structure

An example: LOD view of ITTB token *prosequi*

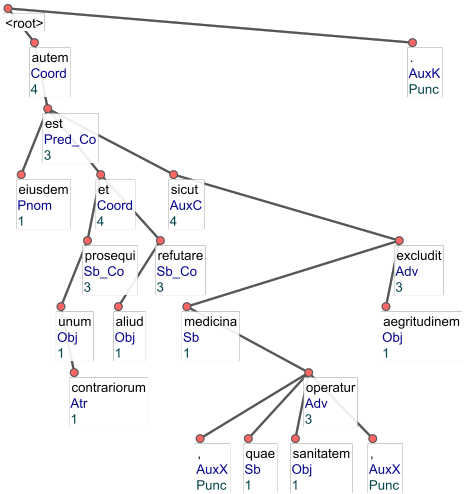


LiLa: Structure

An example: LOD view of ITTB token *prosequi*

eiusdem autem est ***unum contrariorum prosequi et*** *aliud refutare sicut medicina , quae sanitatem operatur , aegritudinem excludit .* (ITTB, 1.1.6)

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)



https://lila-erc.eu/lodlive/

LiLa reflects the annotation granularity of the resources it connects

No data enrichment or further analysis is performed

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

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

- Lemmatised

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

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

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

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

Table of Contents

"[...] computational philology seems to be somewhat decoupled from the recent progress in [Linguistic Linked Open Data]: even though LOD as a concept is gaining significant popularity in Digital Humanities, existing LLOD standards and vocabularies are not widely used in this community, and philological resources are underrepresented in the LLOD cloud diagram [...]." (Chiarcos et al., 2018)

"[...] computational philology seems to be somewhat decoupled from the recent progress in [Linguistic Linked Open Data]: even though LOD as a concept is gaining significant popularity in Digital Humanities, existing LLOD standards and vocabularies are not widely used in this community, and philological resources are underrepresented in the LLOD cloud diagram [...]." (Chiarcos et al., 2018)

"[...] As of yet only a relatively small number of born-digital editions of [...] Latin texts exists [...]." (Fischer, 2017)

> *"[…] computational philology seems to be somewhat decoupled from the recent progress in [Linguistic Linked Open Data]: even though LOD as a concept is gaining significant popularity in Digital Humanities, existing LLOD standards and vocabularies are not widely used in this community, and philological resources are underrepresented in the LLOD cloud diagram […]."* (Chiarcos et al., 2018)

> *"[…] As of yet only a relatively small number of born-digital editions of […] Latin texts exists […]."* (Fischer, 2017)

Of these, only a handful provide (some) data in Linked Data format.

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:
  1. **Textual**, i.e. the transcription (`<body>`, `<ab>`, `<div>`, etc.)

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:
  1. **Textual**, i.e. the transcription (`<body>`, `<ab>`, `<div>`, etc.)
  2. **Bibliographic**, e.g. properties of the edition (`<teiHeader>`)

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:
  1. **Textual**, i.e. the transcription (`<body>`, `<ab>`, `<div>`, etc.)
  2. **Bibliographic**, e.g. properties of the edition (`<teiHeader>`)
  3. **Source**, e.g. date, material, scribe, binding, folio count, size, etc. (`<teiHeader>`)

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:
  1. **Textual**, i.e. the transcription (`<body>`, `<ab>`, `<div>`, etc.)
  2. **Bibliographic**, e.g. properties of the edition (`<teiHeader>`)
  3. **Source**, e.g. date, material, scribe, binding, folio count, size, etc. (`<teiHeader>`)
  4. **Linguistic**, e.g. lemma, etc. (`<app>`)

Why so few Linked Data-compatible editions of Latin texts? Possible reasons:

- Projects lack the know-how and/or have other priorities
- **Scholarly editions are complex objects**. Many layers of information, including:
  1. **Textual**, i.e. the transcription (`<body>`, `<ab>`, `<div>`, etc.)
  2. **Bibliographic**, e.g. properties of the edition (`<teiHeader>`)
  3. **Source**, e.g. date, material, scribe, binding, folio count, size, etc. (`<teiHeader>`)
  4. **Linguistic**, e.g. lemma, etc. (`<app>`)
  5. **Palaeographic,** e.g. abbreviations, ligatures, glyphs, allographs, etc. (`<app>`)

Linked Data support:

Linked Data support:

- ☑ **Bibliographic + Textual**
  - FABiO (FRBR-aligned Bibliographic Ontology)
  - CiTO (Citation Typing Ontology)
  - DC (Dublin Core)

Linked Data support:

- **Bibliographic + Textual**
  - FABiO (FRBR-aligned Bibliographic Ontology)
  - CiTO (Citation Typing Ontology)
  - DC (Dublin Core)
- **Source**
  - DM2E (Digitised Manuscripts to Europeana)
  - FRBRoo (FRBR-object oriented) SAWS (Sharing Ancient Wisdoms)

Linked Data support:

- **Bibliographic + Textual**
  - FABiO (FRBR-aligned Bibliographic Ontology)
  - CiTO (Citation Typing Ontology)
  - DC (Dublin Core)
- **Source**
  - DM2E (Digitised Manuscripts to Europeana)
  - FRBRoo (FRBR-object oriented) SAWS (Sharing Ancient Wisdoms)
- **Linguistic**
  - Ontolex, NIF, POWLA, OLiA

Scholarly Edition
Information layers

Linked Data support:

- ☑ **Bibliographic + Textual**
  - FABiO (FRBR-aligned Bibliographic Ontology)
  - CiTO (Citation Typing Ontology)
  - DC (Dublin Core)
- ☑ **Source**
  - DM2E (Digitised Manuscripts to Europeana)
  - FRBRoo (FRBR-object oriented) SAWS (Sharing Ancient Wisdoms)
- ☑ **Linguistic**
  - Ontolex, NIF, POWLA, OLiA
- ☐ **Palaeographic**
  - Peter Stokes: DigiPal project
  - Paolo Monella: VeDPH seminar, 4th December 2019

Example:

Example:

- Vespasiano da Bisticci, Letters

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Tools:

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Tools:

- TEI-to-RDF converters (e.g. RDF Textual Encoding Framework)

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Tools:

- TEI-to-RDF converters (e.g. RDF Textual Encoding Framework)
- Linked Data support for the *Edition Visualisation Technology* (upcoming talk by Roberto Rosselli del Turco and Paolo Monella at AIUCD 2020)

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Tools:

- TEI-to-RDF converters (e.g. RDF Textual Encoding Framework)
- Linked Data support for the *Edition Visualisation Technology* (upcoming talk by Roberto Rosselli del Turco and Paolo Monella at AIUCD 2020)

Initiatives:

Example:

- Vespasiano da Bisticci, Letters (not lemmatised/PoS-tagged!)

Tools:

- TEI-to-RDF converters (e.g. RDF Textual Encoding Framework)
- Linked Data support for the *Edition Visualisation Technology* (upcoming talk by Roberto Rosselli del Turco and Paolo Monella at AIUCD 2020)

Initiatives:

- Workshop *Scholarly Digital Editions, Graph Data-Models and Semantic Web Technologies* (GraphSDE, 3-4.06.2019)

Scholarly editions

Hypothetical (and brutally simplistic) Corpus + Edition Linked Data scenario

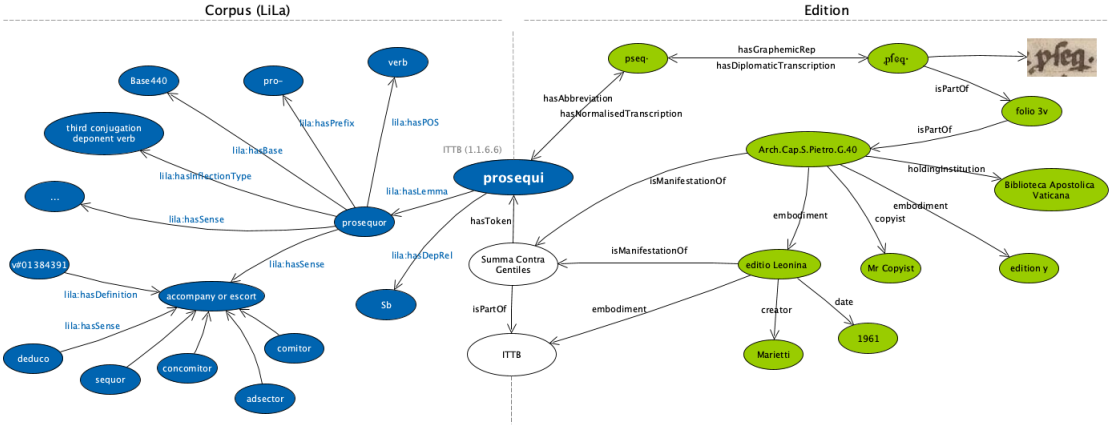


Table of Contents

Conclusion

Scholarly Editions and Corpora: Mutual benefits

Linguistic corpora:

Linguistic corpora:

- provide new forms of access to editions

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Conclusion

Scholarly Editions and Corpora: Mutual benefits

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Scholarly editions:

Conclusion

Scholarly Editions and Corpora: Mutual benefits

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Scholarly editions:

- provide new forms of access to corpora

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Scholarly editions:

- provide new forms of access to corpora
- provide connections to cultural heritage objects

Conclusion

Scholarly Editions and Corpora: Mutual benefits

Linguistic corpora:

- provide new forms of access to editions
- provide the bigger picture, i.e. large and diachronic linguistic context

Scholarly editions:

- provide new forms of access to corpora
- provide connections to cultural heritage objects
- provide philological layer of annotation (textual criticism)

Thanks!
Get in touch

Greta Franzini
CIRCSE, Università Cattolica del Sacro Cuore

greta.franzini@unicatt.it

@ERC_LiLa

https://github.com/CIRCSE

https://lila-erc.eu

Largo Gemelli 1, 20123 Milan, Italy

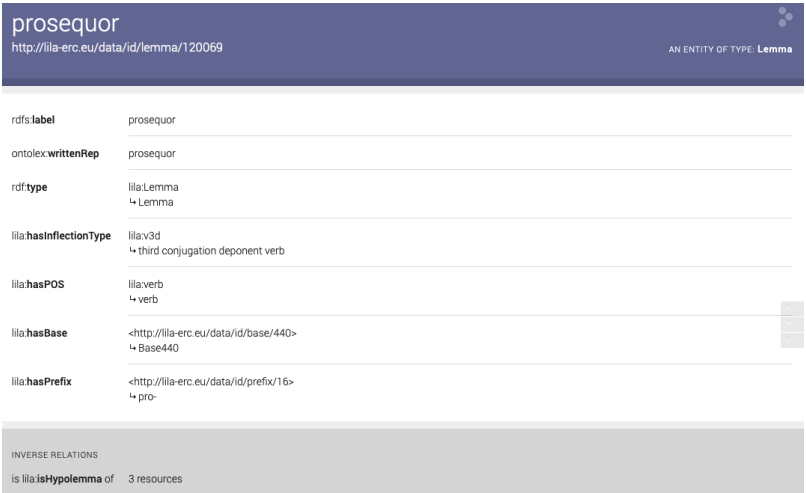
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.

Works cited

- Chiarcos et al. (2018) 'Towards a Linked Open Data Edition of Sumerian Corpora', *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 7-12, Miyasaki, Japan. ISBN: 979-10-95546-00-9*
- Eisner, J. (2016) *How is computational linguistics different from natural language processing?*
- Fischer, F. (2017) 'Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements for Textual Criticism', *Speculum*, 92/S1. DOI: 10.1086/693823
- Mitkov, R. (2004) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press

LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*



prosequor

http://lila-erc.eu/data/id/lemma/120069

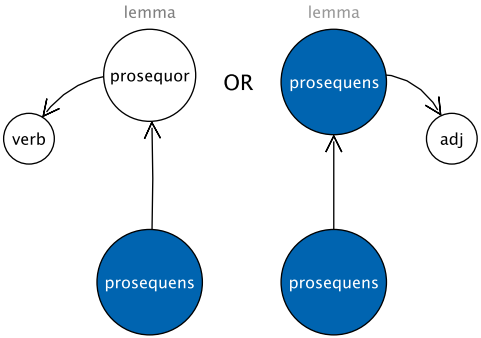
AN ENTITY OF TYPE: **Lemma**

rdfs:**label**	prosequor
ontolex:**writtenRep**	prosequor
rdf:**type**	lila:Lemma ↳ Lemma
lila:**hasInflectionType**	lila:v3d ↳ third conjugation deponent verb
lila:**hasPOS**	lila:verb ↳ verb
lila:**hasBase**	<http://lila-erc.eu/data/id/base/440> ↳ Base440
lila:**hasPrefix**	<http://lila-erc.eu/data/id/prefix/16> ↳ pro-

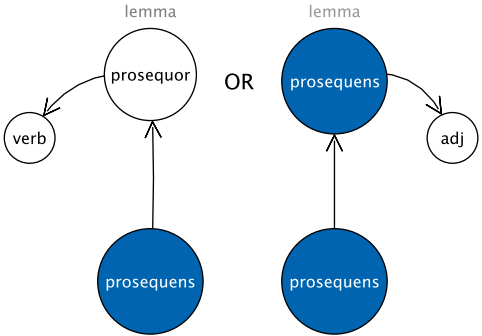
INVERSE RELATIONS

is lila:**isHypolemma** of 3 resources

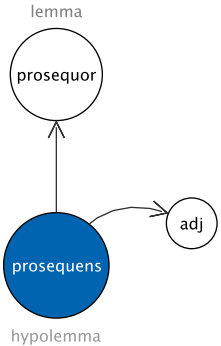
Ambiguity


lemma
prosequor
verb
prosequens
OR
lemma
prosequens
adj
prosequens

Ambiguity



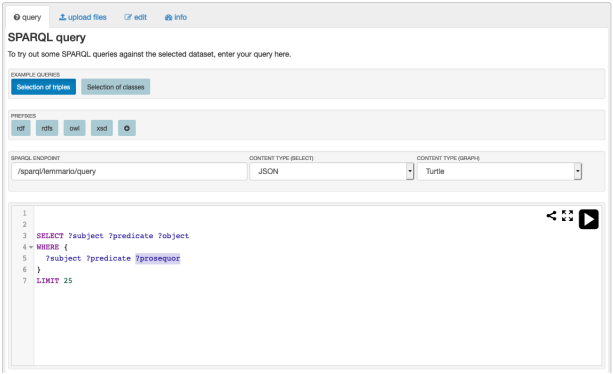
Solution



LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*

SPARQL endpoint with graphical interface to query against the LiLa triplestore.



LiLa: Structure

An example: LOD view of LEMLAT lemma *prosequor*

EvaLatin

- Evaluation campaign designed following a long tradition in NLP (MUC, ACE, SemEval, CoNLL…)
- Shared tasks, shared training and test data, shared evaluation metrics
- 3 tasks:
  1. PoS tagging
  2. Lemmatisation
- 3 sub-tasks for each task:
  1. Basic
  2. Cross-Genre
  3. Cross-Time