Lemmas, Trees and Classes

From the *Index Thomisticus* to Linked Data in the LiLa Knowledge Base of Linguistic Resources for Latin

Marco Passarotti

UNIVERSITÀ CATTOLICA del Sacro Cuore

IRCSE

Dottorato in Letterature Straniere, Lingue e Linguistica
Università di Verona | 23 April 2020

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.

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

Research need
Making sense

To make sense of this quantity of empirical data:

Research need
Making sense

To make sense of this quantity of empirical data:

- to extract maximum benefit from our research investments

Research need
Making sense

To make sense of this quantity of empirical data:

- to extract maximum benefit from our research investments
- to impact and improve the life of Classicists through exploitable computational resources and tools

Research need
Making sense

To make sense of this quantity of empirical data:

- to extract maximum benefit from our research investments
- to impact and improve the life of Classicists through exploitable computational resources and tools

From Information to Knowledge

2018-2023

A collection of interoperable linguistics resources (and NLP tools) described with the same vocabulary for knowledge description

Interlinking as a Form of Interaction

LiLa is based on an ontology made of:

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)
- **Object properties**: ways in which classes and individuals can be related to one another: RDF triples.
  Labels from a restricted vocabulary of knowledge description:
  `hasLemma`, `hasPoS`

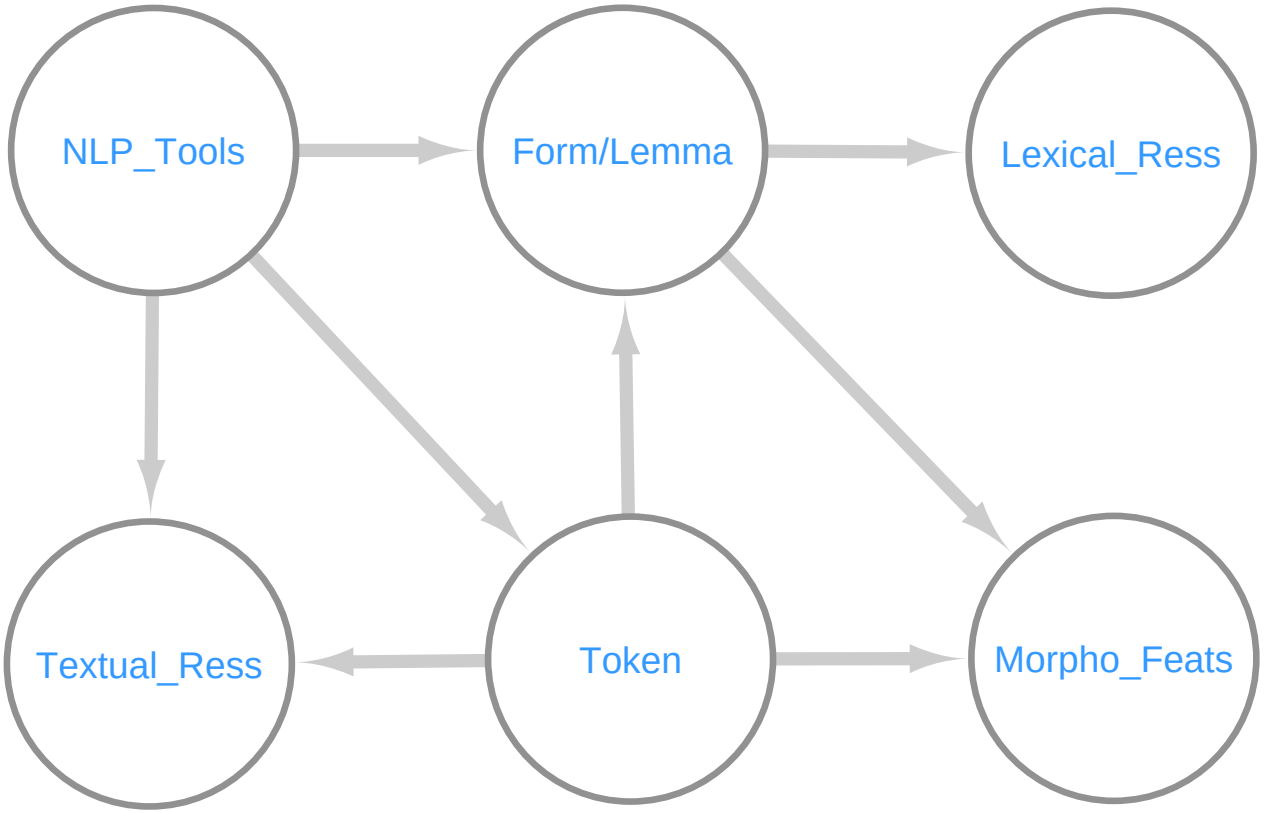
LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)
- **Object properties**: ways in which classes and individuals can be related to one another: RDF triples.
  Labels from a restricted vocabulary of knowledge description:
  `hasLemma`, `hasPoS`

Each component of the ontology is uniquely identified through a URI.

LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa: Overview

Resources connected and upcoming connections

LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles* and more): approx. 350,000 nodes
  - [x] Dante Search (700th death anniversary coming up!): approx. 46,000 tokens
  - [ ] Bibliotheca Croatiae auctorum Latinorum (CroALa)
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics Corpus

LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles* and more): approx. 350,000 nodes
  - ☑ Dante Search (700th death anniversary coming up!): approx. 46,000 tokens
  - ☐ Bibliotheca Croatiae auctorum Latinorum (CroALa)
  - ☐ PROIEL and LLCT treebanks
  - ☐ Computational Historical Semantics Corpus
- **Lexica**
  - ☑ Word Formation Latin: approx. 46,000 lemmas (Classical Latin)
  - ☑ BRILL Etymological dictionary of Latin and the other Italic Languages: approx. 1,400 entries
  - ☑ LatinAffectus (sentiment lexicon for Latin): approx. 2,300 entries
  - ☐ Latin WordNet

LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles* and more): approx. 350,000 nodes
  - ☑ Dante Search (700th death anniversary coming up!): approx. 46,000 tokens
  - ☐ Bibliotheca Croatiae auctorum Latinorum (CroALa)
  - ☐ PROIEL and LLCT treebanks
  - ☐ Computational Historical Semantics Corpus
- **Lexica**
  - ☑ Word Formation Latin: approx. 46,000 lemmas (Classical Latin)
  - ☑ BRILL Etymological dictionary of Latin and the other Italic Languages: approx. 1,400 entries
  - ☑ LatinAffectus (sentiment lexicon for Latin): approx. 2,300 entries
  - ☐ Latin WordNet
- **NLP tools**
  - ☑ LEMLAT (lemma bank): approx. 150,000 lemmas

LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles* and more): approx. 350,000 nodes
  - ☑ Dante Search (700th death anniversary coming up!): approx. 46,000 tokens
  - ☐ Bibliotheca Croatiae auctorum Latinorum (CroALa)
  - ☐ PROIEL and LLCT treebanks
  - ☐ Computational Historical Semantics Corpus
- **Lexica**
  - ☑ Word Formation Latin: approx. 46,000 lemmas (Classical Latin)
  - ☑ BRILL Etymological dictionary of Latin and the other Italic Languages: approx. 1,400 entries
  - ☑ LatinAffectus (sentiment lexicon for Latin): approx. 2,300 entries
  - ☐ Latin WordNet
- **NLP tools**
  - ☑ LEMLAT (lemma bank): approx. 150,000 lemmas
- **TOTAL**: more than 2 million triples

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!

Lemma *admiror*

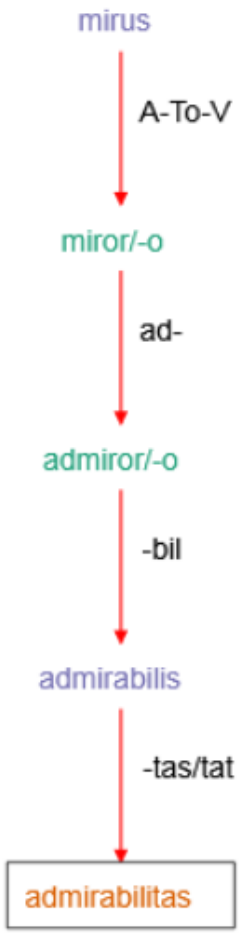
https://lila-erc.eu/data/id/lemma/87541

LiLa: Structure. Word Formation

WFL: Morphotactic and Hierarchical Derivation Trees

LiLa: Structure. Word Formation

WFL: Morphotactic and Hierarchical Derivation Trees

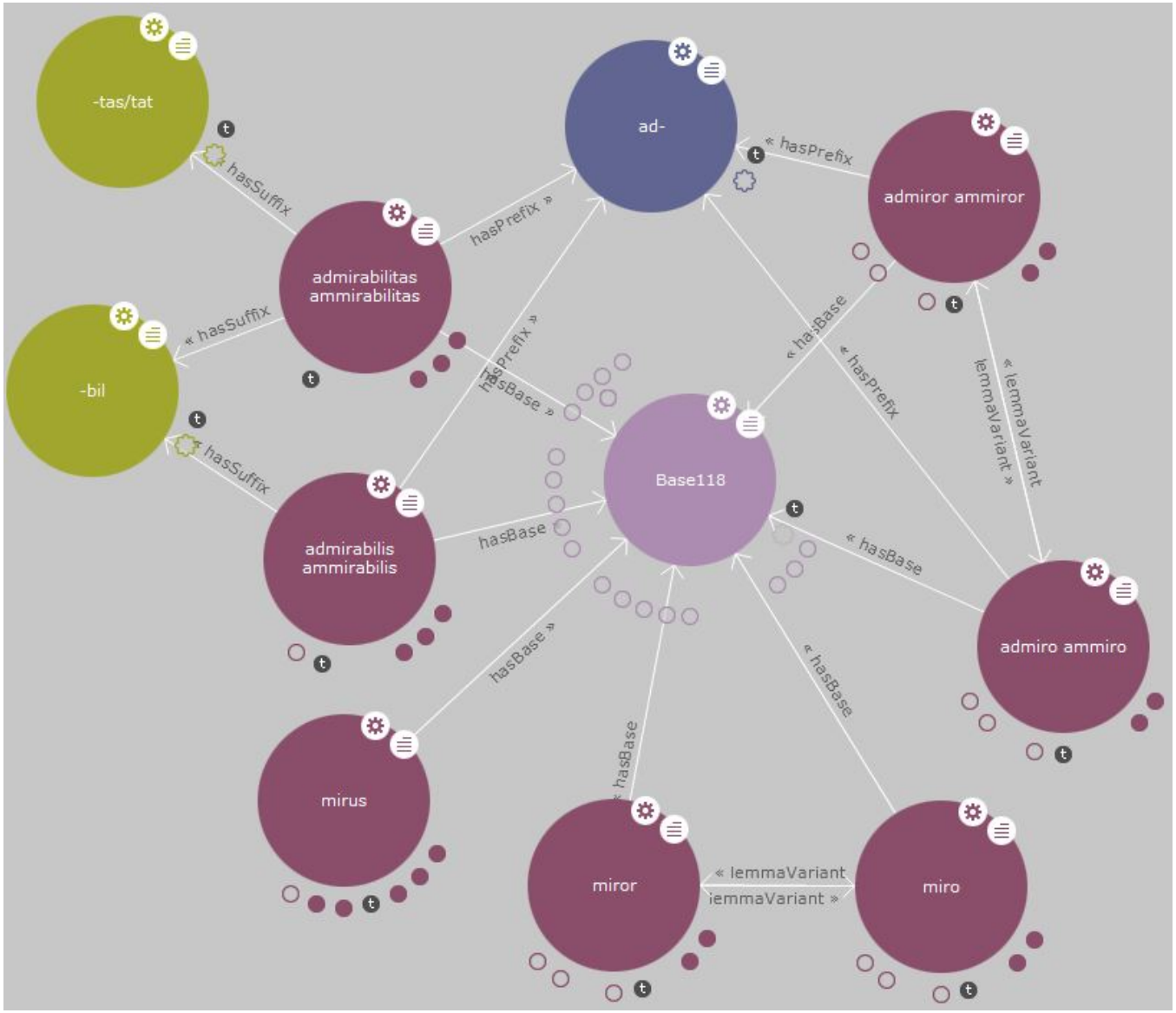


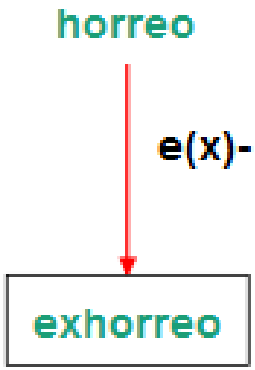
LiLa: Structure. Word Formation

LiLa: Paradigmatic

LiLa: Structure. Word Formation

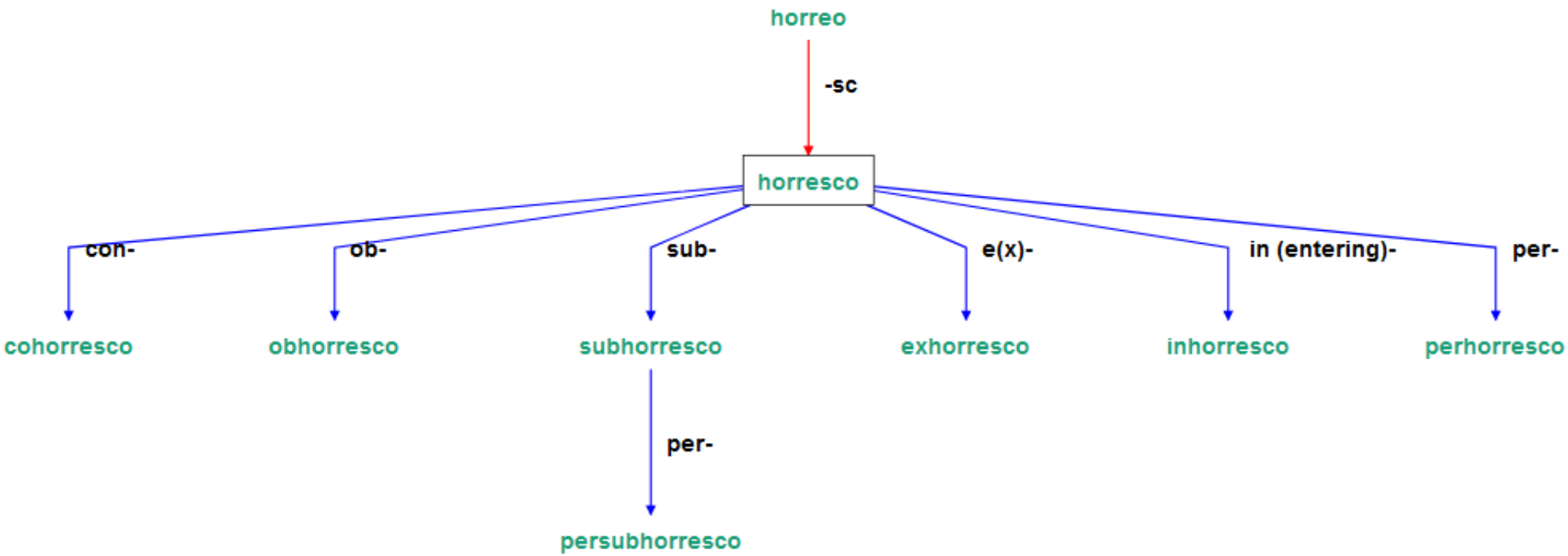
LiLa: Paradigmatic




horreo
e(x)-
exhorreo

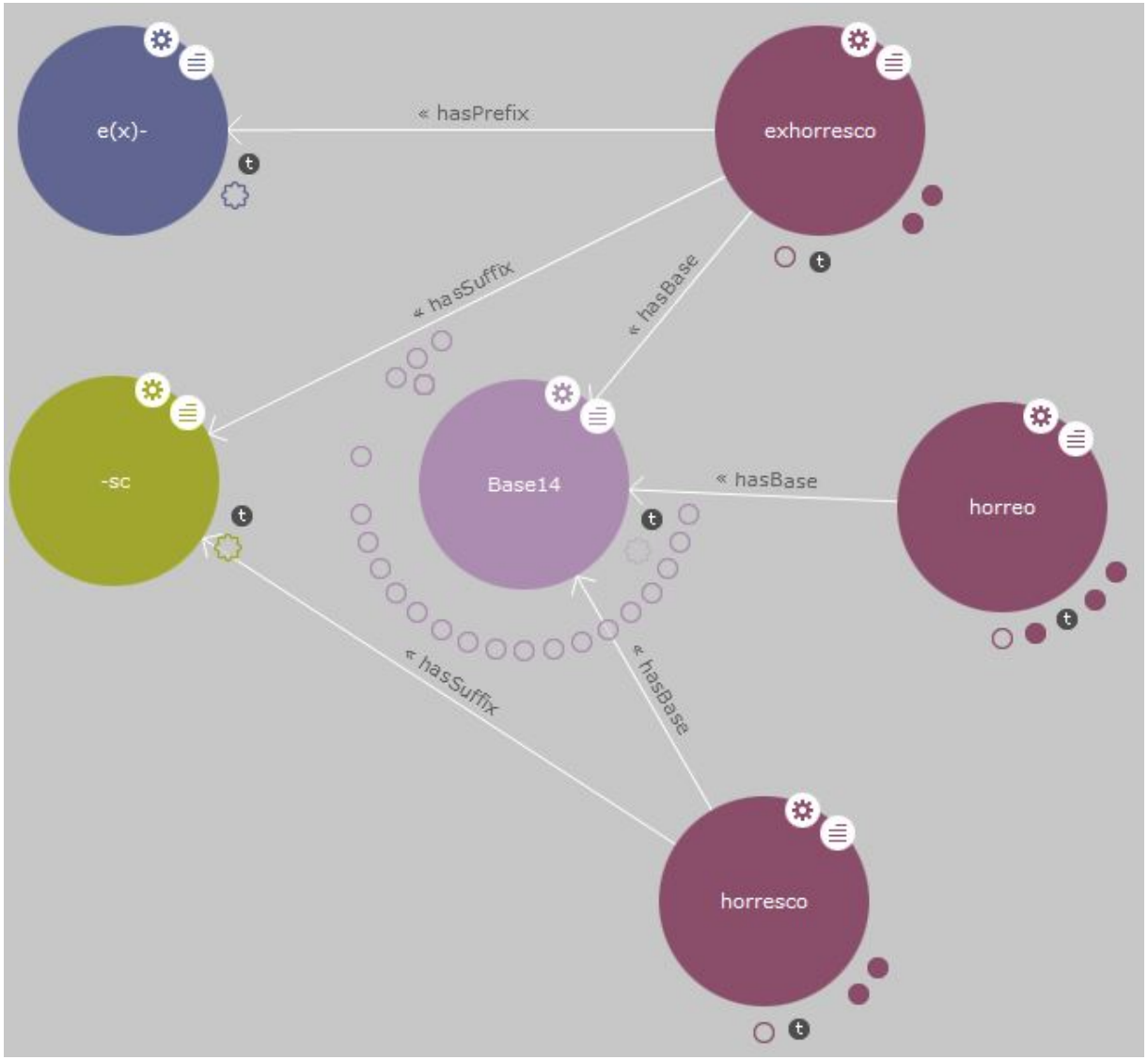
LiLa: Structure. Word Formation

WFL: Morphotactic and Hierarchical Derivation Trees. Who's First?



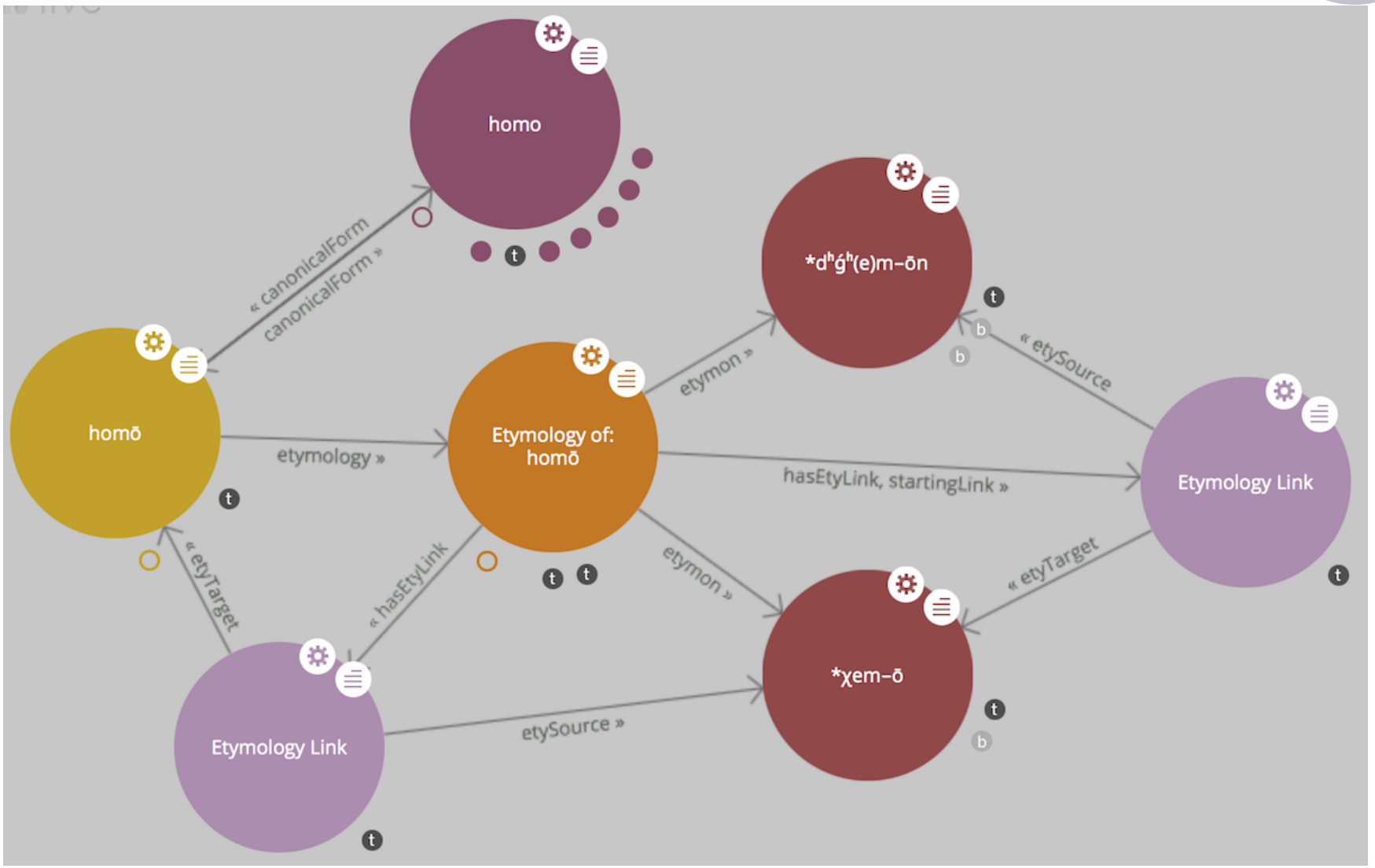
LiLa: Structure. Word Formation

LiLa: Paradigmatic. Nobody's First



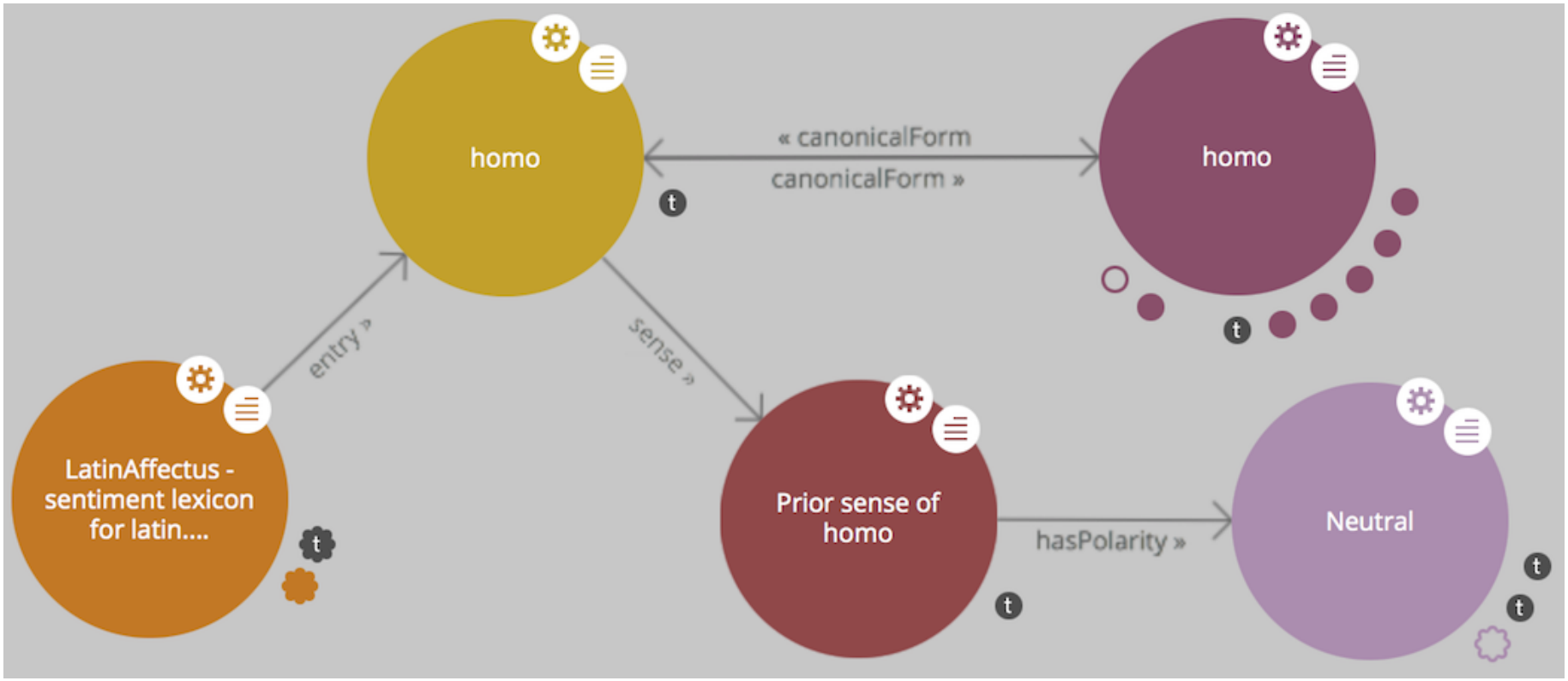
LiLa: Structure. Etymology

Etymological dictionary of Latin and the other Italic Languages



LiLa: Structure. Polarity

LatinAffectus

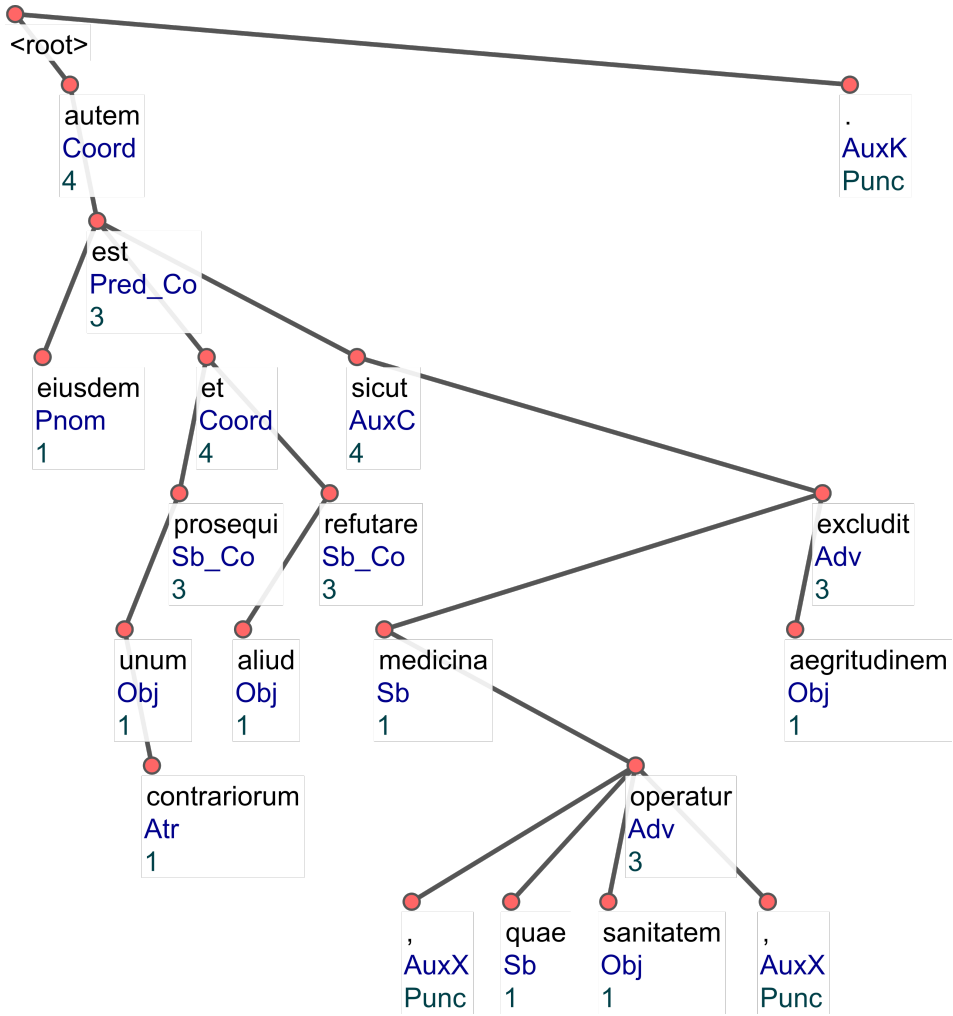


LiLa: Structure. All together now!

Lemma Bank, Word Formation, Etymology and Polarity

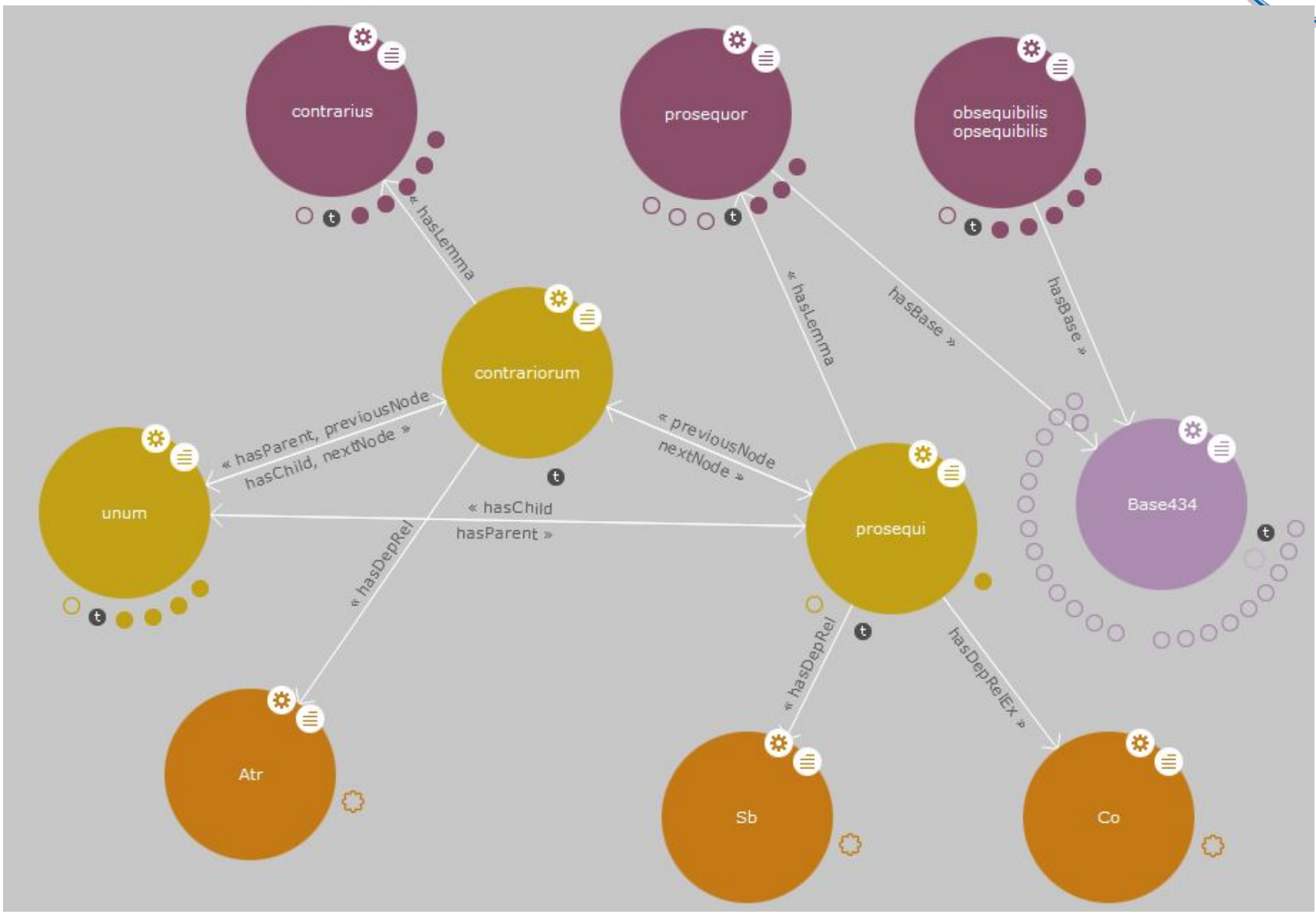
eiusdem autem est ***unum contrariorum prosequi et*** *aliud refutare sicut medicina , quae sanitatem operatur , aegritudinem excludit .* (IT-TB: SCG, lib. 1, cap. 1, n. 6)

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)



LiLa: Structure. Syntax

Phenomena and Noumena



LiLa: Structure

LodLive, Query interface, Corpora and Triplestore

https://lila-erc.eu/lodlive/
https://lila-erc.eu/query/
http://lila-erc.eu/data/corpora/ITTB/id/corpus
https://lila-erc.eu/data/corpora/DanteSearch/id/corpus
https://lila-erc.eu/sparql/

Thanks!
Get in touch

Marco Passarotti
Università Cattolica del Sacro Cuore
CIRCSE Research Centre

marco.passarotti@unicatt.it

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.