
# When Linked Open Data Speak Latin

## The LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti

Elsevier Guest Seminar series
March 24, 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
- The LiLa Lemma Bank

## LiLa now! lexica and corpora

- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

# Table of Contents



Introduction and fundamentals
- LiLa: mission and architecture
- The LiLa Lemma Bank

LiLa now! lexica and corpora
- Lexical Resources
- Textual Resources
- Text Linker
- 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

# 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

**Infra**structure

LiLa
Linking Latin

**Inter**operability

# 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)

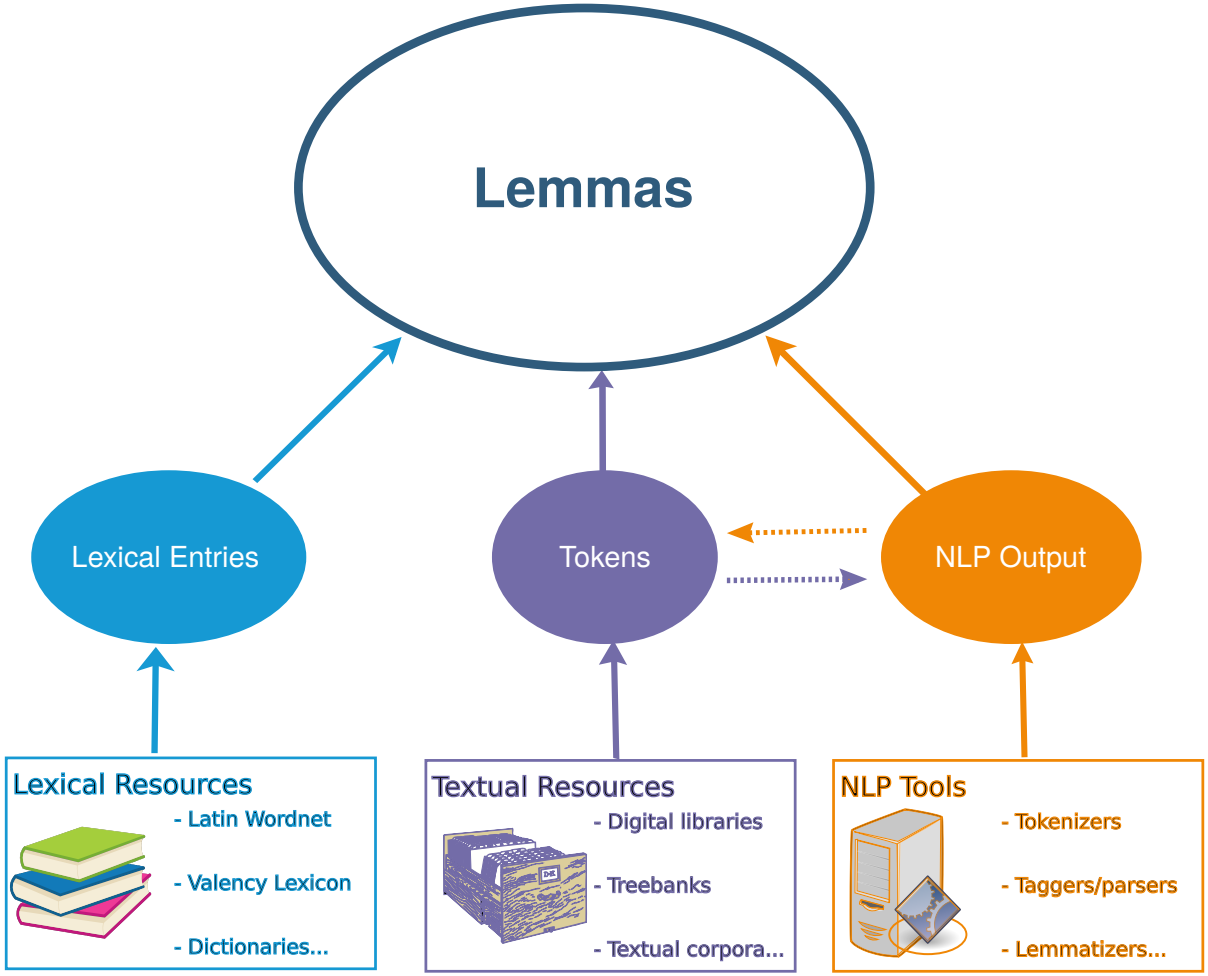


- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF
- Conceptual Interoperability: common ontologies to understand how to use the URIs
- Federation: to combine information from physically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

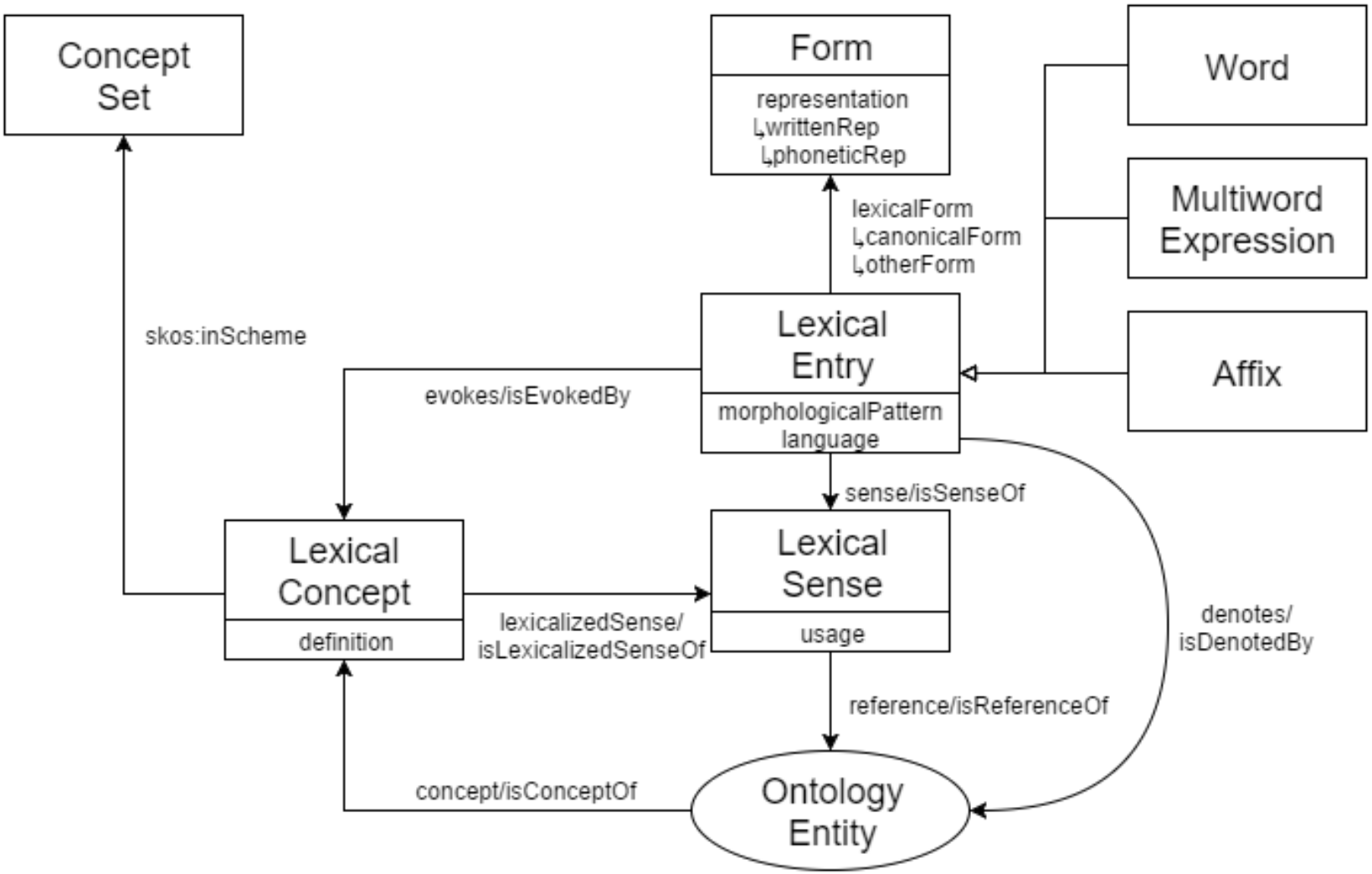
LiLa now! lexica and corpora

- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

# LiLa and Ontolex Lemon

A *de facto* W3C standard for publishing lexical data as LLOD





# The Lemma Bank

The core of LiLa. Source: LEMLAT (extended)



Lemma *admiror* ‘to admire, to respect’

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

# Derivational Morphology

Source: *Word Formation Latin* (CIRCSE Research Centre)

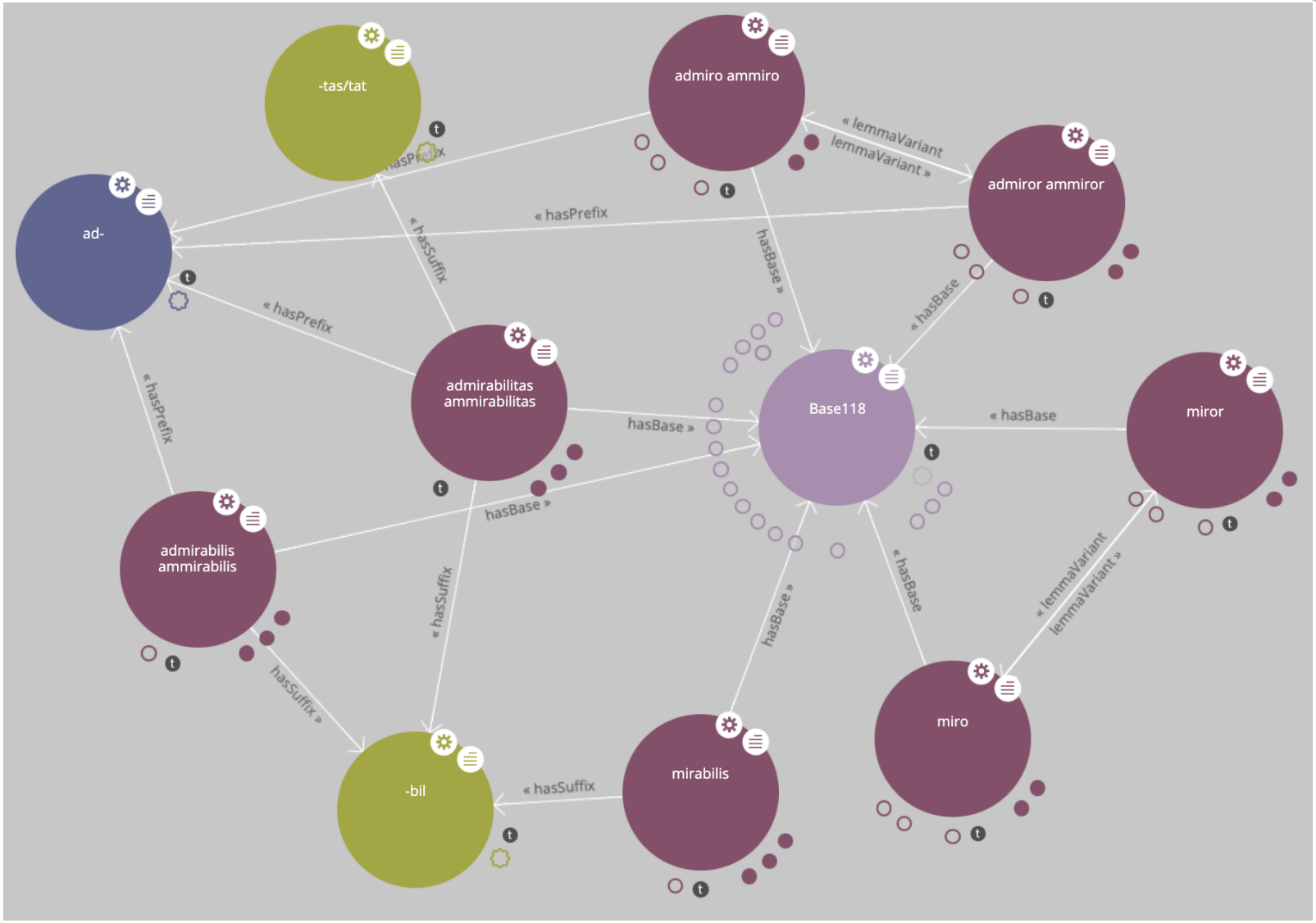


- WFL: directed tree-graphs indicating derivational path of each lemma (procedural)
- New Approach: Construction Morphology (declarative), words analysed in their internal structure
- WFL in LiLa:
  - Three classes of objects:
    1. Lemma
    2. Prefix and Suffix
    3. Base (connectors between lemmas of the same WF family)
  - Connected by three relationships:
    1. hasPrefix
    2. hasSuffix
    3. hasBase

# Derivational Morphology

Source: *Word Formation Latin* (CIRCSE Research Centre)





# Table of Contents

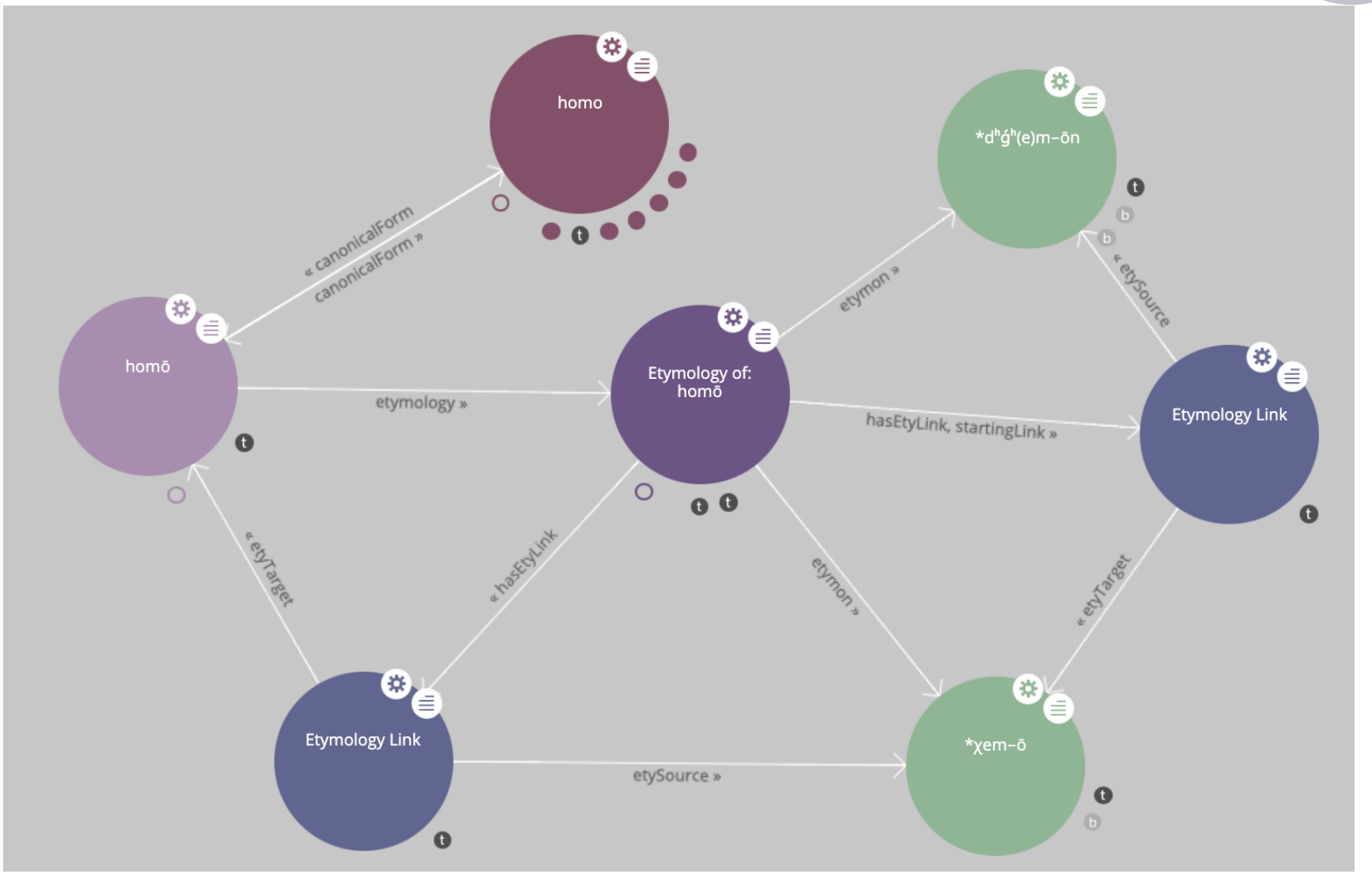
Introduction and fundamentals
- LiLa: mission and architecture
- The LiLa Lemma Bank

LiLa now! lexica and corpora
- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

# Etymology

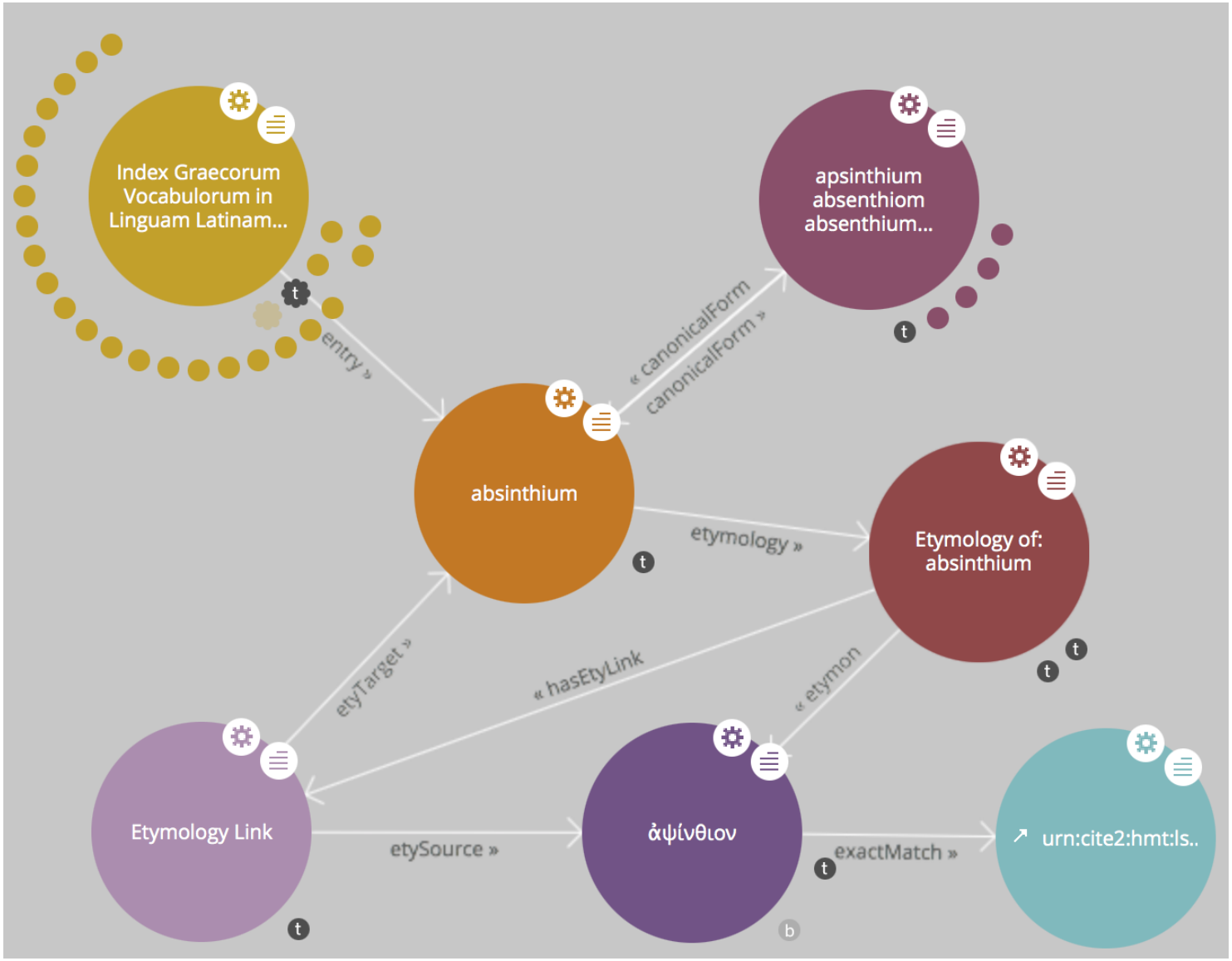
Source: *Etymological dictionary of Latin and the other Italic Languages* (De Vaan, 2008)





# Etymology

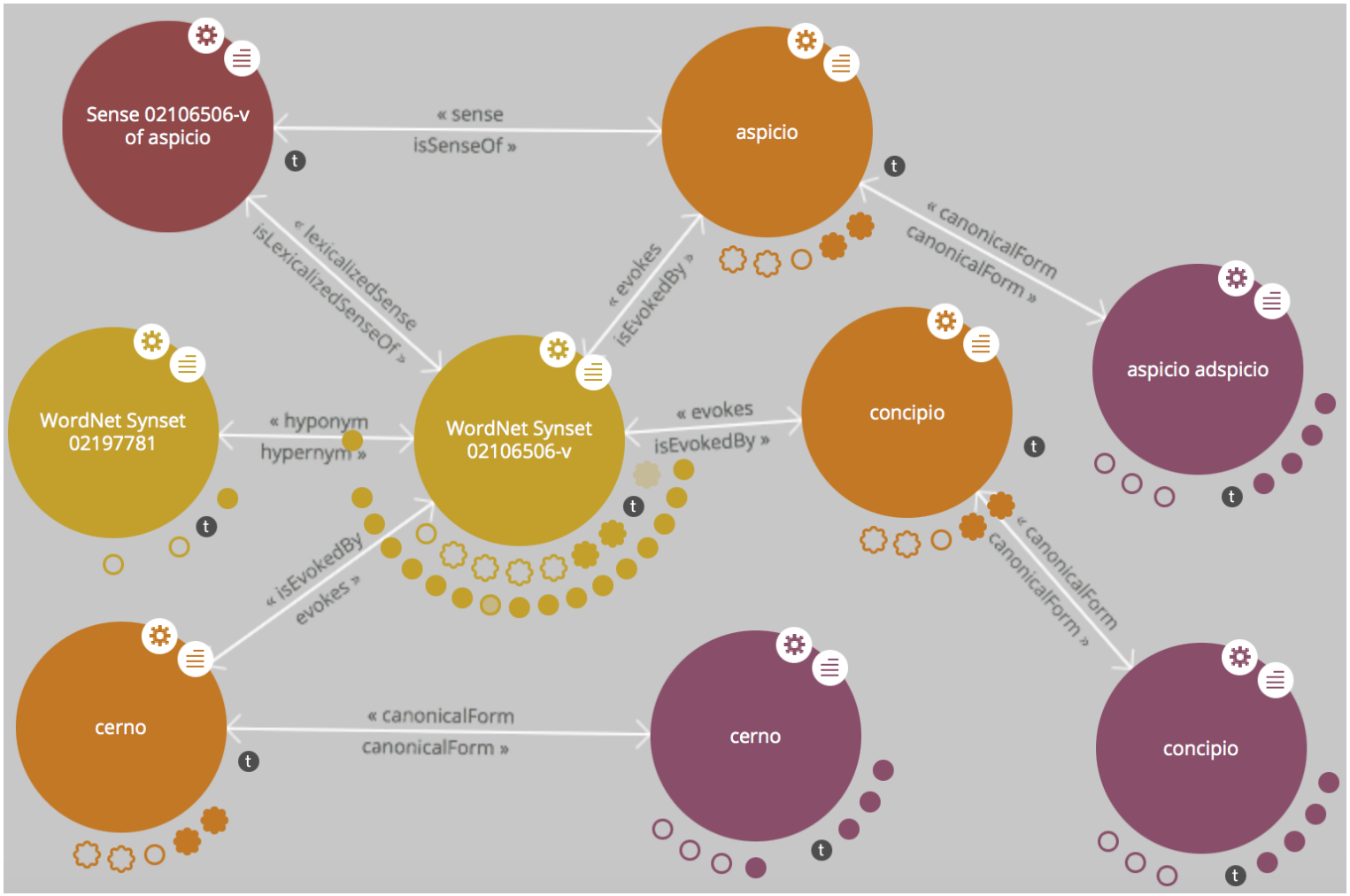
Source: *Index Graecorum Vocabulorum in Linguam Latinam* (Saalfeld, 1874)



# Latin WordNet

## Lexical Entries, Senses and Concepts

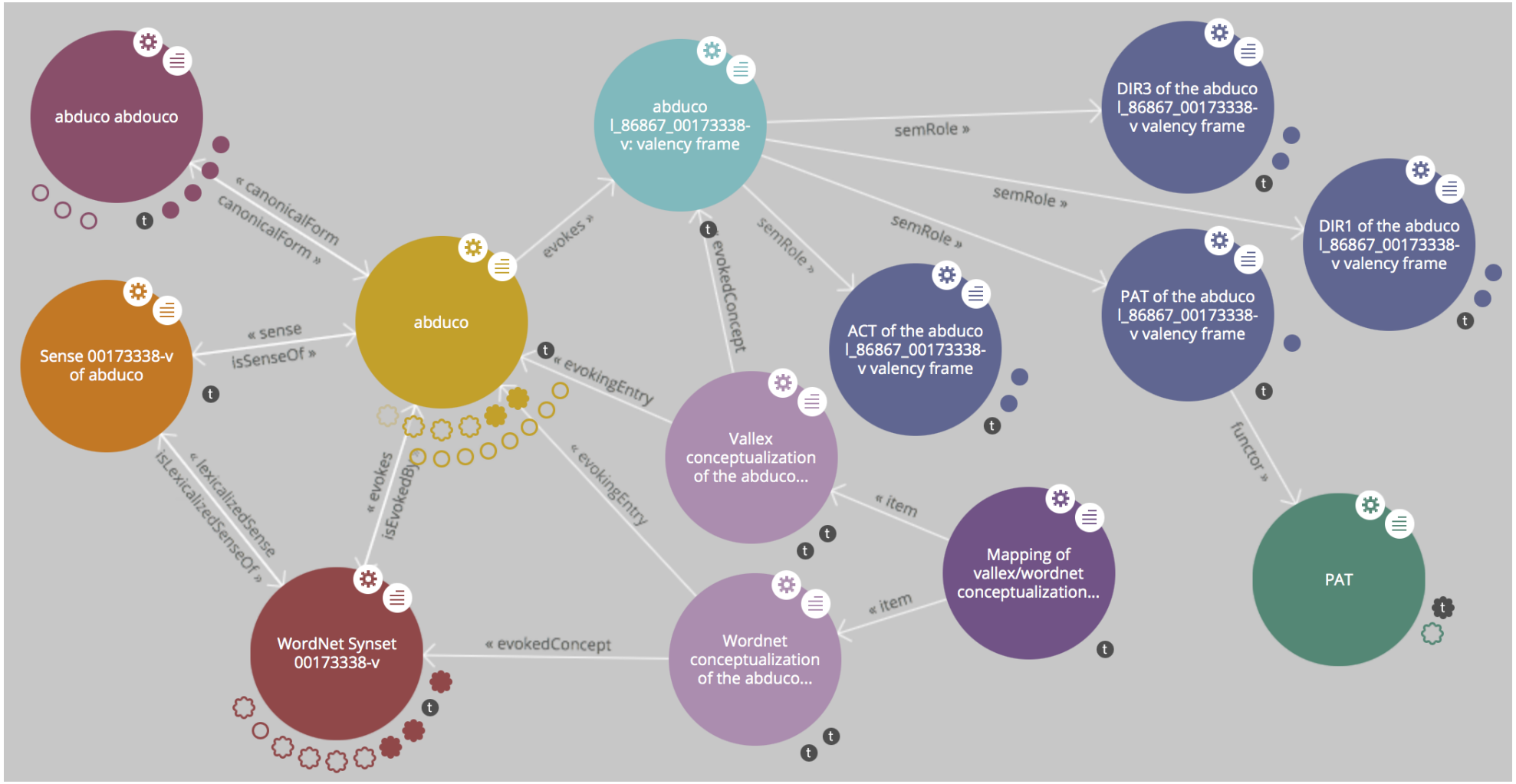




# Latin WordNet and Latin VALLEX

## Mapping Senses and Frames





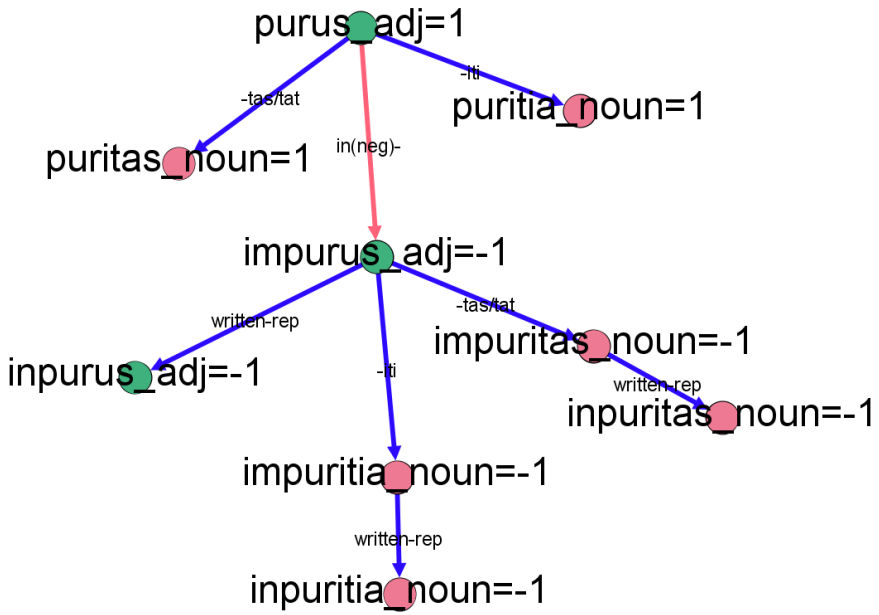
# Polarity

Source: *Latin Affectus* (CIRCSE Research Centre)



List of 2,437 nouns and adjectives associated to their **out-of-context sentiment score**: from -1 (very negative, e.g. *abominatio*) to +1 (very positive, e.g. *amor*)

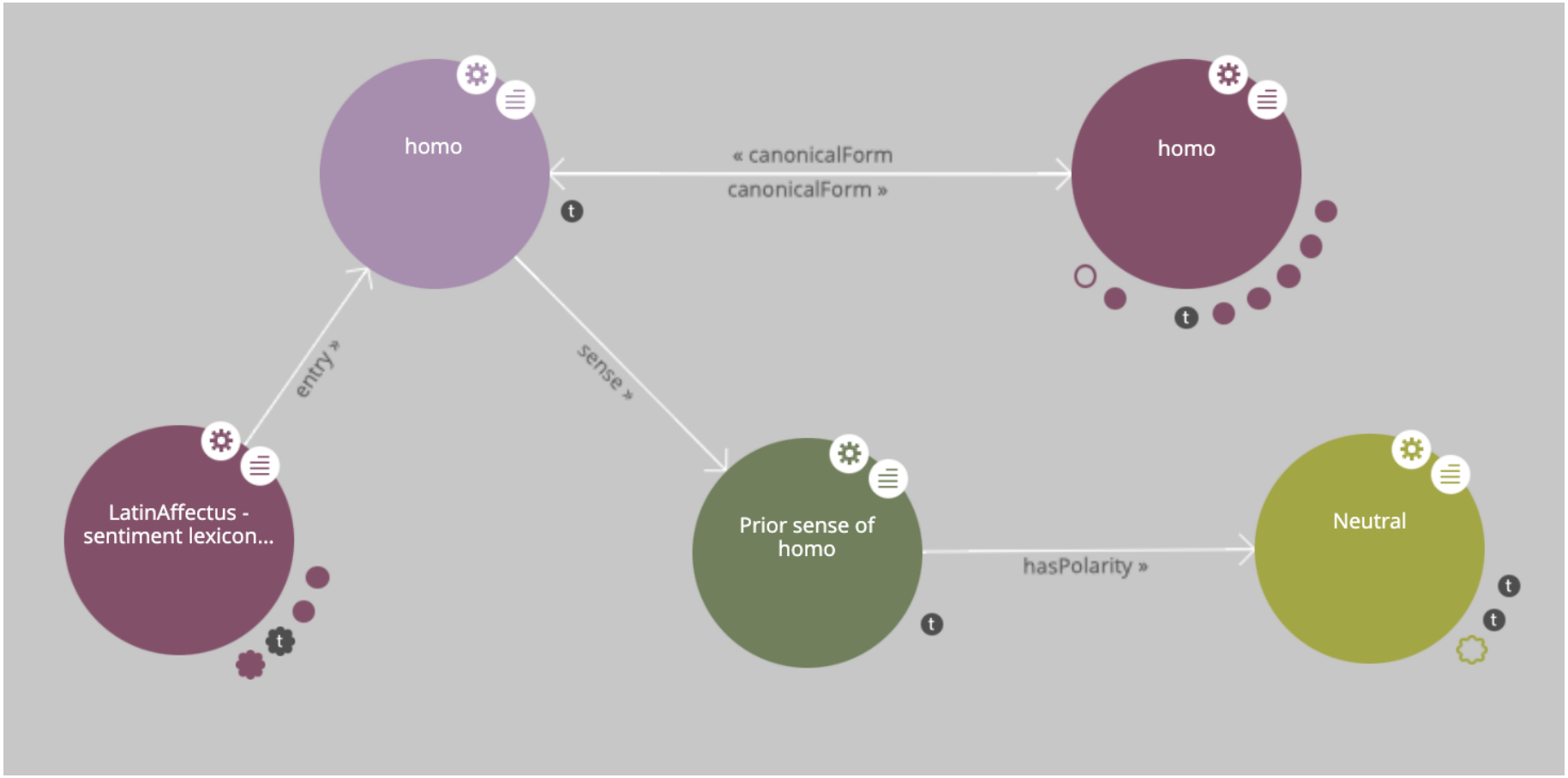
- **Gold Standard**: manually created by 2 Latin language and culture experts + 1 supervisor
- **Silver Standard**: automatically created by deriving new entries from the Gold Standard



# Polarity

Source: *Latin Affectus* (CIRCSE Research Centre)

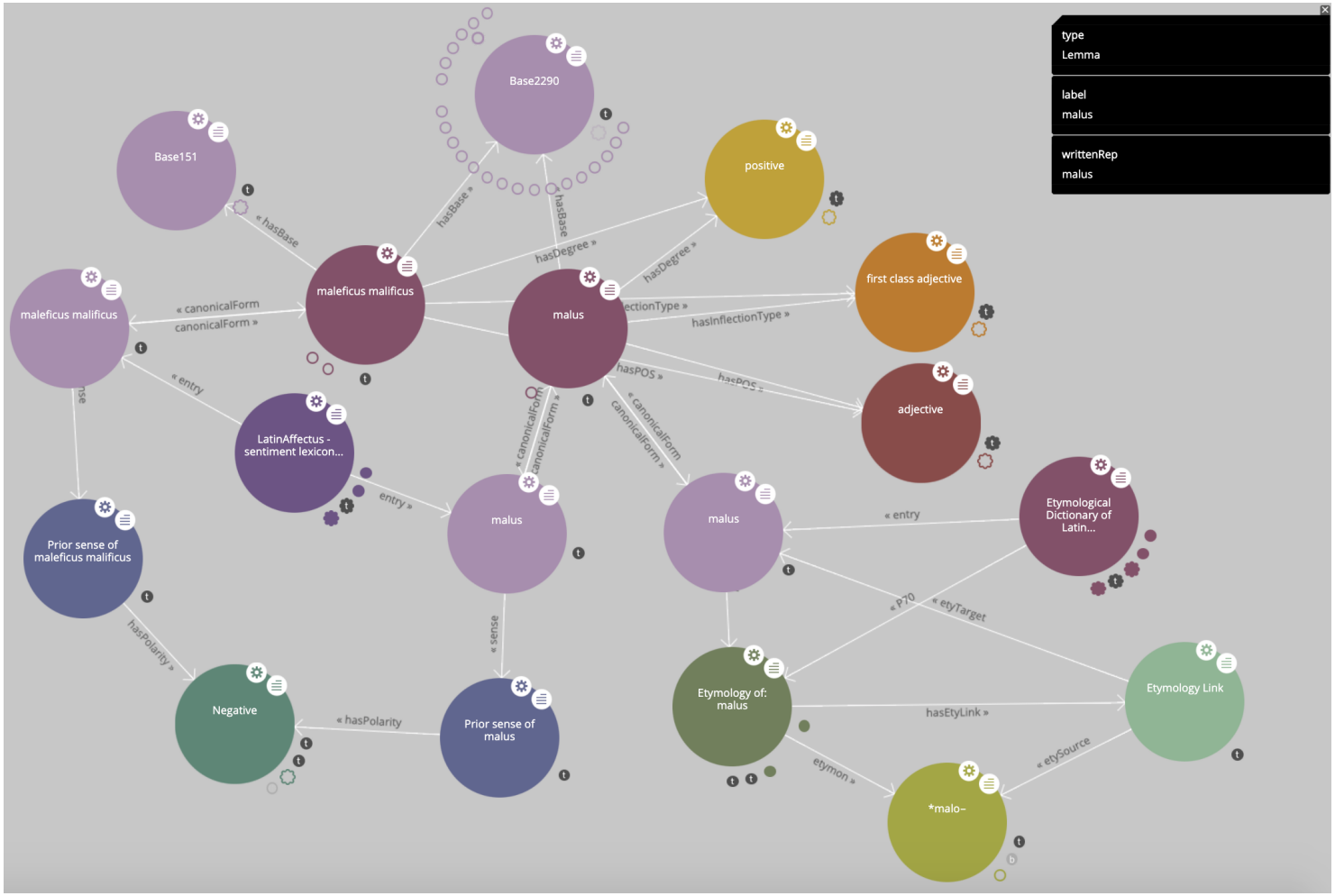




# LiLa: Structure. All together now!

## Lemma Bank, Derivational Morphology, Etymology and Polarity





# Table of Contents



Introduction and fundamentals
- LiLa: mission and architecture
- The LiLa Lemma Bank

LiLa now! lexica and corpora
- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

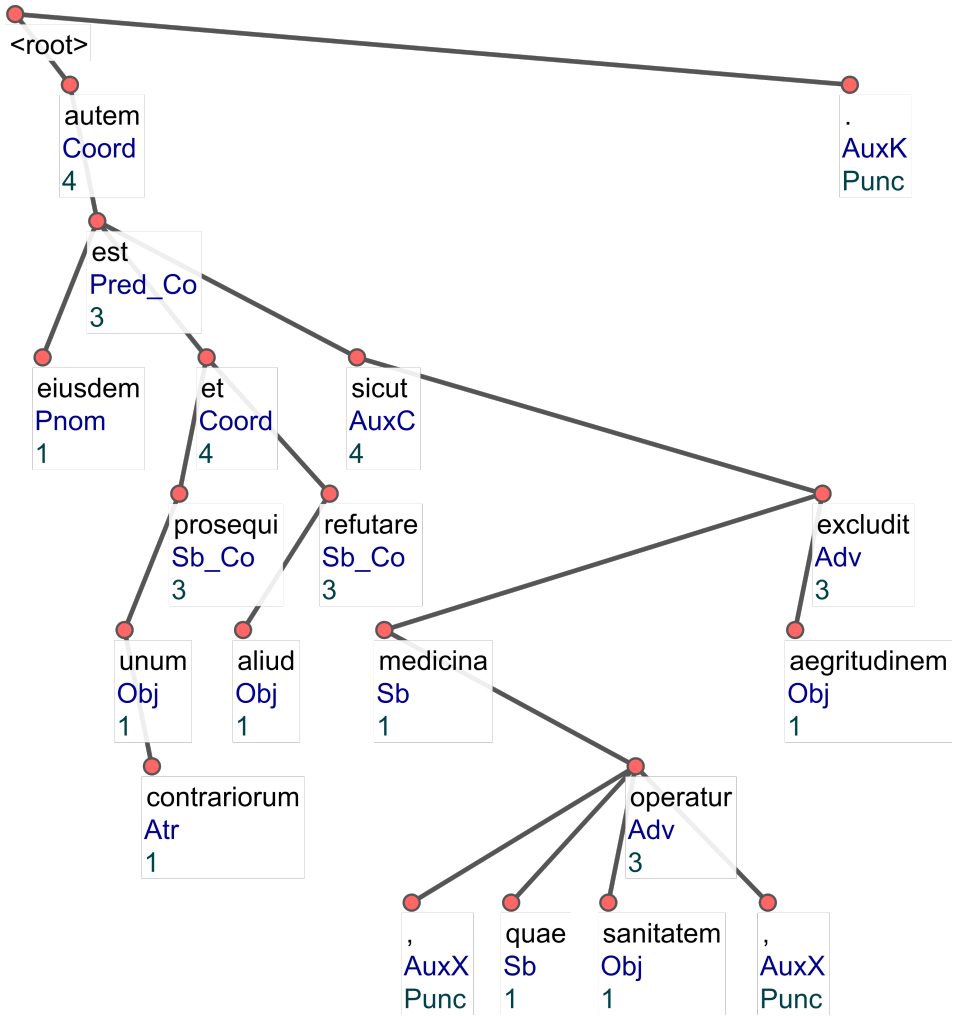
# (Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (CIRCSE Research Centre): Dependency trees



***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)

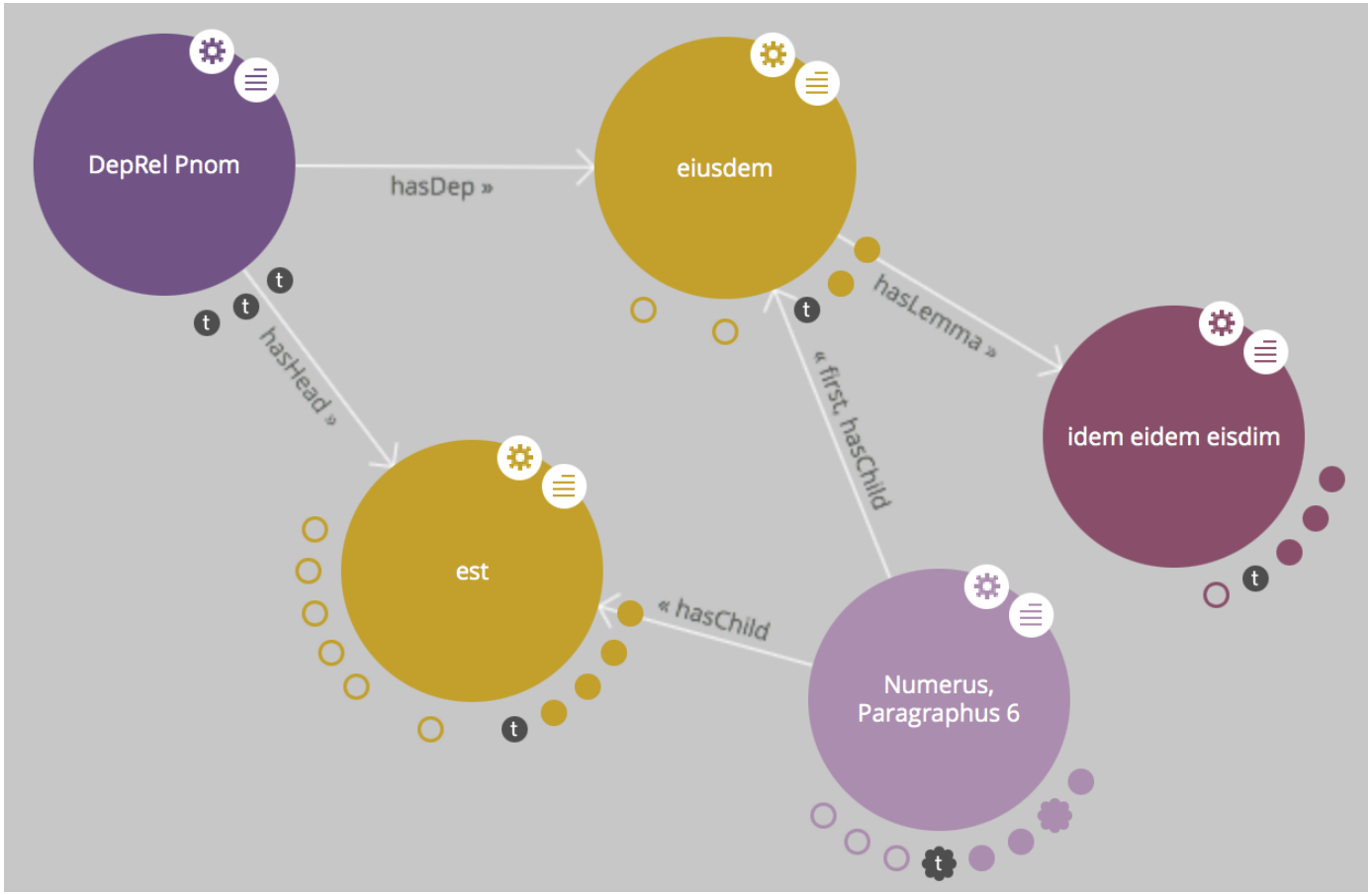
*Now* ***it belongs to the same thing*** *to pursue one contrary and to remove the other: thus medicine, which effects health, removes sickness.* (Trans. Laurence Shapcote)



# Texts, tokens, relations and lemmas

## Phenomena and noumena





# Table of Contents



Introduction and fundamentals
- LiLa: mission and architecture
- The LiLa Lemma Bank

LiLa now! lexica and corpora
- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

# TextLinker
## Welcome screen




Figure: LiLa's Text Linker

# TextLinker
## Processed output



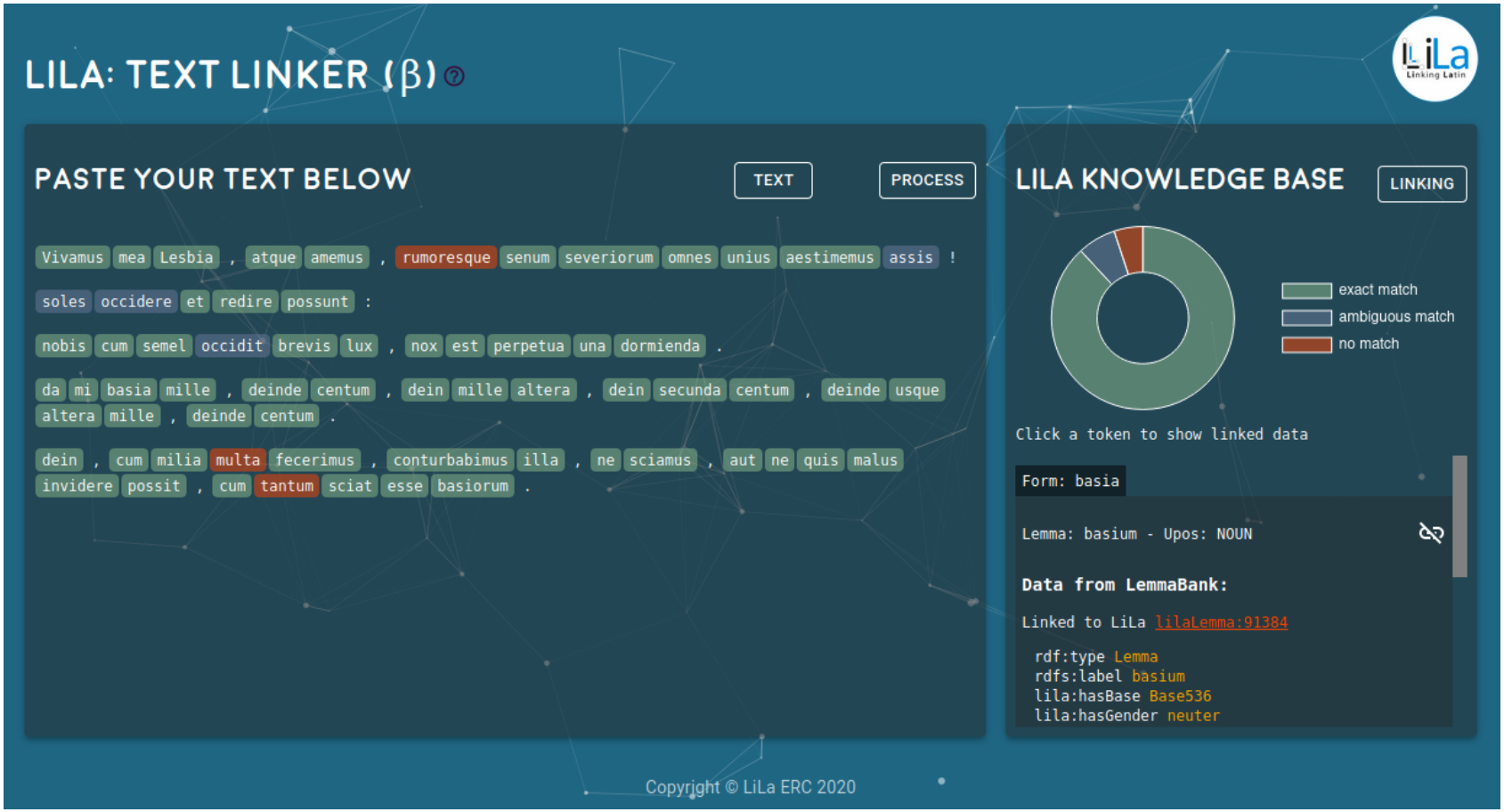



Figure: Text processed against the LiLa Knowledge Base

# Table of Contents



Introduction and fundamentals
- LiLa: mission and architecture
- The LiLa Lemma Bank

LiLa now! lexica and corpora
- Lexical Resources
- Textual Resources
- Text Linker
- To sum up

# LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] Dante Search (700th death anniversary coming up!): ca. 46,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics and LASLA Corpora
- **Lexica**
  - [x] Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
  - [x] Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
  - [x] LatinAffectus: ca. 2,300 entries
  - [x] Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
  - [x] Latin WordNet: ca. 1,000 manually checked entries
  - [x] Latin Vallex 2.0: Valency Lexicon
  - [ ] Lewis & Short Dictionary
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 13 million triples**

## Query Interface, Triplestore and Linker

- https://lila-erc.eu/query/; https://lila-erc.eu/sparql/
- http://lila-erc.eu:8080/LiLaTextLinker/

## Linguistic Resources. Corpora

- https://lila-erc.eu/data/corpora/ITTB/id/corpus
- https://lila-erc.eu/data/corpora/DanteSearch/id/corpus
- https://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia

## Linguistic Resources. Lexica

- https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon
- https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon
- https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon
- http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon
- https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon

# 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.