LiLa: Linking Latin

Building a Knowledge Base of Interlinked Linguistic Resources for Latin

The LiLa team

Conference *Linked Pasts 6*
University of London and British Library
December 2-16, 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.

Overview

Table of Contents

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

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

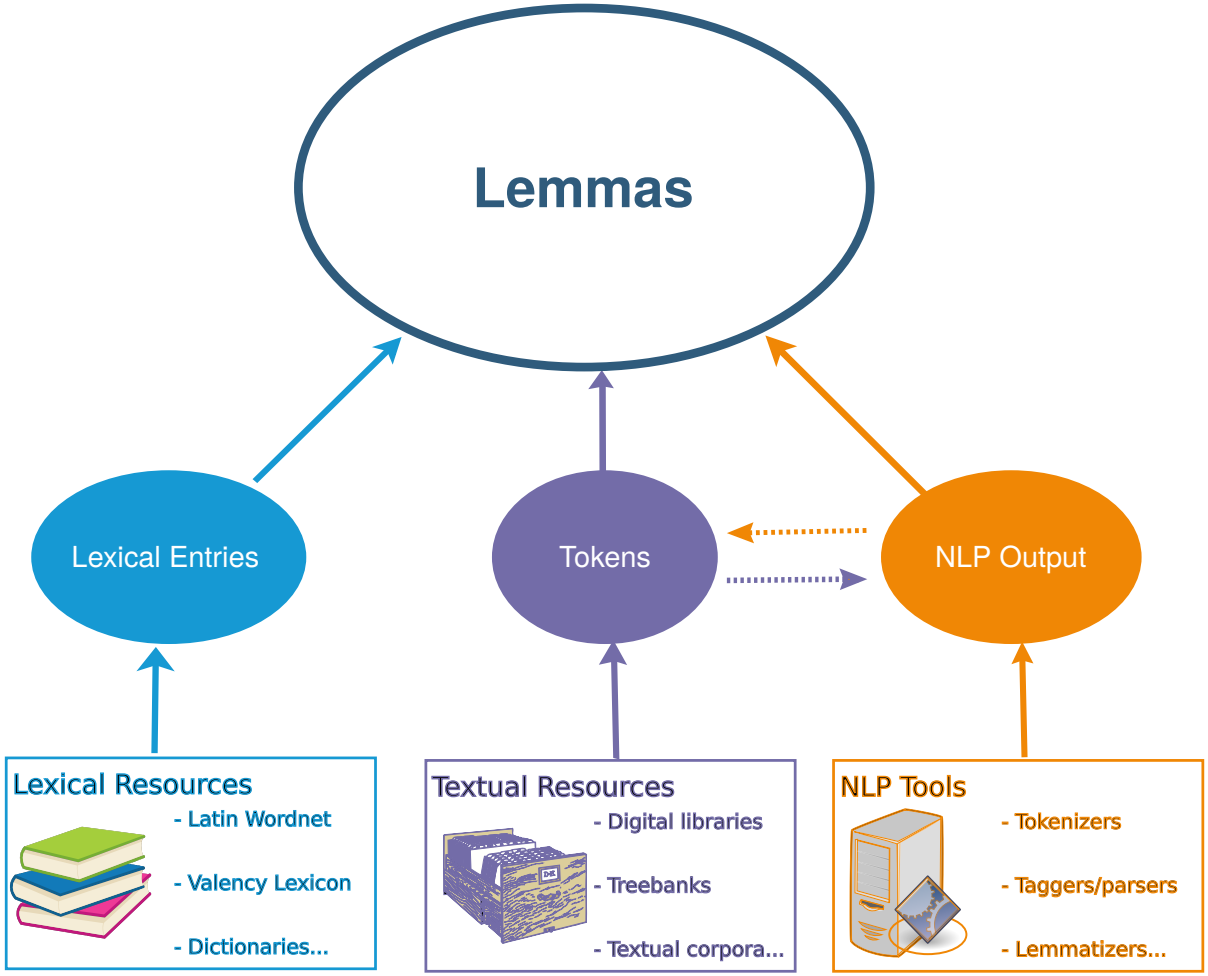
Benefits of Applying LD to Linguistic Resources

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

Lemmatisation & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

Goals

Lemmatisation and **part-of-speech tagging** (POS-tagging) aim to **abstract** some linguistic properties to allow **form-invariant** reference to words/tokens.

?! How can I retrieve all occurrences of a word in a text?

?! How can I know which roles a word plays in a text?

Different contexts harbor different word forms…

- …*his rebus cognitis Caesar Gallorum animos* ***verbis*** *confirmavit*…
  → ablative plural
- …*quod ego si* ***verbo*** *adsequi possem*…
  → ablative singular
- …*ne more iuvencae mugiat, et timide* ***verba*** *intermissa retemptat*…
  → accusative singular
- …

…but each can be referred to a canonical form:

⇒ ***uerbum***
  → nominative singular of neuter II. declension noun

Lemmatisation

is the process of assigning each token in a text a **standardised** corresponding word form, based on

- morphological paradigms (which declension, conjugation?)
- etymology (same lexical base?)
- lexicography (is it registered as an entry in a dictionary?)
- graphic normalisation ($v \rightarrow u$, $j \rightarrow i$; *condi**c**io* $\rightarrow$ *condi**t**io*, ...)

Lemmatisation & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

Words play different (syntactic) roles in sentences:

- ⋆ ***supra***
  - ▶ *. . . ager trecentis aut etiam* ***supra*** *nummorum milibus emptus. . .*
    - → adverb (ADV)
  - ▶ *. . . ille qui* ***supra*** *nos habitat. . .*
    - → preposition (ADP)
- ⋆ ***scribo***
  - ▶ *. . . atque in Thesauro* ***scripsit*** *causam dicere prius unde petitur. . .*
    - → verb (VERB)
- ⋆ ***elephantus***
  - ▶ *. . .* ***elephanto*** *beluarum nulla prudentior. . .*
    - → noun (NOUN)

These roles are predictable and come from a rather small set of alternatives.

Part-of-speech tagging

is the process of assigning each token in a text one of a given set of **roles**, i. e. parts of speech, mainly based on **morphologic** and **syntactic** criteria.

LiLa is oriented towards the formalism of **Universal Dependencies**

16+1 classes: ADJ (adjectives), ADP (pre- & postpositions), ADV (adverbs), AUX (auxiliaries), CCONJ & SCONJ (co-ordinating & subordinating conjunctions), DET (determiners), INTJ (interjections), NOUN & PROPN (common & proper nouns), NUM (numerals), PART (particles), PRON (pronouns), VERB (verbs), SYM (symbols), X (other) + PUNCT (punctuation)

`https://universaldependencies.org`

Lemmatisation & Part-of-Speech Tagging

Putting everything together

On a big set of documents either natively annotated with, or converted into, the UD formalism, such as…

- ☑ *Summa contra Gentiles* (Thomas Aquinas)
- ☑ LASLA corpus
- ☑ CompHistSem corpus
- ☑ Dante's Latin works
- ☑ *Confessiones* (Augustine)
- ☑ …and other texts

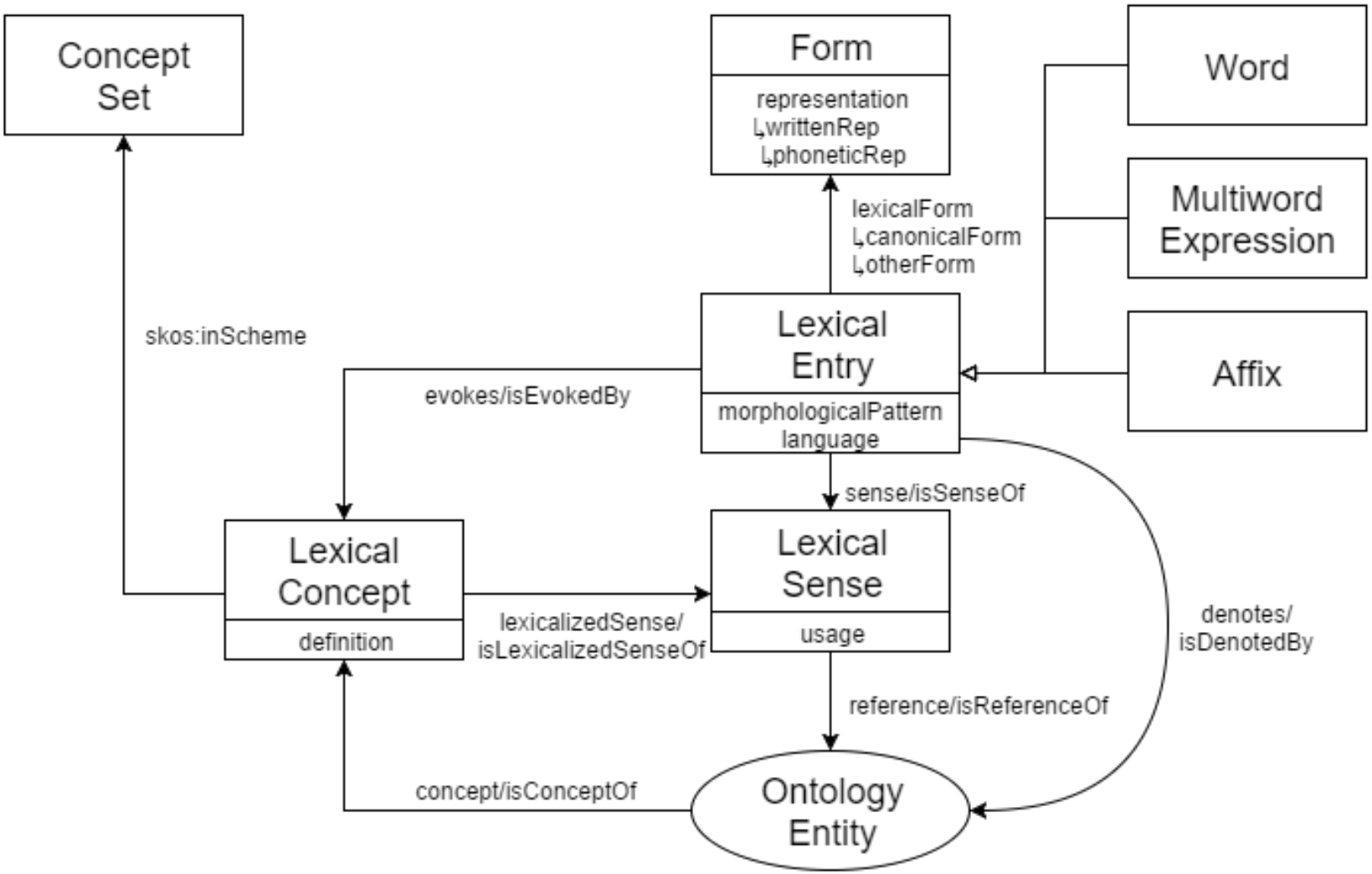
…we used the POS-tagger tool **UDPipe** to train a **big, comprehensive model**!

This **Big Model** is capable of good LiLa-compliant lemmatisation and part-of-speech tagging over a wide variety of Latin documents!

Table of Contents

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)

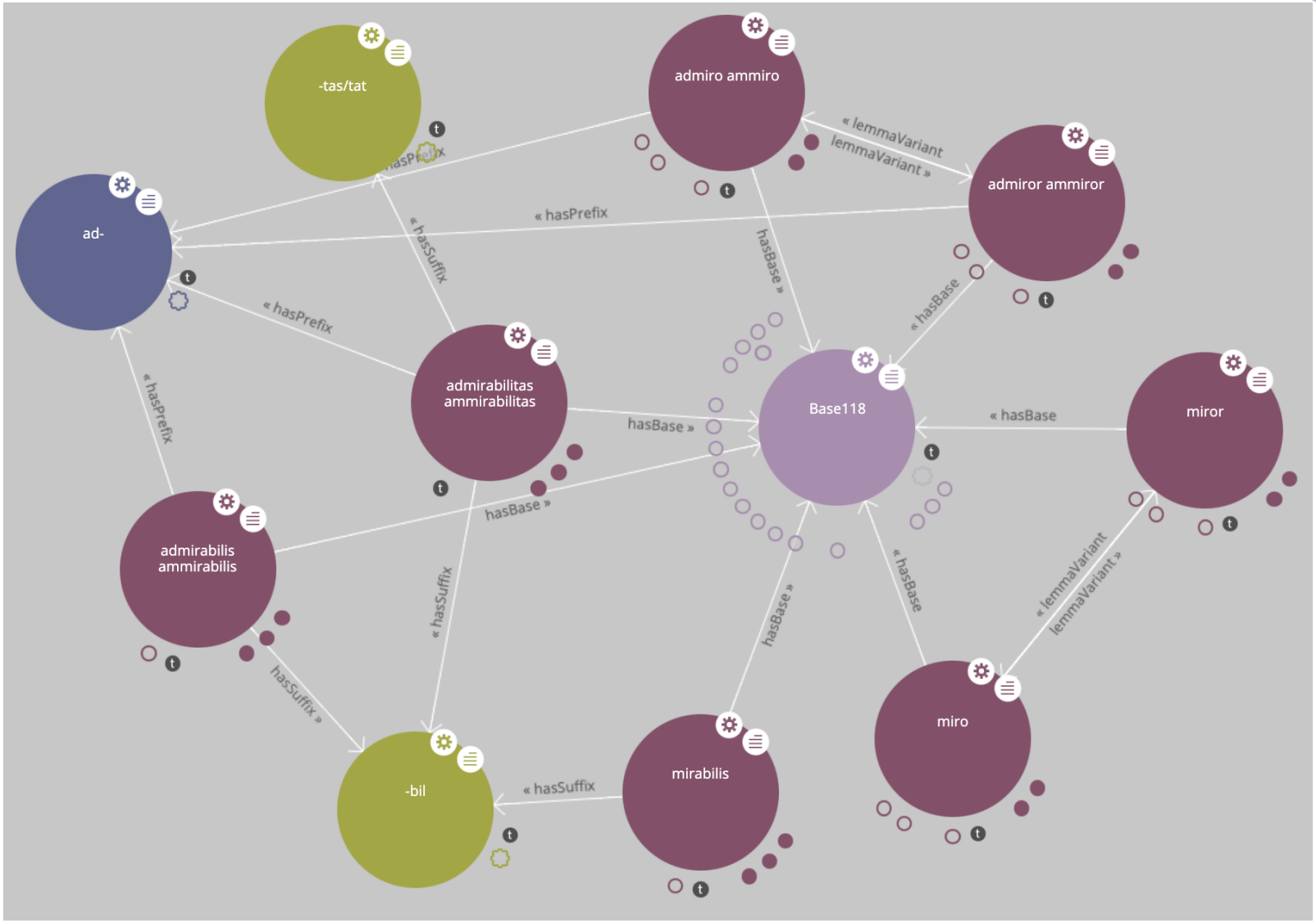
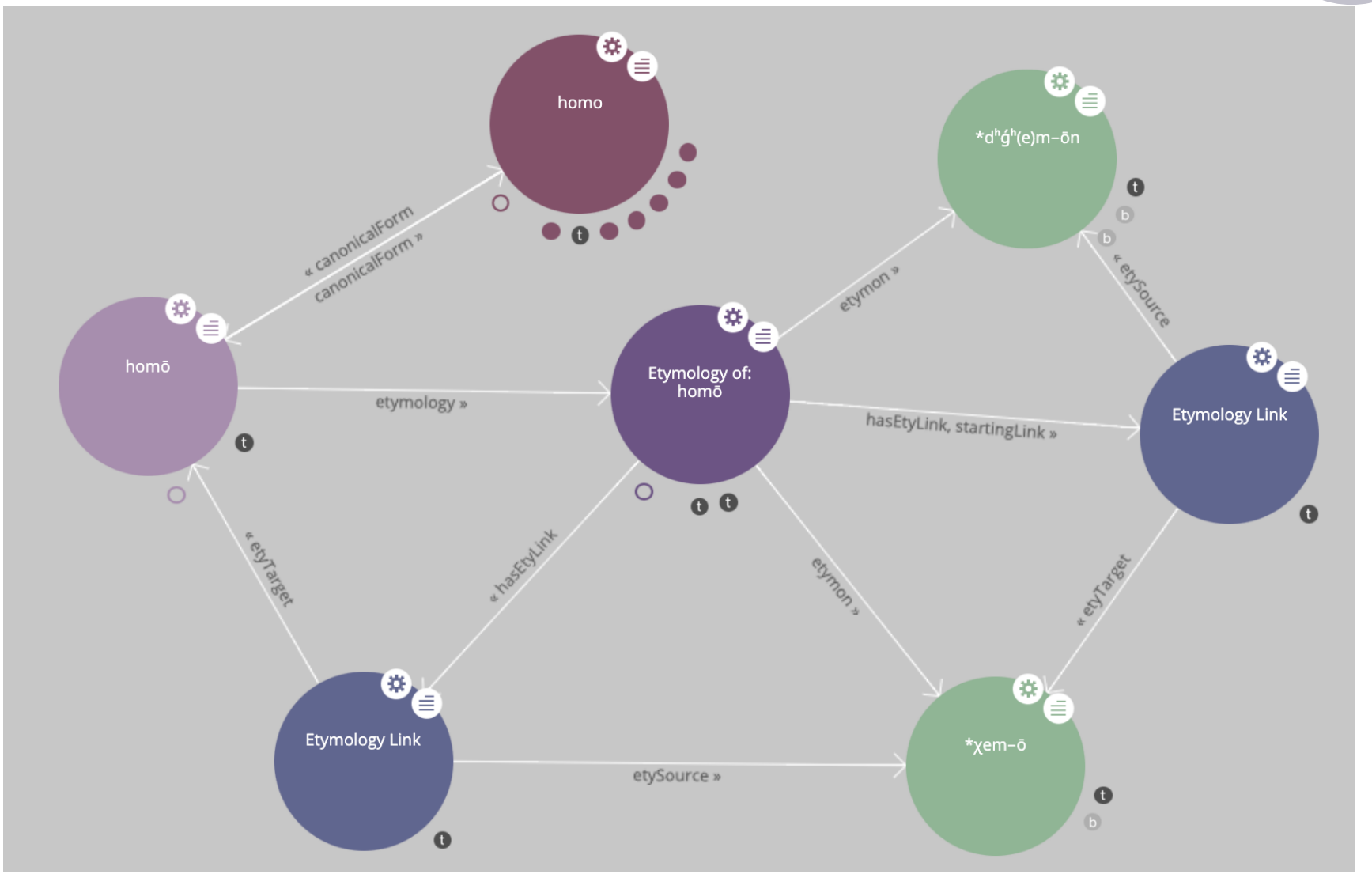


Table of Contents

Etymology

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



Etymology

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

Etymology

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

Borrowing: absinthe (ENG) ← **absinthium (LAT)** ← ἀψίνθιον (GRC)

Etymology

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

Borrowing: absinthe (ENG) ← **absinthium (LAT)** ← ἀψίνθιον (GRC)

- Saalfeld's Index of 1,763 Ancient Greek loanwords (1874)

Etymology

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

Borrowing: absinthe (ENG) ← **absinthium (LAT)** ← ἀψίνθιον (GRC)

- Saalfeld's Index of 1,763 Ancient Greek loanwords (1874)
- Same ontological model as BRILL

Etymology

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

Borrowing: absinthe (ENG) ← **absinthium (LAT)** ← ἀψίνθιον (GRC)

- Saalfeld's Index of 1,763 Ancient Greek loanwords (1874)
- Same ontological model as BRILL
- Mapping to the *Liddell Scott Jones Lexicon* (CITE application)

urn:cite2:hmt:lsj.chicago_md:n18890 Link

ἀψίνθ-ιον

ἀψίνθιον, τό,

A

wormwood, Artemisia Absinthium, Hp. *Morb.* 3.11, *Mul.* 1.74, X. *An.* 1.5.1, Thphr. *HP* 1.12.1, Dsc. 3.23; ἀψινθίῳ κατέπασας Ἀττικὸν μέλι Men. 708 :—also ἄψινθος, ἡ, Aret. *CD* 1.13, but ὁ, *Apoc.* 8.11; and ἀψινθία, ἡ, Alex. Trall. 1.10.

A.II

ἀψίνθιον, = ἀβρότονον, Ps.- Dsc. 3.24.

A.II.2

= **Artemisia monosperma**, Aq. *Pr.* 5.4.

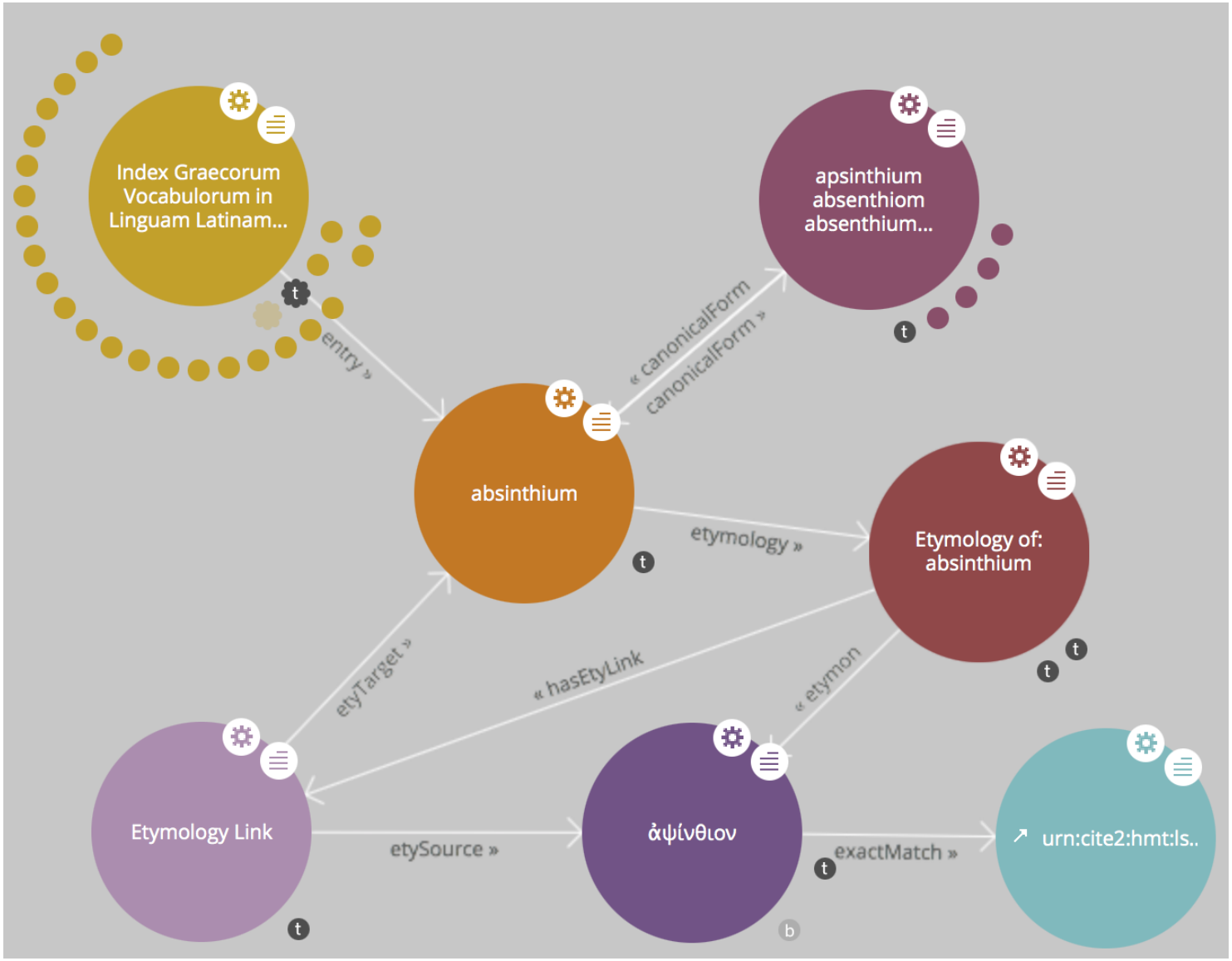
A.II.3

ἀ. θαλάσσιον, = σέριφον, Dsc. 3.23.

Figure: Liddell Scott Jones entry for ἀψίνθιον

Etymology

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



Minozzi's Latin WordNet:

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds
  - relations between synsets: antonymy, hyponymy, meronymy, etc.

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds
  - relations between synsets: antonymy, hyponymy, meronymy, etc.
- automatically derived from the Princeton WordNet (2004, Minozzi)

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds
  - relations between synsets: antonymy, hyponymy, meronymy, etc.
- automatically derived from the Princeton WordNet (2004, Minozzi)
- 9,378 lemmas distributed across 8,973 synsets

Manual removal of noise:

Minozzi's Latin WordNet:

- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds
  - relations between synsets: antonymy, hyponymy, meronymy, etc.
- automatically derived from the Princeton WordNet (2004, Minozzi)
- 9,378 lemmas distributed across 8,973 synsets

Manual removal of noise:

- Senses not applicable to Latin, e.g. *voco*, v
  - send a message or attempt to reach someone by radio, phone, etc.

Minozzi's Latin WordNet:

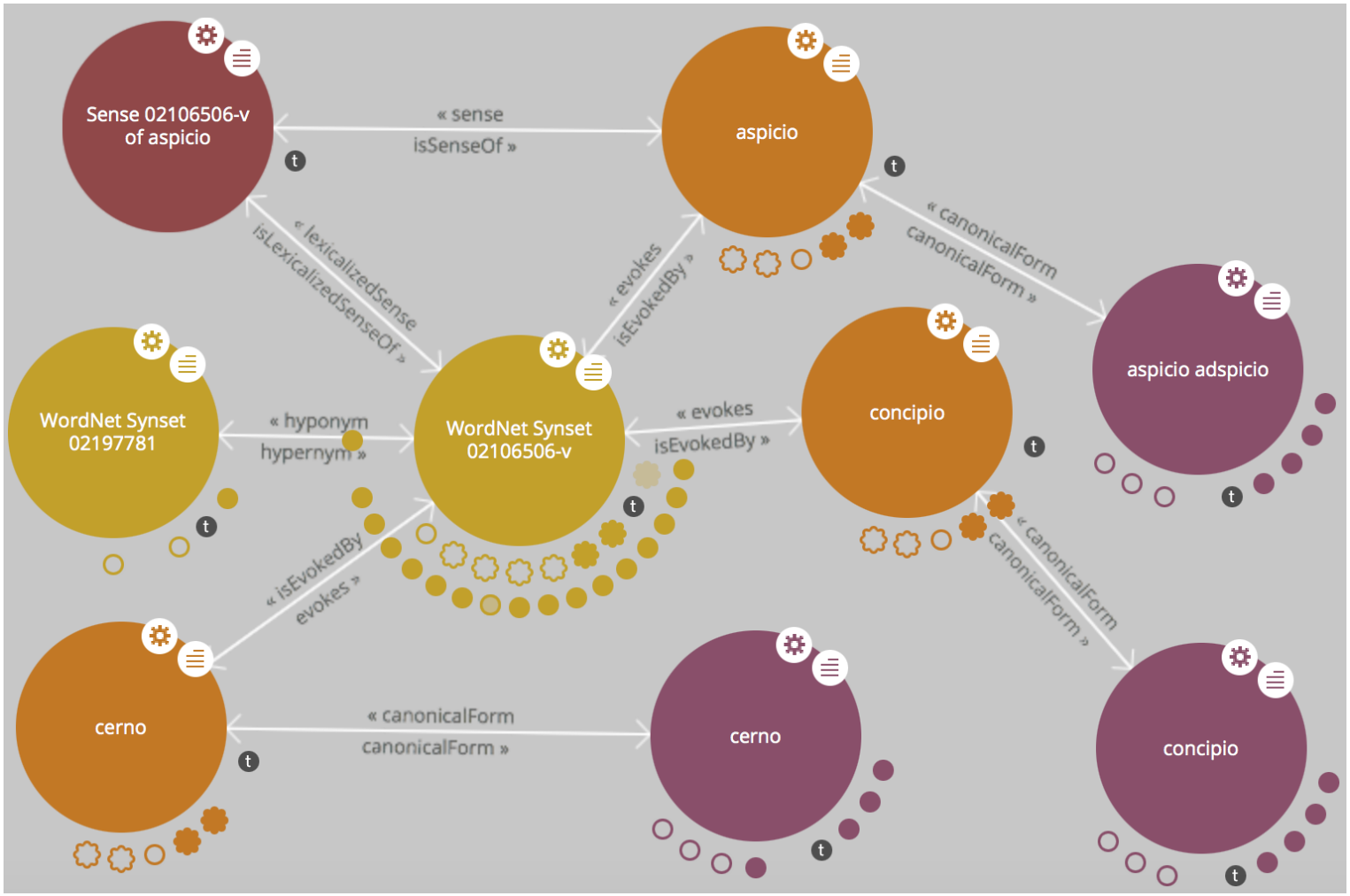
- dictionary of *synsets* (sets of synonymous lemmas sharing a sense)
  - a#00430275 - *nubilosus*, *nubilus* - full of or covered with clouds
  - relations between synsets: antonymy, hyponymy, meronymy, etc.
- automatically derived from the Princeton WordNet (2004, Minozzi)
- 9,378 lemmas distributed across 8,973 synsets

Manual removal of noise:

- Senses not applicable to Latin, e.g. *voco*, v
  - send a message or attempt to reach someone by radio, phone, etc.
- Concrete vs. abstract meaning, e.g. *licentia*, n
  - a legal document giving official permission to do something
  - the act of giving a formal (usually written) authorization
  - leave granted to a sailor or naval officer

Latin WordNet

Lexical Entries, Senses and Concepts

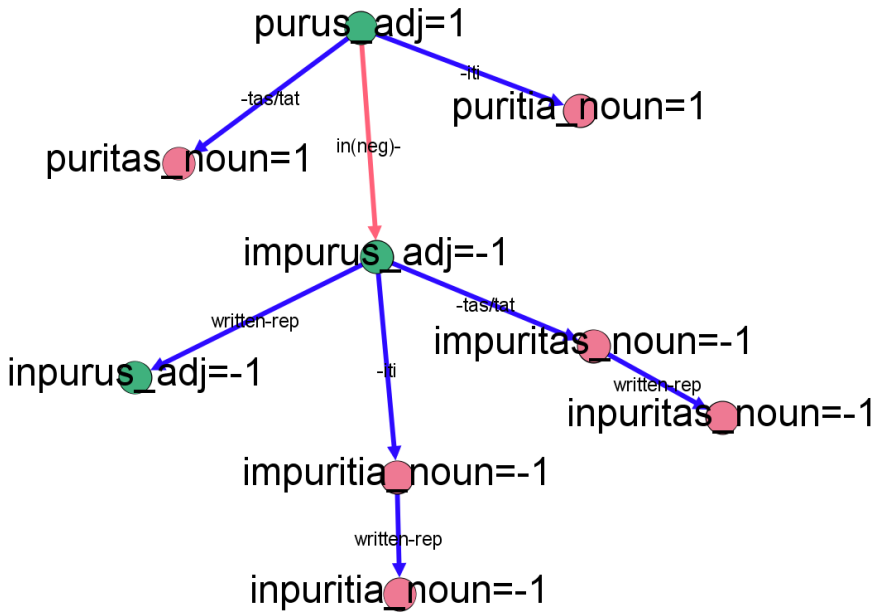


Polarity

Source: *Latin Affectus* (CIRCSE Research Centre)

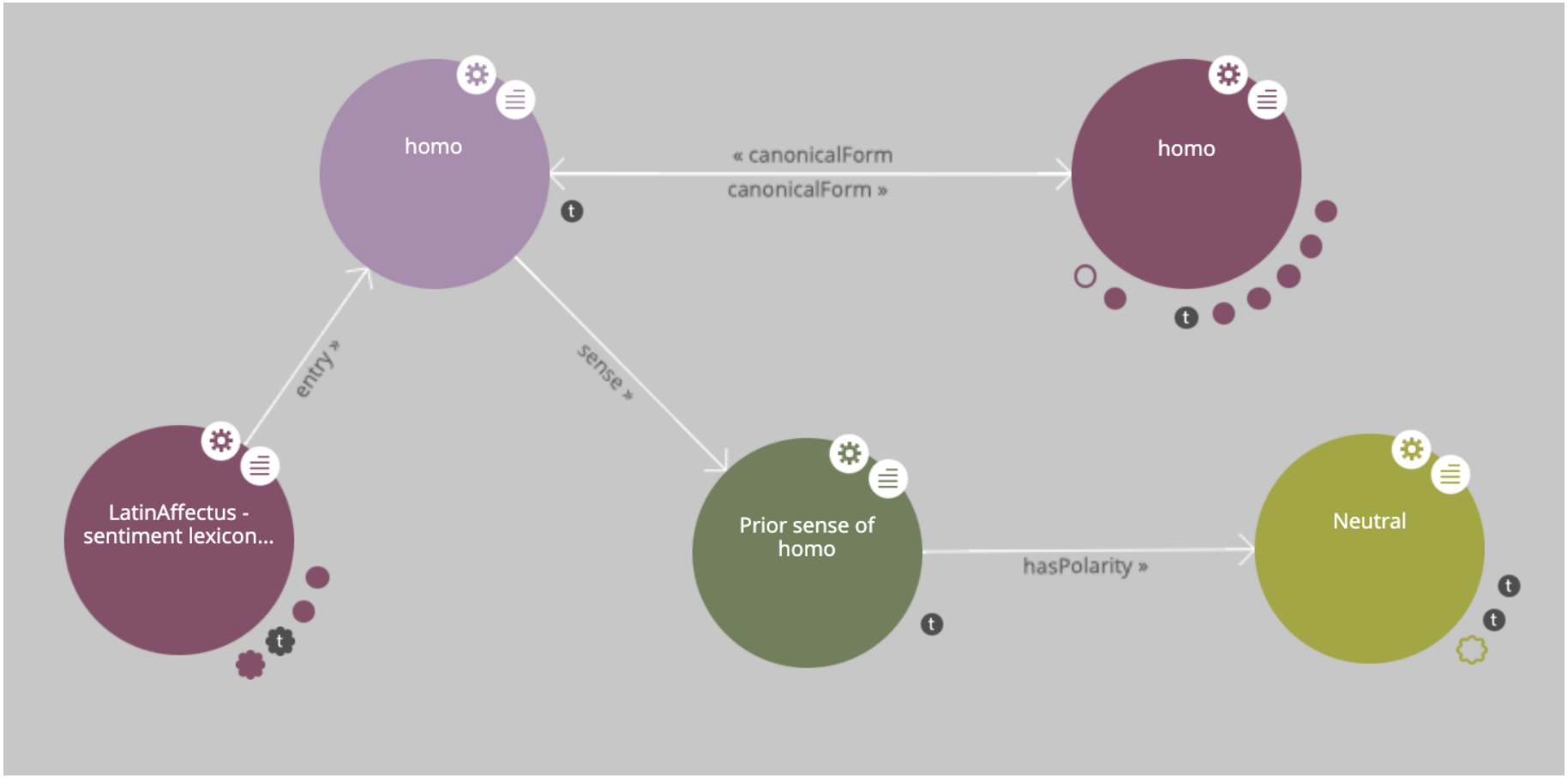
List of 2,437 nouns and adjectives associated to their **out-of-context sentiment score**: from -1 (very negative, e.g. *abominatio*) to +1 (very positive, e.g. *amor*)

- **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

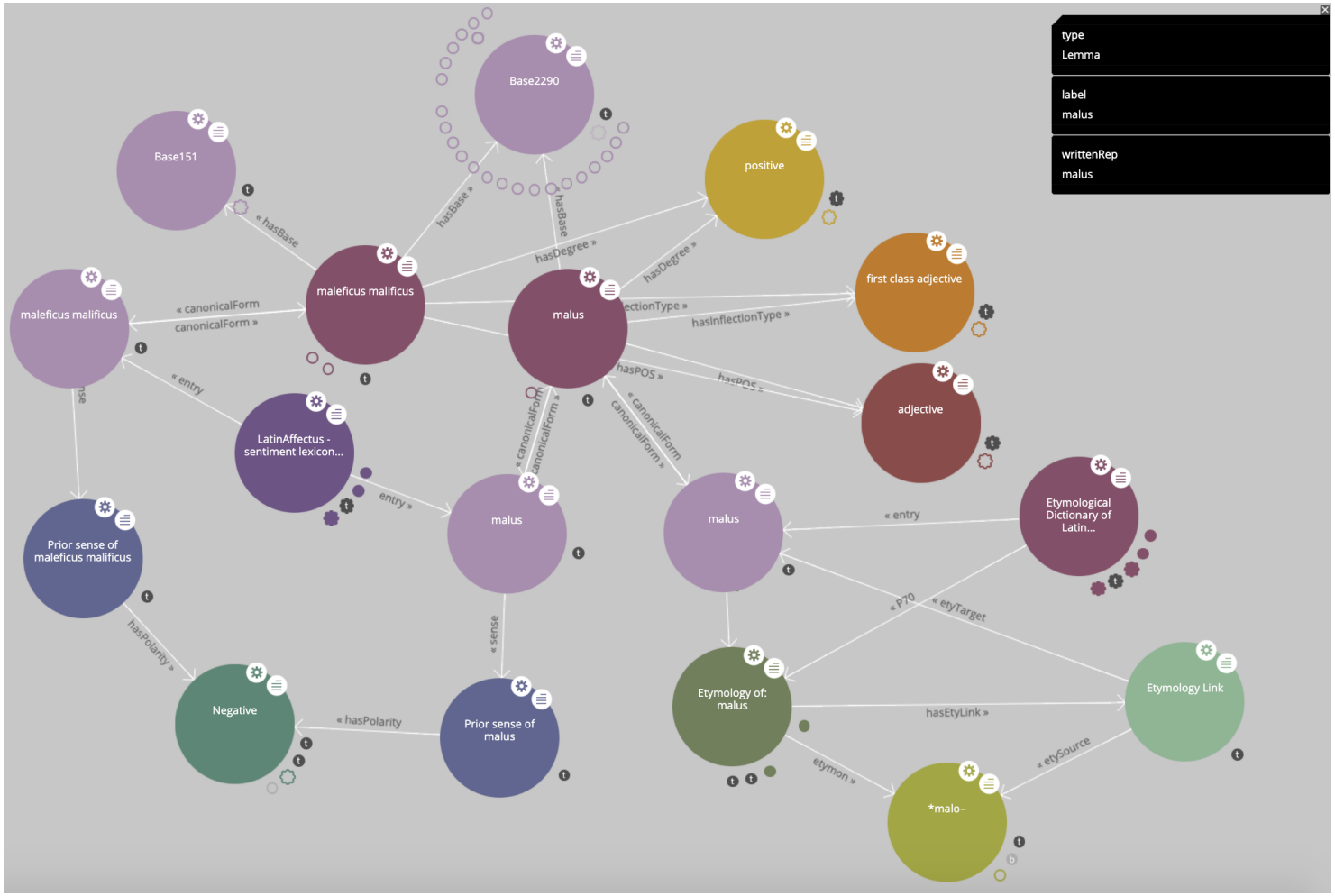


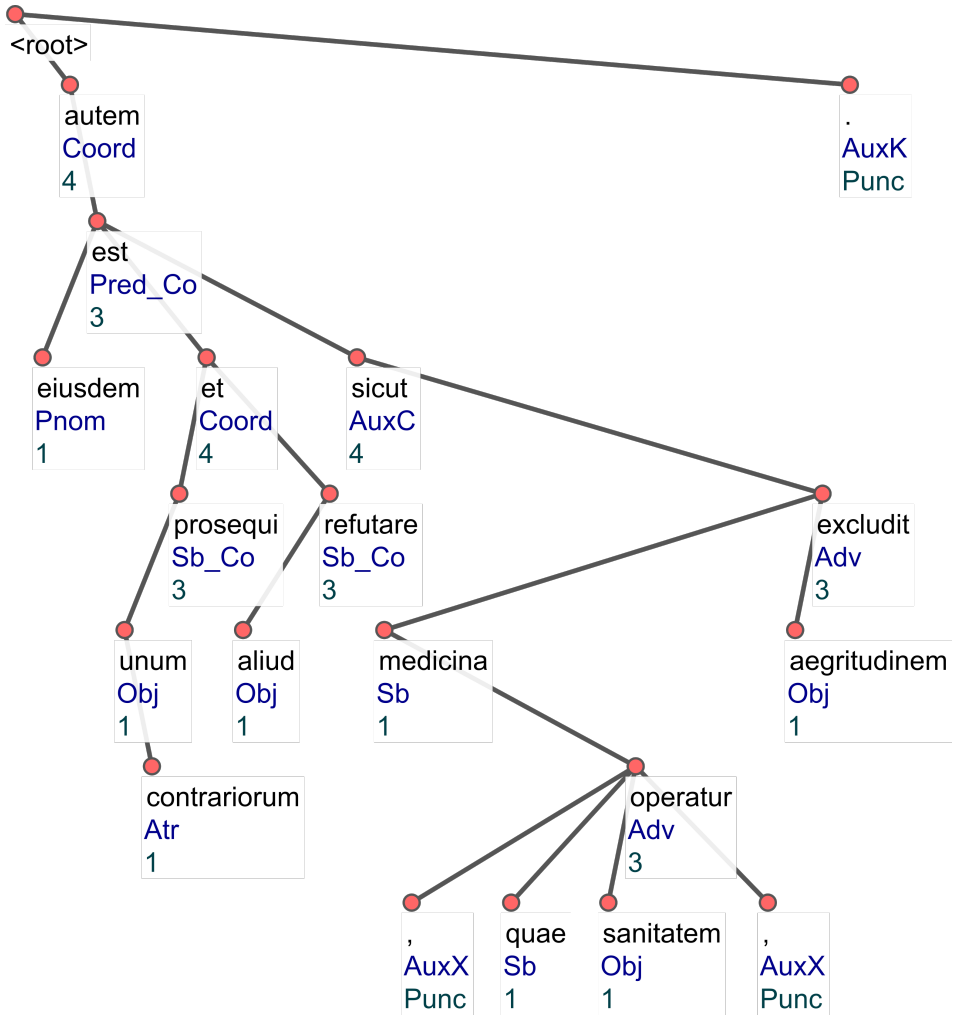
Table of Contents

(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

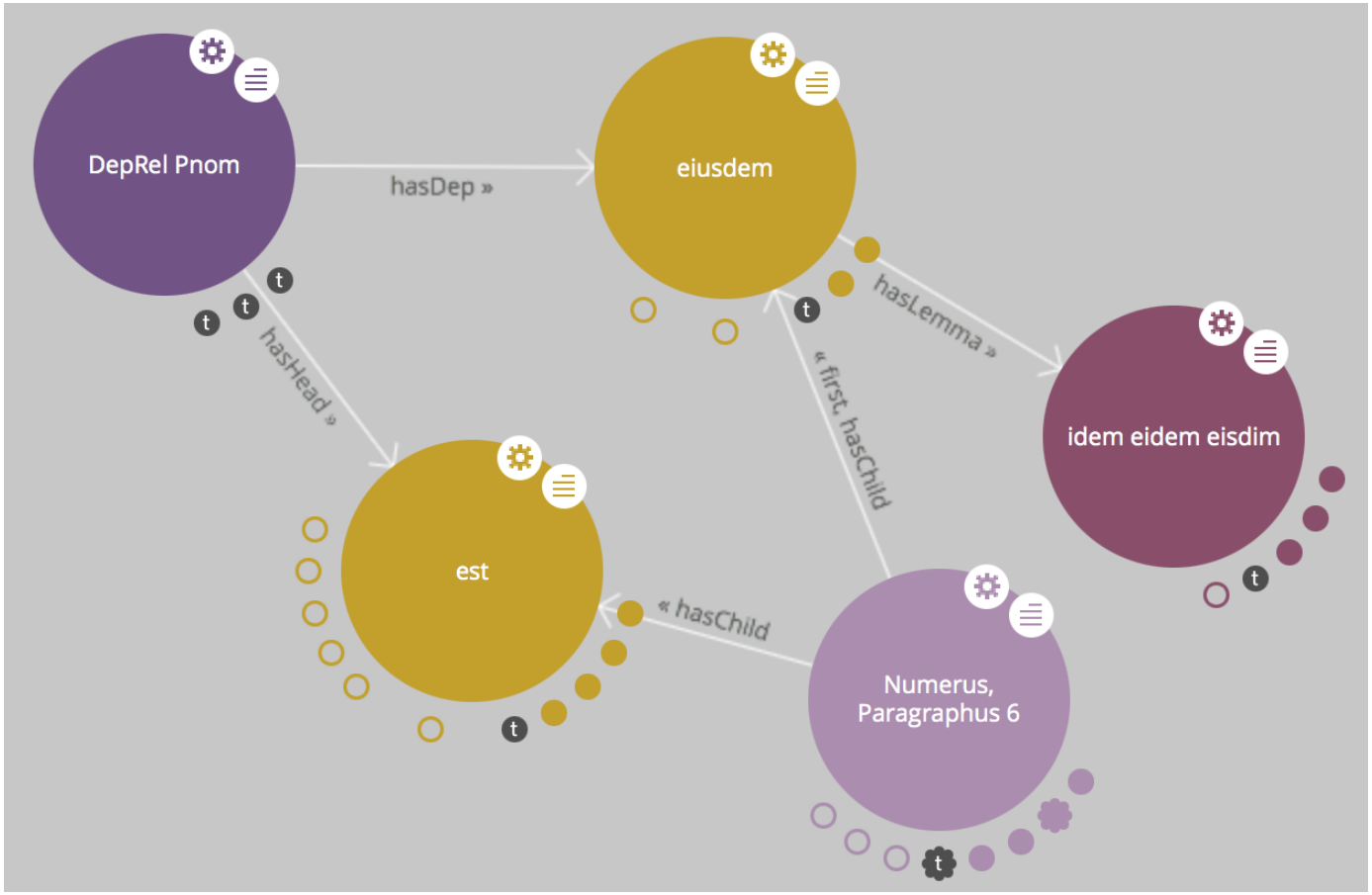


Table of Contents

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 Corpus
- **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
  - [ ] Latin Vallex 2.0: Valency Lexicon
  - [ ] Lewis & Short Dictionary
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 10 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

Table of Contents

LiLa: Linked Pasts 6 Activity

Structure

- **WHAT**: to show the workflow we employ to connect a linguistic resource for Latin to the LiLa Knowledge Base, and to demonstrate the way in which LiLa can be queried.

LiLa: Linked Pasts 6 Activity

Structure

- **WHAT**: to show the workflow we employ to connect a linguistic resource for Latin to the LiLa Knowledge Base, and to demonstrate the way in which LiLa can be queried.
- **HOW**: to teach participants how to perform automatic lemmatisation and RDF-isation (i.e. format conversion) in order to link a Latin text to LiLa.

- **WHAT**: to show the workflow we employ to connect a linguistic resource for Latin to the LiLa Knowledge Base, and to demonstrate the way in which LiLa can be queried.
- **HOW**: to teach participants how to perform automatic lemmatisation and RDF-isation (i.e. format conversion) in order to link a Latin text to LiLa.
- **WHO-1**: anyone who wishes to publish Latin texts on the web.

LiLa: Linked Pasts 6 Activity

Structure

- **WHAT**: to show the workflow we employ to connect a linguistic resource for Latin to the LiLa Knowledge Base, and to demonstrate the way in which LiLa can be queried.
- **HOW**: to teach participants how to perform automatic lemmatisation and RDF-isation (i.e. format conversion) in order to link a Latin text to LiLa.
- **WHO-1**: anyone who wishes to publish Latin texts on the web.
- **WHO-2**: anyone interested in the different aspects involved in the construction of a Linguistic Linked Open Data knowledge base.

LiLa: Linked Pasts 6 Activity

Participants

- **Participants: 41** (35 affiliated, 6 independent)
- **Countries: 12** (Belgium, Brazil, France, Greece, Hungary, Italy, Latvia, Netherlands, Portugal, Switzerland, UK, USA)
- **Expertise:**

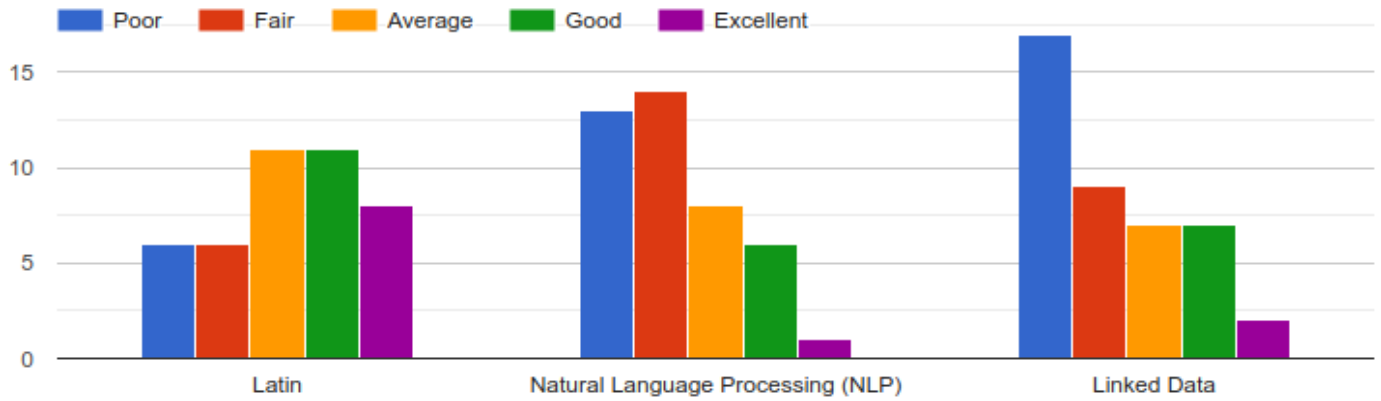


Table of Contents

TextLinker

Welcome screen

Figure: LiLa's Text Linker

TextLinker
Processed output

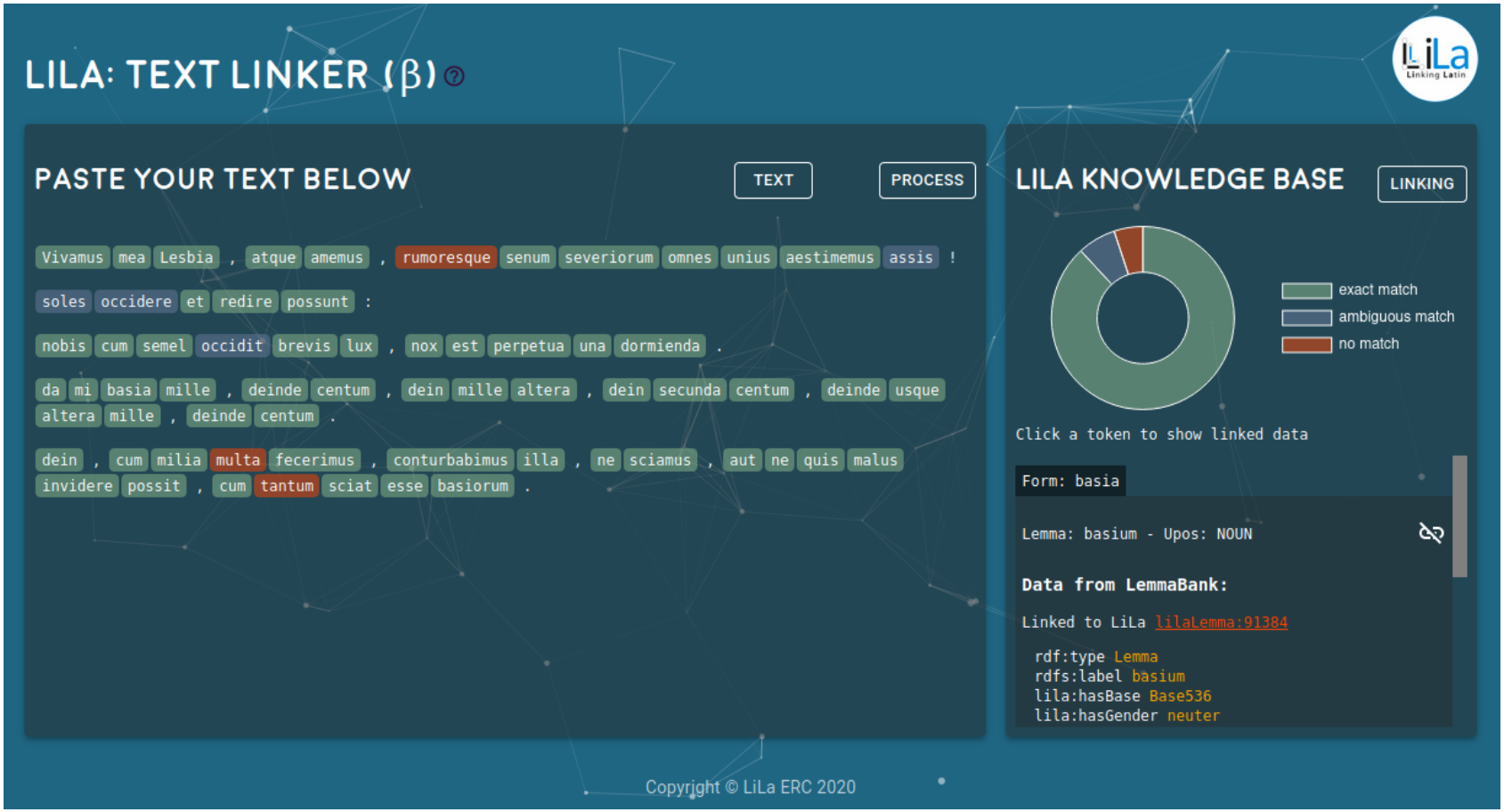


Figure: Text processed against the LiLa Knowledge Base

Table of Contents

Working texts

Texts and passages chosen (out of 8 suggestions):

- Horace (65-8 BC), *Carmina 1.7* and *2.7*
- Pliny the Elder (23/24-79 AD), *Naturalis Historia 2.1-31*
- Giovanni Pico della Mirandola (1463-1494), *Conclusiones secundum Thomam*

Table of Contents

Programme & communication tools

Programme (CET)

- 10:00-12:30: Slide presentation, quiz, hands-on work
- 12:30-14:00: break
- 14:00-17:00: Hands-on work

breaks agreed together as the activity progresses.

Tools

- Microsoft Chat/Raise hand: for quick help
- Google Drive: for collaborative note-taking

The activity will be RECORDED!

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.