Issues in Building the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti, Francesco Mambrini

UNIVERSITÀ CATTOLICA del Sacro Cuore

IRCSE

Conference on
LLOD approaches for language data research and management
(LLODREAM2022)
21-22 September 2022
Mykolas Romeris University, Vilnius, Lithuania

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

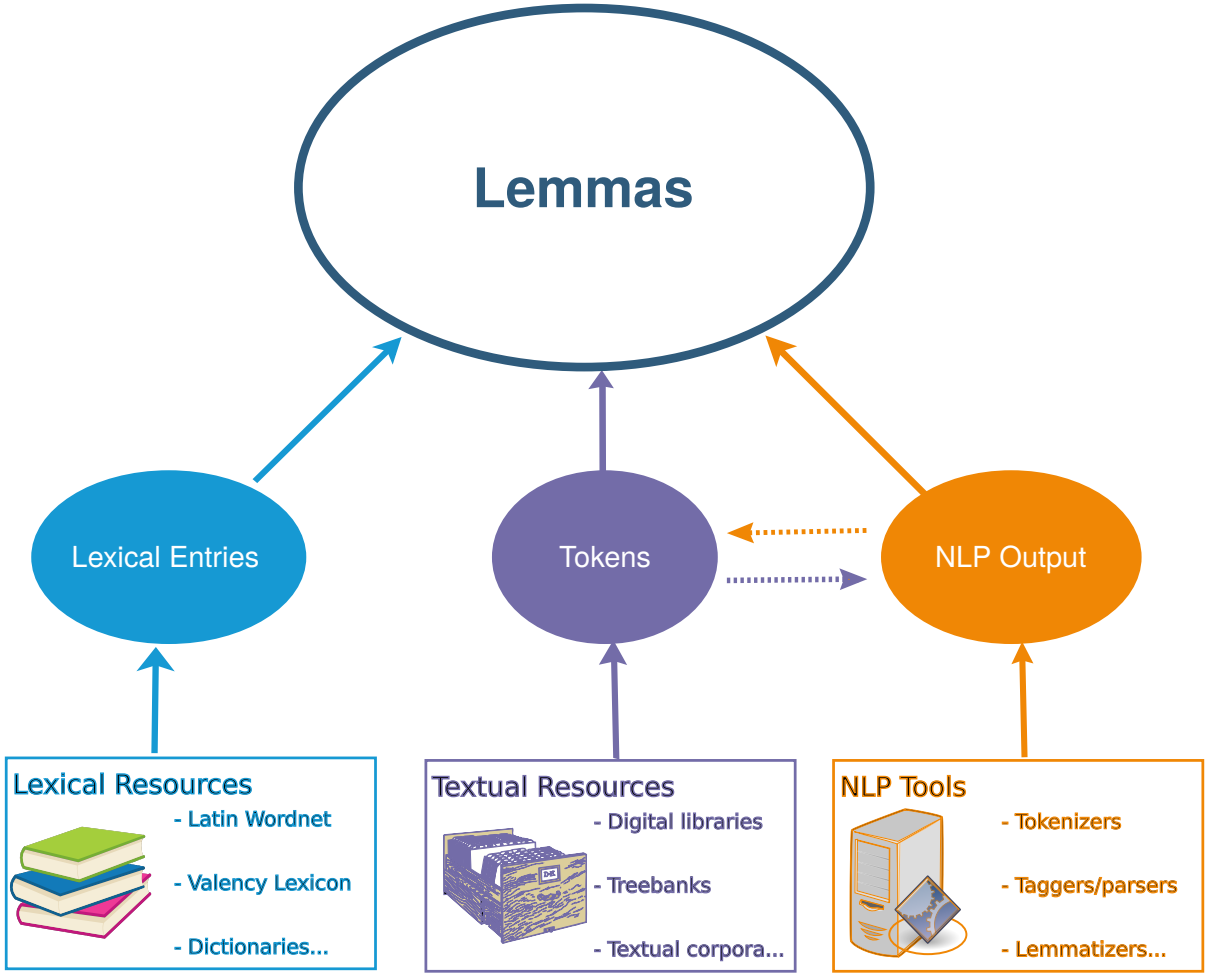
Table of Contents

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)

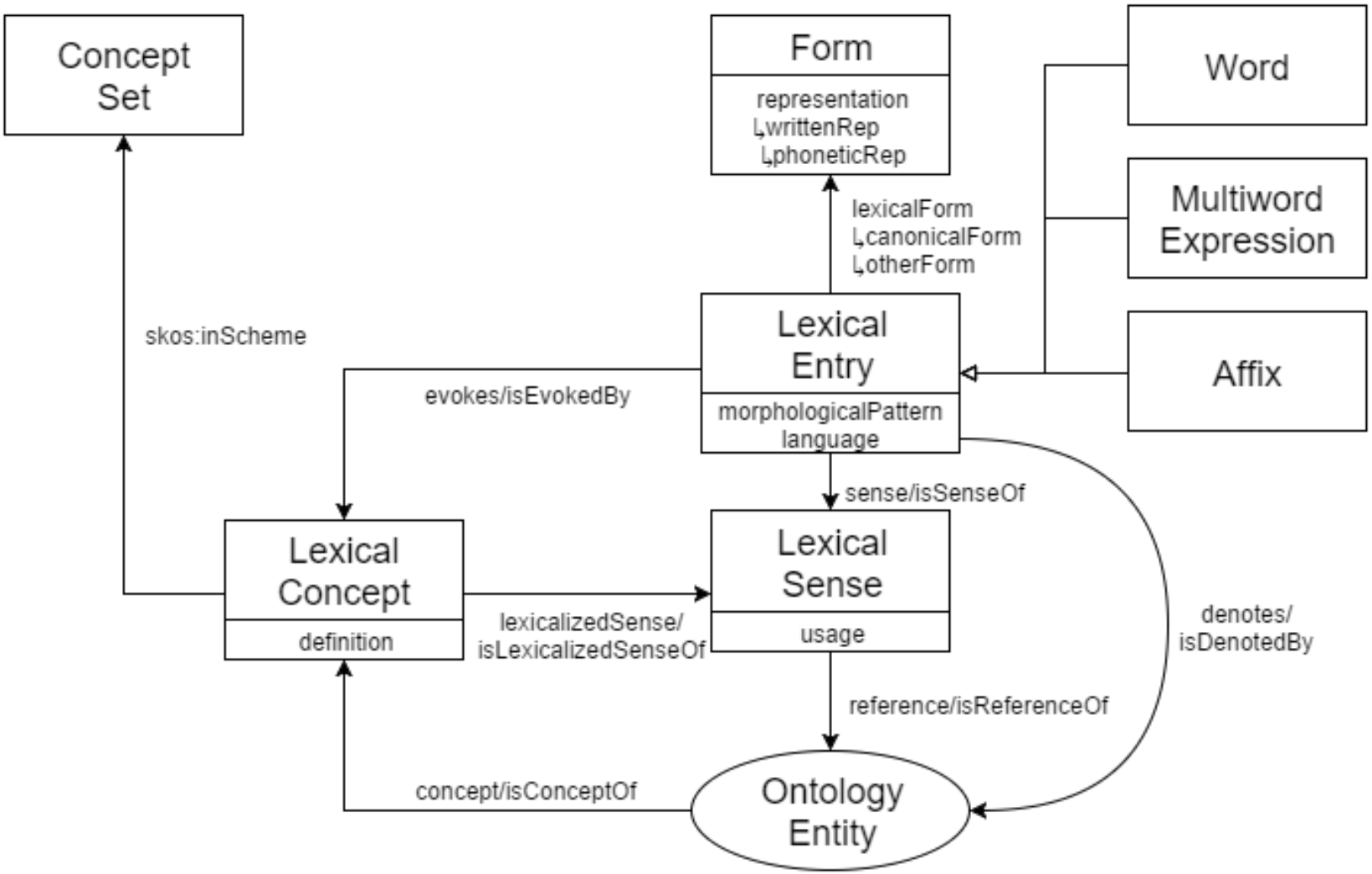
LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

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



LiLa: Overview

Resources connected so far

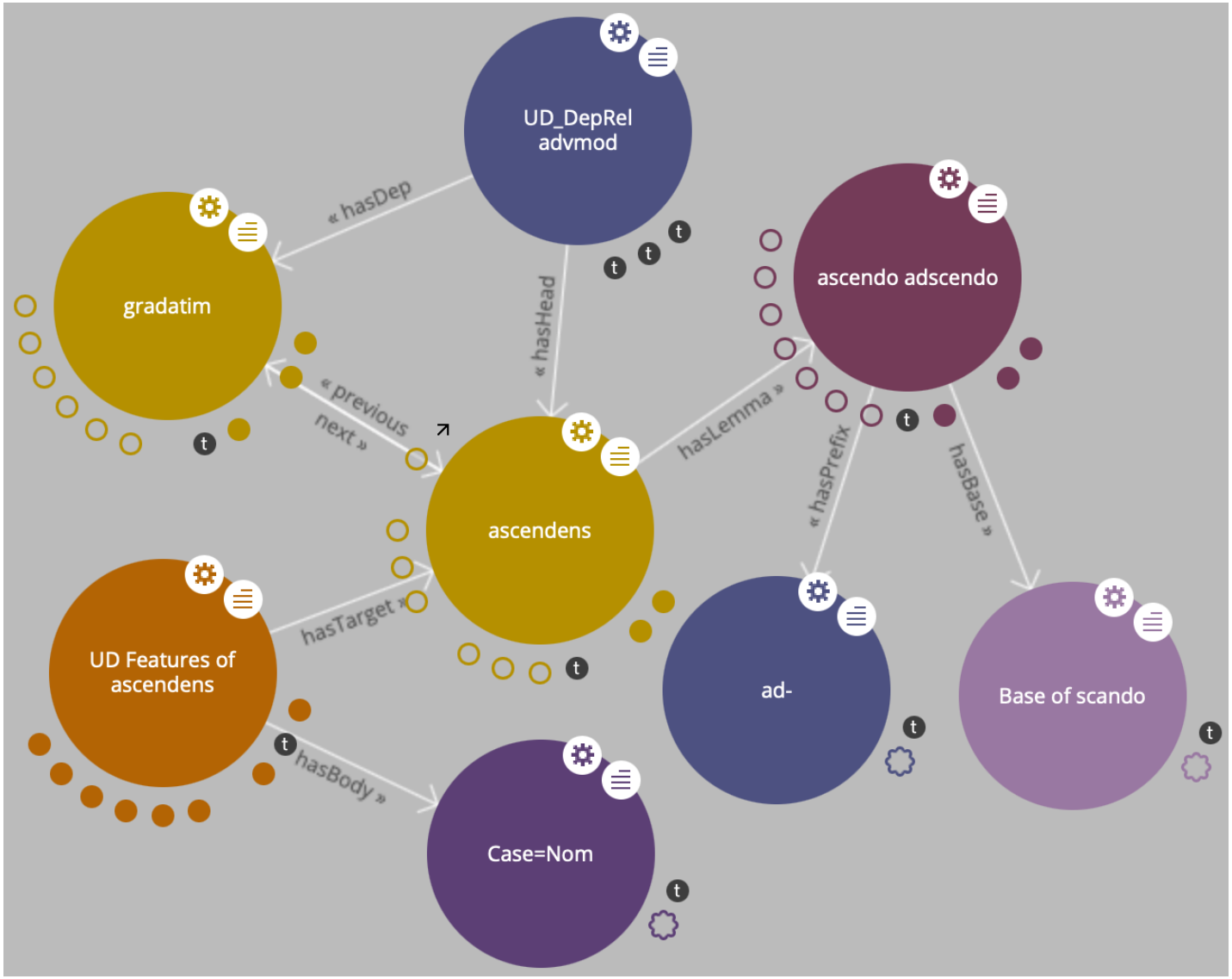
- **Textual Resources**
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 400,000 nodes
  - ☑ UDante Treebank: ca. 46,000 tokens
  - ☑ *Querolus sive Aulularia*: ca. 17,000 tokens
  - ☑ *Liber Abbaci* (ch. VIII) by Leonardo Fibonacci: ca. 30,000 tokens
  - ☑ LASLA Corpus: ca. 1.7 million tokens
  - ☑ Computational Historical Semantics: ca. 1 million tokens
  - ☑ *Confessiones* by Augustinus: ca. 92,000 tokens
- **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

TOTAL: approximately 47 million triples

Table of Contents

Modeling a Textual Resource in LiLa

http://lila-erc.eu/data/corpora/ITTB/id/token/005.SCG*LB4.CP--++1.N.1-7-2.11-1W24



Issues in Publishing Textual Resources in LiLa

Issues in Publishing Textual Resources in LiLa

- Quality of **automatic lemmatization and PoS tagging**

Issues in Publishing Textual Resources in LiLa

- Quality of **automatic lemmatization and PoS tagging**
- **Ambiguity**: to solve or not to solve? See *occido* and *populus*

Issues in Publishing Textual Resources in LiLa

- Quality of **automatic lemmatization and PoS tagging**
- **Ambiguity**: to solve or not to solve? See *occido* and *populus*
- **Missing links**: automatic enhancement of the Lemma Bank? New lemma? New lemma-PoS couple? New written representation?

Issues in Publishing Textual Resources in LiLa

- Quality of **automatic lemmatization and PoS tagging**
- **Ambiguity**: to solve or not to solve? See *occido* and *populus*
- **Missing links**: automatic enhancement of the Lemma Bank? New lemma? New lemma-PoS couple? New written representation?
- **Different lemmatization/PoS criteria** in source corpora: solved by:
  - harmonization (graphical variants + hypolemmas & lemma variants → see next slide)
  - relaxed constraints on PoS-match (PROPN, NOUN, ADJ)

Issues in Publishing Textual Resources in LiLa

https://lila-erc.eu/lodview/data/corpora/Lasla/id/corpus/CaesarBellum20Civile/Caesar_BellumCivile_CaesBC1.BPN_t_0006064

Praeruptus locus erat utraque ex parte derectus ac tantum in latitudinem patebat, ut tres **instructae** cohortes eum locum explerent. (Caesar, *Bellum civile*, 1, 45, 4)

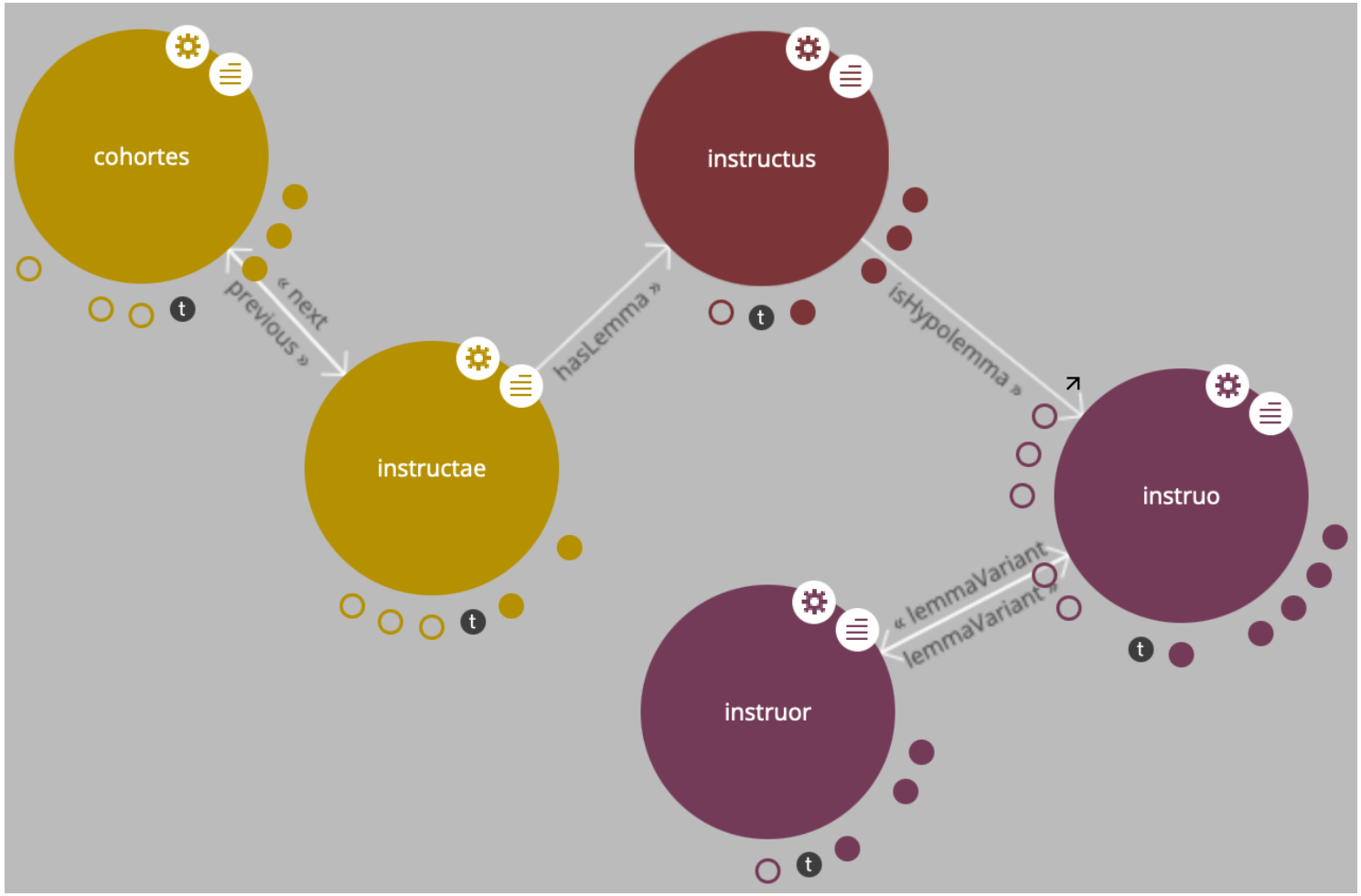
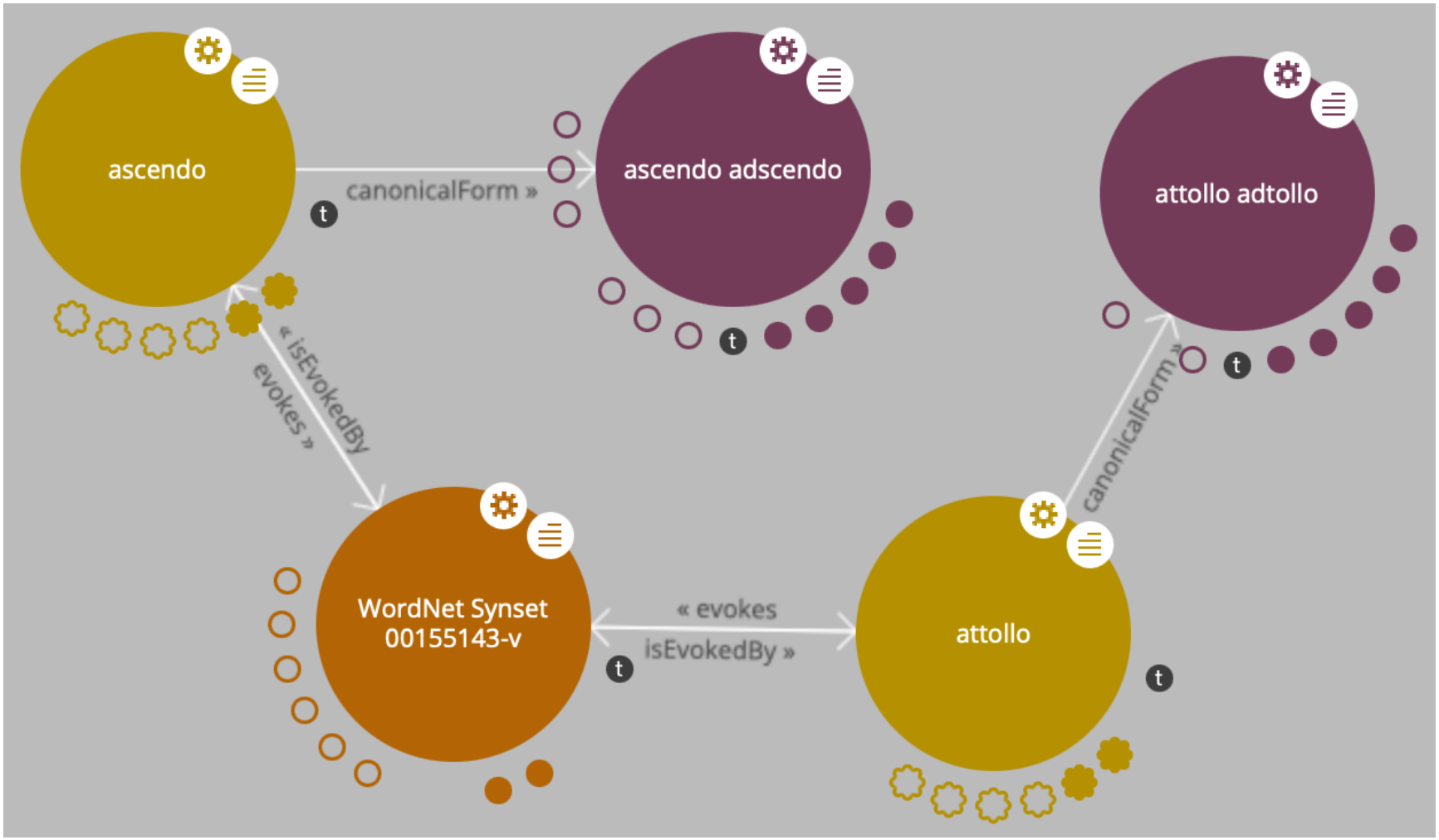


Table of Contents

Modeling a Lexical Resource in LiLa

https://lila-erc.eu/lodview/data/id/lemma/90185



Issues in Publishing Lexical Resources in LiLa

Issues in Publishing Lexical Resources in LiLa

- Modeling and representation of **complex lexical entries**: see the TLL

Issues in Publishing Lexical Resources in LiLa

- Modeling and representation of **complex lexical entries**: see the TLL
- Disambiguation of the linking of **homographs** in the Lemma Bank

Issues in Publishing Lexical Resources in LiLa

- Modeling and representation of **complex lexical entries**: see the TLL
- Disambiguation of the linking of **homographs** in the Lemma Bank
- Lexical resources with **"not canonical" naming of entries** (e.g., infinitive instead of first singular indicative present for verbs): enhancing the Lemma Bank?

Table of Contents

- Lemma Bank Query Interface:
  `https://lila-erc.eu/query/`

- Lemma Bank Query Interface:
  `https://lila-erc.eu/query/`
- SPARQL Access Point:
  `https://lila-erc.eu/sparql/`

Services and Tools
Linking and Querying Resources

- Lemma Bank Query Interface:
  `https://lila-erc.eu/query/`
- SPARQL Access Point:
  `https://lila-erc.eu/sparql/`
- TextLinker:
  `http://lila-erc.eu:8080/LiLaTextLinker/`

- Lemma Bank Query Interface:
  `https://lila-erc.eu/query/`
- SPARQL Access Point:
  `https://lila-erc.eu/sparql/`
- TextLinker:
  `http://lila-erc.eu:8080/LiLaTextLinker/`
- LiLa Query Interface:
  `http://lila-erc.eu:8080/lila-lisp/`

Conclusion

...still a lot of things to do

Conclusion
...still a lot of things to do

- To get rid of PoS tags in the Lemma Bank?

Conclusion
...still a lot of things to do

- To get rid of PoS tags in the Lemma Bank?
- Representation issues: limitations in representing complex lexical entries (like those of TLL); limitations in representing critical apparatus

Conclusion
...still a lot of things to do

- To get rid of PoS tags in the Lemma Bank?
- Representation issues: limitations in representing complex lexical entries (like those of TLL); limitations in representing critical apparatus
- Are we really *linking* Latin? So far we've *published* a lot of Latin LOD

Conclusion
...still a lot of things to do

- To get rid of PoS tags in the Lemma Bank?
- Representation issues: limitations in representing complex lexical entries (like those of TLL); limitations in representing critical apparatus
- Are we really *linking* Latin? So far we've *published* a lot of Latin LOD
- Help more projects publishing Latin data as LOD (and link them to us)

Thank you!

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.