UDante.
Dante's Latin Works
in the LiLa Knowledge Base

Francesco Mambrini, Rachele Sprugnoli

Joint work with: Flavio M. Cecchini, Giulia Pedonese, Marco Passarotti

Annotators: Daniela Corbetta, Federica Favero, Federica Gamba, Martina de Laurentiis, Andrea Peverelli ed Elena Vagnoni

Digital Dante Days
Venice, 16 November 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

The LiLa Knowledge Base

UDante

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

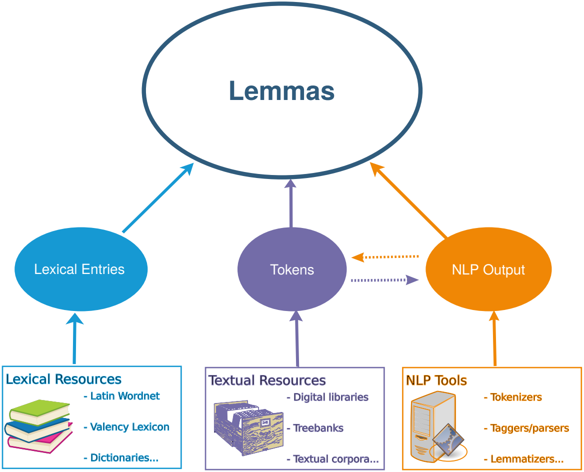
We have built and collected (for Latin and other languages):

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



Lemmas

Lexical Entries

Tokens

NLP Output

Lexical Resources
- Latin Wordnet
- Valency Lexicon
- Dictionaries...

Textual Resources
- Digital libraries
- Treebanks
- Textual corpora...

NLP Tools
- Tokenizers
- Taggers/parsers
- Lemmatizers...

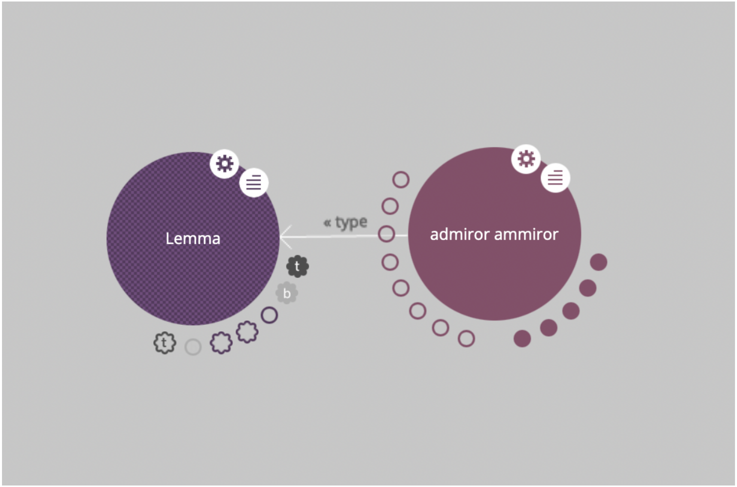
The Linked Data Principles

...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

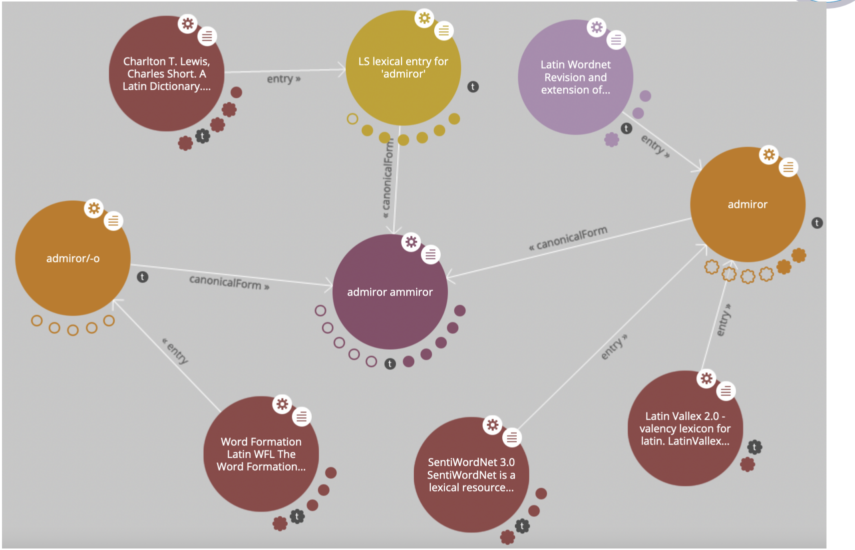
Meet a Lemma!

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



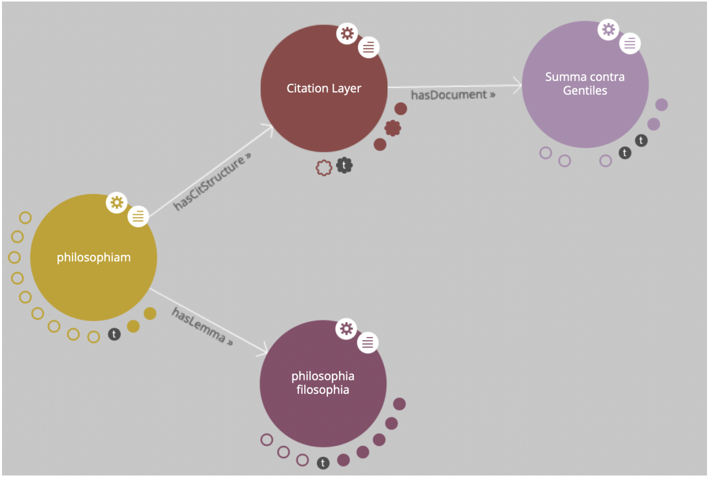
Lemma *admiror* ‘to admire, to respect’

. . . and its Lexical Entries!


Charlton T. Lewis, Charles Short. A Latin Dictionary....
entry »
LS lexical entry for 'admiror'
Latin Wordnet Revision and extension of...
« canonicalForm
entry »
admiror
admiror/-o
« canonicalForm
canonicalForm »
admiror ammiror
« entry
entry »
entry »
Word Formation Latin WFL The Word Formation...
SentiWordNet 3.0 SentiWordNet is a lexical resource...
Latin Vallex 2.0 - valency lexicon for latin. LatinVallex...

. . . and its Tokens!

http://lila-erc.eu/data/corpora/ITTB/id/token/005.SCG*LB1.CP--++1.N.-
1-1.2-4W4



Token *philosophiam* (Thom. Aq., *Summa Contra Gentiles* 1.1.5)

LiLa: Overview

Resources connected and upcoming connections

- **Textual Resources**
  - ☑ Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - ☑ **UDante: ca. 55,000 tokens**
  - ☑ *Liber Abbaci, Chapter VIII*: ca. 30,000 tokens
  - ☑ *Querolus sive Aulularia*: ca. 17,000 tokens
  - □ PROIEL and LLCT treebanks
  - □ Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexical Resources**
  - ☑ Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
  - ☑ Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
  - ☑ LatinAffectus: ca. 4,000 entries
  - ☑ Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
  - ☑ Latin WordNet: ca. 1,000 manually checked entries
  - ☑ Latin Vallex 2.0: Valency Lexicon
  - ☑ Lewis & Short Dictionary
  - □ Lemma Embeddings
- **NLP tools**
  - ☑ LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 15 million triples**

UDante: What is it?

- Corpus of Dante’s Latin works annotated following the ***Universal Dependencies*** (UD) formalism:
  - *Monarchia*, *De vulgari eloquentia*, *Eclogues*, *Letters*, *Questio de aqua et terra*
  - tokenization, sentence splitting, lemmatization, PoS tagging, morphological features, syntactic annotation
  - 55,666 tokens
  - texts taken from *DanteSearch*

UDante: What we have done?

Two activities:

1. linguistic annotation: semiautomatic conversion from DanteSearch + manual syntactic annotation from scratch

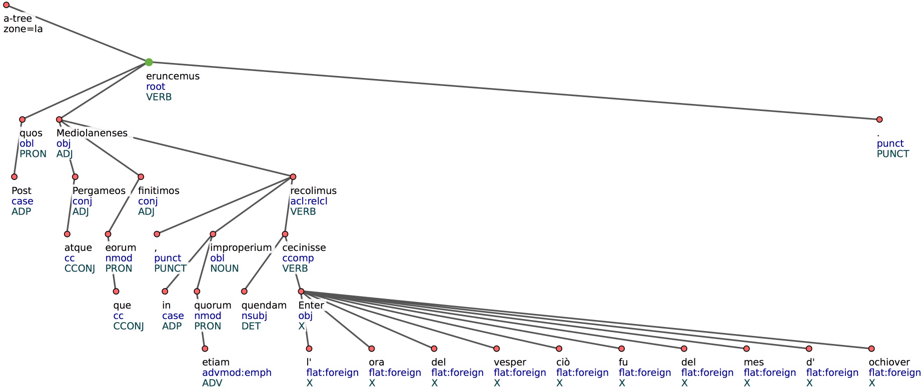
Published in **UD release** v2.8
`https://github.com/UniversalDependencies/UD_Latin-UDante`

2. linking of UDante to the LiLa Knowledge Base

Published in **Linked Open Data**
`https://lila-erc.eu/data/corpora/DanteSearch/id/corpus`

Syntactic Annotation

DVE I xi 5: *Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l'ora del vesper, ciò fu del mes d'ochiover.*



Data Format: CoNLL-U

```
# sent_id = DVE-124
# text = Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l' ora del vesper ciò
fu del mes d' ochiover.
# citation_hierarchy = Liber_Primus,xi,Paragraphus_5
1 Post          post          ADP   e       AdpType=Prep                                                                              2  case         _ _
2 quos          qui           PRON  prepma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                           9  obl          _ _
3 Mediolanenses mediolanensis ADJ   Smp3a   Case=Acc|Gender=Masc|InflClass=IndEurI|NameType=Nat|Number=Plur                           9  obj           _ _
4 atque         atque         CCONJ co      Emphatic=Yes                                                                              5  cc            _ _
5 Pergameos     pergameus     ADJ   Smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|NameType=Nat|Number=Plur                           3  conj         _ _
6-7 eorumque    _             _     _       _                                                                                         _  _            _ _
6 eorum         is            PRON  ddepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|Person=3|PronType=Prs                  8  nmod         _ _
7 que           que           CCONJ co9     Clitic=Yes                                                                                6  cc           _ _
8 finitimos     finitimus     ADJ   smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|Number=Plur                                        3  conj         _ _
9 eruncemus     erunco        VERB  va1cpp1 Aspect=Imp|InflClass=LatA|Mood=Sub|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 0  root         _ SpaceAfter=No
10 ,            ,             PUNCT Pu      _                                                                                         17 punct        _ _
11 in           in            ADP   e       AdpType=Prep                                                                              14 case         _ _
12 quorum       qui           PRON  prepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                           14 nmod         _ _
13 etiam        etiam         ADV   co      Compound=Yes                                                                              12 advmod:emph  _ _
14 improperium  improperium   NOUN  sns2a   Case=Acc|Gender=Neut|InflClass=IndEurO|Number=Sing                                        17 obl          _ _
15 quendam      quidam        DET   dinsma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Sing|PronType=Ind                           16 nsubj        _ _
16 cecinisse    cano          VERB  va3fr   Aspect=Perf|InflClass=LatX|InflClass[noun]=Ind|Tense=Past|VerbForm=Inf|Voice=Act          17 ccomp        _ _
17 recolimus    recolo        VERB  va3ipp1 Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 3  acl:relcl    _ _
18 Enter        enter         X     zi      Foreign=Yes                                                                               16 obj          _ _
19 l'           l             X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
20 ora          ora           X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
21 del          del           X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
22 vesper       uesper        X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
23 ciò          cio           X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
24 fu           fu            X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
25 del          del           X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
26 mes          mes           X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
27 d'           d             X     zi      Foreign=Yes                                                                               18 flat:foreign _ _
28 ochiover     ochiouer      X     zi      Foreign=Yes                                                                               18 flat:foreign _ SpaceAfter=No
29 .            .             PUNCT Pu      _                                                                                         9  punct        _ _
```

Linking

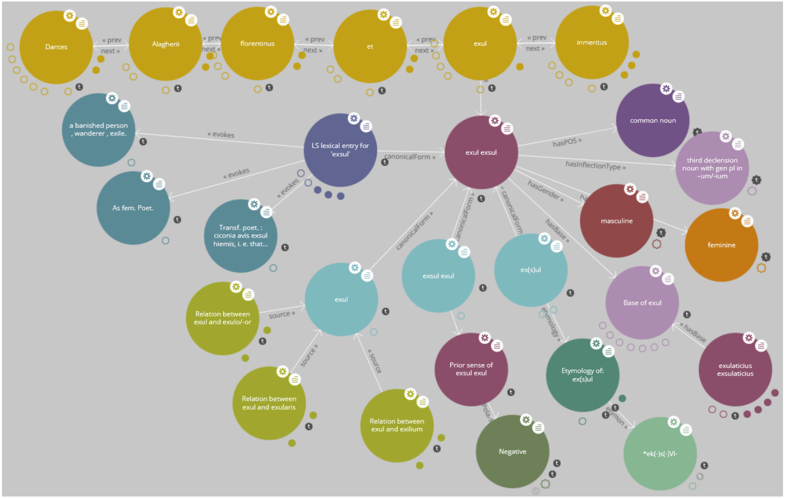
Semi-automatic linking of the lemmas in Dante’s works to their corresponding entries in the LiLa Lemma Bank:

1. automatic matching (~80%)
2. manual linking of:
    - ambiguous lemmas: e.g. *frons*
    - new lemmas: e.g.
        - proper names: *Caualcantis*
        - feminine abstract common nouns: *depauperatio*
        - neuter common nouns: *parysillabum*
        - ethnonyms: *lombardus*
        - neologisms: *turpiloquium*

Linking

Example

- Try it yourself: `https://tinyurl.com/exul-epiVI-1`



Querying Interlinked Resources

- Which are the lemmas used by Dante only in the *De vulgari eloquentia* and not in his other Latin works?
  - We query both UDante and the Lemma Bank with a SPARQL query

LEMMA	FREQ
cantio	65
stantia	30
hendecasyllabum	18
quare	18
syllaba	18
habitudo	17
poetor	16

LEMMA	FREQ
eptasillabum	15
constructio	14
loquela	14
latius	12
desinentia	11
profero	11
curialis	10

- Try it yourself: `https://github.com/CIRCSE/SPARQL-queries/blob/main/distinctivelammas-DVE.rq`

LiLa: Accessing and Querying

Query Interface, Triplestore, Linker, Corpora and Lexica

Query Interface, Triplestore

- Query interface
- Triplestore

Linguistic Resources. Corpora

- Index Thomisticus Treebank
- UDante
- Querolus sive Aulularia
- Liber Abbaci

Linguistic Resources. Lexica

- Word Formation Latin
- Etymological Dictionary of Latin & the Other Italic Languages
- LatinAffectus
- Index Graecorum Vocabulorum in Linguam Latinam
- Latin WordNet
- Latin Vallex 2.0

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.