Sentiment Analysis for Latin: a Journey from Seneca to Thomas Aquinas

Rachele Sprugnoli
rachele.sprugnoli@unicatt.it
Joint works with: D. Corbetta, F. Mambrini, G. Moretti, M. Passarotti, A. Peverelli

Sentiment Analysis in Literary Studies
February 18, 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

1. The LiLa project
2. Creation of sentiment lexicons
   - Methods
   - Evaluation
   - Extension
   - Use case
3. Modelling and linking
   - Method
   - Use case
4. Conclusions

PART I

THE LiLa PROJECT

The *LiLa: Linking Latin* project

Research Question - State of affairs

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

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

REFERENCE: Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P. & Sprugnoli, R. (2020). *Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin*. Studi e Saggi Linguistici, 58(1), 177-212.

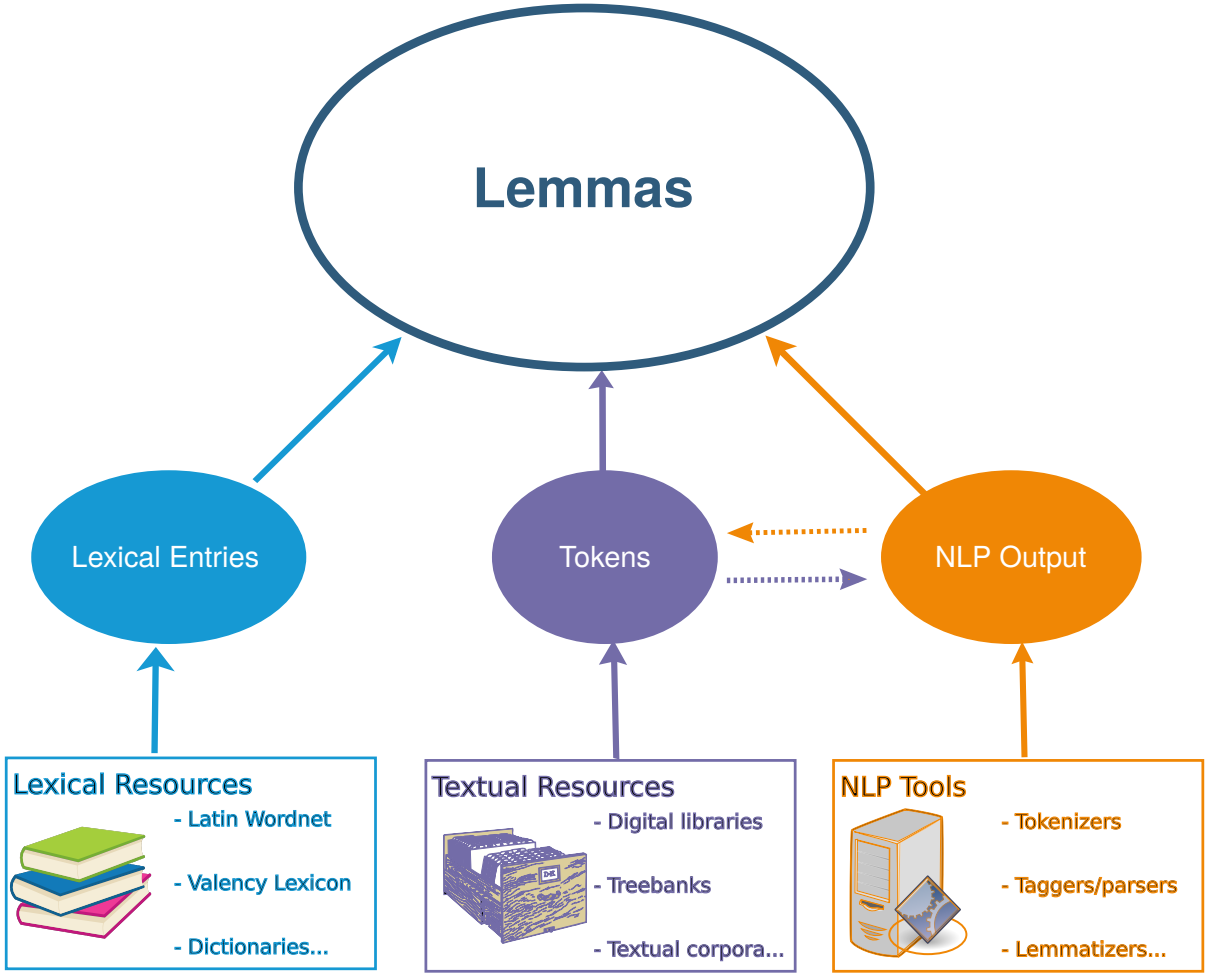
ERC Consolidator Grant 2018-2023

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
KEYWORD: **Interoperability**

LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa: Resources

- **Corpora**
  - Index Thomisticus Treebank (*Summa contra Gentiles*)
  - Dante's Works
  - Querolus sive Aulularia
- **NLP tools**
  - LEMLAT
  - Lemma Embeddings
- **Lexicons**
  - Word Formation Latin
  - Etymological dictionary of Latin & the other Italic Languages
  - Index Graecorum Vocabulorum in Linguam Latinam
  - Latin WordNet
  - Latin Vallex 2.0
  - Polarity Lexicons

LiLa: Resources

- **Corpora**
  - Index Thomisticus Treebank (*Summa contra Gentiles*)
  - Dante's Works
  - Querolus sive Aulularia
- **NLP tools**
  - LEMLAT
  - Lemma Embeddings
- **Lexicons**
  - Word Formation Latin
  - Etymological dictionary of Latin & the other Italic Languages
  - Index Graecorum Vocabulorum in Linguam Latinam
  - Latin WordNet
  - Latin Vallex 2.0
  - Polarity Lexicons

PART II

CREATION OF SENTIMENT LEXICONS

What

- **Sentiment Analysis** = field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language (Bing Liu, 2012)
- **Sentiment lexicon** = list of words associated to scores expressing their (prior) polarity –> essential resource for both machine learning and lexicon-based sentiment analysis systems

Why

- To test the feasibility of Sentiment Analysis methods to less studies text types: NO social media, NO customers' reviews...what about **ancient languages**?
- To **fill a void** in the literature:
  1. NRC lexicons (Mohammad, 2018)
  2. MultilingualSentiment lexicons (Chen & Skiena, 2014)
  3. API of Latin WordNet (`https://latinwordnet.exeter.ac.uk/api`)

How

- Different **methods** applied:
  1. AUTOMATIC: cross-lingual projection
  2. AUTOMATIC: induction from embeddings
  3. MANUAL: gold standard
  4. AUTOMATIC: silver standard
- Only **nouns** and **adjectives**
- `https://github.com/CIRCSE/Latin_Sentiment_Lexicons`

REFERENCE: Sprugnoli, R., Passarotti, M., Corbetta, D., & Peverelli, A. (2020). *Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin*. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 3078-3086).

Cross-Lingual Projection

- Two steps
  1. **translation** of the entries of an English sentiment lexicon using bilingual dictionaries
     - English lexicon: Cho et al. (2014)
     - Bilingual dictionaries: Words & Cassell's Latin dictionary
  2. **projection** of the original decimal score to the Latin translation

- **Output**: lexicon of 10,516 lemmas with decimal scores

ENGLISH	
LEMMA	SCORE
crime	-0.741

LATIN	
LEMMA	SCORE
noxa	-0.741
nefarium	-0.741
….	-0.741

Induction from Embeddings

- Automatic induction from word embeddings starting from a list of seed terms with known sentiment score
  - **seed terms**: most frequent adjs and nouns from “Opera Latina”
  - **embeddings**: pre-trained with word2vec on “Opera Latina” with a LEMMA_PoS representation, e.g. *rosa_noun*, *amo_verb*
  - **algorithm**: `https://github.com/WladimirSidorenko/SentiLex`

N.B. **Opera Latina**: multi-genre corpus of 1.7 million words manually annotated with lemmas, PoS tags,inflectional features

Induction from Embeddings

- Automatic induction from word embeddings starting from a list of seed terms with known sentiment score
  - **seed terms**: most frequent adjs and nouns from “Opera Latina”
  - **embeddings**: pre-trained with word2vec on “Opera Latina” with a LEMMA_PoS representation, e.g. *rosa_noun*, *amo_verb*
  - **algorithm**: `https://github.com/WladimirSidorenko/SentiLex`

N.B. The **k-NN algorithm** calculates the distance between the vectors of seed terms and the vector of a lemma *l* and then assigns to *l* the sentiment score of the seed that is closest to *l* and appears most often as *l*’s neighbor.

Induction from Embeddings

- Automatic induction from word embeddings starting from a list of seed terms with known sentiment score
  - **seed terms**: most frequent adjs and nouns from "Opera Latina"
  - **embeddings**: pre-trained with word2vec on "Opera Latina" with a LEMMA_PoS representation, e.g. *rosa_noun*, *amo_verb*
  - **algorithm**: `https://github.com/WladimirSidorenko/SentiLex`
- **Output**: lexicon of 1,030 lemmas with three-value scores (negative, neutral, positive)

Lemma	PoS	Sentiment
miseria 'misery'	noun	negative
cruciatus 'torture'	noun	negative
optabilis 'desiderable'	adj	positive
beneuolentia 'good-will'	noun	positive
aerumna 'trouble'	noun	negative

Gold Standard

- **Manually created** by two Latin language and culture experts and one supervisor
- **Six-value classification**: 2 (ambiguous), 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative)
  - Semantic (e.g. *pusillus*) versus diachronic (e.g. *regalis*) ambiguity
- Three phases:
  1. **collaborative** annotation of 20 adjectives and 20 nouns
  2. **independent** annotation of all the other lemmas –> IAA
  3. **reconciliation** of cases of disagreement
- **Output**: lexicon of 1,144 lemmas with five-value scores

Gold Standard
Inter-Annotator Agreement

- **IAA** before reconciliation

	ADJ	NOUN	MACRO-AVG
6 CLASSES	0.39	0.49	0.45
4 CLASSES	0.49	0.59	0.56

Evaluation Using the Gold Standard

Workflow:

- conversion of the decimal scores obtained with the cross-lingual projection method
- conversion of the three-value literal classification obtained with the induction method
- calculation of the **accuracy** wrt the Gold Standard (after the riconciliation)

	PROJECTION		INDUCTION
	5 CLASSES	**3 CLASSES**	**3 CLASSES**
ADJ	44.3%	64.9%	86.7%
NOUN	54.8%	66.8%	62.5%
MICRO-AVG	50.61%	66.1%	**74.4%**

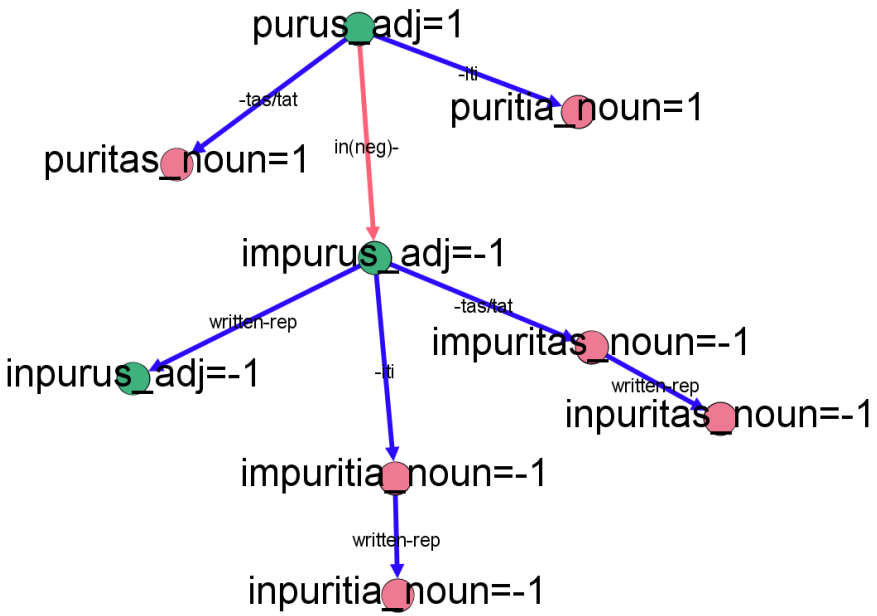
Silver Standard
Gold Standard Extension

New lemmas exploiting 3 resources:

- dictionary of synonyms and antonyms compiled by Skřivan (1890), e.g. *pulcher* ↔ *formosus* ; *beneficium* ↔ *maleficium*
- Word Formation Latin database: 25 prefixal and suffixal relations, e.g. *laetus* ↔ *laetitudo* ; *amaritudo* ↔ *amarus*
- LiLa Lemma Bank: written representations of the same lemma, e.g. *improsper* ↔ *inprosper* ; *tropaeom* ↔ *tropaeum*

Silver Standard
Gold Standard Extension

- Original polarity scores are **propagated** or **reversed** onto the newly derived lemmas

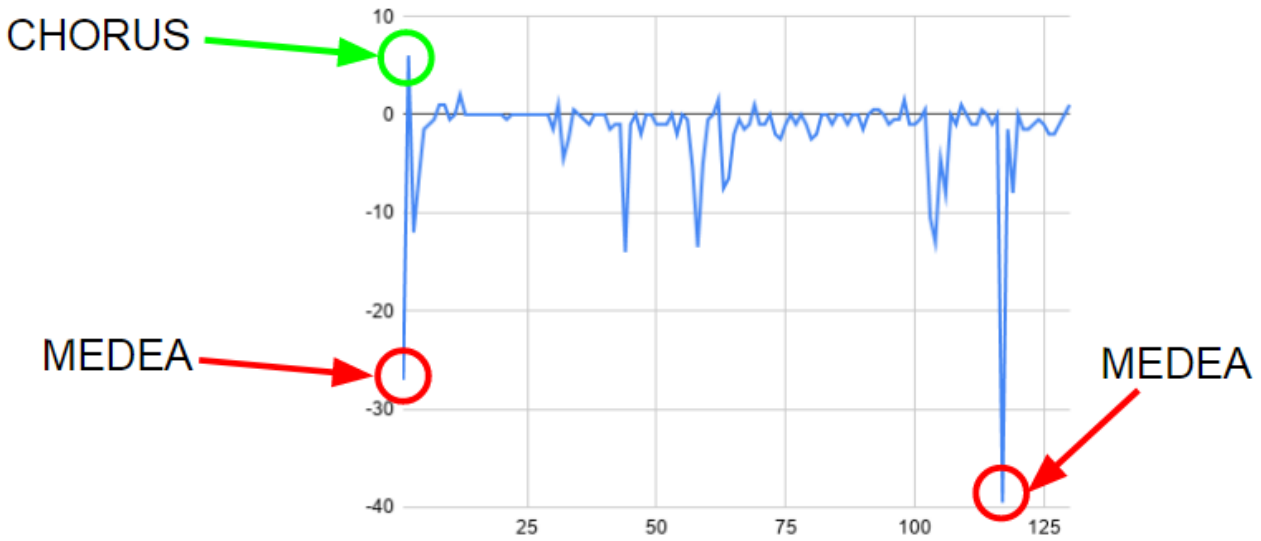


- **Output**: lexicon of 1,293 lemmas with five-value scores

Use Case

Analysis of the tragedy "**Medea**" by Seneca:

- Gold Standard + Silver Standard + Induced Lexicon = 3,253 lemmas
- lemmatized and PoS-tagged text from "Opera Latina"
- lexicon-based analysis at the level of clusters of lines



PART III

MODELING & LINKING A SENTIMENT LEXICON

What

Inclusion of a prior polarity lexicon of Latin lemmas in a Knowledge Base (KB) of interoperable linguistic resources using Semantic Web and Linked Data standards and practices:

- *LatinAffectus*

	PoS		Polarity			
	ADJ	NOUN	POS	NEUT	NEG	**TOT**
Gold Standard	454	690	231	612	301	**1,144**
Silver Standard	512	781	271	689	333	**1,293**
LatinAffectus	966	1,471	502	1301	634	**2,437**

REFERENCE: Sprugnoli, R., Mambrini, F., Moretti, G., Passarotti, M. (2020). *Towards the Modelling of Polarity in a Latin Knowledge Base*. In Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020).

Why

- To **enrich** the information on (a subset of) lemmas of the LiLa KB
- To **enhance the interoperability** taking advantage of past work on formal representations of linguistic resources and services for sentiment analysis

modelling

We re-use **formal representation frameworks** to describe the lexical resource and the sentiment properties of each entry:

- lemon
- lime
- Ontolex
- Marl

Modelling
Lexicon

https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon

LatinAffectus

http://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon

ENTITÀ DI TIPO: **E31**

dcterms:**description**	Gold: prior polarity lexicon of Latin lemmas created by two experts of Latin language and culture following a multi-stage process and an extensive reconciliation phase. It follows a five-way classification: 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative) Silver: prior polarity lexicon built by deriving new entries through synonym, antonym and derivational relations with the entries in the gold standard.
rdfs:**label**	LatinAffectus
dcterms:**creator**	Rachele Sprugnoli, Daniela Corbetta, Andrea Peverelli
dcterms:**title**	LatinAffectus - sentiment lexicon for latin.
dcterms:**contributor**	Francesco Mambrini, Giovanni Moretti
rdf:**about**	https://github.com/CIRCSE/Latin_Sentiment_Lexicons
rdf:**type**	crm:E31
	lime:Lexicon
dcterms:**publisher**	<http://www.wikidata.org/entity/Q89883181>
lime:**entry**	<http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalEntry/lemma_4858>
	<http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalEntry/lemma_6578>
	<http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalEntry/lemma_19360>

Modelling
Entry

malus

http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalEntry/lemma_111418

ENTITÀ DI TIPO: **LexicalEntry**

rdfs:**label**	malus
rdf:**type**	ontolex:LexicalEntry
ontolex:**canonicalForm**	<http://lila-erc.eu/data/id/lemma/111418> ↳ malus
ontolex:**sense**	<http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalSense/lemma_111418> ↳ Prior sense of malus

RELAZIONI INVERSE

è lime:**entry** di	1 risorsa

Prior sense of malus

http://lila-erc.eu/data/lexicalResources/LatinAffectus/id/LexicalSense/lemma_111418

ENTITÀ DI TIPO: **LexicalSense**

rdfs:**label**	Prior sense of malus
<http://www.gsi.dit.upm.es/ontologies/marl/ns#**polarityValue**>	-1.0E0
rdf:**type**	ontolex:LexicalSense
<http://www.gsi.dit.upm.es/ontologies/marl/ns#**hasPolarity**>	<http://www.gsi.dit.upm.es/ontologies/marl/ns#Negative>

RELAZIONI INVERSE

è ontolex:**sense** di	1 risorsa

Linking

Semi-automatic linking of the LatinAffectus lexical entries to their corresponding lemmas in the LiLa KB:

1. automatic matching of 2.084 lemmas (85.5%)
2. manual linking of:
   - 107 Medieval or New Latin lemmas: e.g. *praesuppositio* “assumption”
   - 246 ambiguous lemmas

fīdēs

feminine noun III declension

View the declension of this word

1 chord, instrument string
2 constellation Lyra
3 stringed instrument
4 lyre

fīdēs

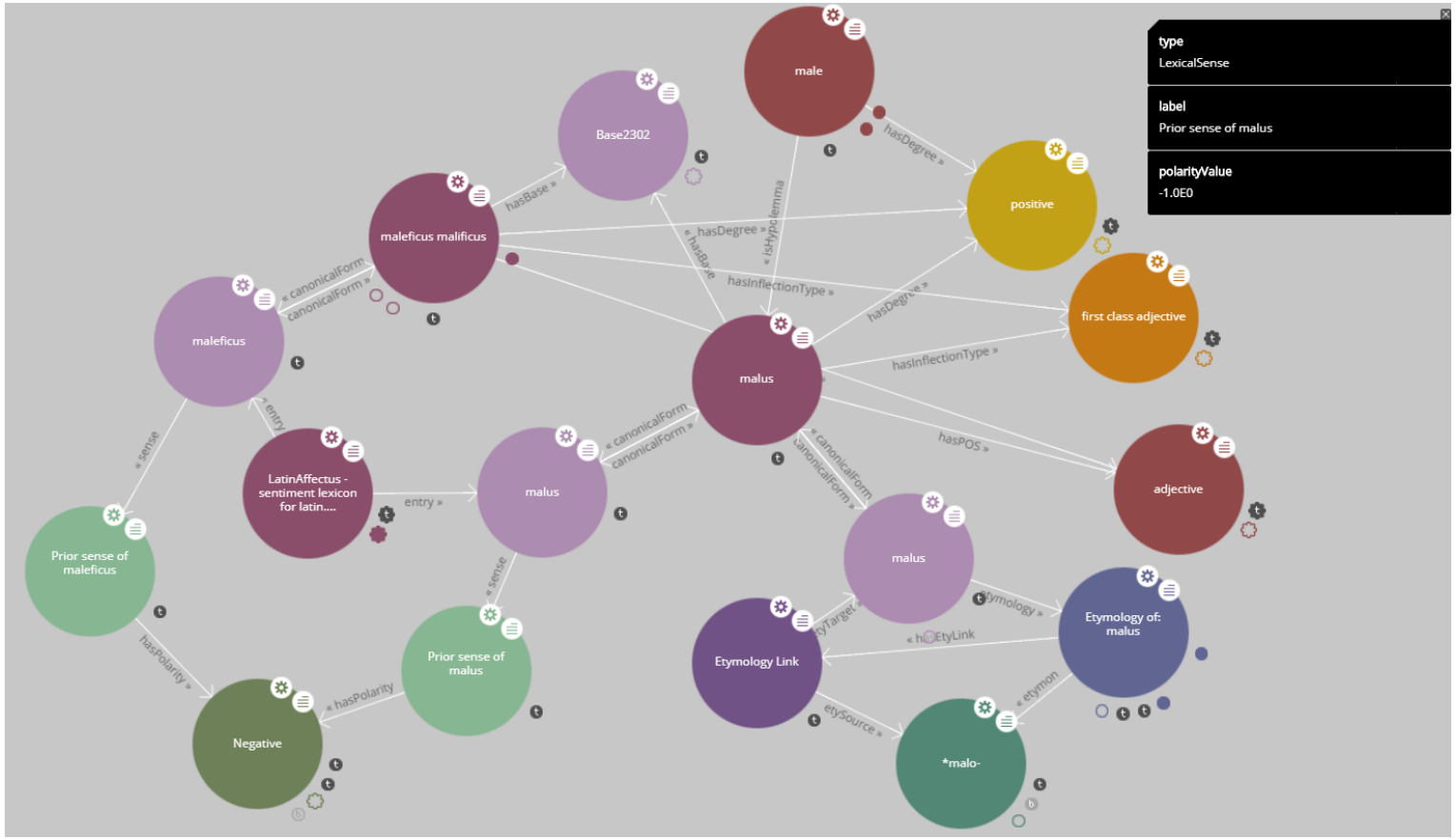
feminine noun V declension

View the declension of this word

1 faith, loyalty
2 honesty
3 credit
4 confidence, trust, belief
5 good faith

Modelling & Linking

DEMO:
https://lila-erc.eu/lodlive/app_en.html?http://lila-erc.eu/data/id/lemma/111418



Use Case

How the interoperability between the resources connected in LiLa, and particularly LatinAffectus, can **support** research in the Humanities?

Use Case

How the interoperability between the resources connected in LiLa, and particularly LatinAffectus, can **support** research in the Humanities?

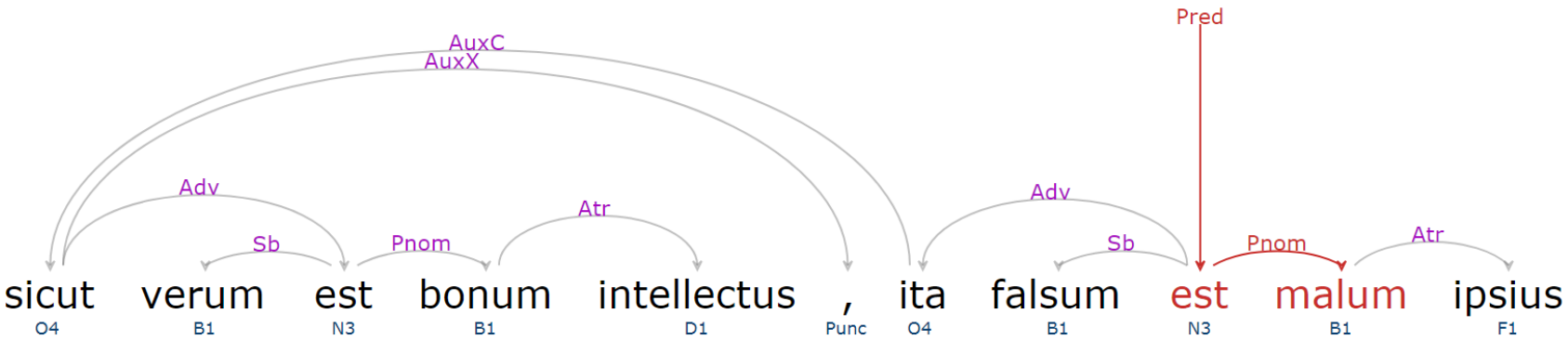
***Index Thomisticus* Treebank** (ITTB):

- four books of the treatise *Summa contra Gentiles* of the philosopher Thomas Aquinas (XIII Century)
- 450,515 tokens
- morphological and syntactic annotation

What is the nature of Evil?

What is the nature of Evil?

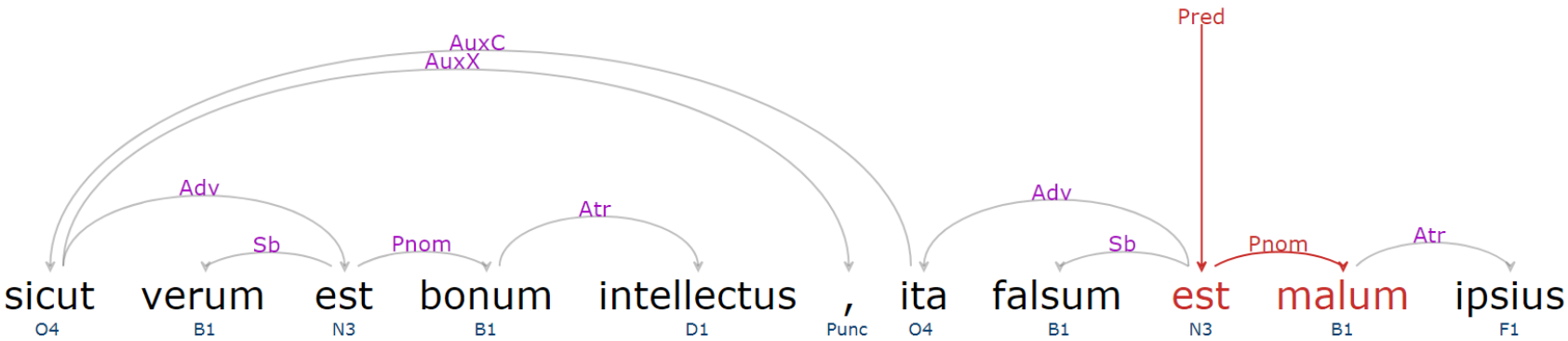
“as the true is the good of the intellect, so the false is its evil”



Summa contra Gentiles, lib. 1, cap. 61, num. 8

What is the nature of Evil?

“as the true is the good of the intellect, so the false is its evil”

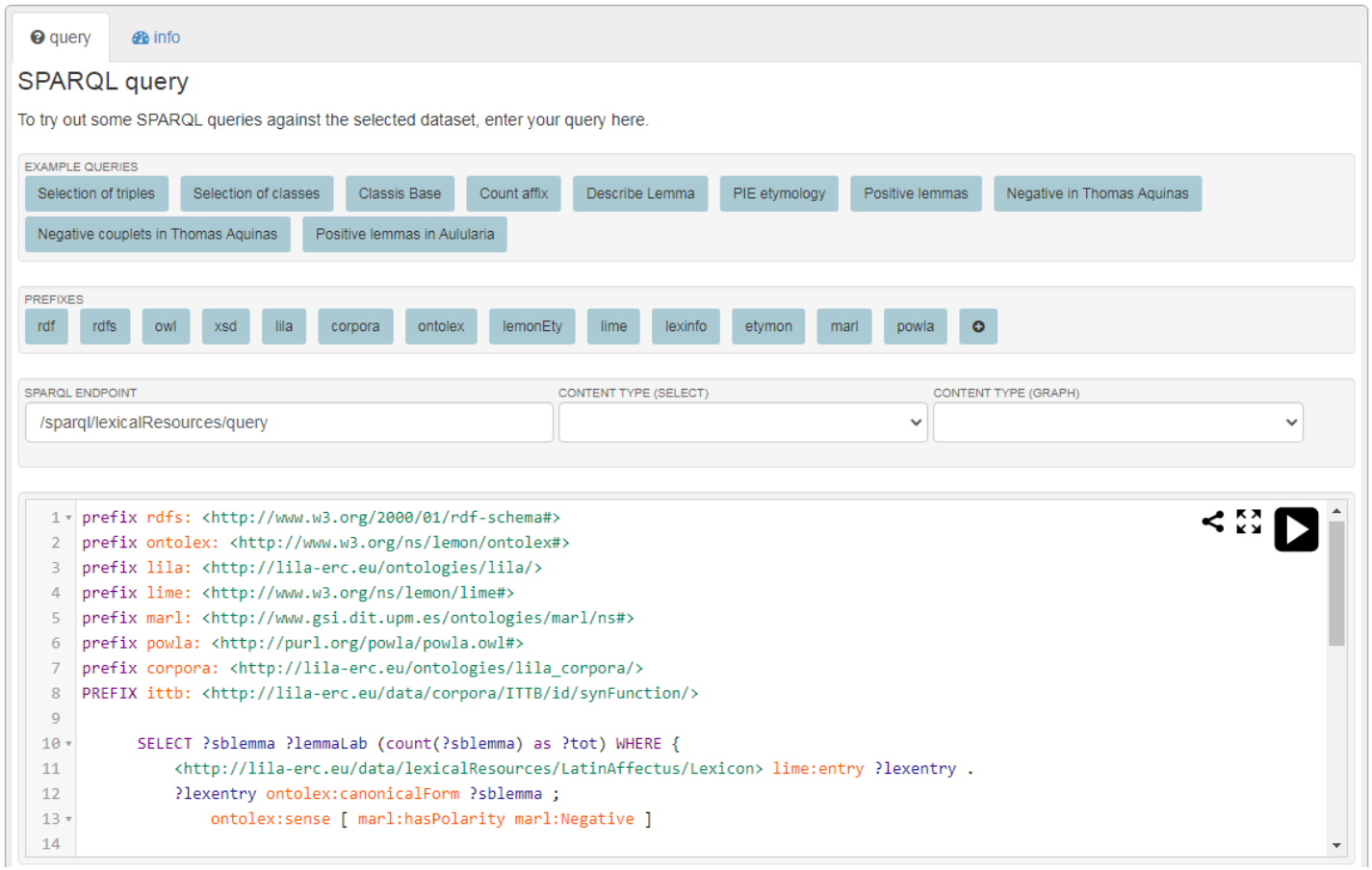


Summa contra Gentiles, lib. 1, cap. 61, num. 8

What are the lemmas with a negative polarity that are the subject of the verb *sum* (“to be”) in a copular construction?

Use Case
SPARQL Query

DEMO:
https://lila-erc.eu/sparql/index.html



The most frequent **negative subjects** of copular constructions in the ITTB

Lemma	English translation	Occurrences
malum	'evil'	38
mors	'death'	9
corruptio	'destruction', 'decay', 'corruption'	8
fornicatio	'fornication'	4

- *malum*: technical word for the concept of evil itself
- *corruptio*: destruction can be a good thing if it implies the destruction of evil
- *fornicatio*: always subject of a copular construction where *peccatum* ("sin") is the predicate nominal

Discussion

- **Coverage** of *LatinAffectus*

  - *privatio* “deprivation” (25 occurrences)
  - *paupertas* “poverty” (9)
  - *peccatum* “sin” (7)
  - *defectus* “defect” (6)

- No need of **pre-processing**

- **Interoperability** leads to a fruitful crossing between different resources to allow a broad range of corpus-based researches:

  - a dependency treebank
  - the *LatinAffectus* lexicon
  - the collection of lemmas of the LiLa KB

PART IV
CONCLUSIONS

Future Works

- To extend the **coverage** –> WORK IN PROGRESS
- To clean **Latin WordNet** and add it to the LiLa KB so to assign different polarities to different synsets –> WORK IN PROGRESS

lemma	synset_id	definition
capitolium	n#06188340	the federal government of the United States
voco	v#00720710	send a message or attempt to reach someone by radio, phone, etc; make a signal to in order to transmit a message [...]

- To generate **time-specific** sentiment lexicons –> PRELIMINARY EXPERIMENT on "Computational Historical Semantics" corpus

What We Have Learned

- Importance of defining good practices to include domain experts (e.g. philologists, classicists) in the development loop

- Combining computational linguistics with historical or classical texts leads us dealing with new challenges:
  - different text genres
  - lack of native speakers
  - abundance *versus* scarcity of data
  - creation of tools and interfaces for not-tech-savvy users

Query Interface, Triplestore

- https://lila-erc.eu/query/
- https://lila-erc.eu/sparql/

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

Lexicons

- 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

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

@RSprugnoli / @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.