Latin Embeddings and the LiLa Knowledge Base of Interlinked Resources for Latin

Marco Passarotti, Rachele Sprugnoli

Computational Approaches to ancient Greek and Latin
Groningen, 2 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

Latin Word Embeddings

- Training and Testing Lemma Embeddings
- Diachronic Analysis
- Induction of Sentiment Lexicons
- Modeling

Conclusions

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

Why LiLa?
State of Affairs

Why LiLa?
State of Affairs

- Resources disconnected from each other (silos of LRs)

Why LiLa?
State of Affairs

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats

Why LiLa?
State of Affairs

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats
- Different representation schemes, query languages, annotation criteria and tagsets

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

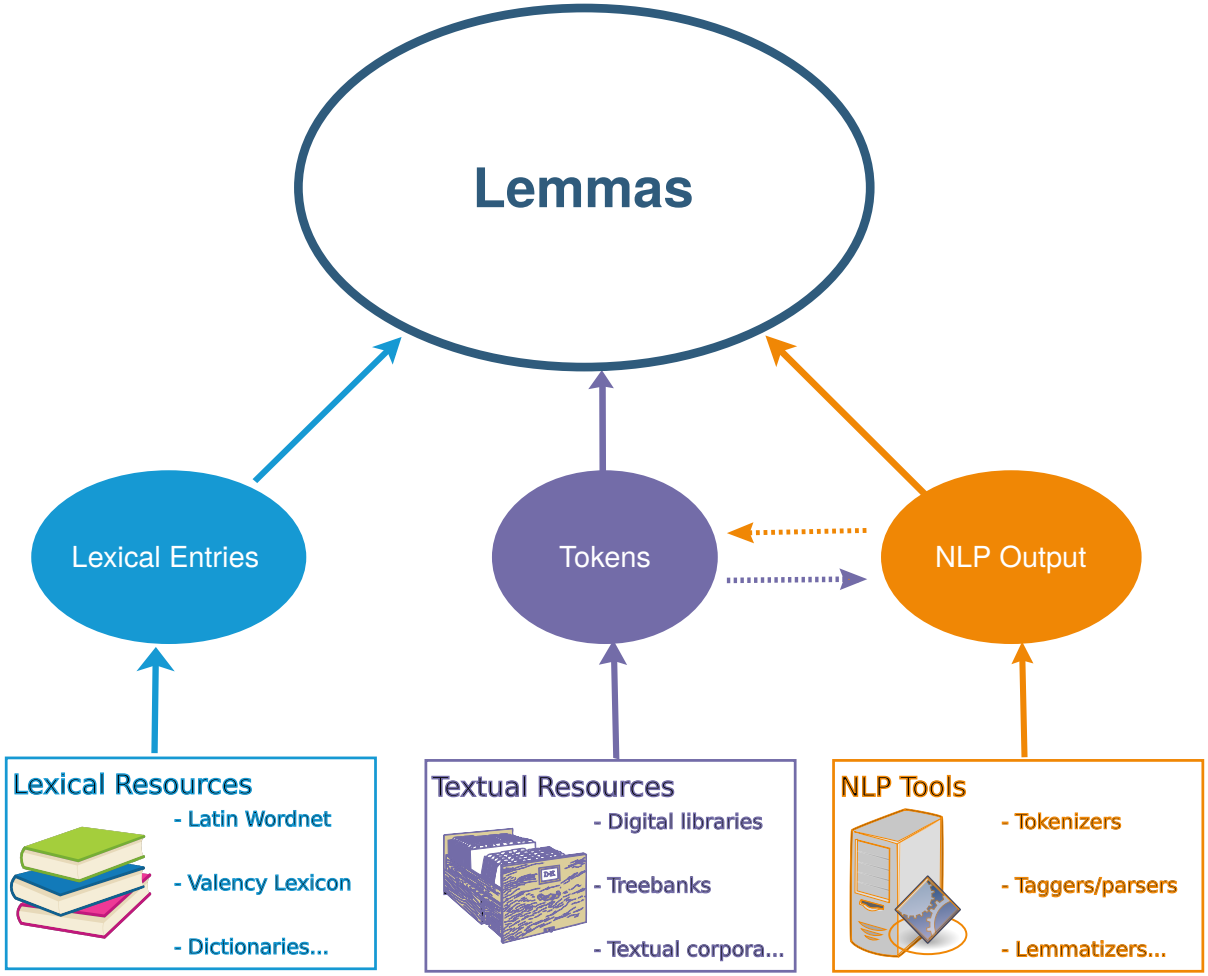
The Linked Data Principles

…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

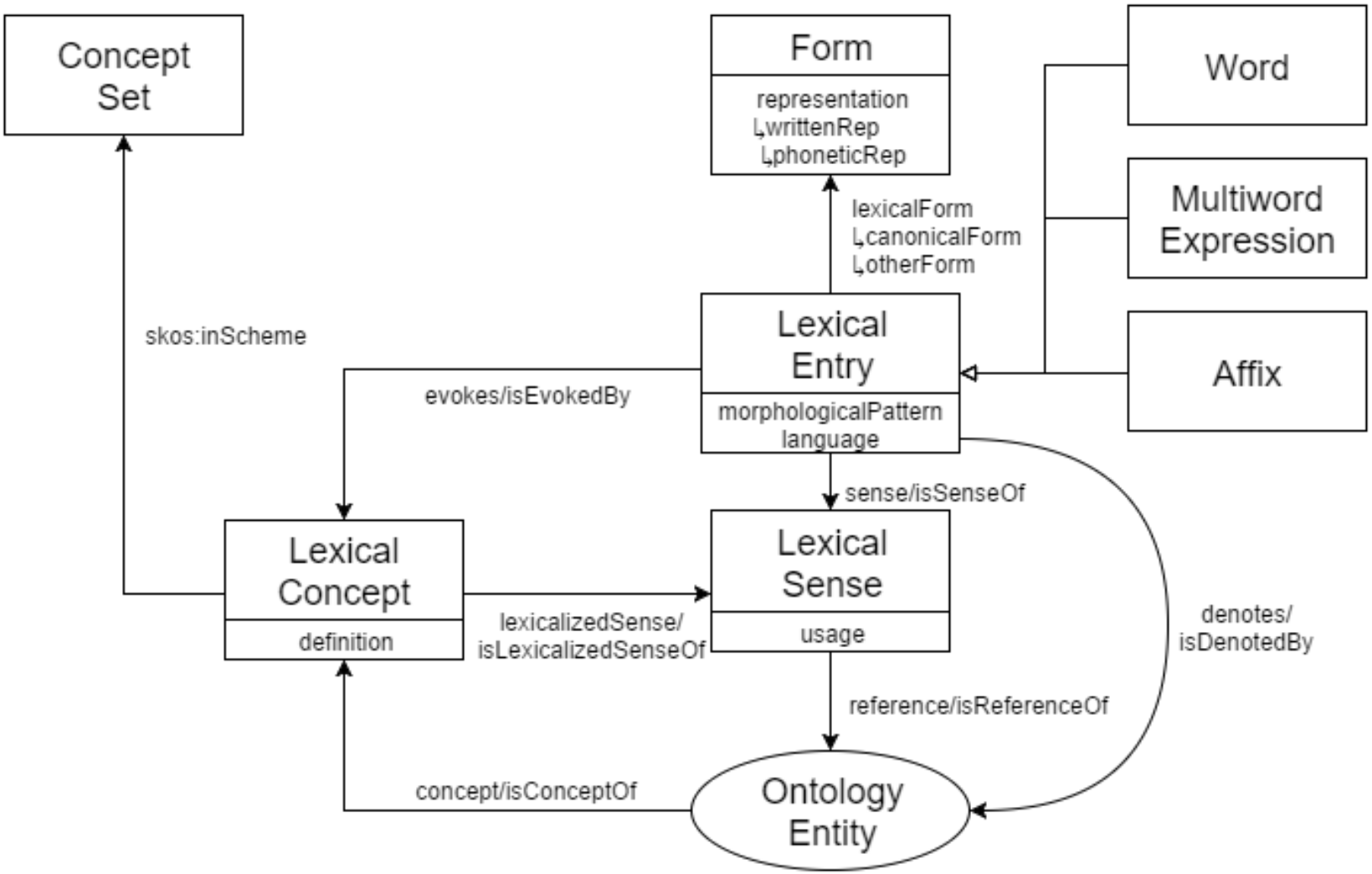
LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

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



Lemma Bank and Lexical Resources

The core of LiLa

Lemma *admiror* ‘to admire, to respect’

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

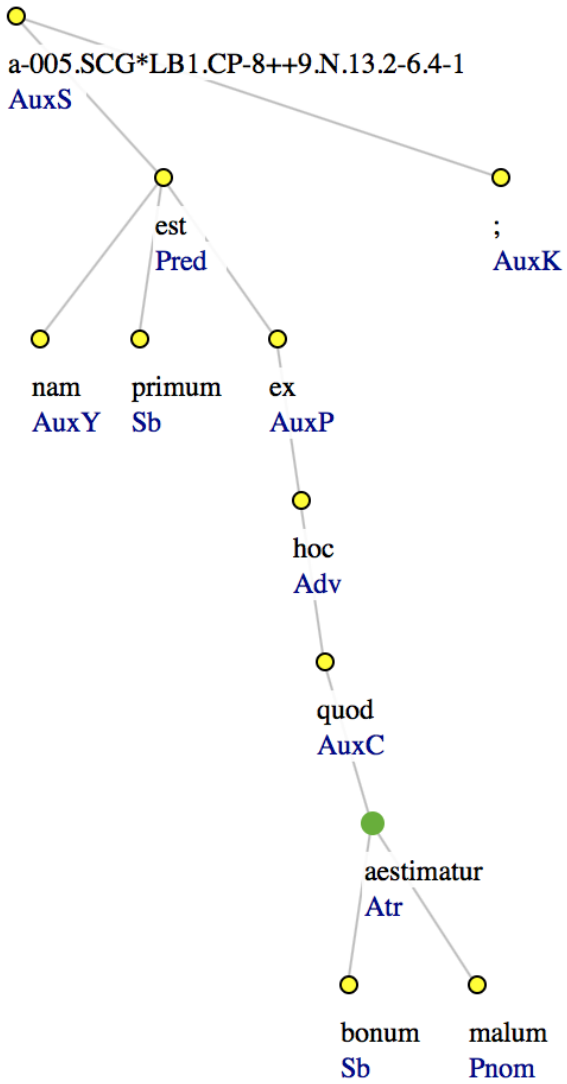
- Lemma Bank
- A derivational lexicon (Word Formation Latin)
- A polarity lexicon (LatinAffectus)
- An etymological dictionary (De Vaan)
- A Valency Lexicon (Latin Vallex)
- A manually checked subset of the Latin WordNet
- A bilingual Latin-English dictionary /Lewis & Short)

Textual Resources

Source: the *Index Thomisticus* Treebank (original scheme)

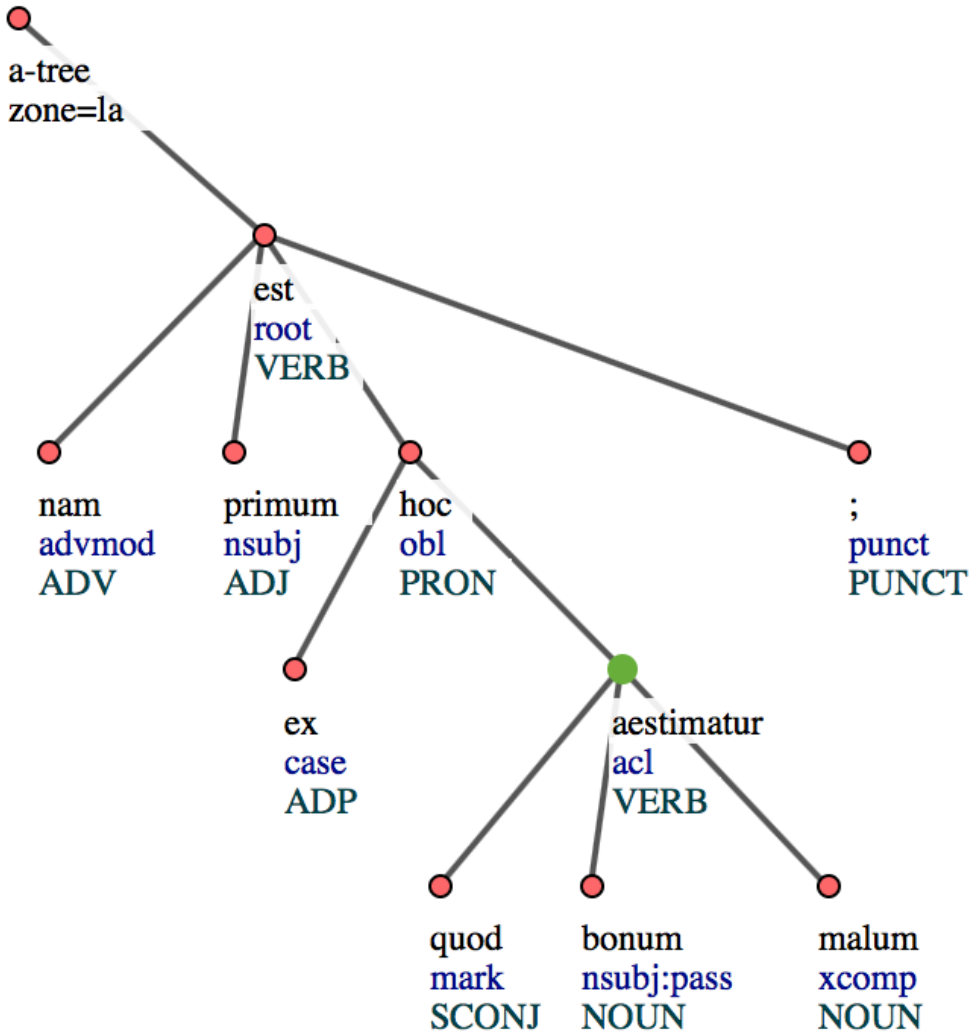
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

for the first arises because the good ***is judged*** *to be evil*; (Trans. Anton C. Pegis)



Textual Resources

Source: the *Index Thomisticus* Treebank (UD scheme)



Textual Resources

A token from the *Index Thomisticus* Treebank in LiLa

Token *aestimatur*

http://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.2-6.4-1W8

LiLa: Overview

Resources connected and upcoming connections

- **Textual Resources**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] Dante Search (700th death anniversary): ca. 46,000 tokens
  - [x] *Liber Abaci*, *Chapter VIII*: ca. 30,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexical Resources**
  - [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. 4,000 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
  - [x] Lewis & Short Dictionary
  - [ ] Lemma Embeddings
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 15 million triples**

Table of Contents

Lemma Embeddings of Classical Latin: Why

1. Supporting data-driven **socio-cultural studies** of the Latin world
2. Fostering the **interdisciplinary collaboration** between Computational Linguistics and Classical Studies
3. **Filling a void** in the literature:

	Word2Vec	FastText	Clean	Download	Evaluation
CoNLL	✓			✓	
Facebook		✓		✓	
Bamman	✓			✓	
CompHistSem	✓	✓	✓	✓	
LiLa	✓	✓	✓	✓	✓

Sprugnoli, R., Passarotti, M., & Moretti, G. (2019). Vir is to Moderatus as Mulier is to Intemperans - Lemma Embeddings for Latin. In Proceedings of the Sixth Italian Conference of Computational Linguistics (CLiC-it 2019).

Lemma Embeddings of Classical Latin

Texts taken from the **LASLA corpus**:

- manually annotated since 1961
- lemmas, PoS tags,inflectional features
- multi-genre
- 158 texts, 20 authors
- 1.7M words

Text pre-processing:

- conversion to CoNLL-U
- extraction of lemmas
- lower-casing
- conversion: v –> u

Lemma Embeddings of Classical Latin

Vector representations:

- Word2vec: treats each word in corpus like an atomic entity
- FastText: treats each word as composed of character ngrams

Models:

- Skip-gram: the distributed representation of the input word is used to predict the context
- CBOW: the distributed representations of context (or surrounding words) are combined to predict the word in the middle

Dimensions:

- 100
- 300

Lemma Embeddings: Intrinsic Evaluation

- **Synonym Selection Task**: select the correct synonym of a target lemma out of a set of possible answers

- **Benchmark**: 2,759 multiple-choice questions each involving 5 terms
  - 1 target lemma
  - 1 synonym of the target lemma taken from Latin synonym dictionaries
  - 3 decoy lemmas

TARGET WORDS	SYNONYM	DECOY WORDS		
exilis/thin	*macer*/emaciated	*moles*/pile	*mortalitas*/mortality	*audens*/daring
globus/ball	*sphaera*/sphere	*patronus*/defender	*breuitas*/brevity	*apex*/cap
cunctor/doubt	*haesito*/hesitate	*uito*/avoid	*conflo*/compose	*pondero*/weigh

Lemma Embeddings: Intrinsic Evaluation

- **Results**:
  - calculate the cosine similarity between the vector of the target lemma and that of the other lemmas
  - pick the candidate with the largest cosine
  - measure the correct-answer accuracy

	word2vec		fastText	
	cbow	skip-gram	cbow	skip-gram
100	81.14%	79.83%	80.57%	**86.91%**
300	80.86%	79.48%	79.43%	86.40%

- **Errors**:
  - meronymy: TARGET: *annalis* - SYN: *historia* - ANSWER: *charta*
  - morphological derivation: TARGET: *consors* - SYN: *particeps* - ANSWER: *sors*

Table of Contents

Lemma Embeddings: Diachronic Analysis

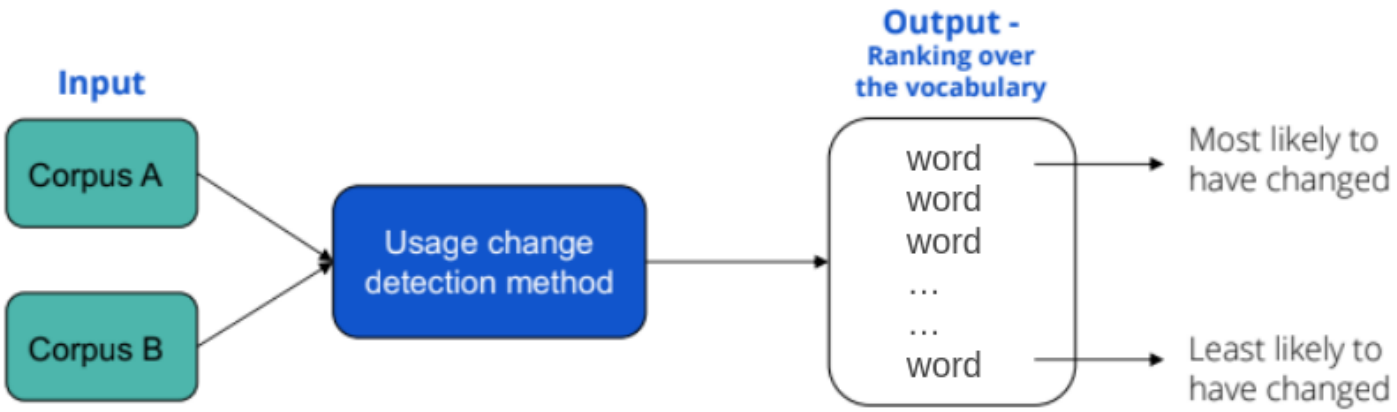
- The use of Latin spans more than two millennia
  - Classical Latin ≠ Medieval Latin

- **New embeddings** to compare:
  - *Opera Maiora* of Thomas Aquinas:
  - philosophical and religious works
  - 13th century
  - manually lemmatized in the *Index Thomisticus* project
  - 4.5 million words

Sprugnoli, R., Passarotti, M., & Moretti, G. Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas.IJCoL. Italian Journal of Computational Linguistics, 6(6-1), 29-45.

Lemma Embeddings: Diachronic Analysis

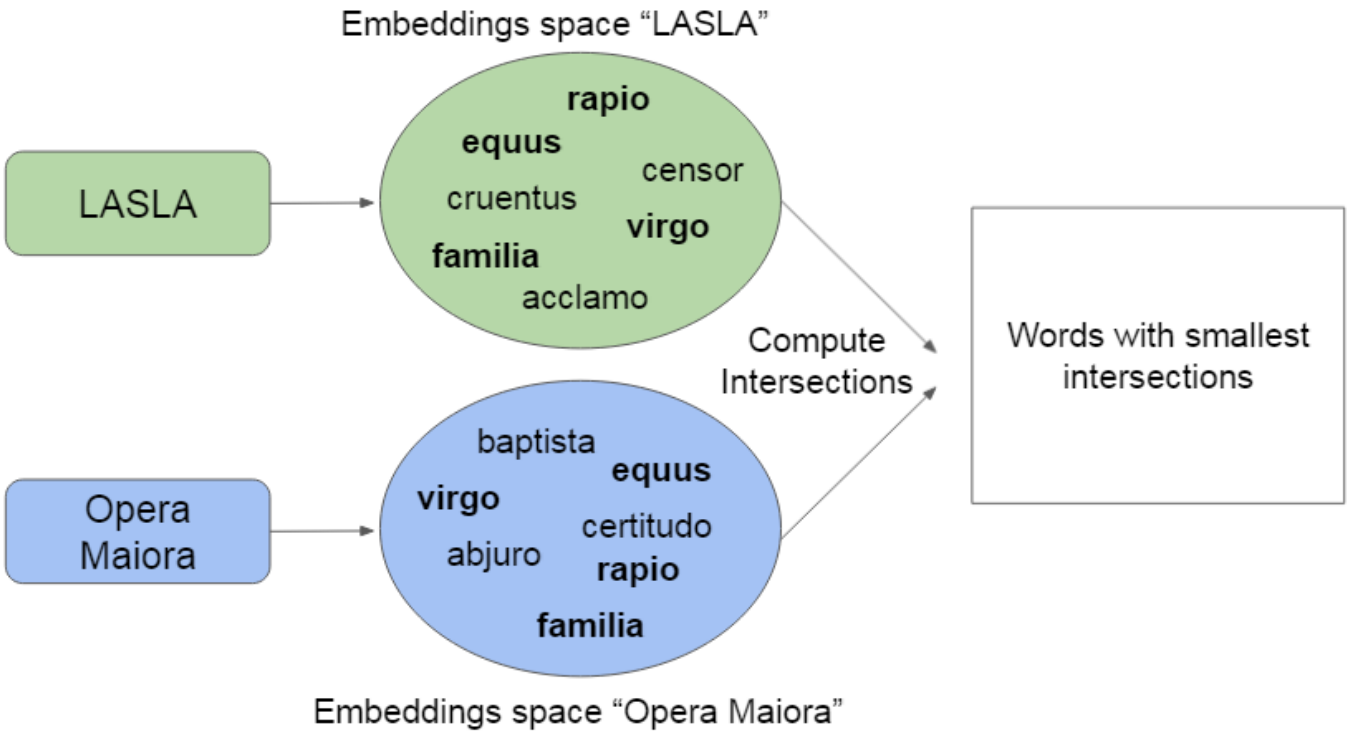
- USAGE CHANGE: identify words that are used differently over time periods



Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 538-555).

Lemma Embeddings: Diachronic Analysis

- Nearest neighbors are a proxy for meaning



Lemma Embeddings: Diachronic Analysis

- Example: *virgo* = girl of marriageable age, virgin

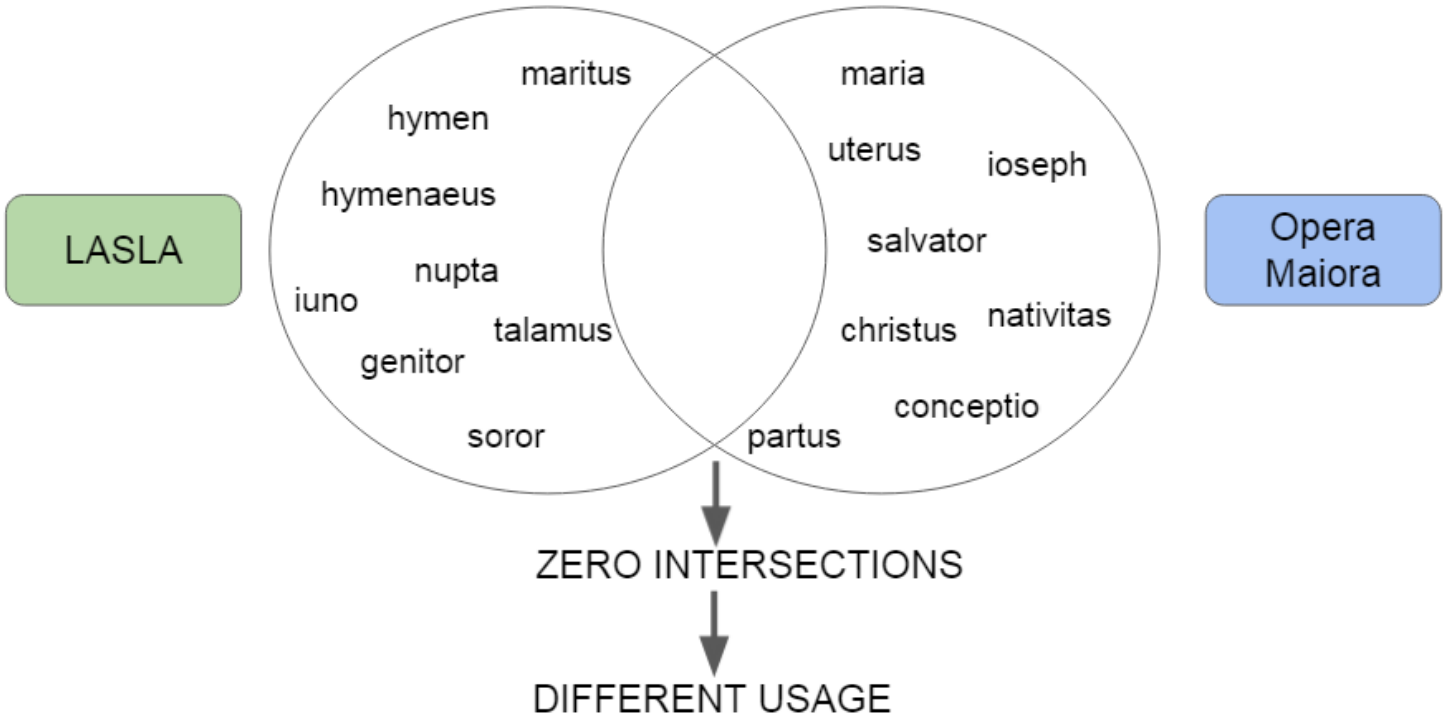


Table of Contents

Lemma Embeddings: Sentiment Lexicon Induction

«Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language.»
Bing Liu, "Sentiment Analysis and Opinion Mining" Morgan & Claypool Publishers, 2012

- Development of sentiment lexicons for Latin = list of words associated to scores expressing their prior polarity:
  - essential resource for both machine learning and lexicon-based sentiment analysis systems
  - set of lexicons created manually or automatically

Lemma Embeddings: Sentiment Lexicon Induction

- Automatic induction from word embeddings starting from a list of seed terms with known sentiment score
  - **seed terms**: 200 most frequent adjs and nouns from the LASLA corpus
  - **embeddings**: pre-trained with word2vec on LASLA corpus 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

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

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

Lemma Embeddings: Sentiment Analysis

Evaluation with respect to a manually annotated gold standard made of 1,144 lemmas:

- calculation of the **accuracy**
- comparison with the results obtained by creating a sentiment lexicon with a translation method

	TRANSLATION	INDUCTION
ADJ	64.9%	86.7%
NOUN	66.8%	62.5%
MICRO-AVG	66.1%	**74.4%**

Lemma Embeddings: Sentiment Analysis

Possibility to generate **time-specific** sentiment lexicons

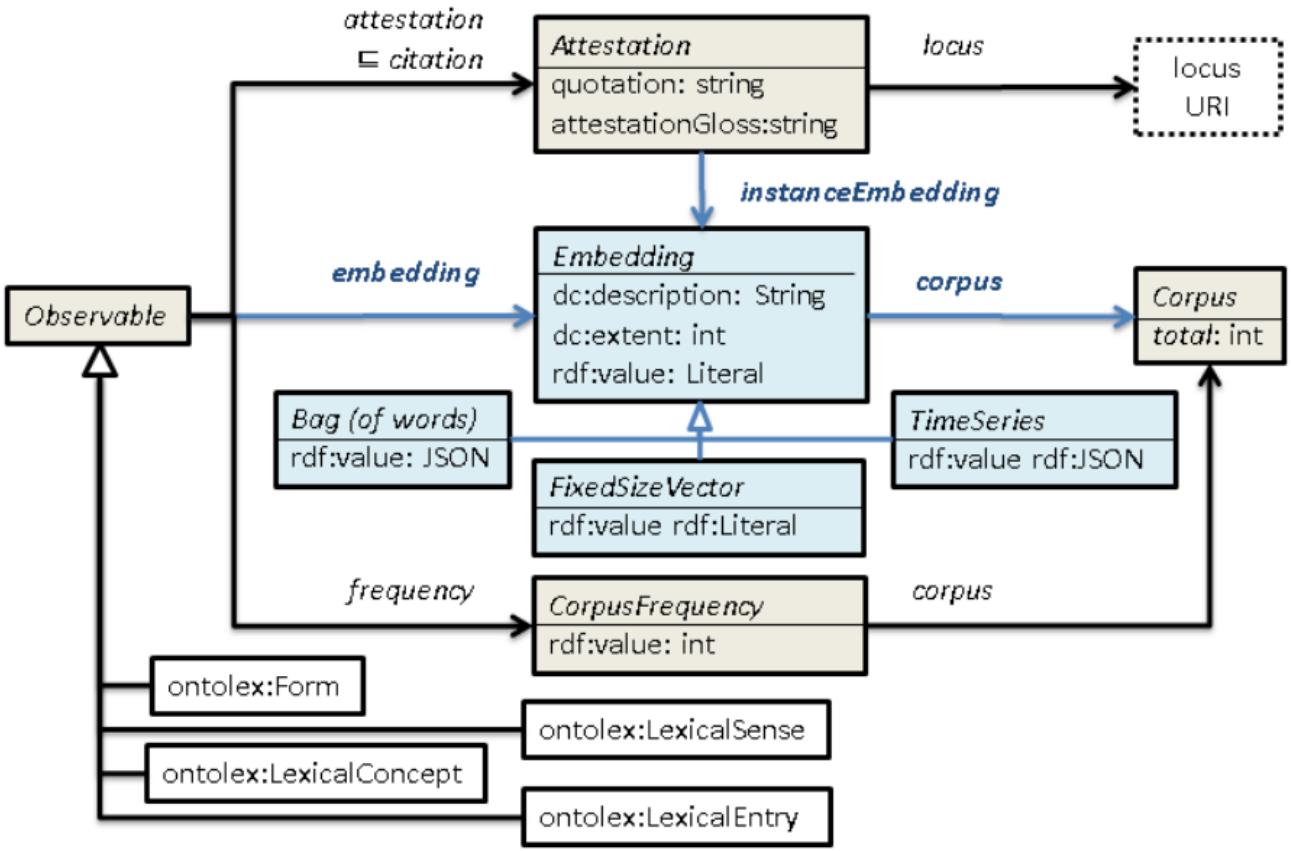
- Induction using embeddings trained on the *Computational Historical Semantics* corpus
  - 904.400 lemmas
  - Latin documentary texts
  - between the 2nd and the 15th century AD

- 71% of the entries in the induced lexicon were different from the ones obtained on the LASLA corpus

LEMMA	PoS	SENTIMENT
archangelus 'archangel'	noun	positive
peccatrix 'female sinner'	noun	negative
criminosus 'guilty'	adj	negative
poenalis 'penal'	adj	negative

Table of Contents

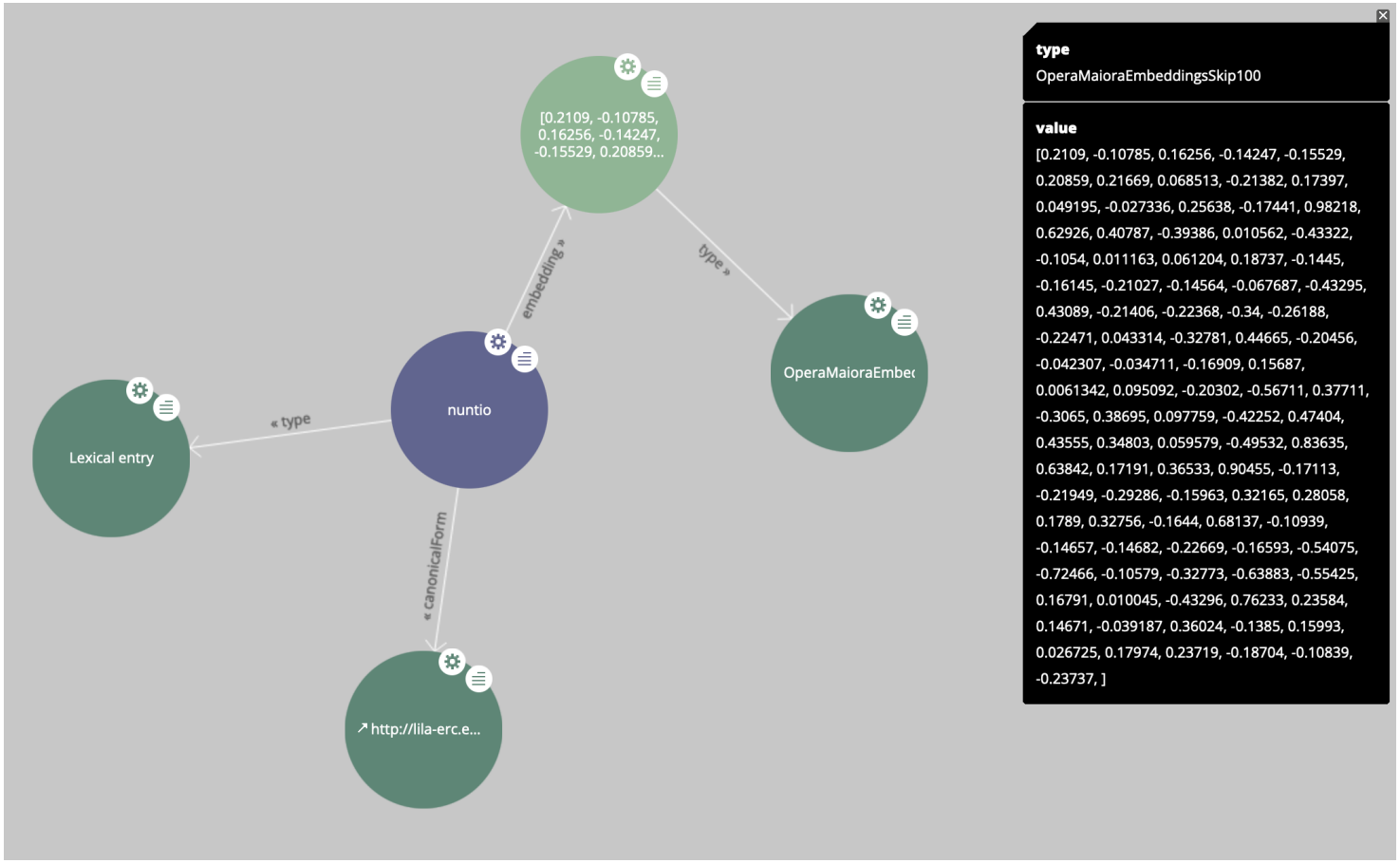
Frequency Attestation and Corpus (FrAC) module



Chiarcos, C., Declerck, T., & Ionov, M. (2021). Embeddings for the Lexicon: Modelling and Representation. In Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6) (pp. 13-19).

Frequency Attestation and Corpus (FrAC) module

- Example: lemma embeddings of *nuntio* 'to announce' pre-trained on *Opera Maiora*, skip-gram model, 100 dimensions



Open Challenges

It is not so easy ...but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

Open Challenges

It is not so easy …but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)

Open Challenges

It is not so easy ...but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.

Open Challenges

It is not so easy ...but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.
- Models still missing for several types of (meta)data: e.g. for critical editions

Open Challenges

It is not so easy …but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.
- Models still missing for several types of (meta)data: e.g. for critical editions
- Community-based effort: persuading resource developers to adopt LOD practices and reaching consensus around shared vocabularies, ontologies, data categories etc.

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.