Vir is to *Moderatus* as *Mulier* is to *Intemperans*

Lemma Embeddings for Latin

Rachele Sprugnoli, **Marco Passarotti**, **Giovanni Moretti**
rachele.sprugnoli@unicatt.it
marco.passarotti@unicatt.it
giovanni.moretti@unicatt.it

CLiC-it 2019 | Bari | November 13, 2019

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.

WHAT

A new set of lemma embeddings for the Latin language:

- trained on a **manually** annotated corpus
- **Classical** era
- quantitative & qualitative **evaluation**
- **diachronic** analysis

WHAT

A new set of lemma embeddings for the Latin language:

- trained on a **manually** annotated corpus
- **Classical** era
- quantitative & qualitative **evaluation**
- **diachronic** analysis

Embeddings, evaluation benchmark, visual exploration interface **all available online**:

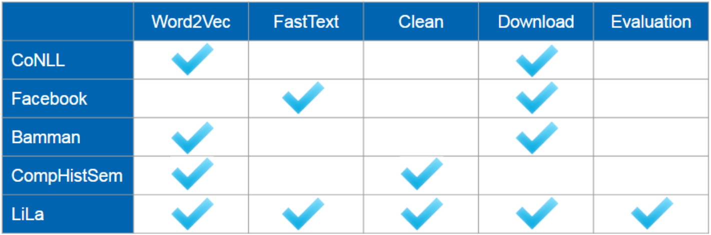
- `https://embeddings.lila-erc.eu`

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:

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:

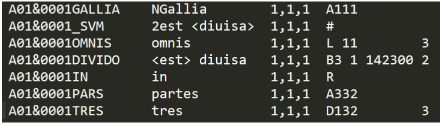


	Word2Vec	FastText	Clean	Download	Evaluation
CoNLL	✓			✓	
Facebook		✓		✓	
Bamman	✓			✓	
CompHistSem	✓		✓		
LiLa	✓	✓	✓	✓	✓

DATASET

“Opera Latina” corpus:

- manually annotated since 1961 by LASLA
  - lemmas, PoS tags, inflectional features
- multi-genre
- 158 texts, 20 authors
- 1.7M words, 24K unique lemmas

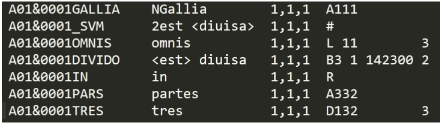


```
A01&0001GALLIA    NGallia         1,1,1  A111
A01&0001_SVM      2est <diuisa>   1,1,1  #
A01&0001OMNIS     omnis           1,1,1  L 11          3
A01&0001DIVIDO    <est> diuisa    1,1,1  B3 1 142300 2
A01&0001IN        in              1,1,1  R
A01&0001PARS      partes          1,1,1  A332
A01&0001TRES      tres            1,1,1  D132          3
```

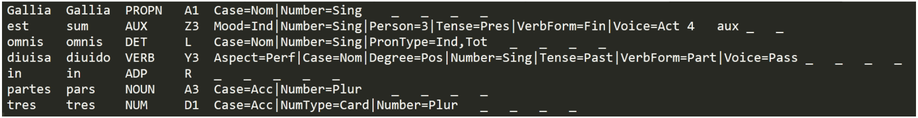
DATASET

“Opera Latina” corpus:

- manually annotated since 1961 by LASLA
  - lemmas, PoS tags, inflectional features
- multi-genre
- 158 texts, 20 authors
- 1.7M words, 24K unique lemmas



```
A01&0001GALLIA    NGallia         1,1,1  A111
A01&0001_SVM      2est <diuisa>   1,1,1  #
A01&0001OMNIS     omnis           1,1,1  L 11           3
A01&0001DIVIDO    <est> diuisa    1,1,1  B3 1 142300 2
A01&0001IN        in              1,1,1  R
A01&0001PARS      partes          1,1,1  A332
A01&0001TRES      tres            1,1,1  D132           3
```



```
Gallia  Gallia  PROPN   A1  Case=Nom|Number=Sing    _   _   _   _
est     sum     AUX     Z3  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 4   aux _   _
omnis   omnis   DET     L   Case=Nom|Number=Sing|PronType=Ind,Tot   _   _   _   _
diuisa  diuido  VERB    Y3  Aspect=Perf|Case=Nom|Degree=Pos|Number=Sing|Tense=Past|VerbForm=Part|Voice=Pass _   _   _   _
in      in      ADP     R   _   _   _   _   _
partes  pars    NOUN    A3  Case=Acc|Number=Plur    _   _   _   _
tres    tres    NUM     D1  Case=Acc|NumType=Card|Number=Plur   _   _   _   _
```

EXPERIMENTAL SETUP

Text pre-processing:

- extraction of lemmas, lower-casing, conversion: v –> u

EXPERIMENTAL SETUP

Text pre-processing:

- extraction of lemmas, lower-casing, conversion: v –> u

Training options:

- 2 vector representations: word2vec, FastText
- 2 dimensions: 100, 300
- 2 models: Skip-Gram, CBOW
- # negative samples: 25
- # threads: 20
- # iterations: 15
- minimal # occurrences: 5
- window size: 10 for Skip-Gram, 5 for CBOW

INTRINSIC EVALUATION

Synonym Selection Task

Select the correct synonym of a target lemma out of a set of possible answers

INTRINSIC EVALUATION

Synonym Selection Task

Select the correct synonym of a target lemma out of a set of possible answers

Creation of a **TOEFL-like benchmark** for Latin = multiple-choice questions each involving 5 terms:

- 1 target lemma
- 1 synonym of the target lemma
- 3 decoy lemmas

INTRINSIC EVALUATION

Synonym Selection Task

Benchmark creation:

- download and parse XML files of 4 digitized Latin synonyms dictionaries
- convert verbs lemmatized under the infinitive form into the 1st pers. sing. present indicative form (e.g. *accingere* –> *accingo*) using LEMLAT
- assemble 2,759 multiple-choice questions: 1 dictionary entry + 1 synonym + 3 random lemmas
- check of the output by a Latin language expert

INTRINSIC EVALUATION

Synonym Selection Task

Benchmark creation:

- download and parse XML files of 4 digitized Latin synonyms dictionaries
- convert verbs lemmatized under the infinitive form into the 1st pers. sing. present indicative form (e.g. *accingere* –> *accingo*) using LEMLAT
- assemble 2,759 multiple-choice questions: 1 dictionary entry + 1 synonym + 3 random lemmas
- check of the output by a Latin language expert

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

INTRINSIC EVALUATION

Synonym Selection Task

Results:

1. calculate the cosine similarity between the vector of the target lemma and that of the other lemmas
2. pick the candidate with the largest cosine
3. 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%

INTRINSIC EVALUATION

Synonym Selection Task

Results:

1. calculate the cosine similarity between the vector of the target lemma and that of the other lemmas
2. pick the candidate with the largest cosine
3. 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: other types of linguistic and semantic relations emerge

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

QUALITATIVE EVALUATION

Rare Lemma Embeddings

quality of the nearest neighbors of lemmas appearing between 5 and 10 times in “Opera Latina”

QUALITATIVE EVALUATION

Rare Lemma Embeddings

quality of the nearest neighbors of lemmas appearing between 5 and 10 times in “Opera Latina”

Examples of the nearest neighbors of rare lemmas: an asterisk marks neighbors that two Latin experts judged as most semantically-related to the target lemma.

Target Lemma	fastText-skip-100	word2vec-skip-100
contrudo/to thrust	***protrudo****/to thrust forward ***extrudo****/to thrust out	*infodio*/to bury *tabeo*/to melt away
frugaliter/thriftily	***frugalis****/thrifty ***frugalitas****/economy	***frugi****/frugal *quaerito*/to seek earnestly
auspicatus/consecrated by auspices	***auspicato****/after taking the auspices ***auspicium****/auspices	*erycinus*/Erycinian *parilia*/the feast of Pales

A DIACHRONIC PERSPECTIVE

The use of Latin spans more than two millennia

Classical Latin $\neq$ Medieval Latin

A DIACHRONIC PERSPECTIVE

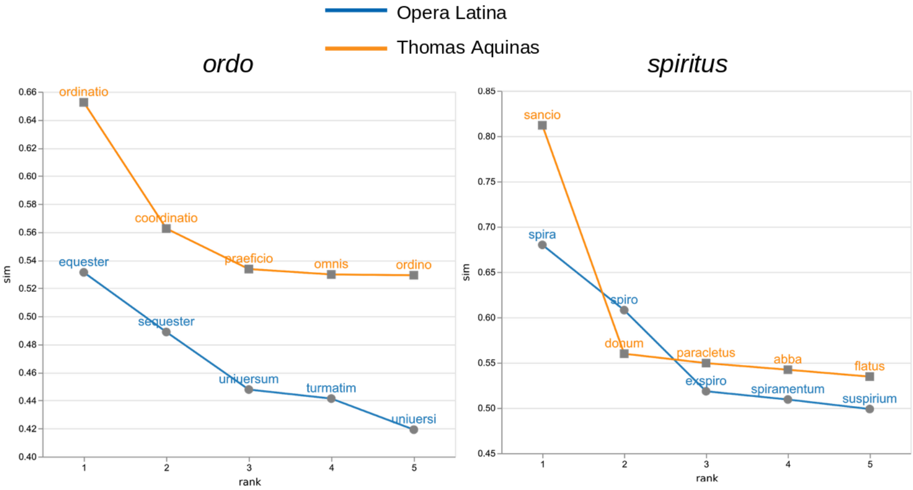
The use of Latin spans more than two millennia

Classical Latin ≠ Medieval Latin

New lemma vectors trained on the **“Opera Maiora”** of Thomas Aquinas (*Index Thomisticus*):

- philosophical and religious works
- 13th century
- manually lemmatized
- 4.5 million words
- trained with fastText skip-gram 100 dimensions
- aligned with Wasserstein Procrustes algorithm

A DIACHRONIC PERSPECTIVE



CONCLUSIONS AND FUTURE WORKS

Our contribution is based on an **interdisciplinary approach**:

- set of new Latin embeddings
- new benchmark for the synonym selection task
- aligned embeddings for diachronic comparison

CONCLUSIONS AND FUTURE WORKS

Our contribution is based on an **interdisciplinary approach**:

- set of new Latin embeddings
- new benchmark for the synonym selection task
- aligned embeddings for diachronic comparison

Future works:

- develop other benchmarks to extend the quantitative evaluation
- extend the diachronic analysis

CONCLUSIONS AND FUTURE WORKS

Our contribution is based on an **interdisciplinary approach**:

- set of new Latin embeddings
- new benchmark for the synonym selection task
- aligned embeddings for diachronic comparison

Future works:

- develop other benchmarks to extend the quantitative evaluation
- extend the diachronic analysis

EvaLatin

Thanks!

Get in touch

Università Cattolica del Sacro Cuore
CIRCSE Research Centre

🌐 https://lila-erc.eu

✉ info@lila-erc.eu / rachele.sprugnoli@unicatt.it

@ERC_LiLa / @RSprugnoli

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.