LiLa: Linking Latin

Building a Knowledge Base of Linguistic Resources for Latin

The LiLa Team
info@lila-erc.eu

First LiLa Workshop: *Linguistic Resources & NLP Tools for Latin*
Milan | 3-4 June 2019

Why, What & How (M. Passarotti)

LiLa Architecture (F. Mambrini)

Resources-1: Derivational Morphology & Valency Lexicon (E. Litta)

Resources-2: Latin WordNet (G. Franzini & A. Peverelli)

NLP-1: Part-of-speech Tagging & Lemmatisation (F. M. Cecchini)

NLP-2: Upcoming Resources in LiLa & a New Initiative (R. Sprugnoli)

Table of Contents

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

- ▶ Textual Resources
- ▶ Lexical Resources
- ▶ NLP Tools

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

- ▶ Textual Resources
- ▶ Lexical Resources
- ▶ NLP Tools

Scattered and unconnected

To make sense of this quantity of empirical data:

To make sense of this quantity of empirical data:

- ► to extract maximum benefit from our research investments

To make sense of this quantity of empirical data:

- ▶ to extract maximum benefit from our research investments
- ▶ to impact and improve the life of Classicists through exploitable computational resources and tools

To make sense of this quantity of empirical data:

- ▶ to extract maximum benefit from our research investments
- ▶ to impact and improve the life of Classicists through exploitable computational resources and tools

From Information to Knowledge

2018-2023

A collection of interoperable linguistics resources (and NLP tools) described with the same vocabulary for knowledge description

Interlinking as a Form of Interaction

LiLa is based on an ontology made of:

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)
- **Object properties**: ways in which classes and individuals can be related to one another: RDF triples.
  Labels from a restricted vocabulary of knowledge description:
  `hasLemma`, `hasPoS`

LiLa is based on an ontology made of:

- **Individuals**: instances of objects (one specific token, lemma etc.)
- **Classes**: types of objects/concepts (token, lemma, PoS etc.)
- **Data properties**: attributes that objects can/must have (morphological features for lemmas/tokens)
- **Object properties**: ways in which classes and individuals can be related to one another: RDF triples.
  Labels from a restricted vocabulary of knowledge description:
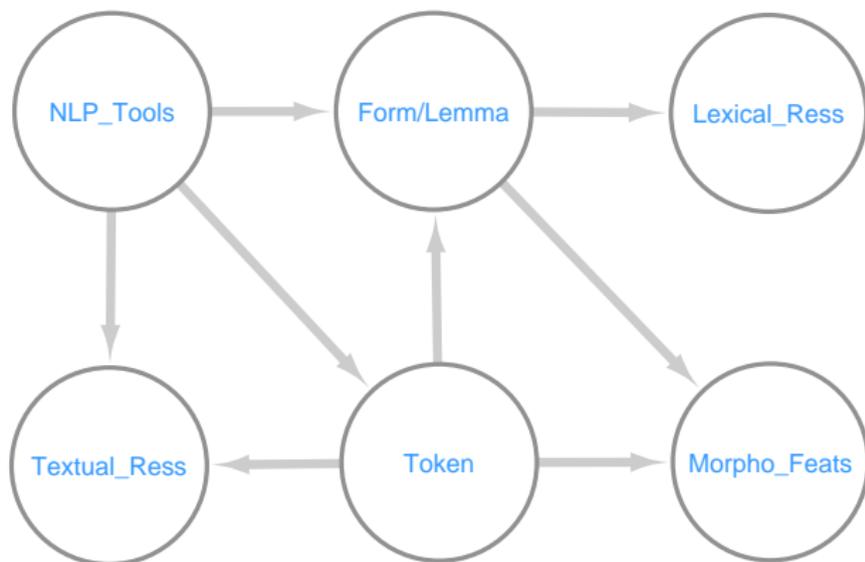 `hasLemma`, `hasPoS`

Each component of the ontology is uniquely identified through a URI.

Table of Contents

The golden rule:

Reuse as many standards as you can.

The golden rule:

Reuse as many standards as you can. *Extend, when you need to.*

The golden rule:

Reuse as many standards as you can. *Extend, when you need to. Create from scratch, if you really must.*

The golden rule:

Reuse as many standards as you can. Extend, when you need to. Create from scratch, if you really must.

LiLa is based on:

- ► the `Ontolex` family, for lexical information
- ► the `OLiA` bundle, for PoS tagging
- ► `NIF` (and `POWLA`?) for corpus annotation
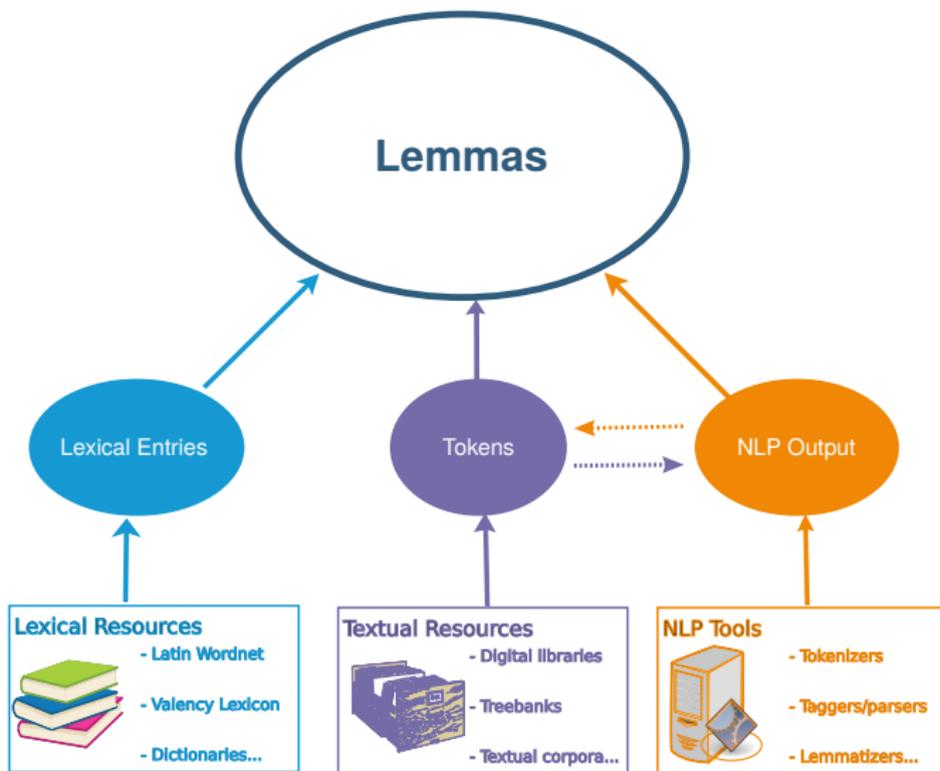
- 43,432 lemmas from Georges, 1913-1918; *OLD* and Gradenwitz, 1904;
- 82,556 lemmas from Du Cange, 1883-1887;
- 26,250 lemmas from Forcellini, 1940.
- WFL added.
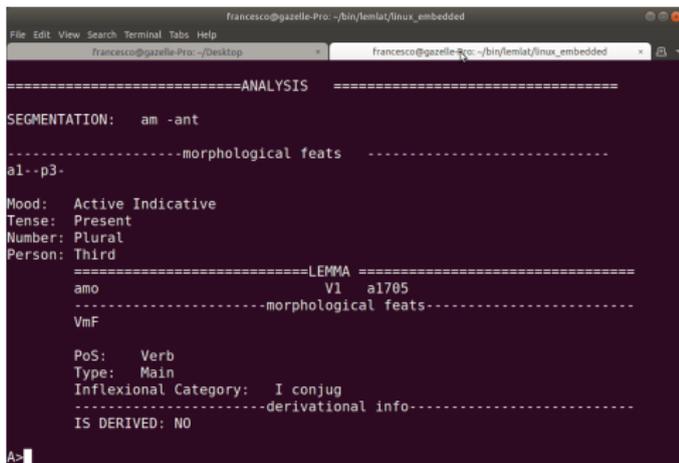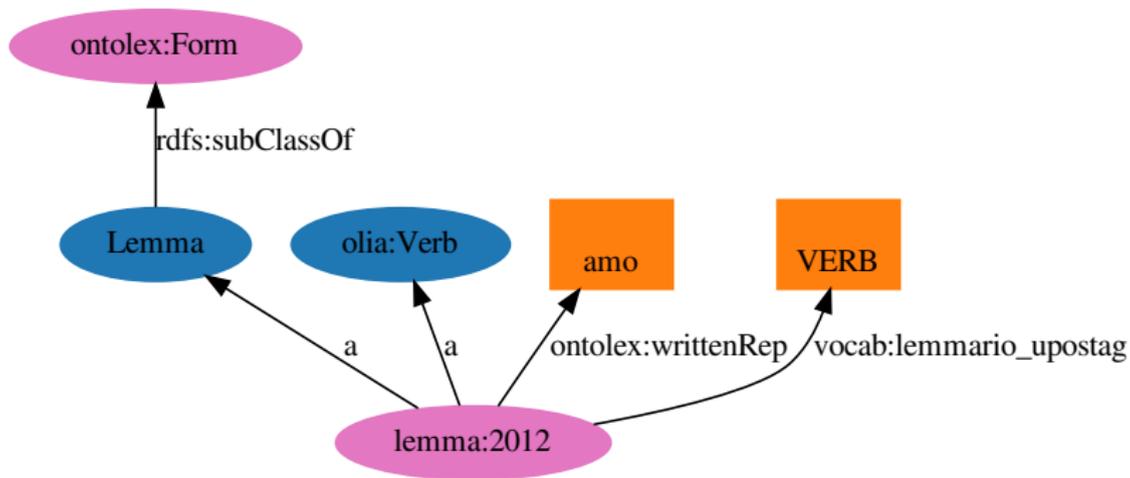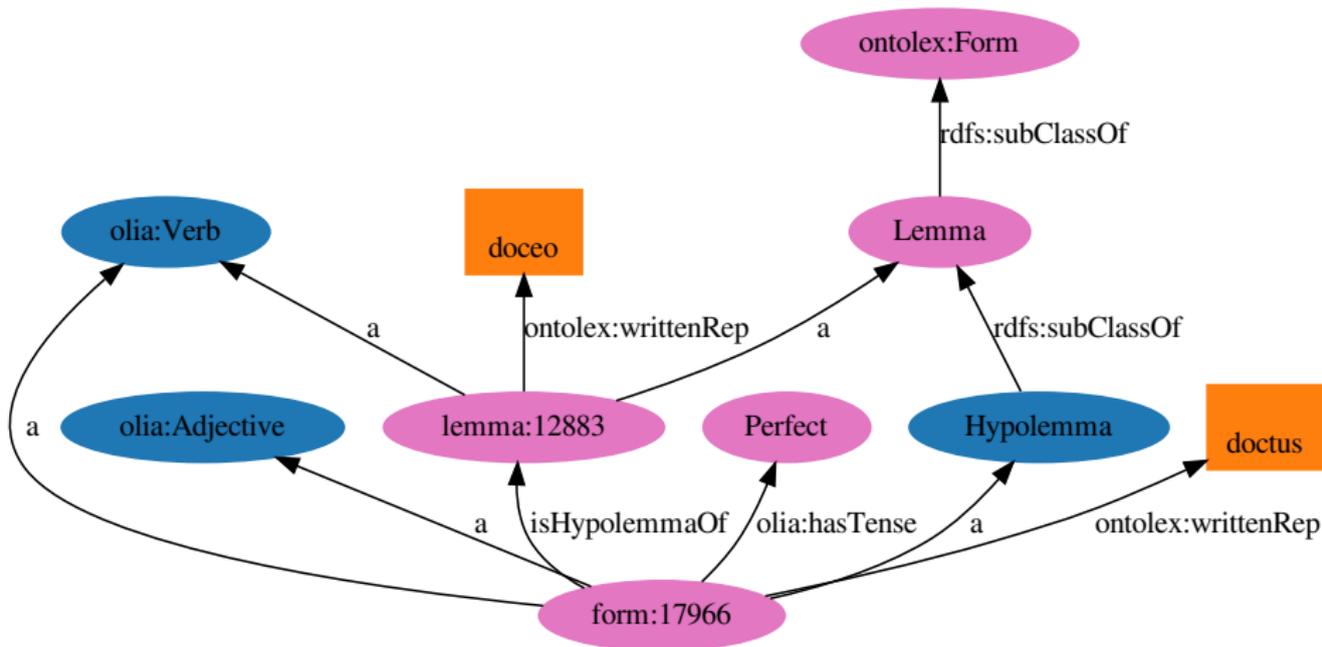
Corpora in LiLa
A token from PROIEL (*Rev.* 1.18)

- ▶ PROIEL (Universal Dependencies)
- ▶ *Index Thomisticus* Treebank (ITTB), both UD and original
- ▶ a portion of the Late Latin Charter Treebank (LLCT) (Timo Korkiakangas)

- ▶ PROIEL (Universal Dependencies)
- ▶ *Index Thomisticus* Treebank (ITTB), both UD and original
- ▶ a portion of the Late Latin Charter Treebank (LLCT) (Timo Korkiakangas)

Try it out!
`https://lila-erc.eu/data/`

1. Include metadata about authors, texts, editions. . .

1. Include metadata about authors, texts, editions. . .
   ▶ Include canonical references

1. Include metadata about authors, texts, editions...
   - Include canonical references
2. link to distributed content (texts are maintained by their providers)

1. Include metadata about authors, texts, editions…
   - ▶ Include canonical references
2. link to distributed content (texts are maintained by their providers)
3. more lemmatisation!

1. Include metadata about authors, texts, editions…
   - Include canonical references
2. link to distributed content (texts are maintained by their providers)
3. more lemmatisation!
   - improve the performance of lemmatisers (Flavio, Rachele)

Open challenges

1. Include metadata about authors, texts, editions...
   - ▶ Include canonical references
2. link to distributed content (texts are maintained by their providers)
3. more lemmatisation!
   - ▶ improve the performance of lemmatisers (Flavio, Rachele)
   - ▶ agree on an annotation scenario with the content managers

Table of Contents

WFL: Word formation-based lexicon for Classical Latin

- ▶ LEMLAT Base lexical basis
- ▶ Word Formation Rules (WFRs) are modelled as directed one-to-many input-output relations between lemmas
- ▶ Relationships between lemmas (nodes) of the same "word formation family" are represented as the edges in a **directed graph** with a hierarchical tree-like structure
- ▶ Compounding is also shown as an intersection between word formation families
- ▶ Can be browsed by WFR, Affix, PoS and Lemma
- ▶ 763 WFRs, 32,428 input-output relations.

WFL: tree-shaped directed graph
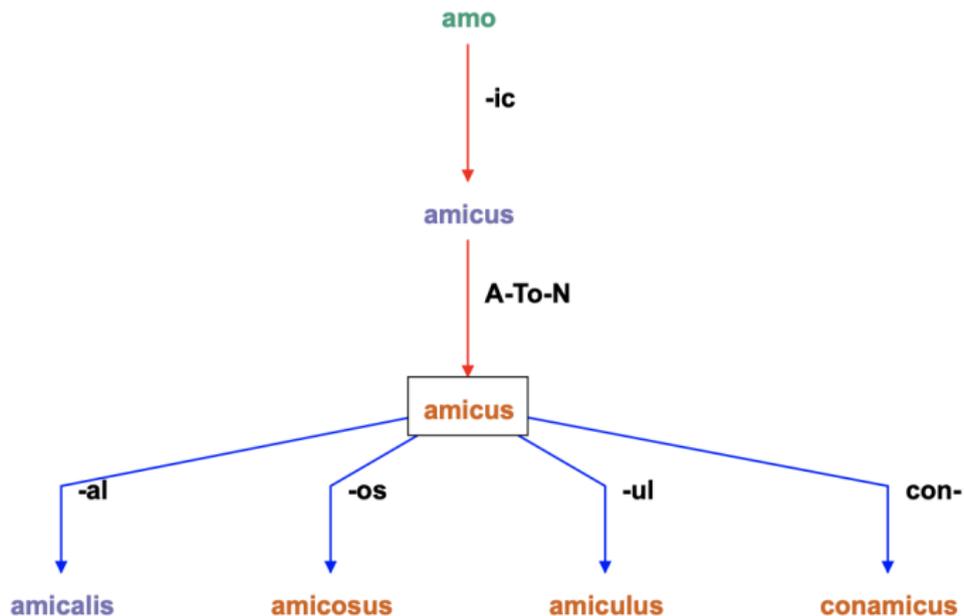
But: **directed graphs** are not completely satisfactory in representing the full range of relationships included within a word formation family.

Main problems:

- ▶ Directionality
- ▶ Non-linear derivations.
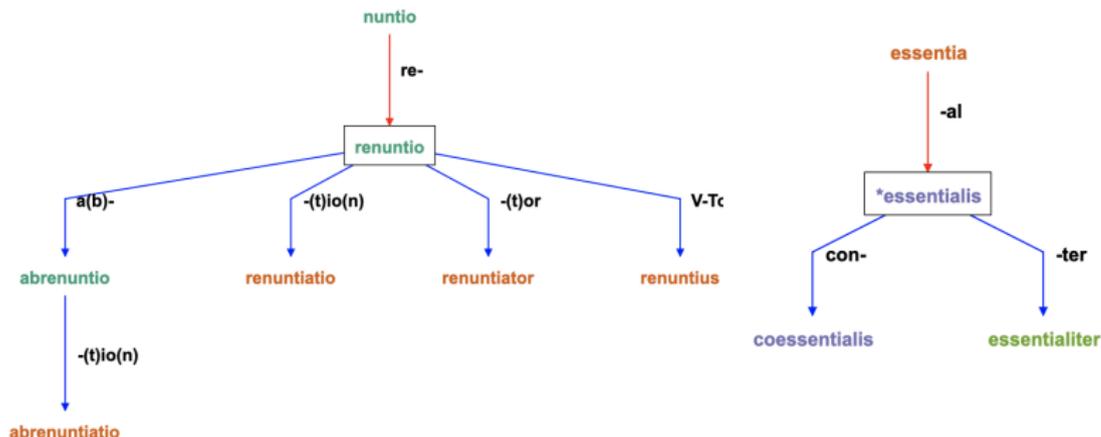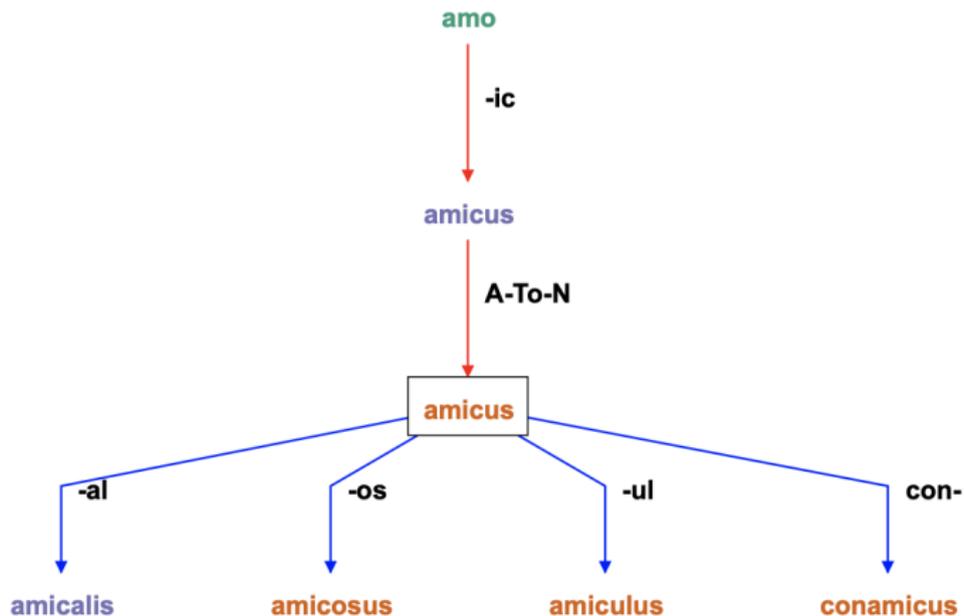
New approach to Word Formation:

- ▶ Structure: **declarative** rather than procedural
- ▶ No directionality
- ▶ No morphotaxis.

Words are described in their formative elements => these are organised in classes of objects in the ontology.

Three classes of objects:

1. Lemmas
2. Affixes (prefixes and suffixes)
3. Bases (connectors between lemmas of the same WF family)

Connected by three possibile relationships:

1. hasPrefix
2. hasSuffix
3. hasBase
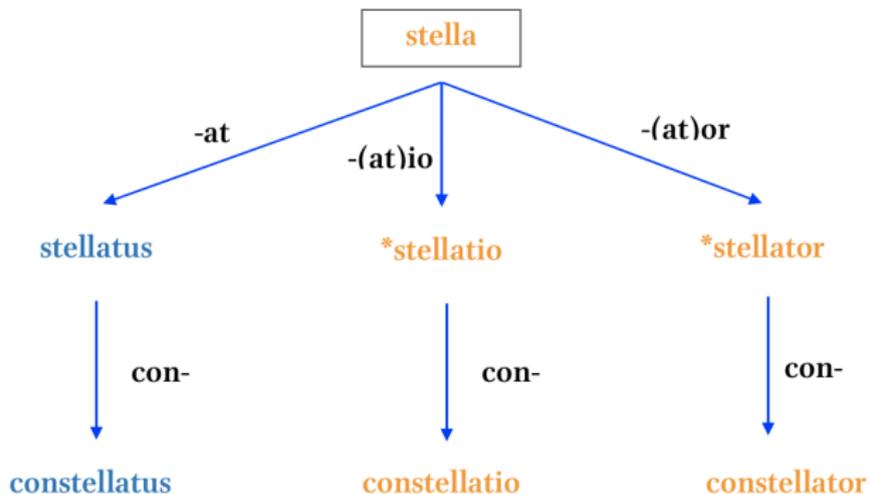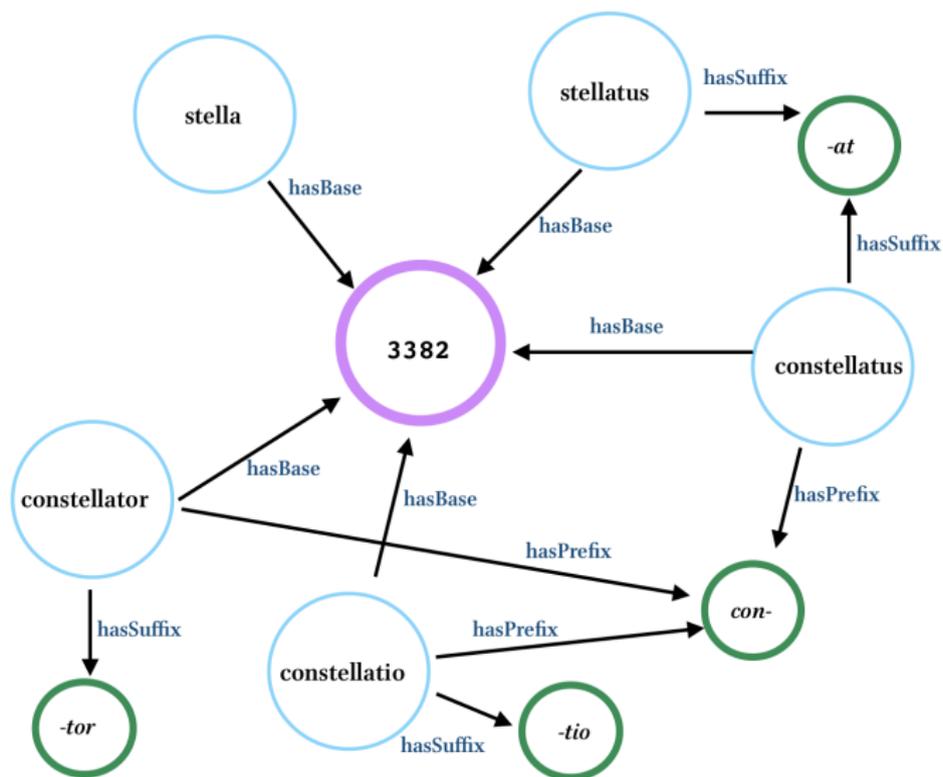
Latin Vallex: Valency Lexicon for Classical Latin

- ► Built in conjunction with the semantic and pragmatic annotation of two Latin treebanks:
  - ► The Index Thomisticus Treebank (Thomas Aquinas),
  - ► The Latin Dependency Treebank (Classical era).
- ► Structure inspired by the Valency Lexicon for Czech *PDT- Vallex*.

termino – V

- Word entries => sequence of frame entries for each lemma.
- Each frame entry => one sense.
- Each frame entry => description of the valency frame + frame attributes.
- Valency frame: sequence of frame slots.
- Frame slot: one complementation of the given lemma.
- Attributes: semantic roles ('functors') used to express types of relations between lemmas and their complementations.

- Frame Entry 1 ('to mark the boundaries of something'):
  - Valency Frame:
    - Frame Slot 1: subj.
    - Frame Slot 2: direct obj.
  - Frame Attributes:
    - Functor 1: ACT
    - Functor 2: PAT
- Frame Entry 2 ('to limit something to something else'):
  - Valency Frame:
    - Frame Slot 1: subj.
    - Frame Slot 2: dir. obj.
    - Frame slot 3: *in*+ dir. obj.
  - Frame Attributes:
    - Functor 1: ACT
    - Functor 2: PAT
    - Functor 3: DIR3

Valency Lexicon
First Steps in LiLa

- ▶ From evidence to intuition-based
- ▶ Cross reference Whitaker's Words definitions with *EngVallex* valency frames (English Valency Lexicon developed at Úfal)
- ▶ Evaluation and Validation (work in progress)
- ▶ Addition of new data.

Table of Contents

WordNet [...] is perhaps **the most widely used electronic dictionary** [...] and serves as the **lexicon for a variety of different NLP applications** including Information Retrieval (IR), Word Sense Disambiguation (WSD), and Machine Translation (MT).

Fellbaum (1998, p. 52)

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition
a#00430275	ENG	cloudy	full of or covered with clouds

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition
a#00430275	ENG	cloudy	full of or covered with clouds
a#00430275	ITA	annuvolato nuvolo nuvoloso	

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition
a#00430275	ENG	cloudy	full of or covered with clouds
a#00430275	ITA	annuvolato nuvolo nuvoloso	
a#00430275	LAT	nubilosus nubilus	

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition
a#00430275	ENG	cloudy	full of or covered with clouds
a#00430275	ITA	annuvolato nuvolo nuvoloso	
a#00430275	LAT	nubilosus nubilus	

Relations between synsets
Hypernymy/hyponymy, meronymy/holonymy, antonymy, entailment, etc.

A **database of synsets** (**sets** of **syn**onymous lemmas)

Synset ID	Lang	Lemma(s)	Definition
a#00430275	ENG	cloudy	full of or covered with clouds
a#00430275	ITA	annuvolato nuvolo nuvoloso	
a#00430275	LAT	nubilosus nubilus	

Relations between synsets
Hypernymy/hyponymy, meronymy/holonymy, antonymy, entailment, etc.

Only **two historical language WordNets**.

- ▶ **Who**: Stefano Minozzi, University of Verona
- ▶ **When**: 2004
- ▶ **How**: **generated** from the MultiWordNet[1]
- ▶ **What**: **limited** coverage
  - ▶ **9,378 lemmas**
  - ▶ **8,973 synsets**
  - ▶ **143,701 relations**
- ▶ **How well**: quite noisy

*La copertura lessicale e **i risultati dell'assegnazione automatica necessiterebbero di una ulteriore fase di valutazione e di controllo***.

Minozzi (2017, p. 130)

---

[1]http://multiwordnet.fbk.eu/english/home.php

1. **Phase 1**: **evaluate** existing LWN data

1. **Phase 1**: **evaluate** existing LWN data
   - ▶ Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.

1. **Phase 1**: **evaluate** existing LWN data
   - ► Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.
   - ► **Test evaluation**: **5 raters independently evaluate the same set of 100 lemmas** (25 per PoS) using a custom app; synsets to evaluate include both LWN data and computed suggestions.[2]

[2]Andrea Peverelli, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri.

The LiLa Team | Università Cattolica del Sacro Cuore, CIRCSE

1. **Phase 1**: **evaluate** existing LWN data
   - Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.
   - **Test evaluation**: **5 raters independently evaluate the same set of 100 lemmas** (25 per PoS) using a custom app; synsets to evaluate include both LWN data and computed suggestions.[2]
   - Calculate the **inter-rater agreement** and the **quality of the evaluations** against a Gold Standard.

—————————————————
[2]Andrea Peverelli, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri.

1. **Phase 1**: **evaluate** existing LWN data
   - ► Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.
   - ► **Test evaluation**: **5 raters independently evaluate the same set of 100 lemmas** (25 per PoS) using a custom app; synsets to evaluate include both LWN data and computed suggestions.[2]
   - ► Calculate the **inter-rater agreement** and the **quality of the evaluations** against a Gold Standard.
   - ► **Compare** the **computed** assignments **against manual** evaluation.

[2]Andrea Peverelli, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri.

1. **Phase 1**: **evaluate** existing LWN data
   - ► Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.
   - ► **Test evaluation**: **5 raters independently evaluate the same set of 100 lemmas** (25 per PoS) using a custom app; synsets to evaluate include both LWN data and computed suggestions.[2]
   - ► Calculate the **inter-rater agreement** and the **quality of the evaluations** against a Gold Standard.
   - ► **Compare** the **computed** assignments **against manual** evaluation.
   - ► **Further automate** where possible, e.g. **remove obvious noise.**

1. **Phase 1**: **evaluate** existing LWN data
   - ► Custom **algorithm** checks Latin resources (Whitaker's Words and Lewis & Short) against MultiWordNet to **propose missing senses**.
   - ► **Test evaluation**: **5 raters independently evaluate the same set of 100 lemmas** (25 per PoS) using a custom app; synsets to evaluate include both LWN data and computed suggestions.[2]
   - ► Calculate the **inter-rater agreement** and the **quality of the evaluations** against a Gold Standard.
   - ► **Compare** the **computed** assignments **against manual** evaluation.
   - ► **Further automate** where possible, e.g. **remove obvious noise.**
2. **Phase 2**: **data-driven enrichment** of the LWN by attaching it to textual tokens in LiLa (effectively performing Word Sense Disambiguation).

---

[2]Andrea Peverelli, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri.

Examples of noise to be removed:

Lemma	Synset	Definition
ager	n#W0021124	in un database, ogni area in cui vengono registrate le singole informazioni che compongono il record (ad esempio nomi, numeri ecc.).
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; Hawaii is calling!; A transmitter in Hawaii was heard calling.

E.g. *velociter*

	S1	S2	S3	S4
Rater 1	1	1	1	O
Rater 2	1	1	1	1
Rater 3	1	1	1	O
Rater 4	1	1	1	1
Rater 5	1	O	O	O

S1 = r#00051957
S2 = r#00082992
S3 = r#00102338
S4 = r#00285860

We measure:

▶ Inter-rater reliability:[3] $A_o = \frac{abs(N_C - N_R)}{N_V}$

  ▶ A_o = **o**bserved **a**greement
  ▶ N_C = n. of **C**onfirmed assignments
  ▶ N_R = n. of **R**ejected assignments
  ▶ N_V = n. e**v**aluations

▶ Quality: correctness against a Gold Standard

---

[3]Percentage of agreement without chance correction.

Figure: Graph rendition of the LWN lemma *abalieno*.

Table of Contents

Lemmatisation and part-of-speech tagging are essential and necessary tasks

- ► for the linguistical analysis of **Latin**…
  - ► rich morphology, ambiguity, …

- ► …and the inclusion of textual resources into **LiLa**!
  - ► the lemma as center stage of its architecture

Lack of annotated resources

Unfortunately, most Latin corpora are not provided with annotation at morphological, grammatical or syntactical level, and not even lemmatisation.

Our goal

To survey the existing tools for Latin lemmatisation and PoS-tagging

To automate annotation of resources and ease their inclusion into **LiLa**

LEMLAT is a powerful morphological analyser for Latin.

Morphological analysis entails lemmatisation.

```
=====================ANALYSIS 6=====================

SEGMENTATION:   aer -e

--------------------morphological feats   ------------------
--bms--

Case:   Ablative
Gender: Masculine
Number: Singular
                 ==================LEMMA =================
                 aer                 N3B  2948  m
                 --------------------morphological feats-----------
                 NpC

                 PoS:    Noun
                 Type:   Proper
                 Inflexional Category:   III decl
                 ---------------------derivational info------------
                 IS DERIVED: NO

=====================ANALYSIS 7=====================
```

aere

. . . Aere (f, PROPN)?

. . . Aer (m, PROPN)?

. . . aer (m/f, NOUN)?

. . . aerus (ADJ)?

. . . aes (n, NOUN)?

However, it can not disambiguate according to context!

> ### Part of speech ⟿ Lemma

We have selected and collected many tools and models for Latin:

CLTK:	TnT, CRF, 1-2-3-gram backoff, all trained on Perseus
Collatinus:	LASLA
Deucalion	LASLA
LaPOS:	Perseus, IT-TB UD 2.3
NLP-Cube:	UD 2.3 Latin treebanks
NLTK:	TnT, CRF, 1-2-3-gram backoff, all trained on IT-TB UD 2.3
MarMot:	Capitula+PROIEL(+Patr. Lat.+Collex-LA) (Eger et al. 2016), IT-TB UD 2.3
RDRPOSTagger:	IT-TB UD 2.3, PROIEL UD 2.3, Perseus UD 2.3
RNNTagger:	IT-TB
TreeTagger:	IT-TB UD 2.3, IT-TB, OMNIA (Bon 2011), Brandolini
UDpipe:	IT-TB UD 2.3, PROIEL UD 2.3, Perseus UD 2.3

. . . and also the lemmatiser **LatMor** (acontextual), based on the Berlin Latin Lexicon.

> ### We primarily focus on existing models rather than training new ones.

Each corpus uses different standards ⇒ Different PoS tagger annotations

perennius 'more lastingly'

- ► ADV - *perennius*
- ► ADV - *perenniter*
- ► ADJ - *perennis*

sanctus 'holy; saint'

- ► ADJ - *sanctus*
- ► NOUN - *sanctus*
- ► VERB - *sancio*

Each annotation standard has its own motivation!
Diachronic changes also have to be taken into account.

We want to be able to compare automated or manual annotations of parts of speech and lemmas wich follow different standards.

LEMLAT as a lexical hub

We exploit its vast coverage of lexicon and orthographical variants to correctly evaluate all possibilities.

affrementissime 'in a most roaring way'

adfrementissime/affrementissime	ADV/D/...
adfrementissimus/affrementissimus	ADJ/A/QLF/...
adfremens/affremens	VERB/V/VBE/... or ADJ/...
adfremo/affremo	VERB/...

will all be accepted as correct analyses!

We adopt the Universal POS Tags of UD (Petrov et al. 2011) as reference

https://universaldependencies.org/u/pos/index.html

De Divinatione by Cicero, 1st c. BC (Gold: LiLa)

PoS:	TreeTagger (Brandolini)	90.7%
	MarMot (Capitula)	88.7%
	UDpipe (PROIEL)	87.1%
Lemmas:	UDpipe (PROIEL)	90.3%
	TreeTagger (Brandolini)	89.9%
	MarMot (Capitula)	89.8%

Confessiones I-III by Augustinus, 4th c. AD (Gold: LiLa)

PoS:	TreeTagger (Brandolini)	93.6%
	MarMot (Capitula)	92.2%
	RDRPOSTagger (PROIEL)	91.6%
Lemmas:	TreeTagger (Brandolini)	95.0%
	MarMot (Capitula)	92.4%
	UDpipe (PROIEL)	92.3%

Hist. Langobardorum Beneventanorum by Erchempertus, 9th c. AD (Gold: Comp. Hist Sem.)

PoS:	MarMot (Capitula)	89.3%
	TreeTagger (Brandolini)	87.7%
	CLTK - CRF	83.9%
Lemmas:	MarMot (Capitula)	85.4%
	UDpipe (PROIEL)	79.6%
	TreeTagger (Brandolini)	79.6%

- ▶ Wide diachronic coverage seems to be more important than sheer size for training

- ▶ Diachronic variations seem to affect lemmatisation more than part-of-speech tagging

Future directions

- ▶ Fine-tuned harmonised evaluation, e. g.
  - ▶ diachronic point of view
  - ▶ evaluation per part of speech

- ▶ Training and evaluation of new models

- ▶ Survey on existing annotation standards and comparisons

- ▶ Automated conversion of annotation standards to UD

Table of Contents

► Lexical resources

► NLP Tools

► Word Embeddings

► Annotated corpora

- ► Valency Lexicon
- ► Latin WordNet
- ► de Vaan, M. (2008). **Etymological Dictionary of Latin**. Leiden, The Netherlands: Brill.

stēlla 'star' [f. *ā*] (Pl.+)

Derivatives: *stēllāns* 'starry' (Lucr.+), *stēllumicāns* 'shining with stars' (Varro), *stēl(l)iō* 'kind of lizard, gecko' (Verg.+).

Plt. *stērlā-.

PIE *h₂stēr-s, *h₂ster- [m.] 'star'. IE cognates: OIr. *ser*, W. *seren* (pl. *ser*), Co. *sterenn* (pl. *steyr*), Bret. *sterenn* 'star' < PCl. *ster-; OIr. *sell* [m.] 'iris' < *stillo-, MIr. *sellaid* 'to see', MW *syllu* 'to gaze', Bret. *sellout* 'to watch' < PCl. *stīrlo-* 'star' < *h₂stēr-lo-; Hit. *hašter(a)-* [c.] 'star' (nom.sg. *hašterza* /Hsterť/) < *h₂stēr-; Skt. *tắraḥ* [nom.pl.], *stŕ̥bhiḥ* [instr.pl.], Av. *star-* [m.] (YAv. acc.sg. *stārəm*, nom.pl. *stārō*, acc.pl. *strāuš* for *strāš*, gen.pl. *strm̥*), Khot. *stāraa-* 'star'; Gr. ἀστήρ, -έρος, Arm. *astl*, Go. *stairno*, ToB *ścirye̦* 'star'.

Schrijver 1995 has established that OIr. *sell* 'iris' can go back to PCl. *stīrlo-*; together with Lat. *stēlla*, this points to an Italo-Celtic formation *stēr-lo/ā-* 'star'.

Bibl.: WH II: 587f., EM 646, IEW 1027f., Schrijver 1995: 421-423, Meiser 1998: 123.

Information about reconstructed Indo-European forms

Proto-Italic

Breton

Avestan

Proto-Celtic

Old-Irish

Welsh

Proto-Indo-European

Cornish

Middle Irish

Sanskrit

Armenian

Hittite

Gothic

ThocarianB

Middle Welsh

Khotanese

Young Avestan

Greek

stēlla 'star' [f. ā] (Pl.+)
  Derivatives: *stēllāns* 'starry' (Lucr.+), *stēllumicāns* 'shining with stars' (Varro), *stēl(l)iō* 'kind of lizard, gecko' (Verg.+).
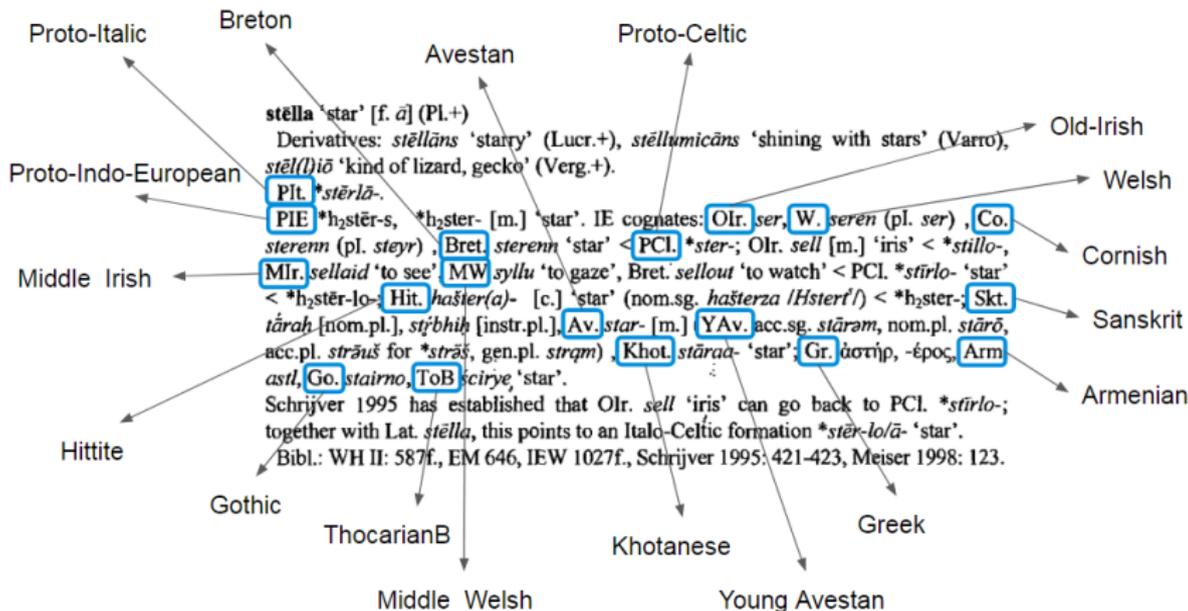
  Plt. *stērlā-.

  PIE *h₂stēr-s, *h₂ster- [m.] 'star'. IE cognates: OIr. *ser*, W. *serēn* (pl. *ser*) , Co. *sterenn* (pl. *steyr*) , Bret. *sterenn* 'star' < PCl. *ster-; OIr. *sell* [m.] 'iris' < *stillo-, MIr. *sellaid* 'to see', MW *syllu* 'to gaze', Bret. *sellout* 'to watch' < PCl. *stīrlo-* 'star' < *h₂stēr-lo- tắrah [nom.pl.], *stŕbhih* [instr.pl.], Hit. *haster(a)-* [c.] 'star' (nom.sg. *hašterza /Hstert'/*) < *h₂ster- tắrah [nom.pl.], *stŕbhih* [instr.pl.], Skt. *stārō*, acc.pl. *strū̆š* for *strắš*, gen.pl. *str̥m*), Khot. *stāraa-* 'star'; Gr. ἀστήρ, -έρος, Arm *astl*, Go. *stairno*, ToB *ścirye* 'star.'
  Schrijver 1995 has established that OIr. *sell* 'iris' can go back to PCl. *stīrlo-*; together with Lat. *stēlla*, this points to an Italo-Celtic formation *stēr-lo/ā-* 'star'.
  Bibl.: WH II: 587f., EM 646, IEW 1027f., Schrijver 1995: 421-423, Meiser 1998: 123.

Models trained on "Opera Latina", a corpus manually annotated by the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) for:

1. Tokenisation
2. PoS Tagging
3. Lemmatisation
4. Inflectional features identification

Models trained on "Opera Latina", a corpus manually annotated by the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) for:

1. Tokenisation
2. PoS Tagging
3. Lemmatisation
4. Inflectional features identification

Models trained on:

1. the whole corpus
2. texts by single authors (i.e. author-specific models)

Word embeddings

Pre-trained word vectors learned on the whole LASLA corpus using:

1. word2vec
2. fastText

▶ Different word representations:

FELIX	
word2vec	**fastText**
beatus	infelix
fortunatus	felicitas
inuideo	feliciter
felicitas	fel
infelix	infelicitas
infelicitas	fortunatus
miser	detestor
bonum	gaudeo

IUDICO	
word2vec	**fastText**
puto	abiudico
sum	diiudico
dico	adiudico
debeo	praeiudico
existimo	iudicatum
ergo	iudicium
sapiens	praeiudicium
delibero	dico

Ancient Latin texts taken from the Perseus Digital Library:

- different authors (Caesar, Seneca, Cicero, Catullus...)
- different genres (treatises, letters, poems...)
- automatically annotated with our new author-specific models

58

How can we promote the development of resources and language technologies for the Latin language?

How can we foster collaboration among scholars working on Latin and attract researchers from different disciplines?

How can we promote the development of resources and language technologies for the Latin language?

How can we foster collaboration among scholars working on Latin and attract researchers from different disciplines?
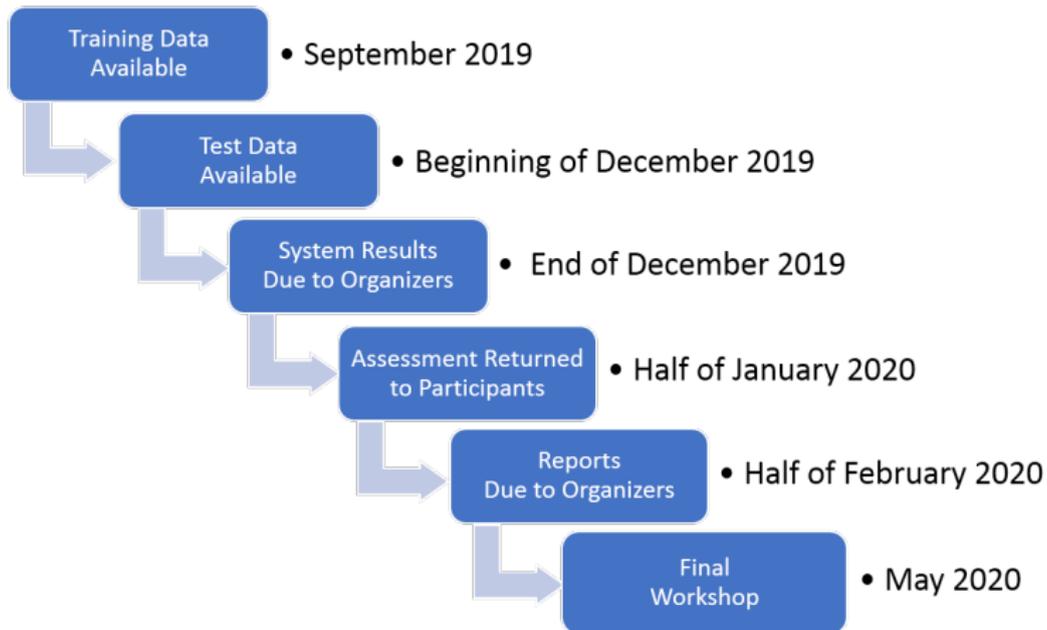
EVALATIN

- ▶ Evaluation campaign designed following a long tradition in NLP (MUC, ACE, SemEval, CoNLL…)
- ▶ Shared tasks, shared training and test data, shared evaluation metrics

EvaLatin

- ▶ Evaluation campaign designed following a long tradition in NLP (MUC, ACE, SemEval, CoNLL…)
- ▶ Shared tasks, shared training and test data, shared evaluation metrics
- ▶ 3 tasks:
  1. PoS tagging
  2. Lemmatisation
  3. Inflectional features identification
- ▶ 3 sub-tasks for each task:
  1. Basic
  2. Cross-Genre
  3. Cross-Time

EvaLatin
Tentative Timeline

Training Data Available • September 2019

Test Data Available • Beginning of December 2019

System Results Due to Organizers • End of December 2019

Assessment Returned to Participants • Half of January 2020

Reports Due to Organizers • Half of February 2020

Final Workshop • May 2020

The LiLa Team
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

▶ Bon, B. (2011) 'OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3)', Bulletin du centre d'études médiévales d'Auxerre (BUCEMA). URL: `http://journals.openedition.org/cem/12015`, DOI: 10.4000/cem.12015

▶ Eger, S., Gleim, R. and Mehler, A. (2016) 'Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art', *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

▶ Fellbaum, C. (1998) 'Towards a Representation of Idioms in WordNet', *Proceedings of the workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL)*, pp. 52–57.

▶ Minozzi, S. (2017) 'Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval', *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, 14, pp. 123–133. DOI: 10.14277/6969-182-9/ANT-14-10

▶ Petrov, S., Das, D. and McDonald, R. (2011) 'A Universal Part-of-Speech Tagset', *arXiv:1104.2086*.