LiLa: Linking Latin

Piecing together a seemingly under-resourced language

Greta Franzini and **Marco Passarotti**
greta.franzini,marco.passarotti@unicatt.it

UNIVERSITÀ CATTOLICA del Sacro Cuore

IRCSE

Computer-Assisted Text Analysis for Resource-Scarce Literatures
Miami, FL | 25 April 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

Resource-scarceness
What does it mean?

Computer-Assisted Text Analysis for Resource-Scarce Literatures

Computer-Assisted Text Analysis for Resource-Scarce Literatures

What does *resource-scarce* mean?

Resource-scarceness
Towards a definition

[...] ***when can one say that a language is properly covered from a resources point of view?*** *There is no broadly accepted definition of what counts as sufficient.* [...] *we intend to arrive at such a definition by introducing the concept of a BLARK: a Basic Language Resource Kit.* (Krauwer 2003, p. 3)

- **When**: 1998
- **Who**: Steven Krauwer in cooperation with ELSNET (European Network in Human Language Technologies) and ELRA (European Language Resources Association)
- **What**: *Minimal set of language resources that is necessary to do any precompetitive research and education* [...]. (Krauwer 2003, p. 4)
- **How**:
  - Resources: data-sets and electronic descriptions that are used to build, improve, or evaluate modules.
  - Modules: basic software components that are essential for developing Human Language Technologies (HLT) applications (e.g. morphological analysis).
  - Applications: classes of applications that make use of HLT (e.g. computer-assisted language learning).

To measure “resourceness" BLARK provides:

1. **Matrix** of minimal set of language components: **availability**
2. **Matrix** of minimal set of language components: **priority**

Distinction

Language vs. Speech Technologies → Historical vs. Modern languages.

> [...] *each language should try to make an inventory of which BLARK components are already available for their language, and which ones are missing*. (Krauwer 2003, p. 5)

Is Latin under-resourced?
The question

Latin a decade ago...

> *Latin can still be considered as a less-resourced language, lacking powerful NLP tools and a broad suite of state-of-the-art language resources* [...] *such as annotated corpora and lexica.* [...] *the less-resourced status affects historical languages in general (because of reasons such as being not commercially interesting or lacking native speakers)* [...]. (Passarotti 2010, p. 27)

Latin today?

Latin BLARK

Availability of language technologies for Latin as of 2019 (speech excluded)

Resources	Availability	Module	Availability
Unannotated corpora	10	Tokenisation	10
Annotated corpora	10	Sentence boundary detection	10
Multilingual corpora	10	Named Entity Recognition	10
Test corpora	10	Spelling correction	10
Monolingual lexicons	10	Lemmatisation	10
Multilingual lexicons	10	Morphological analysis	10
Thesaurus	10	Morphological synthesis	10
		PoS-tagging	10
		Parsers & grammars	10
		Constituency parsing	4
		Semantic analysis	4
		Referent resolution	4
		Word sense disambiguation	4
		Pragmatic analysis	10
		Language dependent translation	3

Table: Availabiity rating: 1 (unavailable) - 10 (Available, obtainable, reusable).

Is Latin under-resourced?
The answer

- In **BLARK terms**: NO

Is Latin under-resourced?
The answer

- In **BLARK terms**: NO
  - *Minimal set of language resources that is necessary to do any precompetitive research and education*

Is Latin under-resourced?
The answer

- In **BLARK terms**: NO
  - *Minimal set of language resources that is necessary to do any precompetitive research and education*
- Relative to **modern languages**: YES

Is Latin under-resourced?
The answer

- In **BLARK terms**: NO
  - *Minimal set of language resources that is necessary to do any precompetitive research and education*
- Relative to **modern languages**: YES
- Relative to other **historical languages**: NO

Latin resources
Lots available

- **Textual Resources**
- **Lexical Resources**
- **NLP Tools**

Latin resources
Lots available

- **Textual Resources**

  Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Latin Dependency Treebank, Index Thomisticus Treebank, PROIEL Latin Treebank, Late Latin Charter Treebank, ...

- **Lexical Resources**

- **NLP Tools**

Latin resources
Lots available

- **Textual Resources**

  Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Latin Dependency Treebank, Index Thomisticus Treebank, PROIEL Latin Treebank, Late Latin Charter Treebank, ...

- **Lexical Resources**

  Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange Glossarium Mediae et Infimae Latinitatis), Thesaurus Lingua Latinae, Oxford Latin Dictionary, Thesaurus Formarum Totius Latinitatis, Dictionary of Medieval Latin from British Sources, & DB of Latin Dictionaries c/o Brepols), Lexicon musicum Latinum medii aevi, ...

- **NLP Tools**

Latin resources
Lots available

- **Textual Resources**

  Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Latin Dependency Treebank, Index Thomisticus Treebank, PROIEL Latin Treebank, Late Latin Charter Treebank, ...

- **Lexical Resources**

  Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange Glossarium Mediae et Infimae Latinitatis), Thesaurus Lingua Latinae, Oxford Latin Dictionary, Thesaurus Formarum Totius Latinitatis, Dictionary of Medieval Latin from British Sources, & DB of Latin Dictionaries c/o Brepols), Lexicon musicum Latinum medii aevi, ...

- **NLP Tools**

  Morphological Analysers (LEMLAT, Whitaker's Words, LatMor), PoS-Taggers (TreeTagger, Collatinus, UDPipe), Dependency Parsers (UDPipe), Chiron Metrical Analysis, Classical Language Toolkit (CLTK), ...

Latin resources
Lots available

- **Textual Resources**

  Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Latin Dependency Treebank, Index Thomisticus Treebank, PROIEL Latin Treebank, Late Latin Charter Treebank, ...

- **Lexical Resources**

  Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange Glossarium Mediae et Infimae Latinitatis), Thesaurus Lingua Latinae, Oxford Latin Dictionary, Thesaurus Formarum Totius Latinitatis, Dictionary of Medieval Latin from British Sources, & DB of Latin Dictionaries c/o Brepols), Lexicon musicum Latinum medii aevi, ...

- **NLP Tools**

  Morphological Analysers (LEMLAT, Whitaker's Words, LatMor), PoS-Taggers (TreeTagger, Collatinus, UDPipe), Dependency Parsers (UDPipe), Chiron Metrical Analysis, Classical Language Toolkit (CLTK), ...

Why is Latin still considered under-resourced?

Latin resources
Lots available

- **Textual Resources**

  Perseus Digital Library, Eurasian Latin Archive, Corpus Grammaticorum Latinorum, Croatiae auctores Latini, Archivio della Latinità Italiana del Medioevo, Musisque Deoque, Patrologia Latina, PHI Classical Latin Texts, Latin Dependency Treebank, Index Thomisticus Treebank, PROIEL Latin Treebank, Late Latin Charter Treebank, ...

- **Lexical Resources**

  Valency Lexica (Vallex, IT-VaLex), Latin WordNet, Dictionaries (Du Cange Glossarium Mediae et Infimae Latinitatis), Thesaurus Lingua Latinae, Oxford Latin Dictionary, Thesaurus Formarum Totius Latinitatis, Dictionary of Medieval Latin from British Sources, & DB of Latin Dictionaries c/o Brepols), Lexicon musicum Latinum medii aevi, ...

- **NLP Tools**

  Morphological Analysers (LEMLAT, Whitaker's Words, LatMor), PoS-Taggers (TreeTagger, Collatinus, UDPipe), Dependency Parsers (UDPipe), Chiron Metrical Analysis, Classical Language Toolkit (CLTK), ...

Why is Latin still considered under-resourced?

Scattered, not visible, not interoperable

LiLa: Linking Latin

Towards interoperability

LiLa: Linking Latin
Towards interoperability

- **Objective**: Knowledge Base of Linguistic Resources & NLP Tools
- **Method**: RDF, Linked Data paradigm (FAIR principles)[1]
- **Purpose**: integration of lexica/resources with ontologies for a deeper study of semantics (train NLP applications to interpret NL with respect to ontologies)
- **Funding**: ERC Consolidator Grant, 2M EUR
- **Duration**: 2018-2023
- **Team**: 14; 15 from 1st June 2019
- **Website**: https://lila-erc.eu

- **Work Package 1 (months 1-24)**: protytping & testing phase

[1] Wilkinson (2016).

LiLa: Linking Latin

Conceptual model

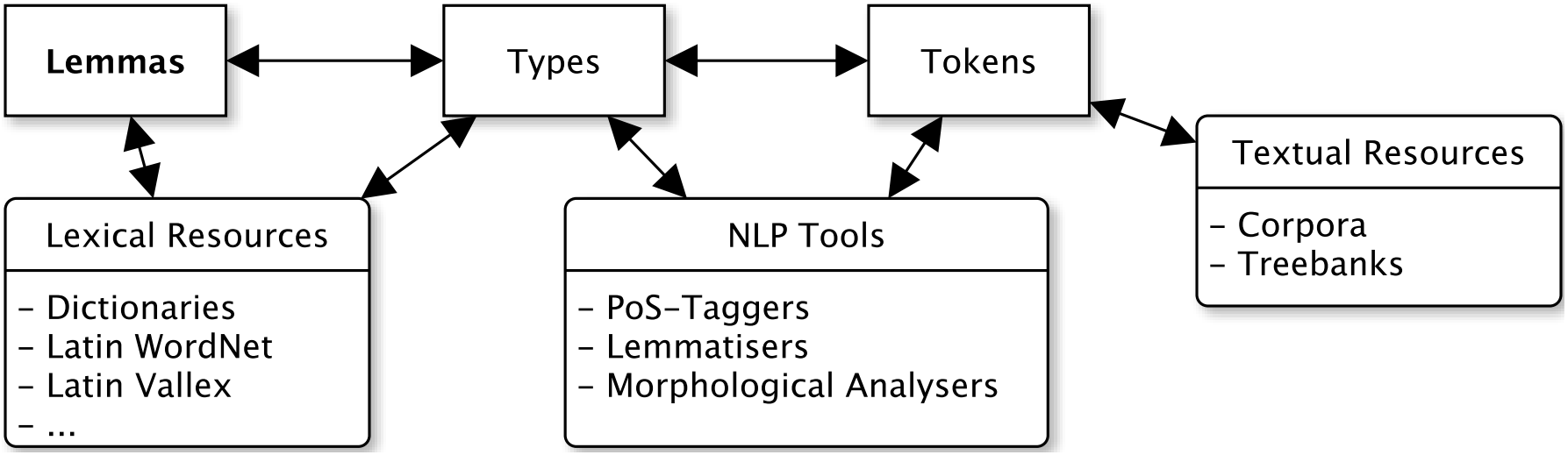


Figure: Simplified conceptual model of LiLa.

Data stays with the provider.

LiLa's **backbone is an RDF ontology**[2] made of:

- Individuals: instances of objects (specific token, lemma, etc.).
- Classes: types of objects/concepts (token, lemma, PoS, etc.).
- Attributes: properties that objects can/must have (morphological features for lemmas/tokens). An attribute can be a class or an individual.
- Relations: ways in which classes and individuals can be related to one another: triples (`SUBJ – PREDICATE – OBJ`). Labels from a dictionary of knowledge description: *has_lemma*, *has_PoS*[3], etc.

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

- Exploring possibilities, e.g. CTS URN notation.[4]

[2] Lemon: https://lemon-model.net/; OLiA: http://nachhalt.sfb632.uni-potsdam.de/owl/
[3] *Universal PoS tags*.
[4] Canonical Text Services: http://www.homermultitext.org/hmt-doc/cite/index.html

Lemma source: *LEMLAT* morphological analyser & lemmatiser[5]

- Added: 43,432 lemmas (Classical Latin)
- To be added: 112,249 lemmas from Du Cange and Forcellini's *Onomasticon*
- different written representations induced from the annotated corpora connected (data-driven approach)
- enhanced lemmas with information on derivational morphology taken from the *Word Formation Latin Lexicon*[6]

[5] Passarotti et al. (2017).
[6] Litta (2018).

LiLa triplestore available at:
https://lila-erc.eu/data/

Resources connected so far:

- Lexical basis of LEMLAT
- Word Formation Latin Lexicon (Classical Latin)
- PROIEL Latin Treebank (Universal Dependencies v. 2.3)
- Index Thomisticus Treebank (original PDT format)[7]
- Index Thomisticus Treebank (Universal Dependencies v. 2.3)

[7] Passarotti (2015).

Resources in the process of being added to the triplestore:

- Ovid[8]
- Texts from the Perseus Digital Library
- Latin WordNet[9]

First LiLa Workshop:

- **Dates**: 3rd-4th June, Milan
- **Objective**: 20+ scholars invited to discuss integration of resources & NLP tools in LiLa
- **Registration**: closes 30th April (70+ people)
- **Website**: https://lila-erc.eu/events

[8] Potential POSTDATA ERC collaboration: `http://postdata.linhd.uned.es/`
[9] Minozzi (2017).

Who does what in LiLa:

- M. Passarotti: Principal Investigator
- E. Litta: Assessing & extending **Latin Vallex** + **Word Formation Latin**
- G. Franzini: Assessing & extending the **Latin WordNet**
- M. Testori: **Gold stndrs** & extending the ***Index Thomisticus*** **Treebank**
- F. Mambrini & R. Sprugnoli: Populating/Testing the **Knowledge Base**
- F. Cecchini: **PoS-tagging** & **harmonising PoS/morph. tagsets**
- P. Ruffolo: **LEMLAT** & **RDF triplestore management**
- New recruit: **Graphical interface** to query LiLa

Thanks!
Get in touch

Greta Franzini and **Marco Passarotti**
Università Cattolica del Sacro Cuore

{greta.franzini,marco.passarotti}@unicatt.it

@ERC_LiLa

https://github.com/CIRCSE

https://lila-erc.eu

Largo Gemelli 1, 20123 Milan, Italy

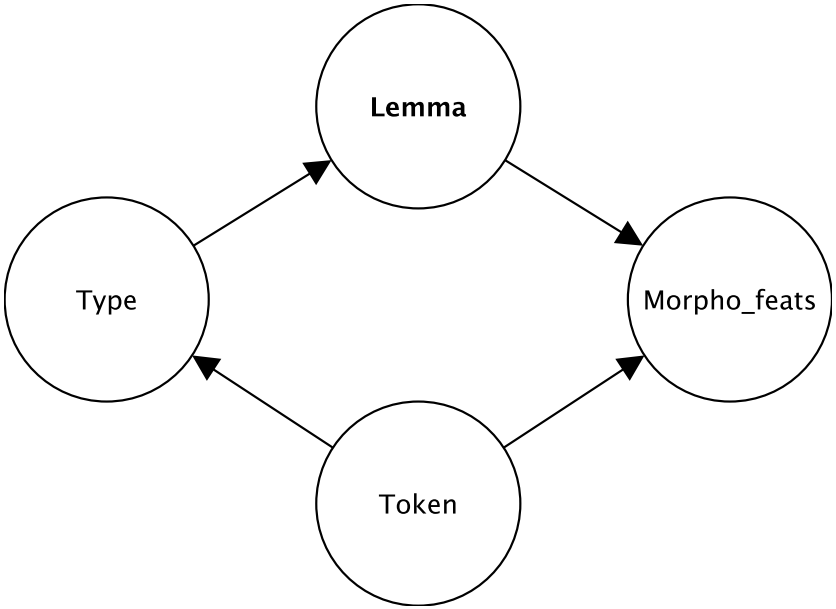
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No 769994

Works cited

- Krauwer, S. (2003) *The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap*. The University of Utrecht.
- Litta, E. (2018) 'Morphology Beyond Inflection. Building a Word Formation Based Lexicon for Latin', in Cotticelli-Kurras, P., Giusfredi, F. (eds) *Formal Representation and the Digital Humanities*. Cambridge Scholars Publishing, pp. 97-114.
- 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
- Passarotti, M., Budassi, M., Litta, E., Ruffolo, P. (2017) 'The Lemlat 3.0 Package for Morphological Analysis of Latin', in Bouma, G., Adesam, Y. (eds) *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Northern European Association for Language Technology (NEALT) Proceedings Series*, Vol. 32, pp. 24-31.
- Passarotti, M. (2015) 'What you can do with linguistically annotated data. From the *Index Thomisticus* to the *Index Thomisticus* Treebank', in Roszak, P., Vijgen, J. (eds) *Reading Sacred Scripture with Thomas Aquinas. Hermeneutical Tools, Theological Questions and New Perspectives*. Brepols, pp. 3-44.
- Passarotti, M. (2010) 'Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the *Index Thomisticus* Treebank', *Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May*, pp. 20–27.
- Wilkinson, M.D. et al. (2016) 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 3.

LiLa: Linking Latin

Terminology: Lemma vs. Type vs. Token



- **Lemma**: lexical entry
  e.g. *puella*
- **Type**: form of the lemma
  e.g. *puellam*
- **Token**: occurrence of the type
  e.g. *puellam*
- **Morphological features**
  e.g. *noun*, *feminine*, *singular*, ...

Issue: how do we lemmatise these?

- Orthographical variation (*voluptas* vs. *uoluptas*)
  - conversion of *V*s to *U*s (preprocessing)
- Orthographical variation (*sulphur* vs. *sulfur* vs. *sulpur*)
  - different `written representations` of the same lemma
- Orthographical variation (*diameter* vs. *diametros* vs. *diametrus*)
  - different `written representations` of the same lemma
- Participles (*ductus*)
  - hypolemmas (subclass of lemma) connected with their main verbal lemma (*duco*) through a subclass of the `form variant` property
- De-adjectival adverbs (*aequaliter*)
  - standalone lemmas
- Homography (*occīdo* vs. *occĭdo*)
  - if this is not disambiguated in the source texts, LiLa provides both options