Morphological resources in the LiLa Knowledge Base

An overview of past, present and future work

Matteo Pellegrini
matteo.pellegrini@unicatt.it

Morphology Reading Group
Université de Paris-CNRS, Laboratoire de Linguistique Formelle
December 17, 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.
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
Outline

Introduction

Linked Data Principles
LiLa: Linking Latin

Word formation
Past
Word Formation Latin (WFL)
Word Formation in the Lemma Bank
Including WFL into the LiLa Knowledge Base

Present
Adding Medieval Latin lemmas to WFL

Future
Building the Latin derivational paradigm

Inflection
Information on inflection in the Lemma Bank
An ontology for Latin principal parts
Discussion and next steps
Nowadays, a lot of linguistic **resources** and NLP **tools** are available for many languages

However, they are characterized by different **conceptual** and **structural models** → interoperability and secondary reuse are difficult to achieve

Be **FAIR**! Make your data **Findable**, **Accessible**, **Interoperable** and **Reusable** (Wilkinson et al. 2016)

Tim Berners-Lee’s principles of **Linked Data**
  - Use **Uniform Resource Identifiers** (URIs)
  - Use the **HTTP protocol** to allow people (and machines) to access URIs
  - Use **web standards** to represent/query (meta)data
  - Include **links** to other URIs

→ **Linguistic Linked Open Data** cloud (Cimiano et al. 2020)
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
Open-ended **Knowledge Base** of interoperable linguistic resources for Latin sharing a common vocabulary for knowledge description

- Use of **web standards** to represent and query data
  - RDF: information is coded in terms of **triples**, connecting a **subject** to an **object** through a **property**
  - SPARQL to query RDF data

- Reuse of **existing ontologies**
  - OLiA (linguistic annotation)
  - NIF, CoNLL-RDF (corpus annotation)
  - OntoLex-Lemon (lexical resources)

- The backbone of the LiLa Knowledge Base is the **Lemma Bank**, a collection of canonical forms (i.e. citation forms) of Latin words
Architecture of the LiLa Knowledge Base

Lemmas

Lexical Entries

Tokens

NLP Output

Lexical Resources
- Latin Wordnet
- Valency Lexicon
- Dictionaries

Textual Resources
- Digital libraries
- Treebanks
- Textual corpora

NLP Tools
- Tokenizers
- Taggers/parsers
- Lemmatizers

Lexical Resources

Textual Resources

NLP Tools
RDF data model

Nunc est bibendum
(Horace, Odes, 1, 37)
Architecture of the OntoLex model
LiLa: connected resources and upcoming connections

Corpora
- Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
- Dante Search (700th death anniversary): ca. 46,000 tokens
- *Querolus sive Aulularia*: ca. 17,000 tokens
- PROIEL and LLCT treebanks
- Computational Historical Semantics, LASLA and CroALa corpora

Lexica
- Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
- Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
- LatinAffectus: ca. 2,300 entries
- Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
- Latin WordNet: ca. 1,000 manually checked entries
- Latin Vallex 2.0: Valency Lexicon
- Lewis & Short Dictionary

NLP tools
- LEMLAT (lemma bank): ca. 150,000 lemmas

TOTAL: approximately 13.5 million triples
Outline

Introduction
- Linked Data Principles
- LiLa: Linking Latin

Word formation
- Past
  - Word Formation Latin (WFL)
  - Word Formation in the Lemma Bank
  - Including WFL into the LiLa Knowledge Base

- Present
  - Adding Medieval Latin lemmas to WFL

- Future
  - Building the Latin derivational paradigm

Inflection
- Information on inflection in the Lemma Bank
- An ontology for Latin principal parts
- Discussion and next steps
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
Derivational lexicon of Latin characterised by a step-to-step morphotactic approach: lexemes that are directly derived from one another are connected via word formation rules (WFRs)

▶ **Derivation** (prefixation / suffixation / conversion) vs. **compounding** rules

▶ Classification based on the **lexical category** of the input and output lexeme

<table>
<thead>
<tr>
<th>input lexeme(s) (PoS)</th>
<th>output lexeme (PoS)</th>
<th>prefix</th>
<th>suffix</th>
<th>WFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FELIX ‘happy’ (A)</td>
<td>INFELIX ‘unhappy’ (A)</td>
<td>in-</td>
<td>-</td>
<td>A-to-A in-</td>
</tr>
<tr>
<td>FELIX ‘happy’ (A)</td>
<td>FELICITAS ‘happiness’ (N)</td>
<td>-</td>
<td>-tas</td>
<td>A-to-N -tas</td>
</tr>
<tr>
<td>MALUS ‘bad’ (A)</td>
<td>MALUM ‘bad thing’ (N)</td>
<td>-</td>
<td>-</td>
<td>A-to-N</td>
</tr>
<tr>
<td>AGER ‘field’ (N); COLO ‘to cultivate’ (V)</td>
<td>AGRICOLA ‘farmer’ (N)</td>
<td>-</td>
<td>-</td>
<td>N+V=N</td>
</tr>
</tbody>
</table>

→ **Lexeme-oriented** perspective (Kyjánek 2020)

Eleonora Litta and Marco Passarotti

*(When) inflection needs derivation: a word formation lexicon for Latin*  
*Words and Sounds*, Volume I, 2020
The hierarchical structure of WFL

Hierarchical structure, representable with a directed tree-graph
Troubles with the rules

- Devised according to an Item-and-Arrangement model of morphology
- Some processes are not easy to fit into this rigidly hierarchical structure (cf. Budassi and Litta 2017)
  - Direction of conversion:
    ADVERSARIUS\textsubscript{A} ‘opposed’ $\rightarrow$ ADVERSARIUS\textsubscript{N} ‘opponent’ or vice versa?
    $\rightarrow$ A decision has to be made even in cases that are not clear-cut
- Parasynthetic formations:
  AQUA ‘water’ $\rightarrow$ *AQUESCO (?) $\rightarrow$ EXAQUESCO ‘become water’
  $\rightarrow$ Even if they are not attested, intermediate steps need to be added
  (“fictional lemmas”)
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
The Lemma Bank provides a limited amount of derivational information, taking it from WFL.

Each lemma is connected:
- to the **affixes** (prefixes/suffixes) it displays;
- to its **base** – an abstract connector between lemmas that belong to the same family.

→ **Family-oriented** perspective (Kyjánek 2020)

Eleonora Litta, Marco Passarotti, Marco Budassi, Marco Pappalepore

*Of nodes and cells. Two perspectives on (and from) Word Formation Latin*

*Lingue antiche e moderne*, 9, 2020
The flat structure of word formation the Lemma Bank
Discussion

- Compatible with more recent Word-and-Paradigm theoretical approaches
  - Construction Morphology (Booij 2010): output-oriented, declarative schemes
- More natural treatment of cases that were problematic for the rigidly hierarchical structure of WFL
  - No unmotivated decisions on the derivational history of lexemes
- However, this means that a lot of potentially useful information of WFL is not represented in the Lemma Bank
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base

Present
  Adding Medieval Latin lemmas to WFL

Future
  Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts

Discussion and next steps
Including WFL into the LiLa Knowledge Base

- Modeling of WFL data into an ontology respecting the Linguistic Linked Open Data standards
- Reuse of classes and properties defined in existing ontologies
  - OntoLex core model
  - OntoLex Variation & Translation module (vartrans)
  - OntoLex Morphology module (morph)
  - LexInfo
  - LiLa
- Definition of new classes and properties specific to the WFL ontology

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, Giovanni Moretti
The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources
Third International Workshop on Resources and Tools for Derivational Morphology, 2021
Architecture of the OntoLex Morphology module (morph)
Architecture of morph: word formation
Architecture of the WFL ontology
Treatment of conversion in the WFL ontology

Matteo Pellegrini | CIRCSE, Università Cattolica del Sacro Cuore
Treatment of prefixation in the WFL ontology
Treatment of compounding in the WFL ontology

Matteo Pellegrini | CIRCSE, Università Cattolica del Sacro Cuore
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
The Lemma Bank currently includes also 63,621 Medieval Latin lemmas, taken from Du Cange’s *Glossarium mediæ et infimæ latinitatis*

WFL only provides information on 44,249 Classical Latin lemmas, taken from Glare 2012, Georges 1998, Gradenwitz 1904

As a consequence, also in the Lemma Bank information on affixes and bases is provided only for such Classical Latin lemmas

Current project:
- including (some of) Du Cange's lemmas into WFL;
- providing derivational information on such lemmas also in the Knowledge Base.
The method

- Identification of the most frequent derivational processes in WFL
- Automatic extraction of pairs of base-derivative candidates

<table>
<thead>
<tr>
<th>process</th>
<th>example</th>
<th>n. pairs in WFL</th>
<th>n. candidate pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-to-V prefixation</td>
<td>DUCO → CONDUCO</td>
<td>4,712</td>
<td>2,194</td>
</tr>
<tr>
<td>V-to-N -(t)io suffixation</td>
<td>ABIURO → ABIURATIO</td>
<td>2,555</td>
<td>381</td>
</tr>
<tr>
<td>V-to-N -(t)or suffixation</td>
<td>AMO → AMATOR</td>
<td>1,419</td>
<td>304</td>
</tr>
<tr>
<td>A-to-N -tas suffixation</td>
<td>PAGANUS → PAGANITAS</td>
<td>623</td>
<td>225</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Manual validation of the extracted pairs (work in progress with Claudia Corbetta and Martina Verdelli)
- Connection of the pairs where both the base and the derivative are Medieval Latin lemmas to existing WFL trees
Open issues

- The source of information is a glossary, not a dictionary
- Derivational information is not systematically provided
  → need for a case-by-case evaluation of the derivational relatedness of candidate pairs
- Not even semantic information is always provided
  → the evaluation is sometimes difficult
- Many entries are very substandard words, in some cases even simply copyists’ mistakes
  → impossible to provide derivational information in such cases
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base

  Present
    Adding Medieval Latin lemmas to WFL

  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base

Present
  Adding Medieval Latin lemmas to WFL

Future
  Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
A paradigmatic approach to word formation

- A **paradigm-oriented** perspective (Kyjánek 2020) to word formation is gaining ground (Van Marle 1984; Bauer 1997; Bauer 2019; Pounder 2000, Štekauer 2014, ...)

- Building **derivational paradigms** for Latin is a promising way to:
  - avoid the issues caused by the rigidly hierarchical, rule-based structure of WFL...
  - ...without the loss of information caused by the flat, family-based structure of the Lemma Bank.

- However, building a systematic derivational paradigm – i.e., capable to account for all the lexemes of a language – is not a trivial task (cf. Bonami and Strnadová 2019)

- What are the cells of the Latin derivational paradigm?
Building the Latin derivational paradigm

Litta and Budassi 2020 → **formal approach**
→ cells are inferred from the rules of WFL and expressed with a CxM-based notation

**TABLE 6.13 Core derivational paradigm for Latin filled with lemmas with bases FAC-, DIC-, AG**

<table>
<thead>
<tr>
<th>BASE [X]</th>
<th>FAC-</th>
<th>DIC-</th>
<th>AG-</th>
</tr>
</thead>
<tbody>
<tr>
<td>[x]V ↔ [SEM]V</td>
<td>facio</td>
<td>dico</td>
<td>ago</td>
</tr>
<tr>
<td>[x]N ↔ [SEM]N</td>
<td>factus</td>
<td>dictus</td>
<td>actus</td>
</tr>
<tr>
<td>[<a href="t">x</a>io]N ↔ [action of SEM]N</td>
<td>factio</td>
<td>dicio</td>
<td>actio</td>
</tr>
<tr>
<td>[<a href="t">x</a>or]N ↔ [one who SEM]N</td>
<td>factor</td>
<td>dictor</td>
<td>actor</td>
</tr>
<tr>
<td>[[x]j os]A ↔ [full of SEM]jA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[[x]j trix]N ↔ [a female who SEM]N</td>
<td>factionarius</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[[x]j ari]N ↔ [[dealer in SEM]N</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matteo Pellegrini | CIRCSE, Università Cattolica del Sacro Cuore
A formal approach

The structure is not fully parallel to inflectional paradigms in that the definition of cells is based on form-meaning pairs.

<table>
<thead>
<tr>
<th>base [x]</th>
<th>[[x]age]_N ↔ [action of SEM]_N</th>
<th>[[x]ment]_N ↔ [action of SEM]_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>laver</td>
<td>lavage</td>
<td>_</td>
</tr>
<tr>
<td>gonfle</td>
<td>_</td>
<td>gonflement</td>
</tr>
</tbody>
</table>

Derivational paradigm

<table>
<thead>
<tr>
<th>base [x]</th>
<th>[[x]us]_N ↔ [SEM.NOM.SG]_N</th>
<th>[[x]a]_N ↔ [SEM.NOM.SG]_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>lup-</td>
<td>lupus</td>
<td>_</td>
</tr>
<tr>
<td>ros-</td>
<td>_</td>
<td>rosa</td>
</tr>
</tbody>
</table>

Inflectional paradigm
A semantic approach

▶ Another option would be to define cells based on meaning alone

<table>
<thead>
<tr>
<th>base [x]</th>
<th>action of SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>laver</td>
<td>lavage</td>
</tr>
<tr>
<td>gonfler</td>
<td>gonflement</td>
</tr>
</tbody>
</table>

Derivational paradigm

<table>
<thead>
<tr>
<th>base [x]</th>
<th>SEM.NOM.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ros-</td>
<td>rosa</td>
</tr>
<tr>
<td>lup-</td>
<td>lupus</td>
</tr>
</tbody>
</table>

Inflectional paradigm

✔ Parallel to the definition of cells in inflectional paradigms
   → it would allow to model competition between derivational processes by the same means that have been deployed to model inflectional predictability (cf. Bonami and Strnadová 2019)

✗ Need for a systematic coding of morphosemantic relations that is currently lacking (and difficult to achieve)
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
The Lemma Bank is a collection of citation forms:
- PRS.ACT.IND.1SG for verbs (PRS.PASS.IND.1SG for deponents, PRS.ACT.IND.3SG for impersonals);
- NOM.(M).SG for nouns (and adjectives) (NOM.PL for pluralia tantum).

Citation forms are linked to their InflectionType through a dedicated property hasInflectionType.

InflectionTypes correspond to the traditional conjugations of verbs, declensions of nouns and classes of adjectives.

In addition, participial forms of verbs are provided as instances of the class Hypolemma, linked to their Lemma through the property isHypolemma.
Verb conjugations

<table>
<thead>
<tr>
<th>Conj.</th>
<th>Sample lexeme</th>
<th>PRS.ACT.INF</th>
<th>PRS.ACT.IND.3PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>LAUDO ‘praise’</td>
<td>laudāre</td>
<td>laudant</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>MONEO ‘warn’</td>
<td>monēre</td>
<td>monent</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>LEGO ‘read’</td>
<td>legere</td>
<td>legunt</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>VENIO ‘come’</td>
<td>venire</td>
<td>veniunt</td>
</tr>
<tr>
<td>mix.</td>
<td>CAPIO ‘take’</td>
<td>capere</td>
<td>capiunt</td>
</tr>
</tbody>
</table>

---

Matteo Pellegrini | CIRCSE, Università Cattolica del Sacro Cuore
Noun declensions

<table>
<thead>
<tr>
<th>Decl.</th>
<th>Sample lexeme</th>
<th>NOM.SG</th>
<th>GEN.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>ROSA ‘rose’</td>
<td>rosa</td>
<td>rosa&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>LUPUS ‘wolf’</td>
<td>lupus</td>
<td>lupí</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>URBS ‘city’</td>
<td>urbs</td>
<td>urbis</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>ARCUS ‘bow’</td>
<td>arcus</td>
<td>arcús</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt;</td>
<td>SPES ‘hope’</td>
<td>spes</td>
<td>speí</td>
</tr>
</tbody>
</table>

Rosa

http://lila-erc.eu/data/id/lemma/122165

<table>
<thead>
<tr>
<th>rdfs:label</th>
<th>rosa</th>
</tr>
</thead>
<tbody>
<tr>
<td>onto:writtenRep</td>
<td>rhosa</td>
</tr>
<tr>
<td>rdf:type</td>
<td>lila:Lemma</td>
</tr>
<tr>
<td>lila:hasBase</td>
<td><a href="http://lila-erc.eu/data/id/base/2896">http://lila-erc.eu/data/id/base/2896</a></td>
</tr>
<tr>
<td>lila:hasGender</td>
<td>lila:female</td>
</tr>
<tr>
<td>lila:hasInflectionType</td>
<td>lila:n1</td>
</tr>
<tr>
<td>lila:hasPOS</td>
<td>lila:noun</td>
</tr>
</tbody>
</table>
Adjective classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Sample lexeme</th>
<th>NOM.M.SG</th>
<th>NOM.F.SG</th>
<th>NOM.N.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>MAGNUS ‘big’</td>
<td>magnus</td>
<td>magna</td>
<td>magnum</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ACER ‘sharp’</td>
<td>acer</td>
<td>acris</td>
<td>acre</td>
</tr>
</tbody>
</table>
However, traditional classifications capture only partly the inflectional behaviour of lexemes → the impact of **stem allomorphy** is disregarded

- Present vs. perfect and third stem of verbs

<table>
<thead>
<tr>
<th>Conj.</th>
<th>Sample lexeme</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND.1SG</th>
<th>SUP.ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ˢᵗ</td>
<td>LAUDO ‘praise’</td>
<td>laudāre</td>
<td>laudāvī</td>
<td>laudātum</td>
</tr>
<tr>
<td>1ˢᵗ</td>
<td>CREPO ‘rattle’</td>
<td>crepāre</td>
<td>crepuī</td>
<td>crepitum</td>
</tr>
<tr>
<td>1ˢᵗ</td>
<td>SECO ‘crack’</td>
<td>secāre</td>
<td>secui</td>
<td>sectum</td>
</tr>
</tbody>
</table>

- Direct vs. oblique cases of 3ʳᵈ decl. nouns and 2ⁿᵈ class adjectives

<table>
<thead>
<tr>
<th>Decl.</th>
<th>Sample lexeme</th>
<th>NOM.SG</th>
<th>GEN.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>3ʳᵈ</td>
<td>URBS ‘city’</td>
<td>urbs</td>
<td>urbis</td>
</tr>
<tr>
<td>3ʳᵈ</td>
<td>CONSUL ‘consul’</td>
<td>consul</td>
<td>consulis</td>
</tr>
<tr>
<td>3ʳᵈ</td>
<td>FLUMEN ‘river’</td>
<td>flumen</td>
<td>fluminis</td>
</tr>
</tbody>
</table>
Latin dictionaries and grammars summarize (almost) the whole inflectional behaviour of lexemes by providing other **principal parts** alongside the citation form:

- **PRS.ACT.INF**, **PRF.ACT.INF.1SG**, **SUP.ACC** for verbs;
- **GEN.SG** for nouns;
- **NOM.F.SG** and **NOM.N.SG** (when these forms are distinct from the citation form) or **GEN.M/F/N.SG** for adjectives.

The notion of principal parts have been recently recovered by theoretically-grounded studies → Stump and Finkel 2013, Bonami and Beniamine 2016
Outline

Introduction
  Linked Data Principles
  LiLa: Linking Latin

Word formation
  Past
    Word Formation Latin (WFL)
    Word Formation in the Lemma Bank
    Including WFL into the LiLa Knowledge Base
  Present
    Adding Medieval Latin lemmas to WFL
  Future
    Building the Latin derivational paradigm

Inflection
  Information on inflection in the Lemma Bank
  An ontology for Latin principal parts
  Discussion and next steps
We are working on a resource listing **principal parts** and providing **fine-grained inflectional information** for Latin verbs, nouns and adjectives.

This resource will be modeled into an ontology linked to the LiLa Knowledge Base.

Principal parts can be useful to accomodate resources that make lemmatization choices different than the ones of the LiLa Lemma Bank:

- Verbs that are sometimes lemmatized under PRF.ACT.IND.1SG → e.g. COEPI (rather than COEPIO) ‘begin’
- Lexical resources that use PRS.ACT.INF as citation form → cf. Du Cange’s *Glossarium*

Fine-grained inflectional information can be useful to allow for more sophisticated queries indentifying lexemes that follow patterns more specific than their traditional conjugation/declension:

- 3rd decl. nouns with NOM.SG in -o and GEN.SG in -inis → e.g. HOMO,HOMINIS ‘man’, VERTIGO,VERTIGINIS ‘whirl’, CALIGO,CALIGINIS ‘mist’, ...

Both can be used to generate full paradigms for Latin lexemes.
Generation of principal parts

- Principal parts are generated from the database of Lemlat, a recently renewed Latin morphological analyzer (Passarotti et al. 2017)
- For each lexeme, Lemlat’s database lists a set of LExical Segments (LES) – roughly corresponding to the stems used in different portions of the paradigm
- To each LES, Lemlat’s database associate a CODLES that provides information on the ending compatible with that LES
- Joint information on LES and CODLES allows to generate 6 principal parts

<table>
<thead>
<tr>
<th>Information in Lemlat</th>
<th>Generated Principal Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>LES</td>
<td>CODLES</td>
</tr>
<tr>
<td>rump</td>
<td>v3r</td>
</tr>
<tr>
<td>rumpere</td>
<td></td>
</tr>
<tr>
<td>rup</td>
<td>v7s</td>
</tr>
<tr>
<td>rupt</td>
<td>n41</td>
</tr>
<tr>
<td>ruptus</td>
<td>n6p1</td>
</tr>
<tr>
<td>ruptur</td>
<td>n6p2</td>
</tr>
</tbody>
</table>

- The stems of PRF.PTCP.NOM.M.SG and FUT.PTCP.NOM.M.SG almost always coincide with the stem of SUP.ACC, but there are a few cases where they are different
  → having them as additional principal parts allows to generate full paradigms for all lexemes
Coding of fine-grained inflectional behaviour

We use **binary alternation patterns** to capture the inflectional behaviour of lexemes (Bonami and Boyé 2014, Beniamine 2018)

<table>
<thead>
<tr>
<th>lexeme</th>
<th>PRS.ACT.IND.1SG</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND.1SG</th>
<th>PRS.1SG-INF</th>
<th>PRS.1SG-PRF</th>
<th>INF-PRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAUDO</td>
<td>laudo</td>
<td>laudare</td>
<td>laudai</td>
<td>o ⇔ are</td>
<td>o ⇔ aui</td>
<td>re ⇔ ui</td>
</tr>
<tr>
<td>CUBO</td>
<td>cubo</td>
<td>cubare</td>
<td>cubui</td>
<td>o ⇔ are</td>
<td>o ⇔ ui</td>
<td>are ⇔ ui</td>
</tr>
<tr>
<td>MONEO</td>
<td>moneo</td>
<td>monere</td>
<td>monui</td>
<td>o ⇔ re</td>
<td>eo ⇔ ui</td>
<td>ere ⇔ ui</td>
</tr>
<tr>
<td>DELEO</td>
<td>deleo</td>
<td>delere</td>
<td>deleui</td>
<td>o ⇔ re</td>
<td>eo ⇔ eui</td>
<td>re ⇔ ui</td>
</tr>
</tbody>
</table>

This procedure has the advantage of being applicable algorithmically ≠ global segmentation in stem vs. endings valid for all paradigm cells (cf. Beniamine 2018)

We use the Qumin toolkit to extract alternation patterns between all pairs of principal parts (without context because our data is in orthographic rather than phonological transcription)
### Treatment of overabundance

Cases of **overabundance**: sometimes more than one form can be generated for the same cell → this would cause a multiplication of alternation patterns

- Due to the availability of different **stem allomorphs**

<table>
<thead>
<tr>
<th>lexeme</th>
<th>PRS.ACT.IND.1SG</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND</th>
<th>SUP.ACC</th>
<th>PRS.INF-PRF.IND</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABALIENO</td>
<td>abalieno</td>
<td>abalienare</td>
<td>abalienaui</td>
<td>abalienatum</td>
<td>re ⇔ ui, i_re ⇔ _ui,</td>
</tr>
<tr>
<td>‘separate’</td>
<td>abaleno</td>
<td>abalenare</td>
<td>abalenai</td>
<td>abalenatum</td>
<td>_re ⇔ i_ui</td>
</tr>
</tbody>
</table>

- Due to the possibility of different **inflection class assignments**

<table>
<thead>
<tr>
<th>lexeme</th>
<th>PRS.ACT.IND.1SG</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND</th>
<th>SUP.ACC</th>
<th>PRS.INF-PRF.IND</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAVO</td>
<td>lauo</td>
<td>lauare</td>
<td>lauui</td>
<td>lauatum</td>
<td>re ⇔ ui, are ⇔ ui,</td>
</tr>
<tr>
<td>‘wash’</td>
<td>lauo</td>
<td>lauere</td>
<td>laui</td>
<td>lautum</td>
<td>ere ⇔ aui, ere ⇔ ui</td>
</tr>
</tbody>
</table>
The notion of flexeme

- Fradin and Kerleroux 2003 → proposal to distinguish between:
  - **lexemes** → lexical units with a unique **meaning**
  - **flexemes** → lexical units with a unique **form** (i.e., a unique inflectional paradigm)

- This distinction was originally introduced to account for mismatches where one flexeme maps to different lexemes:
  - derivatives that select a specific meaning among the ones of the base
e.g. French FILLE ‘girl, daughter’ → FILLETTE ‘small girl’ (*‘small daughter’)

<table>
<thead>
<tr>
<th>flexeme</th>
<th>lexeme</th>
<th>SG</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FILLE</td>
<td>FILLE_1 ‘girl’</td>
<td>fille</td>
<td>filles</td>
</tr>
<tr>
<td></td>
<td>FILLE_2 ‘daughter’</td>
<td>fille</td>
<td>filles</td>
</tr>
</tbody>
</table>

- It has been recently applied to mismatches where one lexeme maps to different flexemes (Thornton 2018, Bonami and Crysmann 2018)
  - overabundance, e.g. Italian gender-alternating noun ORECCHIO_M/ORECCHIA_F ‘ear’

<table>
<thead>
<tr>
<th>lexeme</th>
<th>flexeme</th>
<th>SG</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORECCHIO/-A</td>
<td>ORECCHIO</td>
<td>orecchio</td>
<td>orecchi</td>
</tr>
<tr>
<td></td>
<td>ORECCHIA</td>
<td>orecchia</td>
<td>orecchie</td>
</tr>
</tbody>
</table>
A collection of flexemes

- Our resource is a collection of flexemes rather than lexemes (a “flexicon”)
- This is conceptually motivated by the fact that “inflection is about flexemes” (Bonami and Crysmann 2018: 184)
- This also allows to:
  - avoid the multiplication of patterns;
  - express the connection between forms that display the same stem or belong to the same inflection class.

<table>
<thead>
<tr>
<th>lexeme</th>
<th>flexeme</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND</th>
<th>PRS.INF-PRF.IND</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAVO ‘wash’</td>
<td>LAVO\textsubscript{1st}</td>
<td>lau\textit{are}</td>
<td>lau\textit{au}i</td>
<td>re ⇔ ui</td>
</tr>
<tr>
<td></td>
<td>LAVO\textsubscript{3rd}</td>
<td>lau\textit{ere}</td>
<td>lau\textit{i}</td>
<td>ere ⇔ ui</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lexeme</th>
<th>flexeme</th>
<th>PRS.ACT.INF</th>
<th>PRF.ACT.IND.1SG</th>
<th>PRS.INF-PRF.IND</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAL(I)ENO ‘separate’</td>
<td>ABALIENO</td>
<td>\textit{abalien}are</td>
<td>\textit{abalien}au</td>
<td>re ⇔ ui</td>
</tr>
<tr>
<td></td>
<td>ABALENO</td>
<td>\textit{abal}en\textit{are}</td>
<td>\textit{abal}en\textit{au}i</td>
<td>re ⇔ ui</td>
</tr>
</tbody>
</table>
Architecture of the LatInFlexi ontology
Architecture of the LatInFlexi ontology

▶ Each Flexeme is connected:
   ▶ to the corresponding Lemma in the Lemma Bank through the property ontolex:canonicalForm;
   ▶ to its principal parts through the property ontolex:otherForm;
   ▶ to the Pattern between each of its principal parts through the property latinflexi:hasPattern

▶ Each Form is linked to the Cell it fills through the property latinflexi:fillsCell
▶ Each Pattern is linked to the Cells it relates through the property latinflexi:relates
▶ Each Cell is coded using the schema of the UniMorph project (Sylak-Glassman 2016)
  → an olia annotation model is available for this tagset (Chiarcos, Fäth, and Abromeit 2020)
  that allows for integration with other vocabularies for linguistic description (e.g. ISOCat, GOLD)
▶ The overall inflection (micro)class is not stored, but it can be inferred
  → two flexemes belong to the same class if they share the exact same patterns
Example: LAUDO

Matteo Pellegrini | CIRCSE, Università Cattolica del Sacro Cuore
Example: ABALENO
Flexeme-Lemma mapping

- Overabundance due to the availability of different stem allomorphs → different flexemes map to the same lemma
- Overabundance due to the possibility of different inflection class assignments → different flexemes map to different lemmas (due to modeling choices of the Lemma Bank)
  - Lexemic identity is not lost thanks to the property lila:lemmaVariant between the two lemmas

<table>
<thead>
<tr>
<th>unit</th>
<th>n.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms</td>
<td>39,438</td>
</tr>
<tr>
<td>Flexemes</td>
<td>10,716</td>
</tr>
<tr>
<td>Lemmas</td>
<td>7,973</td>
</tr>
</tbody>
</table>

(only Classical Latin verbs, excluding a few irregulars and their derivatives)
Outline

Introduction
- Linked Data Principles
- LiLa: Linking Latin

Word formation
- Past
  - Word Formation Latin (WFL)
  - Word Formation in the Lemma Bank
  - Including WFL into the LiLa Knowledge Base
- Present
  - Adding Medieval Latin lemmas to WFL
- Future
  - Building the Latin derivational paradigm

Inflection
- Information on inflection in the Lemma Bank
- An ontology for Latin principal parts

Discussion and next steps
An alternative modeling choice: Lexemes as lexical entries
Alternative modeling choices: a comparison

Flexemes as lexical entries

✓ allows to use ontolex vocabulary for the linking with Forms
✗ lexemes have no explicit representation (LiLa’s lemmas are ontolex:forms!)

Lexemes as lexical entries

✓ lexemes have an explicit representation
✗ need for new vocabulary to link Flexemes to Forms
✗ redundancy: for most Lexemes there is a 1:1 mapping with Lemmas
✗ lack of homogeneity: a different treatment is required in cases where several Flexemes correspond to same Lemma
Lexeme-Lemma mapping: 1:2
Lexeme-Lemma mapping: 2:2
Adaptive principal parts

We have seen that we have three different principal parts for SUP.ACC, PRF.PTCP.NOM.M.SG and FUT.PTCP.NOM.M.SG, as there are a few verbs for which these cells are based on different stems.

However, for the overwhelming majority of verbs these three cells are interpredictable.

A more economical solution might be to provide adaptive principal parts (Stump and Finkel 2013), i.e. listing additional forms only for verbs that need them:

<table>
<thead>
<tr>
<th>ABUTOR ‘consume’</th>
<th>MORIOR ‘die’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>form</strong></td>
<td><strong>cell</strong></td>
</tr>
<tr>
<td>abutor</td>
<td>PRS.PASS.IND.1SG</td>
</tr>
<tr>
<td>abuti</td>
<td>PRS.PASS.INF</td>
</tr>
<tr>
<td>_</td>
<td>PRF.ACT.IND.1SG</td>
</tr>
<tr>
<td>abusus</td>
<td>PRF.PTCP.NOM.M.SG</td>
</tr>
<tr>
<td>_</td>
<td>FUT.PTCP.NOM.M.SG</td>
</tr>
</tbody>
</table>
Troubles with flexemes

- We start from the strong assumption that all cases of overabundance should yield the introduction of distinct flexemes ⇒ the entries of our lexicon are never overabundant

- Theoretical issue: previous studies leave this question open (cf. Thornton 2018: 312-3, Bonami and Crysmann 2018: 198)

- Practical issue: in some cases, the identification of flexemes is not straightforward

- Cases of non-systematic overabundance across different cells → multiplication of flexemes

<table>
<thead>
<tr>
<th>lexeme</th>
<th>PRS.ACT.IND.1SG</th>
<th>PRS.ACT.INF</th>
<th>SUP.ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERG(E)O ‘cleanse’</td>
<td>tergeo</td>
<td>tergère</td>
<td>tersum</td>
</tr>
<tr>
<td></td>
<td>tergo</td>
<td>tergere</td>
<td>tertum</td>
</tr>
</tbody>
</table>

⇒ 4 Flexemes

- The situation can be far more complex!
Example: TERG(E)O
Phonological distinctions

Coding phonological distinctions

- Distinction between semivocalic and vocalic <i> (/j/ vs. /i/) and <u> (/u/ vs. /w/)
- Vowel length
  → remarkable impact on alternation patterns!
  e.g. PRS.ACT.IND.3SG *venit vs. PRF.ACT.IND.3SG vēnīt

Source: Lewis and Short 1879?

✓ Available in machine-readable format
✓ Recently added to the Knowledge Base (Mambrini et al. 2021)
✗ Not among the dictionaries on which Lemlat is based
  → only partial overlap with the entries of our lexicon
✗ Vowel length is not marked when it is phonologically predictable
Full paradigms

- Using `morph` vocabulary to express rules to generate all the wordforms of (a selection of) lexemes from the principal parts
- Integration of our ontology with `morph` can be achieved by establishing a sub-class relation between the (dynamically generated?) `latinflexi:MicroClass` and `morph:Paradigm`
Each flexeme can be said to belong to a microclass that is composed of all the patterns that are displayed between the principal parts of that lexeme.

This information can be inferred → two flexemes belong to the same microclass if they share the exact same patterns:
- remarkable reduction of stored triples

However, having this information stored explicitly would allow for:
- simpler queries to extract the overall class;
- an easier integration with morph (by establishing a subclass relation between our microclass and morph:Paradigm)
Corpus frequencies

- Using the `frac` (Frequency, Attestation and Corpus information) module of Ontolex to provide corpus frequencies for the generated wordforms

- Corpora with the appropriate level of granularity:
  - Treebanks
    - Index Thomisticus Treebank (Passarotti 2011)
    - UDante (Cecchini et al. 2020)
    - Latin Dependency Treebank (Bamman and Crane 2011)
    - PROIEL (Haug and Jøhndal 2008)
    - Late Latin Charter Treebank (Cecchini, Korkiakangas, and Passarotti 2020)
  - Computational Historical Semantics corpus
  - LASLA corpus?
    (annotation uninformative on gender for adjectives)
Stored and dynamically generated triples
Matteo Pellegrini
CIRCSE, Università Cattolica del Sacro Cuore

matteo.pellegrini@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.


Bonami, Olivier and Berthold Crysmann (2018). “Lexeme and flexeme in a formal theory of grammar”. In: The lexeme in descriptive and theoretical morphology 4, p. 175.


Works cited III


