Morphological resources in the LiLa Knowledge Base

An overview of past, present and future work

Matteo Pellegrini
matteo.pellegrini@unicatt.it

UNIVERSITÀ CATTOLICA del Sacro Cuore

IRCSE

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

Outline

Linked Data principles

- 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 **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (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

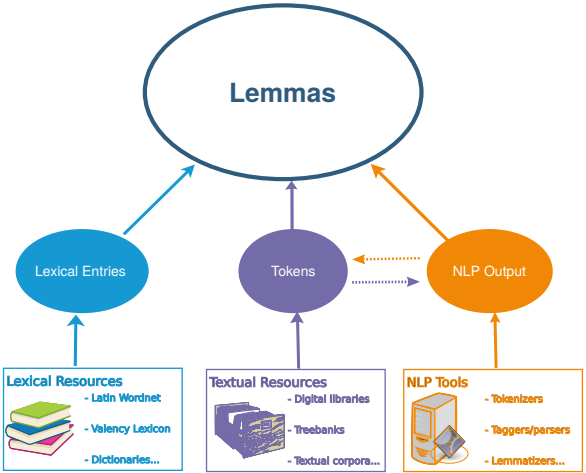
  → **Linguistic Linked Open Data** cloud (Cimiano et al. 2020)

Outline

LiLa: Linking Latin

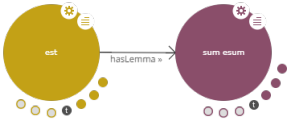
- Open-ended **Knowledge Base** of interoperable linguistic resources for Latin sharing a common vocabulary for knowledge description
- Use of **web standards** to represent and query data
  - RDF: information is coded in terms of **triples**, connecting a **subject** to an **object** through a **property**
  - 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



RDF data model

Nunc **est** bibendum
(Horace, *Odes*, 1, 37)



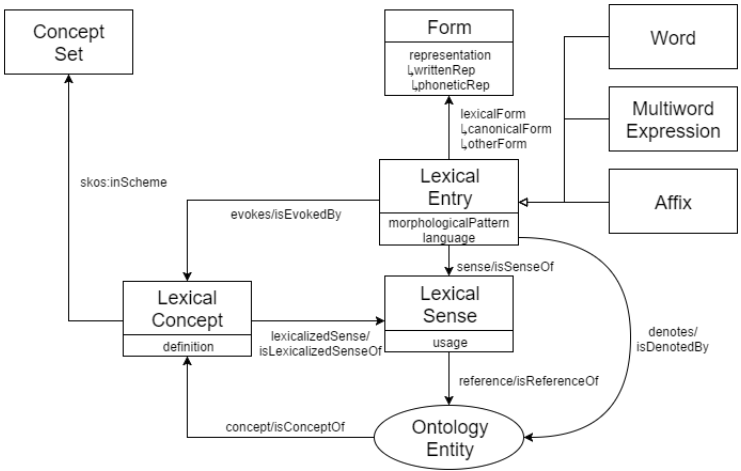
est

http://lila-erc.eu/data/corpora/Horace_Odes_b6ff30d5-891a-4a21-8ecf-cb2b96143db6/id/citationUnit/Horace_Odes/CiteStructure/s-340/t-2

<http://purl.org/powla/powla.owl#**Terminal**>

rdfs:**label**	est
<http://purl.org/powla/powla.owl#**hasStringValue**>	est
rdf:**type**	<http://purl.org/powla/powla.owl#Terminal>
<http://purl.org/powla/powla.owl#**next**>	<http://lila-erc.eu/data/corpora/Horace_Odes_b6ff30d5-891a-4a21-8ecf-cb2b96143db6/id/citation ↳ bibendum
lila.**hasLemma**	<http://lila-erc.eu/data/id/lemma/126689> ↳ sum
<http://lila-erc.eu/ontologies/lila_corpora/**hasCitStructure**>	<http://lila-erc.eu/data/corpora/Horace_Odes_b6ff30d5-891a-4a21-8ecf-cb2b96143db6/id/citation ↳ Citation Layer
<http://purl.org/powla/powla.owl#**previous**>	<http://lila-erc.eu/data/corpora/Horace_Odes_b6ff30d5-891a-4a21-8ecf-cb2b96143db6/id/citation ↳ Nunc

Architecture of the OntoLex model



LiLa: connected resources and upcoming connections

- **Corpora**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] Dante Search (700th death anniversary): ca. 46,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa corpora
- **Lexica**
  - [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. 2,300 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
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 13.5 million triples**

Outline

Outline

Word Formation Latin (WFL)

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

input lexeme(s) (PoS)	output lexeme (PoS)	prefix	suffix	WFR
FELIX 'happy' (A)	INFELIX 'unhappy' (A)	in-	-	A-to-A in-
FELIX 'happy' (A)	FELICITAS 'happiness' (N)	-	-tas	A-to-N -tas
MALUS 'bad' (A)	MALUM 'bad thing' (N)	-	-	A-to-N
AGER 'field' (N); COLO 'to cultivate' (V)	AGRICOLA 'farmer' (N)	-	-	N+V=N

→ **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$_A$ ‘opposed’ → ADVERSARIUS$_N$ ‘opponent’ or *vice versa*?
    → A decision has to be made even in cases that are not clear-cut
  - Parasynthetic formations:
    AQUA ‘water’ → *AQUESCO (?) → EXAQUESCO ‘become water’
    → Even if they are not attested, intermediate steps need to be added (“fictional lemmas”)

Outline

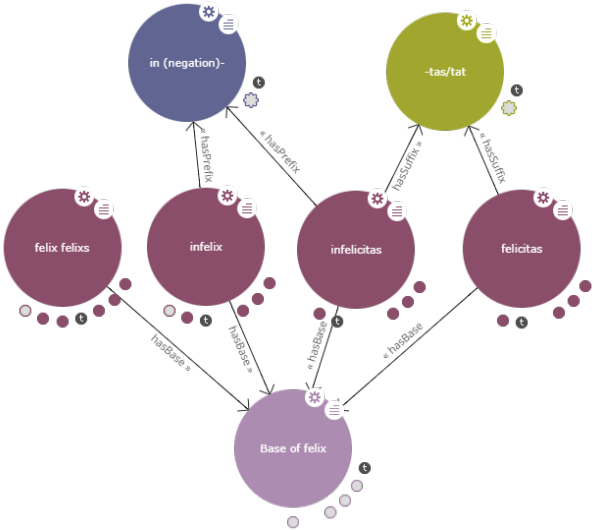
Word formation in the Lemma Bank

- 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

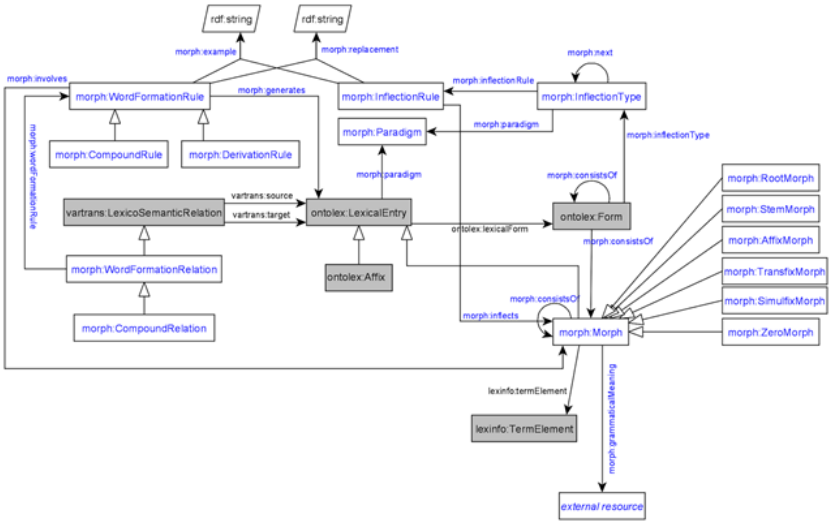
Outline

Including WFL into the LiLa Knowledge Base

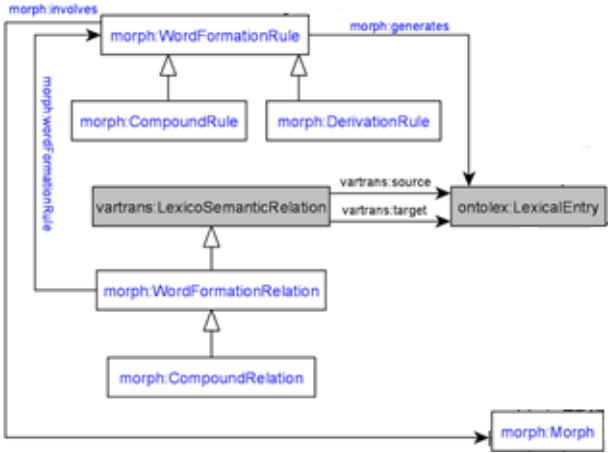
- Modeling of WFL data into an ontology respecting the Linguistic Linked Open Data standards
- Reuse of classes and properties defined in existing ontologies
  - OntoLex core model
  - OntoLex Variation & Translation module (`vartrans`)
  - OntoLex Morphology module (`morph`)
  - LexInfo
  - LiLa
- Definition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

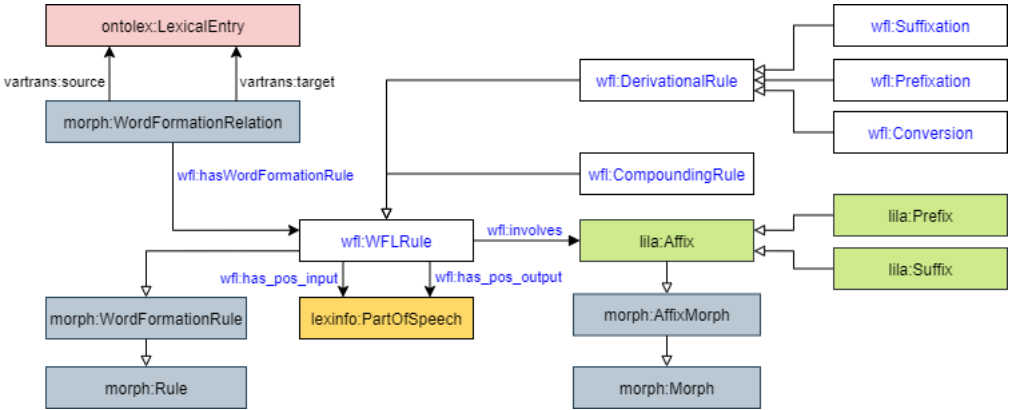
Architecture of the OntoLex Morphology module (`morph`)



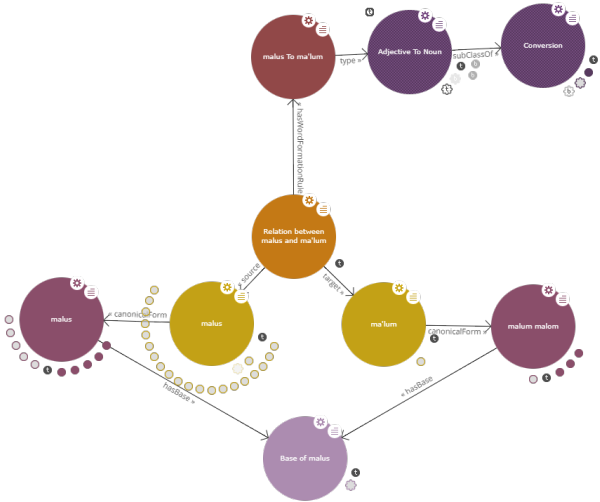
Architecture of `morph`: word formation



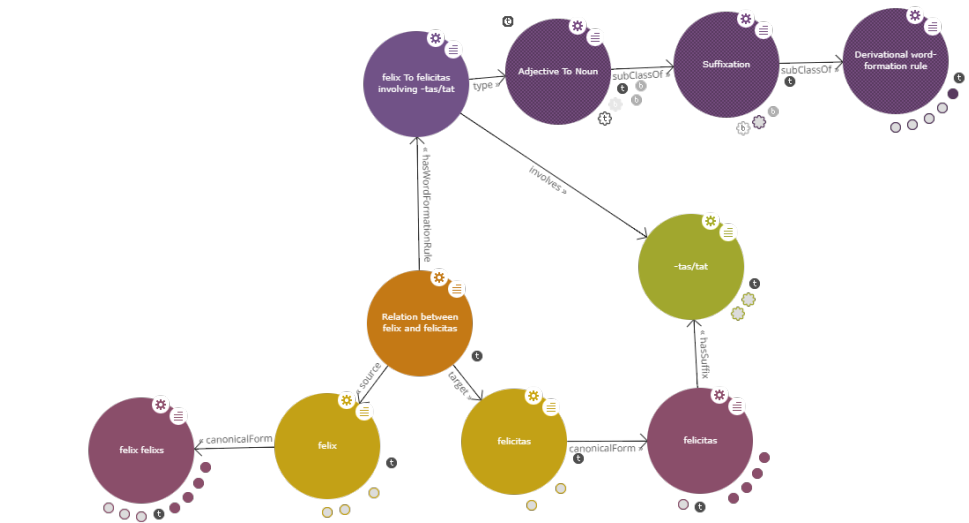
Architecture of the WFL ontology



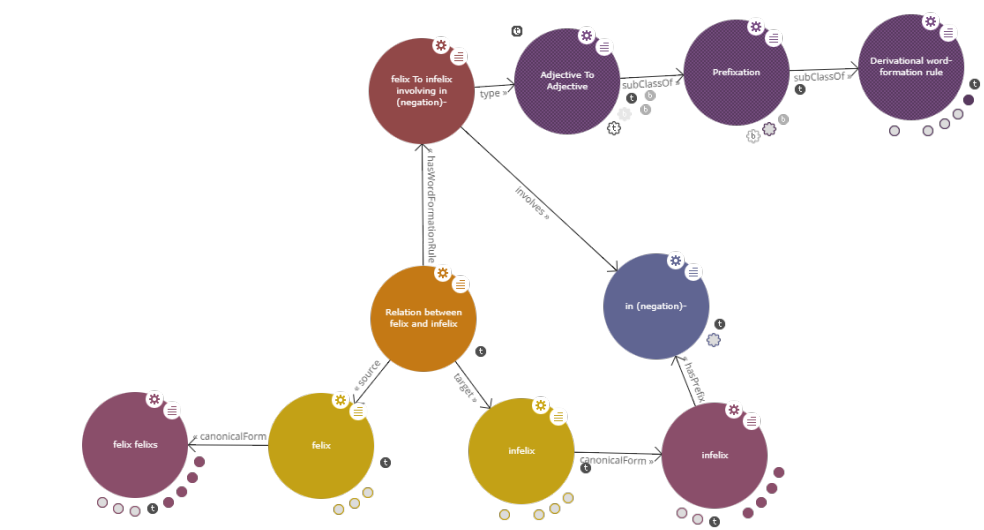
Treatment of conversion in the WFL ontology



Treatment of suffixation in the WFL ontology



Treatment of prefixation in the WFL ontology



Treatment of compounding in the WFL ontology

Outline

Outline

Adding Medieval Latin lemmas to WFL

- 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

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

- 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

Outline

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-*

BASE [X]	*FAC-*	*DIC-*	*AG-*
$[x]_V \leftrightarrow [\text{SEM}]_V$	*facio*	*dico*	*ago*
$[x]_N \leftrightarrow [\text{SEM}]_N$	*factus*	*dictus*	*actus*
	factum	*dictum*	*actum*
$[x]_A \leftrightarrow [\text{SEM}]_A$			
$[[x](t)io]_N \leftrightarrow [\text{action of SEM}]_N$	*factio*	*dicio*	*actio*
$[[x](t)or]_N \leftrightarrow [\text{one who SEM}]_N$	*factor*	*dictor*	*actor*
$[[x]]_j \text{ os}]_A \leftrightarrow [\text{full of SEM}_j]_A$			*actuosus*
$[[x]_j \text{ ari}]_A \leftrightarrow [\text{having qualities of SEM}_j]$			*actuarius*
$[[x]_j \text{ trix}]]_N \leftrightarrow [\text{a female who SEM}_j]_N$	*factrix*		*actrix*
$[[x]_j \text{ ari}]_N \leftrightarrow [[\text{dealer in SEM}_j]_N$	*factionarius*		*actuaria*
			actuarius

A formal approach

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

base [x]	[[x]age]$_N$ ↔ [action of SEM]$_N$	[[x]ment]$_N$ ↔ [action of SEM]$_N$
laver	*lavage*	_
gonfler	_	*gonflement*

Derivational paradigm

base [x]	[[x]us]$_N$ ↔ [SEM.NOM.SG]$_N$	[[x]a]$_N$ ↔ [SEM.NOM.SG]$_N$
lup-	*lupus*	_
ros-	_	*rosa*

Inflectional paradigm

A semantic approach

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

base [x]	action of SEM
laver	*lavage*
gonfler	*gonflement*

Derivational paradigm

base [x]	SEM.NOM.SG
ros-	*rosa*
lup-	*lupus*

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

Information on inflection in the Lemma Bank

- 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 af the class `Hypolemma`, linked to their `Lemma` through the property `isHypolemma`

Verb conjugations

laudo

http://lila-erc.eu/data/id/lemma/110078

rdfs:**label**	laudo
ontolex:**writtenRep**	laudo
rdf:**type**	lila:Lemma ↳ Lemma
lila:**hasBase**	<http://lila-erc.eu/data/id/base/217> ↳ Base of laus
lila:**hasInflectionType**	lila:v1r ↳ first conjugation verb
lila:**hasPOS**	lila:verb ↳ verb

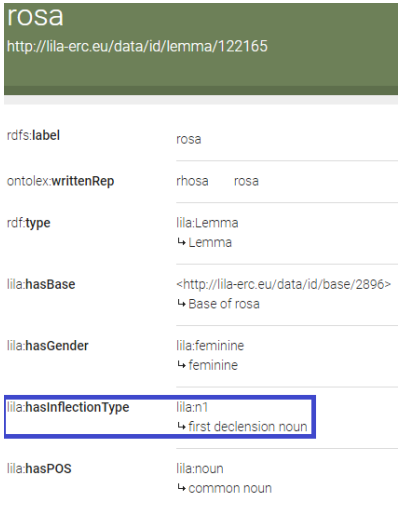
laudatus

http://lila-erc.eu/data/id/hypolemma/25328

lila:**isHypolemma**	<http://lila-erc.eu/data/id/lemma/110078> ↳ laudo
ontolex:**writtenRep**	laudatus
rdf:**type**	lila:Hypolemma ↳ Hypolemma

Conj.	Sample lexeme	PRS.ACT.INF	PRS.ACT.IND.3PL
1st	LAUDO ‘praise’	*laudāre*	*laudant*
2nd	MONEO ‘warn’	*monēre*	*monent*
3rd	LEGO ‘read’	*legere*	*legunt*
4th	VENIO ‘come’	*venīre*	*veniunt*
mix.	CAPIO ‘take’	*capere*	*capiunt*

Noun declensions



Decl.	Sample lexeme	NOM.SG	GEN.SG
1st	ROSA ‘rose’	*rosa*	*rosae*
2nd	LUPUS ‘wolf’	*lupus*	*lupī*
3rd	URBS ‘city’	*urbs*	*urbis*
4th	ARCUS ‘bow’	*arcus*	*arcūs*
5th	SPES ‘hope’	*spes*	*speī*

Adjective classes

magnus

http://lila-erc.eu/data/id/lemma/111319

rdfs:**label**	magnus
ontolex:**writtenRep**	magnus
rdf:**type**	lila:Lemma ↳ Lemma
lila:**hasBase**	<http://lila-erc.eu/data/id/base/2167> ↳ Base of magnus
lila:**hasDegree**	lila:positive ↳ positive
lila:**hasInflectionType**	lila:n6 ↳ first class adjective
lila:**hasPOS**	lila:adjective ↳ adjective

Class	Sample lexeme	NOM.M.SG	NOM.F.SG	NOM.N.SG
1st 2nd	MAGNUS ‘big’ ACER ‘sharp’	*magnus* *acer*	*magna* *acris*	*magnum* *acre*

Missing pieces

- 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

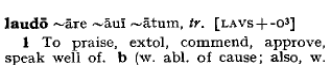
Conj.	Sample lexeme	PRS.ACT.INF	PRF.ACT.IND.1SG	SUP.ACC
1st	LAUDO 'praise'	*laudāre*	*laudāvī*	*laudātum*
1st	CREPO 'rattle'	*crepāre*	*crepuī*	*crepitum*
1st	SECO 'crack'	*secāre*	*secuī*	*sectum*

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

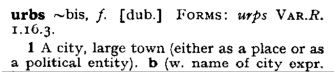
Decl.	Sample lexeme	NOM.SG	GEN.SG
3rd	URBS 'city'	*urbs*	*urbis*
3rd	CONSUL 'consul'	*consul*	*consulis*
3rd	FLUMEN 'river'	*flumen*	*fluminis*

Information on inflection in Latin traditional descriptions

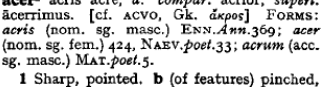
- Latin dictionaries and grammars summarize (almost) the whole inflectional behaviour of lexemes by providing other **principal parts** alongside the citation form:
  - PRS.ACT.INF, PRF.ACT.INF.1SG, SUP.ACC for verbs;
  - GEN.SG for nouns;
  - NOM.F.SG and NOM.N.SG (when these forms are distinct from the citation form) or GEN.M/F/N.SG for adjectives.



laudō ~āre ~āuī ~ātum, *tr.* [LAVS+-O³]
1 To praise, extol, commend, approve, speak well of. **b** (w. abl. of cause; also, w.



urbs ~bis, *f.* [dub.] FORMS: *urps* VAR.*R.* 1.16.3.
1 A city, large town (either as a place or as a political entity). **b** (w. name of city expr.



ācer² ācris ācre, *a. compar.* ācrior, *superl.* ācerrimus. [cf. ACVO, Gk. *ákpos*] FORMS: *acris* (nom. sg. masc.) ENN.*Ann.*369; *acer* (nom. sg. fem.) 424, NAEV.*poet.*33; *acrum* (acc. sg. masc.) MAT.*poet.*5.
1 Sharp, pointed. **b** (of features) pinched,

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

Outline

Principal parts and fine-grained inflectional information in LiLa

- 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

information in Lemlat		generated principal part	
LES	CODLES	form	cell
rump	v3r	*rumpo*	PRS.ACT.IND.1SG
		rumpere	PRS.ACT.INF
rup	v7s	*rupi*	PRF.ACT.IND.1SG
rupt	n41	*ruptum*	SUP.ACC
rupt	n6p1	*ruptus*	PRF.PTCP.NOM.M.SG
ruptur	n6p2	*rupturus*	FUT.PTCP.NOM.M.SG

- 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

lexeme	PRS.ACT.IND.1SG	PRS.ACT.INF	PRF.ACT.IND.1SG	PRS.1SG-INF	PRS.1SG-PRF	INF-PRF
LAUDO	*laudo*	*laudare*	*laudaui*	o ⇌ are	o ⇌ aui	re ⇌ ui
CUBO	*cubo*	*cubare*	*cubui*	o ⇌ are	o ⇌ ui	are ⇌ ui
MONEO	*moneo*	*monere*	*monui*	o ⇌ re	eo ⇌ ui	ere ⇌ ui
DELEO	*deleo*	*delere*	*deleui*	o ⇌ re	eo ⇌ eui	re ⇌ ui

- 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**

lexeme	PRS.ACT.IND.1SG	PRS.ACT.INF	PRF.ACT.IND	SUP.ACC	PRS.INF-PRF.IND
ABALIENO	***abalien****o*	***abalien****are*	***abalien****aui*	***abalien****atum*	re ⇌ ui, i_re ⇌ _ui,
‘separate’	***abalen****o*	***abalen****are*	***abalen****aui*	***abalen****atum*	_re ⇌ i_ui

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

lexeme	PRS.ACT.IND.1SG	PRS.ACT.INF	PRF.ACT.IND	SUP.ACC	PRS.INF-PRF.IND
LAVO	*lau****o***	*lau****are***	*lau****aui***	*lau****atum***	re ⇌ ui, are ⇌ ui,
‘wash’	*lau****o***	*lau****ere***	*lau****i***	*lau****tum***	ere ⇌ aui, ere ⇌ ui

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')

flexeme	lexeme	SG	PL
FILLE	FILLE_1 'girl'	*fille*	*filles*
	FILLE_2 'daughter'	*fille*	*filles*

- 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'

lexeme	flexeme	SG	PL
ORECCHIO/-A	ORECCHIO	*orecchio*	*orecchi*
	ORECCHIA	*orecchia*	*orecchie*

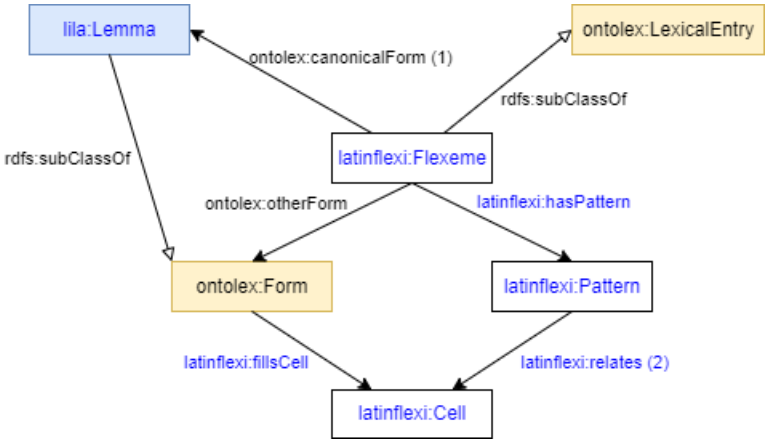
A collection of flexemes

- 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.

lexeme	flexeme	PRS.ACT.INF	PRF.ACT.IND	PRS.INF-PRF.IND
LAVO ‘wash’	LAVO_{1st}	*lau**are***	*lau**aui***	re ⇌ ui
	LAVO_{3rd}	*lau**ere***	*lau**i***	ere ⇌ ui

lexeme	flexeme	PRS.ACT.INF	PRF.ACT.IND.1SG	PRS.INF-PRF.IND
ABAL(I)ENO ‘separate’	ABALIENO	***abalien**are*	***abalien**aui*	re ⇌ ui
	ABALENO	***abalen**are*	***abalen**aui*	re ⇌ ui

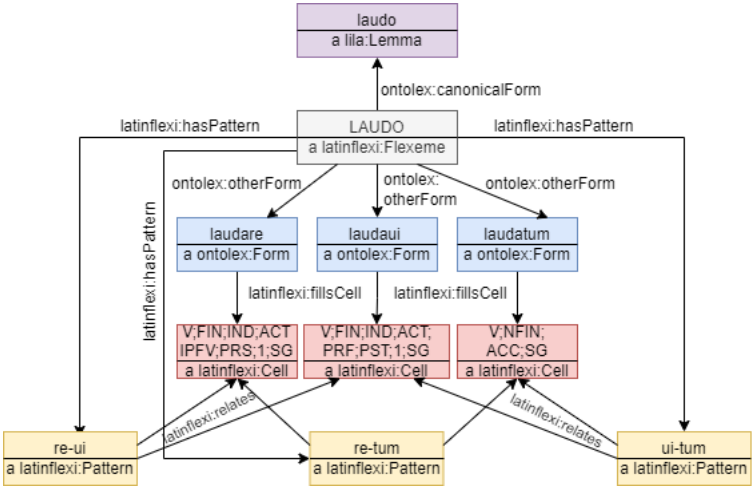
Architecture of the LatInFlexi ontology



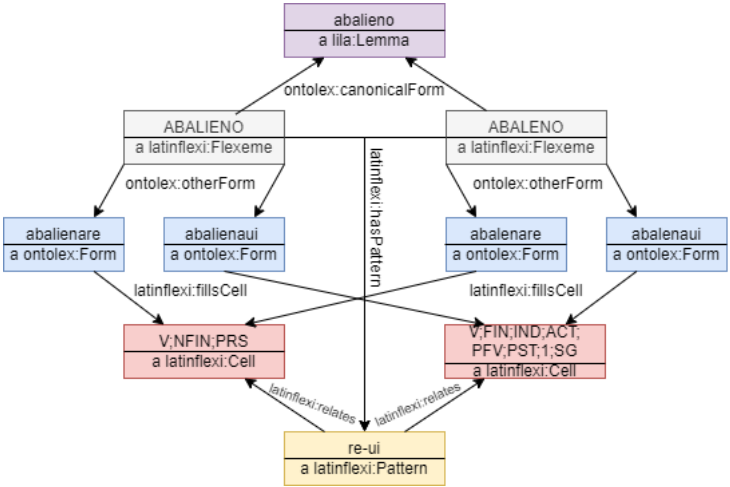
Architecture of the LatInFlexi ontology

- 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)
  → a `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

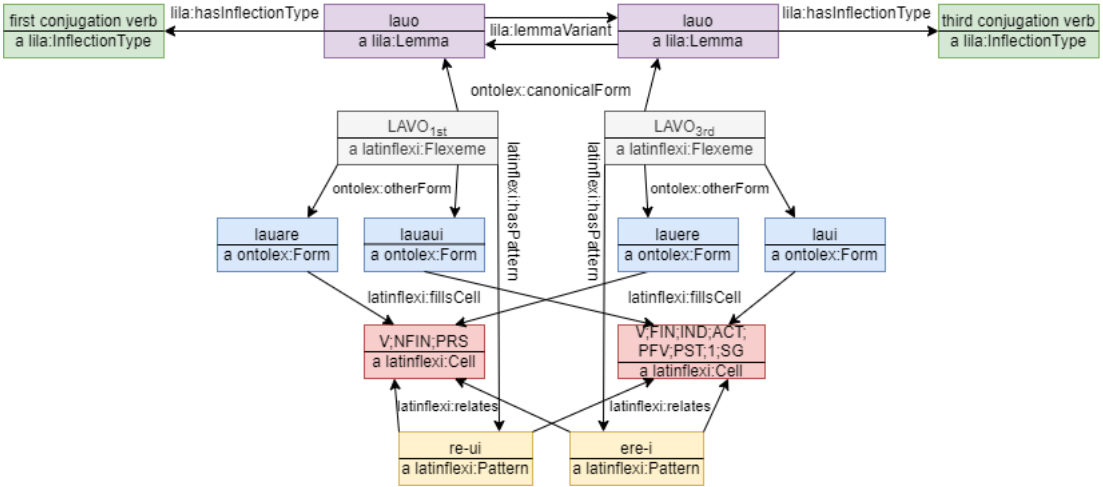
Example: LAUDO



Example: ABALIENO



Example: LAUO



Flexeme-Lemma mapping

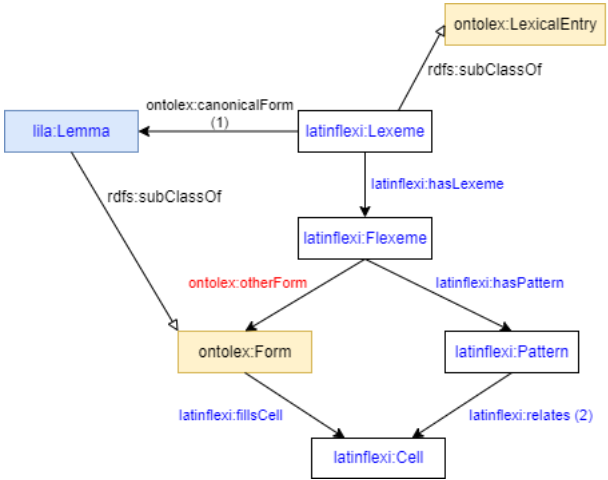
- Overabundance due to the availability of different stem allomorphs
  → different flexemes map to the same lemma
- Overabundance due to the possibility of different inflection class assignments
  → different flexemes map to different lemmas
  (due to modeling choices of the Lemma Bank)
  - Lexemic identity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

unit	n.
Forms	39,438
Flexemes	10,716
Lemmas	7,973

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

Outline

An alternative modeling choice: `Lexemes` as lexical entries



Alternative modeling choices: a comparison

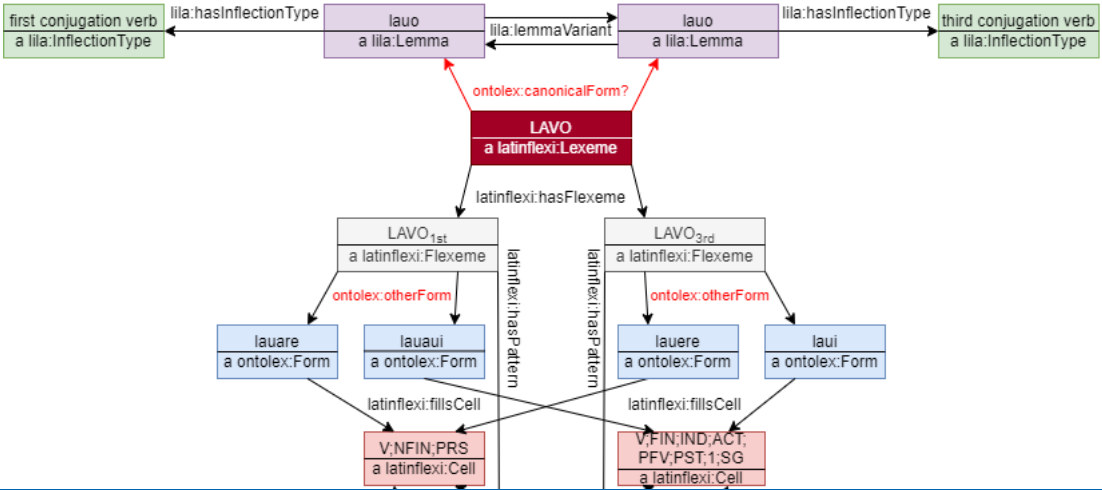
`Flexemes` as lexical entries

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

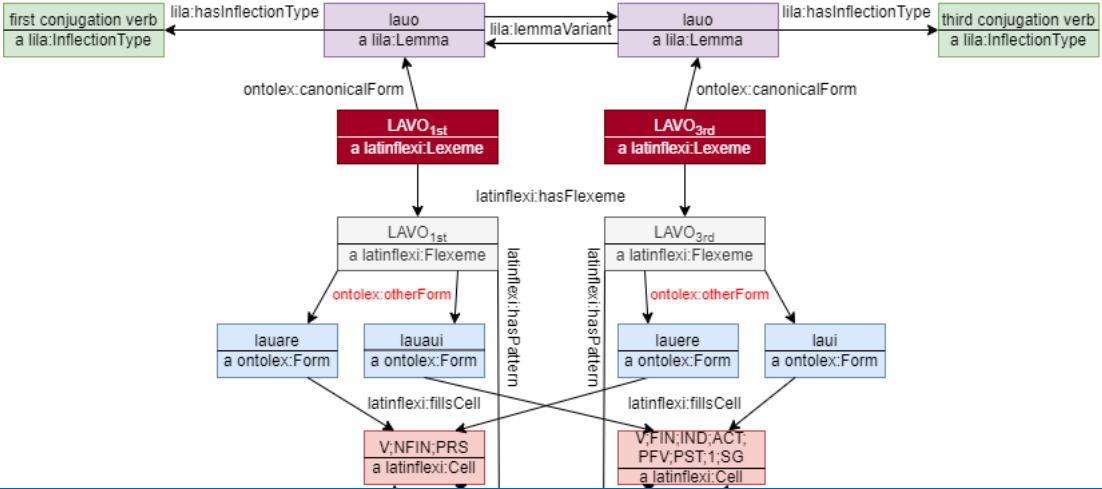
`Lexemes` as lexical entries

- ✓ 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:

ABUTOR 'consume'		MORIOR 'die'	
form	cell	form	cell
abutor	PRS.PASS.IND.1SG	*morior*	PRS.PASS.IND.1SG
abuti	PRS.PASS.INF	*mori*	PRS.PASS.INF
_	PRF.ACT.IND.1SG	_	PRF.ACT.IND.1SG
abusus	PRF.PTCP.NOM.M.SG	*mortuus*	PRF.PTCP.NOM.M.SG
_	FUT.PTCP.NOM.M.SG	*moriturus*	FUT.PTCP.NOM.M.SG

Troubles with flexemes

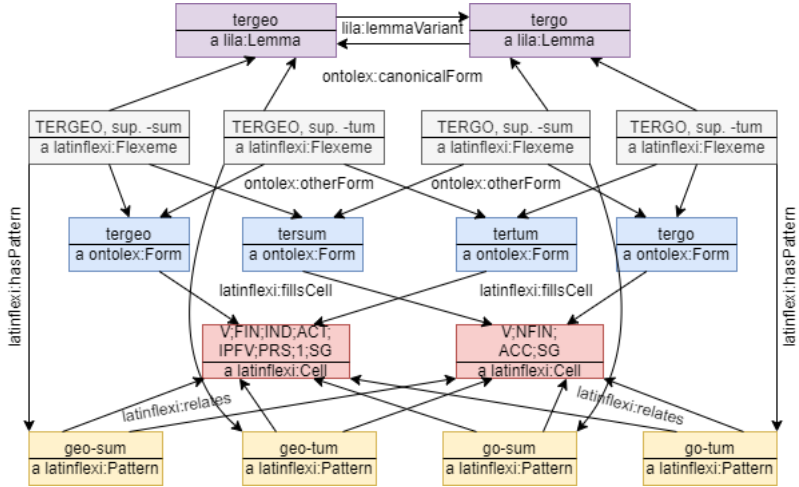
- We start from the strong assumption that all cases of overabundance should yield the introduction of distinct flexemes ⇒ the entries of our lexicon are never overabundant
- Theoretical issue: previous studies leave this question open (cf. Thornton 2018: 312-3, Bonami and Crysmann 2018: 198)
- Practical issue: in some cases, the identification of flexemes is not straightforward
- Cases of non-systematic overabundance across different cells → multiplication of flexemes

lexeme	PRS.ACT.IND.1SG	PRS.ACT.INF	SUP.ACC
TERG(E)O 'cleanse'	*tergeo* *tergo*	*tergēre* *tergere*	*tersum* *tertum*

→ 4 Flexemes

- The situation can be far more complex!

Example: TERG(E)O

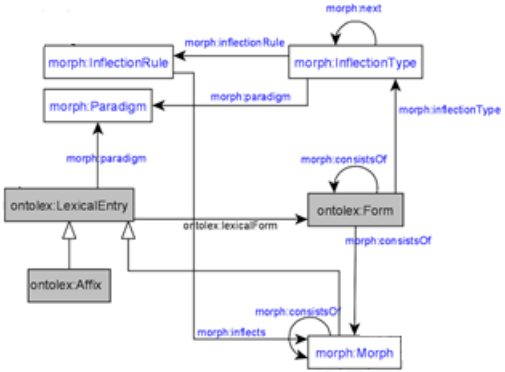


Phonological distinctions

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

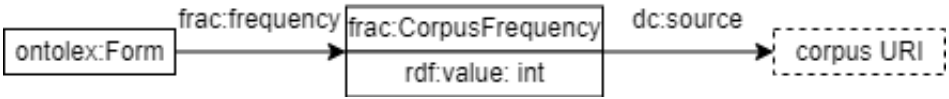


Overall (micro)class: storage or generation?

- 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`)

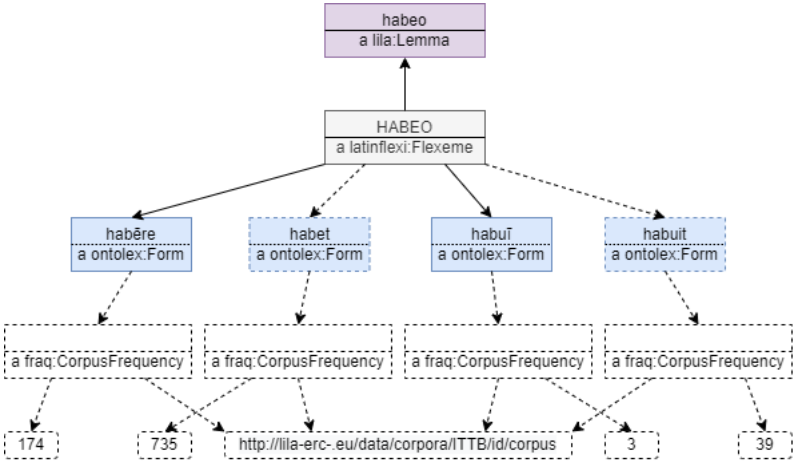
Corpus frequencies

- 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
    - [x] Index Thomisticus Treebank (Passarotti 2011)
    - [x] 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



Credits and acknowledgements

Claudia Corbetta, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Marco Passarotti, Martina Verdelli

Thanks!
Feedback?

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.

Works cited I

Bamman, David and Gregory Crane (2011). "The ancient Greek and Latin dependency treebanks. Selected Papers from the LaTeCH Workshop Series". In: *Language Technology for Cultural Heritage*. Ed. by Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou. Theory and Applications of Natural Language Processing. Berlin/Heidelberg, Germany: Springer, pp. 79–98.

Bauer, Laurie (1997). "Derivational paradigms". In: *Yearbook of morphology 1996*. Springer, pp. 243–256.

— (2019). "Notions of paradigm and their value in word-formation". In: *Word Structure* 12.2, pp. 153–175.

Beniamine, Sacha (2018). "Classifications flexionnelles. Étude quantitative des structures de paradigmes". PhD thesis. Université Sorbonne Paris Cité-Université Paris Diderot.

Bonami, Olivier and S. Beniamine (2016). "Joint predictiveness in inflectional paradigms". In: *Word Structure* 9.2, pp. 156–182.

Bonami, Olivier and Gilles Boyé (2014). "De formes en thèmes". In: *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*. Ed. by Florence Villoing, Sarah Leroy, and Sophie David. Paris: Presses Universitaires de Paris-Ouest, pp. 17–45.

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.

Bonami, Olivier and Jana Strnadová (2019). "Paradigm structure and predictability in derivational morphology". In: *Morphology* 29.2, pp. 167–197.

Booij, Geert (2010). "Construction morphology". In: *Language and linguistics compass* 4.7, pp. 543–555.

Budassi, Marco and Eleonora Litta (2017). "In Trouble with the Rules. Theoretical Issues Raised by the Insertion of -sc- Verbs into Word Formation Latin". In: *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, pp. 15–26.

Cecchini, Flavio M. et al. (2020). "UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works". In: *Seventh Italian Conference on Computational Linguistics*. Bologna: CEUR-WS.org, pp. 1–7. URL: `http://ceur-ws.org/Vol-2769/paper_14.pdf`.

Works cited II

Cecchini, Flavio Massimiliano, Timo Korkiakangas, and Marco Passarotti (May 2020). "A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 933–942. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.117.

Chiarcos, Christian, Christian Fäth, and Frank Abromeit (2020). "Annotation Interoperability for the Post-ISOCat Era". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5668–5677.

Cimiano, Philipp et al. (2020). *Linguistic Linked Data*. Springer.

Fradin, Bernard and Françoise Kerleroux (2003). "Troubles with lexemes". In: *Selected papers from the third Mediterranean Morphology Meeting*, pp. 177–196.

Georges, Karl Ernst (1998). *Ausführliches lateinisch-deutsches Handwörterbuch. Aus den Quellen zusammengetragen und mit besonderer Bezugnahme auf Synonymik und Antiquitäten unter Berücksichtigung der besten Hilfsmittel ausgearbeitet. Unveränderter Nachdruck der achten verbesserten und vermehrten Auflage*. Ed. by Heinrich Georges. Reprint of first edition of 1913–1918, Hannover, Germany: Hahnsche Buchhandlung. Darmstadt, Germany: Wissenschaftliche Buchgesellschaft. URL: http://www.zeno.org/Georges-1913.

Glare, Peter G. W. (2012). *Oxford Latin Dictionary*. Ed. by Alexander Souter. Oxford Languages. Oxford, UK: Oxford University Press. ISBN: 978-0-19-958031-6.

Gradenwitz, Otto (1904). *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Leipzig: Hirzel.

Haug, Dag Trygve Truslew and Marius Jøhndal (May 2008). "Creating a Parallel Treebank of the Old Indo-European Bible Translations". In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Ed. by Caroline Sporleder and Kiril Ribarov. Marrakesh, Morocco: European Language Resources Association (ELRA), pp. 27–34. URL: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf.

Kyjánek, Lukáš (2020). "Harmonisation of Language Resources for Word-Formation of Multiple Languages". MA thesis. Univerzita Karlova, Matematicko-fyzikální fakulta.

Works cited III

Lewis, Charlton Thomas and Charles Short (1879). *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph. D. and Charles Short, LL. D.* Ed. by E. A. Andrews. Oxford, UK: Clarendon Press. ISBN: 978-1-99-985578-9. URL: https://logeion.uchicago.edu.

Litta, Eleonora and Marco Budassi (2020). "What we talk about when we talk about paradigms: representing Latin word formation". In: *Paradigmatic relations in word formation*. Brill, pp. 128–163.

Mambrini, Francesco et al. (Sept. 2021). "Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin". eng. In: *Further with Knowledge Graphs*. Vol. 53. Studies on the Semantic Web, pp. 16–28. DOI: 10.3233/SSW210032. URL: https://zenodo.org/record/5482432 (visited on 09/07/2021).

Passarotti, Marco Carlo (2011). "Language resources. The state of the art of Latin and the Index Thomisticus treebank project". In: *Deuxiéme Colloque International ALIENTO*. ALIENTO, pp. 301–320.

Pounder, Amanda (2000). *Processes and Paradigms in Word-Formation Morphology*. DeGruyter.

Štekauer, Pavol (2014). "Derivational paradigms". In: *The Oxford handbook of derivational morphology*, pp. 354–369.

Stump, Gregory T. and Raphael A. Finkel (2013). *Morphological typology: From word to paradigm*. Cambridge: Cambridge University Press.

Sylak-Glassman, John (2016). *The composition and use of the universal morphological feature schema (UniMorph schema)*. Tech. rep. Department of Computer Science, Johns Hopkins University.

Thornton, Anna M (2018). "Troubles with flexemes". In: *The lexeme in descriptive and theoretical morphology* 4, p. 303.

Van Marle, Jaap (1984). *On the paradigmatic dimension of morphological creativity*. Foris.

Wilkinson, Mark D et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1, pp. 1–9.