The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources

Matteo Pellegrini, Eleonora Litta, Marco Passarotti,
Francesco Mambrini and Giovanni Moretti
{matteo.pellegrini}{eleonoramaria.litta}{marco.passarotti}
{francesco.mambrini}{giovanni.moretti}@unicatt.it

UNIVERSITÀ CATTOLICA del Sacro Cuore

CIRCSE

3rd Workshop on Resources and Tools for Derivational Morphology | September 9-10, 2021

erc LiLa Linking Latin

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

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

Discussion and Conclusions

- Discussion
- Conclusions

Outline

Background

(Linguistic) Linked Data

LiLa: Linking Latin

Word Formation in LiLa

The Word Formation Latin resource (WFL)

Word Formation in the Lemma Bank

Including WFL into the Knowledge Base

Discussion and Conclusions

Discussion

Conclusions

Background

For Latin (and for many other languages) a wealth of electronic resources and tools have been developed in the last decades

- Linguistic resources
    - Textual resources (corpora)
    - Lexical resources (dictionaries, lexicons, etc.)
- NLP tools (morphological analysers, PoS-taggers, etc.)

Such resources and tools are often characterised by different conceptual and structural models, which makes secondary reuse difficult

Going FAIR

Data should be:

- Findable
- Accessible
- Interoperable
- Reusable

Mark D. Wilkinson *et al.*
The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data, 3, 2016

(Linguistic) Linked Data

Tim Berners-Lee’s principles of Linked Data

- Use URIs for things
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data
- Include links to other URIs

Application to language data → **Linguistic Linked Open Data** cloud

Philipp Cimiano, Christian Chiarcos, John P. McCrae, Jorge Gracia
Linguistic Linked Data. Representation, Generation and Applications
Springer, 2020

Outline

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

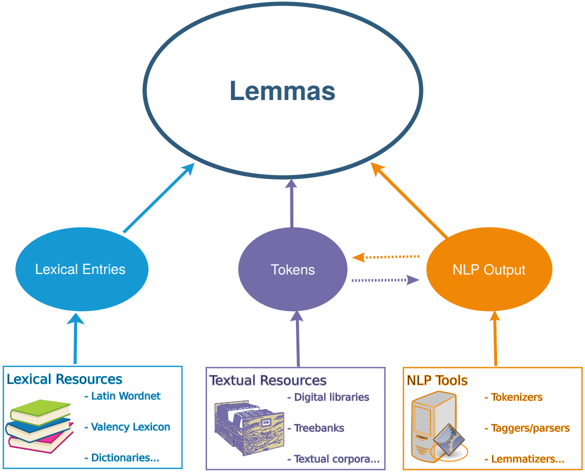
Discussion and Conclusions

- Discussion
- Conclusions

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



Outline

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

Discussion and Conclusions

- Discussion
- Conclusions

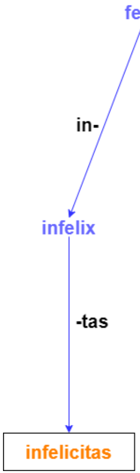
Word Formation Latin (WFL)

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)

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

The hierarchical structure of WFL

- **Hierarchical structure**, representable with a **directed tree-graph**



Outline

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

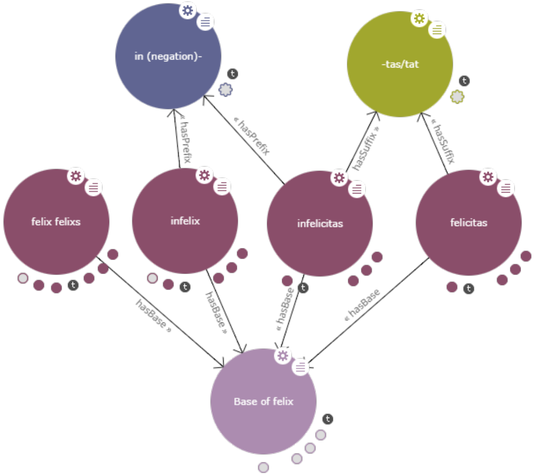
- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

Discussion and Conclusions

- Discussion
- Conclusions

Word Formation in the Lemma Bank

- The Lemma Bank includes only a selection of the derivational information provided by WFL: each lemma is connected to the affixes it displays and to its **base**
- **Flat structure**



The flat structure of derivational information in the Lemma Bank

- The choice of this flat organisation is due to its compatibility with more recent, Word-and-Paradigm theoretical approaches, like Construction Morphology
- Furthermore, it allows for a more natural treatment of cases that were problematic for the rigidly hierarchical structure of WFL
  - Directionality issues in conversion: ADVERSARIUS$_{A}$ ‘opposed’ ↔ ADVERSARIUS$_{N}$ ‘opponent’?
  - Parasynthetic formations: AQUA ‘water’ → EXAQUESCO ‘become water’ (*AQUESCO/*EXAQUO)
- However, this means that a lot of potentially useful information of WFL is not represented in the Lemma Bank

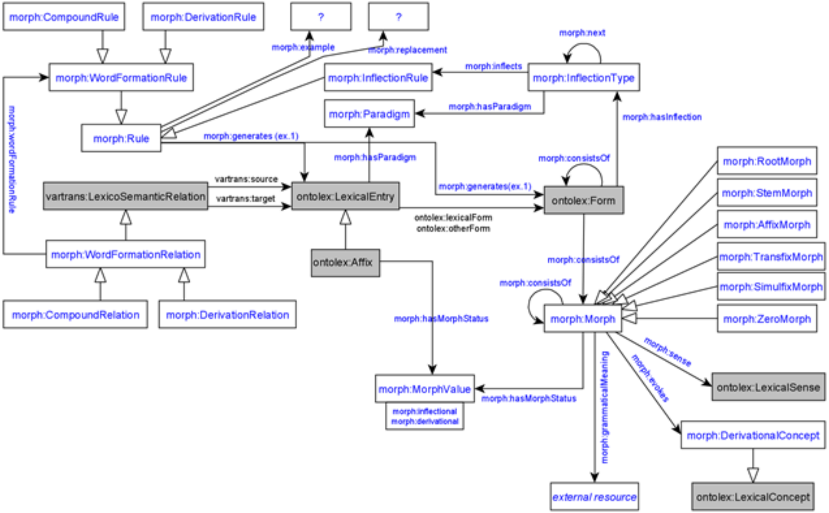
Outline

- Background
    - (Linguistic) Linked Data
    - LiLa: Linking Latin
- Word Formation in LiLa
    - The Word Formation Latin resource (WFL)
    - Word Formation in the Lemma Bank
    - Including WFL into the Knowledge Base
- Discussion and Conclusions
    - Discussion
    - Conclusions

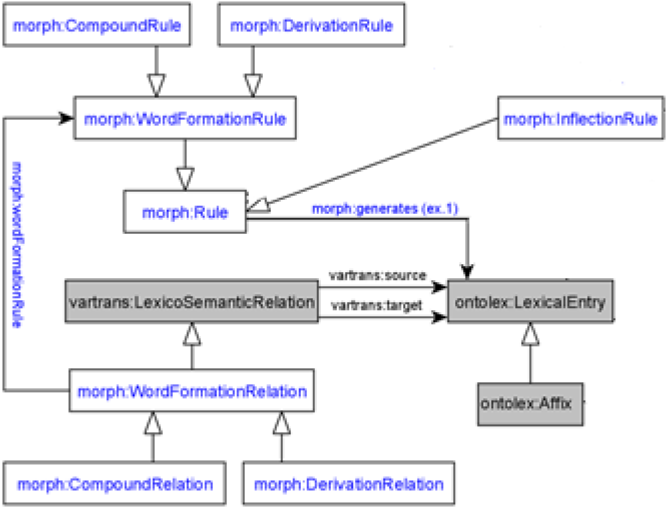
Including WFL into the Knowledge Base

- Modellisation 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

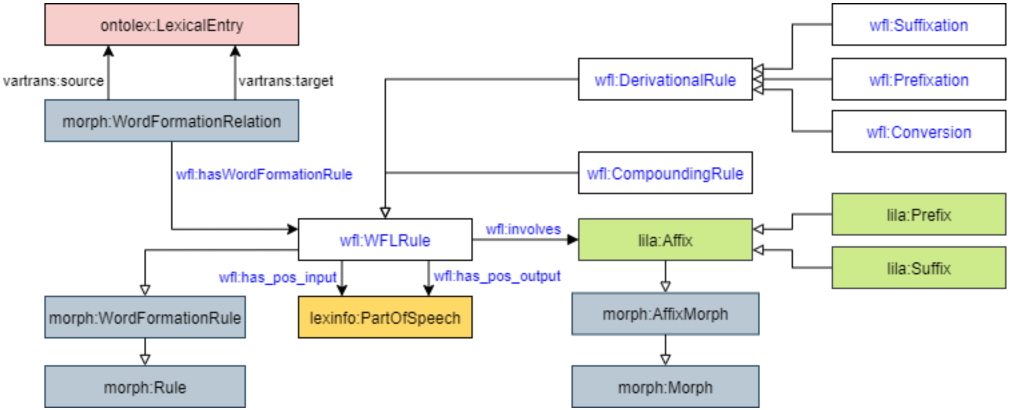
Architecture of the OntoLex Morphology module



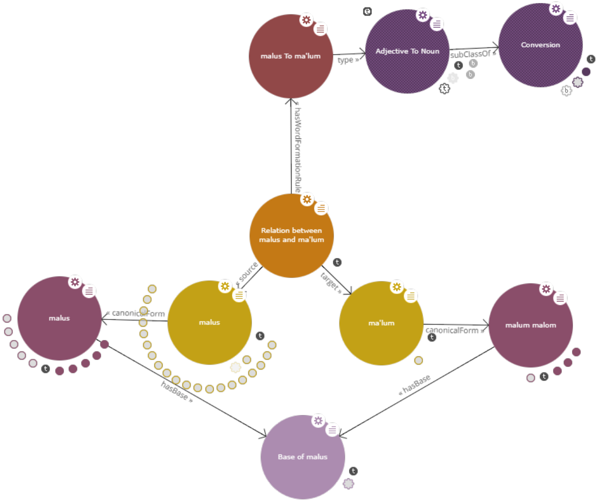
Architecture of the OntoLex Morphology module



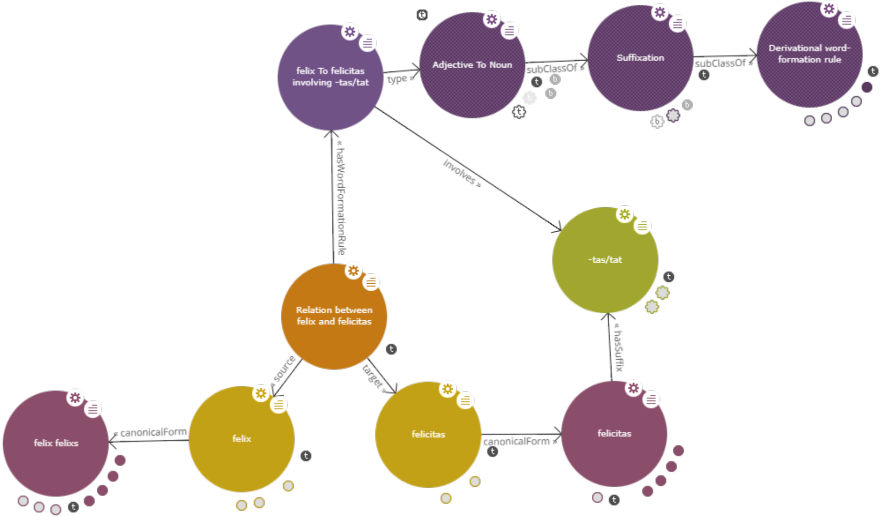
Architecture of the WFL ontology



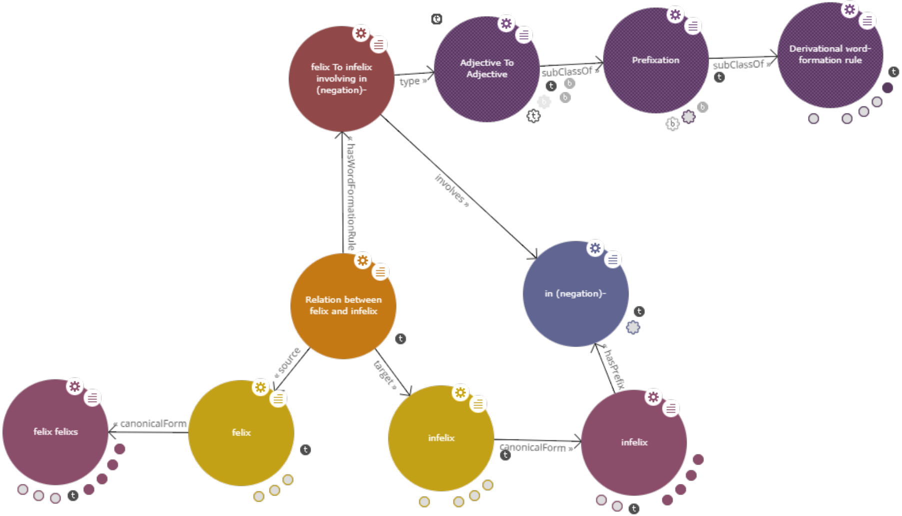
Treatment of conversion in the WFL ontology



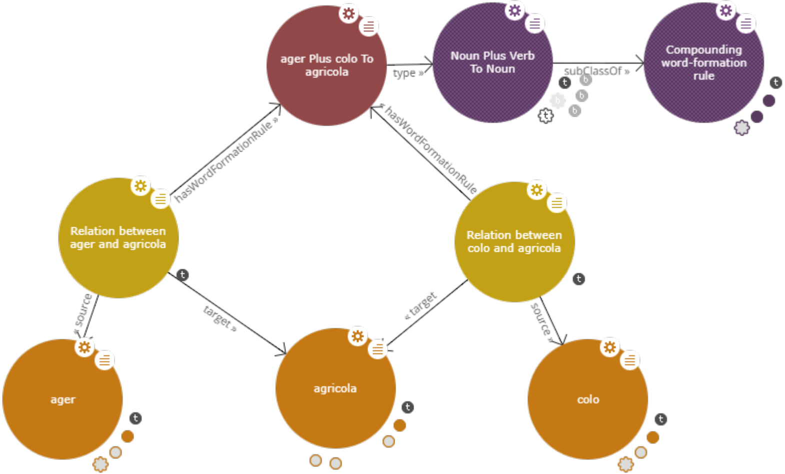
Treatment of suffixation in the WFL ontology



Treatment of prefixation in the WFL ontology



Treatment of compounding in the WFL ontology



Outline

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

Discussion and Conclusions

- Discussion
- Conclusions

Discussion

Different approaches in resources specialised in word formation:

- morpheme-oriented
- **lexeme-oriented → WFL**
- **family-oriented → word formation in the Lemma Bank**
- paradigm-oriented

Lukáš Kyjánek
Harmonisation of Language Resources for Word-Formation of Multiple Languages
2020

Outline

Background

- (Linguistic) Linked Data
- LiLa: Linking Latin

Word Formation in LiLa

- The Word Formation Latin resource (WFL)
- Word Formation in the Lemma Bank
- Including WFL into the Knowledge Base

Discussion and Conclusions

- Discussion
- Conclusions

Conclusions

Both approaches to the organisation of derivational information have their merits

- Lexeme-oriented, hierarchical structure of WFL:
  - allows to focus on smaller, more tightly connected sub-sections of word formation families
  - allows to extract only lexemes that are formed by means of a specific WFR
- Family-oriented, flat structure of derivational information in the Lemma Bank:
  - allows to easily extract all the lexemes that display a given affix, regardless of its position and/or order of insertion in the derivational history

The adoption of Linked Data standards makes both approaches available within a **unified framework**

Thanks!

Get in touch

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini and Giovanni Moretti
[1] CIRCSE, Università Cattolica del Sacro Cuore

{matteo.pellegrini}{eleonoramaria.litta}{marco.passarotti}{francesco.mambrini}{giovanni.moretti}@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.