Interoperable Words

Interlinking Lexical (and Textual) Resources for Latin in the LiLa Knowledge Base

Marco Passarotti

UNIVERSITÀ CATTOLICA del Sacro Cuore

IRCSE

Eighth Conference on Electronic Lexicography in the 21st Century
(eLex 2023)
Brno, June 27–29, 2023

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.

Overview

Introduction and Fundamentals

- LiLa: Mission and Architecture

LiLa now!

- Lemma Bank and Lexical Resources
- Services and Tools

Conclusion

- State of Affairs
- Closing Remarks

Table of Contents

Research question
State of affairs

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

Research question
State of affairs

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

- Textual Resources

Research question
State of affairs

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

- Textual Resources
- Lexical Resources

Research question
State of affairs

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

- Textual Resources
- Lexical Resources
- NLP Tools

Research question
State of affairs

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

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

ERC Consolidator Grant
2018-2023

A collection of multifarious, interoperable linguistic resources described with the same vocabulary for knowledge description (by using common data categories and ontologies)

Interlinking as a Form of Interaction

The Linked Data Principles

…just to be FAIR

The Linked Data Principles
...just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)

The Linked Data Principles
...just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL
- Include links to other URIs

LiLa Knowledge Base

Lexically-based architecture and (meta)data sources

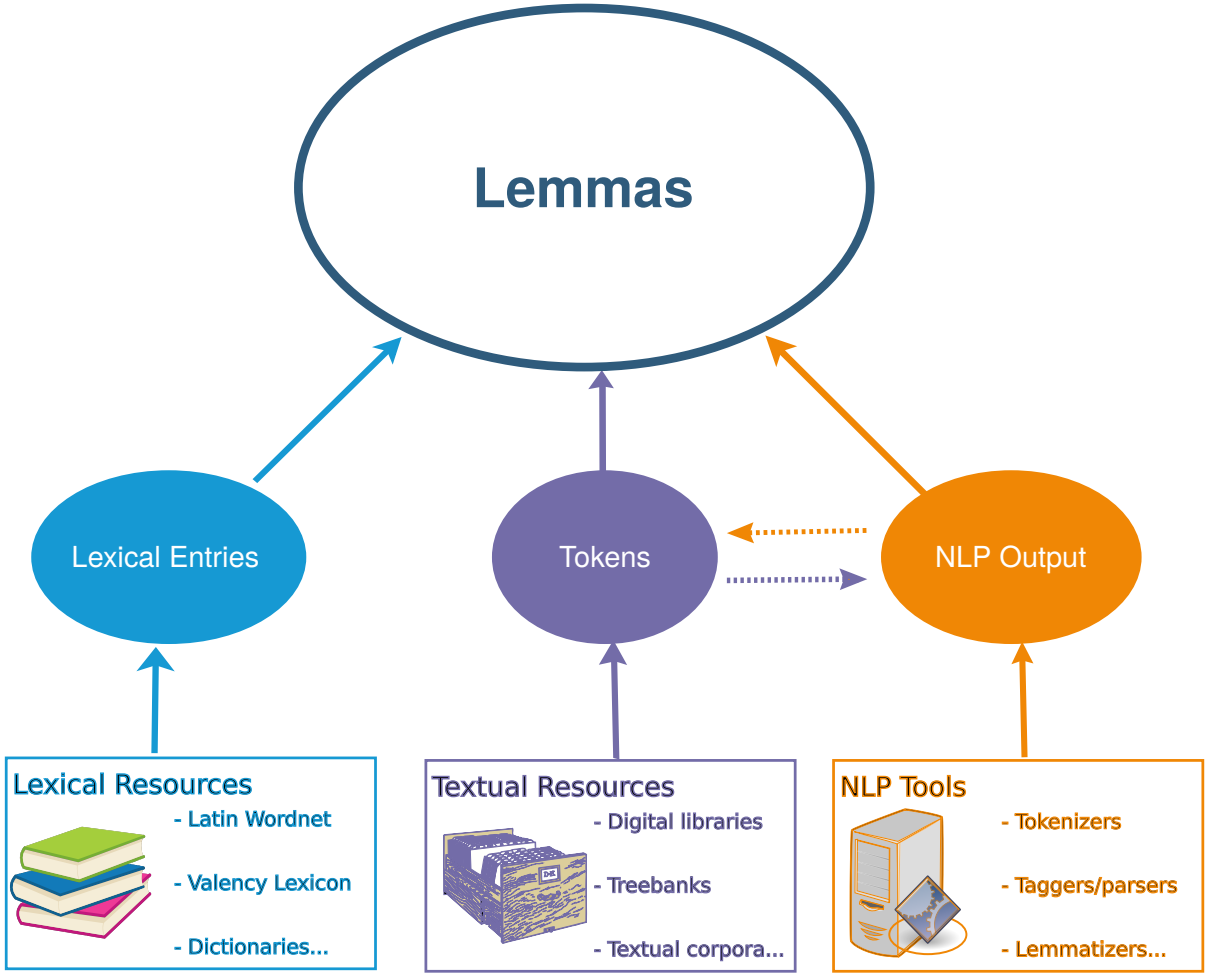
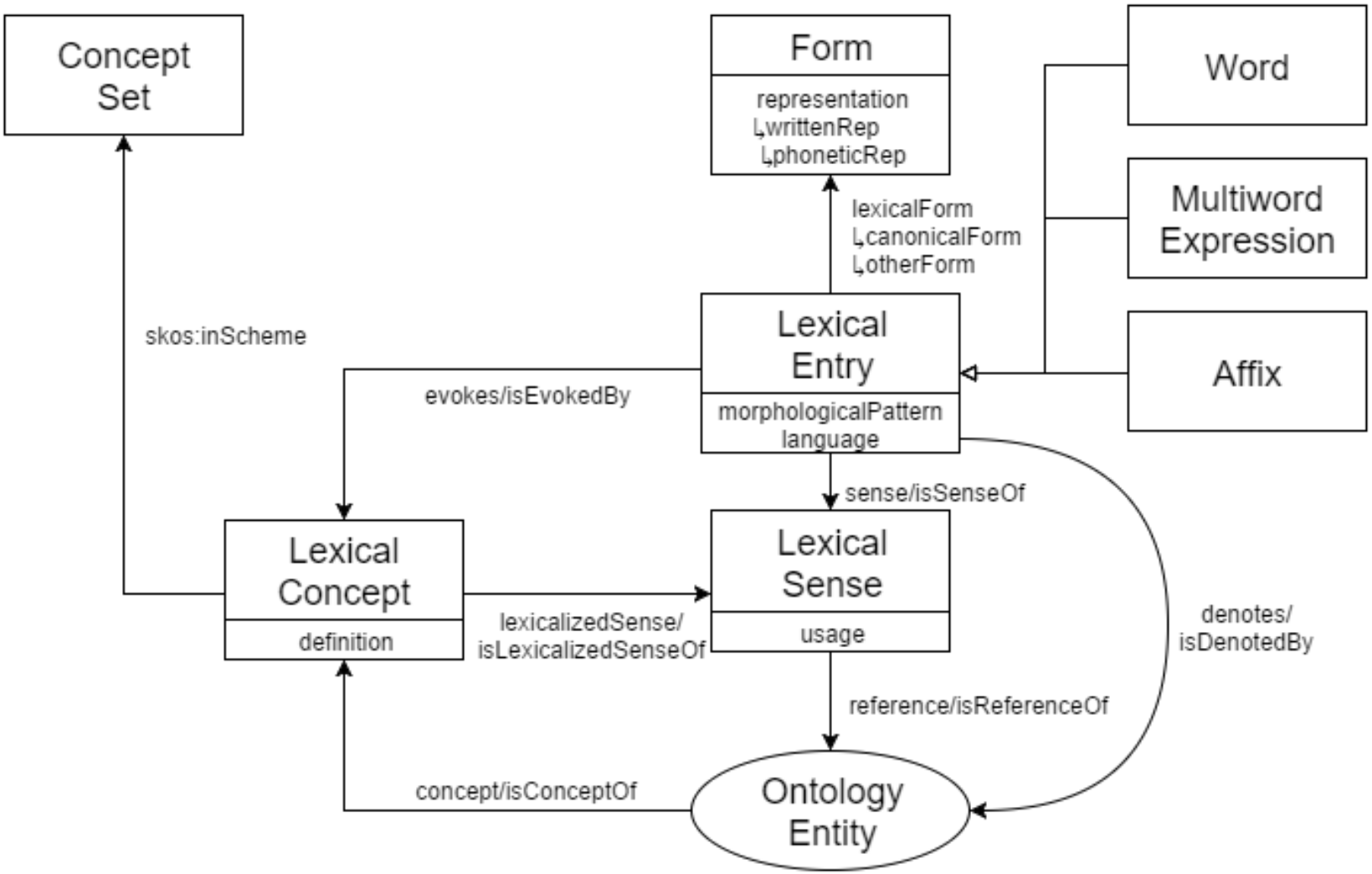


Table of Contents

LiLa and Ontolex Lemon

A *de facto* W3C standard for publishing lexical data as LLOD



Lemma Bank and Lexical Resources

The core of LiLa. Source: LEMLAT (extended)

Lemma *admiror* 'to admire'

`http://lila-erc.eu/data/id/lemma/87541`

- Lemma Bank
- A bilingual Latin English dictionary (Lewis & Short)
- A derivational lexicon (Word Formation Latin)
- A polarity lexicon (LatinAffectus)
- An etymological dictionary (De Vaan)
- A valency lexicon (Latin Vallex)
- A manually checked subsection of the Latin WordNet
- …and many others…

Table of Contents

Lemma Bank Query Interface

https://lila-erc.eu/query/

SPARQL Access Point

https://lila-erc.eu/sparql/

LiLa Search Platform

http://lila-erc.eu:8080/lila-lisp/

TextLinker

TextLinker

http://lila-erc.eu:8080/LiLaTextLinker/

Table of Contents

Linguistic Resources Interlinked in LiLa

...and some numbers

Resources connected so far

`https://lila-erc.eu/data-page/`

- **Canonical Forms** in the Lemma Bank: approx. 215K
- **Lexical Entries** in Lexical Resources: approx. 145K
- **Tokens** in Textual Resources (158 works): approx. 3,5M
- **Triples**: approx. 70M

Table of Contents

Closing Remarks

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)
  - The role of data (and Big Data): unsupervised learning to build lexical resources based on distributional semantics (embeddings) and Large Language Models. This affects the way textual resources (the weaker role of annotation), lexical resources (definitions, translations, synonyms etc.) and NLP tools are built

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)
  - The role of data (and Big Data): unsupervised learning to build lexical resources based on distributional semantics (embeddings) and Large Language Models. This affects the way textual resources (the weaker role of annotation), lexical resources (definitions, translations, synonyms etc.) and NLP tools are built
- But **resources still remain**: unity (and interoperability) is strength!

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)
  - The role of data (and Big Data): unsupervised learning to build lexical resources based on distributional semantics (embeddings) and Large Language Models. This affects the way textual resources (the weaker role of annotation), lexical resources (definitions, translations, synonyms etc.) and NLP tools are built
- But **resources still remain**: unity (and interoperability) is strength!
  - Should electronic lexicography in the 21st century be based on interoperable resources (maybe, following the Linked Data principles)? **BUT** LiLa is a fortunate and successful case: publishing linguistic resources as LOD takes money, time and expertise. We must facilitate wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)
  - The role of data (and Big Data): unsupervised learning to build lexical resources based on distributional semantics (embeddings) and Large Language Models. This affects the way textual resources (the weaker role of annotation), lexical resources (definitions, translations, synonyms etc.) and NLP tools are built
- But **resources still remain**: unity (and interoperability) is strength!
  - Should electronic lexicography in the 21st century be based on interoperable resources (maybe, following the Linked Data principles)? **BUT** LiLa is a fortunate and successful case: publishing linguistic resources as LOD takes money, time and expertise. We must facilitate wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
  - Using interlinked resources to fine-tune LLMs

Closing Remarks

- **Topic of this edition of eLex**: *Invisible lexicography: everywhere lexical **data** is used without users realizing they make use of a "dictionary"*
  - Everything deals with words: strict connection between words in lexical resources and words in textual resources (ideally, in a virtuous circle)
  - The role of data (and Big Data): unsupervised learning to build lexical resources based on distributional semantics (embeddings) and Large Language Models. This affects the way textual resources (the weaker role of annotation), lexical resources (definitions, translations, synonyms etc.) and NLP tools are built
- But **resources still remain**: unity (and interoperability) is strength!
  - Should electronic lexicography in the 21st century be based on interoperable resources (maybe, following the Linked Data principles)? **BUT** LiLa is a fortunate and successful case: publishing linguistic resources as LOD takes money, time and expertise. We must facilitate wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
  - Using interlinked resources to fine-tune LLMs
  - **Beyond Latin**: portability of the LiLa architecture in infrastructures

Thank you

Get in touch!

LiLa: Linking Latin
Università Cattolica del Sacro Cuore
CIRCSE Research Centre

info@lila-erc.eu

https://github.com/CIRCSE

https://lila-erc.eu

@ERC_LiLa

Largo Gemelli 1, 20123 Milan, Italy

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.