Of Families and Occurrences

Derivation and Word Usage in Latin

Marco Passarotti and **Eleonora Litta**

marco.passarotti,eleonoramaria.litta@unicatt.it

DeriMo 2023 | Dubrovnik | 6th October 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.

Table of Contents

Introduction

LiLa: Linking Latin

Word Formation Latin

Using LiLa

Hypothesis evaluation

Discussion

Conclusion

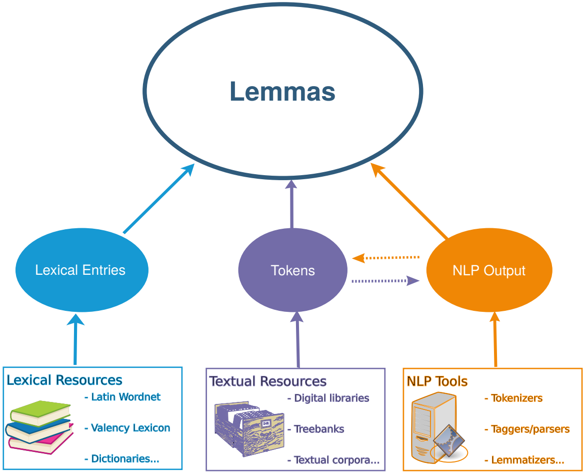
subsection

LiLa: Linking Latin

Some background

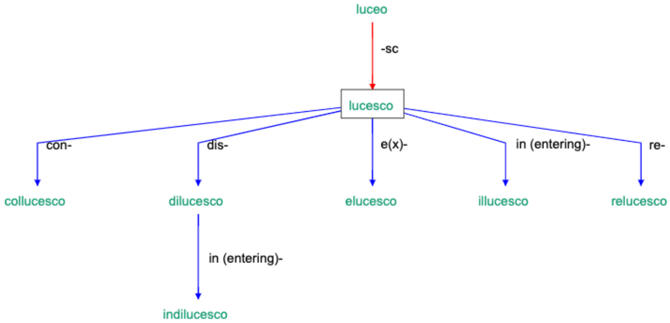
- Open-ended **Knowledge Base** of interoperable linguistic resources for Latin sharing a common vocabulary for knowledge description
- Use of **web standards** to represent and query data, following the **Linked Data** principles
  - RDF: information is coded in terms of **triples**, connecting a **subject**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



Word Formation Latin (WFL)

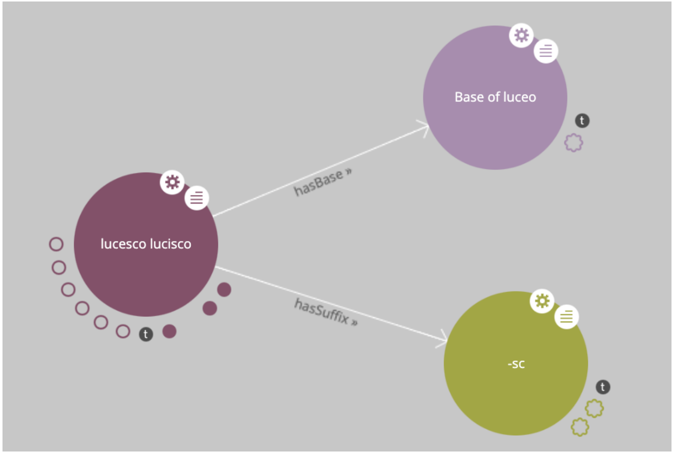
- **Hierarchical structure**, represented through a **directed tree-graph**



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**
- LiLa ontology in two classes:
  - `Affix`
    - `Prefix`
    - `Suffix`
  - `Base`
- Properties:
  - `lila:hasSuffix`
  - `lila:hasPrefix`
  - `lila:hasBase`
- **Flat structure**

Word Formation in LiLa



The Hypothesis

- Given a **derivational family** (set of words sharing the same ancestor/root), the member with the **highest number of occurrences** in texts is **derivationally simple**, and more typically is the root of the family.

The Query

- Selection of families (at least $\geq$ 10 members)= 1,086 families (in-degree via property lila:hasBase of $\geq$10)
- Total number of occurrences in textual resources linked to LiLa $\geq$ 100 = 878 families

The Results

Table: Families with $\geq$ 100 occurrences in LiLa texts

No. of families	Most frequent word
582	root
296	non-root
89	zero-affix
207	1 or more affixes

The Results

Affix	Number of families	Lemma Bank ranking	Example
con-	25	3	cognosco
-i	22	11	substantia
-id	11	36	frigidus
-or	11	4	calor
de-	11	9	detrimentum
ad-	10	10	accipio
-in	9	19	dominus
ex-	9	5	exsulto
in(entering)-	9	8	instruo
-(t)io	8	1	oratio

Table: The 10 most attested affixes in the most frequent derived words of a family.

The Results

Affix	Number of families	Lemma Bank ranking	Example
con-	25	3	cognosco
-i	22	**11**	substantia
-id	11	36	frigidus
-or	11	4	calor
de-	11	9	detrimentum
ad-	10	10	accipio
-in	9	19	dominus
ex-	9	5	exsulto
in(entering)-	9	8	instruo
-(t)io	8	**1**	oratio

Table: The 10 most attested affixes in the most frequent derived words of a family.

The Results

Ranking	*-i* set	*-(t)io* set
1	consilium (2,147)	ratio (3,513)
2	gratia (2,051)	oratio (1,250)
3	substantia (1,697)	opinio (504)
4	sententia (1,606)	fornicatio (179)
5	memoria (1,039)	satisfactio (175)

Table: The 5 most frequent words in the *-i* and *-(t)io* sets.

Lexicalisation

> *“lexicalisation […] concerned with those signs which […] are handled holistically […] to directly grasp the whole without consideration of the parts”*
>
> – Lehmann (2002), p. 1-2

Lexicalisation

> *“lexicalisation […] concerned with those signs which […] are handled holistically […] to directly grasp the whole without consideration of the parts”*
>
> – Lehmann (2002), p. 1-2

↬ substantia "the quality of being real" < substo "to hold one’s ground" + ia = spacial semantic field is lost in lexicalisation: the meaning of the word does not correspond to the sum of its parts.

POS distribution

PoS	Root	Most frequent
adjective	133	114
common noun	364	415
verb	351	291

Table: PoS distribution of root words and most frequent words in derivational families.

Comparing Corpora

- LiLa Corpora: **diverse** periods and genres ↬ **representative** set of data to draw conclusions

Comparing Corpora

- LiLa Corpora: **diverse** periods and genres ↬ **representative** set of data to draw conclusions
- **LASLA** (Classical, 1.7m words) vs **ITTB** (Medieval, 350k words)

Comparing Corpora

- LiLa Corpora: **diverse** periods and genres ↬ **representative** set of data to draw conclusions
- **LASLA** (Classical, 1.7m words) vs **ITTB** (Medieval, 350k words)
- 214 fam with more than 100 members common to both corpora
- 116 of these have the same more frequent word
- 89 have a different most frequent word
- 34 fam with different most fequent word ↬ not root

Lexical differences

Root	Most frequent in LASLA	Most frequent in ITTB
facio	facio	facio
fero	fero	differentia
capio	accipio	principium
ago	ago	actus
verto	versus	universalis
gero	gero	NA
pes	pes	impedio
lego	legio	intellectus
eo	eo	transeo
fluo	flumen	NA

Table: Most frequent word of the 10 largest families in the LASLA and ITTB corpora.

Conclusions

...and Future Work

- Interoperability between resources has proved useful in our investigation
- Exploit the evidence we collected to explore trends (e.g. are conversions always recorded the right way around in dictionaries?)
- Interoperability between languages would be helpful, Latin could play an important role at least for Romance languages

Thanks!

Get in touch

Full name[1] and **Full name**[2]

[1] CIRCSE, Università Cattolica del Sacro Cuore

email1,email2

@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

A. Author.
Introduction to Giving Presentations.
Klein-Verlag, 1990.

S. Someone.
Title of work.
Journal of This and That, 2(1):50–100, 2000.

Credits and acknowledgements

Thank you, Grazie, Gracias, …

Backup slides

It is useful to add slides at the end of your presentation to refer to during audience questions.