Latin Lemmatization & POS Tagging

Issues, Resources, Tools

Francesco Mambrini, Marco Passarotti

IANLS Vacation School: Digital Humanities and Neo-Latin Studies
University of Bonn

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

Lemmatization & Part-of-Speech Tagging

Textual Resources

- Lemmatized Corpora for Latin

Tools and Hands-on

- Tools for lemmatization & POS Tagging
- Try it yourself on the provided text(s)!

Latin in the Semantic Web

- The LiLa Knowledge Base
- The TextLinker

Querying Lemmatized Resources

- Latin Treebanks in Universal Dependencies
- Interlinked Lexical and Textual Resources in LiLa

Goals

Lemmatization and **part-of-speech tagging** (POS-tagging) aim to **abstract** some linguistic properties to allow **form-invariant** reference to types/tokens.

?! How can I retrieve all the occurrences of a word in a text?

?! How can I know which (morphosyntactic) function(s) a word plays in a text?

Lemmatization & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

Different word forms in different contexts…

- …*his rebus cognitis Caesar Gallorum animos* ***verbis*** *confirmavit*…
  → ablative plural (token); dative & ablative plural (type)
- …*quod ego si* ***verbo*** *adsequi possem*…
  → ablative singular (token); dative & ablative singular (type)
- …*ne more iuvencae mugiat, et timide* ***verba*** *intermissa retemptat*…
  → accusative plural (token); nominative, accusative & vocative plural (type)

…but all can be referred to a canonical/standardized citation form (Lemma):

⇒ ***uerbum***
→ nominative singular of neuter II. declension noun

What about *cognitis*, *intermissa* and *timide*?

Lemmatization

Type-based the process of assigning each type (in a text) to one, or more lemma(s)
Token-based the process of assigning each token in a text to a lemma

Different lexicographic criteria:

- inflectional morphology: same paradigm, same lemma? what about participles?
- graphical representation: *voluptas* vs. *uoluptas*
- spelling: *sulphur* vs. *sulfur*
- ending and inflectional type: *diameter* vs. *diametros* vs. *diametrus*
- paradigmatic slot for the lemma: *sequor* vs. *sequo* (see Du Cange: infinitives used)
- homographs: *occido/[caedo|cado]* vs. *occido[1|2]*

Lemmatization & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

Words can play different (morphosyntactic) functions in sentences:

- ⋆ ***supra***
  - ▶ *. . . ager trecentis aut etiam* ***supra*** *nummorum milibus emptus. . .*
    - → adverb (ADV)
  - ▶ *. . . ille qui* ***supra*** *nos habitat. . .*
    - → preposition (ADP)
- ⋆ ***scribo***
  - ▶ *. . . atque in Thesauro* ***scripsit*** *causam dicere prius unde petitur. . .*
    - → verb (VERB)
- ⋆ ***elephantus***
  - ▶ *. . .* ***elephanto*** *beluarum nulla prudentior. . .*
    - → noun (NOUN)

These functions are predictable and come from a rather small set of alternatives.

Part-of-speech tagging

Type-based: the process of assigning each type one, or more morphosyntactic **function(s)**, i. e. parts of speech, from a given set
Token-based the process of assigning each token in a text one morphosyntactic **function**, i. e. part of speech, from a given set

Current standard de facto tagset: **Universal Dependencies**

16+1 classes: ADJ (adjectives), ADP (pre- & postpositions), ADV (adverbs), AUX (auxiliaries), CCONJ & SCONJ (co-ordinating & subordinating conjunctions), DET (determiners), INTJ (interjections), NOUN & PROPN (common & proper nouns), NUM (numerals), PART (particles), PRON (pronouns), VERB (verbs), SYM (symbols), X (other) + PUNCT (punctuation)

`https://universaldependencies.org`

Lemmatization & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

Type-based vs. token-based POS tagging:

- Every ADJ can be NOUN
- Every ADP, CCONJ, SCONJ etc. can be NOUN (like in metalinguistic discourse)
- Every VERB can be NOUN

One or more part-of-speech? Which part-of-speech?

- *italicus*: ADJ? NOUN? PROPN?
- *ubi*: ADV? SCONJ?
- *non*: ADV? PART?
- *aliqui*: PRON? DET? ADJ?

Table of Contents

Lemmatized Corpora for Latin

- LASLA Corpus
- Index Thomisticus
- Computational Historical Semantics
- 5 Latin Treebanks in UD
- CLaSSES
- …and others

Table of Contents

LEMLAT

Type-based: https://github.com/CIRCSE/LEMLAT3


```
Analysed wordform : sulphur

============================ANALYSIS    ================================

SEGMENTATION:   sulphur

--------------------morphological feats 1 --------------------------
--nns--

Case:   Nominative
Gender: Neuter
Number: Singular
--------------------morphological feats 2 --------------------------
--ans--

Case:   Accusative
Gender: Neuter
Number: Singular
--------------------morphological feats 3 --------------------------
--vns--

Case:   Vocative
Gender: Neuter
Number: Singular
        ============================LEMMA ================================
        sulpur                          N3B  s3429 n
        ----------------------morphological feats-----------------------
        NcC

        PoS:    Noun
        Type:   Common
        Inflexional Category:   III decl
```

UDPipe

Token-based: `https://lindat.mff.cuni.cz/services/udpipe/`


```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipe_model = latin-proiel-ud-2.6-200830
# udpipe_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# text = Cui dono lepidum novum libellum arida modo pumice expo
1    Cui    qui    PRON    Pr    Case=Dat|Gender=Masc|Number=Si
2    dono    donum    NOUN    Nb    Case=Abl|Gender=Neut|Nu
3    lepidum    lepidus    ADJ    A-    Case=Acc|Degree=Pos|Gender=
4    novum    novus    ADJ    A-    Case=Acc|Degree=Pos|Gender=
5    libellum    libellus    NOUN    Nb    Case=Acc|Gender=Masc|N
6    arida    aridus    ADJ    A-    Case=Acc|Degree=Pos|Gender=
7    modo    modo    ADV    Df    _    8    advmod    _    TokenR
8    pumice    pumic    NOUN    Nb    Case=Abl|Gender=Masc|Nu
9    expolitum?    expolio    VERB    V-    Case=Nom|Gender=M
SpaceAfter=No|TokenRange=51:61
```

UDPipe

Token-based: `https://lindat.mff.cuni.cz/services/udpipe/`


```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipe_model = latin-evalatin20-200830
# udpipe_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# text = Cui dono lepidum novum libellum arida modo pumice expolitum?
1	Cui	qui	PRON	_	_	_	_	_	TokenRange=0:3
2	dono	donum	NOUN	_	_	_	_	_	TokenRange=4:8
3	lepidum	lepidus	ADJ	_	_	_	_	_	TokenRange=9:16
4	novum	novus	ADJ	_	_	_	_	_	TokenRange=17:22
5	libellum	libellus	NOUN	_	_	_	_	_	SpacesAfter=\r\n|TokenRange=23:31
6	arida	aridus	ADJ	_	_	_	_	_	TokenRange=33:38
7	modo	modo	ADV	_	_	_	_	_	TokenRange=39:43
8	pumice	pumicus	NOUN	_	_	_	_	_	TokenRange=44:50
9	expolitum?	expolito	VERB	_	_	_	_	_	SpaceAfter=No|TokenRange=51:61
```

TreeTagger

Token-based: HMM + Binary Decision Tree

- Download the tool from `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
- Prepare a txt file with a Latin text
- Tokenize the file and prepare the input (one-word-per-line):

```
cd treetagger/cmd
perl utf8-tokenize.perl INPUT-FILE.txt >
OUTPUT-FILE.txt
```

TreeTagger

Token-based: HMM + Binary Decision Tree

- `cd ../bin`
- Linux/Mac:
  `./tree-tagger <parameter-file> <input-file> <output-file> -token -lemma`
  Example (download a parameter file for Latin and put it into the 'bin' folder): `./tree-tagger latin.par input.txt output.txt -token -lemma`
- Windows:
  `tag-LANGUAGE.bat <input-file> <output-file>`
  Example: `tag-latin.bat input.txt output.txt`

Further Tools
All Token-based

- Collatinus Web: `https://outils.biblissima.fr/en/collatinus-web/`
- Deucalion: `https://dh.chartes.psl.eu/deucalion/latin`
- Stanza: three models for Latin. `https://stanfordnlp.github.io/stanza/available_models.html`
- Morpheus: `https://github.com/PerseusDL/morpheus`
- Whitaker's Words: `https://latin-words.com`

Table of Contents

Table of Contents

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

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats
- Different representation schemes, query languages, annotation criteria and tagsets

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF, SPARQL

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF, SPARQL
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF, SPARQL
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from physically separated repositories

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF, SPARQL
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from physically separated repositories
- Dynamicity: to provide access to the most recent version of a resource

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: HTTP, URIs, RDF, SPARQL
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from physically separated repositories
- Dynamicity: to provide access to the most recent version of a resource
- Ecosystem: a large and active community with common tools and practices. Initiatives: (1) COST Action *Nexus Linguarum* (COST Action 2019-2023): European network for Web-centred linguistic data science; (2) *Prêt-à-LLOD* (RIA 2019-2022): Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors; (3) LD4LT (*Linked Data for Language Technology Community Group*): to create a consolidated LOD vocabulary for web (linguistic) annotation

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

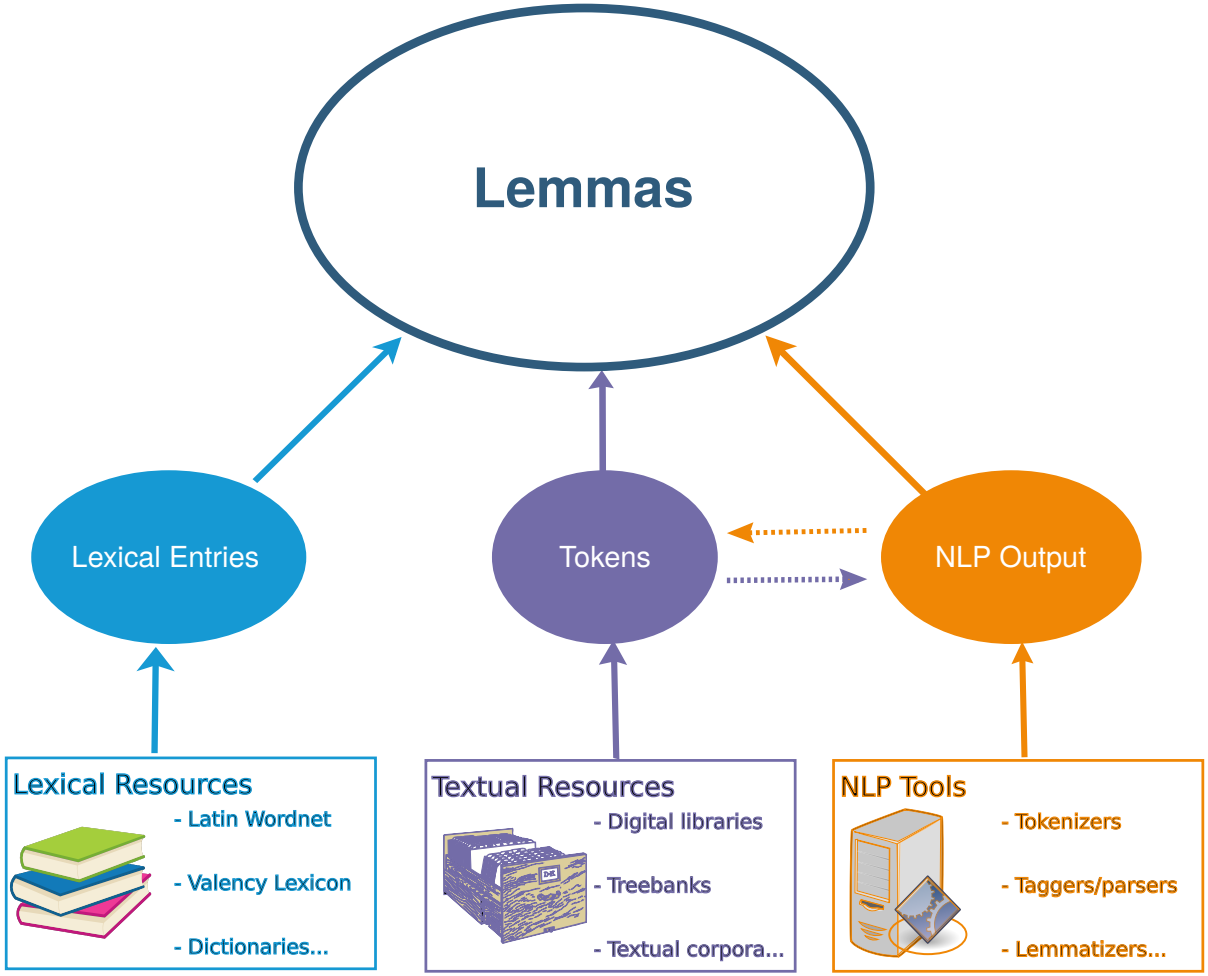
CLARIN
Common Language Resources and Technology Infrastructure

Infrastructure

Interoperability

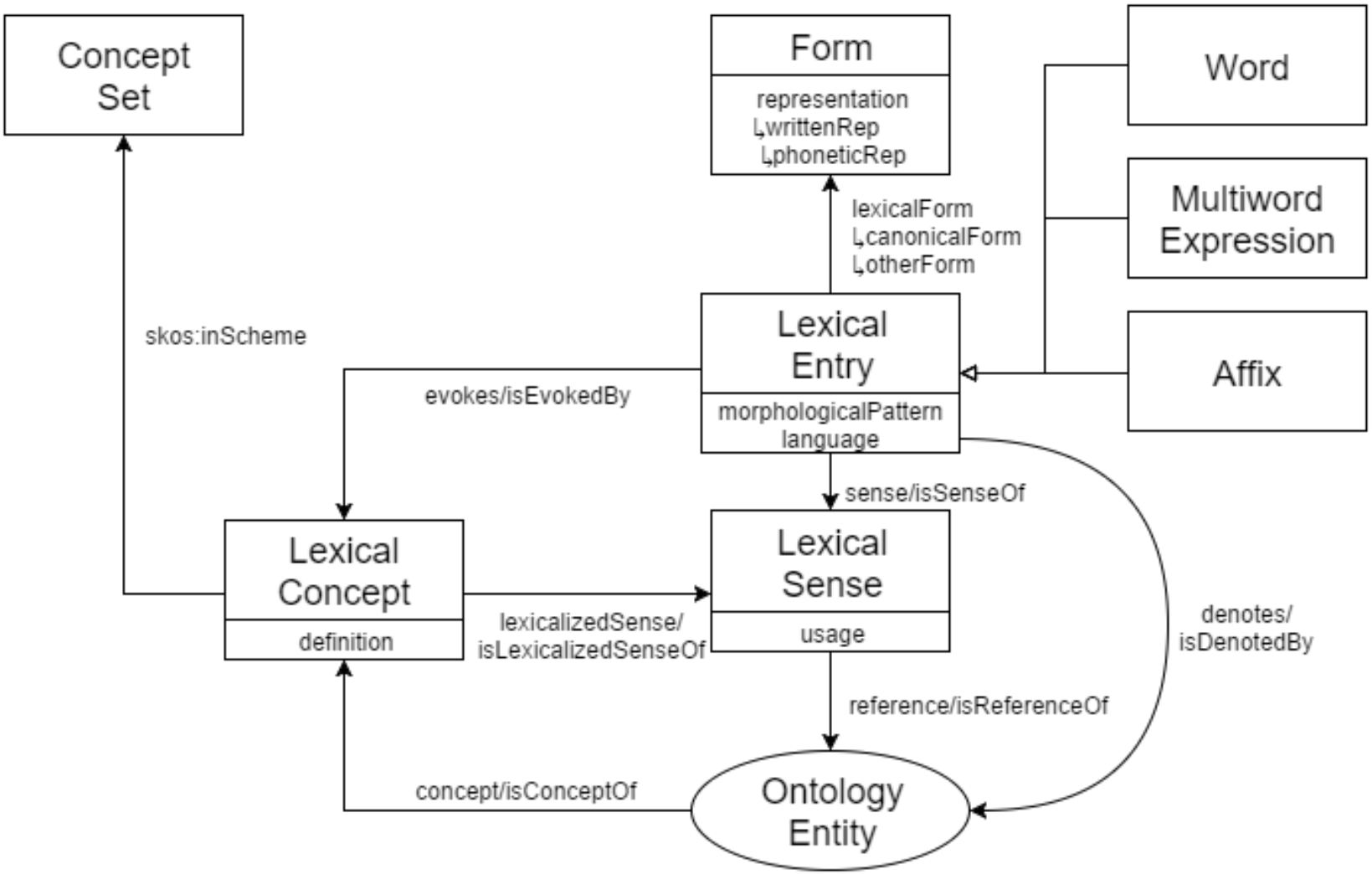
The LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

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



Lemma *admiror* 'to admire, to respect'

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

- Lemma Bank
- A bilingual 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 subset of the Latin WordNet

Lemma Bank Query Interface

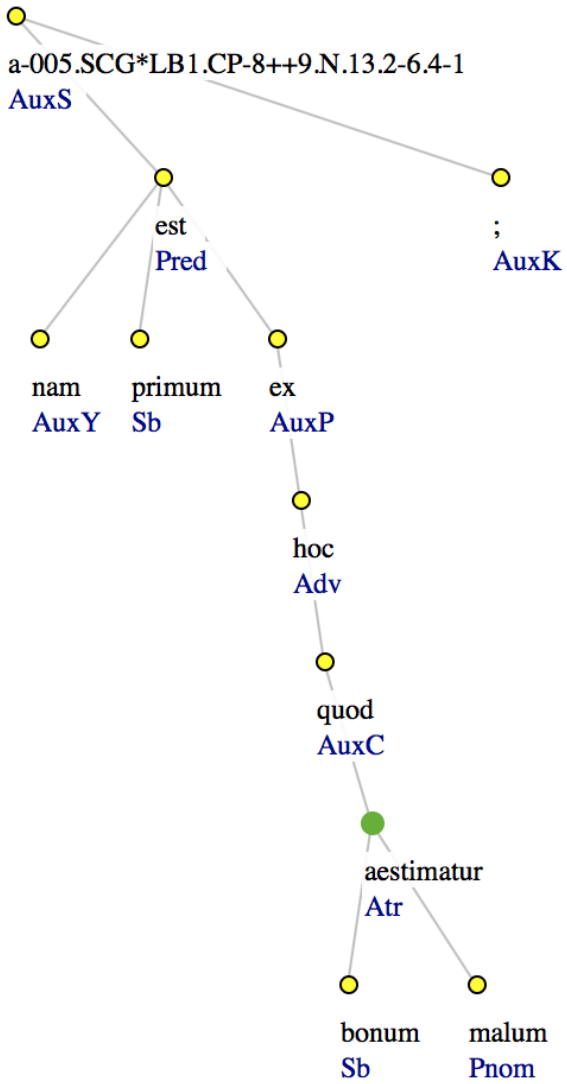
https://lila-erc.eu/query/

Textual Resources

Source: the *Index Thomisticus* Treebank (original scheme)

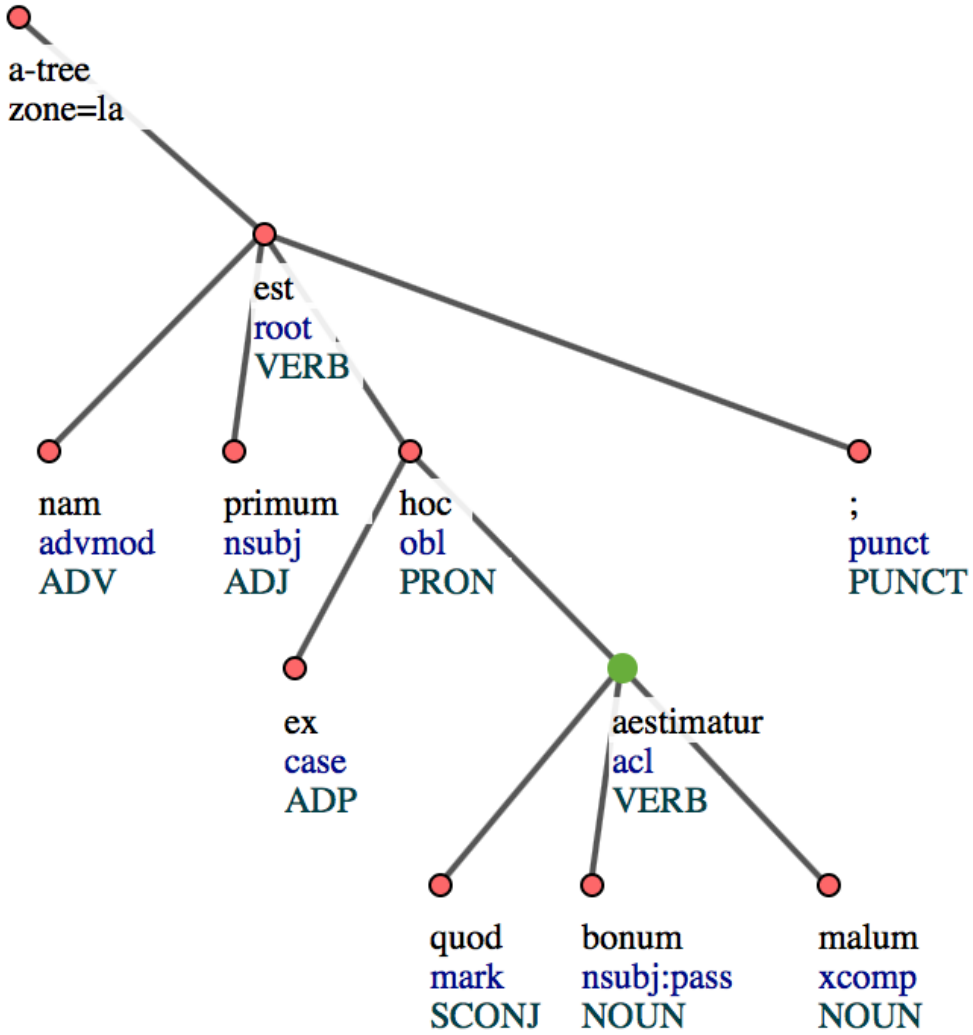
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

for the first arises because the good ***is judged*** *to be evil*; (Trans. Anton C. Pegis)



Textual Resources

Source: the *Index Thomisticus* Treebank (UD scheme)



Textual Resources

A token from the *Index Thomisticus* Treebank in LiLa

Token *aestimatur*

http://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.2-6.4-1W8

LiLa: Overview

Resources connected and upcoming connections

- **Textual Resources**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 400,000 nodes
  - [x] UDante Treebank: ca. 46,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [x] *Liber Abbaci* (ch. VIII) by Leonardo Fibonacci: ca. 30,000 tokens
  - [x] LASLA Corpus: ca. 1.7 million tokens
  - [ ] PROIEL and LLCT treebanks, CompHistSem, CroALa, Musisque DeoQue
- **Lexical Resources**
  - [x] Lemma Bank: ca. 200,000 canonical forms
  - [x] Word Formation Latin: ca. 36,000 lemmas (Classical Latin)
  - [x] Etymological Dict. of Latin & the Other Italic Langs.: ca. 1,500 entries
  - [x] LatinAffectus: ca. 3,300 entries
  - [x] Index Graecorum Vocabulorum in L. Latinam Transl.: ca. 1,800 entries
  - [x] Latin WordNet: ca. 2,500 manually checked entries
  - [x] Latin Vallex 2.0: ca. 2,000 entries
  - [x] Lewis & Short Dictionary: ca. 50,000 entries
  - [ ] Lexikon der Indogermanischen Verben (LIV). Wiktionary, BabelNet

TOTAL: approximately 33 million triples

Table of Contents

TextLinker

Welcome screen: `http://lila-erc.eu:8080/LiLaTextLinker/`

Figure: LiLa's Text Linker

TextLinker

Processed output: `http://lila-erc.eu:8080/LiLaTextLinker/`

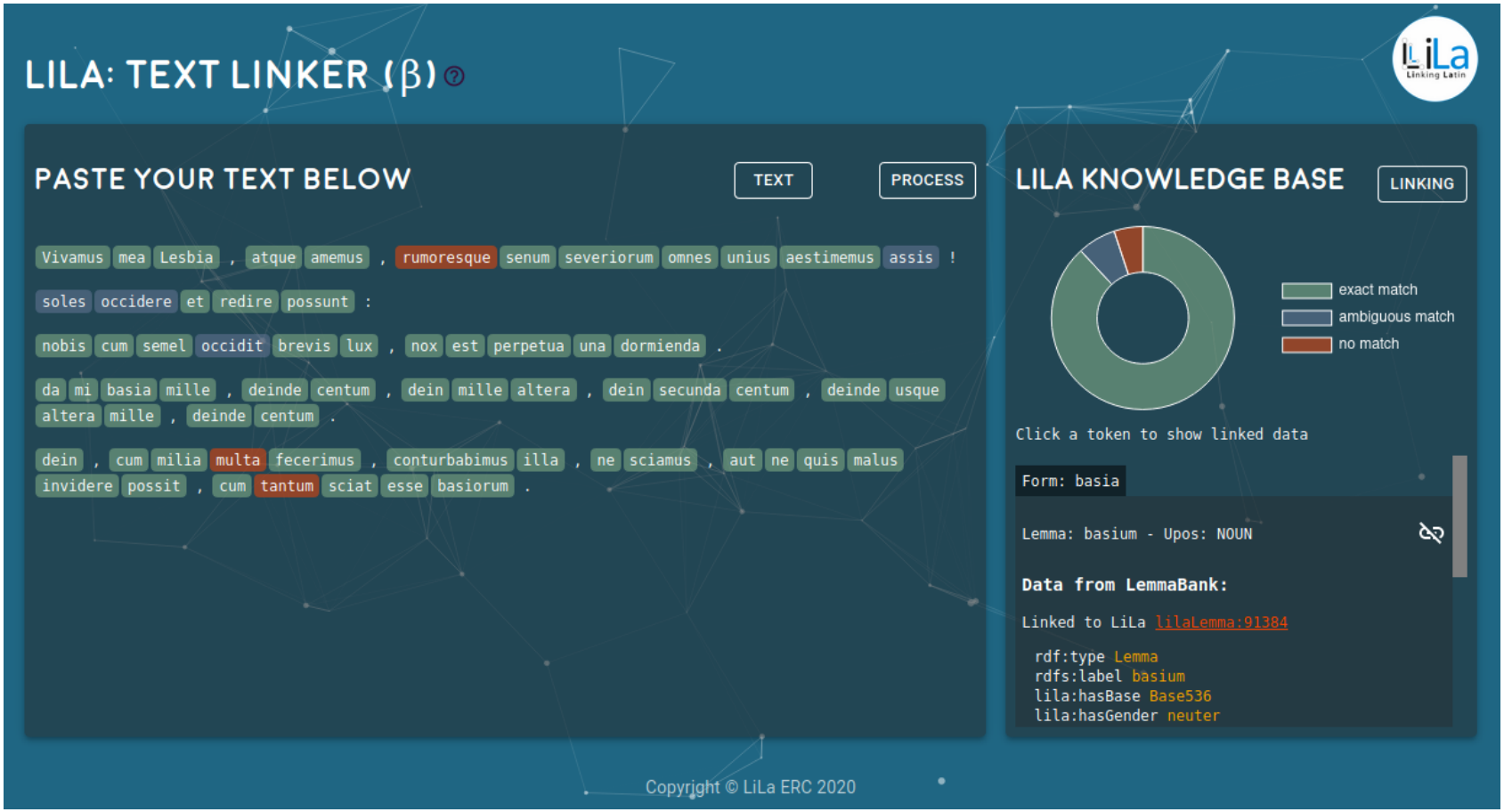


Figure: Text processed against the LiLa Knowledge Base

TextLinker

Hands-on: `http://lila-erc.eu:8080/LiLaTextLinker/`

Try the TextLinker yourself on the provided text(s)!

Table of Contents

SETS Treebank Search

http://depsearch-depsearch.rahtiapp.fi/ds_demo/

Select one UD Latin treebank from the list

- `dicitur`
- `L=dico`
- `NOUN`
- `L=dico & Number=Sing`
- `L=dico &!  Number=Sing`
- `L=dico | L=materia`
- `L=dico >nsubj L=commentator` (ITTB)

KonText

https://lindat.mff.cuni.cz/services/kontext/corpora/corplist

Select collection, language (Latin) and corpus

- Basic: worform (`hominem`) or, if the query is a lemma (`homo`), all the forms of that lemma
- Lemma: `homo`
- Phrase: `forma de`
- Word part: `hom`
- CQL: `[upos="NOUN" & (lemma="homo" | lemma="forma")]`

You can always specify the context
(lemmas co-occurring with query results in a specified window size)

Grew-match

http://match.grew.fr

Select language (Latin) and corpus

- `pattern {N [lemma="homo"]}`
- `pattern {N [upos="NOUN"]}`
- `pattern {V [upos=VERB];} without {V [lemma="sum"]}`
- `pattern {N1 [lemma="forma"]; N2 [lemma="materia"]; N1 < N2 }`
- `pattern {N1 [upos=VERB]; N2 [upos=NOUN]; N1 < N2 }`
- `pattern {GOV -> DEP; DEP [upos=NOUN]}`
- `pattern {N [upos = AUX]}`

PML Tree Query

http://lindat.mff.cuni.cz/services/pmltq/#!/home

Select collection (UD), language (Latin) and corpus

- `a-node $n1:= [ lemma = "homo" ]`
- `a-node $n1:= [ conll/cpos = "NOUN" ]`
- `a-node $n1:= [ conll/cpos = "NOUN" ]`
  `» for $n1.lemma give $1, count(), sort by $2 desc, $1`
- `a-node $n1:= [ conll/cpos = "NOUN", child $n2:= [ conll/deprel = "amod"] ]`
  `» for $n2.lemma give $1, count(), sort by $2 desc, $1`
- `a-node $n1:= [ lemma = "forma", child $n2:= [ conll/deprel = "amod"] ]`
  `» for $n2.lemma give $1, count(), sort by $2 desc, $1`

Table of Contents

SPARQL Access Point

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

Thanks!
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.