Risorse linguistiche per il latino

Lemmatizzazione, PoS Tagging e Interoperabilità

Marco Passarotti

Università di Parma
29 Marzo 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

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated …but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

- The LiLa Knowledge Base
- Services and Tools
- To sum up

Table of Contents

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated …but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

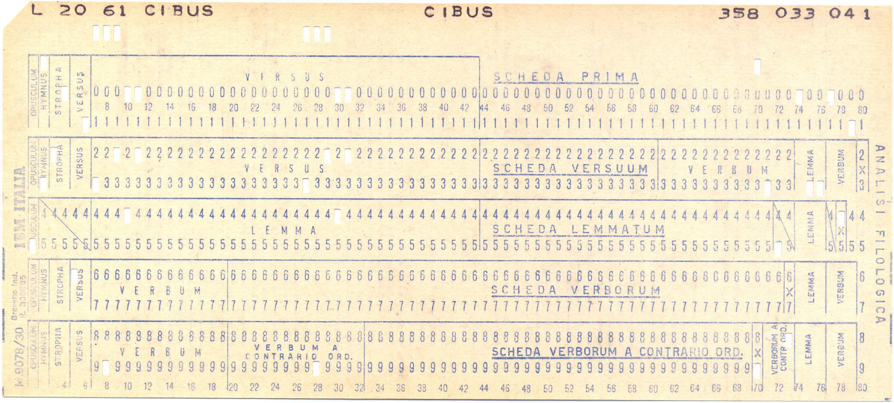
- The LiLa Knowledge Base
- Services and Tools
- To sum up

Millions of Punched Cards

Analogical and isolated ...but Findable

Millions of Punched Cards

Analogical and isolated …but Findable



Millions of Punched Cards

Analogical and isolated …but Findable

[…] in the early 1960s, when computers were scarce, expensive, and cumbersome, using computers for communication was almost unthinkable. […] A scientist who needed to use a distant computer might find it easier to get on a plane and fly to the machine’s location to use it in person.

(Janet Abbate. Inventing the Internet. MIT, 1999, 1.)

Millions of Punched Cards

Analogical and isolated …but Findable

The Result …on Paper

The *Index Thomisticus*

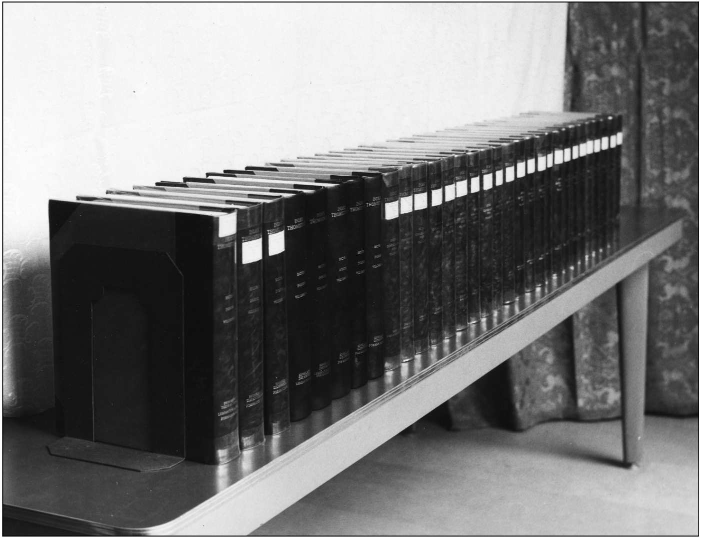


Table of Contents

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated …but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

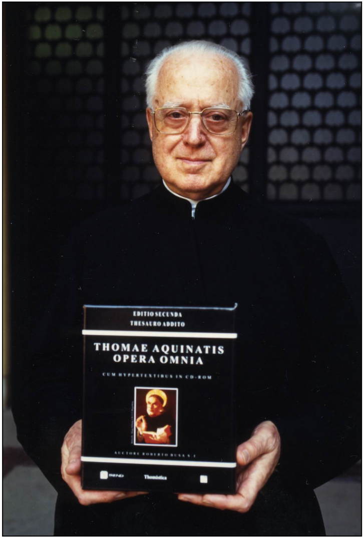
- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

- The LiLa Knowledge Base
- Services and Tools
- To sum up

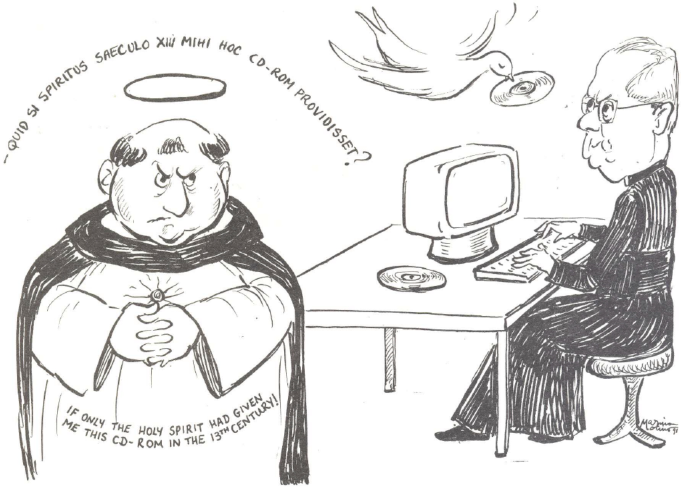
The Digital Turn: (meta)data (partly) accessible

The CD-ROM of the *Index Thomisticus*



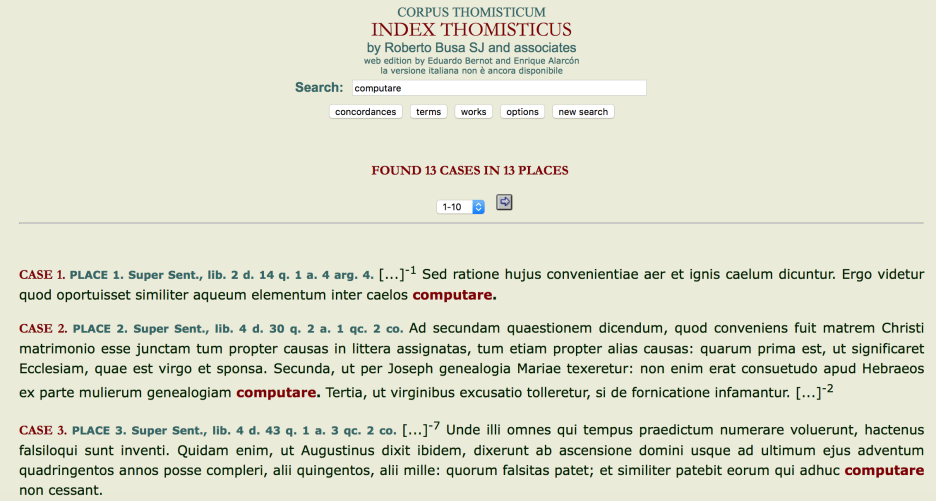
The Digital Turn: (meta)data (partly) accessible

Thanks God for the CD-ROM!


– QUID SI SPIRITUS SAECULO XIII MIHI HOC CD-ROM PROVIDISSET?
IF ONLY THE HOLY SPIRIT HAD GIVEN ME THIS CD-ROM IN THE 13TH CENTURY!

The Digital Turn: (meta)data (partly) accessible

Thanks God for Internet!



CORPUS THOMISTICUM

INDEX THOMISTICUS

by Roberto Busa SJ and associates

web edition by Eduardo Bernot and Enrique Alarcón
la versione italiana non è ancora disponibile

Search: computare

FOUND 13 CASES IN 13 PLACES

CASE 1. PLACE 1. Super Sent., lib. 2 d. 14 q. 1 a. 4 arg. 4. […]$^{-1}$ Sed ratione hujus convenientiae aer et ignis caelum dicuntur. Ergo videtur quod oportuisset similiter aqueum elementum inter caelos **computare.**

CASE 2. PLACE 2. Super Sent., lib. 4 d. 30 q. 2 a. 1 qc. 2 co. Ad secundam quaestionem dicendum, quod conveniens fuit matrem Christi matrimonio esse junctam tum propter causas in littera assignatas, tum etiam propter alias causas: quarum prima est, ut significaret Ecclesiam, quae est virgo et sponsa. Secunda, ut per Joseph genealogia Mariae texeretur: non enim erat consuetudo apud Hebraeos ex parte mulierum genealogiam **computare.** Tertia, ut virginibus excusatio tolleretur, si de fornicatione infamantur. […]$^{-2}$

CASE 3. PLACE 3. Super Sent., lib. 4 d. 43 q. 1 a. 3 qc. 2 co. […]$^{-7}$ Unde illi omnes qui tempus praedictum numerare voluerunt, hactenus falsiloqui sunt inventi. Quidam enim, ut Augustinus dixit ibidem, dixerunt ab ascensione domini usque ad ultimum ejus adventum quadringentos annos posse compleri, alii quingentos, alii mille: quorum falsitas patet; et similiter patebit eorum qui adhuc **computare** non cessant.

Table of Contents

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated ...but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

- The LiLa Knowledge Base
- Services and Tools
- To sum up

Lemmatization & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

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 & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

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.

Lemmatization & Part-of-Speech Tagging

Attaining a standard representation of lexicon and morphosyntax

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

- The Story So Far (through the lenses of the *Index Thomisticus*)
  - Analogical and Isolated …but Findable
  - Digital and (Partly) Accessible
- Lemmatization & Part-of-Speech Tagging
  - What is lemmatization and PoS tagging?
  - Lemmatized Corpora for Latin
- Tools and Hands-on
  - Tools for lemmatization & POS Tagging
- Latin in the Semantic Web
  - The LiLa Knowledge Base
  - Services and Tools
  - To sum up

Lemmatized Corpora for Latin

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

Table of Contents

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated ...but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

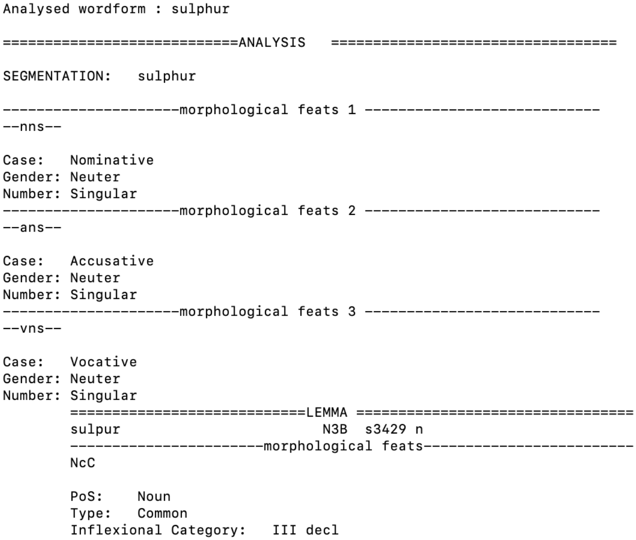
- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

- The LiLa Knowledge Base
- Services and Tools
- To sum up

LEMLAT

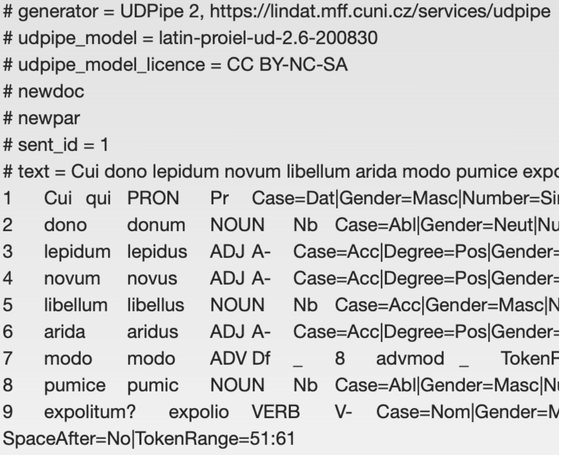
Type-based: `https://github.com/CIRCSE/LEMLAT3`



```
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

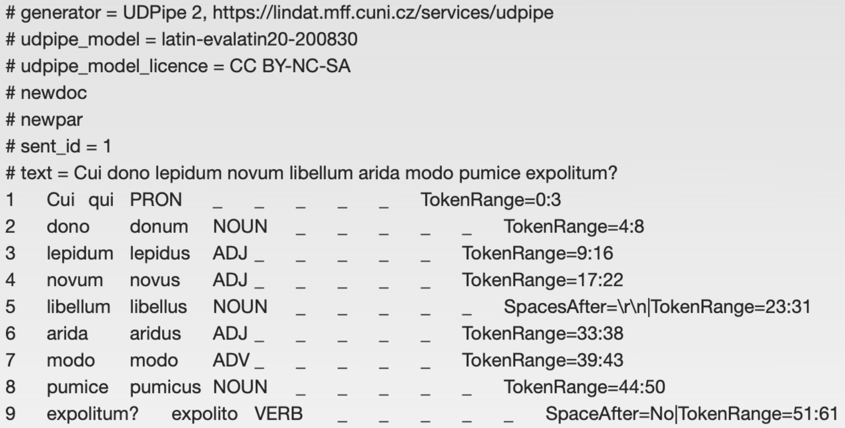
Token-based: `https://lindat.mff.cuni.cz/services/udpipe/`



```
# 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 _      TokenF
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

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated ...but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging

- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on

- Tools for lemmatization & POS Tagging

Latin in the Semantic Web

- The LiLa Knowledge Base
- Services and Tools
- To sum up

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

The LiLa Knowledge Base

Key-word: interoperability

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

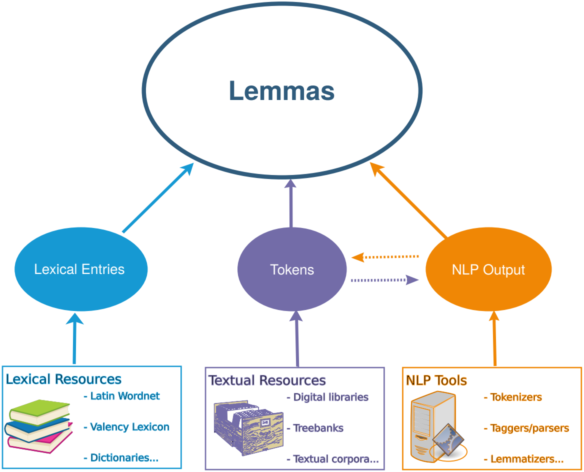
LiLa

Linking Latin

Interoperability

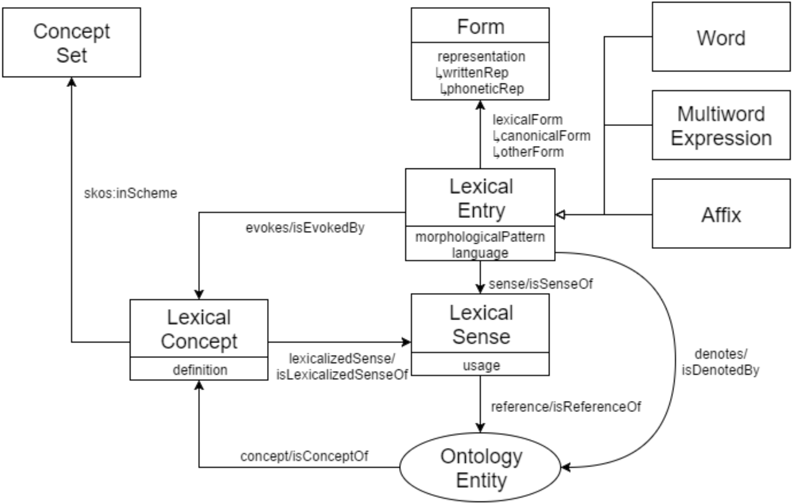
The LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

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



Lemma Bank and Lexical Resources

The core of LiLa

Lemma *admiror* ‘to admire, to respect’

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

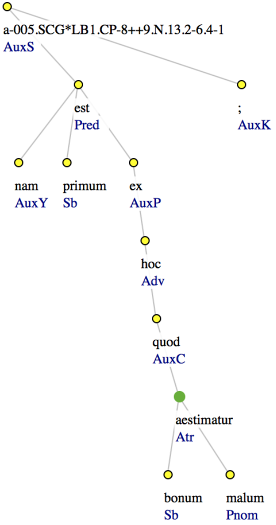
- Lemma Bank
- A bilingual dictionary (Lewis & Short)
- A derivational lexicon (Word Formation Latin)
- A polarity lexicon (LatinAffectus)
- An etymological dictionary (De Vaan)
- A Valency Lexicon (Latin Vallex)
- A manually checked subset of the Latin WordNet

Textual Resources

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

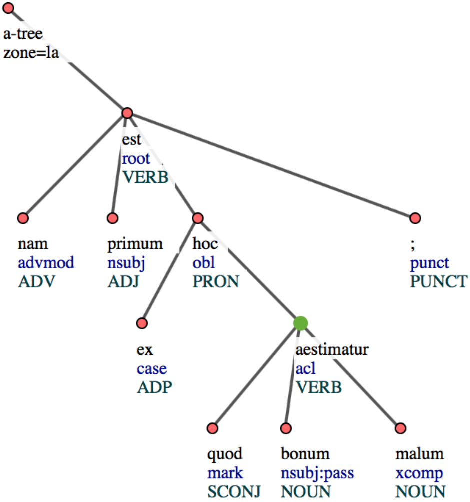
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

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`

Table of Contents

The Story So Far (through the lenses of the *Index Thomisticus*)
- Analogical and Isolated ...but Findable
- Digital and (Partly) Accessible

Lemmatization & Part-of-Speech Tagging
- What is lemmatization and PoS tagging?
- Lemmatized Corpora for Latin

Tools and Hands-on
- Tools for lemmatization & POS Tagging

Latin in the Semantic Web
- The LiLa Knowledge Base
- Services and Tools
- To sum up

Lemma Bank Query Interface

Lemma Bank Query Interface

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

SPARQL Access Point

SPARQL Access Point

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

TextLinker

TextLinker

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

LiLa Search Platform

LiLa Search Platform

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

Table of Contents

- The Story So Far (through the lenses of the *Index Thomisticus*)
  - Analogical and Isolated ...but Findable
  - Digital and (Partly) Accessible
- Lemmatization & Part-of-Speech Tagging
  - What is lemmatization and PoS tagging?
  - Lemmatized Corpora for Latin
- Tools and Hands-on
  - Tools for lemmatization & POS Tagging
- Latin in the Semantic Web
  - The LiLa Knowledge Base
  - Services and Tools
  - To sum up

LiLa: Overview

Resources connected so far

Resources connected in LiLa

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

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.