What's going on in Milan?

Resources and Tools for Latin at the CIRCSE Research Centre

Francesco Mambrini and **Greta Franzini**
francesco.mambrini@unicatt.it, greta.franzini@unicatt.it

'Text Encoding: Latinists Looking for New Synergies' Workshop | Liége |
November 8, 2018

Table of Contents

Lemlat is a morphological analyzer and lemmatizer. Version 3.0 [Passarotti et al. 2017]:

- Extended lexical basis:
  - Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904 → **43,432 lemmas**
  - Forcellini (1771/1940; *Onomasticon*) → **26,250 lemmas**
  - Du Cange's (1678-1887; *Glossarium mediae et infimae latinitatis*) → **85,999 lemmas**
- Connected to *Word Formation Latin* (WFL)
- Strength: Spelling variation
- Performance compared to analogous tools [Springmann et al. 2016][1]:

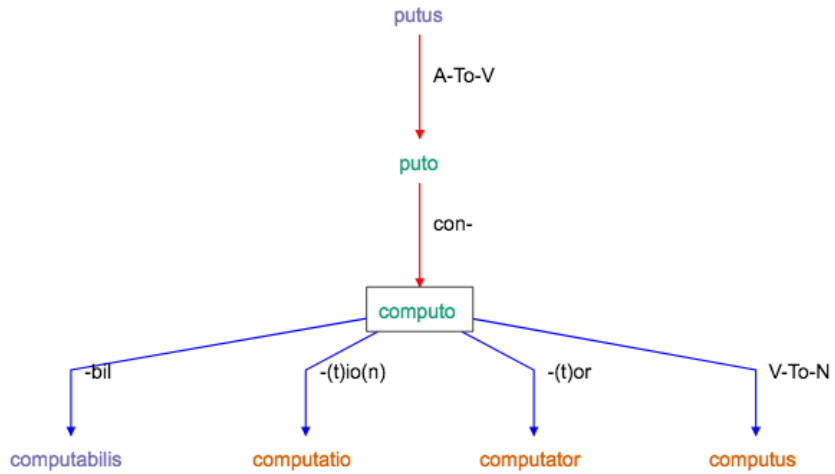
	Caesar		Nepos		Godfrey	
all	type	token	type	token	type	token
PROIEL	70.0	51.6	69.4	47.9	63.1	50.6
Parsley	89.5	95.2	90.0	94.3	86.7	91.7
Words	90.5	96.6	88.1	93.3	93.0	95.4
Morpheus	92.5	93.8	89.0	92.7	87.6	92.7
LEMLAT	98.2	99.0	98.1	99.1	91.0	94.9
LatMor	97.5	99.1	98.1	99.2	96.4	97.5

[1] "If Roman numerals are taken out of account, LEMLAT is in fact the best performing system of all." (p. 390)

```
============================ANALYSIS    ============================

SEGMENTATION:    castigabil -em

--------------------morphological feats 1 ----------------------
--ams-1

Case:    Accusative
Gender: Masculine
Number: Singular
Degree: Positive
--------------------morphological feats 2 ----------------------
--afs-1

Case:    Accusative
Gender: Feminine
Number: Singular
Degree: Positive
        ============================LEMMA ========================
        castigabilis                    N3A   c0772 *
        ----------------------morphological feats------------------
        Af-

        PoS:     Adjective
        Type:    Qualifying
        ----------------------derivational info--------------------
        IS DERIVED: YES
        ----------------------rule id: 38--------------------------
        Lexical Basis:
                castigo                         V1    c0776   VmF
        Derivational Type: Derivation_Suffix
        Derivational Category: V-To-A
        Affix: bil
```

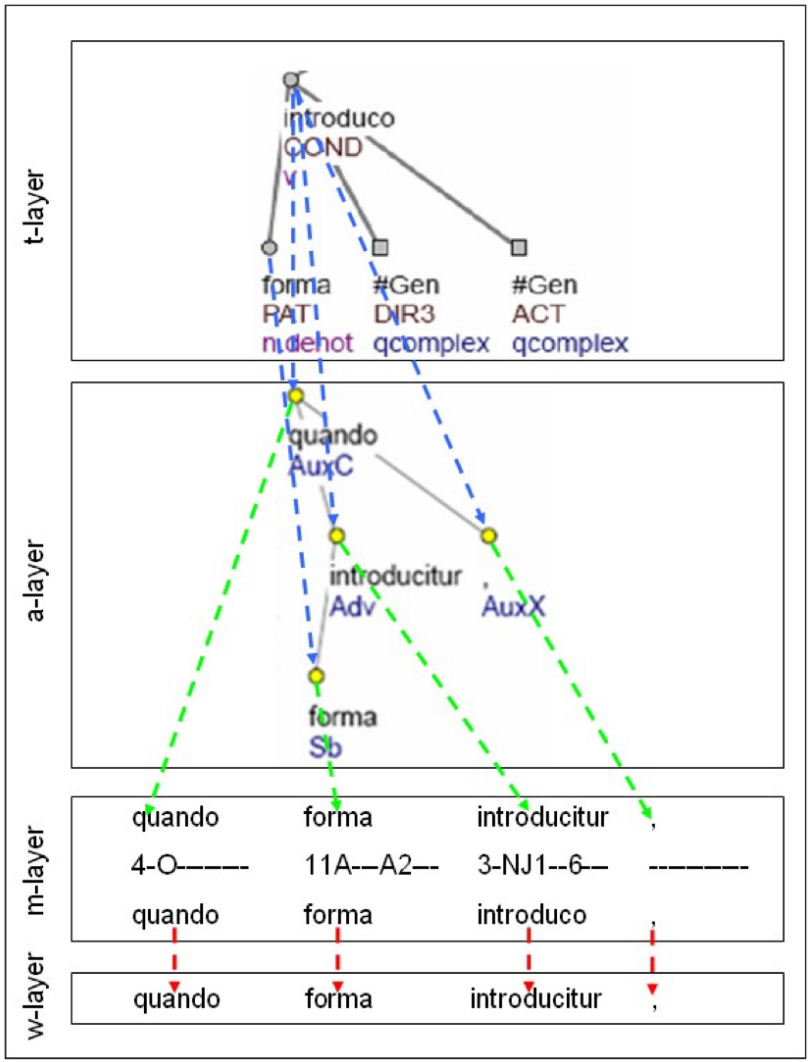
Word Formation Latin (Marie Curie Individual Fellowship)

- Lexical basis:
  - Georges & Georges + Glare + Gradenwitz from Lemlat
- Current state of WFL (**ongoing**):
  - 648 Word Formation Rules (WFRs) applied (derivation; conversion; compounding)
  - 31,644 relations
  - 4,299 morpho-derivational families
- [Litta 2018]



The Index Thomisticus Treebank

http://itreebank.marginalia.it



Meaning

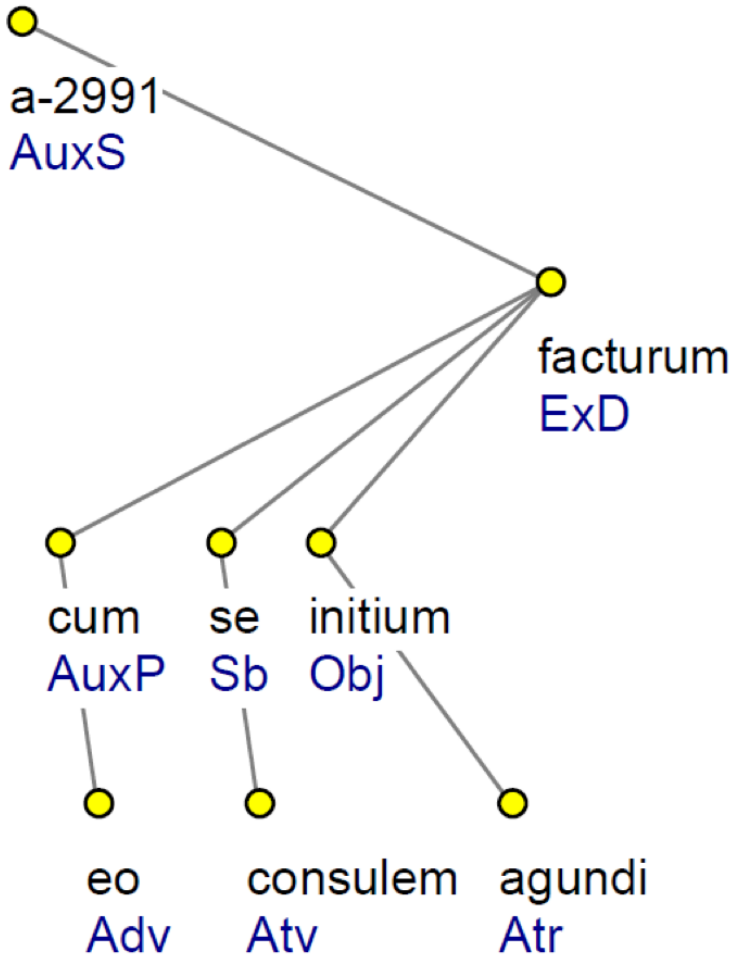
- Approx. 2,000 sentences
- From *Summa Contra Gentiles*
- + excerpts from LDT and entire *Bellum Catilinae*

Form

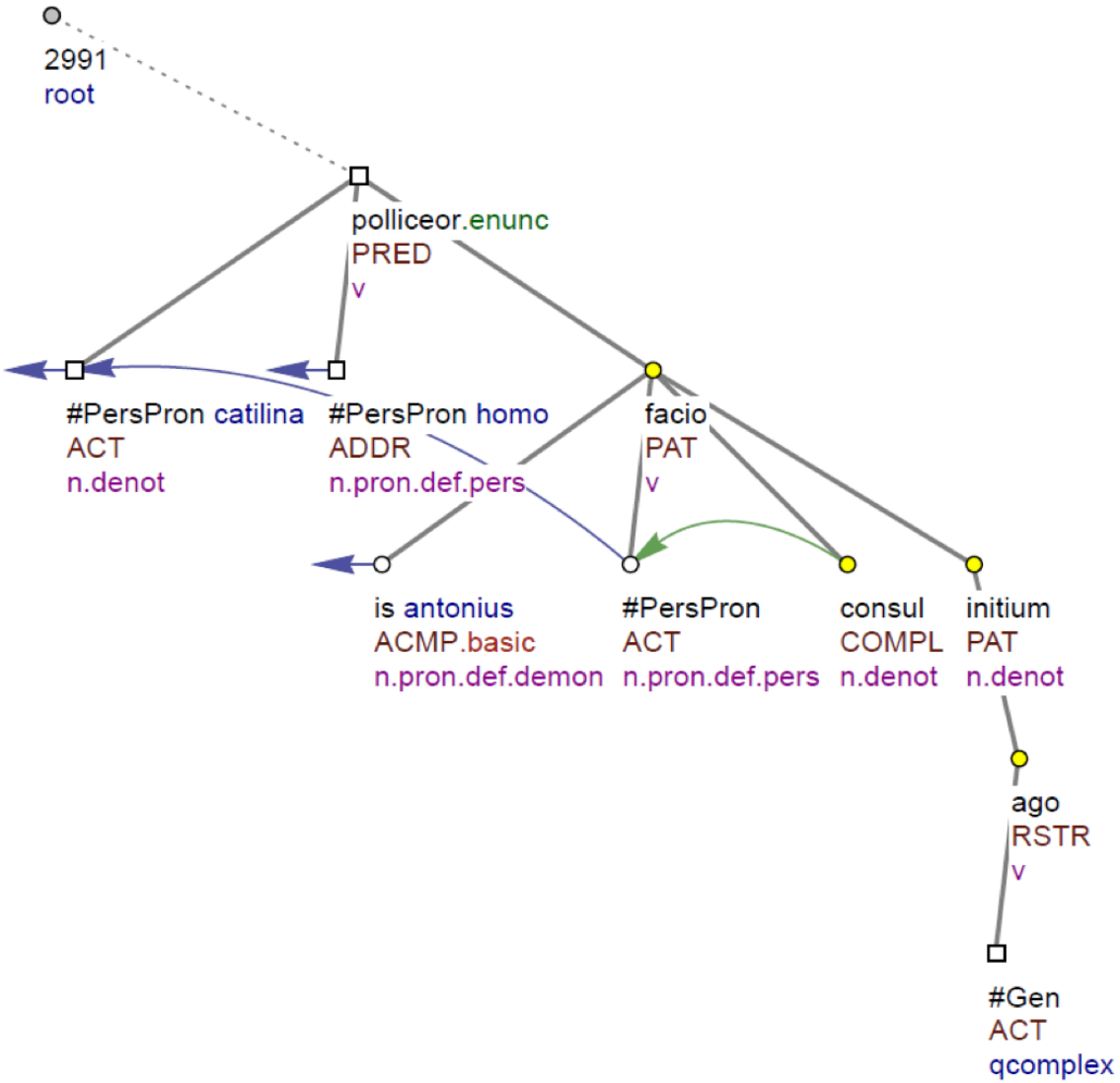
- Approx. 400,000 nodes
- Approx. 20,000 sentences
- From ST, SN, SCG

[Passarotti 2015]

cum eo se consulem initium agundi facturum



cum eo se consulem initium agundi facturum



IT-VaLex

http://itreebank.marginalia.it/itvalex/

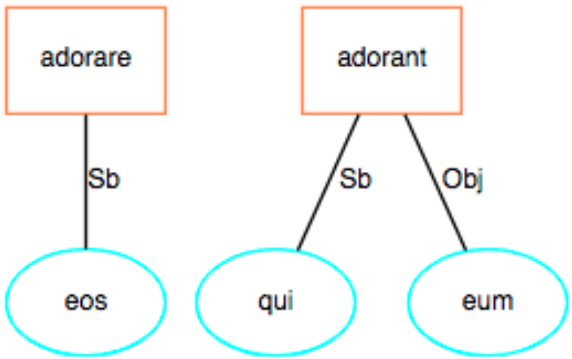
The **ongoing** *Index Thomisticus* Valency Lexicon (IT-VaLex) provides the valency frames of verbs in the IT-TB.

- Distinction between *arguments* and *adjuncts* in Analytical Tree Structure (ATS)
- Dynamically-induced from the IT-TB

- Links every verb to all its arguments textually represented in the IT-TB through a separate lexical item: textual syntactic valency
- Size: **1,276 verbs** (65,535 occurrences)

Summa contra Gentiles ; Liber 1 ; Caput 20 ; Numerus, Paragraphus 36

spiritus **est** deus, et eos qui eum **adorant**, in spiritu et veritate **adorare** **oportet.**



Latin Vallex

http://itreebank.marginalia.it/vallex

Size and Data Source

1,373 lexical entries (verbs, nouns, adjectives, adverbs)
3,406 valency frames

Corpus-based (annotation-driven)

- *Index Thomisticus* Treebank: approx. 2,000 sent (SCG)
- Latin Dependency Treebank: approx. 100 sent
  - Caesar: 100 sent (*De bello gallico*)
  - Cicero: 100 sent (*In Catilinam*)
  - Sallust: 701 sent (entire *Bellum Catilinae*)

Intuition-based (for balance)

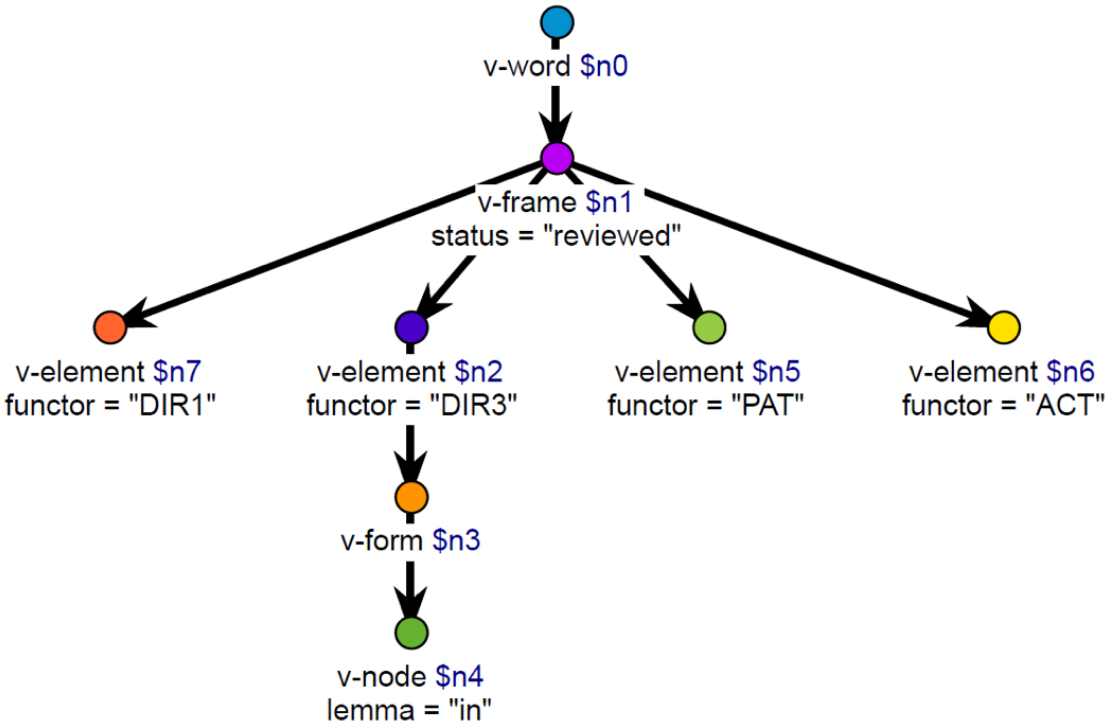
- 163 entries → first 1.000 valency-capable lemmas in Delatte et alii (1981) *Dictionnaire Frequentiel et index inverse de la langue latine*

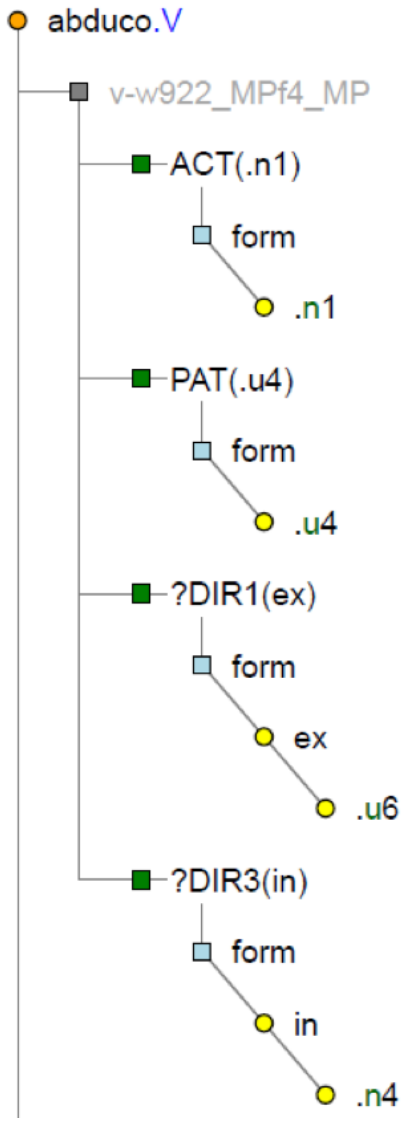
[Passarotti et al. 2016]

Latin Vallex

http://itreebank.marginalia.it/vallex

Lexical entries with a 4-argument valency frame: Actor, Patient, Direction-from, Direction-to (PP introduced by in).





LiLa: Linking Latin

https://lila-erc.eu

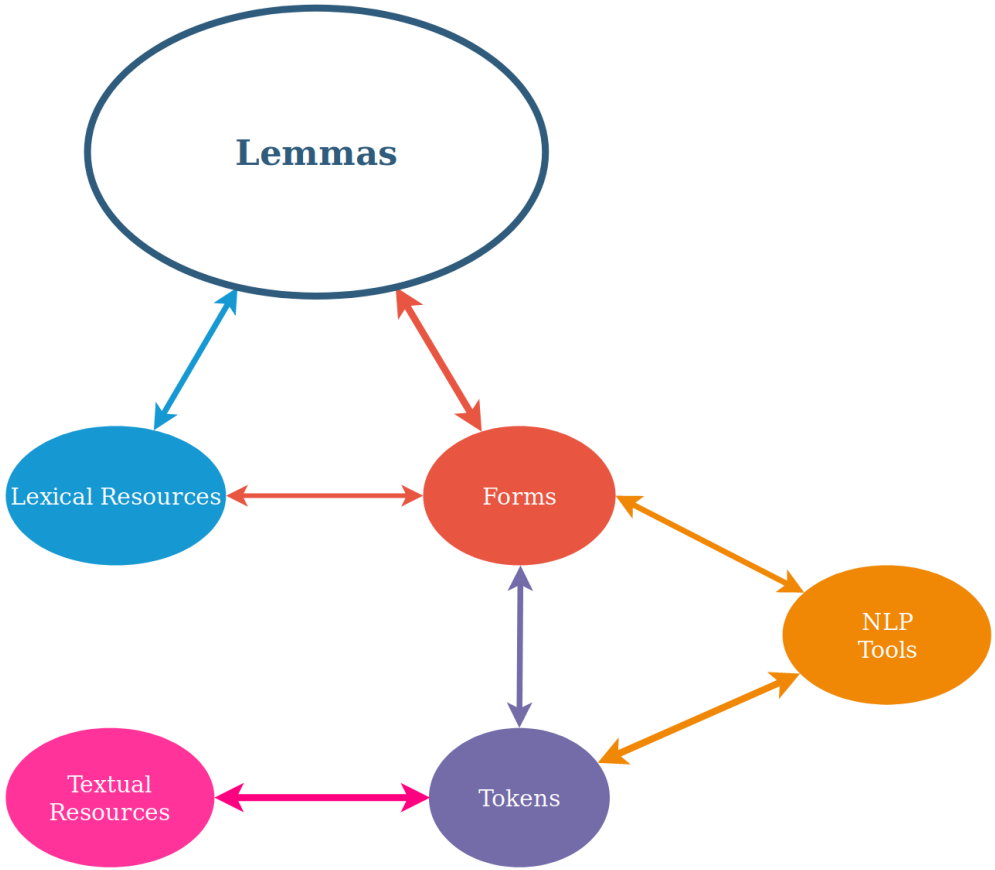
Linking Latin: Building a Knowledge Base of Linguistic Resources for Latin

European Research Council (ERC) Consolidator Grant
H2020 Research and Innovation Programme
N: 769994; Duration: 2018-2023; Award: €2M

LiLa: Linking Latin

https://lila-erc.eu

Connecting and adding linguistic resources to NLP tools and corpora to the *Linguistic Linked Open Data Cloud*: http://linguistic-lod.org/llod-cloud



Gratias vobis agimus

Get in touch!

Francesco Mambrini and **Greta Franzini**
CIRCSE, Università Cattolica del Sacro Cuore

f.mambrini@gmail.com, greta.franzini@unicatt.it

@FrancMambr, @GretaFranzini

https://github.com/CIRCSE

https://centridiricerca.unicatt.it/circse_index.html

Largo Gemelli 1, 20123 Milan, Italy

Works Cited

- Litta, E. (2018) 'Morphology Beyond Inflection. Building a Word Formation Based Lexicon for Latin', in Cotticelli-Kurras, P., Giusfredi, F. (eds) *Formal Representation and the Digital Humanities*. Cambridge Scholars Publishing, pp. 97-114.
- Passarotti, M., Budassi, M., Litta, E., Ruffolo, P. (2017) 'The Lemlat 3.0 Package for Morphological Analysis of Latin', in Bouma, G., Adesam, Y. (eds) *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Northern European Association for Language Technology (NEALT) Proceedings Series*, Vol. 32, pp. 24-31.
- Passarotti, M., González Saavedra, B., Onambele, C. (2016) 'Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin', in Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. May 23-28, 2016, Portorož, Slovenia, European Language Resources Association (ELRA), pp. 2599-2606.
- Passarotti, M., (2015) 'What you can do with linguistically annotated data. From the *Index Thomisticus* to the *Index Thomisticus* Treebank', in Roszak, P., Vijgen, J. (eds) *Reading Sacred Scripture with Thomas Aquinas. Hermeneutical Tools, Theological Questions and New Perspectives*. Brepols, pp. 3-44.
- Springmann, U., Schmid, H., Najock, D. (2016) 'LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity', *Open Linguistics*, 2(1). DOI: https://doi.org/10.1515/opli-2016-0019

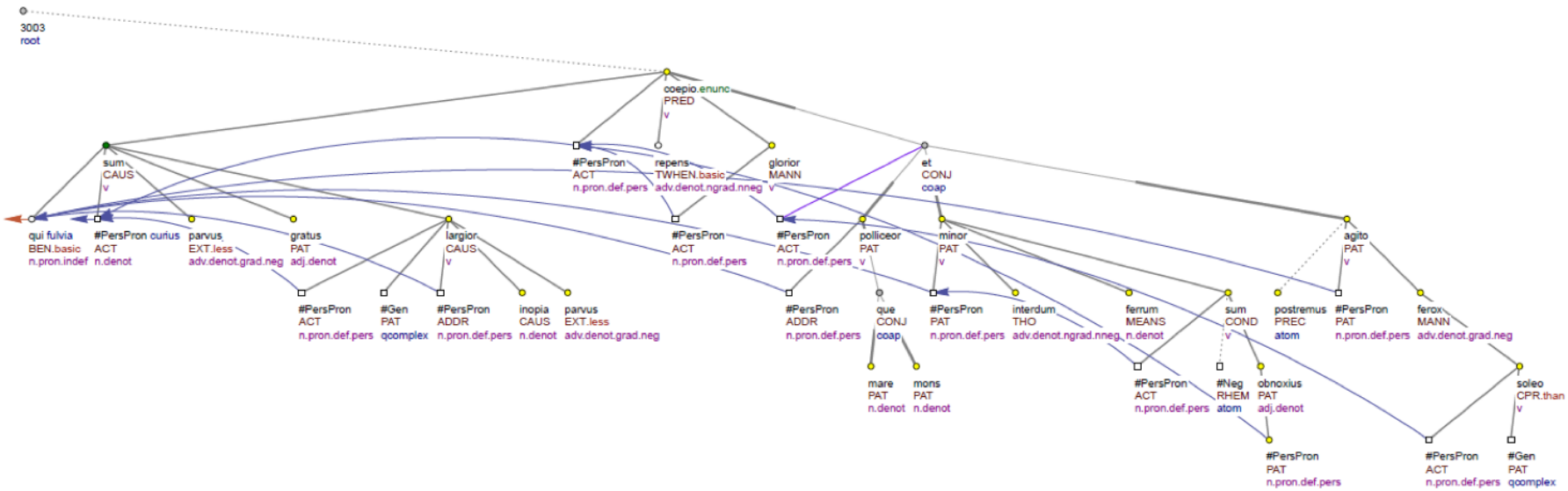
Credits and Acknowledgements

- LaTeX Beamer Feather Theme, modified by Greta Franzini under the terms of the GNU General Public License as published by the Free Software Foundation.

The Index Thomisticus Treebank (IT-TB)

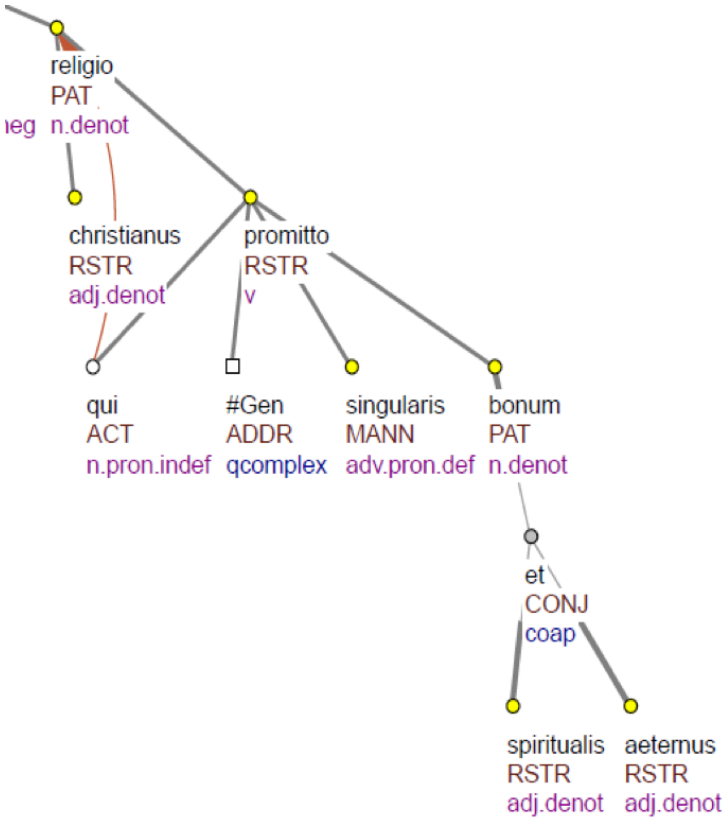
http://itreebank.marginalia.it

Quoi cum minus gratus esset, quia inopia minus largiri poterat, repente glorians maria que montis polliceri coepit et minari interdum ferro, ni sibi obnoxia foret, postremo ferocius agitare quam solitus erat



Treebank → Tectogrammatic Tree Structure (TG-TS)-based Valency Lexicon

[…] christianae religioni […], quae singulariter bona spiritualia et aeterna promittit
(SCG, LB1, CP 5, N. 2)



promitto – V

- Valency Frame 1:
  - Valency:
    - Frame slot 1: .u1
    - Frame slot 2: .n4
    - Frame slot 3: NA
  - Attributes:
    - Functor 1: ACT
    - Functor 2: PAT
    - Functor 3: ADDR