The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, Matteo Pellegrini

Università Cattolica del Sacro Cuore, Milan, Italy

Introduction

In recent times, many treebanks have been published for several languages, including Latin, for which the largest one is currently the **Index Thomisticus Treebank** (ITTB), containing texts from Thomas Aquinas' *Summa Contra Gentiles*.
However, the full exploitation of the available treebanks is limited by:

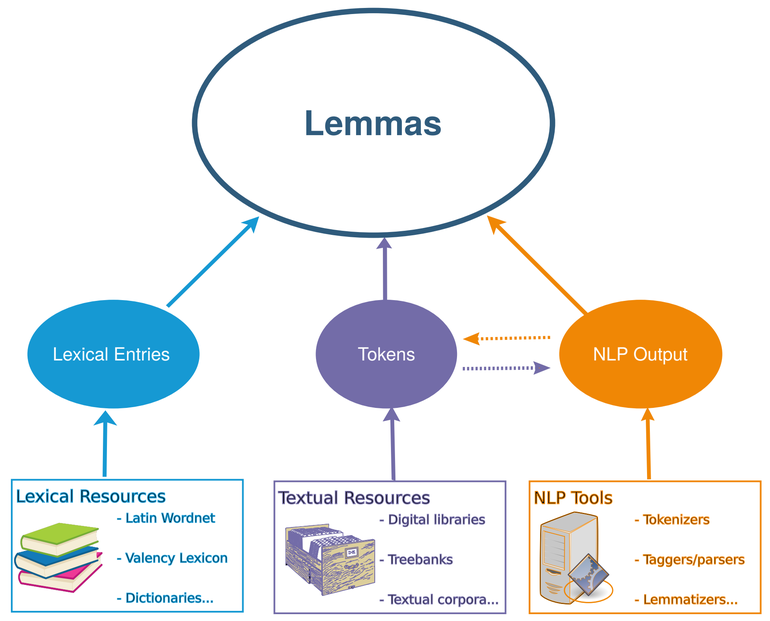
- use of different formats, tagsets and annotation guidelines
  → a successful answer to this shortcoming has been offered by the **Universal Dependencies** (UD) project
- lack of structural connection with other textual and lexical resources
  → we propose to use the principles of **Linguistic Linked Open Data** (LLOD) to address this issue

Universal Dependencies and Linguistic Linked Open Data

Given the growing popularity of UD and LLOD, possible interactions between the two have been explored:

- **CoNLL-RDF** suite of tools to convert from the UD format CoNLL-U to RDF triples
  → at present, the UD corpora included in the LLOD Cloud only display a shallow conversion, without any linking to vocabularies for annotation
- formalization of UD vocabulary with **OLiA** annotation models
  → we use OLiA ontologies to model the annotations of the ITTB, and we connect it to other linguistic resources for Latin by linking its tokens to the lemmas of the **LiLa Knowledge Base**

The LiLa Knowledge Base



8 resources + 5 resources ≈ **33 million triples**

The Index Thomisticus Treebank as Linguistic Linked Open Data

Use of the **POWLA** ontology to model corpus data

- it provides a vocabulary to describe the stratification of a corpus in documents and subsections, which is particularly useful for researchers in Historical Linguistics and Digital Humanities
- it allows corpus providers to group different types of annotations in different layers, making it possible to publish annotations both in the original ITTB format and in its UD conversion

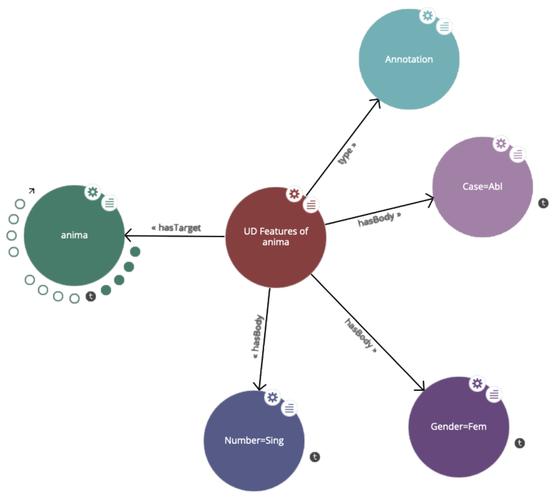
Use of OLiA **UD annotation models**

- classes = general concepts (e.g., the part of speech 'adposition')
- named individuals = language-specific realizations (e.g., adpositions as defined and used in the Italian UD treebanks)

→ as the localized Latin named individuals are not consistently supported by the latest distribution, we had to extend the vocabulary with new terms and classes.

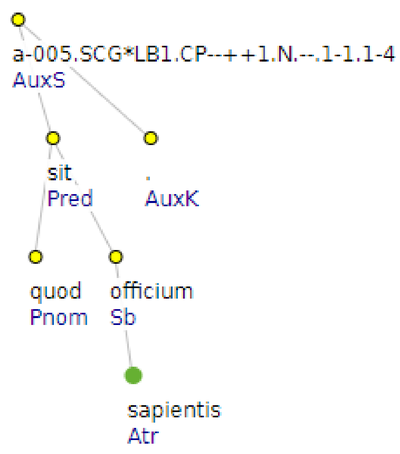
Morphological Features

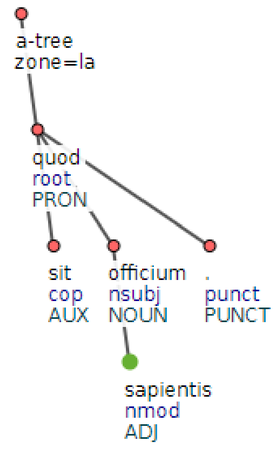
anima, `Case=Abl|Gender=Fem|Number=Sing`

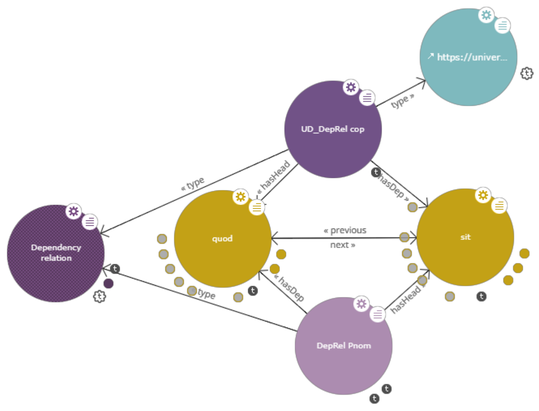


Syntactic Annotation

quod sit officium sapientis. (SCG 1.1.1)

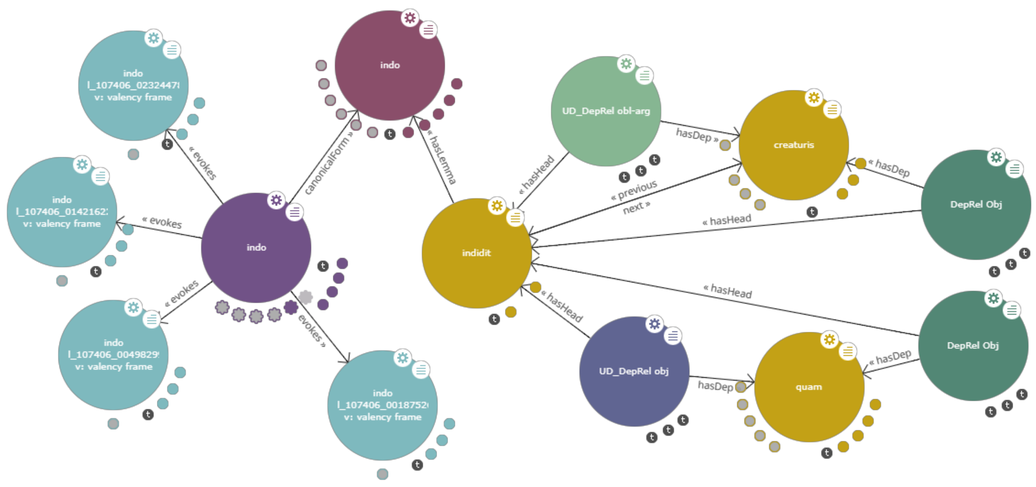






Use Case: Interoperability between Treebanks and Lexical Resources

[...] *quam deus indidit creaturis* [...] (SCG 4.8.10)



Definition of **objects**:

- ITTB: based on the distinction between **arguments** and **adjuncts**
  → verbs with ≥ 3 arguments have more than one of them tagged as `Obj`
- UD: based on the distinction between **core** and **oblique** arguments
  → verbs with ≥ 3 arguments have only one of them marked as `obj`

⇓

Using the valency frames listed in the **Latin Vallex** to check for potential conversion errors

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