LiLa: Linking Latin

Building a Knowledge Base of Linguistic Resources for Latin

Marco Passarotti, Flavio M. Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, Paolo Ruffolo
CIRCSE - Università Cattolica del Sacro Cuore

- RESEARCH ISSUE: wealth of linguistic resources and NLP tools for Latin > **scattered and unconnected**.
- RESEARCH NEED: make sense of this quantity of empirical data > **from information to knowledge**.
- METHOD: combining methods from Information and Communications Technology, Computational Linguistics and Classics with Linked Data technologies > **make resources and tools interact through linking.**

} http://lila-erc.eu {

- OUTCOME: a Linked Data Knowledge Base of resources and NLP tools for Latin connected to the Linguistic Linked Open Data (LLOD) cloud > **adhering to the FAIR principles.**
- IMPACT: speed of retrieval, scale and hold on the data, putting ICT/CL in the hands of Classicists > **no more "Traditional" and "Digital" Humanities.**

RESOURCE ENHANCEMENTS

Latin Vallex
- Assessment of available valency frames.
- Extension of the lexical basis.

Latin WordNet
- Stage 1: (Manual) evaluation of existing Latin WordNet (Classical Latin).
- Stage 2: Conversion of the evaluated WordNet to a Neo4j graph database and expansion of the lexical database using the LEMLAT 3 *lemmario* (ca. 150,000 lemmas).

Index Thomisticus Treebank (IT-TB)
- Analytical annotation of the entire *Summa contra Gentiles*.

EXAMPLE GRAPH OF *DUCO*, 'TO LEAD'

WORD FORMATION

POS TAGGING

Step 1: Gathering of resources
- Collection of Part-of-Speech (PoS) taggers, possibly specifically developed for Latin.
  - If possible, training of a PoS system on the latest IT-TB version of *Universal Dependencies*.
- Collection of tokenised, lemmatised and tagged test sets from documents from different sources (e.g., LASLA, PROIEL) or their production on our part.

Step 2: Putting everything together
- Normalisation of tokenisations to facilitate comparisons.
  - The original tokenisations are retained for further comparisons between different annotation systems.
- All selected PoS taggers are run on each of the selected test sets.

Step 3: Collective evaluation
- All tagging results are evaluated with respect to the gold standard represented by each test set.
- To make the evaluation take into account different annotation systems, we use or develop automatically-inferred mappings between different annotation systems.
- Evaluation measures for clustering, i.e., annotation-agnostic (e.g., BCubed, Rand index).

Step 4: All for one, one for all
- Detection of differences in annotation patterns of selected PoS taggers.
- Conclusions about interaction between PoS annotation systems and automated tagging strategies.
- Creation of an aggregated PoS tagger capable of exploiting such patterns, e.g., by means of deep learning.