EvaLatin 2022

Rachele Sprugnoli, Marco Passarotti
Flavio Massimiliano Cecchini, Margherita Fantoli
rachele.sprugnoli@unipr.it
marco.passarotti@unicatt.it
flavio.cecchini@unicatt.it,
margherita.fantoli@kuleuven.be

@ LT4HALA 2022 | Marseille, France | June 25th 2022

Salve!

Welcome to the **second edition** of EvaLatin!

EvaLatin is an **evalutaion** campaign for **N**atural **L**anguage **P**rocessing tools devoted to **Latin**.

Its research questions:

- How can we **promote** the development of resources and language technologies for the Latin language?
- How can we **foster** collaboration among scholars working on Latin and attract researchers from different disciplines?

Traning and Test sets

Both a **training** and **test** set have been provided for the campaign.

Traning and Test sets

Both a **training** and **test** set have been provided for the campaign.

- The **training** set is constituted by the texts of five Classical prose authors (Caesar, Cicero, Seneca, Pliny the Younger and Tacitus) for a total of more than 300,000 tokens, and different genres (treatises, speeches, letters)

Traning and Test sets

Both a **training** and **test** set have been provided for the campaign.

- The **training** set is constituted by the texts of five Classical prose authors (Caesar, Cicero, Seneca, Pliny the Younger and Tacitus) for a total of more than 300,000 tokens, and different genres (treatises, speeches, letters)
- The **test** set contains tokenized words, but not their tags. The texts are split into three sub-tasks:
  1. ***Classical*** sub-task: one text of the same period and genre as the training set (Livy)
  2. ***Cross-genre*** sub-task: one poetic text (Ovid, *Metamorphoses*) and one technical text (Pliny the Elder)
  3. **Cross-time** sub-task: one Renaissance prose text (Sabellicus, *De Latinae Linguae reparatione*, provided by Timo Korkiakangas)

The Corpus

The source of the corpus is the **LASLA**
(Laboratoire Statistique des Langues Anciennes, Université de Liège)

Its corpus contains nowadays 2,500,000 tokens. For each token, an alphanumeric string following internal LASLA conventions encodes metadata, lemmatization and morphological analysis.

The whole corpus has been converted from the **LASLA** into the **CoNLL-U** format, and:

- consequently, the corpus has been converted into the **Universal Dependencies** formalism
- specifically for **EvaLatin**, only a **subset** of morpholexical features has been retained
  - ⟶ the most stable and/or morphologically grounded ones
  - ↬ `Abbr`, `Aspect`, `Case`, `Degree`, `InflClass`, `Mood`, `Number`, `Person`, `Tense`, `Variant`, `VerbForm`, `Voice`

Tasks and results

- The participants had to perform
  - **Lemmatization**
  - **Part-of-Speech tagging**, and
  - (morphological) **Features Identification**

Tasks and results

- The participants had to perform
  - **Lemmatization**
  - **Part-of-Speech tagging**, and
  - (morphological) **Features Identification**

- **Two teams** participated in all tasks and sub-tasks:
  - `Kraków`, Jagiellonian University, Institute of Polish Language, Enelpol (Poland)
  - `KU-Leuven`, Katholieke Universiteit Leuven, Brepols Publishers (Belgium)

Tasks and results

- The participants had to perform
  - **Lemmatization**
  - **Part-of-Speech tagging**, and
  - (morphological) **Features Identification**

- **Two teams** participated in all tasks and sub-tasks:
  - `Kraków`, Jagiellonian University, Institute of Polish Language, Enelpol (Poland)
  - `KU-Leuven`, Katholieke Universiteit Leuven, Brepols Publishers (Belgium)

- The **results** for the three tasks are provided separately. For the Cross-Genre sub-task, the score corresponds to the macro-average accuracy.

Tasks and results

Lemmatization

Classical	
Kraków-open	97.26
Kraków-closed	96.45
KU-Leuven	85.44
Baseline (UDPipe)	80.36

Cross-Genre	
Kraków-open	95.08 (1.34)
Kraków-closed	91.62 (2.02)
KU-Leuven	86.48 (1.04)
Baseline (UDPipe)	79.03 (1.52)

Cross-time	
Kraków-open	92.15
Kraków-closed	91.68
KU-Leuven	84.60
Baseline (UDPipe)	81.92

Tasks and results

PoS tagging

Classical	
Kraków-open	97.99
Kraków-closed	97.61
KU-Leuven	96.33
Baseline (UDPipe)	78.23

Cross-Genre	
Kraków-open	96.06 (1.01)
Kraków-closed	94.62 (0.22)
KU-Leuven	92.31 (3.32)
Baseline (UDPipe)	76.58 (1.75)

Cross-time	
Kraków-closed	92.97
Kraków-open	92.70
KU-Leuven	92.11
Baseline (UDPipe)	74.26

Tasks and results

Feature Identification

Classical	
Kraków-open	95.46
Kraków-closed	95.42
KU-Leuven	69.91
Baseline (UDPipe)	24.98

Cross-Genre	
Kraków-open	89.43 (0.88)
Kraków-closed	89.32 (0.88)
KU-Leuven	60.55 (3.55)
Baseline (UDPipe)	23.34 (1.16)

Cross-time	
Kraków-closed	86.50
Kraków-open	86.50
KU-Leuven	60.09
Baseline (UDPipe)	27.84

And now, the participants

Gratias vobis agimus!

EvaLatin