AGILe: The First Lemmatizer for Ancient Greek Inscriptions

Evelien de Graaf, Silvia Stopponi, Jasper Bos, Saskia Peels-Matthey, Malvina Nissim
Centre for Language and Cognition Groningen, University of Groningen, The Netherlands

LREC 2022
Marseille
June 20-25

No available lemmatizer for ancient Greek inscriptions

- Ancient Greek: relatively low-resource, morphologically complex
- Lemmatization of inscriptions: potentiality of automatic analysis, e.g. advanced searches
- Few manually lemmatized corpora
- Fundamental texts for knowledge of the ancient Greek world

Challenges for lemmatization

- Durable materials
- Large number of texts

Testing Available Lemmatizers

Lemmatizers for AG trained and tested on literary texts: low performance on inscriptions

- GLEM (Bary et al., 2017)
- CLTK ‘default’ lemmatizer (Johnson et al., 2021): part of a Stanza-based pipeline, trained on PROIEL treebank (Haug and Jøhndal, 2008)
- CLTK ‘backoff’ lemmatizer (Burns, 2020): more lemmatizers in series, token-lemma lexica used
- UDpipe (Straka, 2018): pipeline for ancient Greek trained on Perseus treebank

AGILe: a Lemmatizer for AG Inscriptions

- Based on Stanza (Qi et al., 2020): dictionary-based lemmatizer + neural sequence-to-sequence lemmatizer
  Optional lexicon lookup
  - All entries from Liddell-Scott-Jones Lexicon + gold lemmas from training set
  - If predicted lemma not in lexicon: changed to first lemma in the lexicon, the closest for edit distance

AGILe: Error Analysis

- Manual analysis of ~250 errors over 750
- Difficulties for AGILe:
  - spelling: e.g. ἄρην for ἄρην
  - crisis, e.g. τάξις = τάξι + τάξι
  - low-frequency forms due to complex morphology
  - unique names (locations, persons, months...)
- False negatives:
  - wrong gold standard
  - output lemmas not identical to gold or variants of it, e.g. πρῶτος lemmatized as superl. πρῶτος ≠ gold πρῶτος
  - capitalization and accentuation: e.g. Φάρσαλος lemmatized as Φάρσαλος ≠ gold Φάρσαλος
  - ambiguous forms, more lemmas possible
  - e.g. σάτιον, ambiguous between σάτιον and σάτιον

Accuracy on CGRN

CGRN gold standard: wordforms and lemmas, no punctuation

AGILe best performing on AG inscriptions

- Accuracy: 84.7% dev. set, 85.1% test set
- 82.1% without lexicon lookup (dev. set)
- Comparison with the other lemmatizers, same CGRN test set (5K tokens)

Generalizability of AGILe

Tested on literary data

73.6% on PROIEL (13,314 tokens)
UDPipe obtained ~94% → AGILe specializes on inscriptions

Tested on other inscriptions

Cretan institutional inscriptions (similar timespan to CGRN, various kinds of texts, Vagionakis, 2021) - AGILe: 62.2%; GLEM 51.2%

Error analysis of 838 errors (268 unique):
- 513 false negatives, 61%!
- errors mostly due to different lemmatization conventions e.g. τάξιον lemmatized τάξιον instead of LSJ τάξιον

Hypothetical 85% acc. for AGILe

Future work

Integrate AGILe in a large corpus of inscriptions such as PHI
Improve performance:
- improve the lexicon lookup
- integrate POS-tagging
- retrain on more annotated data
- allow users to choose the correct lemma between options

Data

1. CGRN: 60-20-20 split (train – dev – test)
2. PROIEL treebank, Greek portion (Haug and Jøhndal, 2008): 88-6-6 split, no punctuation

Custom Rules

- h and p ignored
- κτο/ικ and ιτο/ικ converted to κ
- όι/ικ converted to ι

Results

Accuracy on literary texts

<table>
<thead>
<tr>
<th>Lemmatizer</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDPipe</td>
<td>45.0</td>
</tr>
<tr>
<td>CLTK</td>
<td>41.6</td>
</tr>
<tr>
<td>CLTKb</td>
<td>34.8</td>
</tr>
<tr>
<td>GLEM</td>
<td>61.9</td>
</tr>
<tr>
<td>AGILe</td>
<td>85.2</td>
</tr>
</tbody>
</table>

Need for a specific lemmatizer for ancient Greek inscriptions!

CONTACTS: e.de.graaf.6@student.rug.nl; s.stopponi@rug.nl; s.peels@rug.nl; m.nissim@rug.nl

https://github.com/agile-gronlp