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

ENCODE 2023
Leuven
April 24

Problem

No available lemmatizer for **ancient Greek inscriptions**

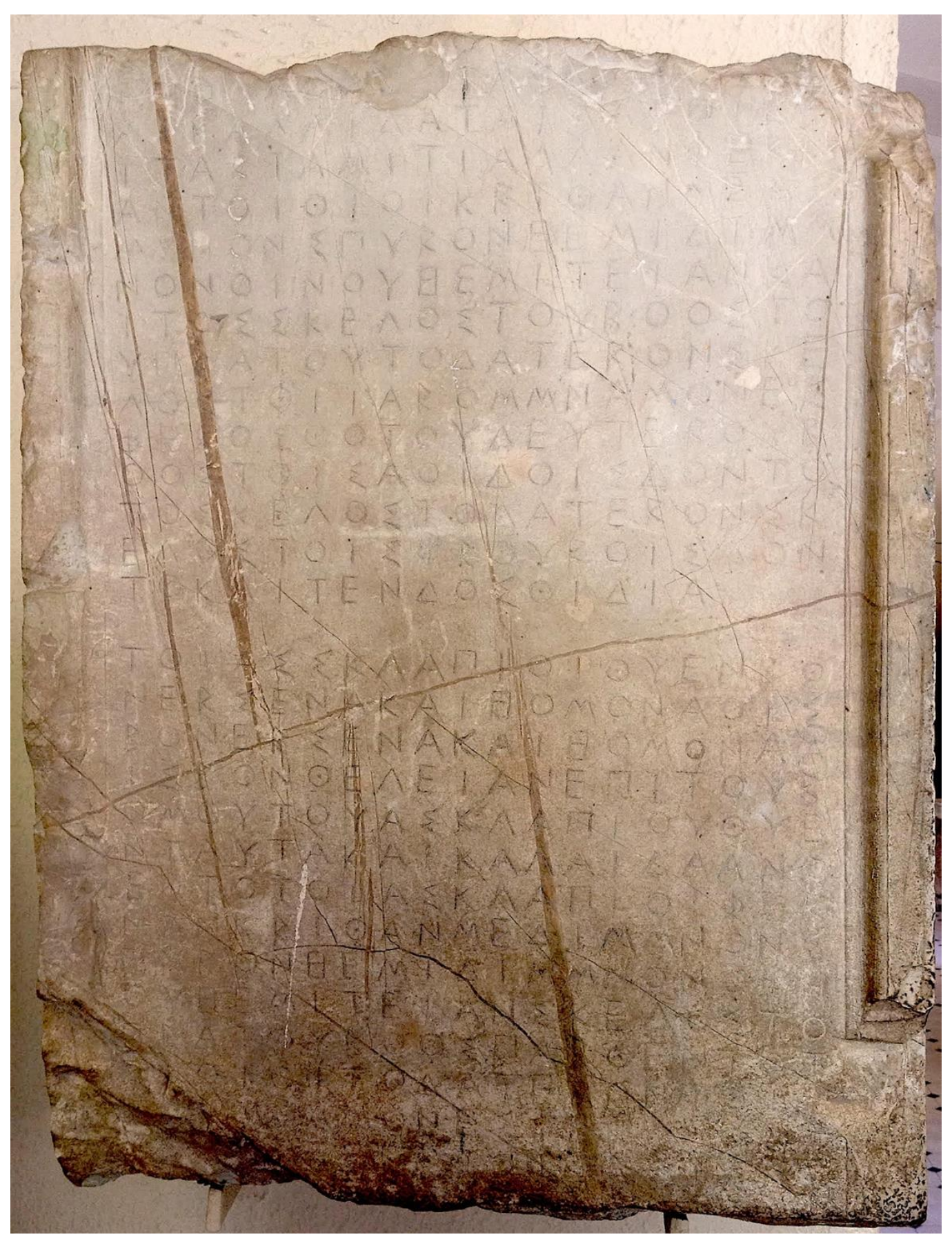
- **Ancient Greek**: relatively low-resource, morphologically complex
- Lemmatization of inscriptions: potentiality of **automatic analysis**, e.g. **advanced searches**
- Few **manually** lemmatized corpora
- **Fundamental** texts for knowledge of the ancient Greek world

Ancient Greek Inscriptions

- **Durable** materials
- **Large number** of texts

Challenges for lemmatization

- **Differ from literary** texts (orthography, morphology, dialectal variation)
- **No standard alphabet** before 4th cent. BCE:
  - cluster **/ks/** spelled as χ, ξ, χσ, κσ
  - characters ***h*** (aspiration) and **ϝ** (sound /w/)
  - no difference between **short** and **long** vowels, e.g. long and short /o/ written as o



Stone with an ancient Greek inscription (CGRN 34, end 5th cent. BCE).

CGRN

A Collection of Greek Ritual Norms (CGRN)

- **225** normative texts
- **Religious rituals**
- **6th** century BCE - **1st** century CE
- Large **topographical** spread
- **TEI XML**, **EpiDoc**-compliant files
- 38K tokens, **25K manually lemmatized**
- Lemmas: base forms from Greek-English Lexicon **Liddell and Scott** (1940)

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 and PROIEL treebank

Reported Accuracy on Literary Texts

Lemmatizer ↓ | Test data →	Herodotus	Thucydides	Homer	Lysias	PROIEL	Perseus
GLEM punctuation (a, b)	**95.7**	93.0	72	81	-	-
GLEM no punctuation (b)	-	-	84	94	-	-
CLTK (a)	78.7	76.6	-	-	-	-
CLTK backoff (b)	-	-	91	97	-	-
CLTK (b)	-	-	65	65	-	-
UDPipe 2.0 (c)	-	-	-	-	94.0	91.9
UDPipe 2.3 (c)	-	-	-	-	93.5	85.0

Accuracy of all lemmatizers on all test data.
Sources: **a.** Bary et al. (2017); **b.** Vatri and McGillivray (2020); **c.** Straka et al. (2019a), Straka et al. (2019b).

Accuracy on CGRN

CGRN gold standard: wordforms and lemmas, no punctuation

System	Accuracy	Wrong	Correct	Missed
UDPipe Perseus	46.3	13,474	11,606	149
UDPipe PROIEL	47.3	13,263	11,912	60
CLTK	46.4	13,390	11,581	258
CLTKb	37.1	15,768	9,292	169
GLEM	**62.5**	9,379	15,650	200

Accuracy of the four lemmatizers tested on the CGRN.

Need for a specific lemmatizer for ancient Greek inscriptions!

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**

Results

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

AGILe best performing on AG inscriptions

Lemmatizer	Accuracy
UDPipe PERS	45.0
UDPipe PRO	46.2
CLTK	41.6
CLTKb	34.8
GLEM	61.5
AGILe	**85.1**

Custom Rules

- *h* and ϝ ignored
- κ+σ/ς and χ+σ/ς converted to **ξ**
- ϕ+σ/ς converted to **ψ**

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

AGILe: Error Analysis

- Manual analysis of ~**250 errors** over 750
- **Difficulties for AGILe:**
  - **spelling**, e.g. ἀρέν for ἀρήν
  - **crasis**, e.g. κἀπί = καί + ἐπί
  - **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. πρώτει lemmatized as superl. πρῶτος ≠ gold πρότερος
  - **capitalization** and **accentuation**
    e.g. Φηραίωι lemmatized as Φηραῖος ≠ gold Φηραίος
  - **ambiguous** forms, more lemmas possible
    e.g. σιωπῆι, ambiguous between σιωπάω and σιωπή

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. τύχαι lemmatized τύχα instead of LSJ τύχη

Hypothetical 85% acc. for AGILe

Future work

Integrate AGILe in a large corpus of inscriptions such as PHI or IG Online

Improve performance:

- improve the lexicon lookup
- testing other models
- retrain on more annotated data (25 new inscriptions added to CGRN)
- testing AGILe on other diverse corpora of inscriptions such as IGCyr, GVCyr, and Inscriptions of Aphrodisias

https://github.com/agile-gronlp

CONTACTS: evelien.degraaf@kuleuven.be; s.stopponi@rug.nl; s.peels@rug.nl; m.nissim@rug.nl