En makroniserare för antik grekiska

Albin Thörn Cleland,
The Swedish Graduate School of Digital Philology,
Centre for Languages and Literature, Lunds universitet
Och tillsammans med Eric Cullhed för AI-delen!

DIGPHIL
THE SWEDISH GRADUATE SCHOOL OF DIGITAL PHILOLOGY

LUNDS UNIVERSITET

Bakgrund

- Latin har sedan Winge 2015 haft en makroniserare, men grekiska har fortfarande ingen!
- För grekiska gäller det att avgöra längden hos de tre “dikrona” vokalerna alfa, jota och ypsilon.

LUNDS UNIVERSITET

Frågeställningar

- Hur väl kan man makronisera ett stort grekiskt korpus algoritmiskt?
- Och går det sedan att förbättra resultatet ytterligare genom maskininlärning?

Metod för algoritmiskt program

- Först morfologisk markup genom modellen odyCy (Kardos et al., Center for Humanities Computing Aarhus)
- 13 moduler av makronisering, av tre typer:
  - Databaser
    - Egen custom-ordlista
    - Wiktionary
    - LSJ
    - Hypotactic
  - Morfologiska regler
    - Nominalformer
    - Verbformer
    - 3 accentregler (σωτῆρᾰ-regeln och proparoxyton-regeln)
    - Prefix
  - Rekursion (ärva vokallängder från mkt liknande ordformer)
    - Dubbel-accent
    - Elision
    - Kasusbyte
    - Grav-till-akut
    - Versal-till-gemen

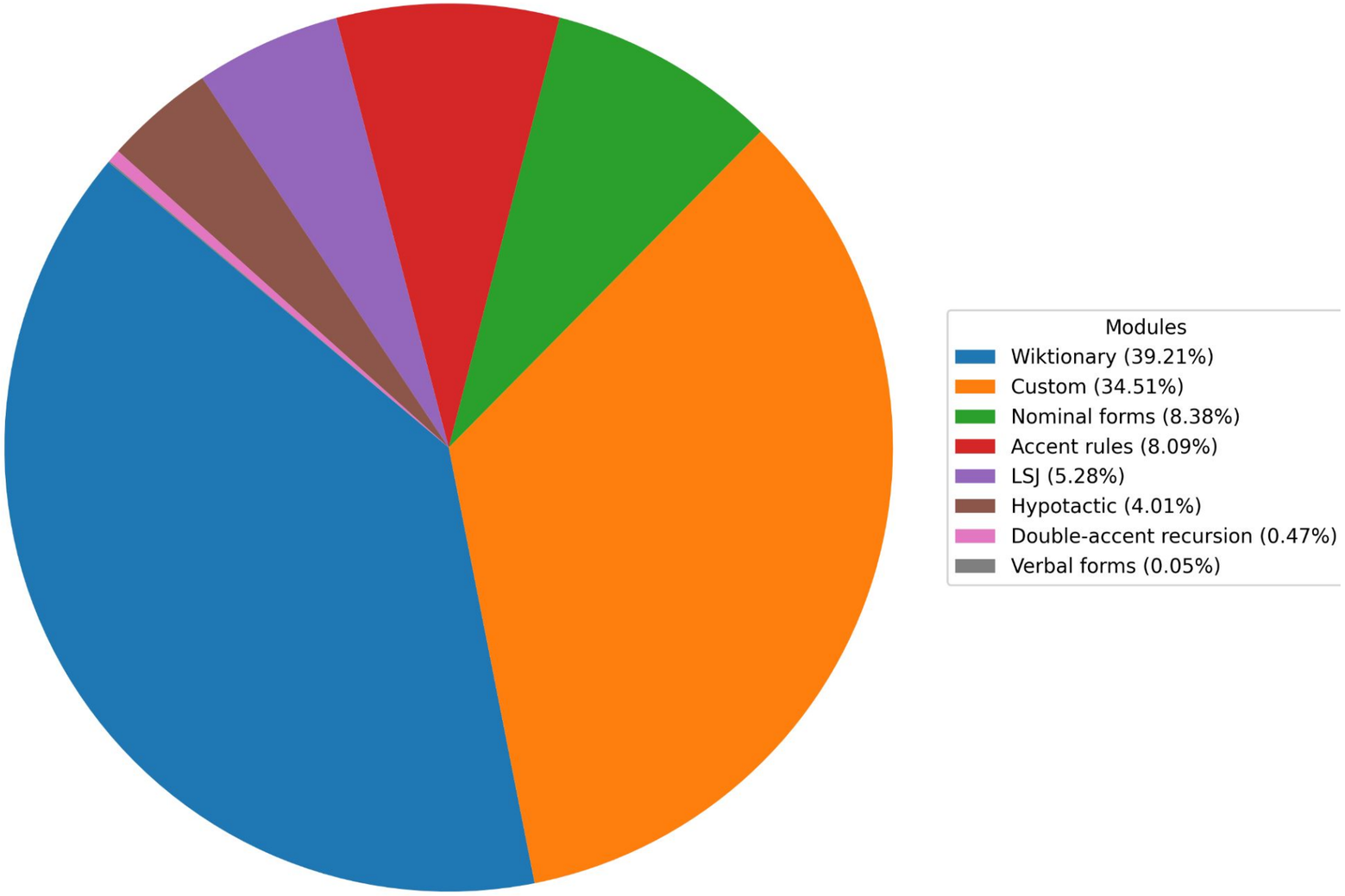
LUNDS UNIVERSITET

Resultat för algoritmiskt program

- På ett mycket stort korpus (*First1KGreek*, *PatristicTextArchive* och Perseus' *canonical-greekLit*) blir **76%** makroniserat. Detta säger dock inget om hur stor del som är **korrekt**!
- Vi introducerar därför *Norma*, ett benchmark för stavelsegränser och vokallängder (makronisering).
- Manuellt (av mig) syllabifierade och makroniserade texter. Runt 20 rader vers och prosa från 14 olika författare, plus runt 1000 rader av Aristofanes sånger.
- Om vi gissar "kort" på allt som ej makroniserats, får vi 91.2% på Norma. Att jämföra med att gissa "kort" på alla längder: 67.8%.

LUNDS UNIVERSITET

Vilka moduler makroniserade mest av korpusen?



LUNDS UNIVERSITET

Metod för maskininlärning

- Träna på redan makroniserad korpus (ignorera de 24% som ej makroniserades). Vi använder Alvis genom varsitt beräknings-grant från NAISS (**National Academic Infrastructure for Supercomputing in Sweden**).
- Hypotes: förträning kommer ge bättre resultat
- Tränar två modeller parallellt: en förtränad större modell (ModernBERT) och en icke-förtränad väldigt liten modell, RoBERTa.

Resultat (Norma) för maskininlärning

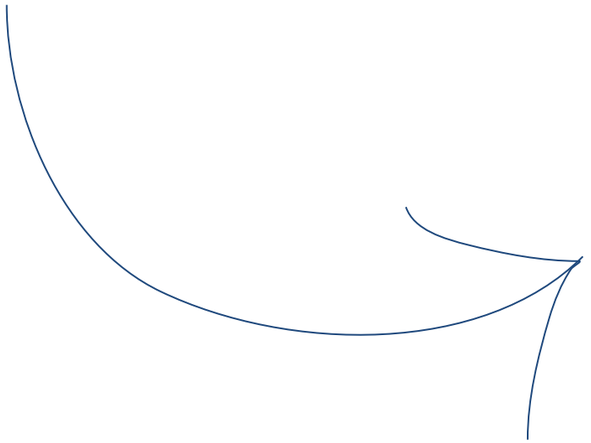
- ModernBERT: 95.2%
- RoBERTa: 95.0%
- Jmfr. igen algoritmiska: 91.2%

Slutsatser

- **Maskininlärning hjälper!** Vid höga siffror är varje procent mycket värd och representerar icke-triviala vokallängder.
- **Förträning ger marginell vinst!** Vilket är glädjande, eftersom icke-förtränade modeller som RoBERTa kan tränas och ge inferens på en vanlig CPU.

Resurser

- https://huggingface.co/Ericu950/SyllaMoBert-grc-macronizer-v1
- https://huggingface.co/Ericu950/macronizer_mini
- Direkt för Python: pip install grc-macronizer
- https://github.com/Urdatorn/grc-macronizer



LUNDS UNIVERSITET

LUNDS
UNIVERSITET