EstNLTK muutuste tuules: 1.4 $\rightsquigarrow$ 1.6

Sven Laur[♢,♠]

https://github.com/estnltk

[♢]Tartu Ülikool

[♠]Tarkvara Tehnoloogia Arenduskeskus OÜ

Mis on EstNLTK projekt?

- Vabavaraline Pythoni teek eestikeelsete tekstide töötlemiseks
  - sõnestamine, lausestamine
  - morfoloogiline ja süntaktiline analüüs
  - märgenduskihid fraaside eraldamiseks
- Eraldiseisvad rakendused standardsete probleemide lahendamiseks
  - tekstide klassifitseerimine (textclassifier)
  - oluliste märksõnade tuvastamine (volcanoplot)
  - näitemärgenduste koostamine (ner-tagger, gap-tagger)
- Keeleresursid ja testandmestikud
  - Lünktestid kontekstide ennustamiseks (word embeddings)
  - Nimeolemite ennustamise testandmestikud (named entity recognition)

EstNLTK 1.4 teek

- Andmeformaat märgenduskihtide salvestamiseks
  - tähe põhised märgendus elemendid
  - märgenduselemendile vastavad attribuudid
  - märgenduselementidest koosnevad märgenduskihid
- Liidesed olemasolevatele analüsaatoritele
  - lausestaja
  - morfoanalüsaator (vabamorf)
  - pindsüntaktiline analüüs (visl, maltparser)
- Uued komponendid
  - fraaside märgendajad (ajaväljendid, verbiahelad, ...)
  - keerulised fakti- ja fraasieraldusalgoritmid

EstNLTK 1.4 märgenduskihid

- **paragraphs**
- **sentences**
- **words**
  - analysis
    - ..., form, lemma, ...
  - clause_index
  - ner_label
- **gt_words**
  - analysis
    - ..., form, lemma, ...
- **verb_chains**
- ...

Puudused ja probleemid

- Sõnadekiht (**words**) ja lausetekiht (**sentences**) võivad olla nihkes

words:	Ta läks kell 14 . 30 ära
sentences:	Ta läks kell 14. 30 ära

- Osad kihid (analysis, clauses, . . .) on esitatud atribuutidena
  - **words**
    - analysis
      - . . .
      - form
      - lemma
      - . . .

  kuid baasaoperatsioonid on defineeritud kihtidel

Puudused ja probleemid

- Süsteemitus mitmesuste esitamisel. Kombinatoorne plahvatus
  - **words**
    - analysis

EstNLTK 1.6 märgenduskihid

- **words**
  - **paragraphs**
  - **sentences**
  - **normalised_words**
  - **morf_analysis**
    - ..., form, lemma, ...
  - **pronouns**
    - type
  - **finite_forms**
    - ∅
  - **verb_chains**
  - ...
- ...

Mitmesuste käsitlemine

- Kihis võib võib ühele märgenduselemendile vastata mitu tõlgendust

tee →

lemma	POS	form	...
tee	S	sg n	...
tee	S	sg g	...
tegema	V		...

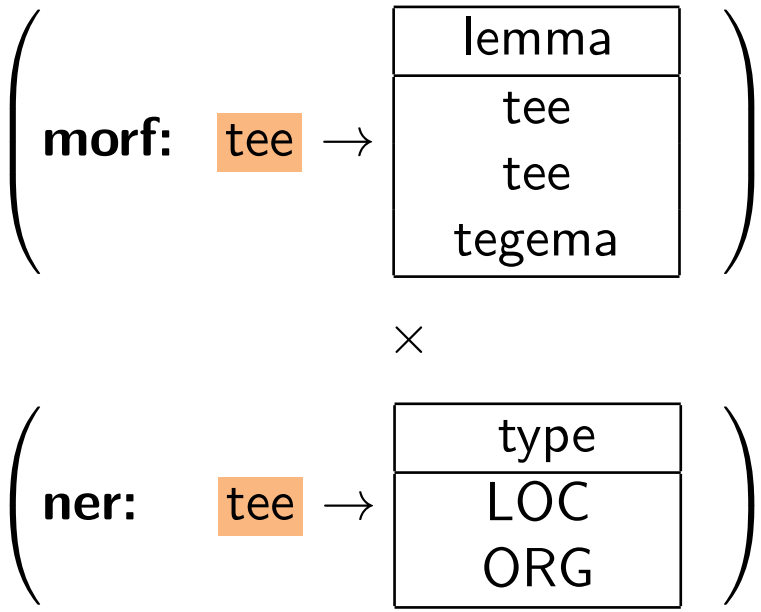
Iga elemendi tõlgendusele vastavad konkreetsed atribuutide väärtused

- Erinevate elementide tõlgendused kihis on kombineeruvad sõltumatult

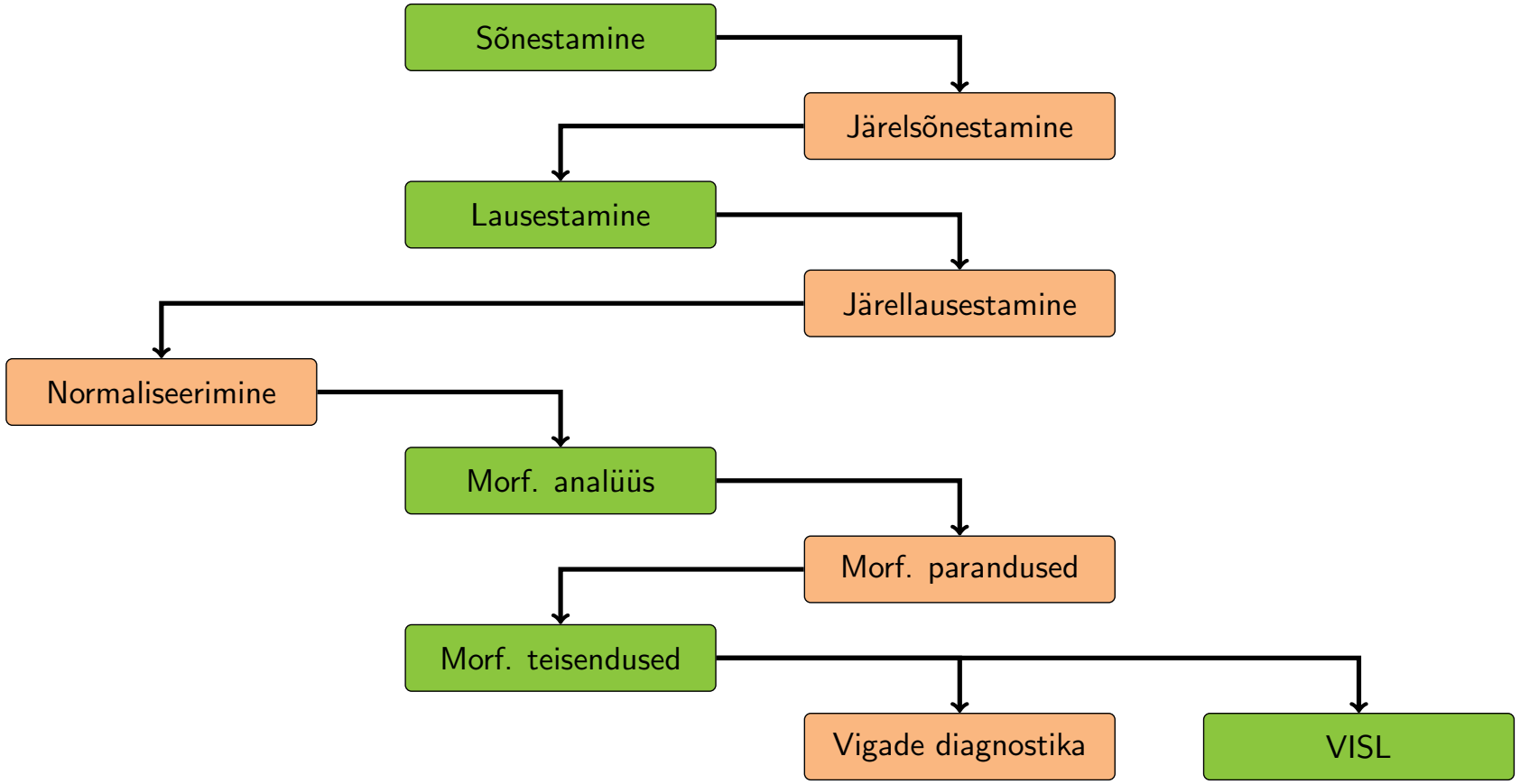
(tee → | lemma | tee / tegema |) × (aja → | lemma | aeg / ajama |)

Mitmesuste käsitlemine

- Erinevate kihtide tõlgendused on kombineeruvad sõltumatult



Robustne tekstianalüüs töövoog



Milleks on sellist analüüsi vaja?

Probleem: Koondkorpuses on üle 10 000 erineva asesõna

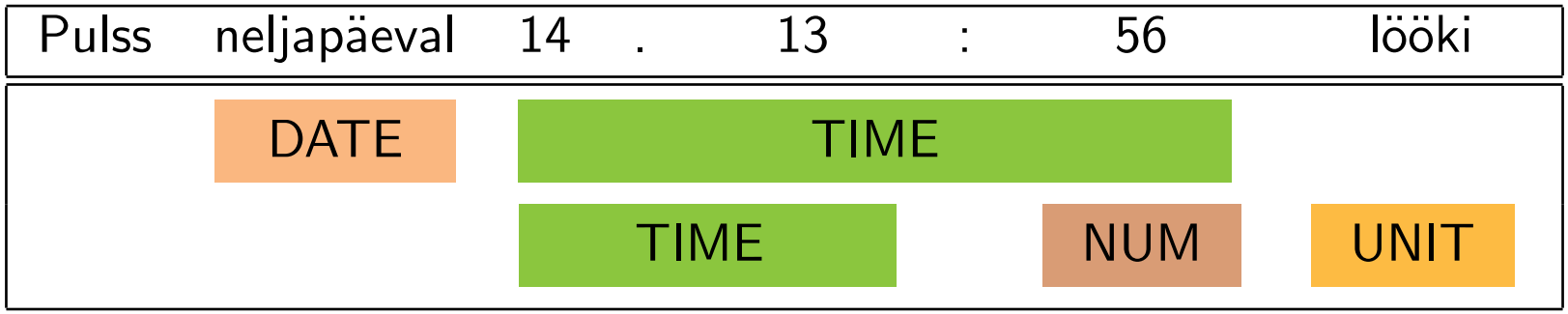
- ▷ tege-ma → lemma: tege-mina
- ▷ 1203-me → lemma: 1203-meie
- ▷ kes-teab-mis → lemma: kes-teab-mis

Olemasolev töövoog ei sisalda standardseid teksti puhastussamme

- ▷ See muudab raskemaks tüüpiliste tekstide analüüsi
- ▷ Iga üks meist leiutab samu tekstide puhastamise võtteid
- ▷ Ei jää aega korpuse spetsiifiliste probleemide tuvastamiseks

Fraaside eraldamine

Pulss neljapäeval 14.13:56 lööki



- ▷ Huvitavad fraasid on kirjeldatavad *lõpliku* grammatikaga
- ▷ Fraaside parsimisel olev sõnestus pole ühene ega ülekatteta
- ▷ Sõnestuse ühestamiseks on mõistlik kasutada grammatikat
- ▷ Reeglite prioriteete on kasulik annoteerida tõenäosustega

Kokkuvõte

Hea

- EstNLTK liides ühtlustub
- EstNLTK kasutamine muutub ettearvatavamaks
- EstNLTK liidestub erinevate veebikeskkondadega
- EstNLTK katab operatsioonid sõnestusest kuni süntaksianalüüsini

Halb

- EstNLTK liides liides muutub
- EstNLTK nõuab Pythoni versiooni 3.5
- Kõik EstNLTK 1.4 olevad komponendid ei jõua EstNLTK 1.6

Täname

Arendajad

- Siim Orasmaa
- Timo Petmanson
- Uku Raudvere
- Dage Särg
- Paul Tammo
- Aleksandr Tkatšenko

Nõustajad

- Heiki-Jaan Kaalep
- Kadri Muischnek
- Tarmo Vaino