

# ESTNLTk

<https://github.com/estnltk/estnltk>

Siim Orasmaa  
Tartu Ülikool

# Mis on EstNLTK?

- Vabavaraline Pythoni teek eestikeelsete tekstide töötlemiseks
  - baastööriistad ja keeletöötluste töövood:
    - nt lausestamine, lemmatiseerimine, nimeüksuste ja ajaväljendite tuvastamine
  - raamistik tekstimärgenduste haldamiseks
    - salvestamine / võrdlemine / visualiseerimine
  - ⇒ “ehitusklotsid” keerukamate rakenduste loomisel:
    - nt faktieraldus / sisuanalüüs / hoiakute-meelestatuse analüüs

# EstNLTK paigaldus ja kasutamine

- Paigaldamine:
  - Anaconda pakett (soovitatav)
  - Värskeim versioon github-ist (peamiselt häkkeritele)
- Litsents - GPLv2
  - Tuletatud töö: kasutada sama vabavaralist litsentsi;
  - Kommertslik kasutus: sisekasutus, nt teenuse/serveri koosseisus;
  - Muud juhud: kontakteeruda Tartu Ülikooliga

# EstNLTK ~~hargnemine~~ arenemine

- versioon 1.4
  - + palju erinevaid tööriistu (sh morf analüüs, süntaks, IE, semantika)
  - ebaühtlane liides; probleemid kihtide ja mitmesuste esitamisel
- **versioon 1.6**
  - + ühtlane liides, (parendatud) tööriistad sõnestusest süntaksianalüüsini;
  - beta staadiumis; osa v1.4 tööriistu veel puudu

# EstNLTK ver 1.6(.3)

- Robustne keeletöötluste töövoog
  - hulk valmiskujul reegleid:
    - sõnestuse parandused, nt
      - numbrid, lühendid, www- ja e-maili aadressid
    - lausestuse järelparandused
  - võimalik lisada ise valdkonnapõhiseid parandusi

# EstNLTK ver 1.6(.3)

- PostgreSQL andmebaasi tugi
  - Kiired päringud info ekstraheerimisel
  - Märgendatud tekstide salvestamine
    - märgenduste versioneerimine
    - kihtide indekseerimine
    - märgenduste võrdlus (tööriistade arendamisel)

# EstNLTK ver 1.6(.3)

- Uus morfoloogiline analüsaator (*HFST*)
  - detailsem liitsõna osade analüüs, nt:
    - virumaalane => *Virumaa + lane*
    - mäesuusatamine => *mägi + suusatama + mine*
- Süntaktilise analüüsi tugi ( *VisICG3* )
- Ajaväljendite tuvastaja uuendatud liides
  - võimalik kasutada omakirjutatud reegleid

# Viiteid

- Lähtekood ja installeerimisjuhised:
  - <https://github.com/estnltk/estnltk>
- Õppematerjalid / kasutusjuhised:
  - <https://github.com/d009/EstNLP>  
(ver 1.4, eesti k)
  - [https://github.com/estnltk/estnltk/tree/version\\_1.6/tutorials](https://github.com/estnltk/estnltk/tree/version_1.6/tutorials)  
(ver 1.6, inglise k)



# Täname

- Projektijuht                      Sven Laur
- Arendajad  
Timo Petmanson              Siim Orasmaa                      Dage Särg  
Paul Tammo                      Aleksandr Tkatšenko              Uku Raudvere
- Nõuandjad  
Heiki-Jaan Kaalep              Kadri Muischnek  
Kairit Sirts                      Tarmo Vaino