Töövahendid ja tekstitöötlus

Kristel Uiboaed, Liina Lindström, Kadri Muischnek

25. jaanuar 2017

Sissejuhatus

- Tööriistad, nende plussid ja miinused
- Tekstitöötluse ABC
- Praktikum: sagedusloendid ja kollokatsioonid

Linuxi shell

- Plussid
  - Lihtne õppida
  - Intuitiivne süntaks
  - Sobib väga hästi lihtsamate teksti puhastamise ja esmase teksti töötluse ülesanneteks
  - Lihtsaid asju saab teha kiiresti ja lihtsalt
- Miinused
  - Keerulisemate ülesannete täitmine raske
  - Piiratud funktsionaalsus
  - Windowsis üsna tülikas kasutada
  - Aeglane

R

- Plussid
  - Võimalik täita kõiki ülesandeid: teksti puhastamine, analüüs, statistika, visualiseerimine jne.
  - Palju statistilise analüüsi võimalusi
  - Suur ja aktiivne kogukond
  - Üsna lihtne õppida
  - Vt ka
- Miinused
  - Harjumatu, kui juba oskad mõnda programmeerimiskeelt
  - Suuremahuliste tekstiandmete töötlemine aeglane
  - Tekstitöötlus on “koodirohke”

Python

- Plussid
  - Programmeerimiskeel
  - Tekstitöötlus kiire
  - Võimalik täita kõiki ülesandeid: teksti puhastamine, analüüs, statistika, visualiseerimine jne.
  - Sobib pea kõige jaoks
- Miinused
  - Õppimisprotsess pikem
  - Faili- ja tekstitöötlus natuke rohkem trükkimist nõudev
  - Visualiseerimine ja statistika tunduvalt tülikam kui nt R-s

Nn valmistööriistad

- Wordsmith
- AntConc
- SketchEngine (tasuline, 30 päeva prooviversioon tasuta). Olemas on morfoloogiliselt analüüsitud eesti keele korpused.
- Ja väga paljud teised . . .

Mis räägib olemasolevate vahendite kasuks?

- Üsna mõistlik valik ingliskeelsete tekstidega töötamisel
- Lihtsamad asjad on võimalik ära teha kiiresti
- Kiiresti õpitavad

. . . ja nende kahjuks?

- Eesti keele morfoloogia (*andma*, *annan*, *antud*, *inimene*, *inimese*, *inimestega*)
- Pea alati on vaja lemmatiseerida (ja lemmatiseeritud teksti saab kasutada ka valmisrakendustega)
- Kodeeringu probleemid pea alati garanteeritud (täpitähed)
- “Seletamatud jamad”, kuna koodi ei näe, siis on väga raske jälile saada potentsiaalsele probleemile ja halvemal juhul ei saa üldse teada, et mingi probleem olemaski on.

Eesti keele jaoks olulisi tööriistu

- Estnltk
  - Morfoloogiline analüüs
  - Kitsenduste grammatika analüsaator
  - Maltparser
- Vabamorf
- Formaadi teisendajad (nt märgendamise kujusid on erinevaid)
- Süntaksi tööriistad

Kust õppida?

- Humanities Data in R
- Quantitative Corpus Linguistics with R: A Practical Introduction
- Text Analysis with R for Students of Literature
- MOOC-id: Coursera, DataCamp, Codacademy jmt.

"Like families, tidy datasets are all alike but every messy dataset is messy in its own way." (Wickham 2014: 2)

Wickham, Hadley (2014). **Tidy data**. Journal of Statistical Software 59(10).

Tekstitöötluse ABC

- Tutvu algmaterjaliga
- Uuri, milliseid sümboleid märke jmt tekst sisaldab
- Tee erinevaid sagedusloendeid
- Kas sõnad on “kokku kleepunud”?
- Kodeering, kodeering!
- . . . ja igasugu muud jamad
- Otsused:
  - Kas suured ja väiksed tähed ühtlustada (pärisnimed, lause algus)?
  - Mida teha nubmritega, punktuatsiooni jm märkidega (*25-aastane, 35%*)?
  - Olemid