

Vallakohtuprotokollide kasutusvõimalustest

Gerth Jaanimäe, Siim Orasmaa, Maarja-Liisa Pilvik
Tartu Ülikool

Rahvusarhiivi uurijatund
12.05.2021

Taust

- Eesti 19. sajandi vallakohtuprotokollid on tänu oma temaatikale, iseloomule ja süstemaatilisusele üheks olulisemaks eesti keele ja kultuuri ajaloo uurimise objektiks.
- Protokollides peegeldub mh
 - tollase talurahva eluolu,
 - majanduslik seisund,
 - kõikvõimalikud üleastumised ja elukorralduse häired,
 - üldine mentaliteet,
 - sotsiaalsed võrgustikud,
 - eesti kirjakeele kujunemise protsess
- 2019. a kevadel Rahvusarhiivis käivitatud ühisloomeprojekt
- 2019. a sügisel TÜ projekt “Ajalooliste tekstide automaatse analüüsi võimalused 19. sajandi vallakohtuprotokollide näitel”

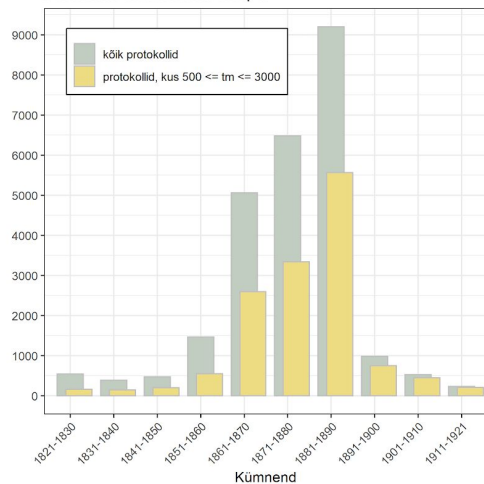
Taust

Projekti põhieesmärgiks on

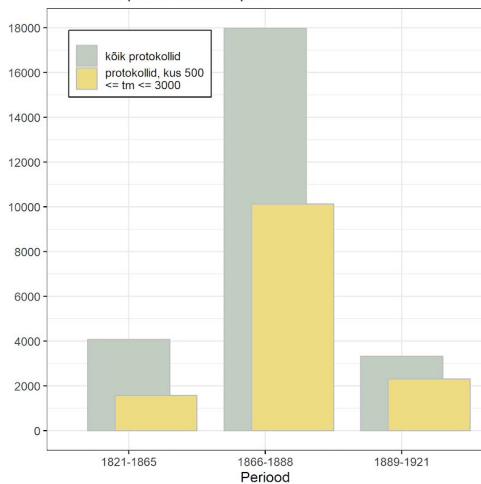
- vallakohtuprotokollide kasutusvõimaluste laiendamine.
 - Hetkel tekstid otsitavad tekstijärjendite järgi, isiku- ja kohanimede põhjal päringute tegemine sõltub puhtalt sisestajate kohusetundlikkusest ja süsteemsusest nimede märkimisel.
 - Projekti käigus lisame lingvistilise märgenduskihi ning automaatselt märgendatud nimeüksused.

Täna räägimegi pisut sellest, kuidas seda teha ja milleks see vajalik on.

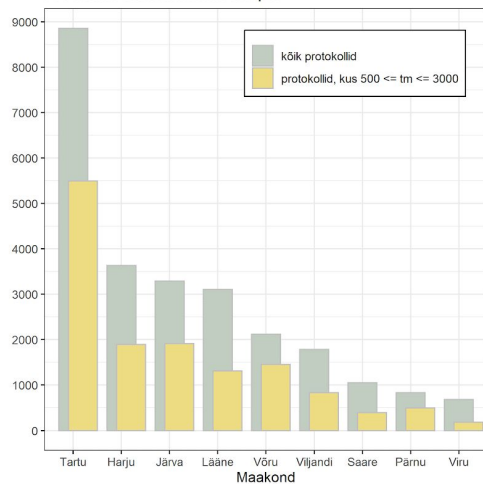
Protokolle kümnendite kaupa



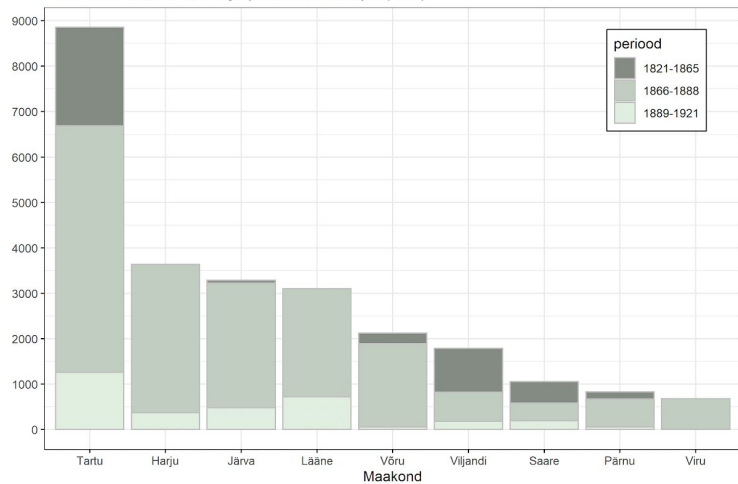
Protokolle perioodide kaupa



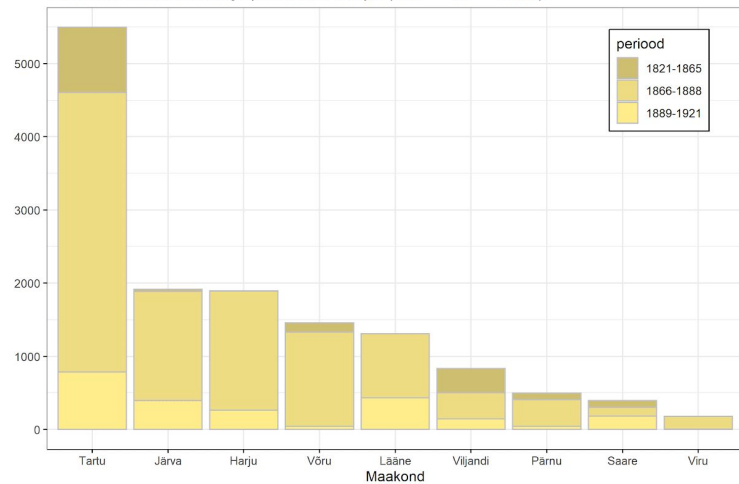
Protokolle maakondade kaupa



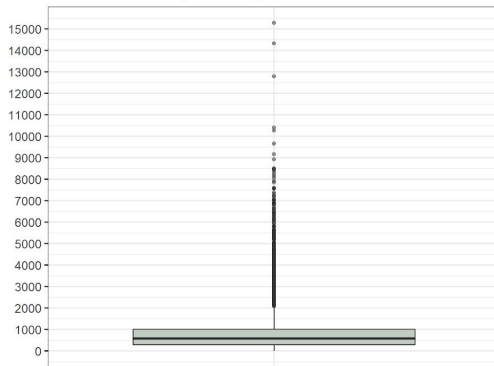
Protokolle maakondade ja perioodide kaupa (kõik)



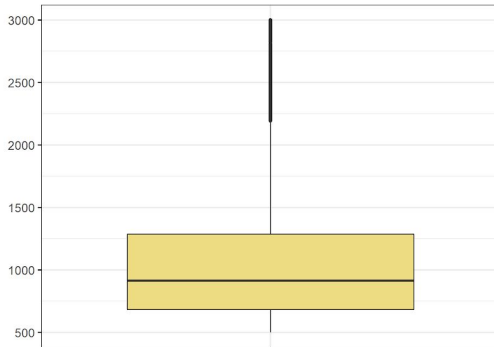
Protokolle maakondade ja perioodide kaupa (500 ≤ tm ≤ 3000)



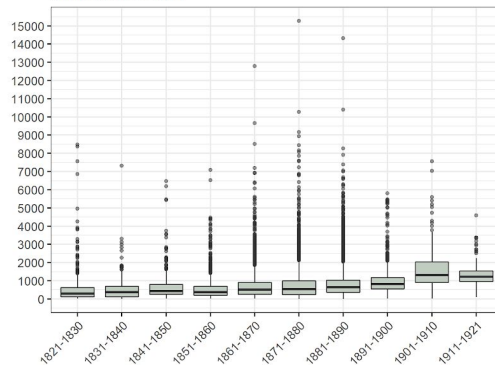
Tähemärkide arvu jaotumine protokollides



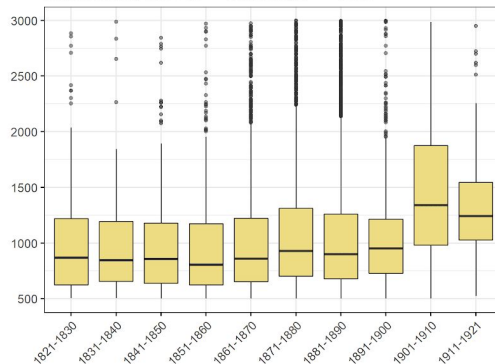
Tähemärkide arvu jaotumine protokollides, kus 500 ≤ tähemärgid ≤ 3000



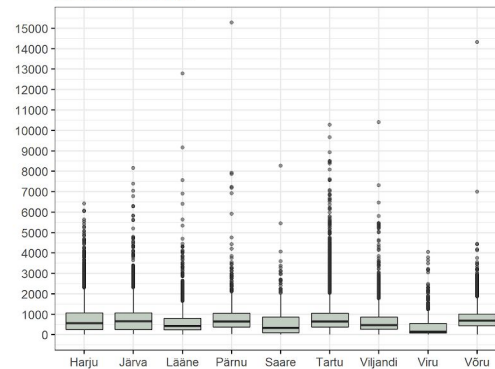
Tähemärkide arvu jaotumine protokollides kümnendite kaupa



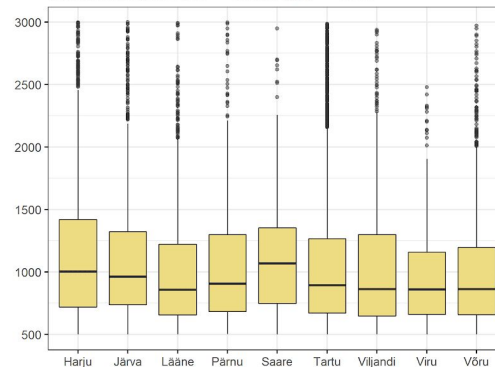
Tähemärkide arvu jaotumine kümnendite kaupa protokollides, kus 500 ≤ tähemärgid ≤ 3000



Tähemärkide arvu jaotumine protokollides maakondade kaupa



Tähemärkide arvu jaotumine maakondade kaupa protokollides, kus 500 ≤ tähemärgid ≤ 3000



Karbi sisse jääb 50% protokollide tähemärkide arv. Paks horisontaalne joon näitab mediaani.

Vurrud katavad 75% protokollide tähemärkide arvu. Punktid näitavad erindeid (antud juhul erandlikult kõrgeid väärtusi).

Ajalooliste tekstide automaatanalüüsist

Ajalooliste tekstide automaatanalüüs

Põhimõtteliselt kaks lähenemist:

- kohandada vahendid tekstide järgi
- kohandada tekstid vahendite järgi -> normaliseerimine

Tekstide normaliseerimine

- Sõnade teisendamine mittestandardset kujult õigekirjanormidele vastavaks.
- Enamasti kasutatakse netikeele või varasemalt kasutusel olnud kirjaviisis tekstide kontekstis.

Milleks kasulik?

- Võimalik tekste tänapäeva keele analüüsivahenditega uurida,
- otsida märksõnu
- jpm.

Tekstide normaliseerimine

Normaliseerimise teeb keeruliseks

- suur varieeruvus (sh vallakirjutajate vahel, sisestajate vahel);
- osa tekste vanas kirjaviisis (nt *sanud*, *temma*, *lubb*, *se*, *luk*),
osa uues kirjaviisis (nt *saanud*, *tema*, *lubas*, *see*, *lukk*),
ning osa n-ö üleminekuperioodist:

Tulli ette Johann Alajaan, kes käes olewal aastal wäe teenistuse kohuse...

- põhja- ja lõunaeesti keel.

Normaliseerimismeetodid

- Tekstide normaliseerimiseks on erinevaid meetodeid:
 - sõnastikud,
 - reeglipõhised lähenemised,
 - masintõlge,
 - jt.

Normaliseerimismeetodid

Masintõlge

- Vana ja uut kirjaviisi käsitletakse eraldi keeltena.
- Sõnu käsitletakse kui lauseid ja tähemärke kui sõnu.
- Nii saab n-ö tõlkida erinevaid täheühendeid ja nendega seonduvaid mustreid.
- Vaja treeningandmeid.

Probleemid

- Treeningandmeid võib olla liiga vähe.
- Vana kirjaviis on mitmeti mõistetav, nt *kalla* tähendab nii *kala* kui ka *kalla!*
 - probleemi lahendamiseks võib kasutada n-gramme või kollokatsioone

Esiialgsed normaliseerimiskatsed

- N-ö baastõlge, keskmine korrektsus 85%.
- Kogu korpus tõlgitud ja morfoloogiliselt analüüsitud (81% vastas tänapäeva keelele).
- Tõlkimata korpuses oli vastav protsent 70%.
- Bigrammid: testhulgal korrektsused väga kõikuvad (69-87%).
- Tõenäoline põhjus: pikemad tekstisõned tekitavad paratamatult rohkem vigu.

Näide probleemsest lausest

Mis võib valesti minna?

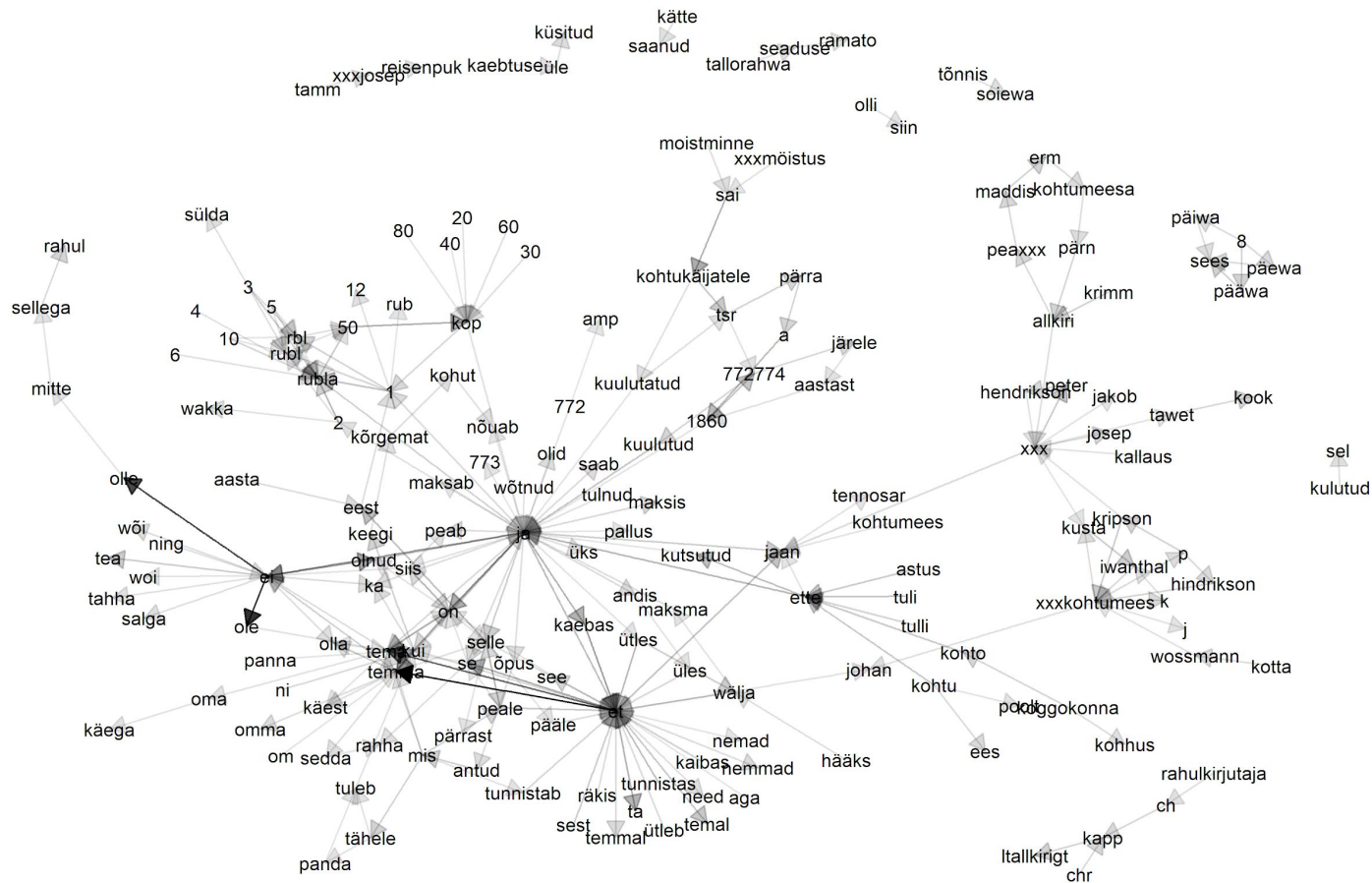
Ette tuli Jakob Kerra ja kaibas et Andri Uggur ollewat tedda ilma asjanda temma krami wälja pillanu nink möllemba naisega tedda pesnud.

Paar näidet analüüsivõimalustest

Tartumaa protokollide bigrammide võrgustik

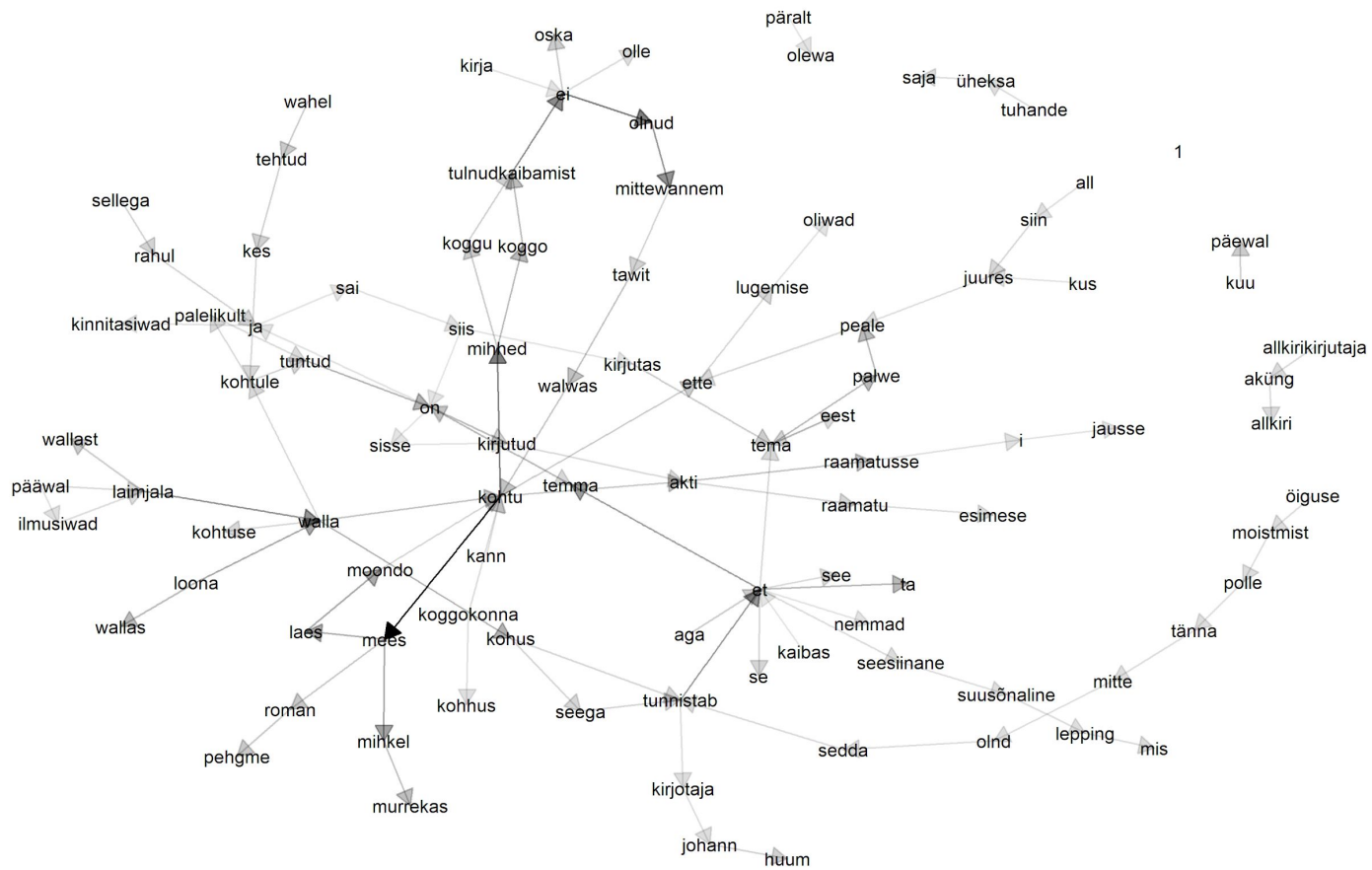
Normaliseerimata algtekstid

Milliseid protokollides
korduvaid vormeleid
on võimalik
tuvastada?



Saaremaa protokollide bigrammide võrgustik

Normaliseerimata algtekstid



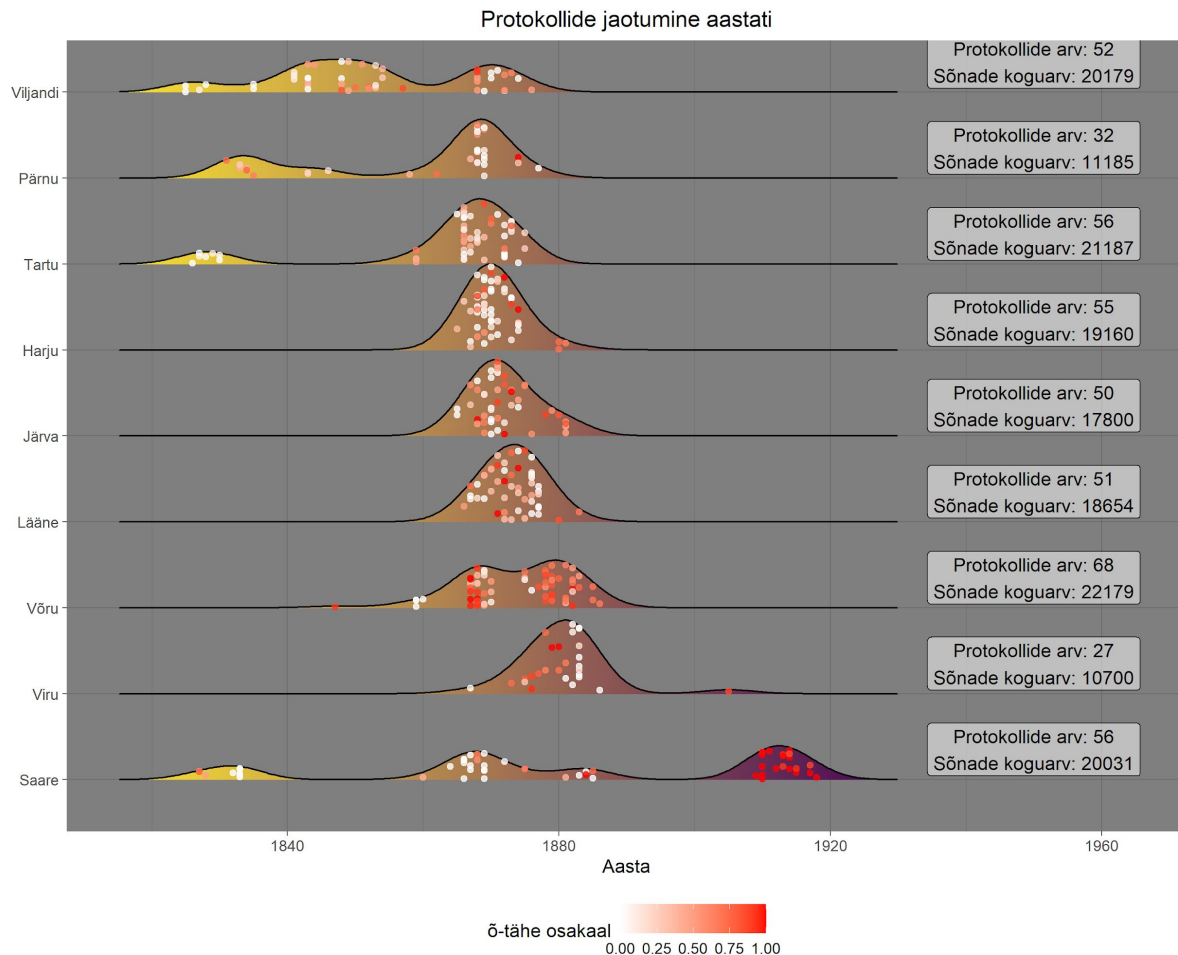
õ-tähe kasutus vallakohtuprotokollides

Automaatselt analüüsitud ja käsitsi
parandatud protokollide tekstid,
nn treeningandmed
(hetkel parandatud 448 protokoll)

Võimalik võrrelda
standardiseeritud sõnakujusid ja
originaali sõnakujusid

Näiteks

Moisa mõisa mõis 0 S sg g



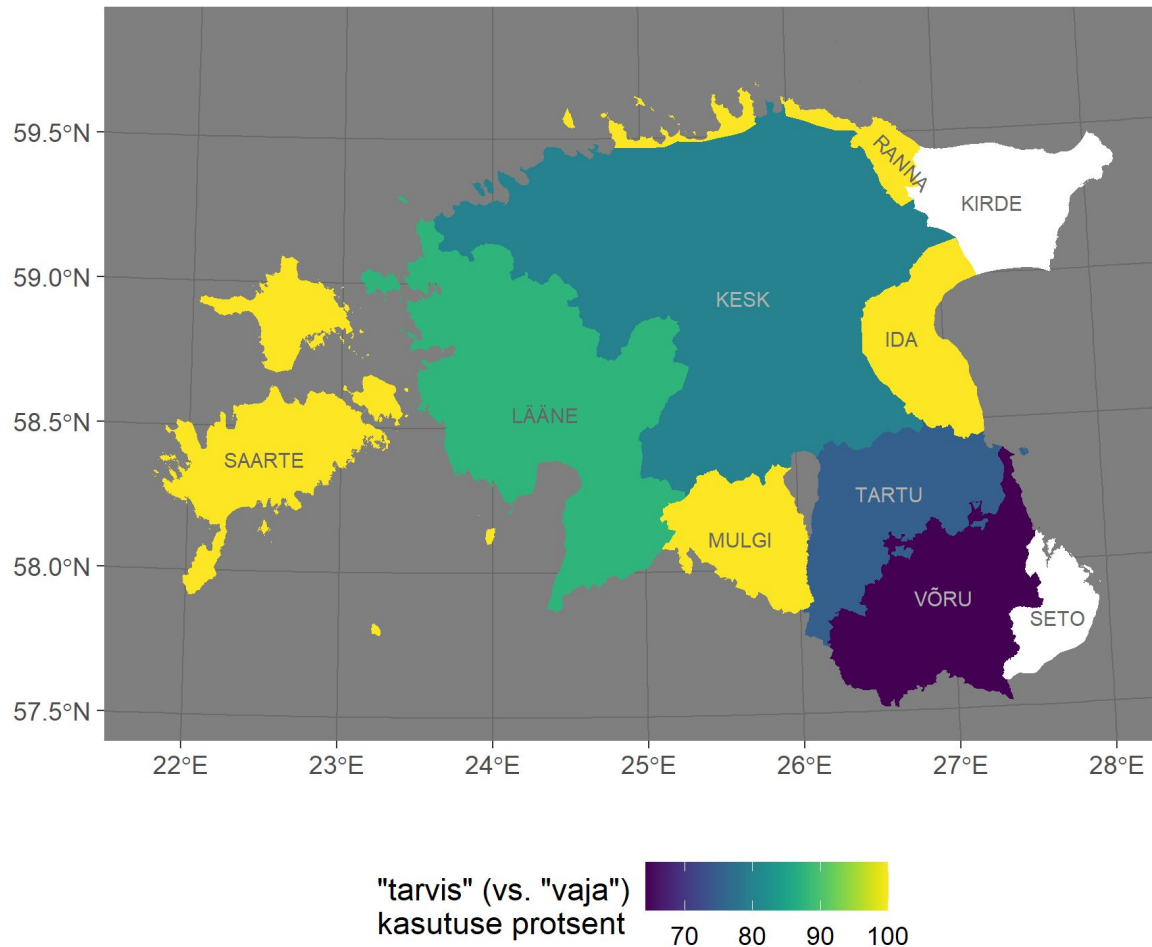
tarvis/vaja kasutus vallakohtuprotokollides

Automaatselt analüüsitud ja käsitsi
parandatud protokollide tekstid,
nn treeningandmed

Võimalik otsida sõna
standardiseeritud vormi või
sõnaraamatukuju (ja sõnaliigi)
põhjal

Näiteks

waea **vaja** **vaja** 0 D



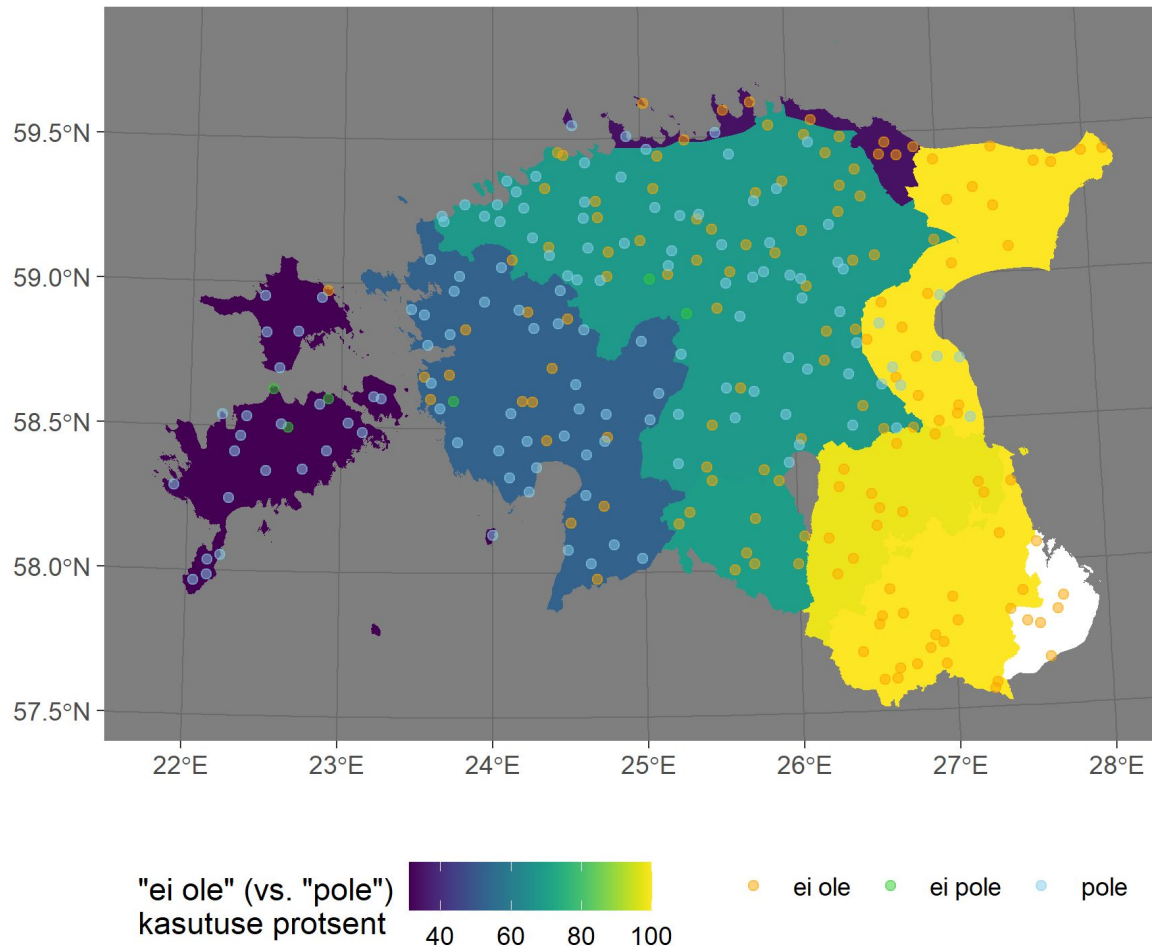
ei ole/pole kasutus vallakohtuprotokollides ja Andrus Saareste “Väikeses eesti murdeatlases” (1955)

Automaatselt analüüsitud ja käsitsi
parandatud protokollide tekstid,
nn treeningandmed

Võimalik otsida sõna
standardiseeritud vormi,
sõnaraamatukuju ja vormiinfo
põhjal

Näiteks

eij	ei	ei	0	V	neg
õle	ole	ole	0	V	o
põlle	pole	ole	0	V	neg o



Nimeüksuste märgendamisest

Nimeüksuste märgendamine

- Eesmärk: leida tekstist nimefraasid ja määrata nime liik
- Näide:

Ette tuli Rummost^{KOHT} Jaan Lustberri^{ISIK} ja kaebas seältsammast Pridik Rosenbergi^{ISIK} peäle, et se temma rahha, 9 rubla, Mihkli päwa aial olli ärrawarrastanud.

- Praegu toimub see ühisloomes käsitsi ja märgenduste kvaliteet on ebaühtlane

Nimeüksuste märgendus - milleks kasulik?

- Paindlikum otsing protokollides:
 - konkreetse koha või isikuga seotud protokollide leidmine
 - dokumentide sirvimine (~ kiire ülevaade igas dokumendis mainitud nimedest)
- Nimede võrgustikuanalüüs
 - Kes, kus ja millal koos kohtus viibisid?
 - nõuab paraku ka nimede samaviitelisuse lahendamist ...

Nimeüksuste automaatmärgendus - kuidas?

- Protsess:
 - Käsitsi nimede märgendamine
 - märgendusjuhiste väljatöötamine
 - kooskõla hindamine
 - Märgendatud nimede põhjal masinõppe algoritmide “välja treenimine”

Nimeüksuste märgendus - probleemid

- Mitmesõnalised nimeüksused (märgenduse ulatus?):
 - Liigsõnadega nimed:
 - ... tuleb [**Radama küllast**] ...
 - ... perremes [**Hans Jürgew**] ...
 - Põimunud nimed:
 - ... [**Jaan Kask**] [**Jüri**] poeg ...
 - ... [**Mihkel Tõnise poeg Reinberg**] ...

Nimeüksuste märgendus - probleemid

- Kas tüüpiline nimeüksuste liigitus (isikud, kohad, organisatsioonid) sobib?
- Koha- ja organisatsiooninimede eristamine:
 - “konkreetsed” kohad:
 - *metsawahi sead oliwad, **Soniste metsas** tustimas käinud*
 - kohad ja ühtlasi ka “organisatsioonid”:
 - *Astus ette **Otsa** perremes Tiets Allikow*

=> Otsa talu => koht
 - *... sealt on **Otsa** rahwas neid kinni wõtnud ...*

=> Otsa talu rahvas => kohaga seotud inimrühm

Nimeüksuste märgendus - probleemid

- Kas tüüpiline nimeüksuste liigitus (isikud, kohad, organisatsioonid) sobib?
- Muud nimeliigid?
 - “esemed”
 - laew “Eduard”, Jaht Klara, bairich õlut
 - sündmused:
 - Nuustaku laadal, Paide Mardilaada
 - seaduseraamatud:
 - Liiwimaa talurahwa seaduse raamatu, Tallorahwa seaduse ramato”, “T. S. R.

Nimeüksuste käsitsi märgendamine I

- Märgendatud korpus:
 - 1500 protokolli, 27538 nimeüksust

Isikunimed	84%
Kohad/organisatsioonid	9,9%
Kohanimed	3,7%
Organisatsiooninimed	1,5%
Muud nimed	0,9%

Nimeüksuste käsitsi märgendamine II

- Märgenduse kooskõla (kui hästi märgendusjuhised “toimisid”?)
 - Kontrollmärgendus 250-I protokollil:
 - kaks sõltumatut märgendajat, keskmine kooskõla: 93%
 - võrdluseks:
 - kooskõla meie märgendajate märgenduste ja ühisloome märgenduste vahel: ~68%

(arvestades ainult isiku- ja kohanimesid)

Nimeüksuste automaatmärgendaja loomine (käimasolev töö)

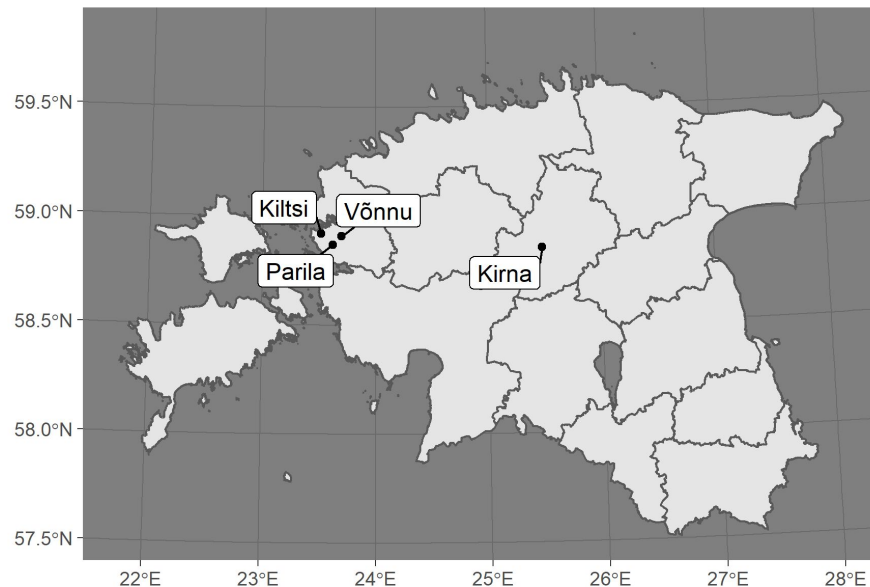
- EstNLTK nimeüksuste tuvastaja kohandamine
=> Kristjan Poska bakalaureusetöö
- Tehisnärvivõrkudel põhinevate Bert-i mudelite katsetamine

Sotsiaalsed võrgustikud

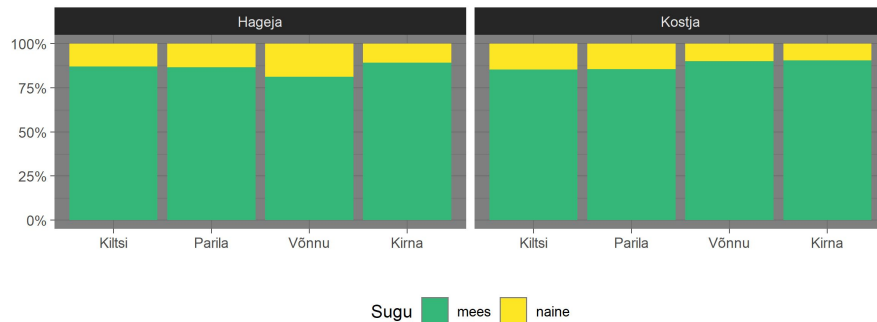
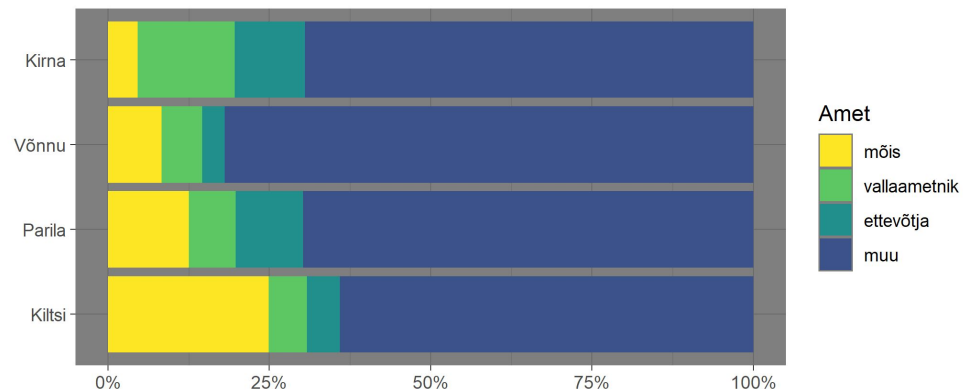
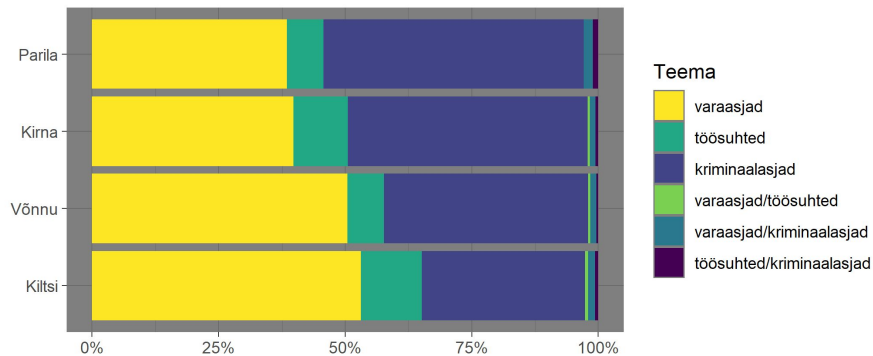
Kes kellega külas tülitises?

Kersti Lust

- 1575 vallakohtuprotokolli
- Kiltsi, Kirna, Parila ja Võnnu
- aastatest 1867-1891
- tuvastanud **hagejad** ja **kostjad**, nende soo, ameti, elukoha ning protokollu teema (vara- ja rendisuhted, töösuhted ning üleastumised ja kuriteod).



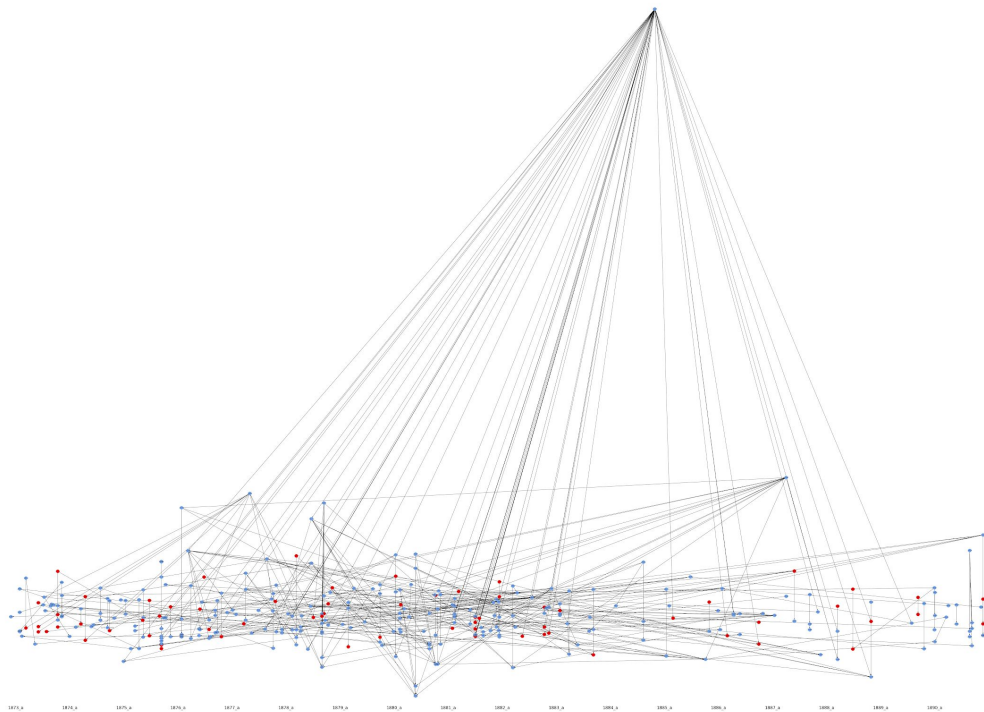
Teemad, hagejad, sood



Võrgustik

Kiltsi 237 protokolli

Mõis kõige aktiivsem
hageja



Aitäh!