doi:10.5128/ERYa15.08

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 15, 139–158

MÖISTUS SAI KUULOTEDU: 19. SAJANDI VALLAKOHTUPROTOKOLLIDE TEKSTIDEST DIGITAALSE RESSURSI LOOMINE

Maarja-Liisa Pilvik, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa, Tõnis Türna

Ülevaade. Artikkel käsitleb digitaalse ressursi loomist aastatest 1866–1890 pärinevatest vallakohtuprotokollidest. Vallakohtuprotokollide tekstiandmebaas sisaldab ligi 420 000 sõna XML-märgendusega failides. Tekstid on keeleliselt mitmekesised, keelise kuju põhilised mõjutajad on uue *vs*. vana kirjaviisi kasutamine, murdelisus ning valla- või kohtukirjutaja hariduslik ning keeleline taust. Samuti mängivad suurt rolli protokollide sisestamisel tehtud ortograafilised valikud. Tekstide keelelise analüüsi ning märksõnastamise jaoks katsetati automaatset morfoloogilist analüüsi ning nimeüksuste tuvastamist EstNLTK vastavate moodulite abil, hinnati väljundi kvaliteeti ning kaardistati analüüsi parandamise põhilised viisid. Vallakohtute protokollide kättesaadavaks tegemine ja otsitavuse parandamine tekstide keelelise ja temaatilise märgendamise abil aitab luua rikkalikku digitaalset ressurssi, mille kasutajaskonna moodustavad väga erineva tausta ja huvidega inimesed.*

Võtmesõnad: keeletöötlus, automaatne morfoloogia, digihumanitaaria, korpuslingvistika, andmebaasid, keeleajalugu, eesti keel

1. Sissejuhatus

Ajalooliste tekstide automaatne analüüs ning keelelise muutuse modelleerimine on aktuaalne suund nii keeletehnoloogias kui ka digihumanitaarias. Üha enam koostatakse ajalooliste tekstide digitaalseid kogusid, mille töötlemiseks aga ei piisa tänapäeva keele jaoks loodud vahenditest (lemmatiseerija, morfoloogiline analüsaator), ja seda mitmel põhjusel: ajalooline tekst on enamasti käsitsi kirjutatud ning vajab eeltööd, et see viia digitaliseeritud kujule; ajalooline tekst võib olla

* Kirjutise valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus), Eesti Haridus- ja Teadusministeerium uurimisprojektiga IUT20-56 ja projektiga EKKM17-458 "1860-80 vallakohtuprotokollid kultuurimälu kandjana (1.1.2017–31.12.2018)".

väga varieeruv, sest kirjalikus keelekasutuses ei ole normid välja kujunenud või on korraga kasutusel mitmeid norme; ortograafiline süsteem võib olla muutunud; keel võib sisaldada murdejooni; tekstide digitaliseerimisel on kasutatud vahendeid, mis teevad vigu (nt tärktuvastamine ehk OCR); sõnade tähendus võib olla muutunud; sõnade muutmine võib olla muutunud; kasutatakse tänapäeval tundmatuid sõnu jne (vt nt Piotrowski 2012: 11–23, Petterson 2016). Kõik need asjaolud on ühelt poolt huvitavad keele muutumise ja varieerumise aspektist, ent teiselt poolt raskendavad ajalooliste tekstide masintöötlemist ning tekstikaevet.

Käesolev artikkel annab ülevaate katsest töödelda ajaloolisi keeleliselt varieeruvaid tekste, täpsemalt 19. sajandi vallakohtute protokolle tänapäeva eesti kirjakeele analüüsiks mõeldud vahenditega (morfoloogiaanalüsaator ning nimetuvastus EstNLTK[1] (Orasmaa jt 2016) vahendite komplektist) ning välja selgitada, kui otstarbekas selline lähenemisviis on ning milliseid kohandusi on vaja teha, et eri valdade tekste edukalt analüüsida. Võrdleme valdade murdetaustast tingitud erisusi ning tekstide sisestamisel tehtud ortograafiliste kohanduste mõju morfoloogilise analüüsi tulemustele. Vallakohtuprotokollide tekstidest koostatakse multidistsiplinaarne digitaalne ressurss ning automaatne keeleline analüüs on vajalik eelkõige selle ressursi mugandamiseks tänapäeva kasutajale, aga ka edasiste analüüsivõimaluste arendamiseks.

1.1. Vallakohtuprotokollidest

Vallakohus oli 19. sajandi alguskümnenditel talurahvaseadustega ellu kutsutud vallakogukonna kui kujuneva talurahva omavalitsusüksuse tuumikuks. Sinna koondusid talupoegadesse puutuvad esmased õiguslikud, politseilised ja halduslikud funktsioonid. Esimestel kümnenditel ei omanud mõisaga tihedalt seotud vallakohus erilist mõjuvõimu, olles pigem mõisavalitsuse abistaja ning vahendaja rollis mitmesugustes talupoegadesse puutuvates küsimustes (Traat 1971: 39, 1980). 1866. aasta reformiga[2] vabanes asutus nii mõisa järelevalvest kui ka haldus- ja majandusasjadega tegelemisest, need läksid põhiliselt üle vallavalitsustele ja -volikogudele. Edaspidi tegeles vallakohus üleastumiste, nõuete, vaidluste ja perekonnasuhetega. Vallakohtud senisel kujul lõpetasid tegevuse 1918. aastal, 1919. aasta algul seisuslikud kohtud kaotati ja vallakohtud kujundati ümber eestkoste- ja hoolekandeasutusteks.

Protokolliraamatud seati sisse kooskõlas 1819. aasta Liivimaa talurahva vabastamisseadusega, kuid neid peeti algul ebajärjekindlalt, sõltuvalt kirjutaja oskustest või paremast äranägemisest (Traat 1980: 62–63). Tüliasju lahendati eelistatult n-ö jalapealt ja haldustoimingute ülestähendamine ei olnud kohustuslik (Hiio 1996). 1819. aasta seadus nägi kogukonnakohtu koosseisus ette kirjutaja, kuid teda lubati palgata juhul, kui ükski kohtunikest ei osanud kirjutada ega arvutada. (Traat 1971: 40) Vallakohtu menetlus oli suuline ja eestikeelne, mis aga ei taganud, et protokollid oleks kirja pandud keeles, milles toimus menetlus (Traat 1980: 62–63). Vallakohtute tegevuse esimestest kümnenditest Lõuna-Eestis on säilinud nii sõnavaeseid konarlikus eesti keeles kui ka ladusas saksa keeles protokolle. Viimaste kirjutajaks on olnud mõisavalitseja, -kirjutaja, -rentnik vms (vt ka Traat 1971: 40,

[1] Täpsemalt EstNLTK versioon 1.4.1, https://github.com/estnltk/ (4.1.2019).

[2] Kui Lõuna-Eestis toimisid vallakohtud 19. sajandi jooksul järjepidevalt (sh Saaremaal, kus vallakohtud asutati küll alles 1820. aastal), siis Põhja-Eestis 1816. aasta talurahvaseadus vallakohut enam ette ei näinud ja selle ülesanded anti kihelkonnakohtule. Vallakohus taastati seal alles 1866. aasta vallaseadusega. (Traat 1971: 39, 1980: 10) Käesolev artikkel käsitleb vaid Lõuna-Eesti alale jäänud Liivimaa kubermangu kuulunud vallakohtute materjale.

1980: 45). 1866. aastast tegeles vallakohtu istungite protokollimise ja muu dokumentatsiooniga vallavolikogu poolt palgatud vallakirjutaja. Eraldi kohtukirjutaja võisid kihelkonnakohtu loal palgata suurema töökoormusega vallad või mitut valda ühendavad vallakohtud.

Rahvusarhiivis on hoiul ligi 450 tsaariaegse vallakohtu materjalid, mis on aga säilinud ebaühtlaselt ning millest kõik ei sisalda protokolliraamatuid. Protokolliraamatute pidamine muutus järjepidevaks pärast 1866. aasta vallareformi. 1889. aastal toimunud kohtureformi järel hakati kohtuasjade kohta pidama eraldi toimikuid ning protokolliraamatud kadusid käibelt.

Vallareformi järgsest ajast säilinud protokollid on oma iseloomu, temaatika, süstemaatilisuse ning selgelt eristatavate isiku- ja kohanimede tõttu ajaloolastele ja etnoloogidele oluliseks allikmaterjaliks, mille põhjal kirjeldada talurahva elu-olu, varanduslikku seisu, erinevaid probleeme talude, maade, varaliste tehingutega, kõikvõimalikke üleastumisi ja elukorralduse häireid, perekonna- ja moraaliprobleeme, arusaamu õiguslikust korraldusest ning ka üldist mentaliteeti tollases külaühiskonnas (vt nt Linnus 1970, K. Must 1998a, Kaaristo 2004, 2006).

Vallakohtute protokollid on ka keeleajaloo uurimisobjektiks, võimaldades hinnata kirjakeele kujunemise protsessi. Kohtuprotokollides seguneb tolleaegne murdekeel ja alles kujunev kirjakeel, sõltudes suuresti kohtukirjutaja haridusest, päritolust ja murdetaustast. Protokollide keelekasutuse põhjal võib seega hinnata sotsiolingvistilist olukorda sel perioodil, aga ka leida murdejooni, mis hilisemates üleskirjutustes või salvestustes on haruldased. Katse hinnata tekstide murdepärasust ning kirjaviisi üleminekut vanalt uuele on näiteks Peetri vallakohtu protokollide põhjal teinud Kristiine Kurema (2013).

Kuna protokolliraamatutesse kanti kõik kohtuasjad kronoloogilises järjekorras ning registreid tehti haruharva, on suurimaks probleemiks vallakohtu protokollide kasutamisel olnud see, et vajalikud andmed on raskesti leitavad. Huvipakkuva info jaoks on tulnud kõik protokollid järjest läbi lugeda. Struktureeritud, märgendatud ja otsitava andmebaasi loomine muudab temaatiliselt hajali oleva teabe kiiresti leitavaks ning analüüsitavaks. Tekstide masinloetavale kujule viimine, temaatiline ja keeleline märgendamine ning ruumiinfoga sidumine avab mitmeid uusi uurimisperspektiive nii ajaloo, etnoloogia kui ka keeleteaduse vaatenurgast. Allika koha- ja isikuloolisus ning jutustav stiil võimaldab uurida kohaajalugu mikrotasandil, mis on huvitavaks ja vajalikuks uurimis- ja õpiobjektiks kohalikele kogukondadele, koolidele ja ühendustele.

1.2. Ajalooliste tekstide automaatsest morfoloogilisest märgendamisest

Ajalooliste tekstide otsitavaks ja analüüsitavaks tegemisel varustatakse tekste nii temaatilise kui ka keelelise infoga (eelkõige tekstisõna vormi kohta). Kui tänapäeva keelte jaoks on välja töötatud kõrge täpsusega automaatsed analüsaatorid, siis nende kasutamisel arhailise teksti analüüsiks mõjutavad teksti eripärad oluliselt tulemuse täpsust.

Vanemat keelekasutust sisaldavate korpuste automaatse morfoloogilise märgendamise katseid tänapäeva keele analüüsiks mõeldud vahendeid kasutades on

tehtud mitmete keelte põhjal ning selle tulemused võivad olla üsna erinevad. Näiteks katsetas Hrafn Loftsson (2013) mitut tänapäeva islandi keele materjalil treenitud masinõppepõhist morfoloogiaanalüsaatorit vanaislandi 13.–14. sajandi tekstide analüüsiks. Väljundteksti korrektsus (ingl *accuracy*; õige analüüsi saanud sõnede osa sõnade koguhulgast) varieerus 83% ja 87% vahel.

Silke Scheible ja kolleegide (Scheible jt 2011) katsed kasutada tänapäeva keele jaoks treenitud Treetaggerit (Schmid 1995) aastatest 1350–1650 pärinevate saksakeelsete tekstide morfoloogiliseks märgendamiseks andsid korrektsuseks 70% ning Marcel Bollmani (2013) katsed rakendada morfoloogilist märgendajat RFTagger (Schmid, Laws 2008) saksakeelsetele 15.–18. sajandi tekstidele korrektsuse 27% kuni 88%, olenevalt teksti kirjutamisajast ja murdelisusest.

Vanemate tekstide morfoloogilise märgendamise kvaliteedi parandamiseks on mitu võimalust. Üks võimalus on kasutada tänapäeva keele jaoks arendatud töövahendit ja väljundit käsitsi parandada. See tee on tööjõumahukas, kuid tekstide käsitsi üle kontrollimine tagab tulemuse hea kvaliteedi. Praktikas sageli kasutatav lahendus on nn normaliseerimine, mille eesmärk on teisendada analüüsimist vajavad ajaloolised tekstid tänapäeva keelekasutusele lähedasemaks. Normaliseerimistehnikaid on põhjalikult analüüsinud nt Eva Petterson (2016: 45–81) ja Michael Piotrowski (2012: 69–83).

On ka võimalik kohandada morfoloogilist analüsaatorit ja ühestajat analüüsitava keelevariandiga. Analüsaatori kohandamiseks piisab tänapäeva keelekasutuses puuduvate sõnade lisamisest morfoloogilise analüüsi leksikoni, kuid ühestaja kohandamiseks on vaja märgendada käsitsi vastava keelevariandi treeningkorpus ning treenida sellel morfoloogilise märgendaja spetsiaalne mudel; lähemalt vt nt (Sanchez-Marco jt 2011, Petterson 2016, Piotrowski 2012).

Eestikeelsete ajalooliste tekstide automaatse morfoloogilise analüüsiga ei ole varem kuigivõrd tegeldud. Suur hulk ajaloolisi tekste on kogutud eesti vana kirjakeele korpusesse (VAKK), ent nende tekstide morfoloogiliseks märgendamiseks on rakendatud abiprogrammi, mis aitab sagedaste sõnade märgendamisel. Põhitöö toimub siiski käsitsi ja on aeganõuev. (Vt nt Prillop 2004).

2. Digitaalse ressursi loomine

Vallakohtute materjale on teadustöödes kasutatud alates 1930. aastatest (Anepaio 2007). 1866. aasta reformi eelsete kohtute ajaloo ja säilinud protokollid töötas põhjalikult läbi August Traat (1980). 1866. aasta reformi järgsetes protokolliraamatutes oleva teabe süstematiseerimine ja kättesaadavaks tegemine sai alguse alles 1990. aastatel, kui Aadu Musta eestvedamisel hakati protokolliraamatute põhjal koostama esmalt temaatilisi sedeleid ning kümnendi keskel tehti algust ka protokollide elektroonilise sisestamisega. Selle tulemusel valmis 1997. aastal juba täistekste koondav internetipublikatsioon, nn Eesti õigusajaloo krestomaatia (A. Must 1997). Pärnumaa tüüpilisemate ja värvikamate protokollide täistekste on avaldanud internetis ka Kadri Must (1998b) ja Reeli Ziius (Türna 2004: 13).

Artikli alusmaterjal on pärit Tõnis Türna Tartu Ülikoolis kaitstud bakalaureusetööst, milles valiti Lõuna-Eestist välja ligi 2000 vallakohtuprotokolli aastatest 1866–1890 ja koostati nende põhjal täistekst-andmebaas. Töö sisaldas tekste

Lõuna-Eesti 11 valla arhiivist. Türna valis Lõuna-Eesti, kus vallakohtute traditsioon oli pikem ja kust pärineb 60% kõigist säilinud protokolliraamatutest. Ajavahemiku valikul lähtuti kahest olulisest ümberkorraldusest vallakohtute tegevuses: 1866. aasta valla- ning 1889. aasta kohtureformist. (Türna 2004: 8)

Andmebaas avaldati veebilehel http://www.history.ee/vallakohus ning sinna lisandusid järgneva paari aasta jooksul veel Tartumaa 11 valla tekstid, ent paraku hävis andmebaas serveri history.ee rikke tõttu. Hinnaline andmestik, mille suuruseks on üle 400 000 tekstisõna, taastati 2016. aastal veebiarhiivide kaudu.

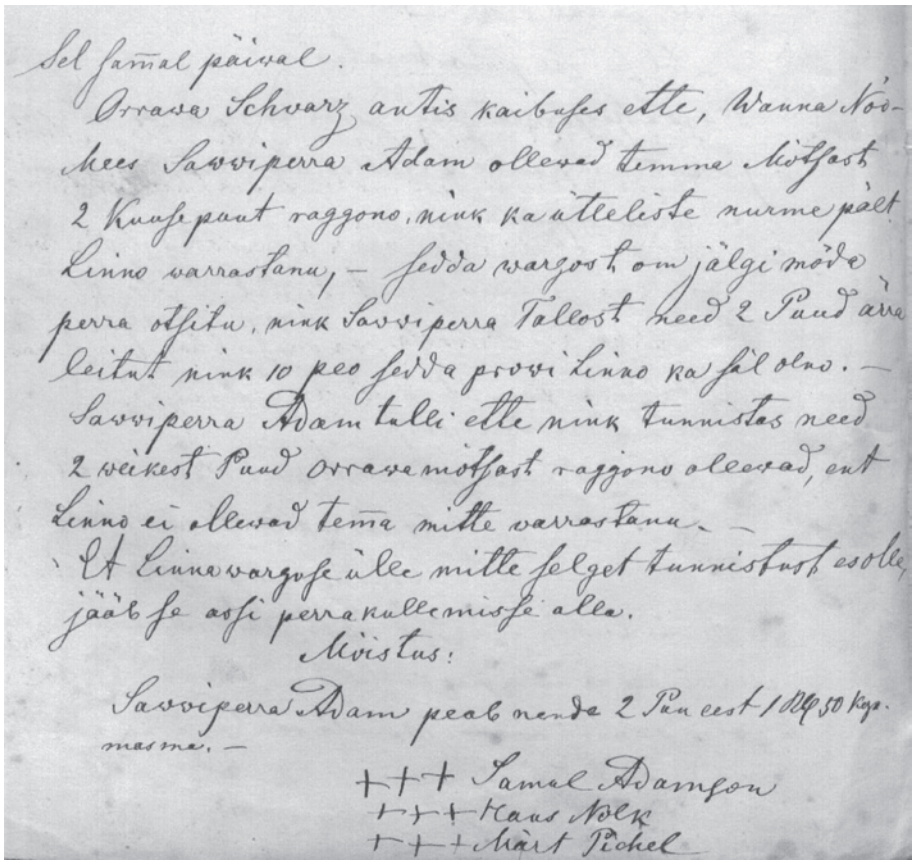
Tabelis 1 on esitatud valdade ja maakondade kaupa taastatud protokollide arv, sõnade arv protokollides (ilma kirjavahemärkideta) ning periood, millest protokollid pärinevad. Esimese 11 valla tekstid sisestas Tõnis Türna bakalaureusetöö käigus ning viimase 11 omad hiljem.

Tabel 1. Protokollide ja nende sõnade arv valdade ja maakondade kaupa

Vald	Maakond	Protokollide arv	Sõnade arv
Alatskivi	Tartumaa	199	24 344
Joosu	Võrumaa	144	18 110
Kahkva	Võrumaa	231	26 878
Kiuma	Võrumaa	166	21 492
Laiuse	Tartumaa	51	12 522
Maasi	Saaremaa	75	14 000
Mihkli	Pärnumaa	76	14 598
Navesti	Viljandimaa	119	18 482
Tarvastu	Viljandimaa	157	14 193
Uue-Suislepa	Viljandimaa	395	54 672
Vastse-Nõo	Tartumaa	194	32 062
Aru	Tartumaa	118	17 345
Haaslava	Tartumaa	160	27 415
Kokora	Tartumaa	59	7725
Kärevere	Tartumaa	46	5979
Laeva	Tartumaa	154	22 188
Luke	Tartumaa	54	5444
Mäksa	Tartumaa	122	19 182
Pangodi	Tartumaa	116	15 412
Suure-Konguta	Tartumaa	53	7714
Valguta	Tartumaa	106	21 368
Väike-Rõngu	Tartumaa	72	17 383
Kokku		2867	418 508

Serveri rikke tõttu kaotatud ning veebiarhiivide kaudu taastatud tekstid olid HTML-kujul. Need protokollifailid viidi üle XML-formaati, võttes lähtefailist üle sisuosa, daatumi, protokolli numbri (kui see oli märgitud), kohtumeeste nimed, sisestaja poolt märgitud pealkirja ning valla nime (mis omakorda seoti geograafiliste koordinaatidega). XML on arendatud andmete kirjeldamiseks, HTML andmete

esitamiseks, seetõttu sobib esimene protokollide sisu struktureerimiseks ja hilisemaks töötlemiseks paremini, võimaldades muuhulgas luua andmetest lähtuvaid märgendeid (nt kohtumeeste märkimiseks) ning tagades andmebaasi konverteeritavuse ja ühildatavuse. Igasse XML-faili on lisatud ka eraldi teemaindeksite failist ekstraheeritud teemavaldkonnad. Kõnealuse perioodi vallakohtute töökorralduse aluseks oli suuresti 1819. aasta seadusele tuginev Liivimaa 1860. aasta talurahvaseadus, mis sätestas ka kohtuprotsessi korralduse ning seetõttu on ka protokolliraamatute struktuur suhteliselt universaalne ja automaattöötluseks hästi sobiv (Türna 2004: 19). Eestikeelsesse protokolliraamatusse tuli iga kohtuasja kohta märkida kohalviibinud kohtumeeste, hageja ja kostja nimed, kohtuasja kokkuvõte, esitatud tunnistused ja ütlused ning kohtuotsus. Mõned aastad hiljem täiendati seda nõudega panna kirja apelleerimist puudutavad seigad, fikseerida otsuse ja selle täitmisega seotud asjaolud ja kuupäevad ning varaloetelude puhul märkida asjade hind, kogus, mõõt või kaal. (Türna 2004: 17–18) Politseilistes asjades tuli lühidalt kirja panna kohtuasja põhisisu. Joonis 1 on näide Luke valla protokolliraamatust. Suur osa vallakohtute materjalidest on Rahvusarhiivis skaneeritud ning veebi kaudu kättesaadavad (Saaga).



Sel samal päival.
Orrasa Schwarz antis kaibuses ette, Wanna Nõokees Sawiperra Adam olleva temma wastast 2 Kuusepuud raggono, mink ka ätteleiste nurme pealt Linno warrastanu, — sedda wargost om jälgi mööda perra otsitu, mink Sawiperra Tallost leed 2 Puud äraleitut mink 10 peo sedda proovi Linno ka fal olno. — Sawiperra Adam talli ette mink tunnistas need 2 wiekest Puud Orrase mõistast raggono olleva, ent Linno ei olleva tema mitte warrastanu. —
Et Linno wargose üllemitte selget tunnistust ei olle, jääb se asi perra kullemisse alla.
Mõistus:
Sawiperra Adam peab nende 2 Puu eest 1 Rbl 50 Kop. maksma. —
+++ Samal Adamson
+++ Kaus Nõlk
+++ Mart Pichel

Joonis 1. Luke valla protokolliraamatu näide Rahvusarhiivist

Algsetes HTML-failides oli järgitud protokollide originaalstruktuuri, mistõttu võis üks lause jätkuda üle mitme rea. XML-failide sisuosa lausestati ehk ühele reale hakkas vastama üks lause. See hõlbustab tekstide automaatset analüüsi (nt nimeüksuste ja loendite tuvastamist). Joonisel 1 kujutatud protokollifaili struktuuri XML-formaadis illustreerib joonis 2.

```
<protokoll>
<head>
<pealkiri>Protokoll härra Schwarzi nõudest Saviperra Adami vastu metsavarguse ja linade varastamise asjas</pealkiri>
<teema>Üleastumised.</teema>
<vald>Luke</vald>
<aeg>10.11.1867</aeg>
<number></number>
<kohtumehed>
<nimi>Samul Adamson</nimi>
<nimi>Hans Nolk</nimi>
<nimi>Märt Pichel</nimi>
</kohtumehed>
</head>
<sisu>
Orrava Schwarz antis kaibuses ette, Vanna Nõo Mees Savviperra Adam ollevad temma Mõtsast 2 Kuusepuut raggono, nink ka üteeliste nurme pält Linno varrastanu, sedda vargost om jälgi möda perra otsitu, nink Savviperra Tallost need 2 Puud ärra leitut nink 10 peo sedda provi Linno ka säl olno.
Savviperra Adam tulli ette nink tunnistas need 2 veikest Puud Orrava mõtsast raggono ollevad, ent Linno ei ollevad temma mitte varrastanu.
Et Linnavargose ülle mitte selget tunnistust es olle jääb se assi perra kullemisse alla.
Mõistus: Savviperra Adam peab nende 2 Puu eest 1 rbl 50 kop masma.
</sisu>
</protokoll>
```

Joonis 2. Luke valla protokollifaili struktuur XML-formaadis

3. Tekstide automaatne analüüs

3.1. 19. sajandi tekstidega seotud problemaatika

Eesti ala on küll territoriaalselt väike, ent siin paiknevad ajaloolised keelevariandid jagunevad kahe hõimukeele – lõunaeesti ja põhjaeesti – murreteks. Seejuures on lõunaeesti keelt peetud keeleajalooliselt vanemaks selles mõttes, et see on eraldunud muudest läänemeresoome keeltest varem. (Viitso 1985, Kallio 2012) Alates 16. sajandist, mil eesti keelt on hakatud süstemaatilisemalt kirja panema, on vastavalt olnud kasutusel kaks kirjakeelt: lõunaeesti ehk tartu kirjakeel ja põhjaeesti ehk tallinna kirjakeel (vt nt Raag 2008: 28). Kaks kirjakeelt elasid paralleelselt Eesti- ja Liivimaal kõrvu veel 19. sajandilgi. Lõunaeesti keele ala oli siiski kõnelejate arvu poolest vähemuses. Aastal 1875 on F. J. Wiedemann sedastanud, et kooli- ja kirikukeelena oli lõunaeesti keel sel ajal kasutusel veel ainult Võru praostkonnas, mujal oli ka koolis ja kirikus kasutusel pigem põhjaeesti kirjakeel. (Wiedemann 2011 [1875]: 65, Raag 2008: 41). Tõenäoliselt oli lõunaeesti keel sel ajal siiski kasutusel ka Lõuna-Tartumaal. Sellele osutab Võru- ja Tartumaa pastorite soovimatus üle võtta uut kirjaviisi, mis seostus eelkõige põhjaeesti keelega (Kask 1970: 186–191).

Lõuna-Tartumaa pastorite vastumeelsusele üle minna põhjaeesti keelele ja uuele kirjaviisile on viidanud ka Mati Hint (2008).

Lisaks kahe keelevariandi paralleelsele kasutusele oli 19. sajandil, eriti selle III veerandil päevakorras kirjaviisi küsimus. Nii tallinna kui ka tartu keelt kirjutati üles vanas saksapärases kirjaviisis, mis aga eesti keele hääldust kuigi hästi edasi ei andnud. Vana kirjaviisi peamised reeglid olid järgmised (Raag 2008):

1) lahtise silbi pikka vokaali (nagu sõnades *maa*, *raamat*, *looja*) märgiti ühekordse vokaalimärgiga ning järgneva ühekordse konsonandiga: *ma*, *ramat*, *loja*;
2) kinnise silbi pikka vokaali märgiti kahekordse vokaalimärgiga: *maalt*, *siis*, *juuste*;
3) lahtise rõhulise silbi lühikest vokaali märgit järgneva kaashääliku topeldamisega: *tühhi* (= *tühi*), *wahhele* (= *vahele*), *innimenne* (= *inimene*).

Alates Eduard Ahrensi 1843. aasta grammatikast võeti järk-järgult kasutusse tänapäevane uus kirjaviis, seda küll pikkade diskussioonidega. Uus kirjaviis pääses trükisõnas võidule 1870.–1880. aastatel. (Kask 1958: 192) Võib arvata, et rahva seas, sh vallakirjutajate hulgas, oli pilt veelgi ebaühtlasem ning siin kestsid edasi nii tartu ja tallinna kirjakeele erisused kui ka vana kirjaviis. Näiteks Fred Puss on analüüsinud 110 koguduse meetrikaraamatut ning leidnud, et aastaks 1891 on neist uuele kirjaviisile üle läinud 69%; seejuures kõige konservatiivsemad olid pastorid Järva-, Tartu- ja Võrumaal (Puss 2018: 180). Tõenäoliselt on suures plaanis samalaadne pilt ka vallakohtuprotokollides. Lisaks säilis tekstides ka peale üleminekut uuele kirjaviisile ortograafiline ebajärjekindlus: näiteks Peetri valla protokollides on küll 1884. aastaks suuresti uuele kirjaviisile üle mindud, ent tekstides leidub sellegipoolest üksjagu varieerumist. (Kurema 2013)

Vallakohtu protokollides on ka tugevaid murdemõjusid, mis on osalt seotud tartu või tallinna keele valikuga, osalt aga lisanduvad neile. Näiteks võib leida palju lõunaeestilisele vormimoodustusele iseloomulikku: *s*-lõpuline konditsionaal *üteldas* ‘öeldakse’ (Joosu), mineviku eituspartikkel *es lubba* ‘ei lubanud’ (Suure-Konguta); partitsiibid *-nud* ja *-tud* esinevad süstemaatiliselt kujul *-nu* ja *-tu-* (nt *ernit olnu maha külvetus* ‘herneid oli maha külvatud’ (Haaslava)) või harvemini veelgi lühemalt (*-n*, *-t*). Ka muidu tekstides üsna harva esinevat kaudse kõneviisi erinevaid moodustusviise võib vallakohtuprotokollidest leida väga erinevaid: järgnevas Aru valla protokolli katkes (näide 1) esinevad nii põhjaeestiline *da*-infinitiivi kujuline kaudne kõneviis (*olla kõnelnud*) kui ka *na*-tunnuseline tartumurdeline kaudne kõneviis (*võina*).

(1) Tulli ette peremees Mihkel Kukkemelk ning andis üles, et tema naabri talu peremees Hans Tillisson, kellega kaebajal piirid koos on rohkem aasta eest tema Mihkel Kukkemelgi lina leo vee lüisi maha lasknu, lina pool ligunemata on kuivas jäänud, ärä tullitanud ja ärä rikkenu ning häbemata teo läbi 340 rubla kahju saanud. Hans Tillissoni naene Liini, **olla** seda tego tema Mihkel Kukkemelgil **kõnelnud**. Seda **võina** kah sulane Mihkel Krootmann tunnistada.

Lisaks murdemõjudele sõltus protokollide keelekasutus väga palju ka vallakirjutaja taustast (haridus, päritolu, kirjutamiskogemused jne) ja individuaalsest eripärast (Linnus 1970: 233, Türna 2004: 30).

Tõnis Türna on bakalaureusetöö materjalide sisestamisel järginud printsiipi, mille eesmärgiks on protokollide puhastamine vana kirjaviisi mõjudest, viies tekstide keelekasutust omaaegsele häälduspärasemale (ja tänapäeva õigekirjale lähemal olevale) kujule. Sealjuures on ta püüdnud säilitada grammatika ning murdekeele eripärasid. (Türna 2004: 29) Normaliseeritud on topeltkonsonandid (*teggi* → *tegi*), lühikesed vokaalid (*kravi* → *kraavi*), kohanimed, kuude nimetused (*Dezembril* → *detsembril*), enamtuntud lühendid (*nr*, *s.a*), *w* on läbivalt asendatud *v*-ga, pikendatud on konsonante (*al* → *all*), lahti on kirjutatud mitmed protokollides esinevad kaalud ja mõõdud (nt *nael* ja *leisikas*), lisatud on kirjavahemärke ja korrigeeritud suure algustähe kasutamist (*Kõrtsmik* → *kõrtsmik*). Samas on püütud hoiduda murdekeele normaliseerimisest ning säilitatud (vähemalt teatud valdade protokollides) mõned hääldusplilti aimavad vormid, nt *olli*, *sis*, *omma* jt. (Türna 2004: 29–30) Hiljem sisestatud tekstides aga ei ole teisendusi enam kas üldse või niisama järjepidevalt tehtud. Samuti ei ole Türna bakalaureusetöös normaliseeritud Võrumaa Joosu ja Kahkva vallakohtu protokollide tekste, milles ta leidis murdekeelsuse olevat märgatavalt suurema (Türna 2004: 30). Seega on automaatsel analüüsil sisendiks kahesugused andmed: kindlate põhimõtete järgi ühtlustatud ja normaliseeritud tekstid ning originaalilähedaselt sisestatud, minimaalselt või üldse mitte normaliseeritud tekstid.

Protokollide tekstide keel ja selle varieerumine sõltub seega viiest erinevast asjaolust: 1) vana ja uue kirjaviisi põhimõtete paralleelsest kasutuselolekust; 2) põhjaeesti ja lõunaeesti kirjakeele paralleelsest kasutusest; 3) kohamurrete mõjust vallakirjutaja keelekasutusele; 4) vallakirjutaja taustast ja oskustest; 5) tekstide sisestamisel tehtud valikutest.

Joonis 3. Vallakohtuprotokollide jagunemine lõuna- ja põhjaeesti murrete, normaliseerimise ning protokollide arvu järgi

Vallakohtuprotokollide jagunemist vallati illustreerib joonis 3. Sümboli toon näitab seda, kas vallas on kõneldud põhja- või lõunaeesti keelt, sümboli kuju viitab sellele, kas tekstid on normaliseeritud või mitte, ning sümboli suurus näitab kõnealusest vallast pärit protokollifailide arvu.

3.2. Tekstide morfoloogiline märgendamine

Automaatne morfoloogiline märgendamine on protsess, mille käigus igale tekstisõnale lisatakse info tema algvormi e lemma, sõnaliigilise kuuluvuse ja sõnavormis sisalduvate grammatiliste kategooriate kohta.

Vallakohtuprotokollide morfoloogiline märgendamine teenib mitut eesmärki. Kui murdelisele ja/või vanapärases kirjaviisis sõnavormile on märgitud lemmaks tänapäeva kirjakeele sõna, saab seda kasutada otsingus, leidmaks kõiki protokolle, kus selle sõnaga tähistatavast nähtusest on juttu. Samuti on morfoloogiline märgendamine vältimatuks eelduseks järgnevatele automaatanalüüsi etappidele. Eesti keele morfoloogiline märgendamine koosneb tavaliselt kahest etapist: morfoloogilisest analüüsist ja ühestamisest. Analüüs on leksikonipõhine, st sõnavara on ette antud sõnastikuna. Lisandub veel produktiivsete tuletiste ja liitsõnade analüüsi moodul ning oletamismoodul (nn oletaja), mis annab tõlgenduse(d) leksikonist puuduvatele sõnadele, mida pole õnnestunud analüüsida tuletiste või liitsõnadena. Oletajat saab sisse ja välja lülitada: kui oletaja on analüüsiprotsessi kaasatud, saab iga sisendteksti sõna mingi analüüsi, tundmatuid sõnu väljundis ei ole; kui oletaja on välja lülitatud, saavad sõnavormid, mida ei ole õnnestunud analüüsida kui leksikonis olevate lemmade vorme, regulaarseid tuletisi või liitsõnu, tundmatu sõna analüüsi. Tundmatu sõna tõlgendusi sisaldav tekst ei sobi järgmise etapi, morfoloogilise ühestamise sisendiks, kuid on väga hea analüüsi vaheetapp, mille põhjal saab teha järeldusi selle kohta, kui suurt osa mingi teksti sõnadest ei ole võimalik analüüsida leksikonipõhiselt. Tundmatu sõna tõlgenduse saanud sõnavormide sagedusloend on lihtsaim viis uurida mingi teksti eripärast (tänapäeva eesti kirjakeeles mitteesinevat) sõnavara või vormistikku ja saada umbkaudset eelinfot selle kohta, milline võiks olla morfoloogilise märgendamise kvaliteet.

Vallakohtu protokollide esmaseks morfoloogiliseks analüüsiks kasutati EstNLTK (Orasmaa jt 2016) koosseisu kuuluvat morfoloogilist analüsaatorit. Selleks, et saada teadmine selle kohta, kui suure osa tekstisõnadest tänapäeva kirjakeele jaoks arendatud morfoloogiaanalüsaator suudab leksikoni põhjal tuvastada, viidi morfoloogiline analüüs läbi ilma oletamiseta ning leiti tundmatu sõna tõlgenduse saanud sõnavormide osakaal valla kaupa.

Tulemused on tabelis 2. Tundmatute sõnade osakaal varieerub ulatuslikult, olles väikseim (u 7%) Laiuse tekstides ja suurim (u 52%) Väike-Rõngu tekstides. Selgelt tuleb välja normaliseerimise mõju: tundmatuid sõnu on enam valdades, mille tekstide sisestamisel tekstikuju ei normaliseeritud. Põhja- ja lõunaeesti keele erinevus nii selgelt esile ei tule, ehkki nii normaliseeritud kui ka normaliseerimata tekstide lõikes näib analüsaatoril olevat rohkem raskusi lõunaeesti tekstidega. Oluline on ka suure algustähega sõnade osakaal tundmatute sõnade hulgas. Need on suure tõenäosusega pärisnimed, millele oletamisrežiimis töötav morfoloogiaanalüsaator suudaks anda pärisnime tõlgenduse (ent mitte tingimata õige lemma). Võrdluseks:

tänapäeva kirjakeele normile vastavates tekstides on tundmatuid sõnu 2,58%, neist suure algustähega 86,97%, st valdav enamus tundmatutest sõnadest on pärisnimed.

Tabel 2. Morfoloogiaanalüsaatori jaoks tundmatute sõnavormide osakaal

Vald	Normaliseeritud	Lõuna- või põhjaeesti murre	Morfoloogia-analüsaatori jaoks tundmatuid sõnu	Neist suure algustähega
Laiuse	Jah	P	6,96%	74,17%
Mihkli	Jah	P	10,26%	55,52%
Navesti	Jah	P	11,19%	38,61%
Uue-Suislepa	Jah	L	13,61%	38,43%
Alatskivi	Jah	P	16,32%	46,75%
Kiuma	Jah	L	16,54%	35,47%
Maasi	Jah	P	16,88%	33,98%
Vastse-Nõo	Jah	L	19,24%	26,86%
Laeva	Ei	P	19,36%	26,86%
Aru	Ei	L	21,14%	34,40%
Tarvastu	Jah	L	25,14%	22,80%
Pangodi	Ei	L	31,08%	26,78%
Kahkva	Ei	L	32,58%	25,71%
Kärevere	Ei	P	38,94%	28,69%
Mäksa	Ei	L	38,97%	20,75%
Joosu	Ei	L	38,98%	24,35%
Haaslava	Ei	L	39,52%	24,28%
Kokora	Ei	P	39,58%	29,15%
Valguta	Ei	L	41,46%	15,87%
Luke	Ei	L	47,30%	27,11%
Suure-Konguta	Ei	L	51,44%	27,09%
Väike-Rõngu	Ei	L	51,80%	18,35%

Lisaks sõna mittetundmisele (st sõnavorm saab tundmatu sõna tõlgenduse) võib morfoloogiline analüsaator tunda tekstisõna küll ära, ent anda sellele vale tõlgenduse. Kuna vale tõlgenduse saanud sõnavorme automaatselt tuvastada ei saa, otsustati väike hulk morfoloogiliste analüüsidega vallakohtu protokolle üle kontrollida käsitsi. Kontrollimiseks valiti igast vallast kuni kolm kõige suuremat faili. Tulemused on esitatud joonisel 4, kus heledama halliga on esitatud morfoloogiaanalüsaatorilt õige analüüsi saanud tekstiüksuste ja sõnade osakaal kõigi tekstiüksuste hulgast, tumedamalt analüsaatorile tundmatuks jäänud tekstiüksuste ja sõnade osakaal. Õige analüüsi all mõeldakse nii seda, kui tekstiüksus või sõna on saanud ühe õige analüüsi (nt *kirjutust*: substantiivi *kirjutus* ainsuse partitiivi vorm), kui ka seda, kui väljastatud on mitu analüüsi, millest üks on õige (nt *protokollis*: substantiivi *protokoll* ainsuse inessiivi vorm ja verbi *protokollima* indikatiivi imperfekti ainsuse 3. isiku vorm). Tekstiüksuste all mõeldakse kõiki tekstis olevaid märgijadasid: sõnu, lühendeid, numbritega kirjutatud arve ja kirjavahemärke. Kahe viimase kategooria

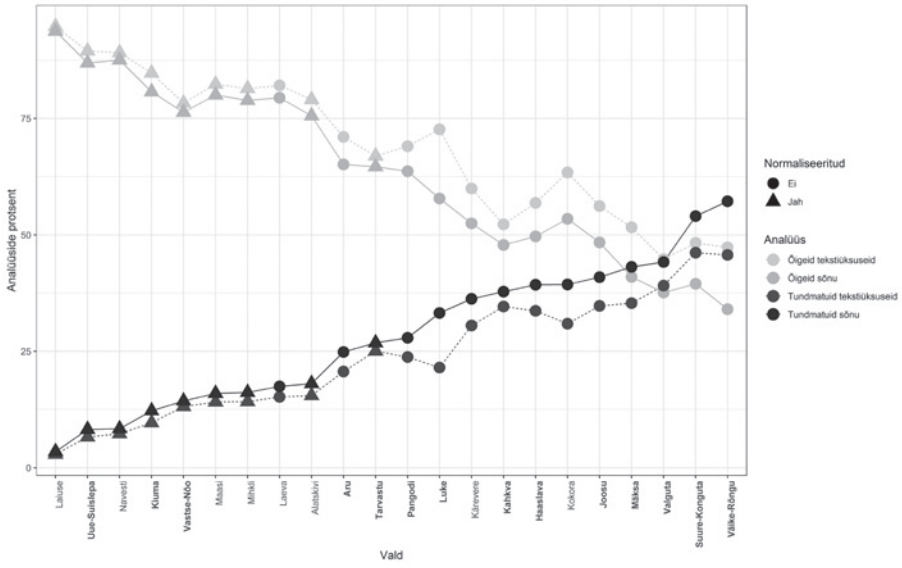
morfoloogiline analüüs on triviaalne ülesanne ja seda ei mõjuta teksti murdelisus ega vanapärasus. Teksti automaatanalüüsi kvaliteeti hinnates on tavaks need siiski tekstisõnade (*token*) hulka arvata. Seetõttu võib vallakohtu tekstide morfoloogilise analüüsi väljundi kvaliteet olla petlikult kõrge, kui käsitsi kontrollitud failide hulka on sattunud palju arve sisaldavaid tekste, nt varaloendeid, vt näide 2 Luke valla kohtust, kus on esitatud oksjonivara.

(2)

Asjade nimmed	Pakkote hind	Osja nimmed
2 poliiritut Sängid	9 rbl	Schulz
1 mõsso Laud	2 rbl 50 kop	Schulz
1 Sohva	13 rbl 90 kop	Schulz

Joonisel 4 ühendavad tekstiüksuste osakaalusid katkendlikud jooned, sõnavormide (sõnade ja lühendite, kuid mitte numbrite ja kirjavahemärkide) osakaalusid pidevad jooned. Ringid märgivad valdasid, mille tekstid ei ole normaliseeritud, kolmnurgad valdasid, mille tekstidest on vana kirjakeele mõjusid püütud süsteemselt kõrvaldada. Tavalises kirjas vallanimed viitavad valdadele, kus kõneldakse põhjaeesti murdeid ja paksus kirjas nimed valdadele, kus kõneldakse lõunaeesti murdeid.

Erinevalt tabelist 2, kus oli leitud kõigi valla tekstide tundmatute sõnade osakaal, on joonisel 4 olevad protsendid leitud käsitsi kontrollitud tekstide põhjal. Jooniselt on jäetud välja juhtumid, kus ei analüsaator ega käsitsi märgendaja ole osanud sõnavormile analüüsi pakkuda.



Joonis 4. Morfoloogiaanalüsaatori õigete analüüside ja tundmatute vormide protsendid käsitsi kontrollitud tekstides

Jooniselt 4 on näha, et kõige paremini sai analüsaator hakkama normaliseeritud tekstidega, kusjuures lõuna- ja põhjaeesti murrete vahel ei tundu väga selgeid erinevusi olevat. Siiski on normaliseeritud tekstidest kõige rohkem tundmatuid vorme ja kõige vähem õigeid analüüse Tarvastu valla materjalides. Normaliseerimata tekstidega on analüsaatoril suuremaid raskusi ning nendes on ka sõnade

ja tekstiüksuste osakaalude erinevus suurem (eriti Luke, Kokora ja Väike-Rõngu vallas). Rohkem tundmatuid sõnu kui õigeid analüüse oli Mäksa, Valguta, Suure-Konguta ja Väike-Rõngu tekstides, mis viitab nende valdade protokollides kasutatud keele kõige suuremale hälbimisele tänapäeva kirjakeelest.

Tabel 3 esitab kõige sagedasemad vale analüüsi saanud või tundmatuks jäänud tekstisõnad normaliseeritud ja normaliseerimata põhja- ja lõunamurretes.

Tabel 3. Kümme kõige sagedamini tundmatuks jäänud või vale analüüsi saanud tekstisõna normaliseerimise ja murderühma kaupa

Normaliseerimata põhjamurded	Normaliseeritud põhjamurded	Normaliseerimata lõunamurded	Normaliseeritud lõunamurded
temma	*luteruse*	*om*	*Ad*
Äint	*Kuresele*	*temma*	*Murro*
sedda	*Vastopä*	*nink*	*pääle*
Tulli	*öige*	*ollev*	*kohto*
vasto	*nimetud*	*se*	*ööse*
se	*naene*	*olle*	*juure*
peäle	*opetaja*	*Rbl*	*Jegorov*
Stamm	*karjasmaal*	*ärra*	*Belov*
Rubl	*Mikku*	*kül*	*keik*
rahha	*Rass*	*päle*	*om*

Tundmatute ja valesti analüüsitud sõnavormide hulgas on sagedased pärisnimed, normaliseerimata tekstides on probleemiks vana kirjaviis (nt *sedda*, *rahha*, *kül*) ja nendes tekstides sagedasti esinevad lühendid. Samuti valmistavad analüsaatorile raskusi murdelised sõnavormid, eriti lõunaeesti keeles (nt *om*, *ollev*, *olle*).

3.3. Nimeüksuste tuvastamine

Nimeüksuste tuvastamine on laiem ülesanne kui morfoloogilise analüüsi käigus tekstisõnale pärisnime tõlgenduse andmine. Nimeüksus võib olla mitmesõnaline, koosnedes kas mitmest pärisnime analüüsiga sõnast (*Peeter Tõruvere*, *Peeter Karpov Baschmakov*) või pärisnime(de)st ja üldnimest (*Kasepää küla*, *Sootaga Kõrts*, *Keiserlik Tartu sillakohus*). Kui nimeüksused on märgendatud, saab kasutajale lubada veelgi paindlikumat otsingut ja sirvimist, võimaldades näiteks leida protokolle, kus mainitakse mingit konkreetset isikut või kohta.

Nimeüksusi märgendati samuti EstNLTK vastava teegi abil, mis tuvastab nimeüksused ja jagab need kolme kategooriasse: isikud, kohad ja organisatsioonid. Selle töö põhineb masinõppel ja treeningmaterjalina on kasutatud tänapäevaseid ajakirjandustekste (Tkachenko jt 2013), milles mainitakse tunduvalt rohkem organisatsioone kui vallakohtu protokollides.

Esialgse katse käigus oli programmi väljundis tuvastatud nimeüksusi kokku *ca* 32 000, millest 86,14% moodustasid isikunimed, 8,42% kohanimed ning 5,44% organisatsiooninimed. Pistelisel kontrollimisel selgus aga, et paljud organisatsiooninimena märgendatud nimeüksused on tegelikult isiku- või kohanimed. Omaette

küsimus on koha ja organisatsiooni vaheline hägune piir: nii mõisa, talu kui ka nt kõrtsi võib olenevalt kontekstist tõlgendada nii ühe kui ka teisena.

Vigu nii nimeüksuste tuvastamisel kui ka morfoloogilisel analüüsil üld- ja pärisnime eristamisel põhjustab suure algustähe tänapäeva mõistes ülekasutus, nt *andis ülles et tema Maamõtja kullu 20 rbl om* (Haaslava); *2 rbl vaeste Ladiko sisse masma* (Luke); *Johan Bergman ei olle Sure Tee ei ka kerriko tee krusa vedanu* (Haaslava).

Teiseks probleemiks on mitme nimeüksuse kõrvuti esinemine: kui kõrvuti on mitu pärisnime või lihtsalt mitu suure algustähega sõna, kaldub nimeüksuste tuvastaja pidama neid sama nimeüksuse osadeks. Nt lauses *et kui mõtsavaht Daniel Perg Tartun käiman ollu* (Kahkva) on tuvastatud isikunimena *Daniel Perg Tartun* või lauses *antis kaibuses ette et Piiri Mihkel Rentirahha 25 rbl võlgo om* (Luke) on organisatsiooninimena tuvastatud *Piiri Mihkel Rentirahha*. Lisa- või kohanimest ning perekonna- ja eesnimest koosnevaid isikute mainimisi (*Jaani Jüri Hans*, *Tullitse Märt Nappasson*, *Karel Läst Allatskivilt*) ongi võimalik tõlgendada mitmeti: kas eraldi koha- ja isikunimena või ühe isikunimena.

Nimeüksuste märgenduse uurimisel jäid silma mitmed süstemaatilised märgendusvead ning seetõttu prooviti heuristikuid märgenduse automaatseks parandamiseks. Valetuvastused eemaldati näiteks sagedasematelt suure algustähega ase-, side- ja määrsõnadelt (*Sel*, *Seejärele*, *Nink*, *Peris*, *Perrast* jm) ning kooloniga lõppevatelt suure algustähega sõnadelt, mis viitasid kas kohtuotsuse väljakuulutamisele (nt *Moisteti*, *Moistetu*, *Mõistm*) või isiku ametile (nt *Vallavanemb*). Kohanimede puhul täheldati, et sageli järgneb kohanimele väiketäheline liigisõna (nt *Järso talo* või *Joosu vald*) ning selliste sõnade nimekirja alusel lisati kohanimemärgendusi ka kontekstidesse, kus eelneval suure algustähega sõnal märgendus puudus või oli vale. Isikunimede puhul oli võimalik teha kindlaid parandusi nn täisnimede (kahe- ja kolme-sõnalised isikunimed, nt *Leonti Semenov Leschnev*) märgendustes.

Kui täisnime oli mainitud mitu korda kas ühe protokolli piires või terve valla protokollides ning vähemalt üks mainimine oli isikunimena tuvastatud, sai märgenduse automaatselt üle kanda ka sama täisnime teistele (märgendamata või valesti märgendatud) esinemistele, kui nimed erinesid väikese teisenduskauguse[3] võrra. Näiteks kanti isikunime märgendus sõnelt *Leonti Semenov Leschnevit* üle sõnele *Leonti Semenov Leschnev*, kuna nende sõnede vaheline teisenduskaugus oli 2 ja see jäi kolmesõnaliste nimede maksimaalse lubatud erinevuse (katseliselt leitud 2 teisenduse) piiresse.

Heuristiliste paranduste tulemusel nimeüksuste koguarv oluliselt ei muutunud, küll aga muutus nimeüksuste liigiline jaotus: isikunimede osakaal tõusis 88,21%-ni, samas langes kohanimede osakaal 7,81%-ni ning organisatsiooninimede osakaal 3,98%-ni. Vähenes ka selliste sõnede hulk, mis olid märgendatud erinevat liiki nimeüksustena (nt sõne oli märgendatud kord isikuna, kord asukohana): kui algul oli selliseid sõnesid 9,1% kõigist unikaalsetest nimeüksustest, siis pärast parandusi 6,19%. Need kaudsed näitajad viitavad sellele, et vähemalt osa automaatparandustest täitis ka oma eesmärki. Täpse hinnangu paranduste kvaliteedile saab aga anda pärast nimeüksuste märgenduste käsitsi kontrollimist, ent see ettevõtmine ei mahtunud käesoleva töö raamidesse.

[3] Teisenduskaugus mõõdab, mitu tähte tuleb minimaalselt muuta (lisada, kustutada või asendada), et saada ühest sõnest teine.

4. Mis saab edasi?

Artiklis kirjeldatud tekstikorpus moodustab vaid tagasihoidliku valimi kõigist Rahvusarhiivis säilitatavatest vallakohtute protokollidest. Kuigi vallakohtute protokolliraamatud on juba aastaid digiteerituna veebis kättesaadavad, ei ole protokollide tekstide ulatuslikumat sisestustööd ette võetud. Tuginedes senistele edukatele kogemustele vabatahtlike erinevatesse nn *crowdsourcing*'u ehk ühisloome projektidesse kaasamisel (Eestlased Esimeses maailmasõjas, Tartu 1867), alustas Rahvusarhiiv 2017. aastal uue vallakohtute ühisloomekeskkonna arendamist, mille eesmärgiks on pakkuda mugavat võimalust vallakohtute protokollide veebipõhiseks sisestamiseks (http://www.ra.ee/vallakohtud/). Keskkond hõlmab kogu vallakohtute vanemat kirjalikku pärandit ehk enam kui 2200 protokolli- ja lepinguraamatut kokku 235 000 digikujutisega. Sisestamiseks sobiva vallakohtuni on võimalik jõuda nii kaardivaates kui ka ajalooliste haldusüksuste loetelu kaudu. Et tulevane tekstimassiiv oleks kasutatav eri valdkondade uurijatele, oodatakse sisestajatelt teksti originaalitruud sisestamist. Lisaks protokolli tekstile saab eraldi väljale märkida ka kohtu koosseisu, daatumi ja protokolli temaatika, mille aluseks on etteantud loetelu. Samuti on juba sisestatud tekstides võimalik märgendada isiku- ja kohanimesid ning raskesti väljaloetavaid sõnu või lõike. Kõik sisestatud tekstid on hiljem muudetavad ja parandatavad teiste kasutajate poolt ning samuti huviliste jaoks allalaetavad. Kuna allikate hulk on väga suur, siis kujuneb projekt kindlasti pikaajaliseks. Käesolevas artiklis kirjeldatud automaatset morfoloogilist analüüsi, mida on kohandatud vallakohtuprotokollide materjalide jaoks, on edaspidi võimalik rakendada ka ühisloome käigus sisestatud tekstidele.

5. Kokkuvõte

19. sajandi vallakohtute digiteeritud protokollide kättesaadavaks ja otsitavaks tegemine ning tekstide varustamine keelelise, geograafilise ja temaatilise infoga loob digitaalse ressursi, mis pakub uurimisainest eri distsipliinidele ning millel on seetõttu väga lai potentsiaalne kasutajaskond. Temaatiline liigitus, kirjakeelestatud algvormi lisamine tekstisõnadele ning nimeüksuste märgendamine võimaldavad paindlikku otsingut ning pakuvad mitmeid võimalusi tolleaegse elu-olu ja sotsiaalsete võrgustike analüüsiks.

Artiklis kirjeldatud katse vallakohtuprotokollide tekste tänapäeva keele loomuliku automaattöötluse vahenditega analüüsida annab infot selle kohta, kui suurel määral tolleaegne, alles tekkiv kirjakeel tänapäeva kirjakeelest erineb, mis on protokollitekstide põhilised iseloomulikud jooned ning kas need jooned erinevad piirkonniti või kirjutajati. Põhiliselt valmistab tänapäeva keelel treenitud morfoloogiaanalüsaatorile raskusi vana kirjaviis, mis tingib selle, et analüsaator kas ei tunne sõnavormi üldse ära või annab sellele vale analüüsi. Protokollides kasutatud murdekeele eripärad ei tule analüüsist nii selgelt esile, ehkki tundmatute sõnavormide ja valede analüüside hulgas esineb sagedasti just lõunaeestilist vormimoodustust. Samuti on, eelkõige suure algustähe ülekasutuse tõttu, morfoloogilise analüüsi käigus keeruline eristada päris- ja üldnimesid. Pärisnimede paremaks tuvastamiseks saab kasutada näiteks automaatset nimetuvastust, mida on

kohandatud teksti eripärast lähtuvalt, samuti on võimalik nimetuvastuse väljundit kaasata morfoloogilisse analüüsi.

Käsitsi kontrollitud materjali hulk ei ole väga suur, ent võimaldab analüsaatorit iga valla tekstide jaoks kohandada, luues tundmatute ning valede analüüside põhjal kasutajasõnastikud ning teisendusmustrid, mille abil saab tekstisõnad automaatselt muuta tänapäeva kirjakeeles kasutatavatele sarnasemaks. Selle tulemusel peaks paranema märkimisväärselt analüüsi kvaliteet. Kohaldatud analüsaatorit saab kasutada teiste sarnaste tekstide, nt ühisloome käigus sisestatud protokollide analüüsiks. Käesoleva projekti käigus analüüsitud tekstid on antud üle Tartu Ülikooli vana kirjakeele korpusele ning on otsitavad lehel http://vakk.ut.ee.

Viidatud kirjandus

Anepaio, Toomas 2007. Vallakohus – kas ainult talurahva kohus? [‘Communal courts – peasant courts only?’] – Ajalooline Ajakiri, 3 (4), 343–368.

Bollmann, Marcel 2013. POS tagging for historical texts with sparse training data. – Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse. August 8-9, 2013 Sofia, Bulgaria. Stroudsburg, PA: Association for Computational Linguistics, 11–18. http://aclweb.org/anthology/W13-2300 (6.3.2019).

Hiio, Ene 1996. Ülevaade vallakohtute materjalidest Eesti Ajalooarhiivis [‘Die Übersicht über die Materialen der Gemeindegerichte im Estnischen Historischen Archiv’]. – Artiklite kogumik Eesti Ajalooarhiivi 75. aastapäevaks. Eesti Ajalooarhiivi toimetised, 1 (8), 143–155.

Hint, Mati 2008. Tartu keele avaliku kasutamise taandareng vajab täpset dokumenteerimist [‘Decline of the public use of the South Estonian language needs precise documentation’]. – Keel ja Kirjandus, 7, 553–556.

Kaaristo, Maarja 2004. Peksmine ja löömine Eesti külas 1868–1911 Nursi vallakohtu protokollide näitel. Õigusetnoloogiline perspektiiv [‘Investigation into the records of the communal court of Nursi: the discussion of the beating cases in Estonian rural village society in 1868–1911 in the perspective of legal anthropology’]. – Mäetagused, 27, 31–46.

Kaaristo, Maarja 2006. Vägivald loomade vastu: inimene ja koduloom Lõuna-Eesti külas 19. sajandi II poolel vallakohtute protokollide näitel [‘Violence towards animals: Humans and animals in South-Estonian villages in the second half of the 19th century on the example of parish court records’]. – Mäetagused, 31, 49–62.

Kallio, Petri 2012. The prehistoric Germanic loanword strata in Finnic. – Riho Grünthal, Petri Kallio (Toim.), A Linguistic Map of Prehistoric Northern Europe. Suomalais-ugrilaisen seuran toimituksia 266. Helsinki: Suomalais-ugrilainen seura, 225–238.

Kask, Arnold 1958. Võitlus vana ja uue kirjaviisi vahel XIX sajandi eesti kirjakeeles [‘Struggle between the old and new spelling in 19th-century Standard Estonian’]. Tallinn: Eesti Riiklik Kirjastus.

Kurema, Kristiine 2013. Kuidas kajastus üleminek vanalt kirjaviisilt uuele Peetri kohtuprotokollide keeles [‘Transition from the old orthography to the new during the second half of the 19th century based on the court protocols of Peetri parish’]. – ESUKA / JEFUL, 4 (3), 55–72. https://dx.doi.org/10.12697/jeful.2013.4.3.03

Linnus, Jüri 1970. 19. sajandi talurahvakohtute materjalid rahvakultuuri uurimise allikana [‘19th century communal court materials as a source to study folk culture’]. – Emakeele Seltsi aastaraamat, 16, 231–242.

Loftsson, Hrafn 2013. Tagging the past: Experiments using the Saga corpus. – Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen (Eds.), Proceedings of the 19th Nordic

Conference of Computational Linguistics (NODALIDA-2013). Linköping: Linköping University Electronic Press, 89–104.
Must, Aadu 1997. http://www.history.ee/ [Eesti õigusajaloo krestomaatia leheküljest Internetis]. – Kleio, 1, lk 64–65.
Must, Kadri 1998a. Tori vallakohtu protokollid ajalooallikana [‘The records of the communal court of Tori as a historical source’]. – Ajalooline Ajakiri, 3, 93–108.
Must, Kadri 1998b. Tori vallakohtu arhivaalid ajalooallikana. http://www.aai.ee/~urmas/tor/kadri.htm (4.1.2019).
Orasmaa, Siim; Petmanson, Timo; Tkachenko, Alexander; Laur, Sven; Kaalep, Heiki-Jaan 2016. ESTNLTK – NLP toolkit for Estonian. – Proceedings of LREC 2016, 2460–2466.
Petterson, Eva 2016. Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Studia Linguistica Upsaliensia 17. Uppsala: Uppsala Universitet.
Piotrowski, Michael 2012. Natural language processing for historical texts. – Synthesis Lectures on Human Language Technologies, 5 (2), 1–157. https://doi.org/10.2200/S00436ED1V01Y201207HLT017
Prillop, Külli 2004. Kuidas märksõnastada vanu eestikeelseid tekste? [‘How to lemmatize old Estonian texts’] – Keel ja Kirjandus, 2, 90–99.
Puss, Fred 2018. Kirjaviisivahetus kirikuraamatutes [‘Change of spelling style in parish registers’]. – Emakeele Seltsi aastaraamat, 63 (2017), 166–200. https://dx.doi.org/10.3176/esa63.08
Raag, Raimo 2008. Talurahva keelest riigikeeleks [‘From the language of peasants to state language’]. Tartu: Atlex.
Rögnvaldsson, Eirikur; Helgadóttir, Sigrún 2011. Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. – Caroline Sporleder, Antal van den Bosch, Kalliopi Zervanou (Eds.), Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, 63–76.
Sánchez-Marco, Cristina; Boleda, Gemma; Padró, Lluís 2011. Extending the tool, or how to annotate historical language varieties. – Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH’11). Portland, Oregon, June 24, 2011. Stroudsburg, PA: Association for Computational Linguistics, 1–9.
Scheible, Silke; Whitt, Richard J.; Durrell, Martin; Bennett, Paul 2011. Evaluating an ‘off-the-shelf’ POS-tagger on Early Modern German text. – Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH’11). Portland, Oregon, June 24, 2011. Stroudsburg, PA: Association for Computational Linguistics, 19–23.
Schmid, Helmut 1995. Improvements in part-of-speech tagging with an application to German. – Proceedings of the ACL SIGDAT-Workshop, 47–50.
Schmid, Helmut; Laws, Florian 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. – Proceedings of the 22nd International Conference on Computational Lingsuistics (COLING 2008). Manchester, United Kingdom, August 18–22, 2008. Stroudsburg, PA: Association for Computational Linguistics, 777–784.
Tkachenko, Alexander; Petmanson, Timo; Laur, Sven 2013. Named entity recognition in Estonian. – Proceedings of the Workshop on Balto-Slavic NLP, 8-9 August 2013, Sofia, Bulgaria. Stroudsburg, PA: Association for Computational Linguistics, 78–83.
Traat, August 1971. Põhijooni vallakohtu arengust Eestis kuni 1866. aastani [‘The key characteristics of parish courts in Estonia till the 1866 reform’]. – Eesti NSV Teaduste Akadeemia Toimetised 20. köide. Ühiskonnateadused 1, 34–45.

Traat, August 1980. Vallakohus Eestis 18. sajandi keskpaigast kuni 1866. aasta reformini ['Parish courts in Estonia from mid- 18th century to the 1866 reform']. Tallinn: Eesti Raamat.
Türna, Tõnis 2004. 1860.–80. aastate Lõuna-Eesti vallakohtute protokollid. Massiliste täistekst-andmebaaside loomise, publitseerimise ja kasutamise metoodika ['South Estonian communal court minute books dating from 1860–1880: methodology of creating a digital resource, publication and use']. Peaseminaritöö. Tartu: Tartu Ülikool.
Viitso, Tiit-Rein 1985. Läänemeresoome murdeliigenduse põhijooned ['Main characteristics of dialect classification of Finnic languages']. – Keel ja Kirjandus, 7, 399–404.
Wiedemann, Ferdinand Johann 2011 [1875]. Eesti keele grammatika. Heli Laanekask (Tõlk.), Ellen Niit (Toim.). Tallinn: Eesti Teaduste Akadeemia Emakeele Selts.

Võrguviited

Eestlased Esimeses maailmasõjas. Rahvusarhiivi ühisloome algatus. http://www.ra.ee/ilmasoda (4.1.2019).
EstNLTK = Open source tools for Estonian natural language processing. https://github.com/estnltk (12.3.2019).
Saaga. www.ra.ee/saaga (4.1.2019).
Tartu 1867. Tartu linna ja Rahvusarhiivi ühisloome algatus (oktoober 2017 kuni juuli 2018). http://www.ra.ee/tartu1867 (4.1.2019).
VAKK = Eesti vana kirjakeele korpus. http://vakk.ut.ee (4.1.2019).
Vallakohtud. Rahvusarhiivi ühisloomerakendus. http://www.ra.ee/vallakohtud (4.1.2019).

CREATING A DIGITAL RESOURCE FROM 19TH CENTURY COMMUNAL COURT MINUTE BOOKS

Maarja-Liisa Pilvik[1], Kadri Muischnek[1], Gerth Jaanimäe[1], Liina Lindström[1], Kersti Lust[2], Siim Orasmaa[1], Tõnis Türna[2]

University of Tartu[1], National Archives of Estonia[2]

This article describes an interdisciplinary attempt to create a digital resource from Estonian communal court minute books dating from 1866–1890, with the focus lying on using contemporary natural language processing tools for analyzing archaic language. The database contains nearly 420 000 tokens in XML-tagged files. The texts are linguistically diverse: the parallel use of old and new spelling systems, dialects, and the background of the parish clerk bring about a lot of language variation. There are also differences in the orthographic choices made during the manual insertion of the texts. For the purpose of linguistic analysis and tagging, automatic morphological analysis and named entity recognition was tested using EstNLTK libraries. A closer examination of the output suggested that it is necessary to use both text normalization and tool adaption for improving the quality of automatic analyses. This would result in tools, which would perform better at analyzing similar texts and which could, therefore, be applied in the automatic analysis crowd-sourced material. Making the communal court minute books accessible and searchable by supplying linguistic and topical information creates a rich digital resource which is subject of interest for many disciplines.

Keywords: natural language processing, automatic morphology, digital humanities, corpus linguistics, databases, language history, Estonian

Maarja-Liisa Pilviku (Tartu Ülikool) teaduslikud huvid on keele varieerumine, eesti murded, korpuslingvistika ja kvantitatiivsed meetodid keeleuurimises.
Jakobi 2-430, 51005 Tartu, Estonia
maarja-liisa.pilvik@ut.ee

Kadri Muischneki (Tartu Ülikool) teaduslikud huvialad on korpuslingvistika, eesti keele süntaktiline struktuur ning automaatne süntaktiline analüüs.
Jakobi 2-426, 51005 Tartu, Estonia
kadri.muischnek@ut.ee

Gerth Jaanimäe (Tartu Ülikool) teaduslikud huvialad on korpuslingvistika ja automaatne tekstianalüüs.
Jakobi 2-430, 51005 Tartu, Estonia
gerthj@gmail.com

Liina Lindströmi (Tartu Ülikool) teaduslikud huvid on süntaks, dialektoloogia, murdesüntaks, keele varieerumine ja muutumine ning keelekontaktid.
Jakobi 2-120, 51005 Tartu, Estonia
liina.lindstrom@ut.ee

Kersti Lust (Rahvusarhiiv) uurib Eesti talurahva ajalugu.
Nooruse 3, 50411 Tartu, Estonia
kersti.lust@ra.ee

Siim Orasmaa (Tartu Ülikool) uurimishuvideks on infootsingu meetodid ja informatsiooni ekstraheerimine vabatekstist.
J. Liivi 2-329, 50409 Tartu, Estonia
siim.orasmaa@ut.ee

Tõnis Türna (Rahvusarhiiv) uurib Eesti talurahva 19. sajandi ajalugu.
Nooruse 3, 50411 Tartu, Estonia
tonis.tyrna@ra.ee