doi:10.5128/ERYa17.03

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 17, 47–59

AJALOOLISTE TEKSTIDE NORMALISEERIMINE

Gerth Jaanimäe

Ülevaade. Ajalooliste tekstide normaliseerimine ehk tänapäevasele kujule viimine võimaldab uurida tekste praeguse keele analüüsivahenditega, otsida tekstidest märksõnu ning võrreldes tänapäevaseid ja vanu kirjakujusid saada paremini aimu keele muutumise kohta. Käesolev artikkel annab ülevaate normaliseerimisest, selle erinevatest meetoditest, mujal maailmas tehtud katsetustest selles valdkonnas ning normaliseerimise põhiproblemaatikast 19. sajandi teisest poolest pärinevate eestikeelsete tekstide näitel.*

Võtmesõnad: loomuliku keele töötlus, normaliseerimine, keeleajalugu, korpuslingvistika, mittestandardne keel, eesti keel

1. Sissejuhatus

Ajalooliste tekstide automaatne analüüs on tänapäeval huvipakkuv ja aktuaalne suund nii keeleteadlastele, ajaloolastele, etnoloogidele, perekonnaloo uurijaile kui ka kõigile teistele, kes kasutavad oma töös digitaalsel kujul ajaloolisi tekstimaterjale. Eesti digiarhiivides on selliseid tekste praeguseks rohkem kui kunagi varem, aga enamasti on need talletatud piltidena, mida ei saa uurida automaatsete tekstianalüüsi vahendite abil. Sageli on tegemist käsikirjaliste tekstidega ja kuna käekirjad on erinevad, ei ole neid võimalik tärktuvastuse (pildifailist automaatse trükiteksti tuvastamise) abil tekstiks teisendada. Seega peab neid tekste ümber trükkima ehk masinloetavaks muutma inimene.

Rahvahanke korras on sel moel Rahvusarhiivi andmebaasi sisestatud rohkem kui 55 000 käsikirjalist eestikeelset vallakohtu protokolli, mis sisaldavad rohkem kui 2,5 mln tekstisõna[1] ning tärktuvastuse abil on digiteeritud vanu ajalehetekste. Kuigi tärktuvastusega on tekstide masinloetavaks muutmine odavam ja kiirem, on probleemiks tekstituvastuse käigus tekkinud vigade suur arv. Vanade tekstide automaatseks lingvistiliseks analüüsimiseks tihti ei piisa nende masinloetavaks muutmisest.

* Artikli valmimist on toetanud riikliku programmi "Eesti keel ja kultuur digiajastul 2019–2027" projekt "Ajalooliste tekstide automaatse analüüsi võimalused Eesti 19. sajandi vallakohtuprotokollide näitel". Autor tänab oma doktoritöö juhendajaid Kadri Muischneki, Siim Orasmaad ja Külli Prillopit.

[1] Ühisloome platvorm https://www.ra.ee/vallakohtud/ (20.1.2021). Autori kasutuses olevas vanemas väljavõttes on sõnu 2 570 000, praegune arv on aga teadmata.

Enamasti ei vasta vanad kirjapanekud tänapäevastele keelereeglitele. Nende tekstide keel varieerub nii kirjutaja kirjaoskuse, eri aegadel kehtinud ortograafiareeglite, sõnavara, morfoloogia, lauseehituse, murdejoonte kui ka üldise keelekasutuse poolest. Variatiivsus teeb materjali lingvistilisest seisukohast väga huvitavaks, kuna annab hea ülevaate keele muutumise kohta aja jooksul, kuid samal ajal teevad vanapärasus ja erisused tekstide automaatse analüüsimise väga keeruliseks, sest enamasti ei ole võimalik nende uurimiseks tänapäevase keele analüüsivahendeid kasutada. (Piotrowski 2012: 12–24)

Vanade tekstide automaatsel analüüsil on valida kahe põhimõtteliselt erineva lähenemise vahel, milleks on 1) praeguse standardkeele analüüsivahendite kohandamine mittestandardse keele tekstide jaoks ning 2) mittestandardsete tekstide kohandamine tänapäeva kirjakeele normile vastavaks (Pettersson 2016: 22). Praeguse keele analüüsivahendite kohandamine võimaldab arvestada ajaloolise perioodi keele eripäradega, nagu näiteks tänapäevasest keelest kadunud muutevormid. Selle lähenemise peamisteks puudusteks on aja- ja ressursimahukus, kuna luua tuleb uus morfoloogilise märgenduse süsteem ja selle alusel ka võimalikult suur käsitsi märgendatud tekstikorpus.

Mittestandardsete tekstide kohandamine tänapäeva kirjakeele normidele vastavaks ehk normaliseerimine, on oma põhimõttelt vastupidine: muudetakse algset teksti, mitte praeguse keele analüüsivahendeid. Normaliseerimiseks ei pea uut märgendust nullist üles ehitama, mistõttu sobib see hästi just niisuguste tekstide jaoks, milles esineb palju keelelist varieeruvust. Lisaks vajavad osad normaliseerimismeetodid oluliselt väiksema mahuga käsitsi märgendatud treeningkorpust ehk treeningandmeid, mille pealt algoritmid õpivad. Mõni normaliseerimismeetod ei vaja üldse treeningandmeid. Normaliseerimist kasutatakse näiteks ka internetikeele kui samuti normist hälbiva keelevariandi analüüsi hõlbustamiseks.

Normaliseerimisel on mitmeid praktilisi väärtusi. Peamine kasu seisneb selles, et tekstidest saab oluliselt paremini otsida erinevaid märksõnu. Näiteks kui on tarvis otsida sõna *hobune*, siis oleks vaja leida ka sellised kirjakujud nagu *obene*, *obune*, *hobbune* jne. Peale selle on normaliseeritud tekste võimalik uurida erinevate praeguse standardkeele jaoks väljatöötatud meetodite abil, nagu näiteks meelestatuse analüüs ja tekstide automaatne klassifitseerimine (van der Zwaan 2015). Lisaks on tekstide vanu ja tänapäevaseid kirjapilte võrreldes võimalik saada parem ettekujutus keele muutumisest aja jooksul. Tuleb muidugi arvestada, et ükski normaliseerimismeetod ei ole täiuslik ja mõningane vigade arv on paratamatu. Seega sõltub kasutajast ja muidugi meetodi korrektsusest, kas väljundit on vaja käsitsi kontrollida või mitte.

Käesolev artikkel keskendub ajalooliste tekstide normaliseerimisele. Kirjeldatakse normaliseerimise põhiproblemaatikat ning erinevaid normaliseerimismeetodeid, mida on mujal maailmas selles valdkonnas katsetatud. Samuti antakse esialgne hinnang nende meetodite sobivuse kohta vanade eestikeelsete tekstide analüüsimiseks. Tekstidena on kasutatud 19. sajandist pärinevaid vallakohtu protokolle, ajalehetekste jne, mis algselt on käsitsi kirjutatud, aga hiljem ümber trükitud või tärktuvastuse abil tekstiks teisendatud.

2. Normaliseerimismeetodid ja nende hindamine

Ajalooliste tekstide normaliseerimiseks kasutatakse mitmesuguseid meetodeid, mida saab tinglikult jaotada kaheks: 1) meetodid, mis nõuavad kasutajapoolseid teadmisi keele kohta ja 2) meetodid, mis neid ei nõua. Sarnast jaotust on oma doktoritöös kasutanud Eva Pettersson (Pettersson 2016: 47–49) Kummalgi rühmal on omad eelised ja puudused.

2.1. Meetodid, mis nõuavad kasutajapoolseid teadmisi

Keele kohta teadmisi nõudvaks lähenemiseks peetakse eelkõige reeglipõhist normaliseerimist (Piotrowski 2012: 75–76). Selle jaoks kirjeldatakse reeglitena ära muutused, mis on keeles või kirjaviisis aja jooksul toimunud või mis on murdelistes vormides standardkeelest erinev. Lähenemise peamine eelis seisneb selles, et reeglite abil on võimalik üsna üksikasjaliselt ära kirjeldada sõna kirjakuju teisendamine vanast kirjaviisist uude. Samal ajal on meetodil ka omad puudused, millest olulisim on ajamahukus, kuna paljude varieeruvuste korral läheb koostatavate reeglite hulk väga suureks. (Piotrowski 2012: 75–76) Probleemiks võib ilmselt olla ka see, et kõiki reegleid ja erandeid on raske ette näha, mistõttu võivad tekkida mitmesugused vead ja vastuolud. Näitena võib välja tuua, et eesti keele tänapäevases kirjaviisis on topeltkonsonandid asendatud ühekordsete konsonantidega, kui häälduses on lühike konsonant, näiteks vanale ortograafiale vastav kirjakuju *wägga* on tänapäeval kujul *väga*, samuti on eelnevas näites tänapäevases ortograafias *w* asendatud *v*-ga. Murdekeelse erisusena võib välja tuua lõunaeesti keeles mineviku partitsiibi tunnuse *nu* kasutamise tunnuse *nud* asemel, nt lõunaeestiline vorm *tennü* põhjaeestilise *teinud* asemel.

Reeglipõhiseid meetodeid on mitmeid. Ühena võib välja tuua lõplikud muundurid, mille abil kirjeldatakse erinevalt mõnest teisest sedaliiki meetodist üksikasjaliselt peale teisendusreeglite ka keele morfoloogia (Beesley, Karttunen 2003).

2.2. Meetodid, mis ei nõua kasutajapoolseid teadmisi

Keele kohta teadmisi mittenõudvate meetodite näidetena võib välja tuua sõnastikupõhise meetodi, teisenduskaugused, müraga edastuskanali mudeli, masintõlke jt. Järgnevalt vaadeldakse neid põgusalt.

Sõnastikupõhine lähenemine (Piotrowski 2012: 74), mida nimetatakse ka mälupõhiseks meetodiks, on oma olemuselt kõige lihtsam. Tööpõhimõte seisneb selles, et vaadatakse paralleelsõnastikust vanu sõnavorme ja nende tänapäevaseid vasteid. Oluline on silmas pidada, et sõnastikus ei pea tingimata olema sõnade lemmad, vaid võib olla ka mõni muu vorm. Meetodi eelis on suhteliselt hea töökindlus ehk vähene võimalus valesti normaliseerimida sageli esinevaid sõnu. Puudusteks on aga vajadus suure sõnapaare sisaldava leksikoni järele ja tõsiasi, et morfoloogiliselt rikkalikes keeltes tuleb iga sõnavormi käsitleda eraldi üksusena. Sõnastikupõhine meetod ei kata sõnavorme, mida sõnastikus ei esine. (Piotrowski 2012: 74)

Sagedamini esinevaid sõnavorme uurinud Juhan Tuldava leidis, et kui näiteks inglise ja prantsuse keeles katavad 2000 sagedamat sõnavormi vastavalt 79% ja

86% tekstist, siis eestikeelses tekstis ainult 59% (Tuldava 1977). Tuldava andmed näitavad, et rikkaliku morfoloogiaga keelte jaoks on keeruline luua kõiki tekstis esineda võivaid sõnavorme hõlmavat sõnastikku.

Probleeme tekitavad ka mitmesused. Näiteks võib sõna *vai* tähendada nii terava otsaga pulka kui ka murdelist varianti sidesõnast *või*.

Teisenduskauguse meetodi (ingl *edit distance*) abil on võimalik mõõta kahe tekstisõne omavahelist sarnasust. Teisenduskauguse leidmise algoritme on mitmeid, kõige tuntum ja laialdaselt kasutatum neist on Levenshteini teisenduskaugus (Levenshtein 1966).[2] Meetodiga võrreldakse omavahel kahte tekstisõnet ja vaadatakse, mitu teisendust tuleb teha, et ühest sõnest saaks teine. Teisenduste hulka loetakse tähemärgi kustutamist, lisamist ja asendamist. Näiteks on sõnede *pesnu* ja *peksnud* teisenduskaugus 2, kuna ühest sõnest teise saamiseks on vaja lisada tähemärgid *k* ja *d*.

Meetodi eeliseks vaatamata selle vanusele on lihtsus ja üsna kergesti rakendatavus piisava sõnavara korral. Puuduseks on aga, nii nagu ka sõnastikupõhisel meetodilgi, oht mitmesuste tekkeks, ehk ühele vanale sõnavormile võib leiduda mitu sama teisenduskaugusega tänapäevast vastet. Näiteks on sõnede *ärra* ja *härra* omavaheline teisenduskaugus 1, aga sama teisenduskaugus on ka sõnedel *ära* ja *ärra*. Järgnevas vallakohtu protokollist pärinevast näitest on konteksti alusel arusaadav, et *ärra* tähendab *ära*.

(1) Kohtu ette tulli kõrtsmik Hans Sults ja ütles, et Jaan Komberg ei olla temma päwad mitte kõrra pärrast ärra teinud; (Juuru, Maidla vald, 6.8.1877)

Kirjeldatud probleemi lahendamiseks saab kasutada kaalutud ehk üldistatud teisenduskaugust (Bollmann jt 2011). Selleks omistatakse teisendustele erinevad kaalud ning sellega vähendatakse osade sõnede omavahelist kaugust. Näiteks kui määrata sõna keskel kahekordse *r*-i asendamisele väiksem kaal, oleks *ärra* ja *ära* omavaheline kaugus väiksem kui sõnedel *ärra ja härra* ja seega oleks teisenduskauguse algoritmi jaoks sarnasemad just *ärra* ja *ära*.

Levenshteini teisenduskaugus moodustab üsna olulise osa müraga edastuskanali mudelis (ingl *noisy channel model*; Brill, Moore 2000, Jurafsky, Martin 2019), mille eesmärk on leida sõnale kirjakeelne vaste ka juhul, kui tähemärgid selles on omavahel segamini läinud. Seepärast kasutatakse seda üsna sageli õigekirja automaatseks korrigeerimiseks (Brill, Moore 2000). Meetod käsitleb kirjakeelenormile mittevastavat sõna n-ö müraga segunenuna. Sõnavormi normaliseerimiseks reastatakse kirjakeele normile vastavad sõnad nende esinemissageduse ja teisenduskauguse järgi. Kõige sagedamini esinev ja kõige väiksema teisenduskaugusega sõna ongi eeldatavasti mittekirjakeelse sõna kirjakeelne vaste. (Jurafsky, Martin 2019)

Statistiline masintõlge (Scherrer, Erjavec 2013) seostub enamasti lausete tõlkimisega ühest keelest teise. Sama meetodit on kasutatud ka vanade tekstide tänapäevasele kujule teisendamiseks. Ainult et paralleelkorpuses käsitletakse sõnu kui lauseid ja tähemärke kui sõnu. Seega fraasimustrite teisendamise asemel õpitakse automaatselt teisendama täheühendeid. Nii on meetodi eeliseks võimalus teisendada ka niisuguseid sõnu, mis on paralleelkorpuses olevast sõnavarast üsna erinevad. Puuduseks on aga vajadus suure treeningkorpuse järele. (Scherrer, Erjavec 2013)

[2] Vt ka https://et.wikipedia.org/wiki/Levenshteini_kaugus (2.4.2021).

Peale statistilise masintõlke on katsetatud ka närvivõrkudel põhinevat masintõlget, mille käigus õpetatakse tehisneuronite ühenduste abil arvutile nii normaliseeritud kui ka normaliseerimata tekstides leiduvaid mustreid. Meetod on enamasti andnud küll paremaid tulemusi, aga sageli on selle rakendamiseks vaja veelgi suuremat treeningkorpust. (Korchagina 2017)

Selleks et aimu saada, kui hästi üks või teine normaliseerimismeetod töötab, on neid vaja kuidagi hinnata. Selleks kasutatakse enamasti mõõdikuna korrektsust või täpsust ja saagist. Korrektsus (ingl *accuracy*) näitab, kui suur osa mingi meetodi ennustustest on õiged. Mõõdiku peamine eelis on lihtsus ja arusaadavus, probleemid tekivad aga siis, kui andmestikus on üht liiki andmeid oluliselt rohkem kui teisi. Näiteks võib tuua rämpsposti kindlakstegemise. Rämpsposti osakaal kõigi kirjade hulgas on enamasti üsna väike, mistõttu mitte-rämpspostiks märgitud kirjad suurendavad korrektsust lihtsalt sellepärast, et sellised kirjad on valdavas enamuses. Seetõttu kasutatakse sageli hindamiseks täpsust (ingl *precision*) ja saagist (ingl *recall*). Täpsus näitab mingi algoritmi ennustustest õigete osakaalu. Saagis näitab, kui suure osa kuldstandardi andmestikust algoritmi korrektsed ennustused ära katavad. (Jurafsky, Martin 2019) Kuldstandard näitab antud kontekstis inimeste käsitsi normaliseeritud tekste.

Normaliseerimise kontekstis tähendab madal täpsus, et algoritm teeb palju vigu, madal saagis, aga, et paljud sõnad jäävad normaliseerimata.

3. Seni tehtud katsed normaliseerida ajaloolisi tekste

Ajalooliste tekstide tänapäevasele kujule teisendamise katseid on tehtud mitmeid. Põhjalikult on vanade tekstide normaliseerimist oma doktoritöös käsitlenud Eva Pettersson (2016), kes katsetas reeglipõhist lähenemist, Levenshteini teisenduskaugust, mälu- ehk sõnastikupõhist normaliseerimist, teisenduskauguse ja mälupõhise normaliseerimise kombineerimist ning statistilist masintõlget. Eksperimendid viidi läbi aastatest 1527–1812 pärinevate rootsikeelsete tekstidega. Kõigepealt normaliseeriti korpus käsitsi ning seejärel katsetati sellega võrdluseks eelnevalt loetletud meetodeid. (Pettersson 2016: 49–50) Saamaks teada, kui palju sõltub meetodi edukus keelest ja konkreetsest korpusest, katsetati samu meetodeid, välja arvatud reeglipõhine lähenemine, veel järgmiste keelekogude normaliseerimisel: ingliskeelne Innsbrucki kirjade korpus aastatest 1386–1698, saksakeelne GerManC korpus aastatest 1650–1800, ungari vanade käsikirjade (koodeksite) korpus aastatest 1440–1541 ning islandi vana kirjakeele korpus IcePaHC 15. sajandist. (Pettersson 2016) Katsetuste tulemusena saadud korrektsused võtab kokku tabel 1.

Tabelist 1 on näha, et kõige paremad olid tulemused ingliskeelsete tekstide puhul, kõige väiksemad korrektsused olid aga ungari ja islandi keele tekste normaliseerides. Esimese kehvemat tulemust seletab peamiselt asjaolu, et ungarikeelsed tekstid pärinesid teiste tekstidega võrreldes varasemast perioodist. Teine põhjus seisneb väga tõenäoliselt selles, et kuna ungari keel on morfoloogiliselt rikkalik, toob see kaasa tekstides oluliselt suurema erinevate sõnavormide arvu. (Pettersson 2016)

Tabel 1. Meetodite edukus erinevate korpuste normaliseerimisel (Pettersson 2016: 82). M2m ja GIZA++ on statistilise masintõlke süsteemid; unigramm ja bigramm märgivad, et masinõppeks on võetud üksikud tähemärgid (unigramm) või tähepaarid (bigramm)

Meetod	Keelekorpus				
	inglise	saksa	ungari	islandi	rootsi
Normaliseerimismeetodit rakendamata	75.8	84.4	17.1	50.5	64.6
Levenshtein	82.9	87.3	31.7	67.3	79.4
Mälupõhine	91.7	94.6	75.0	81.7	86.2
Levenshtein+mälu	92.9	95.1	76.4	84.6	90.8
GIZA++ unigrammid	94.3	96.6	79.9	71.8	92.9
GIZA++ bigrammid	92.4	95.5	80.1	71.5	92.5
m2m unigrammid	90.6	96.0	79.4	71.2	92.3
m2m bigrammid	88.0	95.6	79.5	71.5	92.2

Scherrer, Erjavec (2013) katsetasid normaliseerimist statistilise masintõlke abil 18. ja 19. sajandist pärinevate sloveenikeelsete tekstide peal. Korrektsuseks saadi juhendamata masinõppe meetodiga 35% ja juhendatud masinõppe meetodiga 57%. Juhendamata masinõppe käigus anti masintõlkesüsteemile n-ö õppematerjaliks sõnapaaridena ette sõna vana kirjakuju ja sellest kõige väiksema teisenduskaugusega tänapäevane sõna. Juhendatud masinõppes oli süsteemi sisendiks nii vana kui ka tänapäevane kirjakuju. (Scherrer, Erjavec 2013)

Natalia Korchagina on katsetanud närvivõrkudel põhinevaid meetodeid aastatest 1450–1550 pärinevate saksakeelsete tekstide normaliseerimisel , saades korrektsuseks 81%. Võrdluseks rakendati samade andmete peal statistilist masintõlget, mille korrektsuseks saadi 79%. (Korchagina 2017)

Tehisnärvivõrke on rakendanud Pettersson (2018), eesmärgiga võrrelda meetodi edukust statistilise masintõlkega. Märkimisväärne oli tulemuste paranemine ungarikeelsete tekstide normaliseerimisel, statistilise masintõlke korrektsus oli *ca* 80% ja närvivõrkudel põhineva tõlke korrektsus 92%. Samal ajal vähenes korrektsus veidi rootsikeelsete ja jäi peaaegu samaks saksakeelsete tekstide puhul. Katse autorid selgitavad seda väheste treeningandmetega nendes keeltes. (Pettersson, Tang jt 2018)

Tehisnärvivõrke on rakendatud ka Helsinki ülikoolist 15. –19. sajandist pärinevate ingliskeelsete tekstide (*ca* 183 500 sõna) vanadest kirjakujudest ja nende normaliseeritud vormidest koosneval paralleelkorpusel (Hämäläinen jt 2019). Erinevate tehisnärvivõrkude mudelite korrektsused jäid 36% ja 78% vahele.

Müraga edastuskanali mudelit on katsetatud võrdlemaks baski-, sloveeni- ja hispaaniakeelsete tekstide normaliseerimist. Kuna eeldati, et keeles toimunud muutused on pigem fonoloogilised, siis teisendati enne mudeli rakendamist tähemärgid foneemideks. Baski keele korpusena kasutati 1633. aastal ilmunud raamatut "Gero". Tänapäevase kirjakeelega võrdlemiseks normaliseeriti osa raamatust käsitsi. Täpsuseks saadi 91% ja saagiseks 79%. (Etxeberria jt 2016) Samadel andmetel katsetati ka sõnastikupõhist lähenemist, millega saadi täpsuseks 95% ja saagiseks 39%. Seega, kuigi sõnastikupõhist lähenemist kasutades oli täpsus veidi parem, oli saagis olulisel määral väiksem, mis tähendab, et suur hulk sõnu jäi normaliseerimata.

Sama töö käigus katsetati müraga edastuskanali mudelit keskajast pärinevate hispaaniakeelsete tekstide normaliseerimisel. Täpsuseks saadi 97% ja saagiseks 95%. Sama meetodit katsetati ka 18. ja 19. sajandist pärinevate sloveenikeelsete tekstide paralleelkorpusel. Kuna nimetatud keeles on sel perioodil kasutatud kolme erinevat tähestikku, jagati korpus vastavalt kolmeks. Korrektsused jäid 67% ja 87% vahele. (Etxeberria jt 2016)

Ajalooliste tekstide normaliseerimist on põgusalt katsetatud ka eesti keeleandmete peal. Sõnastiku- ja reeglipõhist poolautomaatset normaliseerimist on rakendanud Külli Prillop 17. sajandist pärinevate Heinrich Stahli ja Joachim Rossihniuse tekstide märksõnastamiseks. Mõlematelt autoritelt valiti korpusesse 4200 sõna. Teisendusreeglite abil suudeti põhjaeestikeelseid Stahli tekste normaliseerida korrektsusega 71%, seevastu reegleid rakendamata vastas tänapäevasele kirjakeele normile 21% sõnadest. Lõunaeestikeelsete Rossihniuse tekstide normaliseerimisel saadi pärast teisendusi korrektsuseks 62%. Kehvema tulemuse põhjuseks oli põhjaeesti keelest erinev sõnavara. Stahli teksti normaliseerimiseks katsetati lisaks reeglitele ka sõnastikupõhist lähenemist, mille korrektsuseks saadi 98%. (Prillop 2004)

Projekti "1860–80 vallakohtuprotokollid kultuurimälu kandjana" käigus uuriti tänapäevase keele morfoloogilise analüüsi vahenditega 418 508 tekstisõnest koosnevat vallakohtuprotokollide korpust (Pilvik jt 2019). Eesmärgiks oli teada saada, kui palju on võimalik nimetatud perioodist pärinevaid tekste praeguse standardkeele analüüsivahendite abil uurida ja ühtlasi saada aimu, kui keeruline nende normaliseerimine tulevikus olla võib. (Pilvik jt 2019) Artiklis kirjeldatud lähenemise edasiarendusena on katsetatud põgusalt sõnastiku- ja reeglipõhist normaliseerimist sagedamini esinevatel sõnadel ja nimeüksustel.[3]

Eesti keele töötlusvahendite komplekt Estnltk 1.6[4] on tehtud kohandatavaks mittestandardse, sealhulgas vanema keelekasutuse analüüsimiseks (Laur jt 2020). Võimalik on kohandada sõnestust ja lausestust, mis vastab varasemale keelele. Lisaks saab sõnadele lisada nende normaliseeritud kirjakujusid ja teostada kasutajasõnastiku abil morfoloogilise analüüsi järelparandusi (Laur jt 2020).

Tabelis 2 on erinevate normaliseerimiskatsete tulemused, välja arvatud need, mis olid juba varasemalt kirjas tabelis 1.

Nagu eelnevalt kirjeldatud katsete kirjeldustest ja tabelitest 1 ja 2 nähtub, võivad sama või sarnaste meetodite rakendamise tulemused olla väga erinevad, isegi samas keeles olevate tekstide puhul. Põhjuseks on tõenäoliselt erinevatest valdkondadest ja ajalistest perioodidest pärinevad keeleandmed ning tekstikorpuste suurus. Samuti on tabelit 2 vaadates väga oluline silmas pidada, et kõrgemad korrektsused ei pruugi näidata, kui hästi tuleb mingi meetod toime mõne teistliiki andmehulgaga. Teisisõnu kui kasutada mõõdikutena korrektsuse asemel täpsust ja saagist, oleks nende täpsus võib-olla kõrge, aga saagis madal. See kehtib just sõnastikupõhise meetodi kohta.

Seega vastust küsimusele, milline normaliseerimismeetod töötab kõige paremini, sõltub suuresti keelest ja normaliseeritavatest tekstidest.

[3] https://bitbucket.org/utDigiHum/vallakohtuprotokollid (20.1.2021).
[4] https://github.com/estnltk/estnltk (20.1.2021).

Tabel 2. Normaliseerimiskatsete tulemused

Meetod	Keel	Korrektsus	Täpsus	Saagis	Viide
Statistiline juhendamata masintõlge	sloveeni	35%			Scherrer, Erjavec 2013
Statistiline juhendatud masintõlge	sloveeni	57%			Scherrer, Erjavec 2013
Närvivõrkudel põhinev masintõlge	saksa	81%			Korchagina 2017
Statistiline masintõlge	saksa	79%			Korchagina 2017
Närvivõrkudel põhinev masintõlge	ungari	92%			Pettersson, Tang jt 2018
Närvivõrkudel põhinev masintõlge	rootsi	91%			Pettersson, Tang jt 2018
Närvivõrkudel põhinev masintõlge	saksa	96%			Pettersson, Tang jt 2018
Närvivõrkudel põhinev masintõlge	inglise	35% – 78%			Hämäläinen jt 2019
Müraga edastuskanal	baski		91%	79%	Etxeberria jt 2016
Sõnastikupõhine	baski		95%	39%	Etxeberria jt 2016
Müraga edastuskanal	hispaania		97%	95%	Etxeberria jt 2016
Müraga edastuskanal	sloveeni	67% – 87%			Etxeberria jt 2016
Reeglipõhine	põhjaeesti	71%			
Reeglipõhine	lõunaeesti	62%			
Sõnastikupõhine	põhjaeesti	98%			

4. Probleemid vanade eestikeelsete tekstide normaliseerimisel

Vanade eestikeelsete tekstide tänapäevasele kujule teisendamine hõlmab endas minu arvates mitmeid probleeme, mida edaspidi käsitlen 19. sajandist pärinevate Rahvusarhiivi andmebaasides olevate vallakohtuprotokollide näitel.

Üheks suurimaks probleemiks on väga suur varieeruvus, mis on tingitud kirjutaja murdetaustast, kirjaoskusest, üldisest haritusest ja üleminekust vanalt kirjaviisilt uuele. Erisusi tekitab ka põhja- ja lõunaeesti keele omavaheline segunemine. Võib tekkida õigustatud küsimus, kui mõttekas ikkagi on lõunaeestikeelseid tekste põhjaeestikeelseteks normaliseerida, arvestades, et lõunaeesti keel oli esimene, mis ühtsest alglähemeresoome keelest lahknes, nii et ajalooliselt on tegemist eraldi keelega (vt Kallio 2014, Pajusalu 2020: 25 jj). Põhja- ja lõunaeesti keelte uuemad ühisjooned on levinud geograafilise läheduse tõttu, viimastel sajanditel ka ühtse eesti kirjakeele kaudu (vt Prillop 2020: 76, Laanekask 2004: 36–42). Vaatame näitelauset 1885. aastal Vastseliinas kirja pandud protokollist:

(2) Ette tulli Peter Härm ja kaibas et Hindrik Toll olewat temä naist pesnud selle perrast et temä hobbene Hindrik Tolli kara raugu päle lännü sis nõwwap 5 Rbla walu rahha ja olewat tunnistaja Johann Sippul.

Lõunaeesti keelele viitavad vokaalharmoonia sõnas *temä*, *lännü* ning häälikuühendi ks asendumine s-iga (*peksnud vs pesnud*). Põhjaeesti keelejoonena võib välja tuua mineviku *nud*-partitsiibi vormi. Ka on vokaalharmoonia üsna ebajärjekindel, esinedes sõnas *temä*, aga sõnas *perrast* see puudub. Seepärast ei oleks niisuguste segunenud tunnustega lõuna- ja põhjaeestikeelsete tekstide lahkulöömine eriti mõistlik. Teine, praktilisem põhjus nende ühtmoodi käsitlemiseks seisneb selles, et kui on vaja protokollidest mingit märksõna otsida, siis oleks hea leida üles nii põhja- kui ka lõunaeestikeelseid vasteid sisaldavad tekstid. Kusjuures nimetatud juhul ei olegi õige grammatilise info tuvastamine primaarne, kõige olulisem on kindlaks teha õige lemma.

Peale piirkondlike variantide tekitavad probleeme ka erisused kirjutaja kirjaoskuses ja erinevused vana ja uue kirjaviisi vahel. Üks ja sama sõna võib olla kirja pandud mitmel erineval moel ja mõnikord ka sama teksti sees. Näites (3) on Alatskivil 1878. aastal kirjapandud protokoll, kus üks nimi on kirjutatud mitmel erineval moel (*Veodor* ja *Feodor*).

(3) Veodor Kromonov, Kallaste küla venelane Alatskivi jao peal, oli kohtu ees tehtud terminid võla asjus mööda minna lasnud, ka mitmekordse kutsumise peale mitte ette tulnud, siis läksivad kohtuliikmed eilsel paeval tema maiasse, tema kraami ülesse kirjutama, resp riisumise teel seda võlga sissenõudma, seal oli aga Feodor Kromonov vastahakanud.

Kirjeldatud probleemid võivad treeningandmete valimise ja erinevate normaliseerimismeetodite rakendamise muuta üsna keeruliseks, kuna eelnevalt nimetatud erisuste tõttu tuleb eri murrete ja võib-olla ka ajaperioodide jaoks luua eraldi andmestikud.

Treeningandmetena saaks potentsiaalselt kasutada eesti vana kirjakeelt sisaldavaid ressursse, nagu Tartu ülikooli vana kirjakeele korpust (VAKK) ja piiblitekstide korpust (Piiblikorpus). Vana kirjakeele korpuses on poolautomaatselt märgendatud 15.–17. sajandi tekstid.

18. ja 19. sajandist on VAKK-i põhiossa võetud valik trükiseid. Märgendatud on 18. sajandist u 45 000 sõnet, 19. sajandi I poolest u 10 000 sõnet, 19. sajandi II poolest samuti u 10 000 sõnet.[5] Piiblikorpuses on vanematest piiblitest kõige hilisem 1739. aasta piibel. Need võivad abiks olla küll normaliseerimismeetodite treenimisel, aga kuna tihtipeale on osad varasemad tekstid, nagu näiteks vallakohtuprotokollid kirjapildilt palju mitmekesisemad ja kirja pandud hiljem, ei pruugi nendest olla nii palju kasu kui esialgu võib tunduda.

5. Arutlus

Järgnevas püüan hinnata, millist normaliseemismeetodit (vt ptk 2) võiks eelistada vanade eestikeelsete tekstide analüüsimisel. Arutlen võimalike probleemide üle, mis seonduvad nende rakendamisel.

[5] https://vakk.ut.ee/tekstid.php (20.1.2021).

Sõnastikul põhinev meetod (Piotrowski 2012: 74) sobib eestikeelsete tekstide puhul tõenäoliselt muutumatute sõnade, nagu määr-, kaas- ja sidesõnade normaliseerimiseks. Käänd- ja pöördsõnad oleks oluliselt problemaatilisemad, kuna sõnastikku tuleb lisada kõik või sagedamini kasutatavad vormid. Seega oleks selle meetodi rakendamine morfoloogiliselt rikkaliku keele puhul, nagu seda eesti keel on, üsna tülikas ja aeganõudev. Samas sai Pettersson näiteks ungari keele normaliseerimisel üsna hea tulemuse (75%). Enne normaliseerimist vastas tänapäevasele keelele vaid 17% sõnadest. Seda võib ilmselt seletada asjaoluga, et sageli esinevate sõnade õnnestunud normaliseerimine suurendab korrektsust üsna olulisel määral. Siiski oli mõõdik võrreldes teiste keeltega veidi väiksem. (Pettersson 2016: 82) Seega võiks eeldada, et sõnastikupõhine meetod toimib mingil määral ka eestikeelsete tekstide normaliseerimiseks, küsimus jääb vaid, kuidas tulla toime suurte variatsioonidega sõnade kirjakujus. Selle probleemi lahendamiseks võib proovida teisenduskaugust. Petterssonil see ungari keele puhul nii häid tulemusi ei andnud, aga kuna andmestikud on erinevad, võivad ka tulemused olla teistsugused.

Statistiline masintõlge on inglise, saksa, ungari, rootsi ja sloveeni keeles kirjutatud vanade tekstide normaliseerimisel üsna hästi toiminud (Scherrer, Erjavec 2013, Pettersson 2016: 82), nii et suure tõenäosusega võib see ka eestikeelsete tekstide peal anda häid tulemusi. Probleem võib tekkida sellega, et meetodi efektiivseks rakendamiseks on vaja üsna palju treeningandmeid. Petterssoni katsetes kasutati selleks vähemalt 50 000 sõnast koosnevat korpust. Kuigi vallakohtuprotokollides on artikli kirjutamise ajal kokku 2,5 mln sõna, on eelnevalt kirjeldatud varieeruvuste tõttu tõenäoline, et kui sellest treeningandmetena mingi osa käsitsi normaliseerida, võib sellest ikkagi väheks jääda.

Reeglipõhine lähenemine (Piotrowski 2012: 75–76) võib eesti keele normaliseerimisel samuti häid tulemusi anda, aga selle rakendamisel on oht, et reeglite hulk muutub üsna suureks ja raskesti hallatavaks. Siiski saab mõne universaalse seaduspärasuse selle meetodiga üsna kergesti ära kirjeldada. Näiteks topelt *g*, *b*, *d* asendamine ühega näib olevat üsna kindel moodus, kuna tänapäevases eesti keeles leidub selliseid tähekombinatsioone sisaldavaid sõnu üsna vähe ning peamiselt vaid liitsõnades (nt *kõrggooti*). Lisaks saab tundmatuks jäänud sõnades asendada *nu*-lõpu *nud*-ga. Ent ainult reeglitele ja eelnevalt kirjeldatud sõnastikupõhisele meetodile lootma jäädes on oht valetuvastuste tekkeks. Näiteks teatud reegleid rakendades saaks näites (4a) esitatud osalausest näites (4b) esitatud osalause.

(4a) need 2 tunnistajad Vana Nõost Peter Peets ja Juhan Enok ja Juhan Org ja Johan Maanus om kiik neljakeste kakkelnu ja üksteist pesnu.
(4b) need 2 tunnistajad Vana Nõost Peter Peets ja Juhan Enok ja Juhan Org ja Johan Maanus on kiik neljakesi kakelnud ja üksteist pesnud.

Selliste valeparanduste automaatne tuvastamine on aga üsna keeruline, kuna morfoloogiliselt on need sõnad igati korrektsed. Kui veel näiteks sõna *kiik* parandada sõnaks *kõik* ja *tunnistajad* sõnaks *tunnistajat*, siis saaks sellest süntaktiliselt ja semantiliselt korrektse, aga pragmaatiliselt ikkagi ebakorrektse lause. Ehk siis, kuigi inimene saab konteksti abil aru, et *pesnu* viitab siinkohal peksmisele, on selle selgitamine algoritmidele keeruline.

Näites (5) võib vastavate taustteadmisteta inimene pidada sõna *töisel* sõna *töine* vormiks, tegelikult aga on tegu sõna *teine* lõunaeestilise kujuga *töine*:

(5) ta näinud küll veel töisel hommikul, kui Jaan Naudi üht hobust kinni keitnud

Lisaks meetodite üksikult rakendamisele tasuks kindlasti proovida nende omavahelist ühendamist. Pettersson (2016) on omavahel kombineerinud vaid sõnastikupõhist meetodit ja Levenshteini teisenduskaugust. Minu arvates tasuks kindlasti katsetada ka muid kombinatsioone: masintõlke ja sõnastikupõhise meetodi ühendamisel saaks tõenäoliselt normaliseerida harvaesinevaid sõnu ning sõnastike ja reeglite ühendamisel saaks parandada tõlkimise käigus tekkinud vigu.

6. Kokkuvõte

Ajaloolised tekstid on tänapäeval huvipakkuv uurimisobjekt nii keeleteadlastele, ajaloolastele, perekonnaloo uurijaile kui ka digihumanitaaridele üldiselt. Selleks, et vanu kirjutisi saaks tänapäevase keele analüüsivahendite abil uurida, tuleb need normaliseerida ehk tänapäevasele kujule viia. Käesolevas artiklis käsitleti erinevaid normaliseerimismeetodeid ja seda, mis mujal maailmas selles valdkonnas nende abil tehtud on. Mitmed meetodid sobiksid hästi ka vanade eestikeelsete tekstide normaliseerimiseks, aga probleemiks on nii tekstides esinevad varieeruvused kirjaviisis ja murretes kui ka eesti keele morfoloogiline mitmekesisus.

Viidatud kirjandus

Beesley, Kenneth R.; Karttunen, Lauri 2003. Finite-State Morphology. Stanford: CSLI Publications.

Bollmann, Marcel; Petran, Florian; Dipper, Stefanie 2011. Rule-based normalization of historical texts. – Cristina Vertan, Milena Slavcheva, Petya Osenova, Stelios Piperidis (Eds.), Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage. Association for Computational Linguistics, 34–42. https://aclanthology.org/W11-4106/ (2.4.2021).

Brill, Eric; Moore, Robert C. 2000. An improved error model for noisy channel spelling correction. – ACL'00: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 286–293. https://doi.org/10.3115/1075218.1075255

Etxeberria, Izaskun; Alegria, Inaki; Uria, Larraitz; Hulden, Mans 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene. – Nicoletta Calzolari et al. (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association, 1064–1069. https://www.aclweb.org/anthology/L16-1169/ (2.4.2021).

Hämäläinen, Mika; Säily, Tanja; Rueter, Jack; Tiedemann, Jörg; Mäkelä, Eetu 2019. Revisiting NMT for normalization of Early English letters. – Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, Stan Szpakowicz (Eds.), Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Association for Computational Linguistics, 71–75. https://doi.org/10.18653/v1/W19-2509

Jurafsky, Daniel; Martin, James H. 2019. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/B.pdf (2.4.2021).

Kallio, Petri 2014. The diversification of Proto-Finnic. – Joonas Ahola Frog, Clive Tolley (Eds.), Fibula, Fabula, Fact: The Viking Age in Finland. Studia Fennica 18. Helsinki: Suomalaisen Kirjallisuuden Seura, 155–170.

Korchagina, Natalia 2017. Normalizing medieval German texts: From rules to deep learning. – Gerlof Bouma, Yvonne Adesam (Eds.), Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. Linköping University Electronic Press, 12–17. https://www.aclweb.org/anthology/W17-0504 (2.4.2021).

Laanekask, Heli 2004. Eesti kirjakeele kujunemine ja kujundamine 16.–19. sajandil ['Formation and Shaping of the Estonian Literary Language in the 16th–19th Centuries']. Dissertationes philologiae estonicae Universitatis Tartuensis 14. Tartu: Tartu Ülikooli Kirjastus.

Laur, Sven; Orasmaa, Siim; Särg, Dage; Tammo, Paul 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Nicoletta Calzolari et al. (Eds.), Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, 7152–7160. https://www.aclweb.org/anthology/2020.lrec-1.884 (2.4.2021).

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. – Soviet Physics Doklady, 10 (8), 707–710.

Pajusalu, Karl 2020. Eesti keel uurali ja läänemeresoome keelena ['Estonian as Uralic and Finnic language']. – Eesti keele ajalugu. Eesti keele varamu 6. Tartu: Tartu Ülikooli Kirjastus, 21–32.

Pettersson, Eva 2016. Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Studia Linguistica Upsaliensia 17. Uppsala: Uppsala University. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-269753 (2.4.2021).

Piiblikorpus = Eesti piiblitõlke ajalooline konkordants. http://www.eki.ee/piibel/index.php (2.4.2021).

Pilvik, Maarja-Liisa; Muischnek, Kadri; Jaanimäe, Gerth; Lindström, Liina; Lust, Kersti; Orasmaa, Siim; Türna, Tõnis 2019. Möistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine ['Creating a digital resource from 19th century communal court minute books']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 15, 139–158. https://doi.org/10.5128/ERYa15.08

Piotrowski, Michael 2012. Natural Language Processing for Historical Texts. Morgan & Claypool. 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.

Prillop, Külli 2020. Eesti keele ajalooline fonoloogia ['Historical phonology of the Estonian language']. – Eesti keele ajalugu. Eesti keele varamu 6. Tartu: Tartu Ülikooli Kirjastus, 74–168.

Scherrer, Yves; Erjavec, Tomasž 2013. Modernizing historical Slovene words with character-based SMT. – Jakub Piskorski, Lidia Pivovarova, Hristo Tanev, Roman Yangarber (Eds.), Proceedings of the 4th Biennial Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics, 58–62. https://aclanthology.org/W13-2409 (2.4.2021).

Tang, Gongbo; Cap, Fabienne; Pettersson, Eva; Nivre, Joakim 2018. An evaluation of neural machine translation models on historical spelling normalization. – Emily M. Bender, Leon Derczynski, Pierre Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, 1320–1331. https://aclanthology.org/C18-1112 (2.4.2021).

Tuldava, Juhan 1977. Sagedussõnastik leksikostatistilise uurimise objektina ['The frequency dictionary as an object of lexicostatistical investigation']. – Töid keelestatistika alalt II. TRÜ toimetised 413, 141–171.

VAKK = Tartu Ülikooli vana kirjakeele korpus. https://vakk.ut.ee (2.4.2021).

van der Zwaan, Janneke M.; Leemans, Inger; Kuijpers, Erika; Maks, Isa 2015. HEEM, a complex model for mining emotions in historical text. – IEEE 11th International Conference on e-Science. IEEE, 22–30. https://doi.org/10.1109/eScience.2015.18

NORMALIZING HISTORICAL TEXTS

Gerth Jaanimäe
University of Tartu

Normalizing historical texts or in other words converting them to modern spelling enables us to analyze them with tools designed for contemporary language. It also makes it possible to search the texts for different keywords and automatically compare the old spelling to contemporary spelling. This article gives a general overview of normalizing, different methods, previously performed experiments and the main problems in the context of the old Estonian texts from the second half of the 19th century.

Keywords: NLP, normalizing, language history, corpus linguistics, computational linguistics, language change, non-standard language, digital humanities, Estonian

Gerth Jaanimäe on Tartu Ülikooli eesti ja soome-ugri keeleteaduse doktorant arvutilingvistika suunal. Uurimisvaldkonnad: korpuslingvistika, loomuliku keele töötlus, ajaloolised tekstid, vallakohtuprotokollid, keele ajalugu, keele muutumine.
Jakobi 2, IV korrus, 51005 Tartu, Estonia
gerth.jaanimae@ut.ee