Keel ja Kirjandus

8–9
2021

DIGIHUMANITAARIA
EESTIS

Peeter Tinits, Jane Klavan, Liina Lindström, Indrek Ibrus, Maximilian Schich, Marek Tamm, Artjoms Šeļa, Maarja Helena Meriste, Peeter Torop, Maarja Ojamaa, Aleksandr Fadeev, Alexandra Milyakina, Merit Rickberg, Margit Langemets, Kristina Koppel, Jelena Kallas, Arvi Tavast, Maarja-Liisa Pilvik, Helen Plado, Fred Puss, Ulrike Plath, Priit Raudkivi, Kaarel Vanamölder, Krister Kruusmaa, Anna Helena Liiv

KOLLEEGIUM
Epp Annus, Martin Ehala, Cornelius Hasselblatt, Tiit Hennoste, Eda Kalmre, Johanna Laakso, Helle Metslang, Sirje Olesk, Karl Pajusalu, Peeter Päll, Raimo Raag, Jaan Undusk, Ülo Valk, Mart Velsker, Märt Väljataga, Haldur Õim

TOIMETUS
Peatoimetaja **Johanna Ross**, johanna@keeljakirjandus.ee
Keeleteadus **Maria-Maren Linkgreim**, maren@keeljakirjandus.ee
Kirjandusteadus **Brita Melts**, brita@keeljakirjandus.ee
Kultuurilugu, folkloristika **Vivian Siirman**, vivian@keeljakirjandus.ee
Tegevtoimetaja **Tiina Hallik**, tiina@keeljakirjandus.ee
Tehniline toimetaja **Astrid Värv**, astrid@keeljakirjandus.ee

Toimetuse
aadress: **Kohtu 6, 10130 Tallinn**
telefon: **6449 126** (üldnumber), **6449 228** (peatoimetaja)
e-post: **kk@keeljakirjandus.ee**

www.keeljakirjandus.ee

Keelt ja Kirjandust on võimalik tellida
Express Posti kaudu telefonil 6177 717
või võrguaadressil **https://tellimine.ee**.
Värsket üksiknumbrit saab posti teel osta
Lehepunkti kaudu võrguaadressil **https://sirvi.eu**.

Ajakiri kuulub Euroopa Teadusfondi ja Norra Teadusandmete Keskuse registrisse ERIH PLUS, andmebaasidesse EBSCO, Linguistics & Language Behavior Abstracts (LLBA, ProQuest), MLA International Bibliography (Modern Language Association), Scopus (Elsevier) ja Ulrich's Periodicals Directory, digitaalarhiividesse CEEOL (Central and Eastern European Online Library) ja DIGAR (Eesti Rahvusraamatukogu) ning Alalise Rahvusvahelise Lingvistide Komitee valikbibliograafiasse „Linguistic Bibliography".

Ajakiri ilmub Eesti riigi ja Eesti Kultuurkapitali toetusel.
Trükk: Tallinna Raamatutrükikoda, Laki 26, 12915 Tallinn.
Kujundanud Kalle Toompere

© Keel ja Kirjandus®, SA Kultuurileht

ISSN 0131-1441
e-ISSN 2346-6014
DOI 10.54013/kk1958

Keel ja Kirjandus

EESTI TEADUSTE AKADEEMIA JA EESTI KIRJANIKE LIIDU AJAKIRI

LXIV AASTAKÄIK

8–9
2021

SISUKORD

PEETER TINITS, JANE KLAVAN, LIINA LINDSTRÖM. **Digihumanitaaria Eestis** 667

ARTIKLID

INDREK IBRUS, MAXIMILIAN SCHICH, MAREK TAMM. **Digihumanitaariast kultuuriandmete analüüsini** 678

ARTJOMS ŠEĻA. **Erinevused, kaugused ja sõrmejäljed. Stilomeetria ja mitmemõõtmelise tekstianalüüsi alused** 696

MAARJA HELENA MERISTE. **Kirjanduskeele kauglugemine. Kirjandust väärindava metodoloogia arendamine tsiteerimisvõtte näitel** 719

MAARJA OJAMAA, PEETER TOROP, ALEKSANDR FADEEV, ALEXANDRA MILYAKINA, MERIT RICKBERG. **Digitaalne lugemine ja humanitaarharidus** 737

MARGIT LANGEMETS, KRISTINA KOPPEL, JELENA KALLAS, ARVI TAVAST. **Sõnastikukogust keeleportaaliks** 755

MAARJA-LIISA PILVIK, HELEN PLADO, LIINA LINDSTRÖM. **Murded, varieerumine ja korpusandmed. Eitussõna paiknemine võru ja seto eituslausetes** 771

FRED PUSS. **Kakasonist Kaljulaks. Perekonnanimede eestistamise andmebaasi analüüs** 797

ULRIKE PLATH, PRIIT RAUDKIVI, KAAREL VANAMÖLDER, KRISTER KRUUSMAA, ANNA HELENA LIIV. **Kuidas kodeerida kliimat? Eesti ajaloolise kliimauurimise digitaalsest pöördest** 819

Koostajad **PEETER TINITS, JANE KLAVAN** ja **LIINA LINDSTRÖM**

CONTENT

PEETER TINITS, JANE KLAVAN, LIINA LINDSTRÖM. **Digital humanities in Estonia** .. 667

ARTICLES

INDREK IBRUS, MAXIMILIAN SCHICH, MAREK TAMM. **From digital humanities to cultural data analytics** 678

ARTJOMS ŠEĻA. **Differences, distances and fingerprints: The fundamentals of stylometry and multivariate text analysis** 696

MAARJA HELENA MERISTE. **Distant reading of literary language: Developing a methodology using quotations as an example** 719

MAARJA OJAMAA, PEETER TOROP, ALEKSANDR FADEEV, ALEXANDRA MILYAKINA, MERIT RICKBERG. **Digital reading and humanities education** 737

MARGIT LANGEMETS, KRISTINA KOPPEL, JELENA KALLAS, ARVI TAVAST. **From a collection of dictionaries to a language portal** 755

MAARJA-LIISA PILVIK, HELEN PLADO, LIINA LINDSTRÖM. **Dialects, variation and corpus data: Position of the negation word in Võru and Seto** 771

FRED PUSS. **Changing Kakason to Kaljula: Analysing database of Estonianizing surnames** 797

ULRIKE PLATH, PRIIT RAUDKIVI, KAAREL VANAMÖLDER, KRISTER KRUUSMAA, ANNA HELENA LIIV. **Kann man das Klima kodieren? Zur digitalen Wende in der historischen Klimaforschung in Estland** 819

Guest editors **PEETER TINITS, JANE KLAVAN** and **LIINA LINDSTRÖM**

Euroopa Liit
Euroopa Regionaalarengu Fond

Eesti tuleviku heaks

CEES Centre of Excellence in Estonian Studies

KEELED:

ingl = inglise; **lad** = ladina; **sks** = saksa; **sm** = soome; **vn** = vene.

https://doi.org/10.54013/kk764a1

Digihumanitaaria Eestis

PEETER TINITS, JANE KLAVAN, LIINA LINDSTRÖM

Keele ja Kirjanduse teemanumbri mõte tekkis seoses digikultuuri aasta toimumisega 2020. aastal. Numbri toimetajad on eest vedanud erinevaid digihumanitaaria-alaseid tegevusi Eestis ning pidevalt pidanud vastama küsimustele: mis on digihumanitaaria ja miks seda vaja on? Teemanumbri eesmärk ongi tutvustada eestikeelsele lugejaskonnale, mis on digihumanitaaria, milline on digihumanitaaria hetkeseis Eestis ja mida selle raames tehakse, ning püüda selle kaudu vastata, mida annab *digi* juurde humanitaariale. Siinsed artiklid kajastavad üsna hästi Eesti digihumanitaaria üldpilti: leidub mitmeid häid näiteid digihumanitaaria meetodite kasutamisest eri distsipliinide sees ja autorite seas on nii üksikuurijaid kui ka töörühmi, kes valdkonna arendamisega tegelevad. Loodetavasti annab käesolev artiklivalik parema arusaama digihumanitaaria meetodite kasutusvõimalustest.

Teemanumbrisse oleme koondanud mitmekesise valiku artikleid keeleteadusest kliimaajalooni. Tutvustatakse mõne eriala arengut digitaliseeruvas maailmas (Ibrus jt; Langemets jt; Ojamaa jt; Plath jt), teisal on fookuses konkreetse uurimisküsimuse lahendamine mõne digihumanitaaria meetodiga ja selle meetodi üldisem tutvustus (Meriste; Pilvik jt; Puss; Šeļa). Tegemist ei ole siiski esimese digihumanitaariale pühendatud väljaandega Eestis: 2020. aasta lõpus ilmus ajakirja Methis teemanumber, mis põhines 2019. aasta Eesti digihumanitaaria konverentsi ettekannetel. Koos võiksid need eesti lugejale pakkuda mitmekülgset teadmist digihumanitaaria võimalustest koos kohalike näidetega.

Mis on digihumanitaaria?

Kuigi *digihumanitaaria* tundub võrdlemisi uus termin, on sel tegelikult pikk ja auväärne ajalugu: see viiakse tagasi esimeste arvutite loomiseni 1950. aastatel või lausa arvutuslike meetodite kasutuselevõtuni XIX sajandi lõpul. Paljud kvantitatiivsed meetodid humanitaarias on välja töötatud esmalt just paberi ja pliiatsiga arvutamiseks, moodne arvutusvõimsus võimaldab lihtsalt seda teha tõhusamalt (vt nt Šeļa siin numbris). Juba 1966. aastal hakkas ilmuma pikaaegne ajakiri Computers and the Humanities (kuni aastani 2004); 1984 alustas teine mõjukas ajakiri Literary and Linguistic Computing, mis kannab praegu nime Digital Scholarship in the Humanities. Seda vedav Euroopa Digihumanitaaria Assotsiatsioon sai alguse, ehkki teise nime all, 1973. aastal.

Märksõna *digihumanitaaria* tekkis 2000. aastate algul distsipliinide vastastikuses defineerimises, omamoodi värskenduskuurina ja programmilise liikumisena. Digirevolutsiooni teema kerkis ühiskonnas laiemalt esile ning ilmselt seetõttu avaldas

termin mõju ka digihuvilistele humanitaarteadlastele. Algusest peale on seda kasutanud ja defineerinud väga erineva taustaga uurijad, kes on sellele ka erinevaid lootusi pannud. Definitsioone on nii palju, et need on kogutud veebileheküljele, mis sisaldab 817 digihumanitaaria definitsiooni aastatest 2009–2014 (vt What Is Digital Humanities?). Defineerimise käigus on kirjutatud ideaalpilti visandavaid manifeste (nt Schnapp, Presner 2009), aga teistele uurijatele on see lihtsalt pragmaatiline märksõna, mille abil leida ühisosa teiste valdkondadega või selgitada rahastajatele töö eesmärke (nt Kirschenbaum 2012). Pikemalt on digihumanitaaria defineerimisest, selle valdkonna kasvuraskustest ja enesemääratlusest juttu teemanumbri esimeses artiklis (Ibrus jt).

Kui püüame leida digihumanitaaria definitsioonides ühisosa, siis enamasti on selleks digitehnoloogia rakendamine, mis eeldab infotehnoloogiaalaste teadmiste-oskuste olemasolu. Digitehnoloogia pakub uuenduslikku võimalust töödelda ja analüüsida suures koguses andmeid – seepärast hõlmab suurem osa digihumanitaaria all tehtavast teadustööst mingil viisil suurte andmehulkade töötlust, eri tüüpi andmete kombineerimist ning väga mitmekesiseid andmete analüüsi ja esitamise viise. Seega võiks digihumanitaaria tuumosa nimetada ka andmepõhiseks humanitaariaks (vt Laak jt 2020). Loomulikult on muidki lähenemisi, mis rakendavad digivahendeid pigem kvalitatiivses analüüsis või mõtestavad laiemalt digikultuuri arengut ja näiteks õppimist digiajastul (vt Ojamaa jt siin numbris).

Digihumanitaaria defineerimine või pigem defineerimatus on toonud kindlasti palju positiivset. Rõhutatud on digihumanitaaria suunatust ka väljapoole akadeemilist maailma, koostöö, avatuse ja mittehierarhiliste suhete väärtustamist ning teadlaste hulgas väljakujunenud normide kritiseerimist (nt retsenseerimis- ja publitseerimispraktikad) (Kirschenbaum 2010: 59; Bode, Arthur 2014: 2). Digihumanitaaria kui valdkonnaga kaasnevate metaküsimuste hulka, mis ei ole otseselt seotud teadusliku uurimistöö sisulise poolega, kuuluvad näiteks andmekaitse ja avatud teadus (vt nt Berry, Fagerjord 2017: 114–135; Gardiner, Musto 2015: 146–165). Need on olulised teemad, mis väärivad pikemat käsitlust, ning just digihumanitaaria andmepõhisus on aidanud algatada vastava arutelu. Kindlasti propageerib digihumanitaaria valjuhäälselt avatud ligipääsuga teadust, nii publitseerimisel kui ka andmete ja andmeanalüüsi skriptide kättesaadavaks tegemisel. Artikli aluseks olevad andmed ja skriptid on teinud andmerepositooriumis kättesaadavaks nt Pilvik jt. Avatud teaduse printsiip püüab tuua teadust lähemale ideaalile, kus kõik uurimistöö sammud on läbipaistvalt teostatud ja korratavad. Avatud teaduse põhimõtete ja vahenditega on võimalik teadustööd teha mitmel viisil paremini korratavaks: näiteks eelregistreeritud uuringute, avatud laborimärkmike, avatud lähtekoodi või avaandmete kaudu. Avaandmete loomine teadustöös või ka näiteks riigiametites on saamas peaaegu normiks, ilma milleta on raske analüüse tõsiselt võtta. Just digitaalsed vahendid teevad selle võimalikuks.

Eestisse ilmus digihumanitaaria (ka digitaalhumanitaaria) temaatiliste konverentside kaudu, mille algatasid Eesti Kirjandusmuuseumi uurijad aastal 2013.[1] Siiski

[1] Konverentside sisu on leitav Eesti Digitaalhumanitaaria Seltsi lehelt (vt DH Estonia).

on ka Eesti digihumanitaarial pikem ajalugu. Arvutitele toetuvast uurimistööst meenub näiteks formaalset grammatikat luua püüdnud Generatiivse Grammatika Grupp (1965–1972), mis viis välja keelekorpuste koostamiseni 1980. aastate lõpus ja tänapäeva arvutuslingvistikani; juba 1960-ndatel tehti Juhan Kahki eestvõttel katsetusi arvutite kasutamiseks ajalooteaduses (vt Plath jt siin numbris) jne. Erinevate valdkondade andmebaaside ja kogude loomise või analüüsiga tõusid esile mitmed uurijad, näiteks Aadu Must ajalooteaduses (vt ka Puss siin numbris) või Arvo Krikmann folkloristikas, keda võib õigusega pidada tänapäeva digihumanitaaria pioneerideks Eestis.

Eesti digihumanitaaria kogukonna hetkeseisust saab hea ülevaate konverentsikavade esinejaid ja teemasid vaadates. Digihumanitaaria Eestis hõlmab üpris erinevaid uurijaid, keda ühendab humanitaarteaduste taust või uurimistemaatika. Juba esimesel konverentsil olid esindatud dialektoloogia, folkloristika, arheoloogiamälestised, tekstitöötlus, geoinformaatika, interaktiivsed keskkonnad ja andmebaaside loomine. Märksõna *digihumanitaaria* on aidanud end seostada rahvusvaheliste traditsioonidega ning leida eeskujusid ja kaasamõtlejaid teistelt uurimisväljadelt. Võib ilmselt väita, et kui esimestel konverentsidel jäid kõlama andmebaaside ja kogude ülesehitamisega seotud küsimused, siis üha enam on laienenud teadlaste ring, kes digihumanitaaria meetodeid teadustöös rakendavad.

Andmepõhine humanitaaria

Igasugune teadustöö toetub üldiselt andmetele ehk millelegi antule (lad *data* 'asjad, mis on antud'). Humanitaarteadustes ei ole tihti seda mõistet tähtsustatud, kuna kasutusel on juba olnud üksikasjalikumad terminid, näiteks *primaar- ja sekundaarallikad* ajalooteaduses, samuti ei ole selge, kuivõrd vaadeldav on „antud" või on see pigem vaatleja loodud (vt Harrower jt 2020). Andmeteks saab miski vaadeldav, kui seda kasutatakse analüüsi alusena (olgu selleks fototõend, tekstikogu või lohk rändrahnus). Digitaalsed vahendid võimaldavad asuda seda vaadeldavat koondama suuremateks organiseeritud andmekogudeks, mis säilitaks vajaliku info ja mida oleks kerge kasutada edasistes uuringutes. Andmepõhises humanitaarias saab töö andmestikega – nende loomine, analüüs ja kriitika – loomuldasa uurimise põhiosaks.

Andmete hulk, mis on hästi ligipääsetav just digitaalselt, kasvab aina kiiremas tempos. Igas minutis kirjutatakse maailmas 500 000 avalikku lühikommentaari, saadetakse 16 miljonit tekstisõnumit, 156 miljonit e-kirja. Suhtlus on kolinud digitaalsesse kanalisse, trükimeedia ja raamatud valmivad esmalt digitaalsel kujul. Kogu seda materjali on võimalik traditsiooniliselt lugeda või luua sellest väiksemaid ja suuremaid andmekogusid, kus võiks saada täielikuma ülevaate kirjutatust. Suhteliselt hea pildi eesti kultuuriajaloost võiks anda näiteks Eesti Rahvusringhäälingu arhiiv, kultuurilehtede arhiivid või ka Keeles ja Kirjanduses ilmunud artiklid.

Viimased paarkümmend aastat on aktiivselt tegeletud maailma kultuuripärandi ülespildistamise ja digiteerimisega. Praeguseks on mõningate hinnangute

järgi ilmunud maailmas üle 130 miljoni raamatu (vt Taycher 2010), millest näiteks Google Books on üles pildistanud üle 40 miljoni nimetuse (vt Lee 2019). Niisamuti kui mujalgi tegeletakse Eestis aktiivselt digiteerimisega: trükipärandist on digiteeritud rohkem kui 25%, alustades tähtsamatest, kanoonilisematest autoritest ja allikatest. Pildistatud allikates tuvastatakse tekst, klassifitseeritakse artikliteks ja tekstiosadeks, lisatakse metainformatsioon kus võimalik: ilmumise aeg, autor, väljaanne jne. Selle tulemusena on võimalik koondada suur hulk tekste ülevaatlike uurimuste tarbeks. Kõik digiteeritud eksemplarid ja kogud ei ole küll kergesti kättesaadavad, aga nende ligipääsetavuse nimel töötavad mitmed asutused. Näiteks Rahvusraamatukogu perioodikakogust on võimalik kohaliku ajakirjanduse vahendusel saada hea pildi siinse ühiskonna arengutest aastatel 1860–1940.[2] Suured kogud on Eesti Kirjandusmuuseumil (sh Eesti Rahvaluule Arhiivil), Rahvusarhiivil jt mäluasutustel. Massdigiteerimine ning kultuuripärandi laiem kättesaadavus on ka Kultuuriministeeriumi „Kultuuripärandi digiteerimise tegevuskava" eesmärk. Kui kogude kättesaadavus pole enam probleemiks, on probleemiks siiski see, millisel viisil kogud on kättesaadavad: kas tekstiandmed on tärktuvastatud (OCR, *Optical Character Recognition*), kas tärktuvastus on kvaliteetne, millises formaadis ja kodeeringus on tekstiandmed esitatud ning millised on piirangud nende kasutamisel. Uurija seisukohast on suurim väljakutse oskus varieeruva ja mitmekesise andmestikuga efektiivselt toime tulla, kasutades erinevaid digimeetodeid ja -vahendeid. Võib üsna julgelt öelda, et mõlemal poolel – nii mäluasutustel kui ka teadlastel – on selles osas veel väga palju õppida.

Kui massdigiteerimine võimaldab andmeid üles pildistada ning tärktuvastusega on võimalik trükitud teksti tuvastada, siis keerukam on käsikirjaliste tekstidega, mida arhiivides on väga palju. Masinõppe meetoditel põhinev käekirjatuvastus, mida digihumanitaaria algatuste käigus on loodud (HTR, *Handwritten Text Recognition*; vt Transkribus), ei ole leidnud laialdaselt teed arhiividesse. Rahvusarhiiv on kasutanud sel eesmärgil üsna edukalt vabatahtlike abi ühisloomeks (ingl *crowdsourcing*) – arhiiv on loonud platvormi, kus vabatahtlike töö ja taibu abil kirjutatakse arhiivimaterjalid süstemaatilisel kujul masintöödeldavaks teksktiks (Rahvusarhiivi ühisloomealgatused, nagu eestlased Esimeses maailmasõjas, Vabadussõjas osalenud, Tartu 1867, tsaariaegsete vallakohtuprotokollide sisestamine; vt Rahvusarhiiv). Ühisloome korras on vanadele fotodele geoinfo märkide lisamisega tegelenud Ajapaik.

Kui suurimad kogud on tekstidest, siis oma filmiandmestikku on üles ehitanud Eesti Filmiarhiiv, teoste ülespildistamisega tegelevad kunstimuuseumid ja arheoloogid – oluline osa on seal ka esemete või hoonete 3D-mudelite loomisel fotode või laserskannerite abil. Näiteks Rahvusraamatukogu digitaalarhiivis on üle 16 000 plakati, 4000 kaardi, 7000 postkaardi (vt Digar). Rahvusarhiivi fotoinfosüsteemi on kantud üle 876 000 kirje, millest rohkem kui 714 000 fotol on olemas ka digipilt (vt Fotis). Digiteeritud esemeid on Eesti Rahva Muuseumis, Eesti Kirjandusmuuseumis, Tallinna Ülikooli Akadeemilises Raamatukogus jm. Fotod, maalid, arheoloogilised

[2] Ülevaadet selle kohta vt http://data.digar.ee/

leiud on koondatud muuseumide infosüsteemi (vt Eesti muuseumide veebivärav), kus need on seotud hulga metainfoga. Ka nende andmetega ringikäimine nõuab spetsiifilisi teadmisi ja oskusi, mis traditsiooniliselt ei ole olnud humanitaarerialade hariduse osa.

Digiteerimise võidukäik kultuuripärandiga ümberkäimisel on toimunud. Arusaadavalt alustati materjalide säilitamisest ja arhiveerimisest ning heal juhul ka andmebaasi moodustamisest. Alles viimastel aastatel on rõhutatud ka selle digiteeritud ja digitaalselt sündinud kultuuripärandi kasutuselevõtu vajadust teadustöös ning loodud selleks ka võimalusi.[3] Just siin pakub digihumanitaaria vahendeid ja oskusi materjale digimeetoditega uurida ning uurijakogukonda, kes mõistab sellise tööga seotud võimalusi, aga ka väljakutseid, mida tuleb asutustevahelises koostöös ületada.

Andmetega töötamiseks on vaja vahendeid ja meetodeid

Erinevat päritolu ning eri „valmidusastmega" andmete töötlemine ja analüüs nõuavad oskusi. Digihumanitaaria arengu käigus on tekkinud väga palju vahendeid, mis enamasti on mõeldud üht tüüpi andmete või ühe andmekogu töötlemiseks ja analüüsiks (neid on vahel püütud ka loetleda, vt nt Gardiner, Musto 2015). Eri tüüpi andmetega toimetulekuks on vaja vähemalt minimaalset programmeerimisoskust. See on kahtlemata praegu digihumanitaaria kõige suurem väljakutse: humanitaarteaduste õpetamistraditsiooni ei ole seni kuulunud programmeerimis- või kvantitatiivsed meetodid. Sellest tulenevalt on käesoleva teemanumbri üheks eesmärgiks lugejatele konkreetsete meetodite tutvustamine, millele on pandud tavapärasest rohkem rõhku (vt nt Pilvik jt ja Šeļa siin numbris).

Digimeetodid võimaldavad andmetepõhist makrotasandi vaadet, nt nn kauglugemise uurimustes (vt Meriste siin numbris). Tasub meeles pidada, et head kvantitatiivset uurimust ei ole lihtsam teha kui traditsioonilisi humanitaaria kvalitatiivseid (lähilugemis)meetodeid kasutavat uurimust. Arvutuslikud meetodid võimaldavad saavutada suuremat täpsust ning kaasata suuremat andmehulka, aga ainult juhul kui meetodite toimimist lähedalt kontrollitakse ning andmestiku iseärasustega arvestatakse. Näiteks tekstide automaattöötluse vahendid on enamasti loodud tänapäeva keelele ja tekstižanridele. Kui nad võimaldavadki suhteliselt hästi kätte saada filmiarvustuse tonaalsust,[4] siis ei ole üldse kindel, et nad sama hästi töötavad XIX sajandi realistliku romaani peatükkide peal. Veelgi suuremat peavalu võib valmistada tänapäeva keele jaoks tehtud vahendite rakendamine XIX sajandi vallakohtuprotokollidele, mis on vajalik Rahvusarhiivi ühisloome käigus täiendatava kollektsiooni analüüsivahendite loomiseks (vt katsetust Pilvik jt 2019). Hea uurimus peab seega kaasama korraga humanitaaria-alast ja tehnilist kompetentsi ning kindlustama, et valitud meetodid ja vahendid oleksid adekvaatsed ning andmetega sobivad. Nõnda on digihumanitaaria-alane uurimistöö juba olemuslikult valdkondadevaheline ning

[3] Näiteks riiklik programm Eesti keel ja kultuur digiajastul 2019–2027.

[4] Vt nt Sentiment Analysis on IMDb, kus võrreldakse eri meetodeid. Kakskümmend parimat meetodit on üle 90% täpsusega.

vajab toimimiseks nii kvantitatiivseid kui ka kvalitatiivseid meetodeid, rakendagu neid üksainus mitmekülgselt pädev uurija või suurem uurimisrühm (Tolonen jt 2020). Kas sellist uurimistööd tuleks teha just digihumanitaaria katusnimetuse all, on eraldi küsimus (vt Ibrus jt siin numbris).

Kui tehnilised valdkonnad püüavad pigem luua aina paremaid algoritme, siis digihumanitaaria eesmärk on leida nende abiga aina põnevamaid uurimistulemusi, mis kannataksid ka igakülgset kriitikat (Nguyen jt 2020). Seetõttu võiks digihumanitaariat pidada ka autonoomseks uurimisväljaks: naaberdistsipliinid ei tee digihumanitaaride eest tööd ära ei algoritmide rakendamisel, ülesannete püstitamisel ega tulemuste kriitilisel tõlgendamisel. Tekstide lugemine toimub üha enam digitaalsete abivahenditega ning andmed tunduvad mängivat ühiskonnas üha kesksemat rolli. Raivo Ruusalepp (2020) näiteks arvab, et kultuuripärandi andmete puhul võiks tulevikus, kui digioskused muutuvad veelgi tavalisemaks, arvestada juba kodanikanalüütikutega. Humanitaarteadlastel on siinkohal võimalik valida, kui palju nad sellistes arengutes kaasa tahavad rääkida.

Digimeetodite integreerimine humanitaarteadlaste ettevalmistusse on väljakutse ennekõike ülikoolidele. Õppimisvõimalused humanitaarerialade üliõpilaste jaoks on loodud Tartu Ülikoolis, Tallinna Ülikoolis, Tallinna Tehnikaülikoolis; alguse on saanud mõned uurimisrühmad, kus digihumanitaaria-alane kompetents on keskne (nt CUDAN Tallinna Ülikoolis). Neli aastat on korraldatud kraadiõppuritele suvekoole „Digitaalsed meetodid humanitaarias ja sotsiaalteadustes", kus umbes sada osalejat aastas on omandanud uusi digioskusi. Konkreetsete distsipliinide sees – näiteks korpuslingvistikas või arvutuslikus arheoloogias – on need meetodid olnud pikemat aega kasutusel, peamiselt tingituna uurimisrühmade või üksikuurijate huvist ja vajadusest. Teema on akuutne ka sotsiaalteadustes ning meetodite ühisosa humanitaarja sotsiaalteadustes on väga suur. Seetõttu sobib õpik „Kuidas mõista andmestunud maailma? Metodoloogiline teejuht" (Masso jt 2020) suurepäraselt ka humanitaarteadlastele. Tuleb küll rõhutada, et edukaks tööks digihumanitaaria vallas ei piisa esmasest tutvumisest mõne meetodiga; süstemaatiliselt on vaja arendada programmeerimis- jm tehnilisi oskusi. Teisalt võib sõltuvalt eesmärgist mõne uue meetodi või vahendi tundmine oluliselt aidata edasi andmete kogumisel, töötlemisel, analüüsil või tulemuste visualiseerimisel. Ka käesolevas ajakirjas näeme üsna erinevaid lähenemisi andmete töötlemisel.

Eesti digihumanitaaria hetkeseisu võib ehk veel kirjeldada kui ärksat algusaega. On näha esimesi märke digioskuste süstemaatilisest kaasamisest ka humanitaarvaldkonnas ning edukaid üksikuurimusi sel teemal. Arhiivkogud ja muuseumid on võtnud suuna andmekogude ligipääsu parandamisele ning uurijad omandavad häid avatud teaduse praktikaid. Näha on alustavaid valdkondadevahelisi uurimisrühmi, mis püüavad keeruliste humanitaaria küsimustega tegeleda, kaasates korraga nii digioskusi kui ka oma kitsama valdkonna traditsioone (nt Plath jt; Ibrus jt; Ojamaa jt siin numbris). Selles, kas seda tööd teha just digihumanitaaria nime all, oleme teemanumbri koostajatena optimistlikumad kui Ibrus jt siin numbris ja leiame, et see märksõna alles kinnitab Eestis kanda ning me vajame seda kasvuraskuste ületamiseks. Ent ükskõik mis sildi all, on kõige tähtsam sisuline töö. Lootus ja võimalus

on, et lähematel aastatel on digioskused humanitaarias nii tavalised, et ei pea enam *digi*-t eraldi esile tooma. Meile endale tundub, et nende oskuste juurutamisest on humanitaarial palju võita.

Ülevaade artiklitest

Teemanumbriga soovime eesti lugejani tuua ülevaateid digihumanitaaria rakendusvõimalustest eri distsipliinides ning näitlikustada neid võimalusi juhtumiuuringutega. Digihumanitaaria hetkeseisu Eestis iseloomustab seegi, et meie üleskutse peale esitati neli korda rohkem teese, kui number mahutab, mistõttu saime esile tuua ainult mõned Eestis viljeletud teemad. Valikutes lähtusime peamiselt sellest, et Keele ja Kirjanduse põhiteemad oleks esindatud, aga samuti tuleks esile valdkonna mitmekesisus.

Kogumikku alustab **Indrek Ibruse**, **Maximilian Schichi** ja **Marek Tamme** kaastöö „Digihumanitaariast kultuuriandmete analüüsini", mis käsitleb digihumanitaaria valdkonna üldist problemaatikat: mis ja miks ikkagi peaks olema digihumanitaaria ja kuidas selles viljeletud teemasid võiks tulevikus uurida. Autorid kutsuvad üles valdkonnaülesele koostööle ja tõstavad esile tärkava kultuuriandmete analüüsi valdkonna, mis võiks ühendada humanitaar- ja teisi teadlasi, kes kultuurivaldkonnaga tegelevad, ning aidata kõige paremini erinevaid meetodeid ühendada.

Kogumiku teises artiklis „Erinevused, kaugused ja sõrmejäljed. Stilomeetria ja mitmemõõtmelise tekstianalüüsi alused" annab **Artjoms Šeļa** ülevaate eesti keeles seni vähe tutvustatud valdkonnast – stilomeetriast, mis püüab kvantitatiivsete vahenditega mõista ja kirjeldada autorite stiile ning tuvastada stiilitunnuseid konkreetsetes tekstides. Autor teeb näidiskatse eestikeelses minikorpuses, tutvustades automaatse autorituvastuse põhimõtteid ja nüansse, millega peab selle juures arvestama ning kuidas mitmemõõtmelise tekstianalüüsi tulemusi lugeda. Lisaks stilomeetria arengu ajaloolisele ülevaatele näitab autor, millist kasu võib kirjandusteadlastel olla tekstide kvantitatiivsetest esitustest nii autorituvastusel kui ka üldiste kirjanduslike suhete (nt žanr, kronoloogiline muutus jne) modelleerimisel.

Kvantitatiivsete meetodite kasutamisest kirjandusteaduses räägib ka **Maarja Helena Meriste**, kes artiklis „Kirjanduskeele kauglugemine. Kirjandust väärindava metodoloogia arendamine tsiteerimisvõtte näitel" esitab kauglugemise vallas tehtud juhtumiuuringu avatud tsitaatide esinemisest valitud eesti ilukirjandusautorite tekstides. Kauglugemine on raamistik, mis ei vaatle kirjandust eraldiseisvate üksikteoste kogumina, vaid tekstuaalse väljana. Artiklis vaadeldakse avatud tsitaati kui teksti tähendusstruktuuri tugevalt mõjutavat stiilivõtet ning leitakse, et avatud tsitaat loob ilukirjandusteksti tähendusreljeefi, mille abil tuleb esile lai seostevõrgustik.

Digitaalse lugemise kui humanitaarteaduste keskse mõiste teemal jätkab teemanumbri neljas artikkel „Digitaalne lugemine ja humanitaarharidus", kus **Maarja Ojamaa, Peeter Torop, Aleksandr Fadeev, Alexandra Milyakina** ja **Merit Rickberg** arutlevad nii sõnaliste tekstide lugemise kui ka tekstide vaatamise ja kuulamise olulisuse üle tänapäeva humanitaarhariduses. Autorid rõhutavad, et üleminek

raamatukeskselt kultuurilt veebikesksele kultuurile on kaasa toonud tekstide ja tähenduste pihustumise. Artiklis tutvustatakse sidususe loomise ja tervikuks seostamise mehhanisme (nt arvestades teksti/platvormi väljenduslikkuse, piiritletuse ja struktureerituse parameetreid), mis on hariduslikust vaatepunktist eriti oluline. Peale teoreetiliste lähtekohtade tutvustamise kirjeldatakse konkreetseid näited digitaalsete vahendite kasutamisest õppematerjalide loomisel, mis toetavad temaatiliste tervikukujutelmade tekkimist.

Kaks järgmist artiklit esindavad keeleteaduse valdkonda, kuid on erineva rõhuasetusega. **Margit Langemets, Kristina Koppel, Jelena Kallas** ja **Arvi Tavast** mõtestavad artiklis „Sõnastikukogust keeleportaaliks" leksikoloogia kui valdkonna üldisi probleeme ja lahendusi. Autorid toovad välja, et digiajastuga on sõnaraamatutöös kaasnenud laiaulatuslikud muudatused, sh tehnilised (nt liikumine ühtse sõnastikusüsteemi ja andmebaasi poole, andmete linkimine, arenenud korpuspäringusüsteemid), leksikograafilised (sõnastike ühendamine ja koostamispõhimõtete ühtlustamine, töö infokihtidega, andmete dünaamilisus) ja töökorralduslikud (kollektiivsuse kasv, avatuse suurenemine). Peale valdkonna üldiste muutuste ja digiajastu väljakutsete kirjeldavad autorid EKI uusimat sõnakogu – EKI ühendsõnastikku – ja selle koostamisel rakendatud Ekilexi andmemudelit.

Teises keeleteaduslikus kaastöös „Murded, varieerumine ja korpusandmed. Eitussõna paiknemine võru ja seto eituslausetes" näitlikustavad **Maarja-Liisa Pilvik, Helen Plado** ja **Liina Lindström** ühe konkreetse uurimisküsimuse lahendamist digihumanitaaria meetodiga. Autorid kirjeldavad ja seletavad keeles esinevat varieerumist (ees- ja tagaeitus lõunaeesti seto ja võru keelevariantides), kasutades meetodeid, mida on võimalik edukalt rakendada ka teistes humanitaaria suundades, nt erinevate klassifitseerimisülesannete lahendamiseks. Artiklis antakse ülevaade mitmemõõtmelisest statistilisest analüüsist – meetod, mis mängib korpuspõhises kvantitatiivses dialektoloogias keskset rolli.

Kaks viimast kaastööd esindavad otsapidi ajaloo valdkonda. **Fred Puss** annab artiklis „Kakasonist Kaljulaks. Perekonnanimede eestistamise andmebaasi analüüs" ülevaate perekonnanimede eestistamise senisest uurimisest, kirjeldades muuhulgas probleeme nimemuutumiste andmestikega töötamisel. Autor rõhutab, et nimede eestistamise andmebaasi kasutusvõimalusi on palju ning olemasolevate naaberandmestike ning digihumanitaarsete oskuste ja võimaluste arenedes tuleb neid veelgi juurde. Kvantitatiivsete uurimismeetodite kasutamise üle ajalooteaduses arutlevad ka **Ulrike Plath, Priit Raudkivi, Kaarel Vanamölder, Krister Kruusmaa** ja **Anna Helena Liiv** artiklis „Kuidas kodeerida kliimat? Eesti ajaloolise kliimauurimise digitaalsest pöördest". Artiklis on juttu kliimaajaloo uurimisest ning sellest, millal, millistes tingimustes, kontekstis ja sihiseadega leidis aset (kliima)ajalooteaduses digitaalne pööre. Autorid asetavad eesti ajaloolaste ja kliimaajaloo uurimise laiemasse digiteerimise konteksti ja näitavad, kuidas on digiteeritud andmeid võimalik kasutada tormisuse mõõtmiseks XIX sajandil.

Artiklid annavad erinevaid sissevaateid digihumanitaaria hetkeolukorda Eestis, pakkudes nii üldisemaid arutlusi digiajastuga kaasnenud väljakutsetest eri valdkondades kui ka konkreetsete uurimisküsimuste lahendusi digihumanitaaria

meetoditega. Kogumikku kaasatud artiklid ei esinda kaugeltki kogu pilti digihumanitaariast Eestis, kuid loodetavasti on meil õnnestunud digihumanitaaria mõiste paremini avada ja panna lugeja kaasa mõtlema sellisel väljakutsuval teemal nagu mida annab *digi* juurde humanitaariale ja vastupidi. Jääme ootama, et tehtav uurimistöö annaks peagi põhjust veel mõneks temaatiliseks kogumikuks.

Artikli valmimist ning teemanumbri koostamist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus).

VEEBIVARAD

Ajapaik. https://ajapaik.ee
DH Estonia. Estonian Society for Digital Humanities. https://dh.org.ee
Digar. Pildimaterjal. https://www.digar.ee/arhiiv/et/pildimaterjal
Eesti keel ja kultuur digiajastul 2019–2027. https://www.hm.ee/sites/default/files/eesti_keel_ja_kultuur_digiajastul_2019-2027.pdf
Eesti muuseumide veebivärav. https://www.muis.ee
Eesti Rahvusraamatukogu avaandmete portaal. http://data.digar.ee
Fotis. Rahvusarhiivi fotoinfosüsteem. https://www.ra.ee/fotis
Kultuuripärandi digiteerimise tegevuskava. https://www.kul.ee/kultuurivaartused-ja-digitaalne-kultuuriparand/digitaalne-kultuuriparand/kultuuriparandi
Rahvusarhiiv. Ühisloome. https://www.ra.ee/kulastajale/uhisloome
Sentiment Analysis on IMDb. https://paperswithcode.com/sota/sentiment-analysis-on-imdb
Transkribus. https://readcoop.eu/transkribus
What Is Digital Humanities? https://whatisdigitalhumanities.com

KIRJANDUS

Berry, David M.; Fagerjord, Anders 2017. Digital Humanities: Knowledge and Critique in a Digital Age. Cambridge: Polity Press.
Bode, Katherine; Arthur, Paul L. 2014. Collecting ourselves. – Advancing Digital Humanities. Toim P. Arthur, K. Bode. London: Palgrave Macmillan, lk 1–12.
Gardiner, Eileen; Musto, Ronald G. 2015. The Digital Humanities: A Primer for Students and Scholars. Cambridge: Cambridge University Press.
Harrower, Natalie; Maryl, Maceij; Biro, Timea; Immenhauser, Beat 2020. ALLEA Working Group E-Humanities. Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities. Berlin: ALLEA – All European Academies.
Kirschenbaum, Matthew G. 2010. What is digital humanities and what's it doing in English departments? – ADE Bulletin, nr 150, lk 55–61.
Kirschenbaum, Matthew G. 2012. Digital humanities as/is a tactical term. – Debates in the Digital Humanities. Toim Matthew K. Gold. Minneapolis–London: University of Minnesota Press, lk 415–428.

Laak, Marin; Viires, Piret; Sarv, Mari 2020. Digikultuuri väljakutseid. Kultuuripärandi kasutamine humanitaarteaduses ja hariduses. – Methis. Studia humaniora Estonica, nr 26, lk 5–16.

Lee, Haimin 2019. 15 years of Google Books. 17. X. https://blog.google/products/search/15-years-google-books (2. IX 2021).

Masso, Anu; Tiidenberg, Katrin; Siibak, Andra (toim) 2020. Kuidas mõista andmestunud maailma? Metodoloogiline teejuht. (Gigantum Humeris.) Tallinn: Tallinna Ülikooli Kirjastus.

Nguyen, Dong; Liakata, Maria; DeDeo, Simon; Eisenstein, Jacob; Mimno, David; Tromble, Rebekah; Winters, Jane 2020. How we do things with words: Analyzing text as social and cultural data. – Frontiers in Artificial Intelligence, kd 3, lk 62.

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. – Eesti Rakenduslingvistika Ühingu aastaraamat, kd 15, lk 139–158.

Ruusalepp, Raivo 2020. Andmestuv kultuuripärand. – Eesti digikultuuri manifest. Toim Indrek Ibrus, Marek Tamm, Katrin Tiidenberg. (Bibliotheca Mediorum et Communicationis.) Tallinn: Tallinn Ülikooli Kirjastus, lk 32–38.

Schnapp, Jeffrey; Presner, Todd 2009. Digital Humanities Manifesto 2.0. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (1. VI 2020).

Taycher, Leonid 2010. Books of the world, stand up and be counted! All 129,864,880 of you. – Google Books Search. 5. VIII. http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html (1. VII 2021).

Tolonen, Mikko; Mäkela, Eetu; Marjanen, Jani; Tahko, Tuuli 2020. Arvutiteaduse kaasamine humanitaarharidusse. – Methis. Studia humaniora Estonica, nr 26, lk 35–50.

Peeter Tinits (snd 1986), MA, Tartu Ülikooli digihumanitaaria ja infoühiskonna keskuse digihumanitaaria spetsialist (Jakobi 2, 51005 Tartu); Tartu Ülikooli ühiskonnateaduste instituudi tekstikaeve spetsialist; Eesti Rahvusraamatukogu digihumanitaaria spetsialist, peeter.tinits@ut.ee

Jane Klavan (snd 1983), PhD, Tartu Ülikooli maailma keelte ja kultuuride kolledži inglise keele kaasprofessor (J. Liivi 4, 50409 Tartu), jane.klavan@ut.ee

Liina Lindström (snd 1973), PhD, Tartu Ülikooli tänapäeva eesti keele professor ning digihumanitaaria ja infoühiskonna keskuse juhataja (Jakobi 2–443, 51005 Tartu), liina.lindstrom@ut.ee

Digital humanities in Estonia

Keywords: digital humanities, data in humanities, digital methods

The article serves as a preface to the special issue on digital humanities of *Keel ja Kirjandus*. The aim of the issue is to familiarize the Estonian readers with the complex and controversial notion of digital humanities by way of examples. The issue gathers overviews of themes and methods in digital humanities along with empirical case studies representing them. We provide a brief discussion on what digital humanities is and why we need it. We consider that the core of digital humanities is the integration of digital data into humanities research. A growing amount of materials of interest to the humanities are now becoming digitized or digitally born. Archival institutions that maintain it have become recently active in facilitating their use for research purposes. Digital humanities can provide the methods, tools and the research community to take advantage of these opportunities. The digitalization of society is happening on many levels and humanities have a chance to contribute to its understanding and benefit from its opportunities. The label digital humanities itself can work as a pragmatic banner at this stage of development, potentially leading to digital skills becoming the norm also in the humanities departments at some point. The articles collected in this special issue give a good overview of the status quo of digital humanities in Estonia, representing various research directions and methods. The articles offer a variety of different perspectives and practical applications in their respective fields related to digital humanities. We believe that the area of digital humanities can advance via practical examples and an understanding of the developments of digital methods in neighbouring fields. We hope that further developments in this field will provide grounds for the compilation of other such thematic issues in the very near future.

Peeter Tinits (b. 1986), MA, University of Tartu, Digital Humanities Specialist in the Centre for Digital Humanities and Text Mining Specialist at the Social Studies Institute (Jakobi 2, 51005 Tartu; National Library of Estonia, Digital Humanities Specialist, peeter.tinits@ut.ee

Jane Klavan (b. 1983), PhD, University of Tartu, College of Foreign Languages and Cultures, Associate Professor of English Language (J. Liivi 4, 50409 Tartu), jane.klavan@ut.ee

Liina Lindström (b. 1973), PhD, University of Tartu, Professor of Modern Estonian and Head of the Centre for Digital Humanities and Information Society (Jakobi 2–443, 51005 Tartu), liina.lindstrom@ut.ee

https://doi.org/10.54013/kk764a2

Digihumanitaariast kultuuriandmete analüüsini

INDREK IBRUS, MAXIMILIAN SCHICH, MAREK TAMM

Digihumanitaaria tõus ja langus

Digihumanitaaria (ingl *digital humanities*) on üks humanitaarteaduste kõige edukamaid lendsõnu viimasel viieteistkümnel aastal. Selle kiiluvees on tehtud ja tehakse mitmekesist ja sisukat uurimistööd ning võib julgelt tõdeda, et digihumanitaaria on kujundanud humanitaarteaduste praegust palet enam kui ükski teine uus lähenemine. Selle eduloo taustal ei saa jätta siiski tähelepanuta, et *digihumanitaaria* on mitmes mõttes probleemne nimetus. Väidame käesolevas artiklis, et kuigi see on olnud väga vajalik lipukiri humanitaarse teadustöö uuendamisel sajandi esimesel kahel kümnendil, siis praeguseks on nimetus oma potentsiaali suuresti ammendanud ning on aeg liikuda edasi täpsemate ja konkreetsemate mõistete ning lähenemiste poole.

Milles seisneb *digihumanitaaria* probleemsus? Kui alustada kõige üldisemast, siis kummitavad digihumanitaariat algusest peale defineerimisraskused. Pea kaks kümnendit väldanud vaidlusi[1] iseloomustab kõige paremini see, kuidas 2004. aastal soovitas valdkonna alustrajav koguteos, „Digihumanitaaria käsiraamat" käsitleda „digihumanitaariat kui omaette distsipliini" (Schreibman jt 2004: xxiii). Kui aga samad autorid koostasid tosin aastat hiljem koguteose täielikult ümbertöötatud versiooni pealkirjaga „Uus digihumanitaaria käsiraamat", siis tunnistasid nad enesekriitiliselt, et „jääb vaieldavaks, kas digihumanitaariat tuleb käsitada kui „omaette distsipliini" või pigem kui omavahel seotud meetodite komplekti" (Schreibman jt 2016: xvii). Vaatamata paljudele katsetele pole digihumanitaarid tänini jõudnud enam-vähemgi konsensusliku arusaamani, kuidas uut valdkonda määratleda ja piiritleda. Väga ilmekas on Jason Heppleri koostatud veebileht What Is Digital Humanities?, mis koondab kokku 817 (*sic!*) digihumanitaaria definitsiooni aastatest 2009–2014.

Üksikasjalikumaks minnes on õigupoolest problemaatilised selle sõnaühendi mõlemad osapooled: *digi* ja *humanitaaria*. Kui alustada viimasest, siis on paradoksaalsel moel tegemist ühtaegu nii liiga avara kui ka liiga kitsa nimetusega, et seda mõttekalt kasutada. Liiga avaraga selles tähenduses, et tänapäeval on vähe teadlasi – kui üldse –, kes identifitseerivad ennast humanitaarteadlasena üldiselt. See on liiga lai katusmõiste, et kirjeldada konkreetset uurimisvaldkonda. Mõistagi saab öelda, et kellegi uurimisvald jääb humanitaariasse, ent sellele peab järgnema selgitus, millisesse humanitaaria valdkonda täpsemalt. Maailmas on vähe humanitaarteaduste professuure, seega ei ole mõistlik neid luua ka digihumanitaaria valdkonnas, mille

[1] Neist saab kõige parema ülevaate, kui lugeda Matthew K. Goldi koostatud antoloogiaid: Gold 2012; Gold, Klein 2016, 2019. Hea sissejuhatuse pakub samuti Melissa Terrase, Julianne Nyhani ja Edward Vanhoutte'i koostatud lugemik „Digihumanitaaria defineerimine" (2013).

tähendusväli pole potentsiaalselt kitsam. Kui vaadata lähemalt digihumanitaarset uurimistööd, siis torkab kiiresti silma, et see ei kata kaugeltki kõiki humanitaaria valdkondi võrdsel määral. Suurem osa digihumanitaariast tegeleb ennekõike tekstiandmetega, filoloogilise uurimistööga selle sõna kõige avaramas tähenduses. Kuid pigem vaeslapse osas on need humanitaaria valdkonnad, kus töötatakse peamiselt pildiliste, esemeliste, ruumiliste, heliliste ja multimodaalsete andmetega. Digitaalset kunsti- ja muusikateadust viljeldakse oluliselt vähem kui digitaalset keele- ja kirjandusteadust, rääkimata näiteks digitaalsest filosoofiast. Seega on *digihumanitaaria* mõneti eksitav nimetus, sest lubab rohkem, kui tegelikult pakub.

Kuid *digihumanitaaria* on ühtlasi liiga kitsendav nimetus, et sobida digitaalselt rikastatud uurimistöö kirjeldamiseks. See aitab elus hoida vana, peamiselt XIX sajandil välja kujunenud vastandust kultuuri- ja loodusteaduste, vaimu- ja reaalteaduste vahel, mida digitehnoloogia võiks aidata ületada, mitte kinnistada. Selmet piirata ennast üksnes *humaniora*'ga, võiks tänapäeva kultuuriuurija püüelda pigem valdkonnaülese koostöö poole sotsiaalteaduste, arvutiteaduse, võrgustikuteaduse, komplekssüsteemide teooria, masinõppe ja tehisintellekti uuringute ning teiste n-ö mittehumanitaarsete erialadega. Niisiis peaks eesmärk olema leida uue, siiani väljakujunemises oleva valdkonna kohta nimetusi, mis ei põlista üle-eelmise sajandi distsipliinipiire, vaid soosivad nende ületamist.

Niisamuti on problemaatiline eesliite *digi-* kasutamine. See tekitab paratamatult vastanduse „analoog-" või traditsioonilise humanitaariaga, mis ei ole kuidagi põhjendatud. Esiteks on rangelt võttes iga humanitaarteadlane „digihumanitaar" selles mõttes, et ta kasutab oma uurimistöös digitehnoloogilisi töövahendeid (olgu selleks või tekstitöötlusprogrammid ja uurimisandmed kõvakettal), küsimus on vaid digivahendite kasutamise määras. Teiseks loob see pettekujutelma, nagu tegeleks digihumanitaaria sootuks teise ainesega kui traditsiooniline humanitaaria, kuigi allikate digiteerimine ei muuda tingimata nende sisu ega olemust. Ainult väike osa digihumanitaariast keskendub n-ö digitaalselt sündinud teabele, valdavalt töötatakse siiski digiteeritud materjalidega, mille kasutamine eeldab üldjuhul samasuguseid kutseoskusi nagu „analooghumanitaarias". Muidugi on digihumanitaaria egiidi all toodud humanitaarteaduslikku uurimistöösse mitmeid uusi meetodeid, ent metodoloogiline uuenemine iseloomustab humanitaariat algusest peale ja see ei eelda tingimata valdkonna ümbernimetamist. *Digi*-eesliite rõhutamine võib tekitada olukorra, kus digihumanitaare käsitatakse klassikalise humanitaarse teadustöö tugiisikutena, tehnilise abipersonalina, kes aitavad arvutada, visualiseerida või mudeldada seda tööd, mida teevad n-ö pärishumanitaarid. Praeguses institutsionaalses töökorralduses on näha, et paljud n-ö digihumanitaaria spetsialistid leiavad tööd ennekõike raamatukogudes ja arhiivides, mitte ülikoolides, kus vastavaid töökohti tuleb endiselt tikutulega otsida.

Digihumanitaaria kritiseerimisel on juba auväärne traditsioon, seda nii valdkonna sees kui ka seda enam väljaspool. See on kindlasti põhjuseks, miks nii mõnedki autorid on nimetuse ebamäärasust püüdnud välja mängida selle tugevusena. Nii nimetab näiteks Matthew Kirschenbaum (2012) digihumanitaariat „taktikaliseks terminiks", mis aitab saavutada praktilisi eesmärke, näiteks hankida rahastust või

luua ametikohti. Sageli räägitakse digihumanitaariast kui katusterminist või kui nn suurest telgist (ingl *big tent*), mille alla mahub palju eriilmelisi tegevusi (nt Svensson 2012; Weingart, Eichmann-Kalwara 2017). James Smithies (2017: 2) viitab aga digihumanitaariale kui „ujuvale tähistajale". Steven Jones (2014: 5) resümeerib asjakohaselt: „Endiselt ei ole selge, kas digihumanitaariat tuleks pidada valdkonnaks või (inter)distsipliiniks või lihtsalt praegu mugavusest kasutatavaks nimetuseks selle kohta, mida kõik humanitaarid üpris varsti tegema hakkavad." Teises mõjukas sedastuses kirjutasid Todd Pressner ja Jeffrey Schnapp oma „Digihumanitaaria manifestis 2.0" (2009): „Digihumanitaaria ei ole ühtne valdkond, vaid rida koonduvaid praktikaid, mis uurivad üht universumit."

Üha enam kohtab katseid täpsustada digihumanitaaria nimetust mõne täiendi toel. Nii on näiteks mõned autorid, kes soovivad näha digihumanitaarias ennekõike revolutsioonilist potentsiaali, mis mängiks ümber senise humanitaarteaduste korralduse, kutsunud üles edendama „transformatiivset digihumanitaariat" (*transformative digital humanities*, vt Balkun, Deyrup 2020). Teised on soovitanud lõimida digihumanitaariat kriitilise teooriaga ja pakkunud välja „kriitilise digihumanitaaria" (*critical digital humanities*) nimetuse (Berry, Fagerjord 2017: 136–150; vt ka Berry 2014).

Kuid need katsed mängida digihumanitaaria ebamäärasus välja selle taktikalise tugevusena või piirata seda uute täiendite lisamisega ei tundu perspektiivikad. Seega teeme ettepaneku käsitada digihumanitaariat selle menule ja populaarsusele vaatamata üleminekuaja nimetusena, ajutise võitlusloosungina, mis on aidanud valdkonnale tähelepanu tõmmata ja huvilisi mobiliseerida, mille raames on tehtud olulist metodoloogilist arendustööd, ent mis praeguses arenguetapis vajab väljavahetamist konkreetsemate ja vanadest vastandustest vabade nimetuste vastu.

Me näeme selget vajadust vabaneda mitmest eristusest, mis on digihumanitaaria projekti sisse kirjutatud. Esiteks tuleb ületada eristus idiograafiliste nähtuste kvalitatiivse ja nomoteetiliste nähtuste kvantitatiivse analüüsi vahel. Teiseks tuleb ületada üksiku ja üldise, osa ja terviku uurimise dihhotoomia. Kolmandaks tuleb ületada digitaalse ja analoogse vastandamine, sest leidub mitmeid asjakohaseid arvutuslikke ja uurimuslikke ülesandeid, mis eeldavad pigem mittedigitaalseid lahendusi. Ja neljandaks tuleb püüda väljuda humanitaarteaduste klassikalistest piiridest ning arendada dialoogi näiteks komplekssüsteemide teooria, masinõppe, süsteemibioloogia, evolutsioonilise majandusteooria ja paljude teiste uurimissuundadega. Pakume välja, et *kultuuriandmete analüüs* võiks olla üks neid uusi uurimisvaldkondi, mida arendada teispool digihumanitaariat ja mis võiks aidata sõnastatud eristusi ületada. Põhjuseid selleks üritame allpool esitada, aga sissejuhatuseks rõhutagem, et mõiste *kultuur* tähendus on küll samamoodi lai nagu mõistel *humanitaaria*, kuid see ei viita mitte institutsionaliseerunud akadeemilisele valdkonnale, vaid uurimisobjektile: uuritakse kultuuri tähendussüsteeme ja praktikaid, samuti kultuuri erinevaid materiaalseid ja mittemateriaalseid esinemisvorme ning nende kõigi muutumist ajas. Mõiste *andmed* rõhutab aga asjaolu, et pärast digitaalset kodeerimist muutuvad kultuurisisu ja -praktikad andmeteks, mida saab uurimisülesanneteks eri viisil ümber organiseerida, nii et see rikastab märkimisväärselt uurimiseks kasutatavat

metoodikat (ennekõike võimaldades teistest distsipliinidest meetodeid laenata). Teisalt tõstatab andmete mõiste selgemalt ka kriitilise refleksiooni vajaduse uurimistöös, sest andmed on pahatihti problemaatilised või eksitavad. Kombinatoorne mõiste *kultuuriandmed* viitab omakorda sellele, et kultuuri saab uurida, ristates eri andmekogusid, mis on kogutud eri viisil ja võivad viidata väga erinevatele modaalsustele, nii kultuuri artefaktidele, sisule kui ka praktikatele. Selline analüüs võimaldab keskenduda kultuuri kompleksse sisesuhete uurimisele, mille selgitamise juurde järgmiseks pöördume.

Kultuuri kompleksuse mõtestamine

Teel uue lähenemise poole, mida nimetame kultuuriandmete analüüsiks, toome esimesena esile kompleksuse mõiste, mida võib pidada ühenduslüliks kultuuri uurimise eri lähenemiste – see tähendab, kultuurinähtuste ja -protsesside nn nomoteetiliste ja idiograafiliste vaatenurkade – vahel. Ajalooliselt on humanitaarteadusi pärssinud eeldus, et need peavad keskenduma üksiknähtustele, sageli vastandudes loodusteadustele, mis vaatlevad väidetavalt ainult üldisi nähtusi. Tavalugejale või asjatundjale võib see tunduda liialdatud vastandus, ent nagu iga valdkonnaüleste huvidega teadlane võib kinnitada, on selle taga karm tegelikkus. See tegelikkus, mida tuntakse „kahe kultuurina" (Snow 2017 [1961]), ei tähenda, et vastandus oleks sisuline – seda ta ei ole –, vaid et seda kasutatakse relvana eelretsenseerimisel, värbamiskomisjonides ja rahastamisotsustes.

Niisiis on arukas üle korrata, kust pärineb kahe väidetavalt vastandliku teaduskultuuri argument. Humanitaaria, seega ka digihumanitaaria ja loodusteaduste ideoloogiline lahutamine põhineb XIX sajandi filosoofial, täpsemalt Wilhelm Dilthey töödel. Temalt pärineb *vaimuteaduste* (sks *Geisteswissenschaften*) ja *loodusteaduste* (sks *Naturwissenschaften*) vastandus (Dilthey 1959 [1883], 1997). Tuleb küll täpsustada, et Dilthey jaoks ei olnud see vastandus nii absoluutne nagu paljudele tema järgijatele. Dilthey käsituses erinevad loodus- ja vaimuteadused teineteisest esmajoones oma uurimisobjektide konstitueerimise poolest: kui vaimuteadusliku teadmise põhjendus tuleb leida sisemise kogemuse analüüsist, siis loodusteaduste pärusmaaks jääb välise kogemuse eritlus. Algusest peale on tähelepanuväärne, et see vaimumaailma ja füüsilise maailma lahutamine mitte ainult ei toimunud enne suuri arenguid kõigis teadustes, vaid oli juba XIX sajandil vastuolus näiteks Goethe esitatud seisukohaga, mis on toodud ära ajakirja Nature esimeses numbris 1869. aastal: „Ka kõige ebaloomulikum on siiski looduse loomulik osa" (Huxley 1869). Samal ajal kui ajakiri Nature hakkas iroonilisel kombel edaspidi kultuuriuurimusi valdavalt ignoreerima, kinnistas teine saksa teadlane Wilhelm Windelband veelgi kindlamalt humanitaar- ja loodusteaduste eraldamise, mõeldes välja kaks terminit, millega iseloomustada nende väidetavalt ületamatut erinevust (Windelband 1993 [1894]). Windelband väitis, et loodusteadused on peamiselt *nomoteetilised*, otsetõlkes „seadust andvad" teadused, mis püüavad mustrite põhjal tuvastada seaduspärasusi. Ilmselge näide oleks see, et füüsilised esemed „lihtsalt" järgivad liikumise seadusi, mille sõnastas

Isaac Newton. Seevastu vaimuteadused, nagu Windelband väitis, on *idiograafilised* ja keskenduvad üksikjuhtumitele. Selle vaate kohaselt on ajaloolised sündmused alati kordumatud, umbes nagu see mitte kunagi täielikult korduv olukord, milles meie seda artiklit kirjutame või teie seda loete.

Kui võtta omaks selline vaatenurk, võib kultuurinähtuste kvantifitseerimine tunduda keelualale tungimisena, sest selle ainus eesmärk saab olla nomoteetiliste kultuuriseaduste sõnastamine. Digikultuuri laiema kontseptualiseerimise kontekstis on selle näiteks olnud Friedrich Kittleri kriitika Niklas Luhmanni süsteemiteoreetilise sotsioloogia aadressil. Kittler töötas välja uue lähenemise (mida tuntakse meediaarheoloogia nime all), et uurida kultuuri muutumist ajaloos vastavalt sellele, millised materjalid ja tehnilised meediaressursid olid parasjagu kättesaadavad. Nii jõudis ta radikaalse historitsismini, mille eesmärk oli lammutada näiteks selliste mõistete nagu *meedia* ning teiste kommunikatsioonivoogusid korraldavate ja tähendusi loovate kultuuriinstitutsioonide universaalsus. Sellest lähtudes oli ta vastu ka laiematele seadusesarnastele ühiskondliku või kultuurilise evolutsiooni teooriatele, kaasa arvatud Luhmanni omadele (vt Kittler 1994; Winthrop-Young 2000: 411).

Ometi on olnud ka vastuliikumisi ja alternatiivseid ettepanekuid, millest kõige silmapaistvama esitas Raymond Williams, kui ta sõnastas vajaduse *kultuuriuuringute* (ingl *cultural studies*) kui uue kultuurianalüüsi meetodi järele. Williams järgis Birminghami nüüdisaja kultuuriuuringute keskuse (Birmingham Centre for Contemporary Cultural Studies) eeskuju – sealgi keskenduti algselt saksa *kultuuriteaduste* (*Kulturwissenschaften*) importimisele ingliskeelsesse akadeemilisse maailma (vt Shuttleworth 1966). 1974. aastal pidas Williams ettekande, kus pakkus välja uue lähenemise, mida ta nimetas kõigepealt *kultuuriteaduseks* ja mille hiljem muutis *kultuuriuuringuteks* (mis tema järgi on korrektne ingliskeelne vaste saksa *Kulturwissenschaft*'ile; Williams 1974). Selles ettekandes lähtus ta Dilthey, Weberi ja Marxi saksa intellektuaalsest pärandist, ent soovis minna kaugemale: määratleda „erinevate praktikate vahelised suhted" XX sajandi teise poole kultuuriteaduse „keskse probleemina" (vt Hartley 2009). Kuigi sellest hoolimata keskendusid kultuuriuuringud edaspidi siiski kultuuri mikrokontekstidele, on nüüd seoses kultuuriandmete analüüsi esilekerkimisega saanud võimalikuks teha algust kultuuris esinevate suhete süstemaatilise analüüsiga ka makrotasandil.

Tuleb rõhutada, et humanitaarteadustest ei ole kunagi puudunud n-ö seadusesarnased üldistused. Tõepoolest, humanitaarteaduste õpikuteadmised ja haridus on täis vaistlikke üldistusi, tüüpe ja kategooriaid, mis eeldavad vähemalt vaikimisi mingit näidetel põhinevat keskmist koondit. Selline üldistus on näiteks gooti kaar 20-eurosel rahatähel. Teisisõnu, võime ette kujutada keskmist gooti katedraali, kuigi sellist keskmist olemas ei ole. Esimene konkreetne katedraal, mis pähe tuleb, on tõenäoliselt silmapaistev, mitte keskmine näide. Musternäidet ennast ümbritseb tavaliselt suurem hulk vähem tüüpilisi näiteid. Sellegipoolest on üldistus loogiline, sest nii on võimalik käsitleda eripära üksikute näidete lõikes. Üksikute konkreetsete tähelepanekute ja üldistuste vaheliste suhete uurimine ja sedastamine on humanitaarteaduste lahutamatu osa, mida tuntakse hermeneutilise ringi nime all – järjekordne Diltheyga seonduv saksa filosoofia mõiste (vt nt Bod 2014: 333–334).

Teisisõnu on selge, et nii üksikjuhtumite uurimine kui ka üldistamine on kultuurinähtuste mõistmiseks hädavajalikud tegevused. Pigem on asi selles, et kultuurinähtused on komplekssed ega järgi tingimata „lihtsaid" kvantitatiivseid seadusi, millel on „lihtsad" sisulised keskmised. See on oluline põhimõte ja tuleks pikemalt lahti seletada. Siinsel juhul ei tähenda *kompleksne* pelgalt keerulist, mis on selle sõna kõige levinum tähendus (Zinoviev 2016). Sõna *kompleksne*, mida me siin kasutame kompleksteaduse mõistes (Mitchell 2009; Complexity Explained; Engelbrecht, Kitt 2020), on mõnes mõttes vastuolus üldlevinud arusaamaga komplitseeritusest. Üksik ajaloosündmus, nagu ka konkreetne maal või visualiseering bioloogiaalases teadustöös võivad olla komplitseeritud, ning meie mõistmise suurendamiseks on vaja teostada nende hoolas vaatlus ja diferentseeritud kvalitatiivne uurimistöö. Teisalt võib suurem hulk näiliselt lihtsaid ja eraldiseisvaid kultuurilisi suhteid anda kokku kompleksse mustri, kus esineb teatavaid korrapärasid, kuid samal ajal ka spetsiifilisi variatsioone. Komplikatsiooni mõistega täiesti lõpuni minnes võime eksida spetsiifiliste, erisuguste detailide massi, samas kui kompleksus võib kujuneda mõnest lihtsast protsessist. Kujundid, mille moodustab kuldnokaparv, kus kõik linnud on ühesugused ja järgivad mingit hulka suhteliselt lihtsaid reegleid, on üks sellise organiseeritud kompleksuse näide (Weaver 1948). Teine on linnastruktuur ja -dünaamika, mis tekib linnaelanike kohalike tegevuste tulemusena (Jacobs 1961; Barthelemy 2019). Sarnaseid nähtusi, kus ühe konkreetse osise lokaalsest tegevusest kujuneb organiseeritud kompleksus, võib leida kõikjal maailmas: tuumafüüsikast arenes molekulaarkeemia, mis viis omakorda rakubioloogia ja organismideni, seejärel teadvuse ja sotsiaalse suhtluseni, ning lõpuks tekkisid materiaalsed ja mittemateriaalsed kultuuritooted (Anderson 1972; Riedl 2000).

Igal tasandil, kus kompleksus tekib kohalikust tegevusest, saab tervik mitte üksnes suurem, vaid ka teistsugune ehk „rohkem on erinev" (ingl *more is different*), nagu Paul Anderson 1972. aastal löövalt sedastas. Selle tagajärjel tekib üllatav võimalus kvalitatiivseks uurimistööks: tõenäoliselt on humanitaarteadustes kvantifitseerimise tulemus mitte lihtsalt kvantitatiivsed seadused, vaid kvaliteedi uudsed vormid. Kohalike spetsiifikate kooskõlast kerkivad esile uued huvitavad klastrid, ka esmapilgul segadusse ajavad mustrid ning erinevate statistiliste mudelite võrdlusvõimalused koos ilmnevate tähenduslike eranditega. Meso- ja makrotasandil leiame äratuntavaid kvalitatiivseid nähtusi, mis pole täielikult ei korrapärased ega ka juhuslikud, meenutades linnatänavate mustreid või maastike topograafiaid, mis saavad nähtavaks lennukõrguselt (Lee jt 2017; Mandelbrot 1982). Suured kogused lihtsaid suhteid, nagu saja tuhande teadaoleva üksikisiku sünni- ja surmakoha omavahel ühendamine, võivad viia kultuuriajaloo metanarratiivi ja kaua otsitud n-ö keskmise kestusega mustrite ilmnemiseni mesotasandil, samal ajal kui makrotasandil saavad ilmsiks demograafilised „rändeseadused" (Schich jt 2014; Ravenstein 1885; Hexter 1972). Süstemaatiline nihe, mis ulatub kaugemale üksikisiku loomulikust kognitiivsest piirist, tuleb esile mitme suure andmestiku võrdleval analüüsil, samas kui kõik eelnimetatu jääb kitsamalt kohalikele eripäradele keskendudes varjatuks.

Samal ajal – ja see on oluline – nihkub fookus sisult funktsioonile, esemetelt suhetele, struktuurilt dünaamikale ning sünteesilt analüüsile (Schich 2019; Cassirer

2010 [1910]. Kuigi võib väita, et humanitaarteadused on alati olnud huvitatud kontekstidest ja konkreetsetest sündmustest, ei saa eitada, et mitmed põhidistsipliinid alates kunstiajaloost ja kirjandusest, lõpetades meediauuringutega määratlevad end endiselt ennekõike sisu kaudu: uuritakse kunstiteoseid, tekste, filme või muid meediume. Sealjuures on aga praeguseks tehtud suuri edusamme nende omavaheliste interaktsioonide mõistmisel, ning k o m p l e k s v õ r g u s t i k e teadus on üha levinum hulgas valdkondades, kus komplekssus kujuneb kohalikust tegevusest – alates füüsikast ja lõpetades kultuuriga (Barabási 2016; Schich 2016; Schich jt 2016; Ahnert jt 2021).

Kultuurisuhetest tärkava komplekssuse kvantifitseerimisel ei ole ohtu, et tavapärane hermeneutiline teadustöö humanitaarteaduses asendataks lühinägelike mehhanistlike või positivistlike kultuurimudelitega. Tekkiva komplekssuse kvantifitseerimine võib hoopis pakkuda võimalust täpsustada kollektiivses hermeneutilises ringis kujunevaid arusaamu, kui humanitaarne loogika ja intuitsioon kohtuvad kvantitatiivsete tõenditega, mis on omakorda mõeldamatud ilma kumulatiivsete tõendite hoolika kvalitatiivse kaardistamiseta.

Üksik ja üldine moodustavad semiosfääri

Üksiku ja üldise, makro- ja mikrotasandi komplekssuste mõtestatud seostamiseks on juba olemas üks, meil Eestis käepärane lähenemine – see on kultuurisemiootika. Pakume, et uue andmeanalüütilise põhjaga kultuuriteadusliku uurimisprogrammi arendamiseks võib just Eesti olla sobilik ja ainulaadne paik, kuna oleme siin kujunenud parimateks ekspertideks Juri Lotmani semiosfääri teooria vallas. Oleme seda ka edasi arendanud ning vorminud selle läbimõeldud raamistikuks, mis lubab analüüsida komplekseid ja dünaamiliselt arenevaid kultuuriprotsesse. Pakume, et Lotmani hilise loomeperioodi teooria kui süsteemiteoreetiline ja holistlik lähenemine kultuuridünaamikale on hea alus nüüdisaegsete digitaalselt vahendatud või võimaldatud kultuuriprotsesside uurimiseks (vt Hartley jt 2021a).

Kultuurisemiootika ja kultuuriandmete analüüsi hea klapp ei ole juhuslik, tuginesju Lotman oma teooria väljatöötamisel suurel määral küberneetikateooriatele (ennekõike Norbert Wieneri ja Ross Ashby töödele) ja mitmesugustele teistele lähenemistele (sh Ilya Prigogine'i (Prigogine, Stengers 2018 [1984]) töödele dissipatiivsetest struktuuridest ja iseorganiseerumisest), mida tänapäeval peetakse nii süsteemikui ka posthumanistlike teooriate eelkäijateks. Tõstkem siinkohal esile ka Jakob von Uexkülli töö (1909, 2012) mõju kultuurisemiootika edasiarendustele pärast Lotmanit (Kull, Lotman 1995; Kull 1998). Digikultuuri süsteemse uurimise kontekstis on oluline nimetada, et ka Ludwig von Bertalanffyd inspireeris *üldise süsteemiteooria* (1951) väljatöötamisel just Uexkülli teadustöö.

Nii eelnimetatuid kui ka kultuurisemiootika nüüdisaegseid arendajaid on mõjutanud ennekõike Uexkülli omailma (sks *Umwelt*) mõiste. Elusloodusele eri liikide, aga ka eri inimeste omailmad võivad olla väga erinevad, sõltuvalt nende tajukanalite (meelte) spetsiifilistest lubavustest. Omailmadevahelistes tõlgetes tekib esmane

jagatud tähendussüsteem. Mida rohkem on selliseid tõlkeid ehk omavahel suhtlevaid üksusi, seda rohkem on ka eraldiseisvaid tähendussüsteeme, mis kõik kokku moodustavad nn semiosfääri – jagatud, ühiselt toodetud tähenduste süsteemi. Seejuures tuleb aga rõhutada, et nii esmalt Uexkülli kui hiljem ka Lotmani jaoks algas igasugune suurema süsteemi, sh kultuuri kujunemine selle allsüsteemide autokommunikatsioonist (Kull, Lotman 2012; Torop 2008). Lotmani jaoks tähendab autokommunikatsioon süsteemi võimet iseendast rääkida, end oma keskkonnas eristada ja defineerida ning seeläbi kommunikatiivselt luua. See on ka võime toota enesekohaseid tekste, seejuures keskkonnast tähenduslikuna näivat süsteemi sisse tõlkida, seda oma seniste koodidega jooksvalt lõimida ja nii omasüsteemi uuendada. Arvestades, et eri süsteemid – näiteks eri kunstižanrid, poliitilised ideoloogiad või idufirmade kogukonnad – moodustavad üksteisele keskkonna, tähendab kultuuri areng protsesse, kus nood süsteemid üksteist tõlgivad ja nõnda omasüsteeme muutuvas keskkonnas taasloovad, ümbritseva keskkonnaga kohandavad. Selliseid süsteeme võib olla väga väikseid (nt paari artisti poolt loodud ja lühiajaliselt kestev muusikažanr), keskmises suuruses (nt Eesti rahvuskultuur) ja väga suuri (kogu inimkultuur või nt YouTube'is viljeldav, miljardite inimeste koosloomes sündiv audiovisuaalkultuur).

Seejuures on aga kultuurisemiootika keskne idee, et kõik tasandid tingivad üksteist. Näiteks ei saaks üksik uus muusikažanr sündida ilma selleta, et kogu ülejäänud kultuuriruum enne eksisteeriks – see on tähenduslik, eriline ja „uus" vaid kogu muu kultuuri kontekstis, tõukub kõigest eelnevast kausaalselt. Kui aga uus kultuurivorm või rühmitus ja seda kandev diskursus sünnivad, nihutavad need omakorda paigast kogu olemasoleva kultuurisüsteemi – ka kõik muud kultuuri allsüsteemid peavad kas otsesemalt või kaudsemalt sellega arvestama, sellega kohanema ja nii muutub lõpuks kogu üleilmne semiosfäär. Eelnevaga seondub veel üks kultuurisemiootika juhtpõhimõte: eri tasandite süsteemid on n-ö isomorfsed, nad on teatud mõttes sarnase struktuuriga, analüüsides üht, saab teha järeldusi teise kohta. Just sellest põhimõttest lähtudes on võimalik näiteks üksikute materiaalsete fragmentide põhjal teha olulisi järeldusi möödanikukultuuride struktuursete seatuste kohta.

Siit johtub, et semiosfääri teooria on ülal kirjeldatud nomoteetiliste ja idiograafiliste lähenemiste üks võimalik kompromiss või ühendaja. Sest selle eeldus on, et üks tingib teise, üht uurides saab teha teise kohta järeldusi, ning mis veelgi olulisem: uurides autokommunikatiivset või iseorganiseerumise dünaamikat kultuuri eri allsüsteemides, nii väga väikestes kui ka väga suurtes, samuti nendevahelisi suhteid või tõlkeid, saab süsteemselt uurida põhjuslikke seoseid kultuuri eri allsüsteemide arengute vahel läbi aegade. Eespool selgitasime, miks on komplekssus oluline mõiste eri tasandite kultuurinähtuste mõistmisel ja kontekstualiseerimisel ning et komplekssust võib vaadelda nii kultuuri mikrotasandil, spetsiifilisi nähtusi uurides kui ka meso- või makrotasandil, kus ilmnevad mikrotasandi protsesside interaktsioonid. Kultuurisemiootilises vaates sünnibki aga komplekssus eri tasandi süsteemide autokommunikatiivsest või modelleerivast tegevusest. Väikesed süsteemid on hõlmatud korraga mitmesse suuremasse kultuurisüsteemi, mis on väiksemate suhtes nomo-

teetilised, kuid mis koosmõjus toovad nende keskkonda/konteksti komplekssust. Samas toodab väikeste süsteemide iseorganiseerumine, autokommunikatsioon ja modelleeriv tegevus uut komplekssust, mis võib viia suured süsteemid muutusteni. Sellist komplekssuse esilekerkimist eri tasandite koosmõjul on võimalik uute kultuuriandmete analüüsi meetoditega uurida. Seni oli selline kultuurisemiootika ideestikust lähtuv uurimisprogramm võimalik vaid hüpoteetiliselt või väikeses mahus, nüüd on sellest suurte andmekorpuste ning kultuuriandmete analüüsi areneva meetodiarsenali toel reaalsus saamas.

Järgnevalt tõstame esile ühe kultuurisemiootilise mõiste, mida saaks ja tuleks erinevatel viisidel kultuuriandmete analüüsis operatsionaliseerida ehk kasulikult rakendada. See ei lähtu küll otseselt kultuurisemiootikast, vaid nn kultuuriteadusest (ingl *cultural science*) – Austraalia õpetlase John Hartley ja tema kolleegide rajatud uurimissuunast, mis tugineb kultuurisemiootikale, aga ka kultuuriuuringutele, komplekssusteadusele ja evolutsioonilisele majandusteadusele (vt Hartley, Potts 2014; Hartley 2020). See mõiste on *deem* (ingl *deme*) ja see viitab autokommunikatiivselt toimivale inimrühmale, keda ühendavad peale ühise enesekohase diskursuse ja seda kandvate tekstide ka ühised meediakanalid, -vormid ja kommunikatsioonivahendid. Näitena tänapäevast saab siin tuua ühised Telegrami või Facebooki grupid või ehk jälle tihedalt lõimitud Twitteri võrgustikud. Vanema aja näited oleksid enne trükipressi ühist kirjavara loonud ja lugenud Euroopa vaimulikud või ka XX sajandi teise poole Eesti kineastid/dokumentalistid, keda ühendasid meedium, žanrid ja tasapisi ajas muutuvad ideed ja diskursused, mida filmides ka väljendati. Twitteripõhiseid deeme on tänapäeval võrgustikuanalüüsi meetoditega lihtne uurida, kuid kasutades näiteks Eesti Filmi Andmebaasi metaandmete korpust, saab sama edukalt analüüsida ka Eesti kineastide ajaloolisi võrgustikke ja deeme (vt Ibrus, Ojamaa 2020).

Peilides aga näiteks Eesti subkultuursete muusikakogukondade Facebooki gruppe ja sealseid autokommunikatiivseid praktikaid ehk enesekohast tekstiloomet, saab uurida, kuidas neis subkultuurides identiteeti luuakse, kuidas end keskkonnast eristatakse ja eriti kuidas seda tehakse pikkade perioodide lõikes (sellega tegeleb praegu oma doktoritöös Madis Järvekülg). Otsides eri gruppide vahelisi linke või neid sidestavaid inimesi, uurides ka neis gruppides aset leidva autokommunikatsiooni sarnasusi ja erinevusi, saab edasi uurida, kuidas on erinevad subkultuurid seotud, kuidas info nende vahel liigub, kuidas nad üksteist tõlgivad. Lähtudes nii Lotmani kultuurisemiootikast kui ka Hartley ja Jason Pottsi kultuuriteadusest, on sama oluline ka deemide eristumispraktikate, sh konfliktide uurimine. Sest eristumises ja konfliktis väljenduvad erinevate ajastute olulised alternatiivid, samuti plahvatuslikud arengud, mille tulemuseks võivad olla uued, teistel alustel toimivad dialoogid või ühendsüsteemid. Näiteks on Hartley, Ibrus ja Maarja Ojamaa (2021b) näidanud konflikti üleilmsete kliimaliikumiste ja vandenõuteooriate internetikogukondade vahel, mis aga samas on paradoksaalselt kujunenud ka dialoogiliseks – need süsteemid on koevolutsioonilised. Eri tasanditel aset leidvate dialoogide, konfliktide, tõlgete ja autokommunikatsiooni praktikaid andmeanalüüsi meetoditega uurides hakkab lõpuks kokku tulema laiema kultuuridünaamika pilt.

Tallinna Ülikoolis arendame parasjagu koos ERR-i ja Postimees Grupiga, samuti Balti Meediamonitooringuga uurimisprojekte, milles püüame eelkirjeldatud mõisteid ja neist lähtuvaid küsimusi rakendada Eesti ajakirjanduslikele andmebaasidele ja nende aastakümneid katvatele tekstikogudele. Metaandmete andmebaasidest saab kätte tekstide/saadete loojad, masinõppealgoritmide ja muude meetodite abil saab neist tuvastada rääkijad ja tegutsejad ning teised olemid, samuti modelleerida tekstide teemasid. Kõiki neid elemente saab ühendada võrgustikeks ja uurida nende muutumist ajas. Saab uurida, kuidas moodustasid ühiskonnas aktiivsed inimesed erinevaid temaatilisi või diskursiivseid deeme, kuidas nad infot/diskursusi/keelt vahetasid ja üksteise vastasmõjus, üksteist tõlkides arenesid.

Kui naasta diskussiooni juurde nomoteetilise ja idiograafilise lähenemise vastuoludest ning kultuurisemiootikast kui kompromissist, siis mainitud uurimistöös soovime näidata, kuidas on laiem kultuuridünaamika tinginud üksikute tekstide (näiteks konkreetsete telesaadete) esilekerkimist ning kuidas need kui ülalkirjeldatud lihtsad mikrotasandi nähtused on omakorda tinginud uute komplekssemate deemide organiseerumist ja edasist muutust terviksüsteemis.

Tuleb aga vaadata ka kaugemale: kas seda tüüpi uurimisprogrammi saaks arendada rakenduslikuks, tulevikku vaatavaks, mitmesuguseid tänapäevaseid väljakutseid (valeinfo levik, kiirelt eskaleeruvad veebikonfliktid jms) käsitlevaks uurimistööks? Kas meie pakutaval uurimisprogrammil saaks olla ka kultuuri- ja meediapoliitiline mõõde? Mõtelgem tagasi meteoroloogia arengule: praeguse täpsuse saavutamiseks kulus ligi kaks sajandit. Alguses peeti seda lootusetuks narruseks, nüüd on see aga üks arenenumaid valdkondi, mis suurte andmemahtude kiirkäitlemise edukust tõendab, pakkudes seejuures kõigile kättesaadavat avalikku hüvet. Kui alguses, XIX sajandi keskpaigas, pilgati esimesi eksperimenteerijaid, siis tänapäeval on tegu maailmamajanduse ja ühiskondade igapäevaelu planeerimisel hädavajaliku teabega. Kliimakriisi arenedes valdkonna tähtsus aina kasvab. Kas ka kultuuriandmete analüüs saaks areneda prognoosivaks (st mitte kindlalt ennustavaks, aga kalkuleerivalt ettevaatavaks) teaduseks?

Selle potentsiaali avamisele suunatud uurimistöö avalik-õiguslikes ülikoolides on seda olulisem, et erasektoris on see käimas nii ehk naa – asjaolu, mis sai selgeks hiljemalt Cambridge Analytica skandaaliga (Isaak, Hanna 2018). Kuid viimane oli siiski suhteliselt vähese mõju ja tegevusulatusega. Enamiku andmeid meie sotsiaalse elu, sidemete ja huvide, kultuurilise tähendusloome kohta koguvad ja salvestavad ülemaailmsed suurplatvormid. Nick Couldry ja Ulises Mejias (2019) nimetavad seda *andmekolonialismiks*. Kui kolonialismi esimene laine tähendas, et lääneriigid annekteerisid tooraine hankimiseks territooriume mujalt, siis uue laine kolonialism seisneb selles, et suurplatvormid ja nendega seotud „andmestamistööstused" mudeldavad ja seeläbi koloniseerivad inimeste elusid, nii erasfääri kui ka sotsiaalset elu, et neist isetahtsi väärtusi ekstraheerida. Koos 5G ja asjade interneti levikuga võib selline elude kaubastamine andmestamise kaudu kujuneda üsna laialdaseks.

Kultuurisemiootika põhise kultuuriandmete analüüsi loomulik eesmärk peaks olema eelkirjeldatud protsesse eirata ja väärata. Siin tuleb silmas pidada, et kultuurisemiootika järgi lisab igasugune mudeldamine, ka teaduslik, laiemasse kultuuri-

ruumi mudeleid juurde. Nagu Peeter Torop (2015) on näidanud, iseloomustab aga kultuurisemiootikat teadlikkus oma tegevuse tulemustest. See aga ei tähenda, et tegevustest peab hoiduma, vaid tuleb panustada kasulike mudelitega. Kui kultuurisemiootika teesiks on olnud, et kultuuri arengu ja hea talitlusvõime eelduseks on selle sisemine mitmekesisus, siis peavad analüütilised tegevused panustama mudelitega, mis seda mitmekesisust tõstavad, avavad nii uusi reaalsusi kui ka osutavad kultuuri allsüsteemide rikkusele.

Semiosfäärilise modelleerimise toel saab näidata, kuidas teadmised ja väärtused kultuuris sünnivad ja levivad paljude inimeste tegevuse toel. Selline modelleerimine on iseenesest valgustav kultuuri toimemehhanismide osas, vähendades potentsiaalselt ebausku, usaldamatust ning gruppidevahelisi hirme. Kuid ühtlasi aitab see ennustada konfliktikoldeid ja turbulentseid perioode. Kui aktsepteerime asjaolu, et inimkultuur ehk semiosfäär mõjutab ja vormib antropotseeni ajastul muid süsteeme (biosfäär, geosfäär, atmosfäär), siis tuleb semiosfääri planetaarset toimimist võimalikult hästi mõista, et hinnata selle riske elukeskkonnale. Nii peaks globaalsete kultuuriliste protsesside prognoosimise eesmärk olema ennekõike a v a l i k u v ä ä r t u s e (Benington, Moore 2011; Mazzucato 2018) loomine – st kõigile osapooltele kättesaadavate teadmiste tootmine ja jagamine. See tähendab ennekõike nn hea valitsemistava huvides prognoosimist.

Kogu eelnevat potentsiaali rõhutades tuleb tunnistada, et sellisel üleilmse kultuuridünaamika modelleerimisel on ka oma olulised kitsaskohad. Meenutagem siinkohal Jorge Luis Borgese lühinovelli „Täpsusest teaduses" („Del rigor en la ciencia"), kus tundmatu impeeriumi kartograafid koostasid üksikasjaliku 1:1 kaardi kogu oma riigi kohta. Kuid selle elanikud leidsid eelduspäraselt, et see on kasutu, ja kaart hävitati. Et 5G mobiilside, asjade internet ja liitreaalsuse rakendused on peatselt tegelikkuseks saamas, mängitakse taas ideedega kogu füüsilise ruumi nn peegelmaailma või digitaalse kaksiku loomisest, Eestiski on majandus- ja kommunikatsiooniministeeriumi eestvõttel sarnane kaksik loomisel (vt Eesti 3D kaksiku uuring). Siiski peab siinses kontekstis meeles pidama, et kui (meta)andmeid kasutatakse sotsiaalse/kultuurilise/füüsilise reaalsuse modelleerimiseks, siis nende loodud mudelid jäljendavad, aga ka tõlgivad ja lihtsustavad ning on sellistena ebatäpsed ja paratamatult ka „valetavad". Viitame siinkohal Umberto Eco (1977: 7) kuulsale väitele, et semiootika uurib kõike, mida saab kasutada valetamiseks. Nagu praeguseks on andmete kallutatuse (nt Milan, Treré 2019; Perez 2019) ja andmeõigluse (nt Heeks, Shekhar 2019; Dencik jt 2019) uuringute põhjal teada, siis andmed tõepoolest pahatihti eksitavad, toovad kaasa eri laadi ebaõiglust ja seeläbi ka uusi konflikte. Kultuurisemiootikal põhinevate andmeuuringute eriline hoiak selliste andmesuhete lahkamisel võiks olla, et kui valetamine on nii-öelda paratamatu, kui mis tahes representatsioonid, sealhulgas andmepõhised representatsioonid, on alati valikulised, kuna kujutavad reaalsusest vaid teatud osa, jättes muu kõrvale, siis tuleks keskenduda sellele, kuidas need valikud mõjutavad kultuuridünaamikat kultuuri eri tasanditel – nii mikro-, meso- kui ka makrokontekstides. Mikrokonteksti näitena tähendaks see küsimust sellest, kuidas metaandmete skeemid kindlates andmebaasides võivad mõjutada kasutajate saadavat ülevaadet ajaloosündmustest (Coleman 2020; Ibrus, Ojamaa

2018); mesotasandil võivad andmekalded mõjutada näiteks rahvuskultuuride võrdlusi (Fischer, Poortinga 2018; Erlin jt 2021), makrotasandil aga üleilmseid kokkuvõtteid, milliseid teostab kõnekalt Lev Manovich (2020).

Kultuuriandmete analüüs Tallinna Ülikoolis

Kõiki eespool välja toodud aspekte arvesse võttes on võimalik esitada uudne süsteemne lähenemine kultuuriandmete analüüsile, mis liigub edasi digihumanitaaria problemaatikast, lõpetab nn kahe kultuuri segregatsiooni ning kasutab ära kultuurisemiootika ja kompleksteaduse ühendamise potentsiaali. Nendest kindlaks tehtud sobivustest lähtudes saame välja pakkuda holistliku lähenemiste skaala, lastes ühelt poolt üksikutel teadlastel jääda oma spetsialiseerumise juurde, kuid võimaldades neil teiselt poolt osaleda süsteemselt loodud võrgustikes, mis püüavad kokku panna mitmetahulist, ent siiski integreeritud ettekujutust semiosfäärilisest dünaamikast.

Tallinna Ülikooli ERA (European Research Area) õppetooli projektis CUDAN (Cultural Data Analytics, vt Cudan Open Lab) rakendatakse seda visiooni praegu praktikas. Tuginedes osaliselt varasematele uurimistöödele,[2] on eesmärk panna alus semiosfääri süsteemsele uurimisele, mis võib kultuurikäsitust muuta sama põhjalikult, kui süsteemibioloogia muutis arusaama elust. Näitena praegusest interdistsiplinaarsest tööst hõlmab CUDAN-i asjatundmine kultuurisemiootikat, kunstiajalugu, kultuuriajalugu, meediauuringuid, poliitikauuringuid, loomemajanduse uuringuid, haridustehnoloogiat, evolutsioonilist majandusteadust, võrguteadust, arvutuslikku sotsiaalteadust, arvutilingvistikat, masinõpet, kasutajakogemuse disaini, kultuurifüüsikat ja generatiivset kunsti. Tuues kokku sedavõrd erisugused meetodid ja uurimissuunad, kasutavad CUDAN-i teadlased kirevat allikmaterjali, mis hõlmab muu hulgas audiovisuaalseid, pildi-, teksti- ja arvandmeid, nii mittestruktureeritud kui ka struktureeritud andmeid andmebaaside ja graafide kujul, mille hulka kuuluvad avaandmed ning andmed avaliku ja erasektori partneritelt. Selle ettevõtmisega kaasnevad ilmselgelt riskid, aga ka suured võimalused. Raskus ja samas väga suur potentsiaal seisneb ühise süstemaatilise lähenemise loomises. Ent metodoloogilist riski on võimalik maandada, kui eri teadlased saavad pakkuda oma valdkonna põhjalikku tundmist ja samal ajal on neil võimalik valdkonnaülese koostöö toel minna kaugemale oma eriala piiridest.

Pakume välja, et sellist valdkonnaülest tööd tehes võib Eesti võtta rahvusvahelises teadusmaailmas süstemaatilise kultuuriteaduse juhtrolli, tuginedes seejuures nn Eesti teooria rikkalikule traditsioonile (Tamm, Kull 2016). See potentsiaal sarnaneb süsteemibioloogia kujunemisega, kus samaväärne kvalitatiivse ja kvantitatiivse uuri-

[2] Kõige olulisema eelkäija ja eeskujuna võib välja tuua kultuurianalüütika (ingl *cultural analytics*) uurimissuuna, millele Lev Manovich pani aluse 2007. aastal (vt ülevaatlikult Manovich 2020). Kui kultuurianalüütika rõhutab eelkõige koostööd arvutiteadustega ja panustab erinevatele arvutuslikele meetoditele, siis kultuuriandmete analüüs peab selle kõrval oluliseks kultuurisemiootilist lähenemist, komplekssete võrgustike analüüsi ja samuti koostööd loodusteadustega (eeskätt süsteemibioloogiaga).

mise lõimimine viis edusammudeni eluprotsesside mõistmises ja tõi ühiskonnale kestvat kasu. Samamoodi võib meie uudne lähenemine kultuuriandmete analüüsile õige suunamise korral tasapisi jõudu koguda, olles lõpuks midagi enamat kui uudse terminoloogia põhine moeröögatus või lühiajaliselt rahastatud projekti piiratud tulem. Väidame, et õige tee on senisest integreeritum lähenemine kultuurianalüüsile, mis seisab kõrgemal kunstlikust nn kahe kultuuri lahutamisest. Meie lähenemise teoreetiline alus on kultuurisemiootika seni veel ebapiisavalt rakendatud potentsiaal (Hartley jt 2021a; Tamm, Torop, ilmumas) ja kultuuri kompleksteadus (Mitchell 2009; Schich jt 2014; Complexity Explained 2019). See tugineb ka nn kultuuriteaduse lähenemisele (Hartley, Potts 2014; Hartley 2020), kus on juba sarnase teoreetilise integreerimise vallas tööd tehtud. Empiiriliselt toetab meie lähenemist kasvav teadustööde hulk, mis ühendab kultuuriajaloo, võrgustike, kõrgema astme topoloogia ja arvutuslike meetodite võimalusi (Schich 2019).

Alates projekti CUDAN käivitamisest on kultuuriandmete analüüsi võtnud kasutusele üha suurem hulk maailma juhtivaid õppeasutusi. 2021. aastal korraldavad Carnegie Melloni ülikool ja Santa Fe instituut kultuuriandmete analüüsi kursuse. Ludwig Maximiliani ülikool Münchenis võttis selle nimetuse kasutusele uue teadusgrupi jaoks. California Berkeley ülikoolis on kultuuriandmete analüüs oluline andmeteadustesse tehtavate investeeringute osa. Kultuuriandmete analüüsi laadne teadustöö toimub samuti sellistes infoteaduste õppeasutustes nagu Indiana ülikool Bloomingtonis, Network Science Institute of Northeastern University Bostonis, Northwestern Institute on Complex Systems of Northwestern University Chicago lähedal ning KAIST Graduate School of Culture Technology Lõuna-Koreas. Kõigis neis ja mujalgi kasvab kultuuriandmete analüüs jõudsalt. Tallinna Ülikooli CUDAN-i Open Lab'i eesmärk on osaleda selles teedrajavas liikumises.

Artikli kirjutamist toetasid Eesti Teadusagentuuri grandid PRG1276 „Digitaalne Liivimaa: keskaja Liivimaa digiajalooline analüüs (u. 1200–1550)" ja PRG1191 „Kultuuri avaandmelahenduste avaliku väärtuse ahelad" ning Euroopa Liidu Horisont 2020 projekt CUDAN (grandi number 810961).

VEEBIVARAD

Complexity Explained. Manlio De Domenico, D. Brockmann, C. Camargo, C. Gershenson, D. Goldsmith, S. Jeschonnek, L. Kay, S. Nichele, J. R. Nicolás, T. Schmickl, M. Stella, J. Brandoff, A. J. Martínez Salinas, H. Sayama. 27. VI 2019. https://doi.org/10.17605/OSF.IO/TQGNW

Cudan Open Lab. https://cudan.tlu.ee

Eesti 3D kaksiku uuring. https://eehitus.ee/timeline-post/3d-kaksiku-uuring

What is Digital Humanities? https://whatisdigitalhumanities.com

KIRJANDUS

Ahnert, Ruth; Ahnert, Sebastian E.; Coleman, Catherine Nicole; Weingart, Scott B. 2021. The Network Turn: Changing Perspectives in the Humanities. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108866804

Anderson, Philip W. 1972. More is different. – Science, kd 177, nr 4047, lk 393–396. https://doi.org/10.1126/science.177.4047.393

Balkun, Mary McAleer; Deyrup, Marta Mestrovic (toim) 2020. Transformative Digital Humanities: Challenges and Opportunities. London–New York: Routledge. https://doi.org/10.4324/9780429399923

Barabási, Albert-László 2016. Network Science. Cambridge: Cambridge University Press.

Barthelemy, Marc 2019. The statistical physics of cities. – Nature Reviews Physics, kd 1, nr 6, lk 406–415. https://doi.org/10.1038/s42254-019-0054-2

Benington, John; Moore, Mark H. 2011. Public Value: Theory and Practice. London: Palgrave Macmillan.

Berry, David M. 2014. Critical Theory and the Digital. London: Bloomsbury Academic.

Berry, David M.; Fagerjord, Anders 2017. Digital Humanities: Knowledge and Critique in a Digital Age. Cambridge: Polity.

Bertalanffy, Ludwig von 1951. General system theory: A new approach to unity of science. – Human Biology, kd 23, nr 4, lk 302–312.

Bod, Rens 2014. A New History of the Humanities. Oxford: Oxford University Press.

Cassirer, Ernst 2000 [1910]. Substanzbegriff und Funktionsbegriff. – E. Cassirer, Gesammelte Werke. Toim Birgit Recki. Hamburg: Felix Meiner Verlag.

Coleman, Catherine Nicole 2020. Managing bias when library collections become data. – International Journal of Librarianship, kd 5, nr 1, lk 8–19. https://doi.org/10.23974/ijol.2020.vol5.1.162

Couldry, Nick; Mejias, Ulises A. 2019. The Costs of Connection: How Data is Colonizing Human Life and Appropriating for Capitalism. Stanford: Stanford University Press. https://doi.org/10.1515/9781503609754

Dencik, Lina; Hintz, Arne; Redden, Joanna; Treré, Emiliano 2019. Exploring data justice: Conceptions, applications and directions. – Information, Communication & Society, kd 22, nr 7, lk 873–881. https://doi.org/10.1080/1369118X.2019.1606268

Dilthey, Wilhelm 1959 [1883]. Gesammelte Schriften. Kd 1. Stuttgart: Teubner.

Dilthey, Wilhelm 1997. Ajaloolise maailma ülesehitus vaimuteadustes. – Filosoofilise hermeneutika klassikat. (Avatud Eesti raamat.) Tlk Andrus Tool. Tartu: Ilmamaa, lk 72–83.

Eco, Umberto 1977. A Theory of Semiotics. Bloomington: Indiana University Press.

Engelbrecht, Jüri; Kitt, Robert 2020. Komplekssüsteemid ehk Tervik on suurem kui osade summa. Tallinn: Postimees Kirjastus.

Erlin, Matt; Piper, Andrew; Knox, Douglas; Pentecost, Stephen; Drouillard, Michaela; Powell, Brian; Townson, Cienna 2021. Cultural capitals: Modeling 'Minor' European literature. – Journal of Cultural Analytics, nr 2, lk 40–73. https://doi.org/10.22148/001c.21182

Fischer, Ronald; Poortinga, Ype H. 2018. Addressing methodological challenges in culture-comparative research. – Journal of Cross-Cultural Psychology, kd 49, nr 5, lk 691–712. https://doi.org/10.1177/0022022117738086

Gold, Matthew K. (toim) 2012. Debates in the Digital Humanities. London–Minneapolis: University of Minnesota Press.

Gold, Matthew K.; Klein, Lauren F. (toim) 2016. Debates in the Digital Humanities 2016. London–Minneapolis: University of Minnesota Press.

Gold, Matthew K.; Klein, Lauren F. (toim) 2019. Debates in the Digital Humanities 2019. London–Minneapolis: University of Minnesota Press.

Hartley, John 2009. From Cultural Studies to Cultural Science. – Cultural Science, kd 2, nr 1, lk 1–16. https://doi.org/10.5334/csci.18

Hartley, John 2020. How We Use Stories and Why That Matters: Cultural Science in Action. New York: Bloomsbury. https://doi.org/10.5040/9781501351662

Hartley, John; Ibrus, Indrek; Ojamaa, Maarja 2021a. On the Digital Semiosphere: Culture, Media and Science for the Anthropocene. London: Bloomsbury. https://doi.org/10.5040/9781501369209

Hartley, John; Ibrus, Indrek; Ojamaa, Maarja 2021b. Emergent self-mediating classes in the digital semiosphere: Coronavirus conspiracies and climate justice movement. – Nordic Journal of Cultural Studies, kd 3, nr 1, lk 118–140. https://doi.org/10.2478/njms-2021-0007

Hartley, John; Potts, Jason 2014. Cultural Science: A Natural History of Stories, Demes, Knowledge and Innovation. London: Bloomsbury.

Heeks, Richard; Shekhar, Satyarupa 2019. Datafication, development and marginalised urban communities: An applied data justice framework. – Information, Communication & Society, kd 22, nr 7, lk 992–1011. https://doi.org/10.1080/1369118X.2019.1599039

Hexter, Jack H. 1972. Fernand Braudel and the Monde Braudellien... – Journal of Modern History, kd 44, nr 4, lk 480–539. https://doi.org/10.1086/240839

Huxley, T. 1869. Nature: Aphorisms by Goethe. – Nature, kd 1, lk 9–11. https://doi.org/10.1038/001009a0

Ibrus, Indrek; Ojamaa, Maarja 2018. Newsreels versus newspapers versus metadata: A comparative study of metadata modelling the 1930s in Estonia. – VIEW: Journal of European Television History and Culture, kd 7, nr 14, lk 123–137. https://doi.org/10.18146/2213-0969.2018.jethc157

Ibrus, Indrek; Ojamaa, Maarja 2020. The creativity of digital (audiovisual) archives: A dialogue between media archaeology and cultural semiotics. – Theory, Culture & Society, kd 37, nr 3, lk 49–70. https://doi.org/10.1177/0263276419871646

Isaak, Jim; Hanna, Mina J. 2018. User data privacy: Facebook, Cambridge Analytica, and privacy protection. – Computer, kd 51, nr 8, lk 56–59. https://doi.org/10.1109/MC.2018.3191268

Jacobs, Jane 1961. The Death and Life of Great American Cities. New York: Vintage Books.

Jones, Steven E. 2014. The Emergence of the Digital Humanities. London–New York: Routledge.

Kirschenbaum, Matthew G. 2012. Digital humanities as/is a tactical turn. – Debates in the Digital Humanities. Toim Matthew K. Gold. Minneapolis: University of Minnesota Press, lk 415–428. https://doi.org/10.5749/minnesota/9780816677948.003.0043

Kittler, Friedrich A. 1994. Wenn die Freiheit wirklich existiert, dann soll sie doch ausbrechen. – Am Ende vorbei. Toim Rudolf Maresch. Wien: Turia & Kant, lk 95–129.

Kull, Kalevi 1998. On semiosis, Umwelt, and semiosphere. – Semiotica, kd 120, nr 3–4, lk 299–310.

Kull, Kalevi; Lotman, Mihhail 1995. Semiotica Tartuensis: Jakob von Uexküll ja Juri Lotman. – Akadeemia, nr 12, lk 2467–2483.

Kull, Kalevi; Lotman, Mihhail 2012. Semiotica Tartuensis: Jakob von Uexküll and Juri Lotman. – Chinese Semiotic Studies, kd 6, nr 1, lk 312–323. https://doi.org/10.1515/css-2012-0122

Lee, Minjin; Barbosa, Hugo; Youn, Hyejin; Holme, Petter; Ghoshal, Gourab 2017. Morphology of travel routes and the organization of cities. – Nature Communications, kd 8, nr 1, lk 1–10. https://doi.org/10.1038/s41467-017-02374-7

Mandelbrot, Benoit 1982. The Fractal Geometry of Nature. Kd 1. New York: WH Freeman.

Manovich, Lev 2020. Cultural Analytics. Cambridge, Massachusetts: MIT Press. https://doi.org/10.7551/mitpress/11214.001.0001

Mazzucato, Mariana 2018. The Value of Everything: Making and Taking in the Global Economy. London: Allen Lane.

Milan, Stefania; Treré, Emiliano 2019. Big Data from the South(s): Beyond data universalism. – Television & New Media, kd 20, nr 4, lk 319–335. https://doi.org/10.1177/1527476419837739

Mitchell, Melanie 2009. Complexity: A Guided Tour. Oxford: Oxford University Press.

Perez, Caroline Criado 2019. Invisible Women: Exposing Data Bias in a World Designed for Men. London: Chatto & Windus.

Pressner, Todd; Schnapp, Jeffrey 2009. The Digital Humanities Manifesto 2.0. http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20

Prigogine, Ilya; Stengers, Isabelle 2018 [1984]. Order Out of Chaos: Man's New Dialogue with Nature. London: Verso Books.

Ravenstein, Ernest George 1885. The laws of migration. – Journal of the Statistical Society of London, kd 48, nr 2, lk 167–235. https://doi.org/10.2307/2979181

Riedl, Rupert 2000. Strukturen der Komplexität: Eine Morphologie des Erkennens und Erklärens. Berlin: Springer. https://doi.org/10.1007/978-3-642-56946-3

Schich, Maximilian 2016. Figuring out art history. – International Journal for Digital Art History, nr 2, lk 41–67.

Schich, Maximilian 2019. Cultural Analysis Situs. ART-Dok eprint.

Schich, Maximilian; Song, Chaoming; Ahn, Yong-Yeol; Mirsky, Alexander; Martino, Mauro; Barabási, Albert-László; Helbing, Dirk 2014. A network framework of cultural history. – Science, kd 345, nr 6196, lk 558–562. https://doi.org/10.1126/science.1240064

Schich, Maximilian; Meirelles, Isabel; Malina, Roger (toim) 2016. Arts, Humanities, and Complex Networks. 5. vlj [Kindle Edition]. Cambridge: MIT Press.

Schreibman, Susan; Siemens, Ray; Unsworth, John (toim) 2004. A Companion to Digital Humanities. Malden, MA–Oxford: Blackwell.

Schreibman, Susan; Siemens, Ray; Unsworth, John (toim) 2016. A New Companion to Digital Humanities. Malden, MA–Oxford: Wiley. https://doi.org/10.1002/9781118680605

Shuttleworth, Alan 1966. Two Working Papers in Cultural Studies – A Humane Culture and Max Weber and the „Cultural Sciences". Birmingham: Centre for Contemporary Cultural Studies, University of Birmingham.

Smithies, James 2017. The Digital Humanities and the Digital Modern. London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-49944-8

Snow, Charles Percy 2017 [1961]. Kaks kultuuri ja teadusrevolutsioon. – C. P. Snow, Kaks kultuuri ja teadusrevolutsioon. Jerome Kagan, Kolm kultuuri. Loodusteadused, sotsiaalteadused ja humanitaarteadused 21. sajandil. (Avatud Eesti raamat.) Tlk Mart Trummal. Tartu: Tartu Ülikooli Kirjastus, lk 11–56.

Svensson, Patrik 2012. Beyond the Big Tent. – Debates in the Digital Humanities. Toim Matthew K. Gold. Minneapolis: University of Minnesota Press, lk 36–49. https://doi.org/10.5749/minnesota/9780816677948.003.0004

Zinoviev, Dmitry 2016. The pain of complexity. – Leonardo, kd 49, nr 5, lk 450. https://doi.org/10.1162/LEON_a_01271

Tamm, Marek; Kull, Kalevi 2016. Toward a reterritorialisation of cultural theory: Estonian theory from Baer via Uexküll to Lotman. – History of the Human Sciences, kd 29, nr 1, lk 75–98. https://doi.org/10.1177/0952695115617407

Tamm, Marek; Torop, Peeter (toim) (ilmumas). The Companion to Juri Lotman: A Semiotic Theory of Culture. London: Bloomsbury.

Terras, Melissa; Nyhan, Julianne; Vanhoutte, Edward (toim) 2013. Defining Digital Humanities: A Reader. Farnham: Ashgate.

Torop, Peeter 2008. Translation as communication and auto-communication. – Sign Systems Studies, kd 36, nr 2, lk 375–397. https://doi.org/10.12697/SSS.2008.36.2.06

Torop, Peeter 2015. Kultuuriteooria kui kultuuri eneseteadvus. – Akadeemia, nr 4, lk 626–649.

Uexküll Jakob von 1909. Umwelt und Innenwelt der Tiere. Berlin: J. Springer.

Uexküll, Jakob von 2012. Omailmad. (Eesti mõttelugu 105.) Koost Kalevi Kull, Riin Magnus. Tlk Mari Tarvas, Krista Räni. Tartu: Ilmamaa.

Weaver, Warren 1948. Science and complexity. – American Scientist, kd 36, lk 536–567.

Weingart, Scott B.; Eichmann-Kalwara, Nickoal 2017. What's under the Big Tent?: A study of ADHO conference abstracts. – Digital Studies/Le Champ Numérique, kd 7, nr 1. https://doi.org/10.16995/dscn.284

Williams, Raymond 1974. Communications as cultural science. – Journal of Communication, kd 24, nr 3, lk 17–26. https://doi.org/10.1111/j.1460-2466.1974.tb00385.x

Windelband, Wilhelm 1993 [1894]. Ajalugu ja loodusteadus. Tlk Marju Luts. – Akadeemia, nr 3, lk 490–505.

Winthrop-Young, Geoffrey 2000. Silicon sociology, or, Two kings on Hegel's throne? Kittler, Luhmann, and posthuman merger of German media theory. – The Yale Journal of Criticism, kd 13, nr 2, lk 391–420. https://doi.org/10.1353/yale.2000.0027

Indrek Ibrus (snd 1974), PhD, Tallinna Ülikooli Balti filmi, meedia ja kunstide instituudi meediainnovatsiooni professor (Narva mnt 27, 10120 Tallinn), indrek.ibrus@tlu.ee

Maximilian Schich (snd 1974), PhD, Tallinna Ülikooli Balti filmi, meedia ja kunstide instituudi kultuuriandmete analüüsi sihtrahastusega professor (Narva mnt 27, 10120 Tallinn), mxs@tlu.ee

Marek Tamm (snd 1973), PhD, Tallinna Ülikooli humanitaarteaduste instituudi kultuuriajaloo professor (Narva mnt 29, 10120 Tallinn), marek.tamm@tlu.ee

From digital humanities to cultural data analytics

Keywords: cultural data analytics, cultural analytics, digital humanities, complexity, semiosphere, cultural semiotics, cultural science

Digital Humanities (DH) has been a success story in the academic world over the last twenty years which has opened many new vistas in the humanities research. However, we argue that it is better to consider DH as a transitory phenomenon that needs to be developed into more specific research fields in order to overcome some fundamental shortcomings of DH. First, it is vital to transcend the division between qualitative inquiry into ideographic phenomena and quantification of nomothetic phenomena. This specifically includes the emergence and analysis of organized complexity, which emerges right in between. Secondly, it is important to surpass the dichotomy of singular and general as research objects and replace this by the notion of semiosphere as a research object, defined as the "smallest" functioning element of culture by Juri Lotman. From this perspective, the singular cultural unit is always conditioned by the whole of the semiosphere, but the whole can be always changed by the singular unit. Thirdly, the denominator "digital" is questionable and needs to be left behind, as there are many other forms of computation, which can be analog and/or digital. Indeed, alternative forms of non-digital computation are already seeing a comeback. Fourthly, the label of "humanities" is at the same time too large and too restrictive. We would argue for a more specific study of meaning making practices in human society, yet without confining ourselves to traditional humanities scholarship but learning from the new developments in system biology, evolutionary economics, complexity science and many more. We propose to call this new transdisciplinary field of study "cultural data analytics", aiming to explore the dynamics of meaning making practices by computational means and looking at a spectrum of materials (textual, sonic, visual, multimodal, etc.) both regarding the longue durée and in real time, if not anticipating the future.

Indrek Ibrus (b. 1974), PhD, Tallinn University, Baltic Film, Media, and Arts School, Professor of Media Innovation (Narva mnt 27, 10120 Tallinn), indrek.ibrus@tlu.ee

Maximilian Schich (b. 1974), PhD, Tallinn University, Baltic Film, Media, and Arts School, Research Professor of Cultural Data Analytics (ERA Chair) (Narva mnt 27, 10120 Tallinn), mxs@tlu.ee

Marek Tamm (b. 1973), PhD, Tallinn University, School of Humanities, Professor of Cultural History (Narva mnt 29, 10120 Tallinn), marek.tamm@tlu.ee

https://doi.org/10.54013/kk764a3

Erinevused, kaugused ja sõrmejäljed

Stilomeetria ja mitmemõõtmelise tekstianalüüsi alused

ARTJOMS ŠEĻA

Arvutuslikud meetodid ja töövahendid on viimasel ajal jõudsalt humanitaarteadustesse levinud. See on taaskäivitanud arutelu uurimisobjekti ning selle matemaatilise esituse või analüüsis kasutatava mudeli vahelise suhte üle. Kas midagi niivõrd kompleksset nagu romaan saab uurida nii lihtsal moel nagu selles esinevate sõnade loendamine? Ehkki mõned uurijad leiavad, et tänapäevased loomuliku keele töötluse tehnikad on kirjandusteaduse ja selle uurimisobjektide kompleksuse jaoks olemuslikult sobimatud (Da 2019), kasutavad teised siiski tekstide lihtsustatud esitusi, et uurida üldisi kirjandusloolisi ja kirjandusteoreetilisi probleeme, nagu ilukirjandužanride eluiga (Underwood 2019), kirjutamisstiilide muutumine ja püsimine ajas (Hughes jt 2012; Storey, Mimno 2020), poeetilise vormi ja semantika suhe (Šeļa jt 2020) ja palju muud.

Käesolev artikkel näitab, millist kasu võib kirjandusteadlastele ja kirjandusloolastele olla tekstide lihtsakoelisest kvantitatiivsest esitusest. Tegemist on kontseptuaalse sissejuhatusega stilomeetriasse kui arvutusliku tekstianalüüsi alamvaldkonda, mis uurib kvantitatiivselt tekstidevahelisi erinevusi (Holmes, Kardos 2003). Kvantitatiivse tekstiesitusega seonduvate küsimuste tutvustamiseks on stilomeetria hea teema kahel põhjusel. Esiteks on sellel traditsiooniline, lihtne eesmärk: selliste tekstiliste mustrite äratundmine, mida saab kasutada konkreetsetele autoritele omaste kirjutamisstiilide eristamiseks. Teiseks hakati stilomeetrias kvantitatiivsele tõendusmaterjalile tuginema varakult, juba enne arvutite tulekut, mistõttu on sellel rikkalik ajalugu tekstide lihtsate, pindmise tasandi matemaatiliste esituste kasutamisel huvitavate probleemide lahendamiseks. Artikkel on mõeldud lugejale, kelle jaoks teema on uus, ega eelda varasemaid teadmisi arvutuslikest meetoditest. Annan ülevaate mitmemõõtmelise tekstianalüüsi põhimõtetest, libisemata samal ajal üle olulistest mõistetest ja põhimistest matemaatilistest tehetest.

Kuna stilomeetria mõiste ja distsipliini areng on olnud tihedalt seotud autorituvastuse probleemiga, tutvustan alustuseks lühidalt stilomeetria varast ajalugu, et selgitada nii uurimisvaldkonna tausta kui ka selle aluseeldusi, mis puudutavad tekste ja autorsust. Seejärel tutvustan mitmemõõtmelise tekstianalüüsi aluseid ja tekstidevahelise kauguse mõistet kui tekstide erinevuse tähistajat. Erinevuse arvutamisel pööratakse põhitähelepanu sõnade sagedusele: infoküllasele ja hõlpsasti ligipääsetavale tunnusele, mida tänapäevases tekstianalüüsis laialdaselt kasutatakse ning mille põhjal võib saada aimu märksa laiaplaanilisematest tekstilistest korrapäradest kui üksnes individuaalsed autorimustrid. Artikkel päädib autorituvastuse proovikatsetega eesti ilukirjanduse korpuse peal, mis testivad paari peamise muutuja käitumist ning on mõeldud üldise kontseptsiooni toimimise tõestuseks.

Üks probleem, mitu lahendusteed: lühiülevaade stilomeetria ajaloost

Stiili mõiste on ajalooliselt olnud seotud kõne ja tekstiga. Alates Aristotelese „Retoorikast" (kus seda nimetati *lexis*'eks) on stiili mõistega kirjeldatud teksti *kuidas*-mõõdet (stiil, vorm), vastandatuna teksti *millest*-mõõtmele (teema, sisu) (Rapp 2010). Häid mõtteid on võimalik edastada halvas stiilis.

Stiil kui normatiivne mõiste, mis reguleerib väljenduse ja suhtlusega seotud tavasid, on tänapäeval endiselt kasutusel (mõelgem näiteks ametlikule või akadeemilisele kirjutusviisile). Kuid sõna *stiil* tähendus on märksa laialivalguvam: sõna leiab kasutust kõiksugu mitteformaalsetes eriteemalistes aruteludes, alates kirjandusest ja kunstist ning lõpetades riiete, arhitektuuri ja eluviisiga. Sõna kiputakse kasutama, kui adutakse mis tahes erinevusi muidu ühte ja samasse raamistikku (nt detektiivromaanid, ämblikuvõrgud, profiilivaates platvormimängud) paigutatud objektide või teguviiside vahel. Stiilist võib mõelda kui mõistest, mis kontseptuaalselt ühendab teatud hulka objekte selliste sidemetega, mis väljendavad mitmesuguseid sarnasuste ja erinevuste määrasid. Stiilist rääkimine loob just sedalaadi kontekstuaalseid süsteeme. Arhitektuuriajaloolane James S. Ackerman on seda väljendanud nii: „Stiili mõiste on vahend tuvastamaks suhteid üksikute kunstiteoste vahel" (1962: 228). Stilomeetrias tuvastatakse sedalaadi suhteid kvantitatiivsetel skaaladel ja need väljenduvad teksti mõõdetavates omadustes, mida nimetatakse t u n n u s t e k s. Traditsiooniliselt on stilomeetriat rakendatud just üht konkreetset laadi tekstidevahelise suhte, nimelt autorsuse mõistmiseks. Alustan sellest.

Kvantitatiivse autorituvastuse ajalugu kujutatakse tihtipeale lineaarsena, algusega XIX sajandi Ameerika matemaatikas ja statistikas (Holmes 1998; Holmes, Kardos 2003; Koppel jt 2009). Lugu alustatakse tavaliselt matemaatik Augustus de Morgani 1851. aastal kirjutatud kirjast, kus ta esitas hüpoteesi, et tekstinäidiste keskmisi sõnapikkusi saab kasutada näidiste autorite eristamiseks. Mõtet rakendas mõni aastakümme hiljem Ameerika füüsik Thomas Corwin Mendenhall, kes vaatles eri teostest pärit ühepikkuste tekstinäidiste sõnade pikkust (Mendenhall 1887, 1901). Ta ei mõõtnud mitte sõnade keskmist pikkust, vaid sõnapikkuste jaotust (mitu ühetähelist, mitu kahetähelist jne sõna tekstis esineb). XIX sajandi spektroskoopia arengutest inspireerituna kutsus Mendenhall neid jaotusi „kompositsioonikõverateks": just nagu vesinikuaatom on alati äratuntav selle mõjul murdunud valguse spektri järgi, ilmneb ka ühe autori tekstides alati teatud laadi sidusus. Selline mõtteviis, mis positivistlikult tõmbab autorite ja keemiliste elementide vahele paralleele, on tänapäevani stilomeetria ja autorituvastuse nurgakivi. Põhiline eeldus on, et sama autori kirjutatud tekstid on üksteisega sarnasemad kui mis tahes muu tekstiga.

Pärast Mendenhalli teedrajavaid töid arenes stilomeetria käsikäes statistikaga ning järgnenud läbimurded võib enamasti panna silmapaistvate statistikute arvele. George Udny Yule tuletas autorituvastuseks erinevaid mõõdikuid, mis põhinesid lausepikkuste jaotusel, tähemärkide esinemissagedustel ning sagedaste sõnade loenditel (Yule 1939, 1944). Käsitledes artiklikogu „The Federalist Papers", üht kuulsaimat vaieldava autorsuse juhtumit, kasutasid Frederick Mosteller ja David Wallace

mitmete funktsioonisõnade kombineeritud informatsiooni (Mosteller, Wallace 1963). Selle uuringuga toimus oluline nihe mitmemõõtmelise analüüsi suunas ning see on jätkuvalt üks tänapäevase autorituvastuse (ja üldisemalt tekstianalüüsi) iseloomulikumaid näiteid.

Eelkirjeldatud n-ö statistikute liin polnud stilomeetria ajaloos aga ainus. Võttes arvesse ka filoloogilistel teadmistel ja meetoditel põhinevaid humanitaarteaduslikke erialasid, joonistub stilomeetria ajaloos välja harali seisev ja tihtipeale omavaheliste ühendusteta mõttevoolude võrgustik (Grzybek 2014). Lühidalt – XIX sajandi lõpus, mil empiristlikud ja positivistlikud hoiakud teksti ja autorsuse suhtes olid haripunktis, tegutses vähemalt kolm eraldiseisvat kvantitatiivse stilistika haru: 1) varaseim Shakespeare'i-uuringute liin Suurbritannias, 2) filosoofide rühmitused Mandri-Euroopas ja Šotimaal, kes mõõtsid Platoni dialoogide stilistilist varieeruvust, 3) ülalkirjeldatud matemaatikud Ameerika Ühendriikides, kes käsitlesid üldisi autorsuse väljendumise küsimusi.

Shakespeare'iga seonduv on uusajast peale olnud põhiline autorsust puudutav vaidlus läänes ning Shakespeare'i näidendid jäävad omamoodi proovikiviks ka tulevaste põlvkondade tuvastusmeetoditele ja käsitlustele. XVIII sajandil kasvas moderniseerunud filoloogiast välja omaette Shakespeare'i autorsuse tuvastamise traditsioon, mis tugines tekstiliste tunnuste kvantifitseerimisele ja võrdlemisele. Tolle haru oluline uuendus, mis võeti peavoolu stilomeetriasse üle alles hiljuti (Plecháč jt 2018), oli draamatekstide uurimisel tuginemine luulele omastele tunnustele, nagu riimivormid ja meetriline struktuur. Uurimisliin, mille algatas Edward Malone (1787) ning mida jätkasid Henry Weberi ja James Speddingi uurimused (Spedding 1850), tõusis esile pärast New Shakespeare Society asutamist ning seda seostatakse tihti selliste uurijatega nagu Frederick James Furnivall, Frederick Gard Fleay, J. Kells jt (Grzybek 2014: 68–69).

Eraldiseisvalt arenes stiili kvantifitseerimine välja Platoni pärandi uurimisele pühendunud filosoofilistes ringkondades. Mõnda aega tegutses kaks rühmitust, mis esialgu polnud teineteisest teadlikud: Lewis Campbelli (1867) järgijad Šotimaal ning Wilhelm Dittenbergi (1881) järgijad Mandri-Euroopas (peamiselt saksakeelsetel aladel). Üks saksa traditsiooni esindajatest oli Wincenty Lutosławski – Poola päritolu filosoof, kes õppis ja kirjutas oma väitekirja Tartus. Just tema tutvustas Mandri-Euroopa Platoni-uurijatele šotlaste töid ning temalt pärineb üldlevinud arvamuse järgi stilomeetria mõiste (kirjutises „Stilomeetriast", 1897), mille all ta pidas silmas „stilistilise suguluse mõõtmise" meetodit (Grzybek 2014: 70–73; Lutosławski 1897a, 1897b).

Platoni uurijate haru tugines stiili kvantifitseerimisel tihtipeale levinuimate sõnade esinemissagedustele, mille põhjal eristusid robustsed kronoloogilised mustrid ühe autori tekstide piires ning tugevad autorsuse märgid eri autorite tekstide puhul. Meetod jõudis hiljem Venemaale (Morozov 1915) ning mõjutas kaudselt Vene meetrikakoolkonna autorituvastuse käsitlust (Tomaševski 1923; Lotman, Lotman 1986; Šapir 2000). Sarnaselt Shakespeare'i uurijatega kasutas ka Vene koolkond vormilisi värsitunnuseid (nagu riimi omadused, rütmi ja sõna piirid meetrilises mustris), et lahendada luuletekstide autorituvastusega seonduvaid küsimusi.

Printsiibid, mis pärinesid vene meetrikast – varaseimalt täielikult väljakujunenud kvantitatiivse uurimise alalt kirjandusteaduses –, jõudsid iroonilisel kombel hiljem tagasi varauusaegse inglise kirjanduse uuringutesse inglise värsiõpetuse silmapaistva spetsialisti Marina Tarlinskaja (1987) tööde kaudu.

Paljud valdkonnad, mille raames stilomeetria algselt välja kujunes, hakkasid XX sajandil kiirelt arenevast statistikast maha jääma, osalt tolleaegsete erialapiiride ning akadeemias pead tõstva humanitaar- ja loodusteaduste vahelise eraldatuse tõttu. Aastatel 1960–1980 nägid kirjandusteadlased, kirjandusloolased, arvutiteadlased ja statistikud palju vaeva, et koondada aastakümneid lahus püsinud traditsioonid, mis keerlesid ometi ühe ja sama küsimuse ümber: kas üksikute autorite keelemustrites on midagi unikaalset?

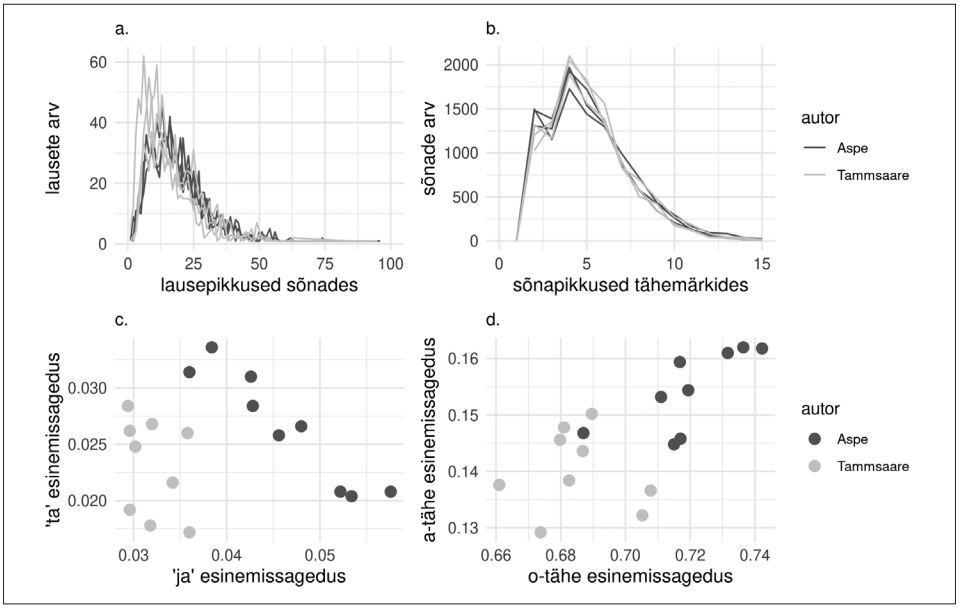
Mitmemõõtmelise stiilikäsituse poole

Varase stilomeetria ajaloos ilmneb tendents, mis võib tänapäeval näida kummaline: miks lahendati stilistiliste erinevustega seotud küsimusi, tuginedes käputäiele mõõdikutele, millel polnud otsest seost sellega, mida üldse tekstiks peetakse? Mida ütleb lause- ja sõnapikkus või grafeemi *o* esinemissagedus Shakespeare'i teoste identiteedi kohta?

Neile küsimustele vastamiseks tasub pöörduda veel ühe XIX sajandil tegutsenud autori isikupärase käekirja uurija poole, seekord mitte tekstide, vaid maalikunsti vallas. Itaalia poliitik ja kunstikriitik Giovanni Morelli avaldas skandaalsed tööd, kus pakkus välja uudse meetodi kunstiteoste autorsuse tuvastamiseks: loobuda tuleks kõige silmatorkavamate ja äratuntavamate osade uurimisest, nagu värvid, kompositsioon, poosid ja stseenid. Need makrotasandi stilistilised tunnused on lihtsasti märgatavad ja pühendunud jäljendajale seetõttu võrdlemisi lihtsasti kopeeritavad. Nende asemel võttis Morelli tähelepanu alla kujutatud figuuride olmelised ja sageli esinevad, pealtnäha ebaolulised pisiasjad, nagu näiteks kõrvade kuju või sõrmed, mida tema väitel kunstnikud maalisid neile ise suuremat tähelepanu pööramata. Nende pisiasjade maalimisel kaalus individuaalne harjumus üles õpitud tehnika ning see võimaldaski eristada algupärase autori kätt võltsija omast. Autorsuse küsimusest sai omamoodi detektiivimõistatus – treenitud silm suutis tuvastada „süüdlase" (autori) näiliselt seosetute pisiasjade ning teistele varjatud või tähenduseta vihjete põhjal (vt Ginzburg 1979).

Varased tekstimõõtmised sarnanesid oma olemuselt Morelli kõrvakujude kataloogiga ning haakusid freudiliku arusaamaga alateadvusest: püüti tabada ilminguid, mis olid varjatud nii publiku kui ka autori enda teadvuse eest. Enne Mostelleri ja Wallace'i uuringut oli stilomeetria püüdnud valdavalt leida ühtainsat n-ö kuldset tunnust, mille abil järjekindlalt paljastada autorite harjumuste erinevusi ning omavahel tugevalt seostada sama autori tekste. See meenutab kohtuekspertiisis kasutatavaid isikutuvastamise tehnikaid, ja tõepoolest – daktüloskoopia ja detektiivkirjanduse populariseerumise taustal muutus autori sõrmejälg stilomeetrilise töö kohta üldlevinud metafooriks.

Kui lühidalt kokku võtta, siis ühtki täiuslikku sõrmejäljelaadset tunnust tekstide puhul leida ei õnnestunud. See ei tähenda, et väljapakutud mõõdikutest üldse mingit abi poleks olnud – nendega sai tabada m õ n i n g a s t stilistilist varieerumist väikses ulatuses. Joonisel 1 esitatud näited kajastavad kahe autori tekstinäidiseid vastavalt teatud tunnuste esinemissagedusele. Edasise statistilise analüüsitagi on selge, et näidiste jaotusi korrastavaks jõuks on autorsus. Sama autori tekstid näivad olevat üksteisele lähemal nii lihtsa kauguse mõttes (kahemõõtmelisel joonisel) kui ka kõverate erinevuse mõttes. Juba pelgalt kahe grafeemi esinemissagedus annab küllaldaselt informatsiooni, et eristada A. H. Tammsaare teksti näidiseid Elisabeth Aspe omadest. Sama kehtib ka ainuüksi kahe kõige sagedasema sõna esinemissageduse kohta (lemmatiseerimata eestikeelsetes tekstides, vt jooniseid 1c ja 1d).



Joonis 1. a. Lausepikkuste jaotus 800-lauselistes näidistes, mis on valitud juhuslikult kolmest Aspe ja kolmest Tammsaare tekstist. **b.** Sõnapikkuste jaotus 800-lauselistes näidistes, mis on valitud juhuslikult kolmest Aspe ja kolmest Tammsaare tekstist. **c.** Kahe kõige sagedasema sõna suhteline sagedus 5000-sõnalistes näidistes, mis on juhuslikult valitud kolmest Aspe ja kolmest Tammsaare tekstist. Igast tekstist on võetud kolm näidist. **d.** Tähtede *a* ja *o* sagedus sõnade arvu suhtes 5000-sõnalistes näidistes, mis on juhuslikult valitud kolmest Aspe ja kolmest Tammsaare tekstist. Igast tektsist on võetud kolm näidist.

Katsed on aga näidanud, et selliste tunnuste kasulikkus kahaneb, kui analüüsi autoreid juurde lisada (Grieve 2007). Keerulisemates liigitamisülesannetes ei anna vähesed tunnused kasvava stilistilise varieeruvuse väljendamiseks enam piisavalt informatsiooni. Seetõttu on tänapäevane stilomeetria võtnud kasutusele mitmemõõtmelise paradigma. Selle asemel et kasutada väikest hulka tugevaid, eristuvaid tunnuseid, võib kasutada kümnete või sadade tunnuste kombineeritud signaali.

Arvutusliku kirjandusteaduse teerajaja John Burrows on väljendanud seda nii: „Suur hulk muutujaid, millest paljud võivad olla nõrgad eristajad, annab peaaegu alati paikapidavamaid tulemusi kui väiksem hulk tugevaid muutujaid. [---] Igatahes moodustub „stilistiline signatuur" tavaliselt paljudest pisikestest pintslitõmmetest." (Burrows 2002: 268)

Burrows pakkus välja tekstidevaheliste sarnasuste arvutamise meetodi, mis hõlmab samaaegselt paljusid muutujaid. Seda tuntakse Burrowsi delta nime all ning sellest on saanud tänapäevase stilomeetria ja autorituvastuse nurgakivi. Meetodi lihtsuse tõttu on tegu hea sissejuhatusega tekstiliste erinevuste kujutamisse kaugustena ja mitmemõõtmelisse tekstianalüüsi üldisemalt.

Burrowsi delta: ülevaade

Üldjoontes väljendab Burrowsi delta kahe teksti vahelist erinevust paarikaupa vaadeldud sõnasageduste kogusummana. Selgitan seda samm-sammult ühe lihtsustatud näite varal. Kaks mitteidentset teksti on kirjutatud keeles, milles on ainult viis sõna: *viuh*, *kilpkonn*, *vapper*, *ninja* ja *põmaki*. Tekstid on erineva pikkusega, seega alustuseks tuletan sõnade suhtelised sagedused (protsentide kujul), et väljendada osakaalu, mis on igal sõnal ühes või teises tekstis (tabel 1).

Tabel 1. Viie sõna suhteline sagedus kahes eri pikkusega tekstis, mis on kirjutatud viiesõnalises keeles.

Sõna	**Tekst 1**	**Tekst 2**
viuh	0,19	0,14
kilpkonn	0,32	0,19
vapper	0,06	0,45
ninja	0,19	0,17
põmaki	0,23	0,04

Nõnda saab mõlemast tekstist arvude kogum, mida on võimalik väljendada ühtsel kujul:

$T1 = (0{,}19; 0{,}32; 0{,}06; 0{,}19; 0{,}23)$
$T2 = (0{,}14; 0{,}19; 0{,}45; 0{,}17; 0{,}04)$

Tegemist on kahe järjestatud kogumiga: väärtused, mis asuvad järjestuses esimesel kohal, vastavad alati sõna *viuh* sagedusele ja nii edasi. Matemaatikas nimetatakse selliseid kogumeid koordinaatvektoriteks või *n*-korteežideks, kus muutuja *n* tähistab elementide arvu vektoris ehk mõõtmeid. Joonisel 1c on igal tekstil kaks koordinaati (*ja* esinemissagedus, *ta* esinemissagedus) ning seda on võimalik esitada kahemõõtmelisel tasapinnal. Inimajul on keeruline kujutleda koordinaatpunkti viies mõõtmes, kuna me sõltume praktiliselt kõiges pelgalt kolmemõõtmelisest tege-

likkuse mudelist. Ent matemaatilised tehted on *n*-mõõtmeliste vektorite puhul ühed ja samad, sõltumata sellest, kas *n*-i väärtuseks on 3 või 300 tunnust.

Kuidas väljendada kahe arvudekogumi vahelist erinevust? Lihtsaim tehe, mida Burrows selleks kasutas, oli lahutamine: ühe teksti vektor lahutatakse teise teksti vektorist. Lineaaralgebras tähendab see, et *T*1 vektori iga element lahutatakse vastavast *T*2 vektori elemendist:

$$T1 - T2 = (\text{viuh}_{T1} - \text{viuh}_{T2}; \text{kilpkonn}_{T1} - \text{kilpkonn}_{T2}; \text{vapper}_{T1} - \text{vapper}_{T2}; \text{ninja}_{T1} - \text{ninja}_{T2}; \text{põmaki}_{T1} - \text{põmaki}_{T2})$$

$$T1 - T2 = (0{,}19 - 0{,}14; 0{,}32 - 0{,}19; 0{,}06 - 0{,}45; 0{,}19 - 0{,}17; 0{,}23 - 0{,}04)$$

$$T1 - T2 = |(0{,}05; 0{,}13; -0{,}39; 0{,}02; 0{,}19)|$$

Tulemuseks saadud arvude kogum on kahe teksti paarikaupa vaadeldud sõnasageduste erinevuste vektor. Selles on viis eri väärtust, millest mõni on negatiivne, ent erinevuse märki ignoreeritakse: pole oluline, kummas suunas erinevusega on tegu, loeb vaid selle suurus. Pole vahet, kas lahutame *T*1 *T*2-st või *T*2 *T*1-st – tulemus kirjeldab sedasama sümmeetrilist suhet.

Arvutuse viimane samm on samuti äärmiselt intuitiivne. Vektoris kajastatud mitmete tunnuste vahelised absoluutsed erinevused võib kokku liita, et väljendada kombineeritud erinevust. See summa annab delta skoori, mis väljendabki kahe teksti vahelist erinevust (Burrows skaleeris lõpptulemust veel kasutatud tunnuste arvuga, kuid see teeb väärtused üksnes väiksemaks ega muuda suhet kui sellist).

$$\triangle(T1, T2) = (0{,}05 + 0{,}13 + 0{,}39 + 0{,}02 + 0{,}19)$$
$$\triangle(T1, T2) = 0{,}78$$

Selle tulemusega üksi pole palju peale hakata, sest puudu on kõige olulisem element stilistikas – kontekst. Kahe asja vahelist suhet ei saa paika panna ilma kolmandat asja mängu toomata. Kui tekste on vähemalt kolm, siis kerkib esile kolmesuunaline suhe: kuivõrd erinev on tekst *T*1 tekst *T*2-st, arvestades erinevusi *T*1 ja *T*3 vahel ning *T*3 ja *T*2 vahel?

Korpus, kontekst ja sageduste standardiseerimine

Eeltoodud näide on lihtsustatud ja tehnilises mõttes ebakorrektne. Üks võtmetähtsusega samm, mille Burrows delta skoori arvutamisel tegi, oli tunnuste standardiseerimine ehk *z*-skooride arvutamine. Viimased väljendavad seda, kui palju iga tunnuse (sõnasageduse) kasutus kõnealuses tekstis erineb korpuse üldisest keskmisest.

Tabel 2. Suhtelised *vs.* standardiseeritud sagedused. *Z*-skoorid on võrreldavas mõõdus ja süsteemsed erinevused autorite vahel on hõlpsasti loetavad (vt esiletõstetud väärtusi).

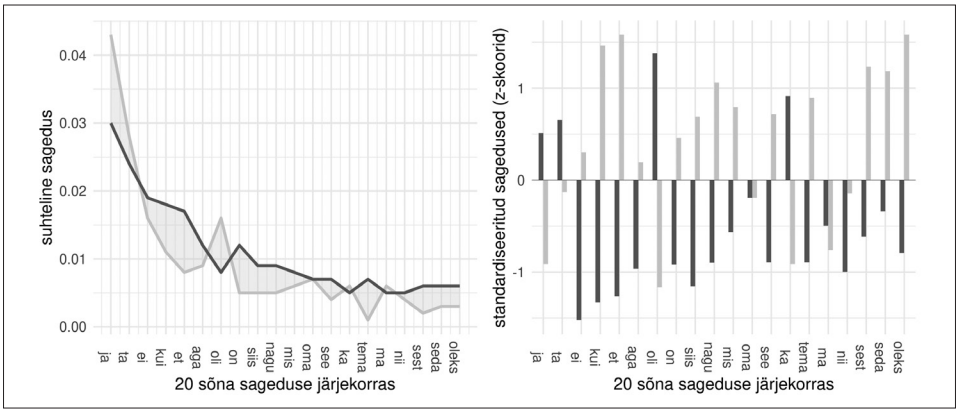
Tekst	Suhtelised sagedused					Standardiseeritud sagedused (*z*-skoorid)				
	ja	**ta**	**ei**	**kui**	**et**	**ja**	**ta**	**ei**	**kui**	**et**
Aspe „Aastate pärast”	0,054	0,021	0,019	0,012	0,011	1,715	–0,721	0,304	**–0,932**	**–0,316**
Aspe „Ennosaare Ain”	0,043	0,028	0,016	0,011	0,008	0,511	0,655	–1,521	**–1,332**	**–1,265**
Aspe „Kasuõde”	0,039	0,033	0,020	0,015	0,010	0,073	1,639	0,913	0,266	**–0,632**
Tammsaare „Kärbes”	0,32	0,023	0,02	0,015	0,012	**–0,693**	**–0,328**	0,913	0,266	0
Tammsaare „Kõrboja peremees”	0,3	0,024	0,019	0,018	0,017	**–0,912**	**–0,131**	0,304	1,465	1,581
Tammsaare „Tõde ja õigus”	0,032	0,019	0,017	0,015	0,014	**–0,693**	**–1,114**	–0,913	0,266	0,632

Tabelis 2 vasakul on näha Tammsaare ja Aspe vastavate terviktekstide viie kõige sagedasema sõna suhtelised sagedused. Paremal on samade tunnuste standardiseeritud skoorid. Need väärtused väljendavad erinevusi iga tunnuse kasutuses kogu korpuse taustal ning on skaleeritud standardhälbe võrra, seega väärtus nagu –1,521 tähendab, et sõna sagedus antud tekstis on 1,521 standardhälbe võrra madalam korpuse keskmisest. *Z*-skoori kujuline sõnasageduste vektor on tegelik arvuline esitus, mida delta jaoks kasutatakse (mis tähendab, et viimane on täiskujul välja öelduna paarikaupa vaadeldud sõnade hälvete erinevuste kogusumma korpuses).

Miks on see tehe tarvilik ning miks peetakse seda tõhusa deltapõhise autorituvastuse jaoks üliolulisek (Evert jt 2017)? Vastus peitub viisis, kuidas sõnade sagedused loomulikus keeles jaotuvad. Iga küllaldaselt suure sõnade arvuga tekst allub nn Zipfi seadusele: sõnade sagedus langeb võrdeliselt nende kohaga sagedusloendis (Zipf 1949). See ei kõla ehk intuitiivselt, aga praktikas tähendab see, et sõnade jaotus on keeles äärmiselt ebavõrdne: leidub väike hulk sõnu, mis on väga sagedased (nn funktsioonisõnad), samal ajal kui valdav osa sõnu on harvad, paiknedes kusagil jaotuse pikas sabas.

Delta meetodit kasutatakse sageli sadade ja tuhandete sõnadega, mis on järjestatud nende sageduse alusel (ingl *most frequent words,* MFW). Sageduse põhjal järjestatud sõnadest alustatakse sellepärast, et kõige sagedasemad neist – artiklid, sidesõnad, eessõnad, asesõnad – võivad oma grammatilise ja süntaktilise funktsiooni kaudu autorite iseloomulike harjumuste kohta nii mõndagi paljastada. Delta põhiprintsiibiks on erinevuste kombineerimine ja siinkohal tekib Zipfi seadusest tulenevalt küllaltki tõsine probleem.

Kui piirduda suhteliste sageduste vaheliste erinevuste loendamisega, siis võib jääda sageduslõksu: suurimad erinevused on kõige sagedasemate sõnade tipuosas ning harvemate sõnade erinevused kahanevad järsult (vt varjutatud piirkonda kahe joone vahel joonisel 2 vasakul). Jaotuse tipp toob mõõtmisse sedavõrd suure



Joonis 2. Erinevused suhtelistes ja standardiseeritud sagedustes Aspe „Kasuõe" ning Tammsaare „Tõe ja õiguse" vahel. Standardiseeritud diagrammi *z*-skoorid arvutati välja terve korpuse suhtes (kuus teksti).

ebavõrdsuse, et erinevused haruldaste sõnade vahel ei saa võrreldaval määral panustada lõplikku erinevuste arvutamisse.

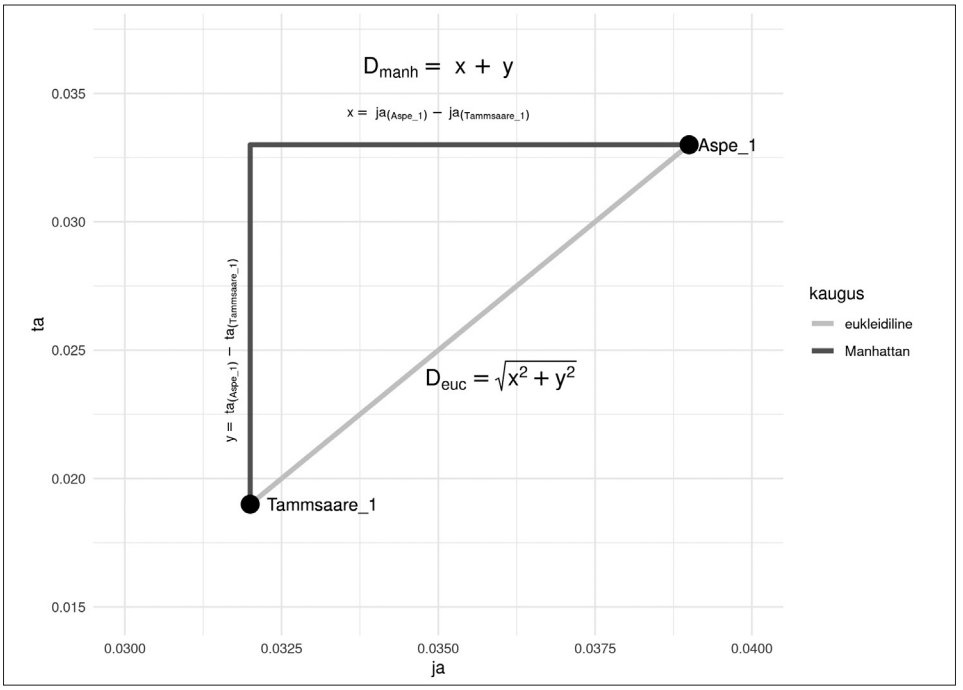
Joonise 2 parempoolsel osal võib näha *z*-skoori kujul esitatud sageduste mõju: kuna nüüd on iga tunnust skaleeritud standardhälbega, on tulemuseks juba paljude tunnuste lõikes võrreldavad erinevusmäärad. Isegi kui sõna esinemissagedus on madal, võib see delta arvutamisse panustada samal (matemaatilisel) alusel nagu väga sagedased sõnad.

Sisuliselt paneb standardiseerimise protseduur delta meetodi sõltuma kontekstist ja korpuse ülesehitusest. Uue teksti lisamine katsesse toob kaasa uue sageduste komplekti, mis muudab keskmiste arvutusi. See mõjutab vastavalt standardhälvet ja *z*-skoore ning seeläbi ka delta lõppväärtusi.

Siit nähtub, kuidas stiili mõistetakse tekstiliste erinevuste mitmemõõtmelises käsitluses: stiil tähendab konkreetse teksti erinevust ümbritsevast korpusest, taustast või n-ö keskmisest keelest (vt lisa Herrmann jt 2015). Iga suhete süsteem, mida stilomeetria käsitleb, sõltub korpusest. Harva on otstarbekas võrrelda varauusaegseid Shakespeare'i ja Marlowe tekste mõne nüüdisaegse novellikoguga. Ainuüksi sellise korpuse ülesehitus, mis toob õigekirjas ja sõnakasutuses sisse suure varieeruvuse, paneb Shakespeare'i ja Marlowe tekstid paistma teineteisega palju sarnasemana, kui nad paistaksid oma kaasaegsete näitekirjanike tööde kontekstis.

Kaugused ja neid kujundavad jõud

Kui matemaatikud Burrowsi delta meetodiga tutvusid (Argamon 2008), tõid nad välja, et Burrows mõõtis sellega intuitiivselt midagi sellist, mida mitmemõõtmelises analüüsis juba vägagi hästi tunti: nimelt kaugusi. Täpsemalt on tegu nõndanimetatud Manhattani või linnakvartali kaugustega. Selle geomeetrilisest põhimõttest saab aimu jooniselt 3.



Joonis 3. Kahe teksti vahelise Manhattani kauguse ja eukleidilise kauguse geomeetriline esitus kahemõõtmelisel diagrammil (*x*- ja *y*-telg tähistavad sõnade *ta* ning *ja* sagedusi).

Vaadates, kuidas punktid kahemõõtmelisel tasapinnal jaotuvad, on enamasti loomulik tajuda nende asukohta üksteise suhtes eukleidilise kaugusena – kahe punkti vahelise sirgjoonena. Manhattani kauguse saamiseks liidetakse seevastu erinevused igas eraldi mõõtmes (ühest punktist teise liigutakse justkui mööda Manhattani tänavaid ja avenüüsid ning kaugust arvestatakse läbitud vahemaa põhjal).

Kauguste mõõtmiseks on märkimisväärne hulk eri viise ning konkreetsetele ülesannetele ja andmetüüpidele vastavaid variatsioone, mis lähtuvad erinevatest geomeetrilistest või informatsiooniteoreetilistest printsiipidest. Paljude stilomeetriliste katsete põhjal on raporteeritud erinevate mõõtmisviiside kõrget efektiivsust autorituvastuse probleemi lahendamisel (nt Smith, Aldridge 2011; Jannidis jt 2015; Kestemont jt 2016; Evert jt 2017), kuid Burrowsi delta on jäänud endiselt lihtsaimaks ja klassikaliseks mõõdikuks, mis seisneb tunnuste standardiseeritud väärtuste põhjal arvutatud Manhattani kauguses. Efektiivne mõõdik on ka koosinus (või koosinusdelta), mis annab „perspektiivse" kauguse: seistes diagrammil koordinaatidel 0,0 ning vaadates mingite punktide (otsekui Manhattani pilvelõhkujate) suunas, arvutatakse nendevaheline kaugus koosinusena vaatesuundi projitseerivate vektorite vahelisest nurgast.

Sümmeetrilistel kaugusemõõdikutel põhinemist võib nimetada objektiivseks lähenemiseks tekstilistele erinevustele. See ei eelda arvutustes mingisugust subjektiivset perspektiivi, kuna teekond A-st B-ni on alati võrdne tagasiteega B-st A-ni (asümmeetriliste mõõdikute kasutamise kohta tekstianalüüsis vt Chang, DeDeo

2020). Heidame tekstid geomeetrilisse ruumi, olles selleks valinud mingi hulga mõõdetavaid tunnuseid, ning jääme lootma, et nad ilmutavad mingit laadi tähenduslikku korrastatust, mis järgib mingit jõudu – klassikalises stilomeetrias konkreetsemalt autorsust. Ilmneb, et tekstid tõepoolest korrastuvad, lausa paljude eri jõudude järgi. Need jõud ilmnevad juba lihtsatest sõnasagedustest, sest sõnade jaotus tekstis on alati seotud keerukate kõrgema tasandi protsessidega.

1. Autorsus. Sama autori kaks eri teksti on muude tingimuste samasuse korral teineteisele lähemal kui mis tahes teise autori tekstile.
2. Kirjandusliigid. Proosa ja luule on silmatorkavalt erinevad ning see erinevus on hõlpsasti tabatav ka kõige sagedasemate sõnade jaotuse põhjal. Meetrikast tulenevad piirangud avaldavad keelelisele süsteemile sedavõrd palju mõju, et luule moodustab proosatekstidega võrreldes peaaegu alati ülimalt isoleeritud eraldiseisva klastri (Storey, Mimno 2020). Suhted ilu- ja muu kirjanduse vahel – mida on püütud Euroopa nüüdiskirjanduse ajaloos aina eristada – on raskemini tabatavad, ent sellegipoolest peegelduvad needki sõnasagedustes (Piper 2017; Underwood 2019).
3. Žanr. Sõnasagedustes peitub piisavalt informatsiooni, et lugeda välja üldisi erinevusi enam väljakujunenud žanride vahel: ulme-, kriminaal-, fantaasiakirjandus jne (Allison jt 2011; Underwood 2017).
4. Kronoloogia. Keel on pidevas muutumises (Croft 2000; Newberry jt 2017). Iga kõnelejate ja kirjutajate põlvkond kasutab keelest versiooni, mis erineb eelmisest õige pisut. Kui arvutada pikast perioodist pärit tuhandete tekstide vahelisi kaugusi, siis ilmneb, et üldiselt on need korrastunud ajavoo alusel: sama ajajärgu tekstid on lähemal üksteisele kui eelnevatele või järgnevatele tekstidele (Jockers 2013; Hughes jt 2012).
5. Sotsiaalsed jõud. Sõnakasutuse põhjal saab ära tunda autori sugu, haridustaset ja sotsiaalset tausta. Paljud uurimused on käsitlenud (enamasti binaarse süsteemina mõistetud) sugude vahelisi stiilierinevusi ja leidnud, et mees- ja naisautoritega saab seostada erinevaid leksikaalseid mustreid (Argamon jt 2009a; Sarawgi jt 2011). Need erinevused pole sugudele kinnistunud ega ole bioloogilise sooga määratud, kuid need on märk olulisest varieeruvusest sotsiaalselt konstrueeritud soorollide lõikes (ehk siis sellest, kuidas mehi ja naisi kultuuriliselt kirjutama õhutatakse) ning naiste ebavõrdsest nähtavalolekust kirjandusväljal. On ka näidatud, et soopõhised vormilised erinevused on aja jooksul kahanenud ja tänapäevastes auhinnatud teostes on sugudevahelisi stiilierinevusi vähem (Koolen 2018).

Stilistilise sõrmejälje metafoor muutub problemaatiliseks, kui püüda sellega järjekindlalt lõpuni minna. Sõrmejälg ei anna ju tuvastamise kõrval isiku kohta sedavõrd palju informatsiooni – sõrmejälgedest ei ilmne tema iseloom, toidueelistused ega popkultuuri huvid. Mitmemõõtmelise stilomeetriaga, mis tugineb samaaegselt paljudele eri tunnustele, võib seevastu tabada selliseid üldisi keelekasutuse mustreid, mis puudutavad palju enamat kui autorsus. Autorisignaali lahtihaakimine teistest mitmemõõtmelisi arvutusi mõjutavatest jõududest pole aga lihtne. Siinkohal tasub veel kord rõhutada korpuse ülesehituse ja katsetingimuste olulisust.

Lähimate naabrite visualiseerimine

Seni on artikkel käsitlenud pigem tekstidevaheliste suhete olemust kui nende suhete põhjal arvutuslikku järeldamist ja otsuse langetamist. Kuidas määrata mõne teksti tõenäolist autorit, kui ette on antud kogum arvulisi väärtusi, mis kirjeldavad kõikide korpuses sisalduvate tekstide vahelisi suhteid?

Paljud autorituvastuse uuringud on ühel või teisel viisil tuginenud (ning tuginevad jätkuvalt) millelegi, mida võib pidada sisuliselt lähima naabri meetodiks. Kui on teatud hulgad väärtusi, mis esindavad autorikandidaate (nt autor A ja autor B), siis milline neist on lähim sihttekstile X? Kui testida ja lihvida nende väärtuste tuletamise meetodit valideerimiskorpuse abil, mille tekstide autorsus on kindel, siis võib eeldada, et väiksema erinevusega arvude hulgad annavad märku stilistilisest vastavusest.

Burrows (2002) arvutas välja delta väärtused (kaugused) sihttekstist kõigi teiste tekstideni, mis sisaldusid tema inglise luule korpuses. Lõppotsuse autorsuse osas langetas ta sihtteksti lähima naabri põhjal. Väikseim delta, lühim vahemaa sihttekstini, annab märku tõenäolisest autorist.

Tekstile autori omistamine lähima naabri põhjal jätab aga varju ülejäänud tekstidevahelised suhted, mida on tohutult. Kui kõik sellised suhted välja arvutada, saab need paigutada ruutmaatriksisse (vt tabelit 3).

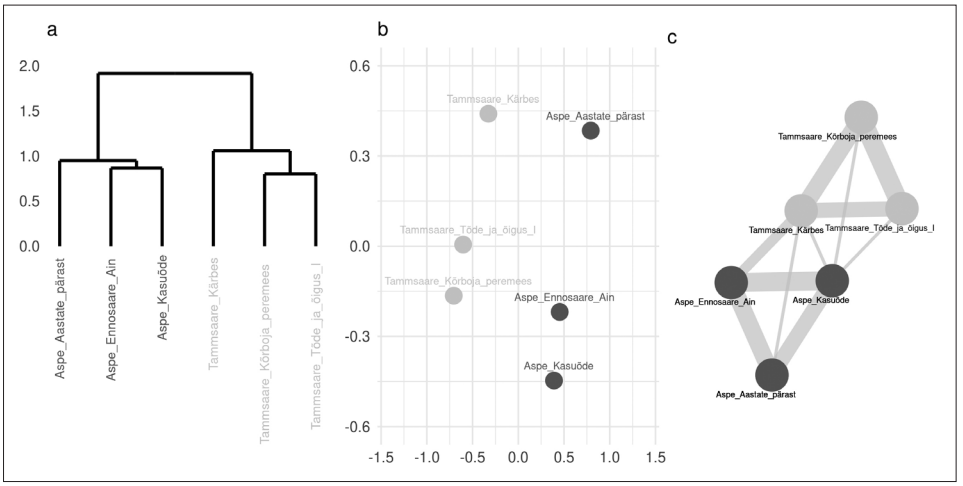
Tabel 3. Kuue teksti vastastikuste kauguste (Burrowsi delta) ruutmaatriks.

Tekst	**Aspe „Aastate pärast”**	**Aspe „Ennosaare Ain”**	**Aspe „Kasuõde”**	**Tammsaare „Kärbes”**	**Tammsaare „Kõrboja peremees”**	**Tammsaare „Tõde ja õigus”**
Aspe „Aastate pärast”	0					
Aspe „Ennosaare Ain”	0,98	0				
Aspe „Kasuõde”	1,02	0,80	0			
Tammsaare „Kärbes”	1,29	1,14	1,22	0		
Tammsaare „Kõrboja peremees”	1,64	1,33	1,27	0,92	0	
Tammsaare „Tõde ja õigus”	1,53	1,31	1,24	0,93	0,87	0

Tabelis 3 on kõik vaatlusalused tekstid esitatud nii maatriksi ridades kui ka veergudes; lahtrid nende ristumiskohtades tähistavad tekstidevahelisi kaugusi. Maatriksil on kaks sümmeetrilist poolt, mida eraldab nullidest koosnev diagonaal (sest kaugus iseendast on 0, st erinevus puudub) ja kogu vajalik informatsioon sisaldub ühes n-ö kolmnurgas. Inimesel on võimalik seda maatriksit lugeda ning selle põhjal tähelepanekuid teha, kuid mida rohkem tekste analüüsi kaasatakse, seda keeru-

lisemaks maatriks muutub. Seetõttu on välja töötatud mitmeid analüüsi- ja visualiseerimismeetodeid, mille abil tuvastatud suhteid mõista ja korrastada. Laiemalt kuuluvad need juhendamata masinõppe meetodite hulka. Juhendamata masinõpe hõlmab tehnikaid, mille abil saab tuvastada andmepunktide vahelisi suhteid ilma eelnevate teadmisteta n-ö klassidest, millesse need punktid kuuluvad. Juhendamata masinõppe meetoditel on oluline roll andmete esmasel uurimisel ning andmestikus toimivate jõudude avastamisel ja teatud mustritesse paigutamisel.

Toon näiteid kolmest stilomeetrias populaarsest meetodist: hierarhiline klasterdamine (joonis 4a), mõõtmete vähendamine (joonis 4b, antud juhul mitmemõõtmelise skaleerimise teel) ning võrgustik (joonis 4c, graaf).



Joonis 4. Kolm näidet kaugussuhete analüüsi ja visualiseerimise meetodite kohta. **a.** Hierarhilisel klasterdamisel põhinev dendrogramm; Wardi meetod klastrite ühendamiseks. **b.** Esialgse maatriksi mitmemõõtmeline skaleerimine, peakoordinaatide analüüs (*Principal Coordinate Analysis*, PCoA). **c.** Iga teksti kolme lähima naabri põhjal moodustatud võrgustik. Joonte paksus väljendab sarnasuse määra.

Kõik kolm meetodit kasutavad infoallikana sedasama kaugusmaatriksit (vt tabelit 3), kuid korrastavad ja teisendavad seda eri viisil. Aglomeratiivne hierarhiline klasterdamine kasutab algoritmi, mis kombineerib andmepunktid sarnasuse põhjal samm-sammult kapselduvatesse hulkadesse – klastritesse –, kuni kõik neist on ühendatud üheksainsaks n-ö hüperklastriks. Harude pikkus vastab antud juhul kaugustele, millelt eri tekstid kokku saavad: sama autori tekstid ühinevad klastriks kõige lühemaid harusid pidi.

Joonisel 4b on tekstid esitatud punktidena kahemõõtmelisel tasapinnal, kasutades selleks mitmemõõtmelist skaleerimist (*Multidimensional Scaling,* MDS). Erinevad MDS algoritmid püüavad paigutada punkte tasapinnale tõhusal viisil, säilitades samal ajal nende kaugustel põhinevad vastastikused suhted. MDS algoritmid kuuluvad mõõtmete vähendamise meetodite laiemasse perekonda, mille otstarve on paljumõõtmeliste andmete teisendamine vähemate mõõtmetega

esituseks. Mõõtmete vähendamist kasutatakse arvutuslikus tekstianalüüsis laialdaselt, kuna tekstid on tihti esitatud rohkete muutujate (näiteks sõnasageduste) kujul. Populaarsed mõõtmete vähendamise algoritmid on veel peakomponentide analüüs (*Principle Component Analysis*, PCA), t-SNE (*t-distributed Stochastic Neighbor Embeddings*) ja UMAP (*Uniform Manifold Projection Maps*). Ehkki need erinevad üksteisest märkimisväärselt, on neil sama eesmärk: säilitada rohkete mõõtmetega andmete omadusi väheste mõõtmetega esituses, mida oleks võimalik kasutada analüüsiks ja pelgalt paaris mõõtmes visualiseerimiseks.

Kaugusmaatriksit võib esitada ka võrgustikuna, kus iga tekst on eraldiseisev sõlm, mille iga ühendus väljendab sõlmedevahelise erinevuse määra. See esitusviis on iseäranis kasulik suure hulga tekstide analüüsimiseks, näiteks uurides suuremastaabilisi stilistilisi mõjusid mõne traditsiooni lõikes, või siis spetsiifiliste tekstiklastrite isoleerimiseks enne põhjalikumat uurimist. See meetod hõlmab kaugusmaatriksis leiduva informatsiooni olulist kärpimist, kuna maatriks kujutab endast sisuliselt üht täielikult ühendatud võrgustiku n-ö klimpi: iga teksti ühendab iga teise tekstiga mingisugune kaugus. Algselt pakuti välja, et kasutada võiks iga teksti üksnes kolme lähimat naabrit ja määrata suhte n-ö skoor selle läheduse alusel (Eder 2015). See skoor annab ühendusele kaalu (mis on joonisel 4c esitatud joone paksusena). Joonisel 4c on kaugusmaatriksi põhjal moodustatud just sedalaadi võrgustik ning ometi väljendab see suuresti sedasama suhete struktuuri, nagu ilmneb ka teiste meetoditega.

Juhendamata masinõpe võimaldab leida tekstilise sarnasuse mustreid ja saada kinnitust aimustele selle kohta, kuidas tekstilised suhted on organiseeritud. Formaalne klasside või sildistuste ennustamine pole aga juhendamata masinõppe paradigmas otseselt võimalik, kuna need meetodid ei ehita ühegi klassi mudelit. Ma ei puuduta siin artiklis juhendatud masinõpet, mis modelleerib klasside eksplitsiitseid omadusi (treeningfaas) ning kasutab neid seejärel selleks, et omistada klass ka seninägemata tekstidele (valideerimis- ja testimisfaas). Sellegipoolest on juhendatud klassifitseerimise strateegia tänapäevases stilomeetrias laialdaselt kasutusel (tugivektorklassifitseerija kasutuse kohta stilomeetrias vt Plecháč jt 2018; klassifikatsiooni- ja hindamismeetodite üldist ülevaadet vt Savoy 2020).

Autorituvastus eesti ilukirjanduses: esmased tähelepanekud

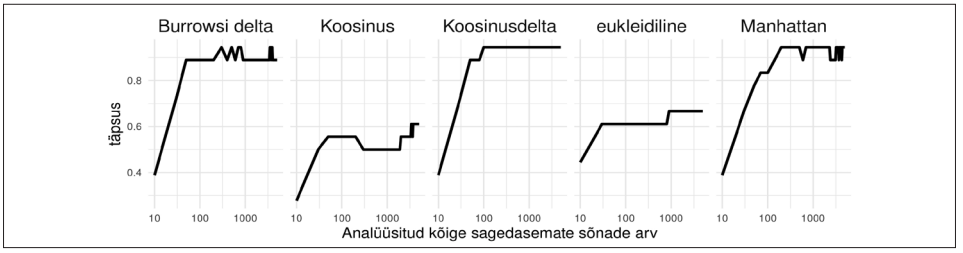
Selleks et demonstreerida kirjeldatud meetodit ning anda aimu selle toimivusest erinevate eesti kirjanduse tekstiliste omaduste puhul, viisin läbi väikese autorituvastuskatse varase eesti ilukirjanduse põhjal. Kasutatud korpus on avatud juurdepääsuga (Uiboaed 2018) ning seda on varem rakendatud stilomeetria ja loomuliku keele töötluse põhialuste näitlikustamiseks (Uiboaed 2017). Tasakaalustasin esialgset tekstivalimit nii, et see sisaldaks võrdse arvu tekste igalt autorilt (kolm juhuslikult valitud teksti kuuelt autorilt: Elisabeth Aspe, Eduard Bornhöhe, Juhan Liiv, Jüri Parijõgi, A. H. Tammsaare, Eduard Vilde). Muid võimalikke korpuse kallutatusi, näiteks žanrit või ajastut, arvesse ei võetud. Lisaks oli korpus esialgu avaldatud nii, et

osa tekste oli standardiseerimata õigekirjas – see aga piirab korpuse kasutust arvutuslikus analüüsis, kuivõrd õigekirjaerinevustel on tugev mõju sagedusloendile ning arhailise ja tänapäevase kirjaviisiga tekstid surutakse loomulikult erinevatesse klastritesse. Standardiseerisin õigekirja minimaalselt, asendades kõik *w*-d *v*-dega, kuid edasine töö korpusega eeldaks suuremat kontrolli.

Katsete põhieesmärk oli üldiselt hinnata autorisignaali äratundmist meetodi selliste üksikute „liikuvate" osade lõikes, mida stilomeetrias traditsiooniliselt testitakse (Juola 2012).

1) Lähima naabri klassifikatsioon, võrdlemaks Burrowsi delta (Manhattani kaugus *z*-skoori kujuliste sagedustega), koosinusdelta (koosinuskaugus *z*-skoori kujuliste sagedustega), lihtsa koosinuskauguse, Manhattani kauguse ja eukleidilise kauguse tulemuslikkust. Oletatavasti toimivad Burrowsi delta ja koosinusdelta paremini kui teised tavapärased kaugusemõõdikud.
2) Katsesse kaasatud sagedaimate tunnuste arv (Eder 2013). Eesmärk oli testida, mitme mõõtmega andmestikku on tugevaks autorituvastuseks vaja. On teada, et autorile jälile jõudmiseks peaks piisama juba käputäiest sagedaimatest sõnadest.
3) Tõhus tekstinäidise maht. Sõnade esinemissageduste kasutamine tekstide võrdlemisel sõltub suurel määral saada oleva tekstinäidise mahust (Eder 2013). Enamasti on halb kasutada väga lühikesi tekste, kuna need ei loo küllalt suurt ruumi, milles stabiilsed sõnakasutusmustrid esile võiks kerkida. Samuti on problemaatiline analüüsida drastiliselt erineva mahuga tekste: tulemuseks saadavad klastrid oleks tugevalt mõjutatud tekstide mahust, mitte eeldatavatest suhetest. Selleks et saada esmast aimu, milline on stilomeetrilise analüüsi jaoks turvaline eestikeelse teksti maht, proovisin tuvastust jooksutada nii, et kasutasin võrdse suurusega juhuvalimit igast tekstist, valimi suurust samal ajal järk-järgult suurendades.

Iga katsekorra jooksul käsitleti iga teksti ükshaaval sihtdokumendina ning selle autor tuvastati lihtsakoelise lähima naabri printsiibi alusel: sihttekstile kõige lähema teksti autor määrati ka sihtteksti tõeliseks autoriks. Kui autorite sildid sihtteksti ja selle lähimate naabrite vahel kokku langesid, läks arvesse tabamus, kui aga mitte, siis möödalask. Iga katsekorra lõpptäpsusena läks arvesse tabamuste kogusumma tekstide koguarvu (18) suhtes. Katseseeriad on näidatud koondatud kujul joonisel 5.



Joonis 5. Kaugusepõhine ristvalideerimine kõige sagedasemate sõnade arvu alusel (10–5000). Kõige sagedasemate sõnade skaala on logaritmiline, et näidata täpsuse järsku hüpet, mis leiab aset 50 ja 100 kõige sagedasema sõna vahel.

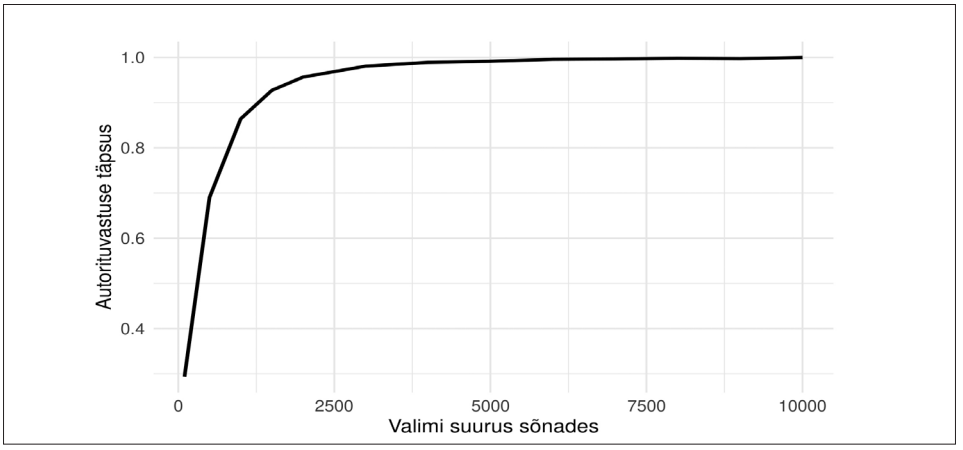
Üldiselt kipub tuvastuse täpsus kõigi kaugusmõõdikute korral tipnema ja stabiliseeruma umbes saja kõige sagedasema sõna kaasamisel. Edasine sõnavektorite mõõtmelisuse kasv ei näi tuvastamise jaoks enam lisainformatsiooni andvat. See kinnitab ootuspäraselt veel kord, et kõige sagedasematel funktsioonisõnadel on üheskoos võime individuaalseid kirjutamisharjumusi väljendada.

Lähima naabri tuvastamine koosinusdelta, Manhattani kauguse või Burrowsi deltaga oli selgelt tõhusam kui eukleidilise või klassikalise koosinuskaugusega. Manhattani kauguse (ehk pelga sõnadevaheliste suhteliste sageduserinevuste summa) hea tulemuslikkus on mõnevõrra üllatav, teades *z*-teisenduse jõudu. See viitab, et autorite lõikes tulevad sõnakasutuse erinevused selles konkreetses korpuses selgelt esile: antud tulemus ei pruugi skaleeruda suuremale ja tasakaalustatumale korpusele. Teisalt on koosinusdelta kõrge töökindlus juba varem kindlaks tehtud (Evert jt 2017).

Kõiki õigesti ja valesti läinud tuvastusi arvesse võttes saab välja selgitada, milliste autorite millised tekstid tuvastamisel enim raskusi valmistavad. Ilmnes, et Liivi ja Vilde tekste ei tuvastatud õigesti umbes pooltel kordadel, samas kui Aspe ja Tammsaare tekste klassifitseeriti kõigi katsekordade lõikes valesti väga harva. Enne kui sellele nähtusele välist seletust otsima hakata, tuleks püüda leida korpusest võimalikke kallutatusi. Selgus, et nii Liiv kui ka Vilde on esindatud korpuse kõige lühemate tekstidega (näiteks Vilde jutustus „Minu esimesed triibulised" on ainult 2300 sõna pikk), mis tõstatab kahtluse, kas põhjus võib olla pikkuses.

Selleks et leida tõhusat tekstinäidise mahtu eestikeelse teksti autorituvastuse jaoks, eemaldasin esmalt katsest Liivi ja Vilde tekstid, et vältida nende pikkusest tulenevat kitsendust. Järele jääv korpus (4 autorit, 12 teksti) muutus varasemast veelgi väiksemaks, kuid kasutasin seda demonstratsiooniks sellegipoolest.

Kuna nüüdseks on teada, et koosinusdelta töökindlus on väga stabiilne, kui kõige sagedasemate sõnade arv > 100, kasutasin seda kaugusetüüpi 200 kõige sagedasema sõnaga, et viia läbi katse näidisemahtudega. Selle asemel et esitada teksti kõigi saada



Joonis 6. Koosinusdelta-põhise tuvastuse täpsus juhuvalimi kasvava suuruse suhtes. 100 juhuvalimit igast tekstist iga valimisuuruse korral; kõver näitab keskmistatud täpsust.

olevate sõnade kujul, esitasin iga teksti sellest juhuslikult valitud sõnade v a l i m i n a. Valimi suurus oli katsekorra piires iga teksti puhul sama, nii et kõik arvutused viidi läbi võrdse suuruse valimiga. Iga suuruse juures (alates sajast kuni kümne tuhande sõnani) võtsin igast tekstist välja sada sõltumatut valimit ning arvutasin iga kord tuvastuse täpsuse. Joonisel 6 võib näha kõverat, mis väljendab lõplikku keskmist täpsust k õ i g i v a l i m i s u u r u s t e k o r r a l.

Koos valimi suurusega kasvas kiiresti täpsus ja autorituvastus jõudis usaldusväärsele tasemele juba ~2000 sõna juures. Stabiilne platoo saabus aga alles ~5000-sõnalise valimiga ja see näitabki autorituvastuse jaoks n-ö turvalise tekstinäidise suurust, mis on ühtlasi kooskõlas varasemate, teiste Euroopa keelte palju suuremate korpuste tulemustega (Eder 2013, 2017). See tõepoolest viitab, et mõned Liivi ja Vilde tekstid ei pruugi olla piisavalt mahukad, et nende autorit õnnestuks järjekindlalt tuvastada. Katse sooritati terviktekstidest juhuslikult võetud sõnavalimitega ning pikemates tekstides on rohkem n-ö ruumi, millest iga kord stiilinäidist võtta. Teksti pikkuse probleemi ei saa sageduspõhises stilomeetrias seega lõplikult lahendada: lühemate tekstide kaasamine analüüsi eeldab alati ettevaatlikku lähenemist.

Kokkuvõte

Käesoleval teemasse sissejuhataval artiklil oli kaks üldist eesmärki. Esiteks püüdsin anda ülevaate elementaarsest kvantitatiivsest tekstianalüüsist, et näidata, kuidas isegi lihtsakoeliseimast tekstiesitusest võib olla kasu igipõliste filoloogiliste küsimuste, nagu autorituvastus, lahendamisel ning laiemate kirjanduslike suhete, nagu žanri, kronoloogilise muutuse jne, modelleerimisel. Teine eesmärk oli demonstreerida kaugusepõhist autorituvastuse meetodit, kasutades selleks väikest eesti ilukirjanduse korpust. Tegemist oli ettevalmistava katsega, mida piiras oluliselt saada oleva korpuse hetkeolukord ning minu teadmised materjali kohta. Katsed näitasid, et Burrowsi delta, koosinusdelta ja Manhattani kaugus annavad häid tulemusi juba siis, kui kasutada analüüsis sadat kõige sagedasemat sõna, kusjuures koosinusdeltal on võrdluses teiste mõõdikutega väike eelis. Samuti oli näha, et stabiilsed tulemused tulevad kahe kuni viie tuhande sõnalise tekstinäidise korral. Need mustrid on kauguspõhises autorituvastuses ootuspärased, kuid selleks et teha mingeid üldistusi eesti keele kohta, oleks vaja siinsetele proovitulemustele lisaks paremini vormistatud katseid.

Ajalooline ülevaade stilomeetria arengust näitas, kuidas varased katsed kvantitatiivse autorituvastuse vallas tuginesid spetsiifilistele eeldustele alateadvuse ja autori objektiivse identiteedi kohta, mille järgi viimane peaks olema leitav autori igast tekstist (või muust teosest), kui valida vaid õiged tekstilised tunnused ja õige analüüsitase. Mõeldes üüratutele tekstimahtudele, mida tänapäeval sotsiaalmeedia vahendusel suheldes päevast päeva toodetakse, võib kerkida mure stilomeetria võimaliku invasiivse kasutuse pärast, olgu selleks siis omavoliline deanonümiseerimine (selle tuntud näideteks on J. K. Rowlingi stiili sidumine Robert Galbraithiga (Juola 2013) ja salapärase Elena Ferrante sidumine Dominico Starnonega (Tuzzi, Cortelazzo 2018))

või tekstipõhine sotsiaalne ja psühholoogiline profileerimine, millele võib tekkida ennustamatuid päriselulisi rakendusi (Argamon jt 2009b; Noecker jt 2013; vt ka Juola 2020: 162–163). Teades, kui vähe inimene kontrollib oma sügavalt juurdunud kirjutamisharjumusi, võib stilomeetriat kasutada mitte üksnes tuvastamismasinate, vaid ka selliste anonümiseerimismasinate loomiseks, mis varjavad autori „sõrmejälge" ja moonutavad stiili (Emmery jt 2021). Miskipärast pole kuigi keeruline kujutada ette tulevikku, kus sedalaadi „võistlev stilomeetria" (Brennan jt 2012) mängib olulist sotsiaalset rolli.

Ingliskeelsest käsikirjast tõlkinud
TAAVI LAANPERE

Uurimistöö teostati projekti „Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics" (NCN 2017/26/E/ HS2/01019) raames. Tänan Peeter Tinitsat eestikeelsete andmetega abistamast ning anonüümseid retsensente helde tagasiside eest.

KOOD JA ANDMED

Korpus ja kood kõigi katsete ja diagrammide jaoks on saadaval varamus: https://github.com/perechen/stylometry_intro_KK. Analüüsi jaoks kasutati tarkvara R 4.1.0. Peamised stilomeetria katsed tuginesid lisaks ka Stylo paketile (Eder jt 2016).

KIRJANDUS

Ackerman, James S. 1962. A Theory of style. – The Journal of Aesthetics and Art Criticism, kd 20, nr 3, lk 227–237. https://doi.org/10.2307/427321

Allison, Sarah; Heuser, Ryan; Jockers, Matthew L.; Moretti, Franco; Witmore, Michael 2011. Quantitative formalism: An experiment. – Literary Lab Pamphlets 1, 15. XI. https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf

Argamon, Shlomo 2008. Interpreting Burrows's Delta: Geometric and probabilistic foundations. – Literary and Linguistic Computing, kd 23, nr 2, lk 131–147. https://doi.org/10.1093/llc/fqn003

Argamon, Shlomo; Goulain, Jean-Baptiste; Horton, Russell; Olsen, Mark 2009a. Vive la Différence! Text mining gender difference in French literature. – Digital Humanities Quarterly, kd 3, nr 2. http://www.digitalhumanities.org/dhq/vol/3/2/000042/000042.html

Argamon, Shlomo; Koppel, Moshe; Pennebaker, James W.; Schler, Jonathan 2009b. Automatically profiling the author of an anonymous text. – Commun. ACM, kd 52, nr 2, lk 119–123. https://doi.org/10.1145/1461928.1461959

Brennan, Michael; Afroz, Sadia; Greenstadt, Rachel 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. – ACM Transaction on Information and System Security, kd 15, nr 3, art 12. https://doi.org/10.1145/2382448.2382450

Burrows, John 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. – Literary and Linguistic Computing, kd 17, nr 3, lk 267–287. https://doi.org/10.1093/llc/17.3.267

Cambpell, Lewis 1867. The Sophistes and Politicus of Plato. Oxford: Oxford Clarendon Press.

Chang, Kent K.; DeDeo, Simon 2020. Divergence and the complexity of difference in text and culture. – Journal of Cultural Analytics, kd 1, nr 1. https://doi.org/10.22148/001c.17585

Croft, William 2000. Explaining Language Change: An Evolutionary Approach. Harlow: Pearson Education.

Da, Nan Z. 2019. The computational case against computational literary studies. – Critical Inquiry, kd 45, nr 3, lk 601–639. https://doi.org/10.1086/702594

Dittenberg, Wilhelm 1881. Sprachliche Kriterien für die Chronologie der platonischen Dialoge. – Hermes. Zeitschrift für klassische Philologie, kd 16, lk 321–345.

Eder, Maciej 2013. Does size matter? Authorship attribution, small samples, big problem. – Digital Scholarship in the Humanities, kd 30, nr 2, lk 167–182. https://doi.org/10.1093/llc/fqt066

Eder, Maciej 2015. Visualization in stylometry: Cluster analysis using networks. – Digital Scholarship in the Humanities, kd 32, nr 1, lk 50–64. https://doi.org/10.1093/llc/fqv061

Eder, Maciej 2017. Short samples in authorship attribution: A new approach. – Digital Humanities 2017. Montreal, Canada, August 8–11, 2017. Conference Abstracts. Montreal: McGill University, lk 221–224.

Eder, Maciej; Rybicki, Jan; Kestemont, Mike 2016. Stylometry with R: a package for computational text analysis. – R Journal, kd 8, nr 1, lk 107–121. https://doi.org/10.32614/RJ-2016-007

Emmery, Chris; Kádár, Ákos; Chrupała, Grzegorz 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. – ArXiv.org. Computer Science. https://arxiv.org/abs/2101.11310 (5. VIII 2021).

Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Reger, Isabella; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten 2017. Understanding and explaining Delta measures for authorship attribution. – Digital Scholarship in the Humanities, kd 32, nr suppl_2, lk ii4–ii16. https://doi.org/10.1093/llc/fqx023

Ginzburg, Carlo 1979. Clues: Roots of a scientific paradigm. – Theory and Society, kd 7, nr 3, lk 273–288. https://doi.org/10.1007/BF00207323

Grieve, Jack 2007. Quantitative authorship attribution: An evaluation of techniques. – Literary and Linguistic Computing, kd 22, nr 3, lk 251–270. https://doi.org/10.1093/llc/fqm020

Grzybek, Peter 2014. The emergence of stylometry: Prolegomena to the history of term and concept. – Text within Text – Culture within Culture. Toim Katalin Kroó, Peeter Torop. Budapest–Tartu: L'Harmattan, lk 58–75.

Herrmann, Berenike J.; Dalen-Oskam, Karina van; Schöch, Christof 2015. Revisiting style, a key concept in literary studies. – Journal of Literary Theory, kd 9, nr 1, lk 25–52. https://doi.org/10.1515/jlt-2015-0003

Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. – Literary and Linguistic Computing, kd 13, nr 3, lk 111–117. https://doi.org/10.1093/llc/13.3.111

Holmes, David I.; Kardos, Judit 2003. Who was the author? An introduction to stylometry. – CHANCE, kd 16, nr 2, lk 5–8. https://doi.org/10.1080/09332480.2003.10554842

Hughes, James M.; Foti, Nicholas J.; Krakauer, David C.; Rockmore, Daniel N. 2012. Quantitative patterns of stylistic influence in the evolution of literature. – Proceedings of the National Academy of Sciences, kd 109, nr 20, lk 7682–7686. https://doi.org/10.1073/pnas.1115407109

Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten 2015. Improving Burrows' Delta – An empirical evaluation of text distance measures. – Digital Humanities Conference 2015. University of Western Sydney, Australia. Konverentsiettekanne. https://www.researchgate.net/publication/280086768_Improving_Burrows'_Delta_-_An_empirical_evaluation_of_text_distance_measures (17. VIII 2021).

Jockers, Matthew L. 2013. Macroanalysis: Digital Methods and Literary History. 1st Edition. Urbana: University of Illinois Press. https://doi.org/10.5406/illinois/9780252037528.001.0001

Juola, Patrick 2012. Large-scale experiments in authorship attribution. – English Studies, kd 93, nr 3, lk 275–283. https://doi.org/10.1080/0013838X.2012.668792

Juola, Patrick 2013. How a computer program helped reveal J. K. Rowling as author of a Cuckoo's Calling. – Scientific American. https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/ (2. VI 2021).

Juola, Patrick 2020. Authorship studies and the dark side of social media analytics. – Journal of Universal Computer Science, kd 26, nr 1, lk 156–170.

Kestemont, Mike; Stover, Justin; Koppel, Moshe; Karsdorp, Folgert; Daelemans, Walter 2016. Authenticating the writings of Julius Caesar. – Expert Systems with Applications, kd 63, lk 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

Koolen, Corina W. 2018. Reading beyond the female. The relationship between perception of author gender and literary quality. (ILLC Dissertation Series DS-2018-03.) Universiteit van Amsterdam.

Koppel, Moshe; Schler, Jonathan; Argamon, Shlomo 2009. Computational methods in authorship attribution. – Journal of the American Society for Information Science and Technology, kd 60, nr 1, lk 9–26. https://doi.org/10.1002/asi.20961

Lotman, Jurij M.; Lotman, Mihail J. 1986. Vokrug desjatoj glavy „Evgenija Onegina". – Puškin: Issledovanija i materialy. Leningrad: Nauka, lk 124–151. [Юрий М. Лотман, Михаил Ю. Лотман, Вокруг десятой главы „Евгения Онегина". – Пушкин: Исследования и материалы. Ленинград: Наука.]

Lutosławski, Wincenty 1897a. The Origin and Growth of Plato's Logic. With an Account of Plato's Style and of the Chronology of His Writings. London: Longmans, Green and Co.

Lutosławski, Wincenty 1897b. On stylometry. – Classical Review, kd 11, lk 284–286. https://doi.org/10.1017/S0009840X00032315

Malone, Edward 1787. A dissertation on parts I, II and III of Henry VI tending to show that those plays were not written originally by Shakespeare. London: From the Press of Henry Baldwin.

Mendenhall, Thomas C. 1887. The characteristic curves of composition. – Science, kd ns-9, nr 214S, lk 237–246. https://doi.org/10.1126/science.ns-9.214S.237

Mendenhall, Thomas 1901. A mechanical solution of a literary problem. – Popular Science Monthly, kd 60, lk 98–105.

Morozov, Nikolaj Aleksandrovič 1915. Lingvističeskie spektry. Sredstvo dlja otličenija plagiatov ot istinnyh proizvedenij togo ili drugogo izvestnogo avtora. – Izvestija otdelenija russkogo jazyka i slovesnoti Imperatorskoj akademii nauk, kd 20, nr 1–4, lk 95–127. [Николай Александрович Морозов, Лингвистические спектры. Средство для отличения плагиатов от истинных произведений того или другого известного автора. – Известия отделения русского языка и словесноти Императорской академии наук.]

Mosteller, Frederick; Wallace, David L. 1963. Inference in an authorship problem. – Journal of the American Statistical Association, kd 58, nr 302, lk 275–309. https://doi.org/10.1080/01621459.1963.10500849

Newberry, Mitchell G.; Ahern, Christopher A.; Clark, Robin; Plotkin, Joshua B. 2017. Detecting evolutionary forces in language change. – Nature, kd 551, nr 7679, lk 223–226. https://doi.org/10.1038/nature24455

Noecker Jr, John; Ryan, Michael; Juola, Patrick 2013. Psychological profiling through textual analysis. – Literary and Linguistic Computing, kd 28, nr 3, lk 382–387. https://doi.org/10.1093/llc/fqs070

Piper, Andrew 2017. Fictionality. – Journal of Cultural Analytics, kd 1, nr 1. https://doi.org/10.22148/16.011

Plecháč, Petr; Bobenhausen, Klemens; Hammerich, Benjamin 2018. Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. – Studia Metrica et Poetica, kd 5, nr 2, lk 29–54. https://doi.org/10.12697/smp.2018.5.2.02

Rapp, Christof 2010. Aristotle's Rhetoric. – The Stanford Encyclopedia of Philosophy. Toim Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2010/entries/aristotle-rhetoric/ (5. VIII 2021).

Sarawgi, Ruchita; Gajulapalli, Kailash; Choi, Yejin 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. – CoNLL '11: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Portland Oregon, June 23–24, 2011. Stroudsburg: Association for Computational Linguistics, lk 78–86.

Savoy, Jacques 2020. Machine Learning Methods for Stylometry. Springer. https://doi.org/10.1007/978-3-030-53360-1

Smith, Peter W. H.; Aldridge, W. 2011. Improving Authorship Attribution: Optimizing Burrows' Delta Method*. – Journal of Quantitative Linguistics, kd 18, nr 1, lk 63–88. https://doi.org/10.1080/09296174.2011.533591

Spedding, James 1850. Who wrote Shakespeare's Henry VIII? – Gentleman's Magazine and Historical Review, kd 34, lk 381–382.

Storey, Grant; Mimno, David 2020. Like two Pis in a pod: Author similarity across time in the Ancient Greek corpus. – Journal of Cultural Analytics, kd 1, nr 1. https://doi.org/10.22148/001c.13680

Šapir, Maksim Il'ič 2000. Fenomen Baten'kova i problema mistifikacii (lingvostihovedčeskij aspekt). – M. I. Šapir, Universum versus: Jazyk, stih, smysl v russkoj poèzii XVIII–XIX vekov. Moskva: Jazyki russkoj kul'tury, lk 335–443. [Максим Ильич Шапир, Феномен Батенькова и проблема мистификации (Лингвостиховедческий аспект). – М. И.

Шапир, Universum versus: Язык, стих, смысл в русской поэзии XVIII–XIX веков. Москва: Языки русской культуры.]

Šeļa, Artjoms; Orekhov, Boris; Leibov, Roman 2020. Weak genres: Modeling association between poetic meter and meaning in Russian poetry. – CHR 2020: Workshop on Computational Humanities Research. Amsterdam: CEUR-WS, lk 12–31.

Zipf, George K. 1949. Human Behavior and the Principle of Least Effort. Cambridge: Addison-Wesley Press, Inc.

Tarlinskaja, Marina 1987. Shakespeare's Verse. Iambic Pentameter and the Poet's Idiosyncrasies. New York: Peter Lang.

Tomaševski 1923 = Boris Viktorovič Tomaševskij, Pjatistopnyj jamb Puškina. – Očerki po poètike Puškina. Berlin: Èpoha, lk 7–143. [Борис Викторович Томашевский, Пятистопный ямб Пушкина. – Очерки по поэтике Пушкина. Берлин: Эпоха.]

Tuzzi, Arjuna; Cortelazzo, Michele A. (toim) 2018. Drawing Elena Ferrante's Profile. Padova: Padova University Press.

Uiboaed, Kristel 2017. Kirjandusteoste automaatanalüüs [Text-mining and Stylometric Analysis of Estonian Novels]. – Tekstikaeve. http://www.tekstikaeve.ee/blog/2017-10-25-kirjandusteoste-automaatanalyys/ (2. VI 2021).

Uiboaed, Kristel (toim) 2018. E-raamatute eeltöödeldud ja lemmatiseeritud failid. https://datadoi.ee/handle/33/76

Underwood, Ted 2017. The life cycles of genres. – Journal of Cultural Analytics, kd 1, nr 1. https://doi.org/10.22148/16.005

Underwood, Ted 2019. Distant Horizons: Digital Evidence and Literary Change. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

Yule, George Udny 1939. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. – Biometrika, kd 30, nr 3–4, lk 363–390. https://doi.org/10.1093/biomet/30.3-4.363

Yule, George Udny 1944. The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press.

Artjoms Šeļa (snd 1989), PhD, Tartu Ülikooli eesti ja üldkeeleteaduse instituudi digihumanitaaria teadur (Jakobi 2, Tartu 51003); arvutusliku stilomeetria lektor Poola Teaduste Akadeemia Poola Keele Instituudis Krakówis, artjoms.sela@ut.ee

Differences, distances and fingerprints: The fundamentals of stylometry and multivariate text analysis

Keywords: stylometry, computational text analysis, authorship attribution, Estonian fiction

The recent rapid expansion of computational methods and tools into humanities have rekindled the conversation surrounding the relationship between a study object and its mathematical representation, or model. The paper serves as a conceptual

introduction to stylometry, a sub-field of computational text analysis that studies differences between texts quantitatively, and shows how simplistic models of texts can be used to uncover their complex relationships.

The very term "stylometry" and the field's development is closely linked to the problem of authorship attribution and identification – the paper briefly introduces the early history of stylometry that highlights its inherited assumptions about texts and authorship. It shows how text analysis methods have shifted from analysis of single features (such as word length) to multivariate computations. The paper explains the basics of modern multivariate text analysis and the notion of a "distance" between texts as a proxy to their difference, focusing primarily on word frequencies as units of analysis.

The paper concludes with a series of preliminary authorship experiments for a small Estonian fiction corpus, which test the behavior of a few foundational variables and serve as a general proof-of-concept demonstration. Experiments show a reliable performance of Cosine Delta distance in the Estonian non-lemmatized corpus, using frequencies of at least 100 most frequent words. The Cosine Delta also achieves stable attribution accuracy for random samples of at least 5000 words, which suggests that texts shorter than this size might not be reliably attributed with the proposed methodology. Both findings are consistent with observations done for other languages, but should be treated as preliminary for Estonian.

Artjoms Šeļa (b. 1989), PhD, Researcher in Digital Humanities in the University of Tartu (Jakobi 2, Tartu 51003), Assistant Professor in Computational Stylometry in the Institute of Polish Language, Polish Academy of Sciences, Kraków, artjoms.sela@ut.ee

https://doi.org/10.54013/kk764a4

Kirjanduskeele kauglugemine

Kirjandust väärindava metodoloogia arendamine tsiteerimisvõtte näitel

MAARJA HELENA MERISTE

Digimeetodite kasutuselevõtt on kirjandusteadusse toomas uurimisviise, mis võivad põhjalikult ümber kujundada senised arusaamad kirjanduskeelest ja kirjandusajaloost laiemalt. Arvutilingvistid on välja arendanud tänuväärseid kvantitatiivse keeleanalüüsi meetodeid ja tööriistu, mis sobivad loomuliku keele uurimiseks ja töötlemiseks. Eesti keeleteaduses on arvutuslikud meetodid juba kanda kinnitanud, ent kirjandusteaduses on neil põhinevaid lähenemisi vaid põgusalt katsetatud. Miks? Keeleteaduses on deskriptiivsed uurimismeetodid (sh arvutustel ja mõõtmistel põhinevad) otseselt kohasemad, sest objekt on inimkeel ise ja eesmärk seda kirjeldada. Kirjanduse keel seevastu on kunstiline keel. Juri Lotmani (1991) traditsioonis mõtestatakse kunstilist keelt teisese modelleeriva süsteemina ehk omaette keelena, mis on tuletatud loomulikust keelest. See teeb kirjandusest märksa keerulisema uurimisobjekti, mis vajab selle eripärale kohandatud uurimisvõttestikku: kirjanduse keel on sihilikult mitmetähenduslik, mänguline, vihjav, varjav, kujundlik, irooniline jne. Tähenduslik ambivalents, tõlgendusvõimaluste paljusus on kunstilise keele essents, mis loob mitmekihilise teksti. Eesti kontekstis on selle hea näide nõukogude aja kirjanduse n-ö ridadevaheline poeetika. Teisisõnu – ilukirjanduses ei pruugi tähendust kandvad elemendid, nendevahelised seosed ja nende struktuur nähtaval olla. Seepärast ei saa keeleanalüüsi meetoditega kirjandusteose tähendust selgitada, ent just tähendusprobleemid on kirjandusteaduse kese. Küll aga võib arvutuslikult tuvastada ilukirjandusest keelekasutusmustreid, mida lugemisel ei haara, kuid mis võivad kultuuriloolisse konteksti asetatuna osutuda tähendusrikkaks ning täiendada või muuta seniseid tõlgendusi.

Sellest tulenevalt on digitaalse kirjandusuurimuse suurim väljakutse sobiva metoodika arendamine vastavalt uurimisprobleemile, mis on loomult kvalitatiivne ning tähenduslikult kompleksne. Käesoleva artikli eesmärk ongi ühe magistritöö (Meriste 2020) varal tuua esile uurimisvõimalusi, mille avab kirjandusteaduse jaoks digitaalne analüüs. Siinkirjutaja arvates on digimeetodite viljakaim kasutuspotentsiaal kirjandusteaduses vormiuurimus. Seda näitlikustatakse artiklis kirjanduskeele ühe spetsiifilise nähtuse, nn avatud tsitaadi mõistmiseks loodud metoodika ja selle rakendamise kaudu. Terminiga *avatud tsitaat* on prantsuse filosoof François Recanati (2001: 637) tähistanud poeetilist võtet, mis seisneb lausungi mingi osa tsitaadiks märkimises nii, et jutumärkidega tõstetakse esile lisa- või moonutatud tähenduskihi (iroonilise, eufemistliku vms) olemasolu. Tsiteerimist on koos muude mänguliste võtetega, nagu paroodia, pastišš ja kollaaž, peetud omaseks 1990. aastate eesti kirjandusele (Pruul 2000: 181), ometi pole selle tegelikku kasutust eesti kirjanduses

uuritud. Selle üheks põhjuseks on arvatavasti olnud asjaolu, et niisugused nähtused võivad ühe teose kontekstis olla liiga marginaalsed, et usaldusväärseks uurimuseks vajalikku hulka näiteid kokku saada: vaja on teosteülest vaadet, mida võimaldavadki digimeetodid. Tänu nendele saab kirjandust vaadelda nähtusekeskselt, nii et üksiku teose või autori eritlemise asemel on rõhk vormiuurimusel ehk sellel, kuidas mõni motiiv, stilistiline element vm võte kirjandusväljal ilmneb, kordub ja teiseneb.

Avatud tsitaadid on kirjanduskeele nähtus, mida on hea sel viisil uurida. Erinevalt paljudest muudest kirjanduslikest võtetest on avatud tsitaatidel kindel formaalne tunnus (jutumärgid), mille abil saab need tuvastada. Artiklis avatakse uurimuse tausta, ettevalmistust, analüüsi käiku ja uurimistulemusi.

Artikli põhitees on, et kunstiuurimusele kohase andmeteaduse arendamisel ei tohi kõrvale heita kvalitatiivset lähenemist, vaid see tuleb ühendada kvantitatiivsega, nii et need teineteist täiendaks. Kirjanduse puhul on arvutuslikud meetodid tulemuslikud stilistiliste üldmustrite tuvastamisel, kuid see üksi jääb vähekõnekaks. Kvalitatiivne analüüs aitab kvantitatiivselt saadud teadmist tõlgenduse ja konteksti asetamise kaudu väärindada. Kvalitatiivsete ja kvantitatiivsete meetodite kombineerimine lubab kirjandusteaduslikele probleemidele läheneda nende loomupärast keerukust lihtsustamata ning sellise uurimisprotsessi kujundamine on digitaalse kirjandusuurimuse esmatähtis oskus.

Vormikoolkonnast kauglugemiseni

Paljusid teoseid ühendava stiilivõttestiku uurimise vastu tekkis huvi ammu enne digiajastut. Kirjandusteooria ajaloos on selle lähenemisviisi silmapaistvaimad arendajad XX sajandi alguse vene vormikoolkonna esindajad. Formalistid pidasid oma ülesandeks vabastada kirjandusteadus traditsioonilistest, ent nende silmis vähemväärtuslikest uurimistavadest, millest üks põhilisi oli teose seletamine autori isiku ja eluloo kaudu. Selle asemel rõhutasid nad vajadust keskenduda just kirjandusteosele ning mõelda sellest kui käsitööesemest, mille uurimisel pole oluline selle valmistanud meister, vaid selles avalduvad võtted. Seda vaadet näitlikustab hästi formalist Ossip Briki 1923. aasta manifest „Nn „Vormimeetod"", kus ta kirjutab:

> Luuletaja sotsiaalset rolli ei saa mõista tema individuaalseid omadusi ja harjumusi analüüsides. **Tarvis on massiliselt tundma õppida poeetilise käsitöö võtteid**, nende erinevusi inimtöö naabervaldkondadest ja nende ajaloolise arengu seadusi. [---] **Luule ajalugu on sõnalise vormimise võtete arengulugu.** [---]
>
> Miks on luuletajad teemade töötlemisel kasutanud just neid ja mitte muid võtteid, mis kutsub esile uue võtte ilmumise ja kuidas vana välja sureb – need küsimused võetakse teaduslikus poeetikas kõige hoolikama uurimise alla. (Brik 2014)

Briki väljendatud idee on erinevates variantides ilmnenud paljude tema kaasaegsete ja hilisemate kirjandusuurijate töödes. Tähelepanuväärseimad neist on

Moskva lingvistikaringi liikmed (nt Boriss Jarho), Praha lingvistikaringi teoreetikud (nt Roman Jakobson) ning prantsuse (post)strukturalistid (nt Gérard Genette ja Roland Barthes).

Oleg Sobchuk (2018: 10) sedastab oma väitekirjas, kus ta katsetab digimeetodite kasutamist kultuuri evolutsiooni uurimisel, et just lingvistiline strukturalism pani kirjandusteadlased otsima mustreid. Niisiis on XX sajandi strukturalistliku kirjandusteaduse pärand nüüdisaegse kvantitatiivse kirjandusuurimise filosoofiliseks ja teoreetiliseks aluseks.

Moodsas kirjandusteaduses tuntakse arvutuslike meetodite abil tehtud vormiuurimusi kauglugemise mõiste all. Kauglugemise looja on USA kirjandusuurija Franco Moretti, kes pälvis tähelepanu julge loosungiga, et uurijad peaksid kirjanduse tegelikuks mõistmiseks lõpetama raamatute lugemise. Sellega osutab Moretti (2013: 48) lähilugemiskeskse kirjandusteaduse kitsaskohale: väheste, tavaliselt kaanonisse kuuluvate teoste põhjal luuakse kirjanduslugu, mis tegelikult pole proportsionaalselt esinduslik.

Moretti arvates tuleb kultuuri läbivate mustrite nägemiseks distantseeruda üksikust tekstist. Selle asemel soovitab ta vaatluse alla võtta paljudest teostest koosneva tekstikorpuse, mis võimaldab uurida üksiku elemendi levimismustreid kirjanduses laiemalt. Suuremas tekstikorpuses ilmnevad teatud nähtuse kasutustendentsid, mis on aluseks üldistusjõulisematele tõlgendustele. (Moretti 2013: 49–50)

Mida suurem on korpus, seda suuremaid kultuuriloolisi, sotsiaalseid vm üldistusi saab materjali põhjal teha. Näiteks ilmnes 5000 ingliskeelse romaani (ilmunud vahemikus 1780–2000) keelekasutust uurides, et raamatud on aja jooksul muutunud aina kergemini loetavaks: laused on aina lühemad, lihtsama struktuuriga, sõnad lühemad (Piper 2018).

Nii Moretti kauglugemise idee kui ka üleüldse digimeetodite kasutamine kirjandusteaduses on viimastel aastatel olnud valdkonna teravamaid aruteluteemasid. Digihumanitaaria skeptikud ei usu, et kauglugemine suudaks anda uut vaadet kirjandusele, heal juhul leiab statistiliselt kinnitust see, mida varem juba teati (vt nt Kirsch 2014). Samuti on palju metoodilisi libastumiskohti seoses korpuse koostamise ning kanoonilise ja mittekanoonilise kirjanduse vahekorraga ning universaalselt toimivaid lahendusi napib (vt nt Wilkens 2012). Korpuse koostamise osas on üks värskemaid ja põnevamaid analüüse uurijatelt (Schöch jt 2021), kes koostavad digitaalset Euroopa kirjandustekstide kogu (ELTeC – European Literary Text Collection).

Nimetatud korpust koostatakse Euroopa teaduskoostöövõrgustiku COST projekti „Distant Reading for European Literary History" raames, mille eesmärk ongi luua üleeuroopaline digitaalne kirjanduskorpus, ühtsustada ja arendada kauglugemise meetodeid ning uurida, kuidas suhestuvad kauglugemisega saadud teadmised traditsioonilise kirjanduslooga.

Kuigi kauglugemine vastandub juba terminoloogiliselt lähilugemisele, oleks digihumanitaarse kirjandusuurimuse seisukohalt kõige kasulikum kaht lähenemist kombineerida. Selle järelduseni on jõudnud mitmed uurijad (nt van de Ven 2017). Eestis tutvustas sellist lähenemist California ülikooli folkloristika, kirjanduse

ja kultuuriuuringute professor Timothy Tangherlini, kes luges 2019. aastal Tartu Ülikoolis kursust makroskoopilistest meetoditest kultuuri uurimisel. Ta rõhutas, et humanitaarteadustes töötatakse harva sellise andmemahuga, mida võiks pidada suurandmeteks (ingl *big data*), pigem töötatakse keskmiselt väikese andmestikuga (*medium small data*). Klassikalised statistilised ja tõenäosuslikud arvutused ei saa väikese korpuse analüüsimisel eriti tulemuslikud olla, kuid humanitaarteadustes võibki mikro- ja makrovaate kombineerimine olla sobivaim. (Tangherlini 2019)

Selline meetod on eriti oluline teose keerukama tähendusega elemente kaardistades, sest sellised elemendid on tekstis tihtilugu varjatud – näiteks intertekstuaalsus, iroonia, vihjelisus. Ka käesoleva artikli kese, tsiteerimisvõte, on tähendusloome keerukuse poolest nendega sarnane, erinedes selle poolest, et tsitaat on tekstis nähtavalt (jutumärkidega) tähistatud.

Mis on avatud tsitaat ja kuidas seda uurida?

Avatud tsitaat on kirjanduses laialt levinud võte, mida iga lugeja on kohanud, kuid uuritud on seda vähe ning mõiste on ka paljudele keele- ja kirjandusuurijatele võõras. Seepärast on alustuseks otstarbekas tuua näide uuritavast korpusest:

> Me oleme ühtmoodi illusioonideta küünikud, kes praktiseerivad „mõningasest metodoloogilisest lahknevusest hoolimata", nagu Adrian ütleb, sedasama maailmavaadet. (Hiram 1999: 17)

Kui akadeemilises ning igapäevakasutuses mõistetakse tsiteerimist valitud materjali sõnasõnalise ümbertõstmisena uude teksti, siis tähenduse uurimise vaatenurgast on tegemist palju komplekssema nähtusega. Tsitaat ei pruugi alati olla seotud viitega teisele kontekstile, mujal varem öeldud või kirjutatud sõnadele. Antud juhul on tsitaadi miinimumtunnuseks teksti mingi osa teadlik eristamine ülejäänust: jutumärgistades valitud sõnad, kutsutakse tsitaati tõlgendama teisiti kui seda ümbritsevat teksti (Capone 2013: 260, 278).

Sellist tsiteerimisviisi nimetatakse avatuks, et seda vastandada suletule. Suletud tsitaadid on täistähenduslikud ja iseseisvad lausungid – näiteks otsekõne esineb tekstis suletud tsitaatidena (Recanati 2001: 649). Otsekõnelt ei saaks jutumärke eemaldada, sest need on ühe diskursusetüübi konventsionaalne tähistusviis. Avatud tsitaat, nagu näites näha, on aga lausesse süntaktiliselt ja tähenduslikult lõimitud. Sellelt jutumärke eemaldades jääks lause keeleliselt korrektseks, aga kaotaks tuntavalt ühe tähenduskihi. Siit ilmnebki, et avatud tsitaadi eesmärk on esile tõsta lisatähenduse olemasolu – pragmaatilist rikastatust (Recanati 2001: 637).

Millised on need võimalikud lisatähendused? Esiteks evidentsiaalsus ehk see, kelle diskursusesse tsiteeritud sõnad kuuluvad – eeltoodud näites juhivad jutumärgid tähelepanu sellele, et märgitud sõnad ei pärine kõnelejalt endalt, vaid teisest allikast. Laiemas mõttes saab niimoodi tähistada võõra allkeele või idiolekti tungimist peadiskursusesse. Peale selle on ilmne distantseerumine: kõneleja sarkastiline

või irooniline hoiak. Samuti võib tsitaadiga rõhutada just sõna tavatähendust, anda sellele ikoonilise jõu (Brendel jt 2011: 5–6).

Nõnda võiks avatud tsitaati mõista kui teksti vastuvõtmise ühtlase, harjumuspärase tempo teadlikku häirimist: lugemiskulg pidurdatakse, lugeja suunatakse osutama tähelepanu esiletõstetud tekstikillule. Mõnesõnaline tsitaat võib keelelise eneseleosutuse kaudu toimida otsekui päästik mõne suurema objekti või nähtuse kujutlusse äratamisel. Tsiteeritud sõnad muutuvad ikooniks mingist laiemast diskursusest, mida nad esindavad (Recanati 2001: 645). Järgnevas näites kasutatakse avatud tsitaati selleks, et osutada ühiskondlikule standardsuhtumisele:

> Tol ajal käis küll koolis palgaline koristaja, aga tema käis kord nädalas ja pesi ainult treppe ja koridoripõrandaid, õpilased pidid oma klassiruumi ikkagi ise koristama ja tüdrukud tegema *naistetöid* nagu tolmupühkimine ja põrandapesu, sest „klass oli ju ometi nende teine kodu". (Viiding 2008: 14)

Niiviisi viitab avatud tsitaat millelegi endast suuremale: suhtumisele, maailmavaatele, kogukonnale, tekstile, kultuurile, teosele vms. Sellega haakub Anneli Mihkelevi seletus allusiivsusest kirjanduses:

> Allusioon saab alguse mikrotasandilt kui vihjav signaal ehk marker, mis asub teksti pindmises kihis. Sellel markeril on esilekutsuv jõud ning käivitab allusiooni makrotasandil, st laia tähenduste ning assotsiatsioonide võrgustiku ehk allusiooni laiemas mõttes. (Mihkelev 2005: 144)

Eesti kirjandusuurimustest puudutavad tsiteerimisvõtteid Mihkelevi eelviidatud väitekiri ja Katrin Puiki doktoritöö irooniast Heiti Talviku ja Betti Alveri luules (2009). Mõlemad eritlevad kvalitatiivselt kirjandusnähtust, mis on olemuselt ambivalentne ja kindlate vormiliste tunnusteta. Käesolevas uurimuses lähenetakse teistpidi: nähtus piiritletakse esmalt formaalse tunnuse järgi ja seejärel vaadatakse, milliseid sisulisi omadusi see kätkeb.

Välismaistest kirjandusuurimustest on tsiteerimisvõttega tihedalt seotud Mihhail Bahtini kontseptsioon võõrast kõnest (Bakhtin 1982) ja Gérard Genette'i (1992: 8) lähenemine intertekstuaalsusele, mis tema järgi esineb tekstis tsitaadi, allusiooni või plagiaadina. Samuti suhestub tsiteerimise teemaga Bahtini ideede edasiarendaja Julia Kristeva (1980: 65–66), kes peab teksti teiste tekstidega suhtlemise dünaamiliseks väljaks, kusjuures teksti all ei pea ta silmas üksnes verbaalset, vaid üleüldse kultuuriteksti.

Seega saab öelda, et avatud tsitaat, mis tekstis esineb pelgalt mõne sõna tüpograafilise esiletõstuna ja mis näiliselt on seotud vaid konkreetse narratiivse hetkega, võib mõõtmatult rikastada teksti tähendust. Seda näiteks intertekstuaalsuse, vihjelisuse, tonaalsuse vm kaudu.

Avatud tsitaatide uurimise eesmärk on kirjeldada teatud ilukirjanduslikku võtet, selle levikut ja mõju ilukirjanduskeelele. Täpsemad uurimisküsimused on järgmised: milline on avatud tsitaatide levik korpuses? Millised on võtte tähenduslikud funkt-

sioonid tekstis? Kas ja millised kasutustendentsid ilmnevad ning kas on märgatavaid erisusi eri autorite/teoste puhul? Mida võib selle põhjal öelda analüüsitud teoste kohta?

Uurimuse ettevalmistus

Tsiteerimisvõtte efektiivse uurimise meetodi väljatöötamisel olid abiks kolm peamist küsimust. Neist tõukunud mõttearendused pole pelgalt tsitaadispetsiifilised – neist võiks olla kasu ka teiste sarnaste uurimuste ülesehitamisel.

Millistest teostest võiks korpus ideaalis koosneda, et uuritavast nähtusest hea ülevaade saada? Siinkohal peab uurija tegema kompromissi kahe vastandpõhimõtte vahel. Ühelt poolt tuleks korpus loomulikult koostada teostest, kus otsitav nähtus esineb. Teiselt poolt on kauglugemise üks põhimõtteid rakendada materjali valikul võimalikult vähe filtreid, et „kõneleda" saaksid ka need teosed, mis tavaliselt uurijate huviorbiidist välja jäävad.

Esimese nõude puhul aitab uurijal fookust seada lugemus ja kirjandusloolised teadmised. Näiteks tsiteerimine kuulub tekstivõtete hulka, mis on iseloomulikud taasiseseisvumisjärgsele eesti kirjandusele. Poliitiline ja sotsiaalne murranguaeg peegeldus kirjanduses mängulisuse, stiliseerituse, intertekstuaalsuse, allkeelterohkuse, iroonia ja erinevate jäljendusvõtete kasvuna (Velsker 2001: 620–622). Seega võis tsiteerimisvõtte esinemist selle perioodi kirjanduses pidada ootuspäraseks, kuigi statistilist kinnitust sellele pole.

Teise põhimõtte puhul tuleb arvestada sellega, et mingi valiku peab uurija niikuinii tegema. Eriti möödapääsmatu on see juhul, kui analüüs hõlmab ka kvalitatiivset osa ja materjali maht peab seda arvestades olema hoomatav. Olgu valitud piirangud ajalised, žanrilised vms, oluline on tulemustest järelduste ja üldistuste tegemisel neid teadvustada. Näiteks ei tohiks teatud perioodi proosakirjanduse ilminguid kergekäeliselt üldistada kogu toonasele ilukirjandusele, mis hõlmab ka luulet.

Kust saada uurimismaterjali? Digitaalse kirjandusuurimuse läbiviimiseks vajab uurija teoste tekstifaile. Eesti kirjanduse puhul piiravad materjali peamiselt kaks asjaolu: autoriõigused ja skaneerimise teel digiteeritud raamatud. Eesti kirjandust digiteeritakse Kultuuriministeeriumi projekti „Kultuuripärandi digiteerimine 2018–2023" (Kultuuriministeerium 2018) raames hoogsalt, Eesti Rahvusraamatukogu digitaalarhiiv Digar täieneb pidevalt. Kahjuks aga saavad raamatukogud autoriõigusega kaitstud teoseid digitaalselt kättesaadavaks teha vaid kohapealseks lugemiseks (AutÕS § 20). Autoriõigused kaitsevad teost autori eluajal ja 70 aastat pärast surma (AutÕS § 38), mis tähendab, et praegu on uurijale kättesaadavad ainult enne 1951. aastat lahkunud autorite teosed. Märkimisväärne osa digitaalselt kättesaadavast vanemast kirjandusest on aga digiteeritud skaneerituna, mis pole tekstitöötlusvalmis formaat. Käesolevaks uurimuseks oli niisiis tarvis teoseid, mis pole vabalt kättesaadavad, mistõttu sõltus korpuse koosseis suuresti sellest, kas kirjanikud ja kirjastajad olid nõus uurimuse jaoks jagama teoste elektroonilisi tekstifaile.

K u i s u u r t k o r p u s t o n v a j a? Korpuse koostamisel on üks nõue selle esinduslikkus: et materjal tõepoolest aitaks vastata uurimisküsimustele. Kauglugemise seisukohalt tuleb esinduslikkuse idee suhtes tähelepanelik olla. Kui teatud perioodi esindusliku kirjanduskorpuse koosseisu arvata ainult teosed, mis kuuluvad kirjandusloo kaanonisse, on juba mööda mindud kauglugemise ühest alusprintsiibist. Sellest lähtuvalt võiks korpusesse hõlmata nii palju teoseid, kui on vähegi kättesaadav. Samal ajal tuleb arvestada ka kättesaadavuse mõjuga korpuse sisule ja sedakaudu uurimistulemustele. Siit järeldub, et praegustes oludes on ideaalset digikorpust koostada raske, kui mitte võimatu. Korpuse koostamise põhimõtted mõjutavad otseselt selle põhjal tehtud uurimusi.

Oluline on teadvustada korpuse koostamisel tehtud valikuid ja piiranguid, isegi kui need on uurijast sõltumatud ja paratamatud, ja olla ka töö tulemusi tõlgendades tähelepanelik väidete tugevuse osas. Tulemuste üldistamine kogu kirjanduspildile pole õigustatud, kuid mööndusi arvesse võttes võib teatavaid suundumusi siiski välja tuua.

Uurimiskorpus sisaldas 35 eesti autori 50 ilukirjanduslikku proosateost (romaanid ja lühiproosa, välja jäeti laste- ja noortekirjandus), mis on ilmunud aastatel 1990–2010 (vt lisa). Korpuse kogumaht oli 2 609 062 sõna.

Analüüsi käik

Järjest enam on saadaval erinevaid pakette tekstianalüüsi teostamiseks, nagu veebipõhine Voyant Tools, Sketch Engine, RStudio programmeerimiskeskkonnas toimiv Stylo (vt Eder jt 2016), mida on kerge kasutada ja mis ei eelda kultuuriuurijalt programmeerimisoskust. Käesolevas uurimuses kasutati Voyant Toolsi korpuse üldiste arvuliste andmete tuvastamiseks (korpuse üldmaht, teoste mahud, proportsioonid jms). Tsitaatide eristamiseks siiski ükski valmistarkvara ei sobinud, selleks tuli sobiv kood ise RStudios kirjutada. Järgnevalt kirjeldan kokkuvõtlikult antud uurimuse tehnilise osa etappe.

1. Esialgne korpus sisaldas teoseid eri failivormingutes. Kõik failid viidi TXT-kujule ning teostelt eemaldati metaandmed (ilmumisaeg, kirjastus jms).
2. Kõik tekstid eraldati üksikuteks lauseteks (seda on kõige lihtsam teha lauselõpumärkide järgi). Sobivas mõõdus lähikontekstiks, milles tsitaati vaadelda, määrati üks lause.
3. Kõikide muude lausete seast eraldati avatud tsitaati sisaldavad laused, kasutades indikaatorina lõpujutumärkide esinemist lause sees (otsekõnes eelneb lõpujutumärkidele kas lauselõpumärk või koma, avatud tsitaadis mitte).
4. Väljafiltreeritud laused kanti edasiseks töötluseks Exceli tabelisse. Seal tuli laused käsitsi puhastada samadele otsingukriteeriumidele vastanud lausetest, mis ei sisaldanud tsitaati, vaid hoopis jutumärgistatud nime, pealkirja vms.
5. Allesjäänud laused ehk kõik korpuses esinenud avatud tsitaatide kasutusjuhud jaotati tähenduspõhistesse kategooriatesse. Selles etapis lisandus uuri-

musele kvalitatiivne lähenemisviis. Kui ühelauselisest kontekstist jäi tsitaadi mõistmiseks väheks, otsiti teose tekstifailis üles algkontekst.

6. Seejärel vaadeldi erinevate kategooriate esinemis- ja kokkulangemissagedust, tähendusloomet ning iseloomustati sedakaudu tsiteerimisvõtte funktsioone kirjanduskeeles.

Just viimase kahe etapi käigus kasvas kvantitatiivsest tekstianalüüsist kirjandusteaduslik uurimus.

Tsitaatide rühmitamise eesmärk oli see, et kategooriad peegeldaksid tsiteerimisvõtte peamisi funktsioone tekstis ja et neist moodustuks mudel, mille abil saaks omakorda muudes, korpusesse mittekuulunud tekstides tsitaadikasutust analüüsida. Avatud tsitaat, mille ülesanne on sõnade algse tähenduse moonutamine või neile lisatähenduse andmine, allub, nagu paljud teisedki kirjanduskeele nähtused, küllaltki raskelt jäikadele jaotustele ja ühestele tõlgendustele. Tähendusliku ambivalentsi säilitamine ongi kirjanduskeele põhiomadusi. Seda silmas pidades peaksid ka selliste nähtuste alusel loodavad mudelid võimaldama tõlgendusruumi. Avatud tsitaatide rühmitamisel väljendus see selles, et üks tsitaat võis kuuluda üheaegselt mitmesse kategooriasse. Nii ei kahandata võtte loomupärast mitmetähenduslikkust.

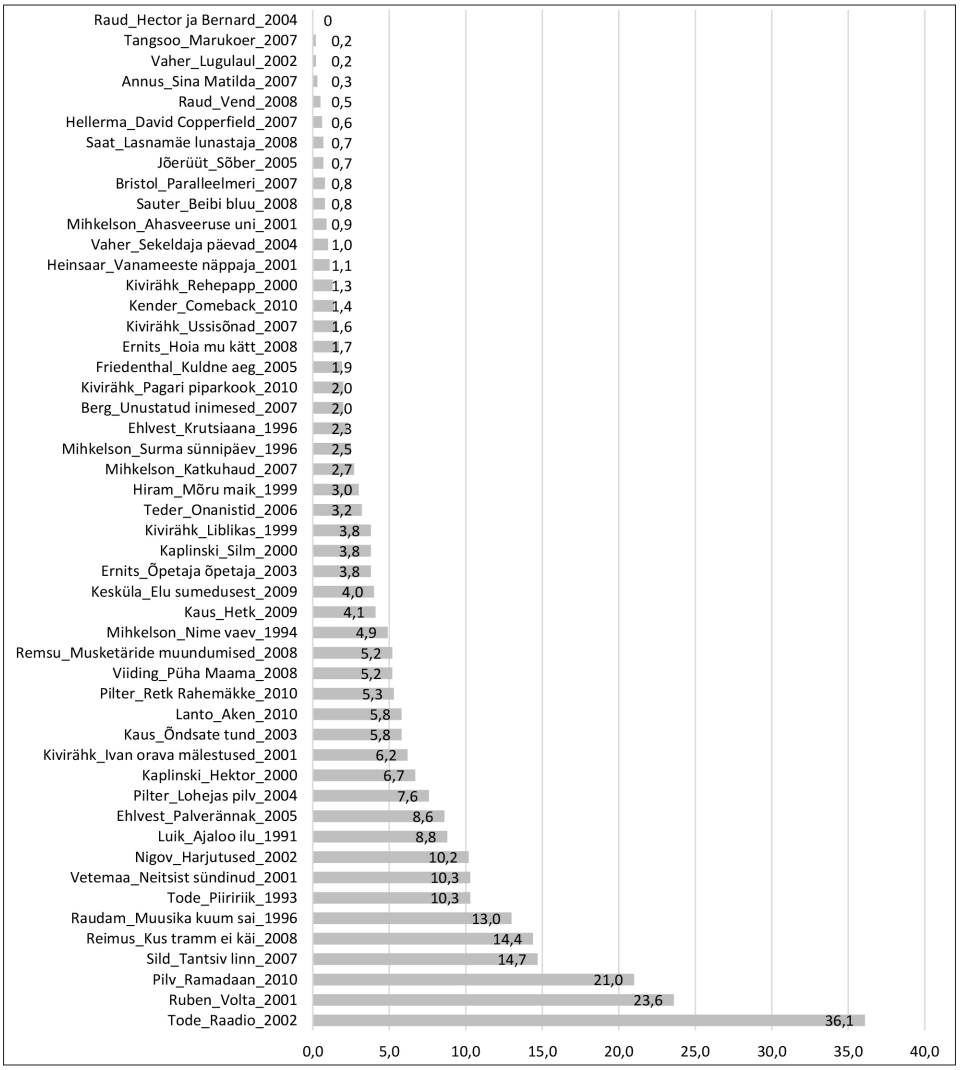
Uurimistulemused

Avatud tsitaate sisaldavaid lauseid oli korpuses kokku 1856. Teoste pikkusest tingituna arvutati tsitaatide esinemissageduse võrdlemiseks nende suhteline sagedus 10 000 sõna kohta (joonis 1). Kümme tsitaaditihedamat teost olid järgmised:

1. Emil Tode, „Raadio" (36,1 avatud tsitaati 10 000 sõna kohta)
2. Aarne Ruben, „Volta annab kaeblikku vilet" (23,6)
3. Aare Pilv, „Ramadaan" (21,0)
4. Ivar Sild, „Tantsiv linn" (14,7)
5. Katrin Reimus, „Kus tramm ei käi" (14,4)
6. Toomas Raudam, „Muusika kuum sai" (13,0)
7. Emil Tode, „Piiririik" (10,3)
8. Enn Vetemaa, „Neitsist sündinud" (10,3)
9. Anton Nigov, „Harjutused" (10,2)
10. Viivi Luik, „Ajaloo ilu" (8,8)

On tähelepanuväärne, et arvulised tulemused hakkasid kvalitatiivset analüüsi suunama autoripõhistes tsiteerimismustrites. Esialgu polnud plaanis tsitaadikasutust uurida teoseti ega autoriti. Statistiliselt ilmnes aga, et eri autorid on võtet kasutanud ootamatult erineval määral.

Näiteks selgus, et Tõnu Õnnepalu kolmes teoses („Piiririik", „Raadio", „Harjutused"), mis moodustasid 9,6% korpuse kogumahust, esines 616 avatud tsitaati, mis on 33% kõigist leitud tsitaadidest. Seevastu skaala teises otsas paiknev Rein Raua „Hector ja Bernard" ei sisaldanud ühtegi avatud tsitaati.



Joonis 1. Avatud tsitaatide suhteline sagedus teoseti 10 000 sõna kohta.

Sellest tuleneb kaks olulist järeldust. Esiteks, antud korpuses toob pelgalt ühe autori looming sisse suure kõikumise, mis annab enam teavet konkreetse autori stiilist kui ajastu üldpildist. Teiseks ärgitabki tsitaadirohkus tõlgendama leidu autori loomingu ja stiili kontekstis. Olles seda asjaolu varem ise uurinud, on siinkirjutaja jõudnud järeldusele, et Õnnepalu teostes on tegevustikku ja tegelaskonda minimaalselt, ent neid asendab intensiivne dialoog minajutustaja ja teiste diskursuste vahel. Kõrgendatud keeleline tundlikkus avaldub ohtras tsiteerimises ja eri stiilide (iroonilises) matkimises ning selle kaudu väljendatakse sageli oma hoiakut tsitaadiga esindatud maailmavaate või teema suhtes (vt lähemalt Meriste 2016).

Avatud tsitaate kvalitatiivselt eritledes tuli esile kaheksa peamist omadust, mille põhjal jaotati tsitaadid kategooriatesse ja kirjeldati neid. Sedakaudu sai täidetud üks

uurimiseesmärkidest: avatud tsitaadi kui kirjanduskeelelise võtte funktsioonide kirjeldamine. Siinses artiklis on kategooriate kirjeldused lühendatud kujul (pikemalt Meriste 2020: 38–50). Kategooriate kirjeldused tuginevad peale kvalitatiivse analüüsi statistilistele tähelepanekutele kategooriate kokkulangevusest ja esinemissagedusest, mis on siin esitatud pärast kategooriate kirjeldusi.

Järgnevalt kategooriatest lähemalt.

1. Avaliku diskursuse kategoorias kasutatakse tsiteerimisvõtet ametlik-poliitilise žargooni, meediakeele või rahvalike käibetõdede rõhutamiseks ja nende suhtes võõristustunde märkimiseks. Sageli kaasneb sellega iroonia või vastumeelsus, tsitaadi abil vastandatakse ja eristatakse kõneleva indiviidi ning kollektiivset, ametlikku, poliitilist mõtlemisviisi. Selle kategooria avatud tsitaadid paljastavad keele ideoloogilise laetuse, mis uuritud kirjanduses seostus enim totalitaarsele režiimile iseloomuliku keelekasutusega.

> Olin valmis otsima vähetasuva, kuid „moraalselt puhta" töö ning elama ainult tema nimel. (Sild 2007: 86)

> Kui tuhahunnikute ja söestunud korstnate keskelt saadad Saksa võimudele palvekirja, et keegi on olnud „kommunist puht teenistuse mõttes", kuid hea pottsepana on teda vaja, siis pisiasjade üle enam ei tingita. (Mihkelson 2007: 179)

2. Rolli kategooria avatud tsitaatide all mõeldakse inimestele omistatud identiteete tähistavaid sõnu, näiteks hüüdnimed, tiitlid, identifitseerivad omadussõnad, sotsiaalsed määratlused jms. Ka rolli kategooriasse kuuluvad tsitaadid märkisid eelkõige vastumeelsust antud rolli või sellega kaasnevate ootuste suhtes.

> Ja siis pidingi mina olema see, kes „ainsa mehena" Mõmmi korrale kutsub, läheb talle metsaserva järele ja viib kurva karu koju. (Kivirähk 2007: 146)

3. Deiksise kategooriasse kuuluvad iseseisva (kontekstist sõltumatu) tähenduseta keelendid, näiteks asesõnad. Deiktikud märgivad tekstisisest või tekstist väljapoole ulatuvat viitesuhet. Seda arvestades oli deiktikute kasutamine tsitaadina või selle osana üllatav leid: milleks rõhutada tsiteerimisvõttega midagi, mille tõlgendamiseks on igal juhul vaja tuge kontekstist? Selgus, et üheks oluliseks funktsiooniks on justnimelt tahe rõhutada, et välditakse otseütlemist ehk et luuakse teadlikult läbipaistmatust.

> Alevi poolt ei julgenud igatahes ükski hing nõmmele lilli korjama ega lambaid karjatama tulla, sest kumu veidrast king-olendist oli seal veel kümme korda jubedamaks paisutatud, kui asi tegelikult oli, nii et isegi Paap Mõrumeelest hoiti parema meelega kaugemale, sest tema ju „tollega" seal mässas. (Heinsaar 2001: 79)

> Naisega koos istugu Valeri autosse ning neid viiakse otsemaid „sinna kuhu tarvis". (Vetemaa 2001: 129)

4. P s e u d o t s i t a a d i k s on siinkohal kategoriseeritud sellised avatud tsitaadid, mis on otsekui ütluse vormis kujutlused kellegi või millegi hoiakutest, olemusest, kuigi neid ütlusi pole tegelikult kõlanud. Sellest tulenevalt on pseudotsitaadid illustratiivsed ning tihtilugu tugevalt stereotüüpsed. Pseudotsitaadiga on võimalik ütlusi omistada isegi kõnevõimetutele objektidele, nagu näoilme või ellusuhtumine.

> Nad lobisesid ning naeratasid üksteisele ühtmoodi võluvalt selle tülgastava „jah, meil on raskeid aegu olnud, aga koos saame me kõigest üle"-ilmega. (Hiram 1999: 332)

> Alles palju hiljem sain aru, et ründasin proua Bintat just selle end väikeseks inimeseks kuulutamise pärast, „keskmised-koju-tulevad"-eluvaate pärast. (Mihkelson 2007: 103–104)

Pseudotsitaadi abil saab üks tegelane esile manada kujutluse teise arvatavast reaktsioonist, mis näitab kujutatud tegelast selgelt stereotüüpses, isegi karikeerivas valguses. Teisisõnu peegelduvad selles pigem tsiteerija kujutlus ja rõhuasetus kui kujutlusobjekti tegelik olemus.

> Aimasin juba ette tema vastuväiteid, mis järgnevad kohe, kui ma ütlen, et olen väsinud ja tahan magama minna – „aga ta on ju su õde", „tema nii hoolitses sinu eest, kui sa väike olid", „me oleme üks pere, me peame üksteist aitama", „me ei tohi jääda võõraks". (Kivirähk 2007: 147)

5. I n t e r t e k s t i k s liigitati avatud tsitaadid, mille lähikontekstis oli selge viide tsitaadi verbaalsele allikale või mis on äratuntavad kultuuriviited. Intertekstuaalse avatud tsitaadi üks dominantseid funktsioone on tsiteeritud diskursuse kohta arvamuse avaldamine ja/või kõneleja distantseerumine sellest. Tavalise otsekõnena see nõnda efektselt ei ilmneks, sest avatud tsitaat on justkui ootamatu tähendusreljeefi tekitav kiil teise tegelase diskursuse sees.

> Ühed teadsid, et see on pikk blond, „ilus puhas eesti mees", nagu üks naine väitis, väga šikilt riides ja ülimalt rafineeritud. (Berg 2007: 158)

6. A l t e r n a t i i v t ä h e n d u s e all peeti silmas avatud tsitaate, mis modifitseerivad sõnade neutraaltähendust. Peamiselt tähistas see eufemisme, irooniat või muul viisil kujundlikkust. Need on head näited kirjanduskeele ambivalentsusest ja mängulisusest, mis jäikade määratlustega analüüsile hästi ei allu. Tsitaadimärkidega rõhutatakse moondunud või lisatähenduse olemasolu, samal ajal ei ole see otsesõnu välja öeldud ning jääb tahes-tahtmata sõltuma lugeja tõlgendusest.

> Sellepärast ma kadestangi vanu ja mõnikord, kui tulevik eriti rõhub, nagu nendel segastel päevadel, kui maailm tundub jälle vaikselt, aga nüri vältimatusega veerevat „õigusemõistmise", ühesõnaga pigem sõja kui rahu poole, kadestan ma ka surnuid. (Nigov 2002: 23)

7. M e t a t ä h e n d u s e g a avatud tsitaatide eesmärk on juhtida otseselt tähelepanu keelekasutusele. Selle abil rõhutatakse teadlikkust sõnade kaastähendustest, nende konnotatsioonidest, sõnade tavatähenduse küsimärgistamist. Enamasti kaasneb seda tüüpi tsitaadiga keelekasutuse otsesõnaline kommenteerimine lähikontekstis. Kuna üksik tsitaat esindab suuremat diskursust (nt poliitilist allkeelt), saab üksikute sõnade kaudu kujutada hoiakut kogu diskursuse suhtes.

> Mida rohkem räägiti sõltumatusest ja demokraatiast, maailmakorrast ja maailmarahust, seda kummalisemaid tähendusvarjundeid omandas sõna „demokraatia". (Mihkelson 2007: 319)

> „Ema", „kummipüksid", „kintsud" ja „külm ilm" kõlavad tema omastarust-juba-vilunud-kõrvale nii süütult, nagu võiksid nad tähendada eriti tumedaid ja kahtlasi asju. (Luik 1991: 92)

8. E p i t e e d i n a kategoriseeriti avatud tsitaadina esitatud täiendeid ja öeldistäiteid. Tsiteerimisvõtte abil tõstetakse rõhulisse positsiooni epiteedid, mis lause tavapärases semantilises struktuuris on teisejärgulised (st mitte põhisõnad). Sellisena tekib omamoodi antonomaasia efekt, mis on levinud lüroeepikas ja ka eesti rahvaluules (Laugaste 1975: 192) – omadus muutub pärisnimega võrdväärseks identifitseerijaks ja võib pärisnime isegi asendada. Uurimiskorpuses esines see valdavalt negatiivse sildistamisena.

> Kui Nõukogude armee asjus „kõlbmatuks", „mittekuuluvaks" ja „väljalülitatuks" tunnistamine kõrvale jätta. (Teder 2006: 169)

> Kas ma tõesti näen välja nii universaalselt „kohalik", et mind on peetud Moskvas venelaseks, Prahas tšehhiks, Hollandis hollandlaseks, Londonis inglaseks? (Pilv 2010: 216)

Enim esines meta- ja alternatiivtähenduse kategooria tunnustega tsitaate (vastavalt 30,9% ja 27,7%). Kõige vähem oli deiksise ja pseudotsitaadi tunnustega näiteid (vastavalt 7% ja 6,8%). Rolli kategooria tuvastati 22,8%, intertekst 21,6%, avalik diskursus 16,9% ja epiteet 11,5% juhtudest.

Kokkulangevust kajastab tabel 1.

Tabel 1. Kahe kategooria koosesinemised absoluut- ja suhtarvudena.

	Avalik	**Roll**	**Deiksis**	**Pseudo-tsitaat**	**Inter-tekst**	**Alter-natiiv**	**Meta**	**Epiteet**
Avalik	**314**	69	11	22	32	87	68	40
Roll	69	**423**	25	5	52	101	79	28
Deiksis	11	25	**130**	3	20	5	41	9
Pseudotsitaat	22	5	3	**127**	24	11	17	6
Intertekst	32	52	20	24	**401**	21	105	33
Alternatiiv	87	101	5	11	21	**514**	42	52
Meta	68	79	41	17	105	42	**573**	35
Epiteet	40	28	9	6	33	52	35	**214**

	Avalik	Roll	Deiksis	Pseudo-tsitaat	Inter-tekst	Alter-natiiv	Meta	Epiteet
Avalik	100%	22%	4%	7%	10%	28%	22%	13%
Roll	16%	100%	6%	1%	12%	24%	19%	7%
Deiksis	8%	19%	100%	2%	15%	4%	32%	7%
Pseudotsitaat	17%	4%	2%	100%	19%	9%	13%	5%
Intertekst	8%	13%	5%	6%	100%	5%	26%	8%
Alternatiiv	17%	20%	1%	2%	4%	100%	8%	10%
Meta	12%	14%	7%	3%	18%	7%	100%	6%
Epiteet	19%	13%	4%	3%	15%	24%	16%	100%

Kokkuvõte

Artiklis uuriti avatud tsitaati, mis on teksti tähendusstruktuuri tugevalt mõjutav stilistiline võte. Seda tehti kauglugemise kontseptuaalses raamistikus, mis ei vaatle kirjandust eraldiseisvate üksikteoste kogumina, vaid tekstuaalsete eneseväljendus-praktikate väljana, kus ilmnevad teatavad mustrid. Sellist vaatepunkti võimaldavad rakendada digimeetodid. Kauglugemise kasuks otsustati uurimisel eelkõige seetõttu, et üksikuid teoseid lähilugedes oleks võimatu koguda uuritava nähtuse kohta nii palju kasutusnäiteid, et nende põhjal saaks teha üldistusjõulise analüüsi. Et keskenduda detailile, tuleb see kontekstist välja filtreerida ja asetada kõrvuti suure hulga sama nähtuse esinemisjuhtudega mujal.

Selgus, et avatud tsitaat on vahend, mis loob ilukirjanduslikku lausesse justkui tähendusreljeefi: tsitaadina esile tõstetud tekstiosa on väliselt vaid väike detail, kuid selle abil tuuakse teksti lai seostevõrgustik, tühistatakse ühemõttelise tõlgendamise võimalus ja rõhutatakse kõrgendatud keelelist teadlikkust. Tsiteerimine võib sõna tähenduse muuta teisejärguliseks või isegi valeks – teisisõnu muuta selle tõeväärtust. Tsiteerimisvõttega saab midagi väita ja samal ajal end sellest distantseerida, näiteks selleks, et märkida irooniat. Üheks huvitavaimaks avatud tsitaadi tähenduskategooriaks osutus pseudotsitaat, mida siinkirjutaja teadmiste kohaselt ei ole varem teoreetiliselt käsitletud. Pseudotsitaat, otsekui verbaalne illustratsioon, asendab tavapärast seletust või kirjeldust. See lähtub tsiteerija kujutlusest, mida keegi või miski võiks öelda, iseloomustades sedakaudu nii illustratsiooni objekti kui ka suhtumist sellesse.

Jutumärkide kui diskursuste selgete eristajate ülesanne võib olla tsitaadiga esindatud inimese, maailmavaate, valdkonna vms võõristamine ja varjatud või ilmse seisukoha väljendamine. Näidetest ilmnes, et selle üks oluline ülesanne taasiseseisvumisjärgses eesti kirjanduses on ühtse väärtussüsteemi eitamine ja tõe relatiivseks muutmine, milles võib näha reaktsiooni eelnenud 50 aastale, nõukogude „ühe tõe" režiimile. Keel annab edasi kategooriaid, mis struktureerivad ka arusaamu maailmast, ning avatud tsitaat võimaldab protestida vaikimisi kinnistunud standardite vastu: olgu nendeks sõnaga omistatud identiteet ja sellega kaasnevad ootused või kollektiivsed käibetõed. Sellisena on avatud tsitaat selgelt seotud subjektikeskse vaate esiletõusuga.

Kvalitatiivset ja kvantitatiivset lähenemist kombineeriva analüüsi tulemusena sai kaardistada ühe stilistilise võtte olemust. Ühtpidi saab sel viisil täiendada seniseid arusaamu uurimismaterjali moodustanud kirjandusteostest. Kuid lisaks saaks neid teadmisi kasutada uue uurimismaterjali juures, mis ei pea olema tingimata ilukirjanduslik. Näiteks ajakirjanduslikus kontekstis kasutatakse avatud tsitaati selliste suhtumiste varjamiseks, mida otsesõnu välja ei soovita öelda, või väite infostruktuuri hägustamiseks.

Kvantitatiivsete ja kvalitatiivsete meetodite kombineerimine annab digihumanitaaria uurimuste alusoskuse: kvalitatiivne hüpotees tuleb kohandada kvantitatiivse lahenduskäiguga ja seejärel jõuda omakorda kvalitatiivsete järeldusteni.

Sellega seoses loodan, et arvutuslikud meetodid saavad peagi humanitaarteaduste õppeprogrammi loomulikuks osaks, kuid veel enam näen vajadust valdkonnaülese koostöö järele. Käesolev uurimus on üksik katseprojekt, mis ei esinda efektiivset ja jätkusuutlikku viisi selles valdkonnas edasi liikumiseks. Vaja on luua igapäevaselt koos töötavaid, mõtlevaid, loovaid uurimisgruppe, kuhu kuuluvad keele-, arvuti- ja kirjandusteadlased ning teised asjakohased eksperdid. See võimaldaks keskenduda mitme formaalse näitaja paralleelsele uurimisele ja näha muutujatevahelist suhet – näiteks antud uurimuse põhjal, kas rohke tsitaadikasutus on korrelatsioonis autori sõnavara rikkalikkuse, lauseehituse keerukuse või muu taolise näitajaga. Usun, et tähendust ei mõjuta niivõrd ühe elemendi esinemisviis, kuivõrd koosesinevate elementide omavaheline suhe ning nende suhete muutumine ajateljel. Digitaalne areng on andnud selliste dünaamiliste tähendusvõrgustike uurimiseks vajalikud tehnilised vahendid, nende kasutusvõimaluste avastamine on uurijate teha.

VEEBIVARAD

Distant Reading for European Literary History. https://www.distant-reading.net
Sketch Engine. http://www.sketchengine.eu
Voyant Tools. http://voyant-tools.org

KIRJANDUS

AutÕS = Autoriõiguse seadus. Vastu võetud 11. XI 1992, jõustumine 12. XII 1992.

Bakhtin, Mikhail 1982. Discourse in the novel. – The Dialogic Imagination: Four Essays by M. M. Bakhtin. Toim Michael Holquist. Austin: University of Texas Press, lk 244–245.

Berg, Maimu 2007. Unustatud inimesed. Tallinn: Tuum.

Brendel, Elke; Meibauer, Jörg; Steinbach, Markus 2011. Exploring the meaning of quotation. – Understanding Quotation. (Mouton Series in Pragmatics 7.) Toim E. Brendel, J. Meibauer, M. Steinbach. Berlin: De Gruyter, lk 1–33. https://doi.org/10.1515/9783110240085.1

Brik, Ossip 2014. Nn „Vormimeetod". – Kirjandus kui selline. Valik vene vormikoolkonna tekste. (Gigantum Humeris.) Koost ja toim Märt Väljataga. Tallinn: Tallinna Ülikooli Kirjastus, lk 46–47.

Capone, Alessandro 2013. The pragmatics of quotation, explicatures and modularity of mind. – Pragmatics and Society, kd 3, lk 259–284. https://doi.org/10.1075/ps.4.3.01cap

Eder, Maciej; Rybicki, Jan; Kestemont, Mike 2016. Stylometry with R: a package for computational text analysis. – The R Journal, kd 8, nr 1, lk 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Genette, Gérard 1992. Palimpsestes: la littérature au second degré. Paris: Seuil.

Heinsaar, Mehis 2001. Vanameeste näppaja. Tallinn: Tuum.

Hiram 1999. Mõru maik. Tallinn: Tuum.

Kirsch, Adam 2014. Technology is taking over English departments: The false promise of the digital humanities. – The New Republic 2. V.

Kivirähk, Andrus 2007. Mees, kes teadis ussisõnu. Tallinn: EKSA.

Kristeva, Julia 1980. Word, dialogue, and novel. – J. Kristeva, Desire in Language: A Semiotic Approach to Literature and Art. Toim Leon S. Roudiez. New York: Columbia University Press, lk 64–91.

Kultuuriministeerium 2018 = „Kultuuripärandi digiteerimine 2018–2023" tegevuskava. Tallinn: Kultuuriministeerium. https://www.kul.ee/kultuurivaartused-ja-digitaalne-kultuuriparand/digitaalne-kultuuriparand/kultuuriparandi (15. II 2021).

Laugaste, Eduard 1975. Eesti rahvaluule. Tallinn: Valgus.

Lotman, Juri 1991. Kunst modelleerivate süsteemide reas. – J. Lotman, Kultuurisemiootika. Tekst – kirjandus – kultuur. Tlk Pärt Lias, Inta Soms, Rein Veidemann. Tallinn: Olion, lk 8–31.

Luik, Viivi 1991. Ajaloo ilu. Tallinn: Eesti Raamat.

Meriste, Maarja Helena 2016. Võõras kõne Tõnu Õnnepalu „Hinnas". Bakalaureusetöö. Tartu Ülikool, kirjanduse ja teatriteaduse osakond.

Meriste, Maarja Helena 2020. Eesti 1990.–2010. aastate kirjanduse kauglugemise katse avatud tsitaatide kasutusmustrite näitel. Magistritöö. Tallinna Ülikool, humanitaarteaduste instituut.

Mihkelev, Anneli 2005. Vihjamise poeetika. (Collegium litterarum 19.) Tallinn: Underi ja Tuglase Kirjanduskeskus.

Mihkelson, Ene 2007. Katkuhaud. Tallinn: Varrak.

Moretti, Franco 2013. Distant Reading. London–New York: Verso.

Nigov, Anton [= Tõnu Õnnepalu] 2002. Harjutused. Tallinn: EKSA.

Pilv, Aare 2010. Ramadaan. Vahtsõliinah, Lillekülas, in Matiano, a Roma. Tallinn: Tuum.

Piper, Andrew 2018. Are novels getting easier to read? – TXTLAB 10. I. https://txtlab.org/2018/01/are-novels-getting-easier-to-read (28. VI 2021).

Pruul, Kajar 2000. Vana sajandi uued luuletajad. – Varjatud ilus haigus. Valik sajandilõpu eesti luuletajaid. Koost K. Pruul. Tartu: Eesti Kostabi Selts, lk 175–207.

Puik, Katrin 2009. Iroonia Heiti Talviku ja Betti Alveri luules. (Dissertationes litterarum et contemplationis comparativae Universitatis Tartuensis 8.) Tartu: Tartu Ülikooli Kirjastus.

Recanati, François 2001. Open quotation. – Mind, kd 110, nr 439, lk 637–687. https://doi.org/10.1093/mind/110.439.637

Schöch, Christof; Erjavec, Tomaž; Patras, Roxana; Santos, Diana 2021. Creating the European Literary Text Collection (ELTeC): Challenges and perspectives. – Modern Languages Open [ilmumas].

Sild, Ivar 2007. Tantsiv linn. Tallinn: Tuum.

Sobchuk, Oleg 2018. Charting Artistic Evolution: an Essay in Theory. (Dissertationes litterarum et contemplationis comparativae Universitatis Tartuensis 18.) Tartu: Tartu Ülikooli Kirjastus.

Tangherlini, Timothy 2019. Macroscopic Methods for Culture Analytics. Loengukursus. Tartu Ülikool, digihumanitaaria ja infoühiskonna keskus, 3.–13. IX.

Teder, Tarmo 2006. Onanistid. Tallinn: Tuum.

van de Ven, Inge 2017. Creative reading in the information age: Paradoxes of close and distant reading. – The Journal of Creative Behavior, kd 53, nr 1. https://doi.org/10.1002/jocb.186

Velsker, Mart 2001. Kirjandus ja ühiskond. – Epp Annus, Luule Epner, Ants Järv, Sirje Olesk, Ele Süvalep, M. Velsker, Eesti kirjanduslugu. Tallinn: Koolibri, lk 608–623.

Vetemaa, Enn 2001. Neitsist sündinud. Tallinn: Tänapäev.

Viiding, Elo 2008. Püha Maama. Tallinn: Tuum.

Wilkens, Matthew 2012. Canons, close reading, and the evolution of method. – Debates in the Digital Humanities.Toim Matthew K. Gold. Minneapolis–London: University of Minnesota Press, lk 249–258. https://doi.org/10.5749/minnesota/9780816677948.003.0026

LISA. TEOSED UURIMISKORPUSES

Annus, Epp 2007. Sina, Matilda. Tallinn: Tuum.
Berg, Maimu 2007. Unustatud inimesed. Tallinn: Tuum.
Bristol, Piret 2007. Paralleelmeri. Tallinn: Tuum.
Ehlvest, Jüri 1996. Krutsiaania. Tallinn: Tuum.
Ehlvest, Jüri 2005. Palverännak. Tallinn: Tuum.
Ernits, Marje 2003. Õpetaja, õpetaja. Tallinn: Tuum.
Ernits, Marje 2008. Hoia mu kätt. Tallinn: Tuum.
Friedenthal, Meelis 2005. Kuldne aeg. Tallinn: Tuum.
Heinsaar, Mehis 2001. Vanameeste näppaja. Tallinn: Tuum.
Hellerma, Kärt 2007. Ma armastasin David Copperfieldi. Tallinn: Tuum.
Hiram 1999. Mõru maik. Tallinn: Tuum.
Jõerüüt, Jaak 2005. Sõber. Tallinn: Tuum.
Kaplinski, Jaan 2000. Silm. Tallinn: Tänapäev.
Kaplinski, Jaan 2000. Hektor. Tallinn: Tänapäev.
Kaus, Jan 2003. Õndsate tund. Tallinn: Tuum.
Kaus, Jan 2009. Hetk. Tallinn: Tuum.
Kender, Kaur 2010. Comeback. Tallinn: EKSA.
Kesküla, Kalev 2009. Elu sumedusest. Tallinn: Tuum.
Kivirähk, Andrus 1999. Liblikas. Tallinn: Tuum.
Kivirähk, Andrus 2000. Rehepapp, ehk, November. Tallinn: Varrak.
Kivirähk, Andrus 2001. Ivan Orava mälestused, ehk, Minevik kui helesinised mäed. Tallinn: Varrak.
Kivirähk, Andrus 2007. Mees, kes teadis ussisõnu. Tallinn: EKSA.
Kivirähk, Andrus 2010. Pagari piparkook: Kivirähk XX sajandil. Tallinn: EKSA.
Lanto, Ari 2010. Aken. Tallinn: Eesti Raamat.

Luik, Viivi 1991. Ajaloo ilu. Tallinn: Eesti Raamat.
Mihkelson, Ene 1994. Nime vaev. Tartu: Ilmamaa.
Mihkelson, Ene 1996. Surma sünnipäev. Tallinn: Tuum.
Mihkelson, Ene 2001. Ahasveeruse uni. Tallinn: Tuum.
Mihkelson, Ene 2007. Katkuhaud. Tallinn: Varrak.
Nigov, Anton 2002. Harjutused. Tallinn: EKSA.
Pilter, Lauri 2004. Lohejas pilv. Tallinn: Tuum.
Pilter, Lauri 2010. Retk Rahemäkke. Tallinn: Tuum.
Pilv, Aare 2010. Ramadaan. Vahtsõliinah, Lillekülas, in Matiano, a Roma. Tallinn: Tuum.
Raud, Rein 2004. Hector ja Bernard. Tallinn: Tuum.
Raud, Rein 2008. Vend. Tallinn: Tuum.
Raudam, Toomas 1996. Muusika kuum sai. Tallinn: Tuum.
Reimus, Katrin 2008. Kus tramm ei käi. Tallinn: Tuum.
Remsu, Olev 2008. Musketäride muundumised. Tallinn: EKSA.
Ruben, Aarne 2001. Volta annab kaeblikku vilet. Tallinn: Tänapäev.
Saat, Mari 2008. Lasnamäe lunastaja. Tallinn: Tuum.
Sauter, Peeter 2008. Beibi bluu. Tallinn: Tuum.
Sild, Ivar 2007. Tantsiv linn. Tallinn: Tuum.
Tangsoo, Jaan 2007. Marukoer. Tallinn: Tuum.
Teder, Tarmo 2006. Onanistid. Tallinn: Tuum.
Tode, Emil 1993. Piiririik. Tallinn: Tuum.
Tode, Emil 2002. Raadio. Tallinn: EKSA.
Vaher, Berk 2002. Lugulaul. Tallinn: Tuum.
Vaher, Berk 2004. Sekeldaja päevad. Tallinn: Tuum.
Vetemaa, Enn 2001. Neitsist sündinud. Tallinn: Tänapäev.
Viiding, Elo 2008. Püha Maama. Tallinn: Tuum.

Maarja Helena Meriste (snd 1992), MA, ajakirja Värske Rõhk toimetaja (Vanemuise 19, 51003 Tartu), maarjahelena.meriste@va.ee

Distant reading of literary language: Developing a methodology using quotations as an example

Keywords: digital humanities, distant reading, quotations, literary style, formalism

This paper aims to take advantage of modern digital methods to address research topics that, until now, have mostly been subjects of assumption and interpretation based on rather few examples.

More specifically, the purpose of this article is to explore some usage patterns of open quotation as a literary device and its impact on meaning. Open quotation, as defined by the French philosopher François Recanati, is a type of quotation where

a string of words in an utterance is marked with quotation marks to imply the presence of multiple or deformed meanings.

Alongside with other stylistic devices like parody, irony and pastiche, open quotation is regarded as a distinctive feature of postmodern literary language, yet it has not been empirically studied. One of the reasons may be the fact that it is difficult to obtain a systematic view of such textual features by means of traditional, i.e., close reading. However, the emergence of digital methods in humanities provide better access to such research problems. In literary studies, application of computational methods is associated with Franco Moretti's approach of distant reading. However, distant reading does not offer a specific method to solve complex research problems in literary studies. Therefore, the second purpose of this study is to create a suitable methodology to best answer quotation-related research questions.

In studies of Estonian culture, the emergence of postmodern aesthetics has been associated with the collapse of the Soviet system at the end of the 1980s. Therefore, the text corpus of this study consists of 50 Estonian novels published in 1990–2010. All cases of open quotation were extracted and divided into 8 categories based on qualitative features, and then explored in detail.

Implying metalinguistic awareness was identified as the most important function of open quotation. Most commonly, it was used to reveal Soviet era related totalitarian ideologies, judgments and values inherent in language. Another interesting type of quotation previously undescribed in theoretical literature was distinguished: the category of pseudo-quotations aiming to illustrate both animate and inanimate objects by attributing to them an imaginary verbal act. In general, the use of open quotations reflects the decay of a monolithic value system by undermining the possibility of there being only one correct interpretation.

Regarding methodology, computational methods proved to be effective in establishing high-level stylistic patterns. However, the patterns alone remained inconclusive, requiring qualitative analysis to reach meaningful conclusions. This suggests a fruitful way to approach complex research questions in humanities: first, a qualitative hypothesis is transformed into a specific quantitative method; secondly, the obtained results are interpreted qualitatively.

Maarja Helena Meriste (b. 1992), MA, editor of literary magazine Värske Rõhk (Vanemuise 19, 51003 Tartu), maarjahelena.meriste@va.ee

https://doi.org/10.54013/kk764a5

Digitaalne lugemine ja humanitaarharidus

MAARJA OJAMAA, PEETER TOROP, ALEKSANDR FADEEV,
ALEXANDRA MILYAKINA, MERIT RICKBERG

Digitaalne lugemine on kujunemas humanitaarteaduste võtmemõisteks. Oluliseks tõukeks sellele on tarkvara arenguga kaasnevad võimalused, mis on humanitaaridele harjumuspärase lähilugemise kõrvale toonud kauglugemise, suurte tekstikorpuste ning metaandmete tõlgendamise jm. Digitaalse lugemise eripäraks on aga asjaolu, et see sisaldab peale sõnaliste tekstide lugemise ka vaatamist ja kuulamist (vt Ojamaa, Torop 2020). Väidame, et üleminek raamatukeskselt kultuurilt veebikeskselt kultuurile on lugemise just sel moel varasemast hõlmavamaks muutnud. Samuti on digitaalses keskkonnas aktuaalseks saanud Juri Lotmani osutus, et lugemist ei peaks käsitlema mitte teksti ammendava tarbimisena, vaid loova suhtlusena teksti kui dialoogipartneriga (Lotman 2016 [1981]: 236). Erinevad teksti tõlgendamise viisid, sealhulgas erinevad modaalsused ja meediumid, on veebikeskkonnas loomulikud ja levinud. See on omakorda kaasa toonud ühe suurima digitaalse lugemisega seotud probleemi: tekstide ja tähenduste pihustumise hoomamatuks paisunud infokülluses. Hariduslikust vaatepunktist on aga tähtis just tervikuks seotud teadmus. Kultuurilise identiteedi seisukohast on eriti oluline seoste nägemine kultuuripärandi ja tänapäeva kultuuriruumi vahel ehk oskus astuda näiteks kirjandusklassikaga dialoogi ka digitaalsete kanalite vahendusel.

Digitaalse lugemise teaduslik analüüs on tihedalt seotud digihumanitaaria üldise arenguga. Kui algselt kaasnes digihumanitaariaga kvalitatiivsete analüüside rikastamine kvantitatiivsete meetoditega ning avarama tausta tekitamisega korpuslingvistika ja andmebaaside kasutamise kaudu, siis digihumanitaaria n-ö teine laine on ümber kujundamas arusaamu kvalitatiivsest analüüsist humanitaarias. Loetavate tekstide digitaalne vahendamine on muutnud nii tekstide olemuse kui ka lugemise käsitlemist. Jeffrey Schnappi ja Todd Presneri „Digihumanitaaria manifest 2.0" lähtub põhimõttest, et digihumanitaaria „tähelepanu komplekssusele, meediumipärasusele, ajaloolisele kontekstile, analüütilisele sügavusele, kriitikale ja tõlgendamisele" (Schnapp, Presner 2009: 2) rikastab humanitaariat, täiendab traditsioonilisi uurimismeetodeid ja digihumanitaariat ei peakski käsitlema millegi eraldiseisvana, vaid traditsioonilise humanitaaria osana (vt ka Svensson 2012). Selline tõlgendus võib küll kaasa tuua metodoloogiliste põhimõtete muutumise (näiteks digiteksti käsitlemisel tekstina) ning distsiplinaarse hübriidsuse kujunemise teaduses ja postdistsiplinaarsuse juurdumise ülikoolides (Berry 2012). Tarkvara kui kirjutamise ja lugemise uudne mõõde paneb analüütikud mõtlema võimalikest muutustest kirjanduskultuuris (Murray 2018), teadustöös (Kembellec, Broudoux 2017) ja hariduses (Tso 2019). Selles protsessis tõuseb taas keskmesse lugemise ja tõlgendamise lahutamatus õppimisest, uute kirjaoskuste omandamisest (Kress 2014: 66). Nii

digitaalne lugemine kui ka tekstide varieerumine ja hajumine digikultuuris tõstatavad analüütiku ees uudse küsimuse: kuidas suhestuvad osalugemine ja tervikulugemine tekstide esitlemise ja pihustumise viisidega?

Artiklis käsitleme tekstide pihustumist kolmest vaatepunktist. Esimeses peatükis selgitame väljakutse kompleksset olemust digitaalse tekstikommunikatsiooni kolme peamise tunnuse kaudu: multimodaalsus (sõnaliste ja pildiliste kanalite kombineerimine), varieeruvus (ühe teksti esinemine paljudes versioonides) ja killustumine (terviktekstide asemel fragmentide lugemine). Nendest tuleneva pihustumise tasakaalustamiseks näeme võimalusi terviklikumalt kontseptualiseeritud digitaalsete platvormide loomises. Teises peatükis tutvustame pihustunud tekstiosade põhjal tähendusliku sidususe loomise mehhanisme, mõtestades digitaalseid platvorme kultuurisemiootilise tekstikäsitluse kontekstis. Täpsemalt väidame, et tervikuloome lähtekohtadena tuleks arvesse võtta teksti/platvormi kolme parameetrit: väljenduslikkust, piiritletust ja struktureeritust. Kolmandas peatükis jõuame konkreetsete näideteni, kuidas oleme digitaalseid vahendeid kasutades toetanud temaatiliste tervikkujutelmade tekkimist oma uurimisrühma loodud õppematerjalides. Siin läheneme probleemile mudeliloome kolme semiootilise põhitüübi ehk mütoloogilise, teadusliku ja mängulise modelleerimistüübi abil.

Tartu Ülikooli semiootika osakonnas tegutseva transmeedia uurimisrühmaga oleme digitaalsete keskkondade ja (keskkooli tasemel) humanitaarhariduse seostega tegelenud nii teoreetilise uurimistöö raames (Fadeev 2020; Milyakina jt 2020; Ojamaa jt 2019; Rickberg 2017; Torop 2019) kui ka praktiliselt, luues ja arendades avatud juurdepääsuga digitaalseid õppeplatvorme (https://haridusekraanil.ee). Nende platvormide keskmeks on olnud mõne (kirjandus)teose ekraniseering, mida tõlgendatakse erinevate kultuuriliste seoste abil. Haridus Ekraanil veebilehe kaudu on seni ligipääsetavad platvormid „Kirjandus ekraanil" (Andrus Kivirähki „Rehepapp ehk November" ekraniseerimise kontekstis), „Ajalugu ekraanil" (Leelo Tungla „Seltsimees laps" mineviku vahendamise kontekstis), „Identiteet ekraanil" (A. H. Tammsaare „Tõde ja õigus" enesekirjeldamise kontekstis), „Loodus ekraanil" (Jaan Tootseni film „Olemise ilu", mis portreteerib Fred Jüssit, looduse vahendamise kontekstis) ja „Kinoklassika ekraanil" (Oskar Lutsu „Kevade" ja tekstide püsimine kultuurimälus). Nende materjalide eesmärk on pakkuda välja sisukaid võimalusi kultuuripärandiga dialoogide aktiveerimiseks, kasutades selleks kaasaegseid tehnilisi vahendeid ning digitaalsele keskkonnale omaseid tekstiloome ja tõlgendamise viise.

1. Digitaalne lugemine: divergents ja konvergents

Digitaliseerumise protsessi iseloomustavad Lev Manovichi (2003) järgi teadmuse digitaalne levik, tarkvara kontroll andmete üle, uute (digitaalsete) ja vanade (analoog-) kultuurikonventsioonide segunemine, algoritmide kiirenenud toime, metameedia ning uus esteetika. Need tunnused ei puuduta üksnes algselt digitaalsetena loodud tekste (nt veebipõhiseid hüpertekste), vaid ka digiteeritud teoseid ja nende täiendatud versioone. Ent ainult tekstid ise ei ole teisenenud, vaid keerulisemaks

on muutunud ka tekstidevahelised suhted. Raamatupõhisele (humanitaar)haridusele on need muutused esitanud väljakutse ning kasvatanud vajadust lõimida formaalsesse õppesse informaalse ehk huvipõhise koolivälise õppimise elemente. Koolivälises elus domineerib praegu just digitaalne keskkond iseomaste ja silmapaistvalt mitmekesiste lugemistavadega. Digitaalses kultuuriruumis aset leidvate muutuste mõtestamise üks mõistepaare on *divergents–konvergents*, mille üks pool märgib lahknemist/eristumist ja teine pool ühendamist/sünteesi. Eeskätt Henry Jenkinsi uurimuste kaudu laiemasse kasutusse jõudnud terminitega osutatakse enamasti sõnumite edastuskanalite mitmekesistumisele olukorras, kus riistvara tasandi meediasüsteemide piirid sulanduvad ning „meedia sisu voolab sujuvalt nende vahel" (Jenkins 2006: 282–287). Nii on nutitelefoni koondunud muusika- ja videoedastamise, teksti- ja pilditöötluse jm masinad. Telefon, televiisor, kirjutusmasin, grammofon, fotoaparaat ja -album, videokaamera ja -ekraan, ajalehed ja raamatud on otsekui kokku sulatatud üheks multifunktsionaalseks digitaalseks seadmeks. Nende seadmete laialdane kättesaadavus on samal ajal aga mitmekesistanud tekstivorme ja nende lugemisviise.

Jenkinsi (2006: 282) argument on, et tänapäeva meediamaastikul domineerivad just konvergeerumisprotsessid. Mõnevõrra nüansseerituma käsitluse on välja pakkunud Gunnar Liestøl (2007), kes vaatleb divergentsi ja konvergentsi kui sümmeetrilisi protsesse, mis seisnevad olemasoleva terviku osadeks lahtivõtmises ning nende ümberkombineerimises uueks tervikuks. See iseloomustab ühtlasi igasugust innovatsiooni. Liestøl kirjeldab neid vastastiksõltuvuses protsesse riistvara, tarkvara ja nn tähendusvara (ingl *meaningware*) ehk teksti ja diskursuse tasandil. Ta möönab ühtlasi, et esimesel kahel tasandil toimuva kohaldamine kolmandale tasandile oleks liialt lihtsustav, kuid sellegipoolest avalduvad ka digitaalsetes tekstides riist- ja tarkvara protsessidega seotud mõjud ja piirangud. Kirjanduse vaatepunktist on aga väidetud (Brillenburg Wurth jt 2013), et just divergents on juhtiv protsess, kuid mitmekesistumine ei seisne üksnes digitaalsete tekstitüüpide plahvatuslikus kasvus. Tekstikultuur laiemalt on kaugel raamatute surmast, mida 1990-ndatel levinud hüpertekstide vaimustuses ennustati, ning paberraamatute vallas on samuti toimunud mitmekesistumine. Seda mitmekesistumist võib tõepoolest näha digitaalse tehnoloogia levikust tulenevana, kuid materjali või tehnoloogia tasandil sulandumise asemel avaldub paberraamatutes kirjaliku teksti visuaalsete ja taktiilsete ning laiemalt kehaliste aspektide kasvav roll tekstiterviku tähendusloomes ja mõjus. Järgnevalt vaatleme lähemalt kolme tegurit, mis suunavad divergeerumisprotsesse digitaalses keskkonnas.

1.1. Multimodaalsus

Tekstide esitamist digitaalses keskkonnas iseloomustab niisiis modaalne mitmekesisus. Kui veebileht vahendabki üksnes sõnalist sisu, on selle visuaalsed aspektid dominantsemad kui sama teksti sisaldava paberkandja puhul. Üldjuhul aga kombineerivad nii veebilehed kui ka kõikvõimalikud teised digikeskkonnad sõnalisi ja pildilisi modaalsusi.

Digitaalsete keskkondade loomuomane multimodaalsus rõhutab ka erinevate meediumide vastastikust seotust ning nende täpse eristamise keerukust. Olukorras, kus kirjandustekst ei ole koos üksnes illustratsiooniga, vaid ka näiteks lugemise taustaks loodud helikujunduse ja samuti terviklike ekraniseeringutega, tekib küsimus teksti piiridest. Hariduslikust vaatepunktist on oluline mõista, kuidas erinevad teineteisest loetu ja vaadatu mäletamine, millised teksti elemendid ja tasandid alluvad meediumidevahelisele ülekandele (näiteks ekraniseerimisele) paremini, millised üldse mitte. Vaadeldes kirjandust mitte suletud süsteemina, vaid laiema meediamaastiku või kultuuriliste vahenduskanalite hierarhilise võrgustiku osana, on võimalik ka kirjanduse õpetamisel kasutada erinevaid meediume. Seda näiteks siis, kui ühe teksti eri versioone mõista üksteist täiendavatena, mitte asendavatena, ehk kui lähtuda digitaalse ja trükiteksti komplementaarsest käsitlusest (vt Pennington, Waxler 2018; Shibata, Omura 2020). Tekstilise ja meediumilise variatiivsuse mõtestamiseks on endiselt omal kohal klassikalised tekstioperatsioonid, nagu lisamine, kustutamine, asendamine ja ümberkombineerimine. Nende kõrval on sama olulised nii meediumide kui ka diskursuste eripärade teadvustamine, sest eeskätt klassikaliste, kultuurikaanonisse kuuluvate kirjandustekstidega tutvutakse sageli just mõne pildilise või helilis-pildilise tõlgenduse kaudu, mitte algteost lugedes. Analoogmeediaga võrreldes pea piiritu mälumaht eri versioonide säilitamiseks ja kõrvutamiseks tähendab, et internetis on võimalik koostada õpirada iga õppija individuaalsetest eelistustest ja teadmistest lähtuvalt. Selle eelduseks omakorda on materjali eelnev korrastamine, mis on üha selgemalt peamine tänapäevane oskus.

Digitaalsete platvormide võimalused lähendada tekstikogemust multimodaalsele ja -sensoorsele maailmakogemusele toovad aga kaasa ohud, mille seas olulisemad on tõlgendusruumi vaesestamine, tähelepanu hajumine liigsete stiimulite tõttu ning informatsiooniline müra või saaste, mis takistab mentaalse terviku loomet (vt ka Toolan 2010: xviii). Üks võimalus neid ohte vältida on digitaalse lugemisega seotud oskuste sidus arendamine (vt lähemalt ptk 2.4).

1.2. Varieerimine (versioonide paljusus)

Tänu laialdaselt kättesaadavale tehnoloogiale on tekstide töötlemine lihtne ning digitaalses keskkonnas on sageli keeruline vahet teha, kas puutume esmalt kokku algteksti või juba muudetud versiooniga. See ei puuduta üksnes nn osaluskultuuri (Jenkins 2009; Burgess, Green 2018), vaid samamoodi lisatakse sageli näiteks DVD-le ametliku stuudioversiooni kõrvale nn režissööri montaaž, mis peegeldab filmilavastaja isiklikku nägemust. Filmiarhiivides on filmide eri versioonide, remikside jms jaoks võetud kasutusele metaandmete standard EN 15907.

Varieeruvus võib tuleneda ka teksti interaktiivsusest ehk lugeja võimalusest ise oma lugemisrada kujundada. Lihtsaim näide on hüpertekstid, mille osade järgnevuse määrab lugeja hüperlinkide kaudu. Keerukamate näidete hulka kuuluvad võimalused teksti ise vabalt muuta, näiteks eri tegelaste vaatepunkti vaheldades või tehisintellektile tugineva vestlusrobotiga dialoogi asudes (nt „Versu", „Tendency to Spill"). Oluline interaktiivne tekstikogemus on ühtlasi märksõnapõhine tekstiotsing,

mis võimaldab kiiresti huvipakkuva tekstikohani jõuda või korduvate märksõnade kontekste kõrvutada. Peale lugemisraja kujundamise pakub digitaalne keskkond avaramaid võimalusi loetu kommenteerimiseks nii arvamust väljendava piltmärgina kui ka sõnalise tagasisidena, sh teistele lugejatele suunatud soovitustena jne. Kommenteeritud versiooni lugemine aga kujundab hoiakuid ja suunab tähelepanu keset ning erineb seega kommenteerimata versiooni lugemisest.

Digitehnoloogia kättesaadavus on muutnud ka traditsioonilist (üksiku) autori ja (kollektiivse) vastuvõtja suhet, mis nüüd kaldub osaluskultuuri poole. Inglise keeles on käibele tulnud uudismõisted *prosumer* (Ritzer, Jurgenson 2010) ja *produser* (Bruns 2009), mis kombineerivad sõnu *tootja* (*producer*), *tarbija* (*consumer*) ja *kasutaja* (*user*). Olgugi lugemine klassikalises arusaamas üksildane tegevus, ei ole selle sotsiaalne mõõde kunagi puudunud, kuid digitaalsel ajastul on see nähtavam (Spink 2014). Lugemises kui ühistegevuses võib eristada kommenteerimist (Genius, Vikipeedia), fänniloomingut ning allikate uuskasutust (meemid, TikTok). Seejuures on lõpptulemusest sageli olulisem just interaktiivne protsess (Scolari jt 2018: 809). Kõige selle mõjul on kirjanduskultuur muutnud lugeja võimekamaks ja „demokraatlikumaks", nagu on hüpertekstide põhjal täheldanud Anne Mangen ja Adriaan van der Weel (2017: 167).

Viimasel ajal on digitaalse lugemise peamisi väärtusi hakatud pidama kitsaskohtadeks, mis õõnestavad oodatud kasu teksti mõistmisel või meelde jätmisel. Vastupidiselt laialdaselt eeldatule ei igatse valdav osa lugejatest aktiivset rolli teksti konstrueerimisel ning interaktiivsed digitaalsed narratiivid mõjuvad pigem eksitavalt ja segadusse ajavalt (Mangen, van der Weel 2017: 173). Otsuste tegemise sund üksikute osade kokkusobitamisel häirib kaasatust ning võib röövida lugemisnaudingu, mis seisneb ennekõike võimaluses loosse sukelduda ja kunstiväline tegelikkus otsekui ära unustada (Mangen, van der Weel 2017: 171).

1.3. Killustatus

Traditsiooniliselt lähtub kirjanduse õpetamine kirjandusteosest kui autonoomsest tervikust, ent digitaalsele keskkonnale omane tekstide ja nende osade pihustumine erinevate meediumide ja modaalsuste vahel esitavad sellisele lähenemisele väljakutse. Kui kirjandustekstide lugemine ja õpetamine eeldab üldjuhul tähelepanu hoidmist ühel objektil, siis digitaalne keskkond soosib ümberlülitumist ja oskust jagada tähelepanu erinevate objektide, modaalsuste ja meediumide vahel. Internetis lugemist on Jenkins (2010) kirjeldanud kui ristmeedialist protsessi, mille käigus lugeja tegutseb otsekui kütt-korilane, kes järgneb loetavale mööda erinevaid meediaplatvorme, kuniks saak koos.

Sellise lugemisviisi eelduseks on tekstide n-ö modulaarne esitus ümberkombineeritavate allosadena, mis samal ajal moodustavad ise madalama taseme tähendusliku terviku. Teisalt tähendab see aga killustumist, mis ongi digitaalse lugemise kontekstis üha sagedamini kõlav märksõna. Kui raamatukultuuri moodustavad eeskätt terviktekstid, siis digitaalne kirjanduskultuur põhineb paljuski fragmentidel, kokkuvõtetel, mille hulgas on ka videoformaadis raamatutreilerid jt ülilühikesed nn

nanotekstid (Scolari 2013). Kõik need on kokkusurutud versioonid, mille loomine algab teksti dominandi määratlemisega valitud vaatepunktist. Dominandist lähtuv versioon on omakorda terviku rekonstrueerimise aluseks ja nii võib ühe teksti põhjal olla kultuuris käibel mitmeid erinevaid, ka omavahel vastukäivaid tervikuid. Sellist põhimõtet on filmide ja videote turundamisel ära kasutanud Netflix, luues ühest filmist erinevad pildid/fragmendid, mille põhjal veendakse eri kasutajaid seda filmi vaatama, lähtuvalt nende varasematest valikutest, ehk taktikaliselt tuuakse esile nt näitlejaid, teemasid, atmosfääri, kunstnikutööd, žanrilist kuuluvust vm (Chandrashekar jt 2017; Riley jt 2018).

Osalussund koos variatsioonide rohkusega võib digitaalse lugemiskogemuse muuta „ei enamaks ega vähemaks kui eraldiseisvaks hoomamatuks kaoseks" (Hunt 2000: 111). Raamatukesksele haridussüsteemile, millel on olukorraga siiani raske toime tulla – pidades silmas üha laienevat lõhet haridussüsteemi ning koolivälise eluviisi ja meediasfääri vahel –, saavad tuge pakkuda digitaalsed haridusplatvormid.

2. Lugemine digitaalsel platvormil

Pihustumise tasakaalustamiseks on sisutasandil holistilise mõõtme esiletoomine vajalik, sest osa kontekstualiseerimine tervikus on igasuguse mõistmise eeltingimus. Kui laias vaates saab tänapäevast õppimist käsitleda liikumisena raamatute juurest platvormide poole (vt Ojamaa, Torop 2020), siis kultuurisemiootika vaatepunktist võib mõlemaid käsitleda tekstiliste nähtustena. Juri Lotman, Boriss Uspenski, Vjatšeslav Ivanov, Vladimir Toporov ja Aleksandr Pjatigorski määratlevad oma „Kultuurisemiootika teesides" teksti kui „tervikliku tähenduse ja tervikliku funktsiooni kandjat" (vt Salupere jt 2013: 109). Järgnevalt avame digitaalsete platvormide keskseid tunnuseid teksti kolme põhitunnuse kaudu. Nendeks on väljendatus, piiritletus ja struktuursus (Lotman 2006 [1970]: 93–96).

Täpsustuseks tuleb esmalt möönda, et digitaalsete õppeplatvormide all peetakse sageli silmas üsna erinevaid nähtusi, mille hulgas on näiteks õpihaldussüsteemid (ingl *learning management systems*, LMS), nagu Moodle, või suuremahulised avatud veebikursused (MOOC). Siinkohal käsitleme spetsiifilist tüüpi platvorme, milliseid oleme ka oma töörühmaga loonud. Määratleme neid veebipõhiste multimodaalsete keskkondadena, mis originaalsisu pakkumise kõrval korrastavad kultuurilised ja hariduslikud allikmaterjalid ühtseks temaatiliseks tervikuks, millele on lisatud juhised ja vahendid materjalide analüüsiks ning loominguliseks eneseväljenduseks erinevates modaalsustes.

2.1. Väljendatus

Tekst vahendab sisu mingisuguses märgisüsteemis ehk keeles. Digitaalsete platvormide puhul oleme rõhutanud nende multi- ja ristmeedialist väljendatust. Ka kirjanduskultuur hõlmab peale kirjalike verbaalsete tekstide näiteks raamatukujundust, turunduslikke paratekste, nagu ülalmainitud raamatutreilerid, kirjandusüritusi,

kõnekeelde tulnud tsitaate, kirjanduslike allikate intersemiootilisi tõlkeid filmideks, teatrilavastusteks jpm. Digitaalsete platvormide väljendusvahendid lubavad sellist pihustatust mõtestada mentaalse tervikuna. Teisisõnu annavad need võimaluse tasakaalustada tekstide transmeedialist pihustumist sihtotstarbelise (nt pedagoogilise eesmärgiga) ristmeedialise tegevusega, valida välja teatud versioonid ja need omavahel suhestada (Torop 2019: 2222). Peale selle võimaldavad digitaalsed vahendid astuda kirjandustekstiga dialoogi läbi erinevate modaalsuste (st mõtestades teksti nt suulise esituse või pildi kaudu). See tagab teosele n-ö ligipääsu erinevat õppimisviisi eelistavatele lugejatele ja arendab ühtlasi loovust.

Olulise panuse erinevate märgisüsteemide kui keelte mõtestamisse hariduse ja pedagoogika vaatepunktist on andnud Gunther Kress oma uue meedia ajastu kirjaoskust puudutavate uurimustega. Kress eristab maailma vahendamisel kaht põhimoodust: sõnastatud (*as told*) ja näidatud (*as shown*) maailma, mis vastab kultuurisemiootilisele sõna ja pildi binaarsusele. Esimene põhineb aja ja järgnevuse loogikal, teine samaaegsuse loogikal. Üks on kohaseim vahendama tegevuste ja sündmuste jadasid, teine suhteid objektide vahel (Kress 2003: 151–156). Digitaalses keskkonnas domineerib eeskätt näidatud maailm. Ühelt poolt on piltide hulk kasvanud kogu kultuuriruumis, teisalt aga mängivad tähendusloomes üha olulisemat rolli kirjaliku teksti visuaalsed tunnused, nagu šrift ja värv, ka küljendus jne. See on teinud vajalikuks disaini märksõna ja toonud kirjaoskuse oluliste komponentidena esile selliste võtete valdamise nagu valik (*selection*), korrastamine (*arrangement*) ning ümberkujundamine (*transformation*) erinevates modaalsustes ja pidevalt muutuvates keskkondades (Kress 2003: 171). Haridus Ekraanil tüüpi platvormid toetavad sedalaadi pädevuse omandamist.

2.2. Piiritletus

Raamatu piiriks on selle kaaned, mis ühtlasi määratlevad teksti alguse ja lõpu ning lineaarse lugemisraja. Platvormi piiri võiks aga kõrvutada pildiraamiga, mille sees paikneva lugemiseks on erinevaid võimalikke ja samavõrra õigeid radu. Ka pildi lugemise protsess ei ole kindlasti konventsioonivaba, ent võimalike valikute ruum on selle puhul siiski avaram. Nii on haridusplatvormidelgi tervikuloome ennem ruumilise kui lineaarse iseloomuga. Tervik on ühtlasi assotsiatiivsem ning selle moodustumise aluseks ei ole üksikosad, mis üldkehtivate süntaktiliste reeglite põhjal jadadeks vormistatakse, vaid protsess toimub sarnaselt mosaiikpildi koostamisega, milles osad sobitatakse omavahel haakuma tervikkujutelmast lähtuvalt, ehk osade asukoha määrab antud tervik. Hajususest hoolimata on digitaalset tervikut võimalik kirjeldada näiteks arhitektooniliselt, eritledes sõnalisi, pildilisi, helilisi jt komponente, aga ka kompositsiooniliselt, seletades osadevahelisi üleminekuid ja kasutusel olevaid sidususe loomise võtteid.

Kultuurisemiootika vaatleb piire „tõlkefiltritena" (Lotman 1999 [1984]: 12), millel on ühtaegu eraldav ja ühendav toime. Ühelt poolt sulgeb piir iseseisva terviku ümbritsevast välja, teisalt võimaldab ümbritsevate süsteemidega dialoogi. Kui raamatute puhul on dialoogid teiste tekstide, metatekstide ja kontekstidega implitsiitsed

ja eeldavad eelteadmist, siis digitaalses ruumis avalduvad dialoogipunktid hüperlinkide kaudu. Olgu näiteks viited täiendavatele allikatele, aga ka kirjeldaval märksõnal põhinevad teemaviited, mille kaudu on üks sotsiaalmeediapostitus selle eri tunnuste (nt kujutatud või käsitletud objekt, üldteema, pildi puhul koloriit jne) põhjal lülitatav väga erinevatesse tervikutesse.

Humanitaarteadusliku platvormi tervikut, mille kese on mõni kunstitekst, ajalooline sündmus või periood, saab käsitleda transmeedialise looilmana (Herman 2009). See on mentaalne mudel, mis põhineb arusaamal, et ka kõige väiksem lugu – seda vahendavast meediumist sõltumata – leiab aset mingisuguses maailmas, millel on ajalised ja ruumilised tunnused ning milles asetsevad tegelased, kes toimivad antud maailma (füüsikalistest, moraalsetest jm) reeglitest lähtudes (vrd Klastrup, Tosca 2004). Lev Manovichi (2012 [2001]) käsitlus andmebaasist kui sümbolilisest vormist, mille loogika vastandub narratiivi omale, on andnud tõuke käsitleda kogu digitaalset keskkonda sellest loogikast lähtuvalt. Siin käsitletavate digitaalsete haridusplatvormide eesmärk on aga just tasakaalustada eraldiseisvate (diskreetsete) andmeühikute kogumeid, pakkudes andmeid sidustavaid kontseptuaalseid tervikuid.

2.3. Struktuursus

Digitaalsete platvormide sisemise hierarhia küsimus on samuti küllalt kompleksne ja dünaamiline võrreldes raamatuga, kus ühtne struktuur luuakse peatükkide ja (ala)pealkirjadega. Esiteks võimaldab platvormide modulaarne ülesehitus kasutaja vajadustest lähtuvalt sisu ümber järjestada, teatud osi vahele jätta, mõne juures hüperlinkide kaudu süvitsi minna jne. Teiseks saab allikteksti kõrvutada selle meta-, para- ja intertekstidega. Need aga mõjutavad üksteise tõlgendamist ja segunevad, mistõttu allikteksti domineeriv positsioon võib mällu jäänud koosluses sootuks hajuda. See kehtib ühtviisi nii kultuuri- kui ka individuaalse mälu puhul. Näiteid selle kohta leidub hulgaliselt: ekraniseeringud, mis on kirjanduslikust allikast oluliselt tuntumad, nagu Coppola „Ristiisa" ja Kromanovi „Viimne reliikvia", või Kristjan Raua illustratsioonid „Kalevipojale", millega on tuttavad ka paljud neist, kes eepost lugenud ei ole. Digitaalsed platvormid oma ristmeedialiste vahenditega toovad selle eriti esile. Seda võib kõrvutada ka Anton Popoviči ja Francis Macri (1977: 124) käsitlusega kirjanduslikest kollektsioonidest, mis dialektilise paigutuse kaudu toovad algtekstis esile tunnused, mis on selle varasemate käsitluste keskmest välja jäänud. Näiteks oleme platvormil „Identiteet ekraanil" käsitlenud Tammsaare teoses koduloomade ning soode ja rabade kujutamist võrdlevas perspektiivis tänapäevaste diskursustega, et tuua kirjandusklassikas esile keskkonnahariduslik ja ökokriitiline mõõde.

Kahtlemata ei ole selline käsitlusviis välistatud ka digitehnoloogiaeelses kultuuris, ent on liikunud perifeersest positsioonist lugemisvõtete tuuma poole. Samuti on selles seoses põhjust meenutada Lotmani (2002 [1978]: 2659) stereoskoopilisuse mõistet, mis tähendab võimalust „saada sama reaalsuse [või teksti – *M. O.*] täiesti erinev projektsioon, tema tõlge hoopis teise keelde". Nii nagu pole võimalik ammendavalt kirjeldada meid ümbritsevat reaalsust, ei ole võimalik „ära mõtestada" kunstilisi

tekste. Iga uus lugemine juhib tähelepanu varem märkamata jäänud aspektidele, seostele, tähendustele. Sellistes mittetäieliku informatsiooni tingimustes kompenseeribki „info vajakajäämise selle stereoskoopilisus" (Lotman 2002 [1978]: 2659). Selline lähtekoht võiks ühtlasi leevendada ärevust lugemiskogemuse killustumise osas, ent ei pisenda seejuures tervikuloome ja sidustamise püüdluste olulisust.

2.4. Kompetentne lugeja

Vahekokkuvõttena tuleks püstitada küsimus kompetentsest lugejast digitaliseerunud kommunikatsioonikeskkonnas (Sukovic 2016: 5; Tso 2019: ix). Võime kirjalikku teksti mõista ja kasutada ei too automaatselt kaasa digitaalse lugemise pädevust (Lim, Jung 2019: 83). Trükiteksti ja digitaalse teksti lugemisel on kognitiivses plaanis küll ilmselgeid ühisosi, kuid digitaalseks lugemiseks vajalike oskuste kooslus tuleneb paljuski just muutunud lugemiskeskkonna ülalkirjeldatud eripäradest. See puudutab kõigepealt informatsioonile ligipääsemise oskust, mille keskmesse on liikunud võime jälgida lugu või informatsiooni lähtuvalt erinevatest modaalsustest ja meediumidest (Scolari jt 2018: 804). Seejuures peaks taolise navigeerimisoskusega kaasas käima võime eri allikaid kriitiliselt hinnata (Lim, Jung 2019: 84) ning nende sisu terviklikuks teadmiseks sünteesida. Viimane peab lähtuma ka antud situatsioonist ehk teadvustamisest, millist tüüpi informatsiooni või millist konkreetse loo või temaatilise terviku osa järgmiseks vajatakse (vt Salmerón jt 2018: 93). Digitaalsele keskkonnale omaste rohkete tähelepanu hajutajate kontekstis on see keskse tähtsusega ning digitaalsed õppeplatvormid saavad selles olulist tuge pakkuda.

Digitaalse lugeja tõlgendamisoskus peaks üha selgemalt tuginema vahendatuse ja meediumi eripärade, aga ka multimeedialiste, tekstidevaheliste, intermeedialiste ja interdiskursiivsete aspektide teadvustamisele. Samuti on, eriti ülalkirjeldatud osaluskultuuri kontekstis, oluline oskus „kujutleda, kavandada, luua, toimetada ja/või taaskasutada meediasisu erinevatel meediaplatvormidel ja keeltes (sh verbaalsed, helilised, helilis-pildilised, koodipõhised...)" (Scolari jt 2018: 805). Tekstiosade uuskasutus ei ole tõhus meetod üksnes loovuse arendamiseks, vaid ka kompositsiooniprintsiipide tundmaõppimiseks ja ühtlasi võib seda pidada õppijakeskse õpetamise peamiseks tunnuseks (Scolari jt 2018: 809). Kuivõrd pädevusi, mida digitaliseerumine kultuuriliikmetelt eeldab, on palju ja erisuguseid, siis on osutunud keeruliseks nende sünteesimine terviklikuks kirjaoskuse käsitluseks. Ühe lahenduse on välja pakkunud Suzana Sukovic oma ühildava kirjaoskuse (ingl *transliteracy*) mõistega, mille põhitunnuseks on „liikumise voolavus kogu välja ulatuses – erinevate kontekstide, modaalsuste, tehnoloogiate ja žanride vahel" (2016: 207). Seejuures tulebki rõhutada eri meediumide vahel liikumisest tekkinud tähenduste kaardistamise võimet, mis erineb üksikute meediumide lugemisoskuste summast (Ipri 2010: 532). Ehkki ühildava kirjaoskuse kontseptsioon on pigem üldine raamistik kui konkreetne programm, on see kindlasti samm lähemale digitaalse lugemise jaoks vajalike kompetentside sidusa arendamisega seotud metodoloogiliste ja pedagoogiliste väljakutsete lahendamisele.

3. Modelleerimine humanitaarhariduses

See, kuidas õpilased jõuavad eri tekste ja nende tähendusi uurides mingist teemast tervikliku arusaama kujundamiseni, on digiõppe põhiküsimus. Ühelt poolt on võimalus kõrvuti käsitleda suurel hulgal eriliigilisi informatsiooniallikaid, mis peegeldavad erinevaid vaatepunkte, kahtlemata digiõppe suureks eeliseks õpikukeskses klassiruumis õppimise ees. Teiselt poolt on selline tekstide paljusus üks keerulisemaid väljakutseid ja võib õpiprotsessi hoopis pärssida (vt Hartman jt 2018; Rau 2020). Seetõttu on äärmiselt oluline, et mitmeid allikaid ja vaatepunkte koondavad digitaalsed õppematerjalid toetaksid õpilaste võimet luua erinevate tekstide vahelisi seoseid. Materjali esitusviis, selgitavad kommentaarid ning ülesannete kirjeldused suunavad õpilasi nende ees avanevat keerulist infokooslust korrastama ning tervikpilti tajuma.

See, kuidas õppematerjal suunab õpilasi tervikpilti looma, võib põhineda informatsiooni süstematiseerimise eri viisidel. Lähtudes Juri Lotmani semiootikateooriast, on võimalik neid infokorrastusviise kirjeldada kui erinevat tüüpi modelleerivaid süsteeme. See omakorda lubab välja pakkuda raamistiku, mille abil analüüsida õppematerjalides esinevaid sidususmehhanisme. Pretendeerimata kõikehõlmavusele, keskendume siinkohal kolmele modelleerimistüübile: mütoloogilisele, teaduslikule ja mängulisele modelleerimisele.[1] See tüpoloogia võimaldab analüüsida transmeedialiste tekstide sidustamise viise õppetöös ning mõtestada seeläbi erinevaid digitaalse lugemise praktikaid. Järgnevalt esitame lühidalt nende modelleerimistüüpide kirjeldused ning toome näiteid sellest, kuidas need toimivad Haridus Ekraanil õppematerjalides.

Mütoloogiline modelleerimine on omane sellist tüüpi infokorrastusviisile, mille peamine tähendust loov mehhanism on samastamine kui mütoloogilise maailmataju keskne tunnus (vt Lotman, Uspenski 1999 [1973]; Lotman, Mints 2016 [1981]). Kuigi Lotmani sõnul on mütoloogiline mõtlemine omane just lastele, ei kao see ka täiskasvanueas, vaid inimese teadvuse struktuuri olulise osana jätkab funktsioneerimist assotsiatsioonide tekitajana ja toimib ühena aktiivsetest modelleerivatest mehhanismidest (Lotman 2002 [1978]: 2649; vt ka Lepik 2007: 36). Tekstides väljendub mütoloogilisus ennekõike objekttasandi ja metatasandi suhestatuses. Mütoloogilised kirjeldused on metatekstid, kus kirjeldatav objekt ja kirjeldav metatekst kuuluvad ühte ja samasse keelde (Lotman, Uspenski 1999 [1973]: 187–188). Lotman ja Uspenski vastandavad selliseid tekste mittemütoloogilistele ehk deskriptiivsetele kirjeldustele, mis toimivad metakeelena ehk objekttasandist erineva süsteemina, ja järeldavad, et kui deskriptiivseid kirjeldusi mõistetakse ühest keelest teise tõlkimise kaudu, siis mütoloogiliste kirjelduste vahendusel saavutatakse mõistmine ära-

[1] Kuigi Lotman ei kasuta mõistet *mütoloogiline modelleerimine*, on ta oma tekstides kirjeldanud müütilisel mõtlemisel põhinevate semiootiliste mudelite eripärasid (vt Lotman 2000 [1990]; Lotman, Mints 2016 [1981]; Lotman, Uspenski 1999 [1973], 2013 [1971]; vt ka Lepik 2007), mis käesolevas artiklis on koondatud ühtse nimetaja alla. Samuti on oluline välja tuua, et artiklis „Kunst modelleerivate süsteemide reas" (Lotman 2016 [1967]) kirjeldab Lotman peale teadusliku ja mängulise modelleerimise ka kunstilist modelleerimist, mis struktuurilt vastab mängulisele modelleerimisele, kuna mõlemat tüüpi mudelite domineerivaks tunnuseks on mitmetähenduslikkuse võimaldamine (Lepik 2007: 176).

tundmise või identifitseerimise abil (Lotman, Uspenski 1999 [1973]: 187–188). Nii võime öelda, et mütoloogilise modelleerimisega korrastatakse inforuumi samasusprintsiibi alusel, vaadates erinevaid tekste ühe ja sama loo variatsioonidena. Õppematerjalide ülesehituse kontekstis tähendab see, et kesksele tekstile omistatakse metateksti funktsioon ja selle tähendust esitatakse muutumatuna kõigi teiste tekstide suhtes. Sellise modelleerimise näide Haridus Ekraanil õppematerjalides on platvormi „Ajalugu ekraanil" interaktiivne kaart „Seltsimees lapse maailm". Selle kaardi hariduslik eesmärk on tutvustada Stalini-aegse maailma nähtusi, nagu propaganda, isikukultus, hirmuõhkkond, pioneerid jm. Kaardil olevad punktid koondavad väga erinevat materjali: ekspertide videoloenguid, katkendeid kunstitekstidest, mälestustest ja teatmeteostest, õpimänge, teste ja selgitusi. Nagu kaardi nimi ka vihjab, on kesksel kohal „Seltsimees laps" – nii Leelo Tungla romaani kui ka Moonika Siimetsa samanimelise ekraniseeringuna. Iga teemapunkt algab kas tsitaadiga romaanist või videoklipiga filmist, milles on kujutatud mõnda eespool mainitud kultuurinähtust. Need katked ei selgita otseselt õpilastele, mis on ühiskondlik hirmuõhkkond või isikukultus, vaid kirjeldavad seda esitatud loo kaudu (Stalini aja hirmuõhkkonna väljendus on väike hirmunud Leelo oma pimedas magamistoas). Olles asetatud metateksti rolli, on „Seltsimees laps" samastatud nii kirjeldatava nähtuse enda kui ka kõigi teiste selle õppematerjali osa tekstide tähendusega, mis omal moel kordavad sedasama lugu Stalini-aegsest maailmast (nt hirm, mida tunneb Leelo, on sama hirm, millest kirjutab Solženitsõn).

Teaduslikku modelleerimist võib samastada ka eespool nimetatud deskriptiivsete kirjeldustega ehk metakeeleliste kirjeldustega (Lotman, Uspenski 1999 [1973]: 187–188). Peamine teaduslike mudelite tunnus on Lotmani jaoks see, et „teaduslik mudel taasloob näitlikus vormis objekti süsteemi. Ta modelleerib uuritava süsteemi „keele"." (Lotman 2016 [1967]: 20) See tähendab, et sellist tüüpi mudelid on vaadeldavad tekstide loomise reegleid kirjeldava süsteemina (Lotman, Uspenski 2013 [1971]: 221) – nende n-ö grammatikana. Teadusliku modelleerimise puhul on seega infovälja terviklikkus saavutatav selles asuvate tekstide kirjeldamise kaudu ühtse reeglitesüsteemi kontekstis. Teadusliku modelleerimise näitena võib tuua platvormi Kirjandus Ekraanil õppeteema „Kirjandus ja film kultuuris". Selles on vaatluse all Andrus Kivirähki romaani „Rehepapp" ja Rainer Sarneti filmi „November" seosed eesti rahvaluule tekstide ja Akira Kurosawa filmiga „Unenäod". Seoste mõistmiseks tutvustatakse selles õppematerjalis kultuurisemiootika mõisteid *prototekst*, *metatekst* ja *intertekst*. Sellise metakeele omandamine võimaldab õpilastel luua süsteemse pildi kultuuriruumis paiknevate tekstide kooslustest ning mõista nendevahelisi tähendusülekandeid. Samuti põhineb teaduslikul mudelil sama platvormi teine teema „Kirjandus filmis", mis avab ühe loo transformatsioone erinevates kunstikeeltes. Tutvudes kunstiteksti struktuuri kirjelduskeelega (algus ja lõpp, faabula ja süžee, jutustuse aegruum jm), omandavad õpilased oskuse vaadelda ja analüüsida erinevaid kunstitekste ühtsete parameetrite alusel.

Mänguliste mudelite eripära seisneb Lotmani jaoks selles, et nende elementidele on omane kahe- või mitmekordne tähenduslikkus (Lotman 2016 [1967]: 30–31). Ta selgitab, et mäng eeldab alati kahe erineva käitumise üheaegset realisatsiooni:

mängija peab meeles pidama, et ta osaleb tinglikus situatsioonis, ja samal ajal peab ta suutma selle ka unustada. See omadus defineerib mänguliste mudelite tähendusloomet: „Mänguefekti mehhanism ei seisne eri tähenduste jäigas üheaegses kooseksisteerimises, vaid selle pidevas teadvustamises, et peale praegu tajutavate tähenduste on võimalikud ka teised tähendused" (Lotman 2016 [1967]: 29).

Peale mitmetähenduslikkuse on mängulisusele omane ka seaduspäraste ja juhuslike protsesside tasakaal (Lotman 2016 [1967]: 20). See tähendab, et mängul on korduvad reeglid, mis jätavad samal ajal piisavalt vabadust ja ruumi kõrvalekalleteks ja uute tähenduste tekkimiseks. Seega mängulistes mudelites on ühtsust loovaks elemendiks samuti reeglid, kuid erinevalt teaduslikest mudelitest ei ole need mõeldud kirjeldama kogu süsteemi tervikuna, vaid suunavad pigem „mängus osaleja" käitumist, võimaldades seejuures luua vabalt uusi tõlgendusi ja tähendusi. Mängulise modelleerimise näitena võib tuua Eesti kinoklassika platvormi õppematerjali „Loo oma „Kevade"". Selles antakse õpilastele võimalus koostada „Kevade" tekstist oma versioon. Valida saab erinevate tänapäeva ümberjutustamisviiside vahel, nagu meem, reaktsioonivideo, remiks, infograafik jm. Materjalis tutvustatakse lühidalt igale formaadile omaseid tekstiloome reegleid, ent samal ajal jääb õpilastele vabadus otsustada, millist lugu nad selle klassikalise teksti vahendusel jutustada tahavad. Selle õppeprotsessi eesmärk on ilmestada, kuidas iga ümberjutustus, luues küll uusi tähendusi, on alati dialoogis eelnenud tekstiversioonidega, ning arendada õpilaste võimet samaaegselt kultuuriruumis eri tähenduskihte tajuda. Kuna mängulisele modelleerimisele on omane see, „et iga mõtestamine moodustab omaette sünkroonse lõike, kuid säilitab seejuures mälestuse eelnevatest tähendustest ja teadmise võimalikest tulevastest tähendustest" (Lotman 2016 [1967]: 29–30), aitab selline lähenemine õpilastel tajuda teksti tähenduste pidevat muutumist kui loomulikku protsessi kultuuris.

Kokkuvõte

Kirjandusteosed ja lugemine on digitaalse keskkonna domineerimise tingimustes läbi teinud muutused, mis on otseselt mõjutanud kirjanduse õpetamist ja kooliharidust laiemalt. Artiklis vaatlesime probleemi tekstide ja tähenduste pihustumise vaatepunktist. Esmalt kirjeldasime teksti- ja lugemiskogemuste mitmekesistumist multimeedialisuse, variatiivsuse ja killustatuse näitel. Ühe teksti erinevad modaalsused osalevad üha aktiivsemalt teksti tähendusloomes ning ühe teksti variatsioonid eri meediumides ja diskursustes on üha levinumad. See on teinud komplekssemaks tekstide piiritlemise ehk terviku küsimuse. Digitaalsed hariduslikud platvormid on aga oluline vahend hoomamatuks paisunud variatiivsuses ja üldises infokülluses tervikupotentsiaali esiletoomiseks. Kultuurisemiootiline tekstikäsitlus pakub selle mõtestamiseks aluse, seda digitaalsete platvormide väljendatuse, dialoogilise piiritletuse ja struktureerituse eripärade avamise kaudu. Siia alla kuuluvad arusaamad, kuidas erinevate meediumide väljendusvahendid tervikliku tähenduse vahendamisel üksteist toetavad või takistavad; et digitaalsetel platvormidel tekkival tervikul on ennem ruumiline kui lineaarne iseloom (arusaam maailmalisusest); ning et digitaal-

sete võimalustega saab asetada teksti dialoogi teiste tekstide ja diskursustega, mis võivad tervikuloome keskmesse tuua teksti varem ebaolulistena tajutud tähenduskihid. Digitaalse keskkonna vormiliste tunnuste ja tehniliste võimaluste kõrval on oluline mõtestada aga ka õppesisu tasandil terviku loomise võimalusi. Siinkohal lähenesime sellele semiootilise modelleerimise tüpoloogia kaudu, näitlikustades oma uurimisrühma loodud materjalide varal erinevaid vahendeid ja nende kombinatsioone mütoloogilise, teadusliku ja mängulise modelleerimise toetamiseks. Haridus Ekraanil platvormide arendamise põhimõtted toetuvad kultuurisemiootika ja transmeedialiste uuringute arusaamadele, mille keskmes on kultuuri sidususe ja mitmekesisuse vastastikune sõltuvus. Seda näitlikustabki samade tekstide eri meediaversioonide koosolu kultuuris. Just kultuuripärandisse kuuluvaid tekste ümbritseb tänapäeval mahukas vahendav ja tõlgendav tekstikooslus, mis aga õpikutes üldjuhul ei kajastu. See tekstikooslus loob transmeedialise looilma, mis peegeldab läbi klassika tõlgendamise mitmekesisuse kultuuri enesemõistmist, tõstab klassika nähtavust kultuuri erinevatel tasanditel ja õpetab tundma mitmesuguste meediumide (teatri, kino, muusika jne) võimet klassikat vahendada, arendades seeläbi õpilaste oskust kultuuripärandit väärtustada.

Uurimistööd on finantseerinud Eesti Teadusagentuur (PSG675 „Transmeedialisuse kultuuriväärtus ja digitaalsete hariduslike platvormide arendamine (1.01.2021–31.12.2024)").

VEEBIVARAD

https://haridusekraanil.ee

KIRJANDUS

Berry, David M. 2012. Introduction: Understanding the digital humanities. – Understanding Digital Humanities. Toim D. M. Berry. Basingstoke–New York: Palgrave Macmillan, lk 1–20. https://doi.org/10.1057/9780230371934_1

Brillenburg Wurth, Kiene; Espi, Sara Rosa; van de Ven, Inge 2013. Visual text and media divergence: Analogue literary writing in a digital age. – European Journal of English Studies, kd 17, nr 1, lk 92–108. https://doi.org/10.1080/13825577.2013.757014

Bruns, Axel 2009. From prosumer to produser: Understanding user-led content creation. Ettekanne konverentsil „Transforming Audience" Londonis 3.–4. septembril 2009.

Burgess, Jean; Green, Joshua 2018. YouTube: Online Video and Participatory Culture. Cambridge: Polity Press.

Chandrashekar, Ashok; Amat, Fernando; Basilico, Justin; Jebara, Tony 2017. Artwork personalization at Netflix. – Netflix Technology Blog 7. XII. https://netflixtechblog.com/artwork-personalization-c589f074ad76 (26. VI 2021).

Fadeev, Aleksandr 2020. Acquisition of artistic literacy in multimodal learning via intersemiotic translation. – Punctum. International Journal of Semiotics, kd 6, nr 1, lk 133–159. https://doi.org/10.18680/hss.2020.0007

Hartman, Douglas K.; Hagerman, Michelle Schira; Leu, Donald J. 2018. Toward a new literacies perspective of synthesis: Multiple source meaning construction. – Handbook of Multiple Source Use. Toim Jason L. G. Braasch, Ivar Bråten, Matthew T. McCrudden. New York: Routledge, Taylor & Francis Group, lk 55–78. https://doi.org/10.4324/9781315627496-4

Herman, David 2009. Narrative ways of worldmaking. – Narratology in the Age of Cross-Disciplinary Narrative Research. Toim Sandra Heinen, Roy Sommer. Berlin: Walter de Gruyter, lk 71–87.

Hunt, Peter 2000. Futures for children's literature: Evolution or radical break. – Cambridge Journal of Education, kd 30, nr 1, lk 111–119. https://doi.org/10.1080/03057640050005807

Ipri, Tom 2010. Introducing transliteracy: What does it mean to academic libraries? – College & Research Library News, kd 71, nr 10, lk 532–533, 567. https://doi.org/10.5860/crln.71.10.8455

Jenkins, Henry 2006. Convergence Culture: Where Old and New Media Collide. New York–London: New York University Press.

Jenkins, Henry 2009. Confronting the Challenges of Participatory Culture: Media Education for the 21st Century. Cambridge (MA)–London: The MIT Press. https://doi.org/10.7551/mitpress/8435.001.0001

Jenkins, Henry 2010. Transmedia education: The seven principles revisited. – Confessions of an Aca-Fan 21. VI. http://henryjenkins.org/2010/06/transmedia_education_the_7_pri.html (26. VI 2021).

Kembellec, Gérald; Broudoux, Evelyne (toim) 2017. Reading and Writing Knowledge in Scientific Communities: Digital Humanities and Knowledge Construction. London: ISTE, Hoboken: Wiley. https://doi.org/10.1002/9781119384410

Klastrup, Lisbeth; Tosca, Susana 2004. Transmedial worlds: Rethinking cyberworld design. Proceedings of the International Conference on Cyberworlds 2004, IEEE Computer Society. Los Alamitos, CA. https://www.itu.dk/people/klastrup/klastruptosca_transworlds.pdf (13. VII 2021).

Kress, Gunther 2003. Literacy in the New Media Age. London: Routledge. https://doi.org/10.4324/9780203299234

Kress, Gunther 2014. Reading, learning, and „texts" in their interaction with the digital media. – Tidsskriftet Viden om Literacy, kd 16, lk 56–67.

Lepik, Peet 2007. Universaalidest Juri Lotmani semiootika kontekstis. (Tartu Semiootika Raamatukogu 6.) Tartu: Tartu University Press.

Liestøl, Gunnar 2007. The dynamics of convergence and divergence in digital domains. – The Ambivalence of Convergence. Toim Tanja Storsul, Dagny Stuedahl. Gothenburg: Nordicom, University of Gothenburg, lk 165–178.

Lim, Hyo Jin; Jung, Hyekyung 2019. Factors related to digital reading achievement: A multilevel analysis using international large scale data. – Computers & Education, kd 133, lk 82–93. https://doi.org/10.1016/j.compedu.2019.01.007

Lotman, Juri 1999 [1984]. Semiosfäärist. – J. Lotman, Semiosfäärist. (Avatud Eesti raamat.) Tlk Kajar Pruul. Tallinn: Vagabund, lk 11–25.

Lotman, Juri 2000 [1990]. Universe of the Mind: A Semiotic Theory of Culture. Bloomington: Indiana University Press.

Lotman, Juri 2002 [1978]. Kultuuri fenomen. Tlk Silvi Salupere. – Akadeemia, nr 12, lk 2644–2660.

Lotman, Juri 2006 [1970]. Kunstilise teksti struktuur. (Avatud Eesti raamat.) Tlk Pärt Lias. Tallinn: Tänapäev.

Lotman, Juri 2016 [1967]. Kunst modelleerivate süsteemide reas. – J. Lotman, Kultuurisemiootika. Tekst – kirjandus – kultuur. Tlk Pärt Lias, Inta Soms, Rein Veidemann. Tallinn: Tänapäev, lk 8–31.

Lotman, Juri 2016 [1981]. Kultuurisemiootika ja teksti mõiste. – J. Lotman, Kultuurisemiootika. Tekst – kirjandus – kultuur. Tlk Pärt Lias, Inta Soms, Rein Veidemann. Tallinn: Tänapäev, lk 234–240.

Lotman, Juri; Mints, Zara 2016 [1981]. Kirjandus ja mütoloogia. – J. Lotman, Kultuurisemiootika. Tekst – kirjandus – kultuur. Tlk Pärt Lias, Inta Soms, Rein Veidemann. Tallinn: Tänapäev, lk 273–298.

Lotman, Juri; Uspenski, Boriss 1999 [1973]. Müüt – nimi – kultuur. – J. Lotman, Semiosfäärist. (Avatud Eesti raamat.) Tlk Kajar Pruul. Tallinn: Vagabund, lk 187–218.

Lotman, Juri; Uspenski, Boriss 2013 [1971]. Kultuuri semiootilisest mehhanismist. – B. Uspenski, Vene kultuuri jõujooni. (Avatud Eesti raamat.) Koost Peet Lepik, B. Uspenski; tlk P. Lepik. Tartu: Ilmamaa, lk 207–239.

Mangen, Anne; van der Weel, Adriaan 2017. Why don't we read hypertext novels? – Convergence. The International Journal of Research into New Media Technologies, kd 23, nr 2, lk 166–181. https://doi.org/10.1177/1354856515586042

Manovich, Lev 2003. New media from Borges to HTML. – The New Media Reader. Toim Noah Wardrip-Fruin, Nick Montfort. Cambridge–London: The MIT Press, lk 13–25.

Manovich, Lev 2012 [2001]. Uue meedia keel. Tlk Ellu Maar. Tallinn: Eesti Kunstiakadeemia.

Milyakina, Alexandra; Ojamaa, Maarja; Pilipovec, Tatjana; Rickberg, Merit 2020. Literature on screen: Teaching adaptations with a multimodal course. – Media Practice and Education, kd 21, nr 3, lk 229–239. https://doi.org/10.1080/25741136.2019.1673029

Murray, Simone 2018. The Digital Literary Sphere: Reading, Writing, and Selling Books in the Internet Era. Baltimore: Johns Hopkins University Press.

Ojamaa, Maarja; Torop, Peeter; Fadeev, Aleksandr; Milyakina, Alexandra; Pilipovec, Tatjana; Rickberg, Merit 2019. Culture as education: From transmediality to transdisciplinary pedagogy. – Sign Systems Studies, kd 47, lk 152–176. https://doi.org/10.12697/SSS.2019.47.1-2.06

Ojamaa, Maarja; Torop, Peeter 2020. Kultuuripärand ja digitaalne lugemine: raamatu ja platvormi vahel. – Methis. Studia humaniora Estonica, nr 26, lk 51–68. https://doi.org/10.7592/methis.v21i26.16910

Pennington, Martha C.; Waxler, Robert P. 2018. Why Reading Books Still Matters. The Power of Literature in Digital Times. New York–London: Routledge. https://doi.org/10.4324/9781315210247

Popovič, Anton; Macri, Francis M. 1977. Literary synthesis. – Canadian Review of Comparative Literature, kd 4, nr 2, lk 118–132.

Rau, Martina A. 2020. Cognitive and socio-cultural theories on competencies and practices involved in learning with multiple external representations. – Handbook of Learning from Multiple Representations and Perspectives. Toim Peggy Van Meter, Alexandra

List, Doug Lombardi, Panayiota Kendeou. New York: Routledge, lk 17–32. https://doi.org/10.4324/9780429443961-3

Rickberg, Merit 2017. Kahest vastandlikust tendentsist ajalooharidu ses. – Acta Semiotica Estica, kd 14, lk 46–68.

Riley, Madeline; Machado, Lauren; Roussabrov, Boris; Branyen, Tim; Bhawalkar, Parth; Jin, Eugene; Kansara, Apurva 2018. AVA: The Art and Science of Image Discovery at Netflix. – Netflix Technology Blog 7. II. https://medium.com/netflix-techblog/ava-the-art-and-science- of-image-discovery-at-netflix-a442f163af6 (13. VII 2021).

Ritzer, George; Jurgenson, Nathan 2010. Production, consumption, prosumption: The nature of capitalism in the age of the digital „prosumer". – Journal of Consumer Culture, kd 10, nr 1, lk 13–36. https://doi.org/10.1177/1469540509354673

Salmerón, Ladislao; Strømsø, Helge I.; Kammerer, Yvonne; Stadtler, Marc; van den Broek, Paul 2018. Comprehension processes in digital reading. – Learning to Read in a Digital World. Toim Mirit Barzillai, Jenny Thomson, Sascha Schroeder, P. van den Broek. Amsterdam: John Benjamins, lk 91–120. https://doi.org/10.1075/swll.17.04sal

Salupere, Silvi; Torop, Peeter; Kull, Kalevi 2013. Beginnings of the Semiotics of Culture. Tartu: University of Tartu Press.

Schnapp, Jeffrey; Presner, Todd 2009. Digital Humanities Manifesto 2.0. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (27. VI 2021).

Scolari, Carlos A. 2013. Lostology: Transmedia storytelling and expansion/compression strategies. – Semiotica, kd 195, lk 45–68. https://doi.org/10.1515/sem-2013-0038

Scolari, Carlos A.; Masanet, Maria-José; Guerrero-Pico, Mar; Establés, María-José 2018. Transmedia literacy in the new media ecology: Teens' transmedia skills and informal learning strategies. – El Profesional de la Información, kd 27, nr 4, lk 801–812. https://doi.org/10.3145/epi.2018.jul.09

Shibata, Hirohito; Omura, Kengo 2020. Why Digital Displays Cannot Replace Paper. The Cognitive Science of Media for Reading and Writing. Singapore: Springer. https://doi.org/10.1007/978-981-15-9476-2

Spink, Bec 2014. Social & community driven learning. – CSU ThinkSpace 3. VIII.

Sukovic, Suzana 2016. Transliteracy in Complex Information Environments. Amsterdam jt: Chandos Publishing. https://doi.org/10.1016/B978-0-08-100875-1.00004-2

Svensson, Patrik 2012. The digital humanities as a humanities project. – Arts and Humanities in Higher Education, kd 11, nr 1–2, lk 42–60. https://doi.org/10.1177/1474022211427367

Toolan, Michael 2010. Foreword: Lost and found in the reading machine. – Literary Education and Digital Learning: Methods and Technologies for Humanities Studies. Toim Willie van Peer, Sonia Zyngier, Vander Viana. New York: Information Science Reference, lk x–xxiv.

Torop, Peeter 2019. Kutuuridünaamika ja kultuurisemiootika väljavaated. – Akadeemia, nr 12, lk 2217–2231.

Tso, Anna Wing-bo (toim) 2019. Digital Humanities and New Ways of Teaching. (Digital Culture and Humanities 1.) Singapore: Springer. https://doi.org/10.1007/978-981-13-1277-9

Maarja Ojamaa (snd 1983), PhD, Tartu Ülikooli semiootika osakonna teadur (Jakobi 2, Tartu 51003), maarja.ojamaa@ut.ee

Peeter Torop (snd 1950), PhD, Tartu Ülikooli semiootika osakonna kultuurisemiootika professor (Jakobi 2, Tartu 51003), peeter.torop@ut.ee

Aleksandr Fadeev (snd 1989), MA, Tartu Ülikooli semiootika osakonna doktorant (Jakobi 2, Tartu 51003), aleksandr.fadeev@ut.ee

Alexandra Milyakina (snd 1991), PhD, Tartu Ülikooli semiootika osakonna nooremteadur (Jakobi 2, Tartu 51003), alexandra.milyakina@ut.ee

Merit Rickberg (snd 1991), MA, Tartu Ülikooli semiootika osakonna doktorant (Jakobi 2, Tartu 51003); Tallinna Ülikooli Juri Lotmani semiootikavaramu juhataja, rickberg.merit@gmail.com

Digital reading and humanities education

Keywords: semiotics of culture, reading comprehension, digital culture

The paper focuses on the problems of reading and teaching literature in the context of digitalization and offers solutions for facilitating a meaningful dialogue with cultural heritage. Both divergence and convergence of media pose a challenge to the conventional ways of teaching literature based on source texts, as both processes entail diversification of media channels and dispersion of content across them. Reading in the digital age is characterized by an increasing multimodality, variability and fragmentariness of texts. The best practices of digital learning platforms allow synthesizing heterogeneous versions of the original into a coherent whole, which results in a new type of text characterized by expression in a (multimodal) language, boundedness and structuredness. The systematization of varied resources on a digital platform is supported by mythological, scientific and play-type modelling.

The project *Education on Screen* developed by the Transmedia research group at the University of Tartu provides a methodological and technological framework for teaching literature under the circumstances of digitalization. The project offers a selection of open digital learning platforms aimed at secondary school students and a general audience. Each platform explores a key text of Estonian culture and their film adaptations against the background of universal cultural processes: Literature on Screen based on *Old Barney or November* by Andrus Kivirähk, History on Screen based on Leelo Tungal's trilogy *The Little Comrade*, Identity on Screen based on A. H. Tammsaare's *Truth and Justice*, and Estonian Film Classics based on Oskar Luts' novel *Spring* and other texts.

The pedagogical approach is informed by the semiotics of culture and transmedia research and implies that each artistic text exists in a series of possible versions in different discourses and media, which are rarely taken into account by school education. The digital learning platform supports the dialogue with cultural heritage

through a variety of versions, while taking into account the vernacular skills and literacies acquired by students in the digital environment.

Maarja Ojamaa (b. 1983), PhD, University of Tartu, Department of Semiotics, Research Fellow (Jakobi 2, Tartu 51003), maarja.ojamaa@ut.ee

Peeter Torop (b. 1950), PhD, University of Tartu, Department of Semiotics, Professor of Cultural Semiotics (Jakobi 2, Tartu 51003), peeter.torop@ut.ee

Aleksandr Fadeev (b. 1989), MA, University of Tartu, Department of Semiotics, Doctoral student (Jakobi 2, Tartu 51003), aleksandr.fadeev@ut.ee

Alexandra Milyakina (b. 1991), PhD, University of Tartu, Department of Semiotics, Junior Researcher (Jakobi 2, Tartu 51003), alexandra.milyakina@ut.ee

Merit Rickberg (b. 1991), MA, University of Tartu, Department of Semiotics, Doctoral student (Jakobi 2, Tartu 51003); Head of Juri Lotman Semiotics Repository at Tallinn University, rickberg.merit@gmail.com

https://doi.org/10.54013/kk764a6

Sõnastikukogust keeleportaaliks

MARGIT LANGEMETS, KRISTINA KOPPEL, JELENA KALLAS,
ARVI TAVAST

XXI sajandil on traditsiooniline sõnaraamatutöö oluliselt muutunud, tuues kaasa nii hõlbustusi kui ka katsumusi sõnaraamatu koostajale ja kasutajalegi. Siin artiklis keskendume digipöördele, mis sai Eesti Keele Instituudis (EKI) alguse 2017. aastal uue sõnastikusüsteemi Ekilex loomisega. Ent seda ei saa sugugi pidada sõnaraamatutöö ainsaks digipöördeks. Esimeseks digipöördeks võiks paradoksaalsel moel lugeda õnnelikku juhust, et 1960-ndatel alustatud „Eesti kirjakeele seletussõnaraamat" (EKSS) ei jõudnud ilmuma hakata varem kui 1980-ndate lõpul – ajal, mil instituudis olid arvutid juba kasutusel. Nõnda sattusime kohe n-ö arenenud leksikograafiamaade hulka, kel juba 1990/2000-ndatel oli elektroonilisel kujul olemas oma keele tähtsaim ja mahukaim sõnakogu (enam-vähem struktureeritud tekstina). Lihtsama ülesehitusega 1976. aasta õigekeelsussõnaraamat oli samuti arvutis olemas. Digipöördeks võib pidada ka oma sõnastikusüsteemi EELex (Langemets jt 2010; Jürviste jt 2011) ehitamist alates 1990-ndate lõpust. Kasutuskõlblikuks sai EELex 2003. aastal esimesena eesti-vene sõnaraamatu (EVS) jaoks, misjärel süsteemi laiendati ja arendati paljude sõnastike tarvis kuni 2015. aastani. Oma hiilgeaegadel oli EELexis ligi 70 valminud või töös olevat sõnastikubaasi. Ühtlasi oli tolleks ajaks saanud selgeks, et samal viisil jätkamine ei ole jätkusuutlik: internet oli plahvatuslikult levinud, tarkvara ja andmeformaadid olid muutunud, leksikograafia kogu Euroopas ja maailmas oli tugevasti oma suunda muutmas.

Sarnaste probleemidega seisid silmitsi ka teiste Euroopa riikide leksikograafid: suurem osa leksikaalsest teadmusest oli talletatud vaid paberile, sõnastikud olid suures osas struktureerimata, nad ei olnud masinloetavad ega sobinud seetõttu ei loomuliku keele töötluse arendamiseks ega muude keeleandmetel põhinevate rakenduste jaoks. Nagu Eestis, nii ka mujal Euroopas alustati hoogsalt järgmise põlvkonna sõnastikusüsteemide ehitamist, millega kaasnes laiaulatuslik pabersõnastike digiteerimine. Algul kasutas enamik sõnastikubaase XML-vormingut, kuid varsti sai selgeks, et hädavajalik on välja töötada andmeformaat, mis võimaldaks keeleinfot kirjeldada ja andmeid vahetada (nii eri keelte kui ka andmebaaside vahel). Muutunud olud tõukasid taas tagant uut digipööret, mille tunnussõna on andmete ühendamine ja linkimine, mis sõnaraamatumaailmas tähendab leksikaalse info koondamist.

Värskelt ilmunud elektroonilise leksikograafia valge raamatu „The Future of Academic Lexicography" (Steurs jt 2020) soovitus riiklikele leksikograafiaasutustele on mitte jätkata üksikute sõnastike koostamist, vaid ehitada üleriigiline taristu – ühtne andmebaas, mis koondab seni koostatud ressursse ning mida saab linkida teistega nii oma riigi sees kui ka rahvusvaheliselt. 35 Euroopa riiki ühendav Horisont 2020 projekti „Euroopa leksikograafiataristu" (ELEXIS) eesmärk on ühendada linkandmete

formaadis eri keelte sõnastikud ja keeleandmed ning pakkuda kõigile sõnaraamatutööga seotud asutustele uusi tööriistu ja teenuseid.[1] Sellele on pühendunud mitmed Euroopa koostööprojektid, kus osaleb ka Eesti Keele Instituut. Seni on loodud pilvepõhine PDF- ja XML-vormingute konverteerimisteenus Elexifier, RDF-põhine linkimistööriist NAISC ja korpuspäringusüsteemiga Sketch Engine (vt ka Kilgarriff jt 2004; Kilgarriff jt 2014) ühendatud sõnastikusüsteem Lexonomy (vt ka Měchura 2017).

Järgnevas puudutame samu võtmeküsimusi, mida eespool mainisime: liikumist ühtse andmebaasi poole, andmete linkimist ja üldist sõnaraamatutöö muutumist. EKI-s loodud sõnastikusüsteemi Ekilex ja keeleportaali Sõnaveeb oleme tutvustanud Euroopa konverentsidel (Tavast jt 2018; Koppel jt 2019a; Tavast jt 2020), siin kirjeldame ka uusimat EKI sõnakogu – EKI ühendsõnastikku – ja selle koostamisel kasutatud Ekilexi andmemudelit. Sõnaraamatutöö on pöördeliselt muutunud mitmes mõttes: tehniliselt (ühtne sõnastikusüsteem ja andmebaas, arenenud korpuspäringusüsteemid), leksikograafiliselt (sõnastike ühendamine ja koostamispõhimõtete ühtlustamine, töö infokihtidega, andmete dünaamilisus) ja töökorralduslikult (kollektiivsuse kasv, avatuse suurenemine). Arutame nende asjaolude üle lähemalt.

1. Kolm sõnastikusüsteemi, sadakond sõnakogu

Ajalooliselt esimene sõnastikusüsteem oli Eesti Keele Instituudis EELex (kasutusel alates 2003. aastast), selle kõrvale lisandusid teised ning ühel hetkel kasutasime üht sõnastikusüsteemi ja kaht terminibaasi, igal oma head ja vead. EELex oli mõeldud üldkeele sõnaraamatute jaoks, terminoloogid kasutasid terminibaasisüsteeme Termeki (pole enam kasutusel) ja MultiTerm.

EELexi andmemudel (Langemets jt 2010; Jürviste jt 2011) on üles ehitatud sõnast lähtuvalt, esindades nii semasioloogilist mudelit. Töörühmad olid (ja on) autonoomsed, igaüks töötas oma soovide järgi kujundatud sõnastikubaasis. Tulemus oli, et kõik baasid (kokku umbes 70) on „natuke struktureeritud", igaüks ise viisil, kõigis on mingis ulatuses korduvat infot (nt märksõnad, morfoloogiline info, tähenduste seletused), mis samuti on esitatud mitmel eri moel; uued sõnad lisandusid igasse üksikusse baasi eraldi. Algusaegade XML-andmebaas kasvas teatavaks XML-i ja relatsioonilise andmebaasi segamudeliks. Hiljem lisandus ka mõistepõhine ehk onomasioloogiline andmemudel, mis siiski väga jõudsat kasutust ei leidnud. Iga valminud sõnaraamat avaldati omaette veebilehel.

Termeki süsteemi arendas Werkdata Ltd aastatel 2007–2015, aastal 2012 omandas EKI õiguse seda Eesti terminitöö jaoks kasutada. Termekis töötasid eeskätt instituudivälised terminikomisjonid ja üksiktegijad. Termeki on relatsiooniline andmebaas osaliselt modifitseeritava onomasioloogilise andmemudeliga. Selles loodi umbes 40 terminibaasi ja üks kakskeelne üldkeelesõnastik (läti-eesti). Iga valminud sõnaraamat tehti kasutajale elektrooniliselt kättesaadavaks Termeki suhteliselt tehnilises keskkonnas ning omaette baasina.

[1] https://elex.is/tools-and-services (10. II 2021).

MultiTermi süsteemi kasutas kaks suurt terminibaasi: mitmekeelne Esterm ning sõjanduse ja riigikaitse termineid koondav Militerm[2]. MultiTerm on samuti relatsiooniline andmebaas osaliselt modifitseeritava onomasioloogilise andmemudeliga.

EELexi tüüpi süsteemi piiratust kogesime kõige teravamalt 2016. aastal, kui lõime poolautomaatselt (kolme olemasoleva baasi põhjal) eesti-soome sõnaraamatu baasi (vt Langemets jt 2017), milles eesti osa aluseks sai samal ajal töös olnud uus eesti keele sõnaraamat (EKS 2019, mõlemad sõnaraamatud ilmusid 2019). Rööbitine töö nende sõnaraamatutega sisaldas väga palju dubleerimist: kui eesti poolel mingi otsus tehti ja esitust muudeti, tuli sedasama käsitsi korrata ka eesti-soome baasis. Oli vaja kardinaalselt teistsugust, andmete vahetust ja taaskasutust võimaldavat, ühtset sõnastikusüsteemi.

Terminitöös seati ühtne üldkasutatav terminibaas eesmärgiks juba 2008. aastal algatatud riiklikus terminoloogiaprogrammis (Valge 2013). Kümme aastat hiljem polnud me päriselt sinna veel jõudnud: olid olemas „natuke ühtne" Termeki ja eraldiseisev Multiterm, aga ühtset baasi ei olnud. Füüsikud küsisid otse: „Kas on olemas plaan, kuidas lõpuks tekib üks päringuaken, kuhu sõna sisestades pöördutakse kõigi erialasõnastike poole?" (Kuusk, Reivelt 2019) Füüsikute arupärimise ajal oli keeleinfo ühendamine tegelikult EKI-s juba alanud. Üks päringuaken tekkis esimesena 2019. aastal üldkeeleinfo jaoks, aasta hiljem ka erialasõnastike jaoks. Sõnastikuinfo koduks sai uus keeleportaal Sõnaveeb.

2. Kogu keeleinfo ühes kohas

Sõnastikuinfo ühendamine algas EKI-s 2017. aastal, kui koostöös tarkvarafirmaga TripleDev OÜ alustasime EKI uue andmebaasi Ekilex loomist (Tavast jt 2018). Ühtse baasi selgrooks ehk tuumikuks kujundasime „Eesti keele sõnaraamatu 2019" (EKS 2019; Tavast jt 2018). Oleme järginud digimaailma põhimõtet, et andmed ja andmete esituskuju seisavad üksteisest lahus (vt ka Gorjanc jt 2017: 11; Langemets jt 2018; Tavast jt 2018).

Sõnastiku- ja terminibaasisüsteem Ekilex on leksikograafide ja terminoloogide profitööriist, mida kasutavad nii EKI enda kui ka majavälised sõnakogude koostajad ja toimetajad (kokku üle 300). Baasis tegutsejate arv kasvab pidevalt koos uute keelte lisamise ja uute terminibaaside loomisega. Eesti keelendeid on Ekilexi andmebaasis kokku ligi 300 000 (seisuga aprill 2021) ehk kaks korda rohkem kui omaaegses suurimas, eesti keele seletavas sõnaraamatus (EKSS 2009). Nii üldkeele- kui ka oskussõnastikes esinevaid keelendeid on 60 000. Üksnes terminibaasides esinevaid termineid on 120 000 (kokku on üle 80 terminibaasi).

Keeleportaal Sõnaveeb on koht, kus kasutajad saavad ühe päringuakna kaudu kogu infot uurida. Keeleõppijale on mõeldud vähema sõnavalikuga (umbes 6000 sõna) ja lihtsama ülesehitusega Keeleõppija Sõnaveeb. Sõnaveebi kasutab kuus üle 32 000 inimese (Google Analyticsi andmed, aprill 2021). 2019. ja 2020. aasta võrd-

[2] Alates märtsist 2021 jätkub Militermi koostamine Ekilexis.

luses paistab kasv püsiv ja ühtlane, mitte hüppeline: aasta jooksul on kasutajate arv (ühe kuu kohta) kasvanud 10 000 võrra.

EKI ühendsõnastik (Sõnaveebis) on eesti keele mahukaim sõnakogu, kuhu oleme koondanud mitmest EKI tänapäeva keele sõnastikust pärit info. Ajapikku hakkab seda täiendama ka õigekeelsussõnaraamat. Ühendsõnastiku tuumik on eesti keele sõnaraamat (EKS 2019), koos teiste sõnastikega nimetame seda alates 2020. aastast EKI ühendsõnastikuks. Sõnastike ühendamine üheks ühiseks andmebaasiks on keeruline protseduur, mida täielikult automatiseerida ei saa. Andmete ühendamisel oleme kasutanud spetsiaalselt loodud tähenduste ühendamise tööriista. Ühendamise rõõmustav pool on, et eri sõnakogudest korjub kokku väärtuslikke andmeid, näiteks sõnu ja tähendusi, mida andmebaasis veel pole registreeritud, samuti mitmekesist infot sünonüümide, rektsioonide jm kohta. Tülikam pool on, et leksikograafid peavad kõvasti vaeva nägema, et andmeid korrastada ning puhastada (eri sõnakogudest tulnud) liiasest infost, kordustest ja ebajärjekindlustest. Näiteks eesti-vene sõnaraamatus (EVS) on eesti keele tähendusjaotust osaliselt painutatud sihtkeele (vene keele) järgi. Kui eesti *nägu* tähistab nii inimese kui ka looma kehaosa, siis vene keeles jaguneb mõiste kaheks: inimesel on *лицо* ja loomal *морда*; sama on sõnaga *ämm*: eesti keeles on ühendmõiste 'mehe- või naiseema', aga vene keeles eraldi mõisted 'meheema' (*свекровь*) ja 'naiseema' (*тёща*). Nüüdset Ekilexi selgroogu hoiame kindlalt paigas eesti keele järgi (sealhulgas tähendused võivad muutuda), ent see nõuab vene keele osas ulatuslikku toimetamis- ja korrastustööd. Oleme väitnud enesekriitiliselt, et alahindasime impordi võimekust andmete harmoniseerimisel (Tavast jt 2018). Õppetunni saanuna ei impordi me edaspidi tervet järjekordset sõnakogu, vaid analüüsime põhjalikult, millist uut infot sellest kogust Ekilexi juurde tuua. Nii toimime näiteks alates 2020. aastast töös oleva õigekeelsussõnaraamatu ÕS info puhul, kus oleme alustanud ÕS-i soovituste ülevaatamist ja ühendsõnastikus esitamist (Langemets, Päll 2020).

Võib öelda, et oleme pikkade sammudega liikunud nn tõelise e-sõnaraamatu põhipunktide poole, mis on välja toodud ligi kümme aastat tagasi (Fuertes-Olivera, Bergenholtz 2011; Langemets 2012). Nüüd kinnitame ka meie, et moodne sõnaraamat 1) on info juurde pääsemise abivahend, infotehnoloogiline tööriist (info on päritav ka API ehk rakendusliidese kaudu); 2) on (relatsiooniline) andmebaas, kus iga andmeelement on ainukordne; 3) on allikas, millest erinevate linkide ja seoste abil luuakse (virtuaalne) sõnaraamat või valikuline väljavõte soovitud infost (vt ka Bajčetić, Declerck 2020); 4) arvestab andmete esitamisel kasutajate eri tasemete ja vajadustega. Kui enamik kasutajaid on ilma suuremate nõudmisteta tavainimesed, siis eksperdid soovivad võimalikult detailset infot. Näiteks saab Sõnaveebis morfoloogilist infot vaadelda kolmel tasemel: põhivormide ulatuses, täisparadigmana ja morfofonoloogilises transkriptsioonis (koos välte, rõhu jt märkidega) esitatud täisparadigmana.

Sõnaveebi tagasiside kaudu teame, et paljud kasutajad soovivad mugavamat võimalust kasutajaliidest enda käe järgi kohandada, infot detailsemalt otsida ning enda jaoks paremini organiseerida. See töö seisab meil alles ees.

3. Muutuv sõnaraamatutöö

Moodne aeg on leksikograafiasse ja leksikograafide jaoks kaasa toonud uusi vaatenurki, millest nii mõnigi on nõudnud harjumist, et muutus omaks võtta. Uut moodi on korraldatud sisuline töö (infokihid, korpuspõhisus, andmete dünaamilisus) ja omavaheline koostöö. Kõik on pidanud kohanema uue ühtse sõnastikusüsteemiga, aga ka aru saama andmemudelist, nii selle võimalustest kui ka piiridest.

3.1. Töö infokihiga, mitte sõnastikuga

Kui varem töötas iga töörühm oma sõnastikuga, siis nüüdse aja leksikograafid töötavad ühtses baasis, aga enamik neist keskendub teatavale andmekihile (vt lähemalt vt Tavast jt 2020), näiteks sünonüümidele, teiste keelte vastetele, ÕS-i soovitustele, sõnaliikidele (vt lähemalt Paulsen jt 2020), kohanimedele, teatava valdkonna terminitele jne. Samal ajal näevad leksikograafid sõna või mõiste kohta kogu infot ja vajadusel täiendavad seda. Uusi andmekihte on võimalik luua jooksvalt, vastavalt vajadusele. Leksikograafid on pidanud piltlikult öeldes hakkama kohe sõitma vahendil (Ekilex), mida samal ajal ise ehitame ja arendame.

EKI ühendsõnastiku ja ühtse baasi sünniga on leksikograafid pääsenud suurest hulgast korduvast tööst, mis varem oleks kuulunud vältimatult iga uue sõnastikuprojekti juurde: näiteks märksõnaloendi loomine, häälduse ja morfoloogilise info esitus, eesti keele tähendusjaotus.

Iga andmekiht toob mõistagi kaasa probleemid, millega tuleb eraldi tegelda. Nii oleme näiteks sünonüümikihi koostamisel muutnud andmemudelit mitu korda. Kui algul koostasime automaatselt genereeritud valiku najal sünonüüme ühepoolselt (a = b), siis alates 2020. aasta lõpust teeme seda kahesuunaliselt (kui a = b, siis b = a). Teisisõnu: kui algul sai *koer* käsitsi koostamise ajal sünonüümiks *peni*, siis vastupidi seost automaatselt ei tekkinud, aga nüüd tekib (Tavast jt 2020). Teiste keelte lisandudes tuleb lahendada keeltevaheliste tähenduserinevuste esitamise probleem: näiteks vaste võib olla eesti sõnast/terminist laiem või kitsam (nt eesti *ämm* jaguneb vene keeles kaheks mõisteks: *meheemaks* ja *naiseemaks*). Andmemudel võimaldab küll tähistada säärast tähendusseost, ent kasutajale me mõistlikul viisil tulemust veel esitada ei saa. See on seotud sellega, et tänapäeva leksikograafias seisavad andmebaas ja info esituskuju (kasutajaliides) üksteisest lahus, on tehniliselt eri asjad. Keeltevahelised tähendusseosed on plaanis lahendada lähitulevikus.

ÕS-i soovituste andmekihi ühendame välislinkide abil teiste õigekeelsusallikatega, eeskätt (samuti töös oleva) EKI teatmikuga ehk õigekeelsuskäsiraamatuga. Nii on võimalik kohe tutvuda põhjalikuma taustamaterjaliga.

3.2. Tekstikorpused ja deskriptiivne keelekirjeldus

Leksikograafilise kirjelduse põhiallikad on tänapäeva leksikograafias mahukad elektroonilised tekstikogud ehk tekstikorpused. Korpuste analüüsimiseks kasutatakse korpuspäringusüsteeme (nt Sketch Engine, KORP), mis võimaldavad arvandmetele

toetudes sõnade kasutust ja kasutuse muutumist mitmekülgselt analüüsida. Traditsiooniliselt on korpusi loodud hoolikalt valitud, kindla päritolu ja kvaliteediga allikatest: nii on loodud nt eesti keele koondkorpus.[3]

Alates 2000. aastatest on järjepidevalt kasvanud veebitekstide maht ning neid on hakatud kasutama veebikorpuste loomiseks (Grefenstette, Nioche 2000; Cavagliá, Kilgarriff 2001; Pomikálek jt 2012; Suchomel, Pomikálek 2012). Veebitekstid kogutakse kokku spetsiaalse programmiga (ingl *crawler*) SpiderLing (vt lähemalt Suchomel, Pomikálek 2012; Suchomel 2020), mis mööda veebilinke liikudes laeb alla sealse tekstilise materjali. Tekst kodeeritakse UTF-8 vormingusse ning puhastatakse (nt muukeelsetest tekstidest), samuti eemaldatakse identsed või väga sarnased dokumendid. Seejärel tekstid lemmatiseeritakse, märgendatakse morfoloogiliselt[4] ning laetakse üles korpuspäringusüsteemi (EKI-s Sketch Engine'isse)[5]. Veebikorpuste puhul kerkivad esile uued probleemid: näiteks on keeruline tuvastada või eristada üht keelt teisest sarnasest keelest (nt vene keelt ukraina või valgevene keelest, vt Koppel jt 2019b: 776), tuvastada duplikaate, veebispämmi ja masintõlkelisi tekste (vt lähemalt Suchomel 2020). Peale korpuse sisu (allikate valiku) ja mahu probleemide on sagedad ka lemmatiseerimise, morfoloogilise märgenduse ja (osa)lausestamise ning mitmesõnaliste üksuste tuvastamise vead, (vormi)homonüümia ja polüseemia (Koppel 2020: 56–63).

Kuigi veebikorpusi on kritiseeritud selle poolest, et neis pole tagatud lingvistiline variatiivsus (st kaetud pole teatud žanrid, nt ilukirjandus ja suuline keel), kaalub nende suurus puudused üle (Cvrček jt 2020): mahukast korpusest on võimalik kasutusnäiteid leida ka väga madala sagedusega keelenähtuste kohta (Pomikálek jt 2009).

Eesti keele jaoks pole kunagi varem olnud nii suuri (ja reaalajas täienevaid) tekstikorpusi, kui on viimase kümne aasta jooksul sündinud EKI ja tarkvarafirma Lexical Computing Ltd. koostöös. Suurim korpus – eesti keele ühendkorpus 2019 (1,5 mld sõna) – koosneb umbes 88% ulatuses just veebist kogutud tekstidest. Ühendkorpus on žanriliselt mitmekesiseim allikas tänapäeva eesti keele uurimiseks, mille allkorpustest saab omakorda eraldi päringuid teha.

Tekstikorpustest avaneb leksikograafile pilt tegelikust keelekasutusest. Mida automaatsemalt korpust analüüsida, seda deskriptiivsem (objektiivsem?) on leksikograafiline keelekirjeldus. Juba on võimalik genereerida märksõnaloendeid, tuvastada definitsioone, kollokaate, tähendusseoseid, näitelauseid ja muid üksusi (vt lähemalt

[3] Koondkorpus on loodud Tartu Ülikoolis ning see sisaldab ajakirjandus-, ilukirjandus- ja teadustekste, riigikogu stenogramme ning Eesti ja Euroopa seadustekste aastatest 1990–2008. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (31. I 2021).

[4] Eesti keele korpuste lemmatiseerimiseks ja märgendamiseks kasutatakse tekstianalüsaatorit EstNLTK. Uusimad (alates 2019. aastast loodud) korpused on märgendatud versiooniga 1.6, mille teek on kättesaadav aadressil https://www.keeletehnoloogia.ee/et/estnltk-1-6-teegi-stabiilne-versioon (31. I 2021).

[5] Märgendamata kujul on korpus allalaadimiseks kättesaadav CC-BY-SA litsentsi alusel Eesti Keeleressursside Keskuse repositooriumis Entu.

Koppel 2020: 11–14).[6] Korpusleksikograafia alusepanija John Sinclair (1991: 39, 60–61) on rõhutanud, et puhtobjektiivne keelekirjeldus *per se* pole piisavalt üldistatud: andmed luuakse küll automaatselt, ent korpusandmeid hindavad ja kokkuvõtte teevad ikka leksikograafid.

Järgmist õigekeelsussõnaraamatut koostades puutume kokku deskriptiivse (korpuspõhise) ja preskriptiivse (ÕS 2018-st pärit) keelekirjelduse vastuoludega. Seni paiknesid eri tüüpi sõnaraamatud (EKSS 2009 ja ÕS 2013; EKS 2019 ja ÕS 2018) üksteisest lahus, omaette veebilehel. Nüüd otsime lahendusi, kuidas kasutajale arusaadaval moel esitada ühendsõnastikus mõlemat: nii tegeliku keele kirjeldust kui ka keelekorralduslikke ÕS-i soovitusi. Oleme alustanud ÕS-i soovituste ülevaatamist teemavaldkondade kaupa. Ka Sinclair (1991: 60–61) on hoiatanud, et preskriptiivsed väited ei tohiks keeleandmeid ignoreerida või neist liiga eemalduda, muidu võivad ettekirjutused kaotada tõsiselt võetavuse.

3.3. Märgatav kollektiivsuse kasv, info dünaamilisus

Varem eraldi töötanud väiksemad eri sõnastike töörühmad (nt ÕS-i koostajad, seletava sõnaraamatu koostajad, naabersõnastiku koostajad) on nüüd koondunud üheks suureks EKI meeskonnaks, kes kõik tegutsevad alates 2020. aastast ühes andmebaasis. Iga sõnaraamat on enamasti olnud kollektiivse töö tulemus, mis toetub projekti jaoks loodud ja aeg-ajalt täienevale stiiliraamatule või koostamisjuhendile. Kõik leksikograafid teavad, kui keeruline on eri inimestel sarnast joont hoida, et subjektiivsust ei õnnestu täielikult vältida: igas kollektiivselt sündinud keelekogus on annus ebaühtlust. Ühendsõnastiku baasis, kus töötab üheskoos üle 40 inimese (neist suur hulk instituudiväliseid kaastöötajaid) ning mille maht kasvab pidevalt, on ühtsuse hoidmine tunduvalt keerulisem. Arvatavasti peame – eeskätt leksikograafid ise – edaspidi harjuma n-ö alatise teelolemisega, mõnikord ka asjaolude muutumisega teekonnal ideaalse (saavutamatu) lõpp-punkti suunas. Aga suur muutus on see ka kasutajate jaoks: polda harjutud, et töö käib kõigi silme all. Kui varem ilmus sõnastik siis, kui ta oli lõpuni viimistletud, käsikiri korrektuurini üle loetud, siis nüüd ilmutame uut infot jooksvalt: EKI ühendsõnastik uueneb iga päev, samuti kõik terminibaasid, kus parasjagu töö käib. Kasutaja kaaluda on, mis on parem: kas lõpliku kuju saanud staatiline sõnastik (mis hakkab vananema alates ilmumisele järgnevast päevast) või pidevalt uuenev, ühiskonnas aktuaalset keelekasutust registreeriv või keelemurele lahendust otsiv keeleportaal.

Koos rahvahanke algatustega muutub kogu sõnaraamatutöö aina laiapõhjalisemaks, selles osaleb üha rohkem inimesi: näiteks võib vabatahtlikke rakendada eri keelte vastete lisamisel, etteantud vastete või sünonüümide sobivuse hindamisel. Ses osas on Eestil arenguruumi küllaga. Seda tüüpi kaasamise pea ainus näide on 2019. aastal ilmunud eesti assotsiatsioonisõnastik (Vainik 2019), kus esitatavad seosed on kogutud testimise teel tavalistelt keelekasutajatelt.

[6] 2018. aasta üleeuroopalisest küsitlusest selgus, et Euroopas luuakse poolautomaatselt ligikaudu 31% ja täisautomaatselt umbes 7,5% leksikaalsetest andmebaasidest (Kallas jt 2019).

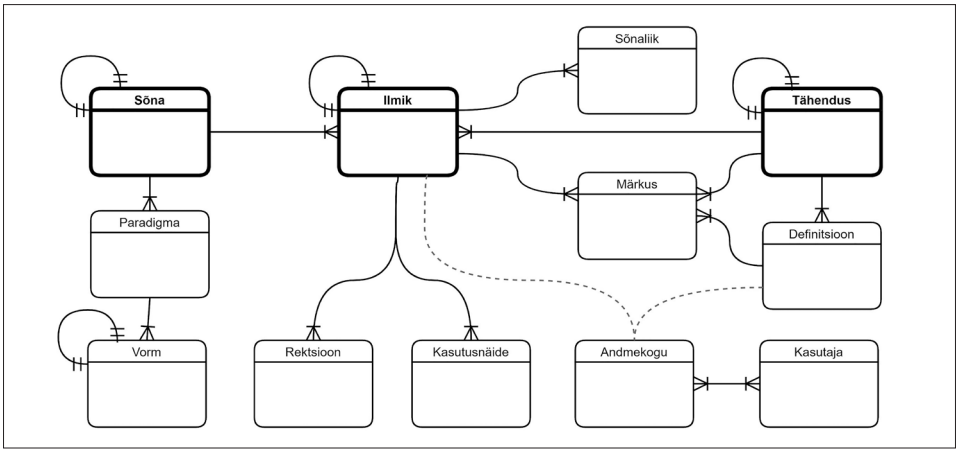
Digiajastu heade külgede ja vigade üle arutlenud Jaan Undusk ja Marek Tamm (2021) on osaluskultuuri ja vabatahtlikku koosloomet hinnanud digiajastu positiivseks küljeks – need on uued vormid ühistes platvormides, mis tihti aitavad asju kiiremini edendada. Spetsialistidele, sh leksikograafidele jääb järelkontrolli roll.

3.4. Infotehnoloogilist: andmemudel ja API

Leksikograafiliste andmete kirjeldamiseks on pakutud mitu erinevat mudelit, näiteks OntoLex-Lemon (McCrae jt 2017) ja selle leksikograafiline moodul Lexicog, TEI Lex-0 (Bański jt 2017; Tasovac jt 2018), LMF ISO-24613 (Francopoulo jt 2006), Eestis Ekilex (Tavast jt 2018).

Ekilexi andmemudel erineb varasematest sõnastikusüsteemidest (EELex, Termeki, Multiterm) seose tüübi poolest. Kolmes varasemas andmemudelis on vormi (sõna/keelendi) ja tähenduse vahel üks mitmele seos 1:n, st ühel sõnal on mitu tähendust (semasioloogilises mudelis) ja ühel mõistel on mitu terminit (onomasioloogilises mudelis). Nende andmemudelite puuduste kohta vt Tavast jt 2018. Ekilexi andmemudelis kasutame vormi (sõna/keelendi) ja tähenduse vahel mitu mitmele seost m:n, st ühel sõnal võib olla mitu tähendust ja üks tähendus (mõiste) võib olla seotud mitme sõnaga, ehk mõlemat pidi on lubatud palju seoseid. Kuna mitu mitmele seose puhul ei saa luua otseseost, siis seob neid vahetabel (vt joonist 1, vahetabeli rollis on ilmik, selle kohta allpool). Sellist seost võib kirjeldada ka kui sõna ja tähenduse andmete pidevat taaskasutamist: relatsioonilistes andmebaasides ühendavad olemeid mitmesed seosed (tabelid), m:n iga binaarse seose jaoks luuakse eraldi seos (tabel).

Andmemudelis nimetame seda mitu mitmele seosega olemit *lekseemiks*, mis lahtiseletatult tähendab 'see sõna selles tähenduses' (seos sõna ja tähenduse vahel; joonisel 1 tähistab seda seos sõnast tähenduseni). Kui algul oli see pelgalt tehniline



Joonis 1. Ekilexi andmemudeli skeem 2021.

seos, siis arenduse käigus on sellest seosest saanud andmemudeli võtmeüksus, mis omakorda seob teisi olemeid: enamik andmeüksusi (nt sõnaliik, rektsioon, õppekommentaarid, kasutusnäited) on lekseemi parameetrid (vt joonist 1).[7]

Ekilexi andmemudel koondab ühtsesse baasi kogu keelelise info, mille algkodu on olnud paljudes eraldiseisvates sõnakogudes. Sõna ja tähenduse mitu mitmele seose vahetabeli rolli täidab olem, mida andmemudelis nimetame *ilmikuks* (vt joonist 1), mis lahtiseletatult tähendab 'see sõna selles tähenduses selles sõnakogus'. Teoreetilise panusena leksikograafiasse saame lekseemi toel kirjeldada sõna ja tähendust kui keeleinfot, mitte kui sõnastikuinfot (Tavast jt 2018). Ehk teisisõnu, püüame ühe sõnatähenduse kirjeldamiseks pakkuda kasutajale variante, nt detailset seletust, vähemat tähendusvihjet, lihtsustatud keeles seletust. Eri sõnastikud on seni kasutanud kõiki variante läbisegi.

Kolme aasta jooksul oleme Ekilexi mudelit mitu korda muutnud, ka nüüd on juba näha, et muutusi tuleb veel. Algne mudel ei olnud seevõrra rekursiivne nagu nüüdne: erinevus seisneb selles, et nüüd on ühtsustatud sõnaks/keelendiks kõik vormiga seotud olemid (sõna/keelend, kollokatsioon (naabersõnad), kasutusnäide, võib-olla edaspidi isegi definitsioon). Toona kaalusime sama, aga otsustasime teisiti: kuigi neil üksustel on mõni sarnane omadus (nt eriti tähendus), aga siiski paljud omadused erinevad, siis jäi iga nimetatu omaette olemiks (vt joonist Tavast jt 2018: 753). Tol hetkel tundunuks see samm ka liiga radikaalne. Kolmeaastase arenduse tulemusel oleme otsustanud ikkagi mudelit muuta rekursiivsuse suunas: kui enne olid kollokatsioonid eraldi olemid, mis olid lekseemiga seotud, siis nüüd on kollokatsioon ise sõna/keelend, mille juures saab (sõnakoguspetsiifiliste) ilmikuseostega esitada, millistest komponentidest ja kuidas see koosneb (vt joonisel 1 omaseost ilmiku kasti küljes). Kollokatsiooni komponendid on sõnad oma mingis tähenduses. Ilmikuseosega on andmemudelis seotud näiteks allmärksõna (vt EKI ühendsõnastikus *aadress* : *aadressil*) ja õppekommentaar (vt Keeleõppija Sõnaveebis *ükssada*). Tavalise (mitmesõnalise) märksõna ja kollokatsiooni vahe seisneb selles, et viimasel esineb (peale muude ilmikuseoste) ka kollokatsioonitüüpi ilmikuseoseid (nt info kollokaadi sõnaliigi ja grammatilise suhte kohta). Oleme jõudnud samale järeldusele, mida 2019. aasta e-leksikograafia konverentsil väljendas üks OntoLex-Lemoni arendajaid Thierry Declerck: „I don't like the idea of lexicographic entry" ('mulle ei meeldi mõte sõnaartiklist'). Ekilexis oleme teel ühtse universaalse struktuuri poole, mis sobiks kõigile keeleandmetele.

Kahtlemata on leksikograafide jaoks Ekilexi andmemudel (vt joonist 1) keeruline. Samal ajal on sellel vaieldamatult suur mõju leksikograafilistele tööprotsessidele (Paulsen jt 2020: 180–181), näiteks sõnaliigi määramisel (lekseemi parameetrina) on mõistlik olla kooskõlas toimivate tekstitöötlusvahendite (nt EstNLTK) liigitusega. Kui keeleteaduses on näiteks sünonüümiat, antonüümiat ja kaashüponüümiat tõlgendatud ühtmoodi kui leksikaalseid suhteid, siis on raske mõista, kuidas siin mudelis käituvad nad erinevalt. Leksikaalsed suhted on mudelis defineeritud kahel viisil: 1) täissünonüümid ja täisvasted on sama tähendusega sõnad (ühes keeles või eri

[7] OntoLex-Lemoni mudelis (McCrae jt 2017) vastab lekseemile leksikaalne tähendus („Lexical Sense"), mis töötab põhimõtteliselt samal viisil („mapping from a word to a concept").

keeltes), nad on seotud ühe tähendusega, st nende jaoks pole eraldi lisaseost; 2) kõik teised leksikaalsed suhted (nt kaashüponüümid, vastandid, osasünonüümid[8]) on mõistelised seosed eri tähenduste vahel (mitte sõnade vahel).

Ekilexi programmiliides (API) on mõeldud arendajatele, kes soovivad kasutada Ekilexi andmeid eri tüüpi rakenduste loomiseks, näiteks liitsõnaraamatuks, õppematerjaliks, äpiks jne. API võimaldab andmeid küsida JSON-vormingus. Seni on EKI API-t kasutatud eelkõige Horizon 2020 projekti ELEXIS raames.

4. Kokkuvõtteks

Läinud 20 aastat on näidanud, et digimaailma normaalne olek ongi pidevalt pöördeid teha. Ühelt poolt kogeme kustumatut uudishimu leiutada uusi võimalusi keele uurimiseks ja kirjeldamiseks, teiselt poolt peame pidevalt tegelema olemasolevate süsteemide järeleaitamisega. Sõnaraamatutöö on muutunud väga interdistsiplinaarseks: leksikograafid peavad tegema koostööd arvutilingvistide ja keeletehnoloogidega, linkandmete, semantilise veebi ja tehisintellekti spetsialistidega (Steurs jt 2020). Samuti ollakse suuremaks kasvanud oma riigist: hulk töövahendeid on kõigile vabalt saadaval üle Euroopa. Kindlasti tehakse koostööd veebiliideste spetsialistidega. Tänapäeva leksikograafi pädevus ulatub lingvistilistest teadmistest väljapoole: tal on vaja oma igapäevatöö jaoks tehnilist, n-ö andmebaasilist mõtlemist ja tehnilisi oskusi. Leksikograafi töö muutub ühelt poolt üha rohkem järeltoimetamiseks, teiselt poolt üha laiapõhjalisemaks ja demokraatlikumaks, hõlmates vabatahtlikke kaastöölisi ja uusi töövorme uutes ühisplatvormides. See on veel paljuski avastamata maa.

Eestis oleme (Ekilexis ja Sõnaveebis) saanud juba kasutada mitmeid keeletehnoloogilisi lahendusi ja rakendusi, muuhulgas kõnesünteesi (näitelausete ettelugemiseks), kõnetuvastust ja hääljuhtimist (sõnaotsingul), automaatset kollokatsioonide ja veebilausete tuvastamist ning kuvamist, sünonüümi- või vastekandidaatide pakkumist.

Kuhu leksikograafia edasi areneb? Võib tunduda ootamatu, aga ... nähtamatuse suunas. Gilles-Maurice de Schryver (2019) ja teised keeletehnoloogid (nt Tasovac 2010) on kirjeldanud leksikaalseid andmeid kui üksusi, mis on märkamatult integreeritud mitmesugustesse rakendustesse, näiteks keeleõpperakendustesse, masintõlkeplatvomidesse, tekstitöötlusprogrammidesse. Ehk teisisõnu, tuleb aeg, kui tekst sisaldab sõnastikku samavõrra kui praegu sõnastik sisaldab teksti.

Artikkel on valminud Haridus- ja Teadusministeeriumi projekti EKKD64 „Eesti keele sõnavara ja korraldus: deskriptiivne ja preskriptiivne vaatenurk", Euroopa Liidu programmi Horisont 2020 projekti INFRAIA-02-2017 „European Lexicographic Infrastructure" ja Eesti Keele Instituudi baasfinantseerimise toel.

[8] Osasünonüüme ehk peaaegu sarnase tähendusega sõnu on nimetatud ka lähisünonüümideks. Siin kasutame terminit osasünonüümid.

VEEBIVARAD

EELex. Eesti Keele Instituudi sõnastikusüsteem. https://eelex.eki.ee
Eesti keele ühendkorpus 2019. https://dx.doi.org/10.15155/3-00-0000-0000-0000-08565L
Ekilex. Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem. https://ekilex.eki.ee
Ekilex API. https://github.com/tripledev/ekilex/wiki/Ekilex-API
EKI ühendsõnastik 2021. Eesti Keele Instituut. Sõnaveeb, 2021. https://sonaveeb.ee/collections
EKS 2019 = Eesti keele sõnaraamat 2019. Eesti Keele Instituut. Sõnaveeb, 2019. www.eki.ee/dict/eks; https://doi.org/10.15155/3-00-0000-0000-0000-08240L
Elexifier. https://elexifier.elex.is
ELEXIS = European Lexicographic Infrastructure. https://elex.is
Entu. https://entu.keeleressursid.ee
Esterm. https://termin.eki.ee/esterm
EstNLTK = Estonian Natural Language ToolKit. Kogumik teeke eestikeelsete tekstide töötluseks. Versioon 1.6.2. https://estnltk.github.io
EVS = Eesti-vene sõnaraamat 2019. 2., täiendatud ja kohandatud veebiväljaanne. Eesti Keele Instituut. Sõnaveeb, 2019. https://www.eki.ee/dict/evs
Keeleõppija Sõnaveeb. Eesti Keele Instituudi keeleportaal. Versioon 1.21. https://sonaveeb.ee/lite
KORP. Korpuspäringusüsteem. https://korp.keeleressursid.ee
Lexicog. https://www.w3.org/2019/09/lexicog
Lexonomy. https://www.lexonomy.eu/
MultiTerm. https://www.trados.com/products/multiterm-desktop/
NAISC. https://github.com/insight-centre/naisc
Sketch Engine. Korpuspäringussüteem. https://www.sketchengine.eu
Sõnaveeb. Eesti Keele Instituudi keeleportaal. Versioon 1.21. https://sonaveeb.ee

KIRJANDUS

Bajčetić, Lenka; Declerck, Thierry 2020. Interlinking Slovene language datasets. – Proceedings of XIX EURALEX Congress: Lexicography for Inclusion. Kd I. Toim Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras. Democritus University of Thrace, lk 73–80.

Bański, Piotr; Bowers, Jack; Erjavec, Tomaz 2017. TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 485–494.

Cavagliá, Gabriela; Kilgarriff, Adam 2001. Corpora from the Web. – Proceeding of the Fourth Annual CLUCK Colloquium, Sheffield, UK, lk 120–124.

Cvrček, Václav; Komrsková, Zuzana; Lukeš, David; Poukarová, Petra; Řehořková, Anna; Zasina, Adrian Jan; Benko, Vladimír 2020. Comparing web-crawled and traditional corpora. – Language Resources and Evaluation, kd 54, nr 3, lk 713–745. https://doi.org/10.1007/s10579-020-09487-4

de Schryver, Gilles-Maurice; Chishman, Rove; da Silva, Bruna 2019. An overview of digital lexicography and directions for its future: An interview with Gilles-Maurice de Schryver. – Calidoscópio, kd 17, nr 3, lk 659–683. https://doi.org/10.4013/ld.2019.173.13

EKSS = Eesti kirjakeele seletussõnaraamat. Kd I–VII. Tallinn: Eesti Keele Instituut, 1991–2007.

EKSS 2009 = Eesti keele seletav sõnaraamat. Kd I–VI. „Eesti kirjakeele seletussõnaraamatu" 2., täiendatud ja parandatud tr. Toim Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll. Tallinn: Eesti Keele Sihtasutus. http://www.eki.ee/dict/ekss (5. I 2021).

Francopoulo, Gil; George, Monte; Calzolari, Nicoletta; Monachini, Monica; Bel, Nuria; Pet, Mandy; Soria, Claudia 2006. Lexical markup framework (LMF). – Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). Genoa: European Language Resources Association, lk 233–236.

Fuertes-Olivera, Pedro; Bergenholtz, Henning (toim) 2011. e-Lexicography: The Internet, Digital Initiatives and Lexicography. Great Britain: Continuum.

Gorjanc, Vojko; Gantar, Polona; Kosem, Iztok; Krek, Simon (toim) 2017. Dictionary of Modern Slovene: Problems and Solutions. Ljubljana: University of Ljubljana, Faculty of Arts. https://doi.org/10.4312/9789612379131 https://doi.org/10.4312/9789612379131

Grefenstette, Gregory; Nioche, Julien 2000. Estimation of English and non-English language use on the WWW. – Recherche d'Information Assistée par Ordinateur (RIAO). 6th International Conference, College de France, France, April 12–14. Proceedings. Toim Joseph-Jean Mariani, Donna Harman. Paris, lk 237–246.

Jürviste, Madis; Kallas, Jelena; Langemets, Margit; Tuulik, Maria; Viks, Ülle 2011. Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. – Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10–12 November. Toim Iztok Kosem, Karmen Kosem. Ljubljana: Trojina, Institute for Applied Slovenian Studies, lk 106–112.

Kallas, Jelena; Koeva, Svetla; Langemets, Margit; Tiberius, Carole; Kosem, Iztok 2019. Lexicographic practices in Europe: Results of the ELEX survey on user needs. – Proceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim I. Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 519–536.

Kilgarriff, Adam; Rychlý, Pavel; Smr, Pavel; Tugwell, David 2004. The Sketch Engine. – Proceedings of the XI EURALEX International Congress. Toim Geoffrey Williams, Sandra Vessier. Lorient, France: Université de Bretagne Sud, lk 105–115.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít 2014. The Sketch Engine: Ten years on. – Lexicography: Journal of ASIALEX, kd 1, nr 1, lk 7–36. https://doi.org/10.1007/s40607-014-0009-9

Koppel, Kristina 2020. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Dissertationes linguisticae Universitatis Tartuensis 38.) Tartu: Tartu Ülikooli Kirjastus.

Koppel, Kristina; Tavast, Arvi; Langemets, Margit; Kallas, Jelena 2019a. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. – Pro-

ceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Toim Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 434–452.

Koppel, Kristina; Kallas, Jelena; Khokhlova, Maria; Suchomel, Vít; Baisa, Vít; Michelfeit, Jan 2019b. SkELL corpora as a part of the language portal Sõnaveeb: Problems and perspectives. – Proceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 763–782.

Kuusk, Piret; Reivelt, Kaido 2019. Füüsika eestikeelsest terminivarast riiklikul oskuskeelekuul. – Sirp 31. V.

Langemets, Margit 2012. Lingvistilisest kolonialismist sõnaraamatus. – Keel ja Kirjandus, nr 8–9, lk 598–613.

Langemets, Margit; Loopmann, Andres; Viks, Ülle 2010. Dictionary management system for bilingual dictionaries. – eLexicography in the 21st Century: New Challenges, New Applications. Toim Sylviane Granger, Magali Paquot. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, lk 425–429.

Langemets, Margit; Hein, Indrek; Heinonen, Tarja; Koppel, Kristina; Viks, Ülle 2017. From monolingual to bilingual dictionary: The case of semi-automated lexicography on the example of Estonian–Finnish Dictionary. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 155–171.

Langemets, Margit; Uibo, Udo; Tiits, Mai; Valdre, Tiia; Voll, Piret 2018. Eesti keel uues kuues. Eesti keele sõnaraamat 2018. – Keel ja Kirjandus, nr 12, lk 942–958.

Langemets, Margit; Päll, Peeter 2020. Kust vaadata kirjakeele normi? EKI keelekool. – Postimees. Arvamus ja Kultuur 19. XII.

McCrae, John P.; Bosque-Gil, Julia; Gracia, Jordi; Buitelaar, Paul; Cimiano, Philipp 2017. The OntoLex-Lemon Model: Development and applications. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 587–597.

Měchura, Michal B. 2017 Introducing Lexonomy: An open-source dictionary writing and publishing system. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 662–679.

Paulsen, Geda; Vainik, Ene; Tuulik, Maria 2020. Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 16, lk 177–202. https://doi.org/10.5128/ERYa16.11

Pomikálek, Jan; Rychlý, Pavel; Kilgarriff, Adam 2009. Scaling to billion-plus word corpora. – Advances in Computational Linguistics, nr 41, lk 3–13.

Pomikálek, Jan; Jakubíček, Miloš; Rychlý, Pavel 2012. Building a 70 billion word corpus of English from ClueWeb. – Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC'12). Toim Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. European Language Resources Association, lk 502–506.

Sinclair, John 1991. Corpus Concordance and Collocation. Oxford: Oxford University Press.

Steurs, Frieda; Schoonheim, Tanneke; Heylen, Kris; Vandeghinste, Vincent (toim) 2020. The Future of Academic Lexicography – A White Paper. Version 1.2 https://ivdnt.org/wp-content/uploads/2021/02/The-Future-of-Academic-Lexicography-A-White-Paper.pdf (25. IV 2021).

Suchomel, Vít 2020. Better Web Corpora For Corpus Linguistics And NLP. Doctoral Theses. Brno: Masaryk University, Faculty of Informatics.

Suchomel, Vít; Pomikálek, Jan 2012. Efficient web crawling for large text corpora. – Proceedings of the 7th Web-as-Corpus Workshop (WAC7). 17 April, Lyon, France. Toim Adam Kilgarriff, Serge Sharoff. Lyon, lk 39–43.

Tasovac, Toma 2010. Reimagining the dictionary, or why lexicography needs digital humanities. – Digital Humanities. King's College London, 7th–10th July. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-883.html (5. II 2021).

Tasovac, Toma; Romary, Laurent; Banski, Piotr; Bowers, Jack; de Does, Jesse; Depuydt, Katrien; Erjavec, Tomaž; Geyken, Alexander; Herold, Axel; Hildenbrandt, Vera; Khemakhem, Mohamed; Petrović, Snežana; Salgado, Ana; Witt, Andreas 2018. TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (5. II 2021).

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina 2018. Unified data modelling for presenting lexical data: The case of EKILEX. – Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July. Toim Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana: Ljubljana University Press, Faculty of Arts, lk 749–761.

Tavast, Arvi; Koppel, Kristina; Langemets, Margit; Kallas, Jelena 2020. Towards the super-dictionary: Layers, tools and unidirectional meaning relations. – Proceedings of XIX EURALEX Congress: Lexicography for Inclusion. Kd I. Toim Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras. Alexandroupolis: Democritus University of Thrace, lk 215–223.

Undusk, Jaan; Tamm, Marek 2021. Digimaailm ja mõnda. – ERR, Ööülikool 16. I. https://podcastid.ee/ooulikool/ooulikool-jaan-undusk-ja-marek-tamm-digimaailm-ja-monda/ (1. II 2021).

Vainik, Ene 2019. Eesti keele assotsiatsioonisõnastik. Eesti Keele Instituut. https://www.eki.ee/dict/assotsiatsioonid (25. IV 2021).

Valge, Jüri 2013. Kaks terminoloogiaprogrammi: 2008–2012 ja 2013–2017. – Õiguskeel, nr 4. https://www.just.ee/sites/www.just.ee/files/juri_valge._kaks_terminoloogiaprogrammi_2008-2012_ja_2013-2017.pdf (5. II 2021).

ÕS 2013 = Eesti õigekeelsussõnaraamat ÕS 2013. Toim Maire Raadik. Koost Tiiu Erelt, Tiina Leemets, Sirje Mäearu, M. Raadik. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus. http://www.eki.ee/dict/qs2013 (26. VIII 2021).

ÕS 2018 = Eesti õigekeelsussõnaraamat ÕS 2018. Toim Maire Raadik. Koost Tiiu Erelt, Tiina Leemets, Sirje Mäearu, M. Raadik. Eesti Keele Instituut. Tallinn: EKSA. https://www.eki.ee/dict/qs2018 (26. VIII 2021).

Margit Langemets (snd 1961), PhD, Eesti Keele Instituudi juhtivleksikograaf (Roosikrantsi 6, 10119 Tallinn), margit.langemets@eki.ee

Kristina Koppel (snd 1985), PhD, Eesti Keele Instituudi vanemarvutileksikograaf (Roosikrantsi 6, 10119 Tallinn), kristina.koppel@eki.ee

Jelena Kallas (snd 1976), PhD, Eesti Keele Instituudi vanemarvutileksikograaf (Roosikrantsi 6, 10119 Tallinn), jelena.kallas@eki.ee

Arvi Tavast (snd 1969), PhD, Eesti Keele Instituudi direktor (Roosikrantsi 6, 10119 Tallinn), arvi.tavast@eki.ee

From a collection of dictionaries to a language portal

Keywords: e-lexicography, corpora, dictionary writing system, language portal, data model, API

The article aims to describe some major changes that have taken place in e-lexicography in recent decades in Europe generally and in Estonia in particular. Digital changes have permeated not only the dictionary compilation process but also whole workflow from lexicographic content creation to publication. The focus has shifted from building specific dictionaries to building a central database and infrastructure that can be adapted for further user and NLP applications.

We describe methods and technologies used to better integrate lexicographic data (several tools have been developed within the Horizon 2020 project European Lexicographic Infrastructure), and to better access lexicographic information.

As a turning point for digital change in Estonian lexicography, we consider the start of the development of the new Dictionary Writing System Ekilex and its user interface Sõnaveeb in 2017. The long-term goal is to have a single data source to provide consistent information about the Estonian language. In connection with Ekilex and Sõnaveeb, we discuss several issues: the theoretical foundations of the Ekilex biggest lexicographic dataset, the EKI Combined Dictionary, improvements in lexicographic workflow, and the Ekilex data model and API. The EKI Combined Dictionary contains information layers imported from several monolingual explanatory dictionaries, bilingual dictionaries, a collocations dictionary, and an etymology and

morphology database. The improvements in lexicographic workflow include working in one general database, more cooperation between research groups in the institute and more active involvement of external users.

The Ekilex data model meets the requirements for treating both words and meanings as independent entities and for representing both semasiological and onomasiological data. Created data are stored in Ekilex's PostgreSQL database and comply with all current standards of data exchange. As of April 2021, Ekilex contains approx. 300,000 headwords from general-language dictionaries and more than 90 terminological databases.

Margit Langemets (b. 1961), PhD, Institute of the Estonian Language, Leading Lexicographer (Roosikrantsi 6, 10119 Tallinn), margit.langemets@eki.ee

Kristina Koppel (b. 1985), PhD, Institute of the Estonian Language, Senior Computational Lexicographer (Roosikrantsi 6, 10119 Tallinn), kristina.koppel@eki.ee

Jelena Kallas (b. 1976), PhD, Institute of the Estonian Language, Senior Computational Lexicographer (Roosikrantsi 6, 10119 Tallinn), jelena.kallas@eki.ee

Arvi Tavast (b. 1969), PhD, Institute of the Estonian Language, Director (Roosikrantsi 6, 10119 Tallinn), arvi.tavast@eki.ee

https://doi.org/10.54013/kk764a7

Murded, varieerumine ja korpusandmed

Eitussõna paiknemine võru ja seto eituslausetes

MAARJA-LIISA PILVIK, HELEN PLADO, LIINA LINDSTRÖM

Varieerumine on keele olemuslik osa: keel võib varieeruda nii kõnelejate vahel kui ka ühe inimese kasutuses, sõltuvalt registrist (nt teaduskeel *vs*. ajakirjanduskeel) ja tekstitüübist (nt uudis *vs*. arutlus), suhtluskanalist (nt suuline *vs*. kirjalik suhtlus) ja suhtlusolukorrast (nt ametlik *vs*. mitteametlik suhtlus), suhtlejate vanusest, soost, haridusest, tegevusalast, päritolust ja omavahelistest suhetest, aga ka inimese kognitiivsetest võimetest (nagu mälu), vahetust keelelisest ümbrusest jne. Kõige tuntum on keele murdeline varieerumine: sama keelenähtus võib geograafiliselt eri punktides olla väljendatud erinevalt (nt ainsuse 1. isiku *n*-lõpuline variant *ma olin* on levinud laialdaselt põhjaeesti murretes, lõputa variant *ma oli* lõunaeesti murretes ning osaliselt Lääne-Eestis ja saartel, vt kaarti Saareste 1955: 71). Keelenähtus võib varieeruda ka murdeala sees (näiteks võib 1. isiku pöördelõpp vahel esineda, vahel mitte) ning selle kasutust võivad mõjutada erinevad tegurid. Grammatiliste, eriti süntaktiliste joonte kohta ongi leitud, et nende levikupiirid ei ole reeglina aredad, vaid toimub ulatuslik varieerumine, mis on oma olemuselt pigem tõenäosuslik kui kategoriaalne (Kortmann 2010). Selliste varieeruvate nähtuste kasutusmustrite selgitamisel on abiks sagedusandmed, mis võimaldavad mõõta varieerumise ulatust ning varieerumist mõjutavate tegurite rolli kvantitatiivsete meetoditega.

Artiklis vaatleme lähemalt üht varieeruvat nähtust – ees- ja tagaeituse kasutamist (nt *ei olõq ~ olõ-iq*) – lõunaeesti seto ja võru keelevariantides ning rakendame varieerumise analüüsi (ingl *variationist linguistics*) meetodeid, et välja selgitada eituskonstruktsioonis esineva varieerumise ulatus ning seda mõjutavad tegurid vaadeldaval keelealal. Uurimus esindab korpuspõhist kvantitatiivset dialektoloogiat (vt ülevaateid Lindström, Pilvik 2018; Szmrecsanyi, Anderwald 2018): uurimuse andmed pärinevad eesti murrete korpusest (EMK) ja loodavast seto korpusest (SetKo) ning eitusmustrite analüüsiks kasutame kvantitatiivseid, sagedusandmetel põhinevaid meetodeid. Selle uurimuse kitsam eesmärk on kirjeldada ja seletada keeles esinevat varieerumist, ent oma andmekesksuse ja analüüsivahendite valiku poolest panustab see laiemalt digihumanitaaria valdkonda ning artiklis kirjeldatud meetodeid on võimalik edukalt kasutada ka teistes humanitaaria suundades, nt erinevate klassifitseerimisülesannete lahendamiseks. Seetõttu on meetoditel ja nende kirjeldamisel artiklis ka tavalisest kesksem roll.

Artikli juhtumiuuring – eitussõna asendi varieerumine põhisõna suhtes – on huvitav nii läänemeresoome keelte uurimise perspektiivist kui ka laiemalt, sest ühel suhteliselt kitsal alal on kõrvuti kasutusel kaks sõnajärjevarianti, mis kõne kognitiivse töötlemise mõttes on väga erinevad. Tagaeitus on haruldasem nii keeletüpoloogiliselt kui ka Uurali keelkonna sees, eeseitus on aga tavaline läänemeresoome keeltes ja

naaberkeeltes ning domineerib ka keeletüpoloogiliselt (vt Krasnoukhova jt 2021). Esiteks huvitab meid, kuidas tagaeitus võru-seto keelealal on levinud ning kui järsk või sujuv on üleminek tagaeituselt eeseitusele. Teiseks huvitab meid, millest sõltub ees- või tagaeituse valik alal, kus on ühtaegu kasutusel mõlemad variandid, ning kas see valik toimub võru ja seto alal sarnaselt või erinevalt.

Artikli esimeses osas anname ülevaate varieerumise analüüsi kujunemisest keeleteaduse meetodiks ning kvantitatiivsete meetodite rakendamisest selle sees. Artikli teises osas tutvustame eituse moodustamise viise seto ja võru keeles, kolmandas ja neljandas osas kirjeldame vastavalt juhtumiuuringu aluseks olevat materjali ja meetodeid. Artikli viiendas osas on esitatud tulemused eitussõna asendi varieerumise kohta võru ja seto keeles.

1. Varieerumise analüüsist keeleteaduses

Varieerumise analüüs keskendub ühe keelelise joone (milleks meie näites on eitussõna asukoht põhiverbi suhtes) kahe (või enama) variandi varieerumist mõjutavate piirangute ja tegurite väljaselgitamisele. Varieerumise analüüs sai alguse sotsiolingvistikast 1960-ndatel USA-s. Selle pioneeriks võib pidada William Labovit, kelle huviorbiidis oli ennekõike häälduslik varieerumine ning selle seotus kõnelejate sotsiaalsete omadustega, sh sotsiaalse klassiga. Labovi esimene vastavateemaline artikkel „The social motivation of a sound change" (1963) käsitles häälikumuutusi Martha's Vineyardi saarel. Ta näitas, et pika aja jooksul toimuvad häälikumuutused võivad sünkrooniliselt avalduda varieerumisena sõnade häälduses ning selle varieerumise taga võivad olla sotsiaalsed tegurid, mis mõjutavad variantide valikut ja selle kaudu ka laiemalt keele muutust (Labov 1963; vt ka Hazen 2011). Labov rõhutas sotsiaalse varieerumise seisukohast olulise uurimisobjektina suulist igapäevast keelt (ingl *vernacular*), mida ta kirjeldas kui „stiili, kus oma kõnele pööratakse minimaalset tähelepanu" (Labov 1972: 208) ning mis just seetõttu annab kõige paremini infot sotsiolingvistilise varieerumise kohta ja võimaldab selle kaudu prognoosida keelemuutuse suunda. Labov on tuntud ennekõike New Yorgi väga erineva sotsiaalse ja etnilise taustaga elanikkonna keelekasutuse uurijana (vt Labov 1972).

Varieerumise analüüs algab varieeruva nähtuse, muutuja (kasutatud on ka terminit *variaabel*, ingl *variable*) tuvastamisest. Sotsiolingvistilise varieerumise analüüsi põhiprintsiip on olnud, et keelelised muutujad varieeruvad mingis keelekogukonnas süstemaatiliselt, mitte juhuslikult, ning keeleuurija ülesanne on üles leida, millest valik sõltub. Varieerumist on võimalik modelleerida – kvantitatiivseid meetodeid kasutades selgitada välja, kuidas varieerumine toimub ning mis keelelisi valikuid mõjutab (Tagliamonte 2012: 2; Walker 2013). Keelelise muutuja mõiste defineerimine oli Labovi varasemates töödes üsna lihtne, ta kirjeldas seda kui sama asja ütlemist kahel (või enamal) erineval viisil (Labov 1972: 188). Eesti keeles võib öelda näiteks *lihtsalt* või *lissalt*, *janu* või *jänu*, *teinud* või *teind*, *julgetud* või *juletud*. Sealjuures on oluline, et iga keelelise muutuja puhul oleksid k õ i k v a r i a n d i d arvesse võetud (ingl *principle of accountability*; Tagliamonte 2012: 9–10) ning et variandid

esineksid võrreldavas kontekstis ja oleksid vastastikku asendatavad. Morfoloogiliste ning süntaktiliste muutujate määratlemine on keerukam, sest vastastikune asendatavus ei ole alati nii selge (vt ülevaadet muutuja mõiste arenemisest Tagliamonte 2012: 15–19). Võib näiteks küsida, kas fraasid *mees, kes jooksis üle tee* ning *üle tee jooksnud mees* on tekstides alati vastastikku asendatavad (tähendus ja funktsioonid ei muutu, kui asendada üks teisega), ehkki mõlemas laiendatakse sama nimisõna samasisulise propositsiooniga.

Meetodid, kuidas varieerumist sotsiolingvistilises traditsioonis uuriti, võeti üle sotsiaalteadustest: lingvistikasse jõudis statistiline andmeanalüüs, mida rakendati selleks, et kahe samas positsioonis esineva, vastastikku asendatava keelendi valikut statistiliselt modelleerida, võttes arvesse erinevaid valikut mõjutada võivaid tegureid. Sotsiolingvistikas otsitakse neid tegureid kõnelejate sotsiaalsetest omadustest (sugu, vanus, haridus, etniline rühm, sotsiaalne klass vms), ent need võivad olla ka keelelised (nt hääliku hääldamist võib mõjutada sellele eelnev või järgnev häälik). Mida enam on varieerumise analüüs liikunud lingvistika muudesse valdkondadesse, seda enam on meetodit kasutatud pigem selleks, et välja selgitada lingvistilisi, tekstilisi, kognitiivseid vm tegureid, mis funktsionaalselt sarnaste konstruktsioonide valikut mõjutavad (vt nt ülevaadet Szmrecsanyi 2017). Näiteks eesti keele lühikese ja pika sisseütleva käände kasutuse uurimisest on ilmnenud, et pikka sisseütlevat (nt *majasse*) kasutatakse pisut teisiti kui lühikest (*majja*): pikka vormi eelistatakse pärisnimedega (nt *Tartusse* vs. *Tartu*). Samuti mängib pika ja lühikese vormi valikul rolli mitmete morfofonoloogiliste ja morfosüntaktiliste tegurite koosmõju (Siiman 2018).

Aastakümneid on varieerumise analüüsi peamiseks meetodiks olnud logistiline regressioonanalüüs, mis võimaldab modelleerida, kuidas iga vaadeldav tegur mõjutab uuritava muutuja variandi valikut (Paolillo 2002: 15). Koos arvutiajastuga loodi selle hõlpsamaks kasutamiseks juba 1970-ndatel, ammu enne digihumanitaaria mõiste sündi ja kvantitatiivset pööret keeleteaduses, spetsiaalne abiprogramm VARBRUL (Cedergren, Sankoff 1974) ja hiljem selle edasiarendatud variandid, nt Goldvarb X (2005) ja Goldvarb Z (2018). Programm oli loodud spetsiaalselt keeleandmete analüüsiks ja võimaldas vajadusel andmeid mugavalt ümber kodeerida, rühmitada jne ning rakendada logistilist regressiooni, mis paljudes muudes statistikaprogrammides polnud sel ajal kättesaadav.

Selle sajandi alguses toimunud kvantitatiivne pööre keeleteaduses (vt nt Kortmann 2021; Veismann jt 2018) on oluliselt muutnud keeleteaduse olemust. Tekkinud on uued kasutuspõhise keeleteaduse (ingl *usage-based linguistics*, vt nt Diessel 2017) suunad, mis seovad keeles toimuvaid protsesse ja muutusi keeleüksuste ja kontekstide esinemissagedusega. Selle taga on arusaam, et keel ei ole oma olemuselt kategoriaalne ning keele kirjeldamiseks sobivad seetõttu pigem statistilised tendentsid (tõenäosused) (vt nt Bod jt 2003). Selle käigus on oluliselt suurenenud ka keele varieerumise uurimine kvantitatiivsete meetodite toel. Laienenud on uurimissuunad: räägitakse näiteks korpuslingvistilisest varieerumise uurimisest (Szmrecsanyi 2017), mis on kasutusel kognitiivses ja laiemalt kasutuspõhises lingvistikas (vt nt Klavan, Divjak 2016; Klavan 2018). Tekkinud on uurimissuund, mida nimetatakse tõenäosuslikuks grammatikaks ning mis vaatleb keeleliste joonte varieerumist ning

väidab, et kõikidel kõnelejatel on oma tõenäosuslik keelemudel, mida pidevalt vastavalt keelelisele sisendile kohandatakse (Bresnan 2007; Bresnan, Ford 2010; Grafmiller jt 2018). Kõigi nende suundade ühisosa on keelelise varieerumise uurimine ja statistiline mudeldamine, ent vastuseid küsimusele, miks mingi nähtus varieerub, ei otsita enam niivõrd sotsiaalsetest faktoritest, kuivõrd keelelisest kontekstist, teksti tüübist, inimese kognitiivsetest omadustest (nt mäluga seotud piirangutest) jne. Ka meetodite pagas on oluliselt laienenud (vt nt Tagliamonte, Baayen 2012): tavaliseks on saanud otsustuspuud, juhumetsad, segamudelid jm mitmemõõtmelise statistika meetodid, mille kasutamiseks on mitmeid vabavaralisi ja üldisema funktsionaalsusega tarkvarasid (nt programmeerimiskeelte R ja Python paketid). Koos meetodi laienemisega on toimunud ka nihe terminoloogias. Sotsiolingvistilises traditsioonis kasutatud *muutuja* või *variaabli* asemel kasutatakse üha enam statistikast üle võetud termineid *uuritav* (*sõltuv*) *tunnus* (inglise keeles ka *response*) ning varieerumist seletavate tegurite jaoks terminit *seletav* (*sõltumatu*) *tunnus* (inglise keeles ka *predictor*), millel on omakorda väärtused, tasemed või klassid. Uuritava ja seletava tunnuse termineid kasutame ka siinse artikli metoodika kirjeldamisel ja tulemuste esitamisel, et viidata vastavalt analüüsis operatsionaliseeritud muutujale ja selle varieerumist mõjutavatele teguritele.

Kui võrrelda varieerumise analüüsi sotsiolingvistilist traditsiooni ja hilisemaid korpuslingvistikas levinud praktikaid, siis võib leida neis nii sarnasusi kui ka erinevusi. Benedikt Szmrecsanyi (2017: 687) on välja toonud üldprintsiibid, mida järgides võib korpuslingvistilisi uurimusi pidada varieerumise analüüsi suuna esindajateks: varieerumisena käsitletakse eri viise sama asja väljendamiseks; defineeritakse muutujad ja nende variandid; tähelepanu keskmes on keeleliste valikute tegemise protsess, mitte niivõrd sagedus tekstis; kasutatakse kvantitatiivseid meetodeid ja statistilist mudeldamist, et varieerumist tingivate tegurite mõju mõõta. Kahe suuna oluliste erinevustena on seevastu välja toodud, et korpuslingvistikas on kasutusel palju mitmekesisem materjalide valik (erinevad piirkonnad, registrid, ajaperioodid jne); sotsiolingvistilises traditsioonis on esikohal häälduse uurimine, korpuslingvistikas on aga võimalik varieerumist palju mitmekesisemalt uurida (leksikas, grammatikas, konstruktsioonides); olemasolevaid korpusi kasutades on sotsiolingvistilised tegurid sageli raskesti kättesaadavad ning seetõttu saavad need korpuslingvistikas vähe tähelepanu. Erinevused on ka kvantitatiivsete meetodite rakendamisel – sotsiolingvistikas domineerib logistiline regressioonanalüüs, korpuslingvistikas on meetodeid oluliselt rohkem (Szmrecsanyi 2017: 688–690).

Eestis on varieerumise analüüsi kasutatud meetodina alates 1990-ndatest. Eesti murretega seoses on sotsiolingvistilist varieerumise analüüsi ja VARBRUL-i kasutatud näiteks Karksi sõnalõpulise [e] häälduse (Pajusalu 1996), võru inessiivi lõpu (*-h* ~ *-n*) vaheldumise (Pajusalu jt 1999), eesti murrete ainsuse 1. isiku pronoomeni väljajätu (Lindström jt 2009) ning võru kõnekeele varieerumiste ja muutuste uurimisel (Mets 2010). Kirjutatud keele põhjal on varieerumise analüüsi meetodiga uuritud näiteks sama või sarnase tähendusega konstruktsioonide valikut (*laua peal* või *laual*, Klavan 2012, 2021; Klavan, Veismann 2017), adessiivis või allatiivis kogeja ja selle väljajätu varieerumist (Lindström, Vihman 2017), demonstratiivpronoomenite

ja -adverbide vaheldust nimisõna täiendina (Reile jt 2020; Hint jt 2021; Taremaa jt 2021), osastava ja sisseütleva käände vormide varieerumist (Siiman 2019) jne.

Murdekorpuse põhjal on võimalik varieerumise analüüsi meetodit edukalt kasutada (nt Klavan jt 2015; Lindström, Uiboaed 2017; Lindström jt 2018; Lindström jt 2021), ent enamasti mitte päris sotsiolingvistilises võtmes, sest info kõnelejate sotsiaalsete tunnuste kohta pole sageli piisav. Samuti on murdekorpuse (ja klassikalise murdeuurimise) keelejuhid üldjuhul valitud sarnastel alustel: vanemas eas kohaliku keele kõnelejad, kes on elanud pikalt samas piirkonnas, elu jooksul vähe liikunud ning kelle haridustee on olnud pigem napp (vt ka Lindström 2001). Kombineerituna uuemate andmetega on see teisalt hea võimalus modelleerida muutust või võrrelda eri piirkondade keelt. Uute andmete lisamine võimaldab keelemuutuste lähtekohana tuua enam esile sotsiolingvistilisi tegureid. Käesoleva artikli aluseks ongi nii vanemad, Tartu Ülikooli eesti murrete korpuse andmed (EMK)[1], kus sotsiolingvistiline info on kohati puudulik, ning uuemad, 2010. aastatel lindistatud seto korpuse andmed (SetKo), milles kõnelejate taust on paremini dokumenteeritud ning mis võimaldavad seega eitussõna asukoha varieerumise seletamisel esile tuua ka kõnelejate sotsiolingvistilisest taustast tingitud erinevusi.

2. Eitusmustrid seto ja võru keeles

Seto ja võru on lõunaeesti keelevariandid. Eesti traditsioonilises murdekäsitluses on keelekuju, mida setod kõnelevad, peetud eesti keele lõunaeesti peamurde Võru murde alla kuuluvaks Setu murrakuks (vt nt Viitso 1998: 115; Keem, Käsi 2002: 21; Pajusalu jt 2002: 52–57; Viikberg 2020) ning on väidetud, et võru ja seto peamine erinevus on mõnes hääldusliku ja sõnavaralises joones, mis on tingitud peamiselt tugevast vene mõjust (Keem, Käsi 2002: 23; Pajusalu jt 2002: 188–189). Keelelistest erinevustest olulisemateks on peetud religioosseid ning kultuurilisi erinevusi: setod on traditsiooniliselt õigeusklikud ning võrukesed luterlased ning neil on erinev identiteet (vt ka Paas 1927: 8–9; Jääts 2000: 651). Jüri Viikberg (2020: 23) on pidanud Setu murde eristamist Võru murdest, nagu seda tehakse Tartu Ülikooli eesti murrete korpuses (vt Lindström 2015), peamiselt sotsiolingvistilistel põhjustel tehtud otsuseks. Siiski on uurimused näidanud, et vähemalt võru ja seto (morfo)süntaktiliste konstruktsioonide kasutussageduses on olulisi erinevusi (vt nt Plado 2015: 199–200; Lindström, Uiboaed 2017: 331–333; Lindström jt 2019: 170–174; seto keele süntaktiliste omapärade kohta vt Lindström, ilmumas).

Kui traditsiooniline murdeuurimine käsitleb mõlemat keelevarianti murdena või lausa murrakuna (seto puhul), siis keeleajalooliselt on viimastel aegadel kõneldud siiski lõunaeesti keelest (vt nt Kallio 2007, 2014), millest tänapäeval on kõige elujõulisemad võru ja seto keel. Ka kõnelejad ise eelistavad kõnelda keelest, mitte murdest (Eichenbaum, Pajusalu 2001: 484). Käesolevas käsitluses kasutatakse lihtsuse huvides mõlema keelevariandi kohta terminit *keel*, diskuteerimata murde ja

[1] Kasutajaliides on kättesaadav aadressil http://www.murre.ut.ee/mkweb (2. VII 2021).

keele vaheliste suhete üle. Samal põhjusel kasutame artiklis väikest algustähte (*seto*, *võru*).

Käesolevas artiklis rakendatakse sagedusandmeid ning varieerumise analüüsi võimaldavaid statistilisi meetodeid ees- ja tagaeituse leviku ning kasutustingimuste väljaselgitamiseks. Eeseituses paikneb eitussõna põhiverbi ees (*ei olõq*), tagaeituses põhiverbi järel (*olõ-õi(q)*). Tagaeitus on võru-seto keelealale iseloomulik ja väljapaistev morfosüntaktiline joon (Iva 2002: 582–587, 2007: 102) ning eesti kirjakeelest selgelt eristuva joonena on see võetud kasutusele ka võru kirjakeeles (näide 1).

(1) *Ma* ***tiiä-ei*** *külh üttegi umaiälist inemist, kiä* ***olõ-õi*** *koolin üten klassiga tiatrin käünü*. (Uma Leht 25. III 2021)

Keeleajalooliselt on alguurali keelde rekonstrueeritud eitusverb, millele lisandusid aja-, kõneviisi- ja isikutunnused (Janhunen 1982). Eitusverb on siiani olemas paljudes tänapäeval Uurali keelkonda kuuluvates keeltes (vt Miestamo jt 2015: 17). Läänemeresoome keeltes on eituse moodustamisel toimunud hulgaliselt muutusi, ent enamikus on eitussõnal säilinud mingi verbi omadus (Miestamo jt 2015: 17), kõige tüüpilisemalt isikutunnus (nt sm *en lue* 'ma ei loe', *et lue* 'sa ei loe' jne). Põhjaeestil põhinevas eesti kirjakeeles on eitussõnal kadunud kõik verbi omadused ning vana eituskonstruktsiooni meenutab vaid spetsiaalne konnegatiivne verbivorm olevikus (*ma/sa/ta ei* **loe**), mineviku moodustamiseks kasutatakse mineviku kesksõna (*ma ei* **lugenud**). Lõunaeestis (sh võrus ja setos) on aga eitussõnal säilinud aja markeerimine: olevikus toimib eitussõnana *ei* ja mineviku eitussõnana *es*, põhiverb on spetsiaalses konnegatiivses vormis nii olevikus kui ka minevikus. Samuti võib minevikuvormides esineda nii ees- kui ka tagaeitus (vastavalt 2a ja 2b).

(2a) *ma/sa/tä/mi/ti/nä* ***es süüq***
(2b) *ma/sa/tä/mi/ti/nä* ***süü es*** ~ ***süü-üs*** ~ ***süü-s***

Eesti (ja kogu uurali) keelealale on üldiselt iseloomulik eeseitus. Tagaeitus on ka maailma keeltes oluliselt haruldasem nähtus kui eeseitus. Eeseitust esineb u 52%-l maailma keeltest, tagaeitust aga u 28%-l (Dryer 2013). Eitussõna tendents paikneda verbi ees on keeletüpoloogias tuntud printsiip, millel arvatakse olevat kognitiivne/psühholoogiline põhjus: kuna eitussõna muudab radikaalselt sõnumi sisu, annab selle paiknemine enne verbi kuulajale selgema aluse järgneva sõnumi tõlgendamiseks, ning aitab vältida segadust sõnumi mõistmisel (Krasnoukhova jt 2021).

Seto-võru keelealal on tagaeitust seostatud enam idapoolse võru keelega, mille hulka on traditsiooniliselt loetud ka seto keelt (Lindström 1997: 144; Keem, Käsi 2002: 53; Iva 2007: 103). Karl Pajusalu jt (2002: 189) ning Paul Hagu (Hagu, Pajusalu 2020: 19) peavad seto keelele omaseks just tagaeitust. Sama on näidanud ka Liina Lindströmi jt (2021) statistiline uurimus idaseto keelepruugi kohta. Läänevõru keeles esineb tagaeitus aga pigem harva.

Võru ja seto keeles moodustatakse standardeitus eitussõna *ei* (olevikus) või *es* (minevikus) ja nn konnegatiivse verbivormi abil. Eeseituse konnegatiivne verbi-

vorm koosneb verbitüvest ja sellele liituvast larüngaalklusiilist (3), mis võib fonoloogilistel põhjustel jääda välja hääldamata või mis võib kokku sulada järgneva sõna alguskonsonandiga (larüngaalklusiili hääldamise kohta võro keeles vt nt Iva 2013: 106–107). Tagaeituse konnegatiivne verbivorm on verbitüvi (4). Võru ja seto keeles on eitusvorm kõikides isikutes sama. Tagaeituse puhul võib eitussõna olla kas verbist eraldiseisev (4a) või sulada eelneva verbiga kokku (4b–4c).

(3) *ma/sa/tä/mi/ti/nä* ***ei süüq***
(4a) *ma/sa/tä/mi/ti/nä* ***süü eiq***
(4b) *ma/sa/tä/mi/ti/nä* ***süü-üiq***
(4c) *ma/sa/tä/mi/ti/nä* ***süü-iq***

Peale ees- ja tagaeituse võib nii võru kui ka seto keeles kasutada topelteitust, kus esimene eitussõna paikneb vahetult kas konnegatiivse verbivormi ees (5a, 5b) või järel (5c) ja teine eitussõna eitusfookuse lõpus (Lindström 1997: 150–153). Seto materjalist leidsime ka mõne topelteituse kasutusjuhu, kus mõlemad eitussõnad paiknevad enne verbi (5d), mille puhul pole selge, kas tegemist on suulise keele töötlemise eripärast tingitud erandlike kasutusjuhtudega või süsteemse kasutusviisiga.

(5a) ***eij*** = *olõ* = *ma säänest tennüq* ***ei*** (Vas)
(5b) *ma=****s*** *soovigi* ***ess*** *avaldaq* (ISe)
(5c) *Vinnemaal üldse olõ-****i*** *säänest* ***eiq*** (ISe)
(5d) *`näime hobõst* = ***õss*** *`kätte* = ***ss*** *saaq* (PSe)

Ka imperatiivis võib kasutada nii ees-, taga- kui ka topelteitust, kuid ees- ja tagaeituses kasutatakse erinevat eitussõna: eeseituses eitussõna *är(q) ~ ar(q)* (***är*** *võtkuq*) ja tagaeituses eitussõna *ei(q)*, mis võib eelneva verbiga kokku sulada (*võtku* ***eiq*** *~ võtku-****uiq*** *~ võtku-****iq***). Nii ees- kui ka tagaeituses on verbil sufiks *-gu(q)/-ku(q)*. Topelteituses kasutatakse mõlemat eitussõna (***är*** *võtku-****iq***). Imperatiivis on topelteituse kasutamine harv, vähemalt idasetos (Lindström jt 2021).

Varasemas artiklis, mille fookuses oli ees- ja tagaeituse varieerumine idaseto alal (Lindström jt 2021), leidsime mõningate kihelkondade andmetele tuginedes, et tagaeituse kasutussagedus langeb järsult võru-seto keeleala piiril. Siinses artiklis täpsustame ees- ja tagaeituse vahekorda seto ja võru alal ning modelleerime võimalikku varieerumist. Samuti laiendame vaadeldavat geograafilist piirkonda kogu seto-võru keelealale ning võrdleme eituskonstruktsiooni valikut tingivaid tegureid eri piirkondades.

3. Materjal

Uurimuses võrdleme nii vanemaid korpusandmeid, milles on info kõnelejate sotsiolingvistilise tausta kohta lünklik, kui ka uuemaid andmeid, mis võimaldavad analüüsi tõlgendamisel arvestada ka sotsiaalsete faktoritega.

Vanem seto materjal (põhjaseto, lõunaseto, idaseto) on pärit eesti murrete korpusest (EMK). Seto alalt pärit intervjuud on lindistatud aastatel 1964–1996. Eitusvorme otsisime morfoloogiliselt märgendatud tekstidest. Põhjaseto materjalist kaasasime analüüsi 487, lõunasetost 707 ning idasetost 345 eituse kasutusjuhtu.

Uuem seto materjal on pärit ainult idaseto alalt, mis jääb praeguse Vene Föderatsiooni Petseri rajooni territooriumile. Materjalid on kogutud välitööde käigus aastatel 2010–2013 ning neid hoitakse Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis (Murdearhiiv). Need on ka loodava seto korpuse (SetKo) osaks. Intervjuud viidi läbi vabas vormis ning need toimusid seto keeles. Analüüsis oleme kasutanud materjali kaheksalt kõnelejalt. Uurimuse andmestiku korjasime intervjuude transkriptsioonidest käsitsi, kokku kaasasime analüüsi 1083 eituse kasutusjuhtu.

Võru materjal on samuti pärit EMK-st ning on salvestatud ajavahemikus 1960–1993. Eitusvorme otsisime morfoloogiliselt märgendatud tekstidest. Uurimuse andmestikus on kaheksa kihelkonna andmed: Rõuge (280 eituse kasutusjuhtu), Vastseliina (438), Urvaste (41), Räpina (193), Põlva (589), Hargla (382), Karula (574) ja Kanepi (255). Eitusjuhtude arv murrakuti erineb sõltuvalt sellest, kui palju teksti on igast piirkonnast EMK-sse kaasatud.

4. Meetod

Nagu eelnevalt selgitatud, on käesoleval juhtumiuuringul kaks põhilist eesmärki. Esiteks huvitab meid see, kuidas on tüpoloogiliselt haruldane tagaeitus võru-seto keelealal levinud. Selleks võrdleme ees- ja tagaeituse ning topelteituse suhtelisi kasutussagedusi korpuste materjalides ning visualiseerime tagaeituse levikut sageduskaardil, et saada parem ülevaade sellest, kas üleminekud eeseituse kasutamiselt tagaeituse eelistamisele on pigem järsud või sujuvad.

Teiseks keskendume nende piirkondade keelele, kus esineb arvestataval määral ees- ja tagaeituse varieerumist, ning selgitame välja, millised tegurid sel juhul eitussõna paiknemist enim mõjutavad. Kuivõrd võru ja seto andmed erinevad oluliselt ees- ja tagaeituse proportsioonide poolest, huvitab meid, kas varieerumist mõjutavad tegurid neis kahes keelevariandis toimivad sarnaselt või erinevalt. Analüüsi kaasatud tegurite ehk seletavate tunnuste nimekiri on esitatud alapeatükis 5.2.1.

Kasutame logistilise regressiooni segamudeleid (vt nt Pinheiro, Bates 2000; Baayen 2008; Baayen jt 2008; Gries 2015, 2021; Winter 2020), et kontrollida, milline roll on eri teguritel võru ja seto keelevariantides. Regressioonimudel püüab ennustada uuritava nähtuse ehk ees- ja tagaeituse kasutamise tõenäosusi matemaatilise lineaarfunktsiooni kaudu, mis võtab korraga arvesse kõikide seletavate tunnuste kombineeritud mõju, ent lubab samal ajal hinnata iga tunnuse muutumise mõju eraldi, hoides teiste seletavate tunnuste mõju samal ajal fikseerituna. See võimaldab pea katselaadsetes tingimustes testida iga üksiku tunnuse olulisust ja mõju suunda muutuja variantide ehk uuritava tunnuse väärtuste ennustamisel, kui tegelike katseandmete asemel on kasutada hoopis korpuste vaatlusandmed (Hinrichs, Szmrecsanyi 2007: 459). Näiteks saame regressioonimudeli abil testida, kas idasetos ees-

eituse kasutamise tõenäosus kasvab või kahaneb, kui vaatleme oleviku eitussõna *ei* asemel hoopis mineviku eitussõna *es*, ent hoiame ülejäänud konteksti (nt grammatilise isiku ja eelmise eituse vormi) samana. Regressioonimudeli eesmärk on niisiis tuvastada suure hulga tegeliku keelekasutuse näidete põhjal kontekste, milles ees- või tagaeituse valiku tõenäosused oluliselt erinevad.

Segamudel on regressioonimudeli edasiarendus, mis võimaldab ees- ja tagaeituse valikut tingivate tunnuste mõju (ehk fikseeritud mõjude) hindamisel arvestada ka nn juhuslike mõjudega. Juhuslikeks mõjudeks nimetatakse selliseid tunnuseid, mille üksikud väärtused valimis esindavad juhuslikku valikut suuremast populatsioonist. Näiteks meie korpusvalimites esindatud kõnelejad on juhuslik hulk kõikidest seto ja võru keele kõnelejatest ning eitatud tegusõnad on (võrdlemisi) juhuslik hulk kõikidest võimalikest tegusõnadest, mida lauses võiks eitada. Seevastu näiteks eituskonstruktsiooni eitussõna leksikaalse kuju väärtused valimist ei sõltu: ei ole tõenäoline, et leiaksime mõnest teisest juhuslikust valimist peale *ei* ja *es* eitussõna ka mingeid teisi lekseeme. Seetõttu saame selle mõju kontrollida erinevates valimites ühtmoodi. Juhuslike mõjude kaasamine võimaldab varieerumise seletamisel kaudselt arvestada aga teguritega, mille täpset olemust on raske määrata või operatsionaliseerida, ning samal ajal leevendada statistiliste mudelite olulist eeldust vaatluste ehk siinses uurimuses üksikute korpustes esinevate eitusjuhtude sõltumatuse kohta.

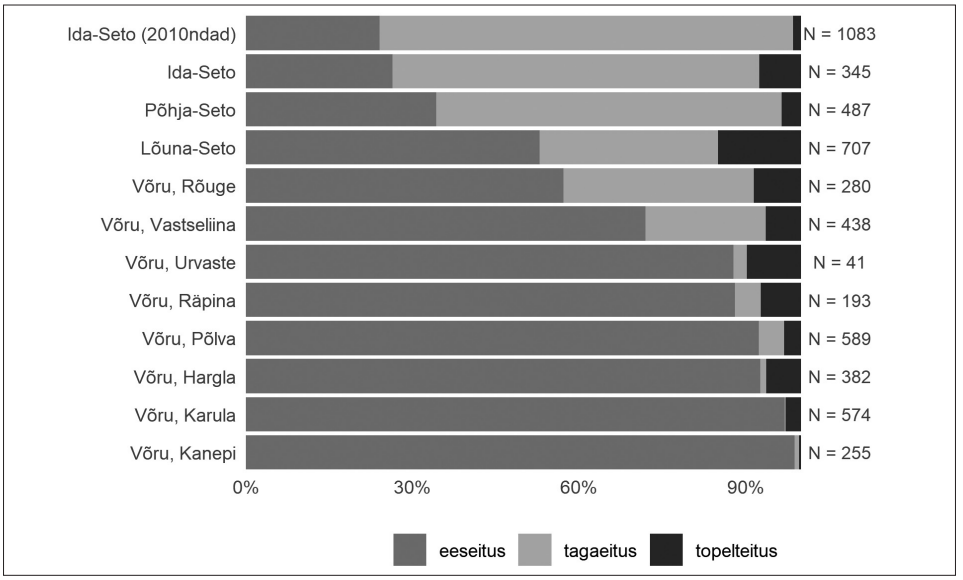
Materjali analüüsisime rakendustarkvaraga R (R Core Team 2020). Mudelite tegemiseks kasutasime R-i paketti *lme4* (vt Bates jt 2015), tulemuste visualiseerimiseks pakette sjPlot, ggeffects (vt Lüdecke 2018) ja ggplot2 (vt Wickham 2016), sageduskaardi jaoks ka ruumiandmete analüüsi paketti *sf* (vt Pebesma 2018). Kasutatud andmestikud ja analüüsi skriptid on avalikult kättesaadavad Open Science Frameworki (OSF) platvormil (Pilvik jt 2021).

5. Tulemused

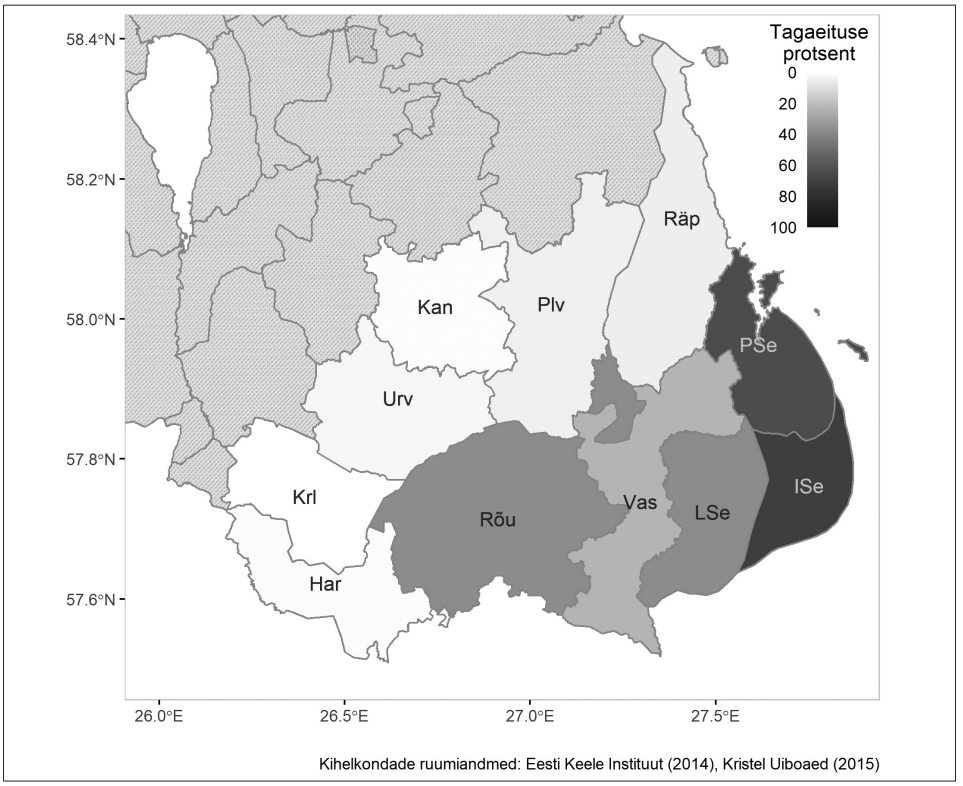
5.1. Tagaeituse kasutamine võru-seto keelealal: sagedusandmed

Esmalt vaatleme suhteliste kasutussageduste põhjal, mis on ees-, taga- ja topelteituse vahekorrad võru-seto keelealal (joonis 1). Tagaeituse osakaalu illustreerib ka sageduskaart (joonis 2), mis sisaldab vaid EMK andmeid, st välja on jäetud idaseto 2010-ndate andmestik.

Nii joonis 1 kui ka joonis 2 osutavad selgelt, et tagaeitus on ennekõike seto keelele iseloomulik joon, mida esineb vanemas materjalis, st EMK põhjal enim idasetos (66%) ning põhjasetos (62%), kusjuures idasetos on tagaeituse osakaal aja jooksul koguni mõnevõrra kasvanud (SetKo 2010-ndate andmestikus 74%). Tagaeituse kasutamine väheneb idast läände liikudes hüppeliselt, olles marginaalne Karula (0,2%), Kanepi (0,8%), Hargla murrakutes (1%) ning väga väike ka Urvastes (2,4%), Põlvas (4,6%) ja Räpinas (4,7%). Üleminekuala seto ja võru vahel moodustavad lõunaseto ala (32%), Vastseliina (22%) ning Rõuge (34%). Seega võib kinnitada, et tagaeitus on iseloomulik ennekõike seto keelele ja teatud määral levi-



Joonis 1. Ees-, taga- ja topelteituse osakaalud võru ja seto keelealadel (eitusjuhtude koguarv *N* on esitatud tulpadest paremal). Andmed EMK-st, idaseto 2010-ndate andmed SetKo-st.



Joonis 2. Tagaeituse leviku osakaal protsentides seto-võru keelealale jäävates kihelkondades. Andmed EMK-st.

nud ka Võrumaa kaguosas. Üllatav on tagaeituse üsna kõrge kasutussagedus Rõuges, mis ei piirne (vähemalt tänapäeval) otseselt seto alaga. Sarnasel põhjusel üllatab ka tagaeituse madal kasutussagedus Räpinas, mis külgneb põhjaseto kõnelemise alaga.

Joonise 1 põhjal on võimalik vaadelda ka topelteituse kasutussagedust kogu keelealal. Selle kõrgeim kasutussagedus on lõunasetos (15%), millele järgnevad Urvaste (9,8%), Rõuge (8,6%), idaseto (vanemad andmed, 7,5%), Räpina (7,3%), Vastseliina (6,4%), Hargla (6,3%). Idasetos on topelteituse juhtumeid tänapäeva andmetes (SetKo) oluliselt vähem (2%) ning nende kasutus on vähem süstemaatiline kui EMK andmetes (vt täpsemalt Lindström jt 2021).

5.2. Ees- ja tagaeituse varieerumise analüüs

Sagedusandmete põhjal valisime edasiseks analüüsiks välja kaks ala: seto keelt esindavad idaseto 2010-ndate andmed (mida analüüsisime ka artiklis Lindström jt 2021) ning võru keelt Vastseliina ja Rõuge EMK andmed, mille aluseks on põhiliselt 1960.–1970. aastatel tehtud lindistused. Idaseto uuema materjali kohta on olemas ka põhjalikud sotsiolingvistilised andmed, mis toetavad tulemuste tõlgendamist. Valisime Rõuge ja Vastseliina, kuna neis domineerib eeseitus, ent varieerumist esineb piisaval määral (vt joonist 1), et selle analüüs oleks mõttekas ning samade analüüsivahenditega ka võimalik.

5.2.1. Tunnused

Analüüsis seletame eitussõna asendit (ees- või tagaeitus) viie tunnuse kaudu, mis kõik iseloomustavad üht aspekti kontekstist, kus ees- või tagaeitus esineb. Seletama oleme valinud põhiliselt grammatilised tunnused: EITUSSÕNA leksikaalne kuju väljendab eituse ajavormi (olevikus *ei*, minevikus *es*); TOPELTEITUS väljendab seda, kas tegemist on topelteitusega või mitte; SUBJEKT näitab seda, millises vormis on lause alus ehk see, kelle tegevust või kogemist eitatakse; ISIK väljendab seda, milline on kõnealuse olendi või asja osalus kõnesituatsioonis (kas *mina/meie*, *sina/teie*, *tema/nemad* või kasutatakse hoopis umbisikulist tegumoodi). Peale grammatiliste tunnuste hindame ka ühe mäluga seotud nähtuse, nn tunnetusliku või strukturaalse praimimise mõju eitussõna paiknemisele (vt Szmrecsanyi 2005; Torres Cacoullos, Travis 2019: 674; eesti keeles nt Tulving 2002). Praimimine tähendab tendentsi korrata suhtluses aktiveeritud struktuure ning siinses uurimuses operatsionaliseerime seda kõnehetkele viimati eelnenud eituskonstruktsiooni sõnajärje kaudu (EELMINE). Nagu eespool öeldud, arvestame mudelites ühtlasi sellega, et kõnelejad võivad oma individuaalse keeletunnetuse ja -kogemuse tõttu kontekstist sõltumata kalduda kasutama pigem ees- või tagaeitust (KÕNELEJA) ning et mingite konkreetsete tegusõnade korral võidakse eituskonstruktsioonis eelistada pigem üht või teist sõnajärge (LEMMA). Analüüsi kaasatud seletavad tunnused on esitatud tabelis 1.

Esialgses seletavate tunnustega kodeeritud uuemas idaseto andmestikus oli 1083 vaatlust ning võru andmestikus 718 vaatlust. Jätsime andmestike puhastamise käigus välja juhud, kus aluse ehk subjekti vormi või isikut ei olnud võimalik määrata,

Tabel 1. Logistilise regressiooni segamudelisse kaasatud eitussõna asukoha varieerumist seletavad tunnused.

Tunnus	Väärtused	Näited
EITUSSÕNA	*ei* *es*	*ma ütle=**i** midägi* *tä ütle=**ess** midägi mullõ*
EELMINE	*ees* *taga*	*ja härrä **iss** tii es's' `mitte midägi* [*es arvõsta*] *`üttegi midägi* *kuq `müildü = **üss** siss (.)* [*anda =es*] *tap- `viinagiq*
TOPELTEITUS	*ei* *jah*	*millegiberäst **ei** olõ näid* ***ei** olõ vist säänest **eiq***
SUBJEKT	*ei* *muu* *nom* *prt*	*tiijä=ei tuud joht* *miil=ei olõq (...) **muijalõ tettä**=ku `taahha* *pühä+bäävä es=tii **pere+miiss** ka es's' tüüd* *puu+`luidsaga `süüd'i (.) es=olek=**`kahvlit***
ISIK	*1* *2* *3* *imps*	*muud **ma**=ei mõistaq tuust pulma as'ast selettäq* *egä **tii**=ei taha vast nüüd tsuidsutta* ***ron'giq** es käüq viil* *ja nüüd=ei **pandak'k'i** noid `her'n'it inämp*
KÕNELEJA	$N_{võru}$ = 12 $N_{idaseto}$ = 8	
LEMMA	$N_{võru}$ = 17 $N_{idaseto}$ = 22	

ning laused, mis asusid lindistuste algul või pärast katkestusi ning seega polnud võimalik tuvastada vahetult eelneva eituse vormi. Muudel juhtudel märgendasime viimati kasutatud eituskonstruktsiooni sõnajärje, olenemata sellest, kui kaugel see kõnehetke suhtes paiknes. Samuti otsustasime jätta välja käskiva kõneviisi juhud (nt *ära `jäl'g'i `jätku `perrä*, *küsügü=üi*), mida esines väga harva.

Artiklis Lindström jt (2021) mudeldasime sama idaseto andmestikku, kaasates eraldi seletava semantilise tunnusena ka verbitüübi (nt kas tegusõna on olemisverb, kognitsiooniverb, liikumisverb vm), ent ei kaasanud tegusõna ennast juhusliku mõjuna. Leidsime, et eitussõna (*ei*, *es*) ning verbitüübi vaheline koosmõju aitas eitussõna asendit küllalt hästi ennustada: kognitsiooniverbidega oli eeseituse esinemine oluliselt tõenäolisem, kui eitussõna oli *ei* ja ajavorm seega olevik (nt *ei tiijäq*, *ei mälehtä*). Selle põhjustena nägime kognitsiooniverbide suurt rolli parajasti käimasoleva diskursuse vahendajatena ning neid sisaldavate eituskonstruktsioonide sagedamat kinnistumist fikseeritud sõnajärjega kõneüksustena. Viimastel on kontaktsituatsioonis suurem ülevõtmise tõenäosus. Teisisõnu võivad tihedamad kontaktid eesti või vene keelega tingida ka eesti või vene keele pärase sõnajärjega eituskonstruktsioonide sagedama kasutamise. Selles artiklis toimime vastupidi, kaasates juhusliku mõjuna tegusõna lemma, ent jättes fikseeritud mõjuna välja verbitüübi, mis klassifitseeriks semantiliselt sama tüüpi, ent erineva esinemissagedusega tegusõnad samasse kategooriasse ning uuriks semantilise klassi kui terviku mõju eitussõna paiknemisele. Semantilise klassi asemel verbilemmade kasutamiseks pidime aga andmestikust välja jätma tegusõnad, mida esines väga vähe, vastasel juhul oleks mudelil konkreetse tegusõna ees- või tagaeituse lembust raske hinnata (vt Gries 2021). Seega eemaldasime mõlema keele vaatluste hulgast laused, milles eitatud tegusõna esines andmestikus vähem kui 5 korda. Idaseto andmestikku jäi

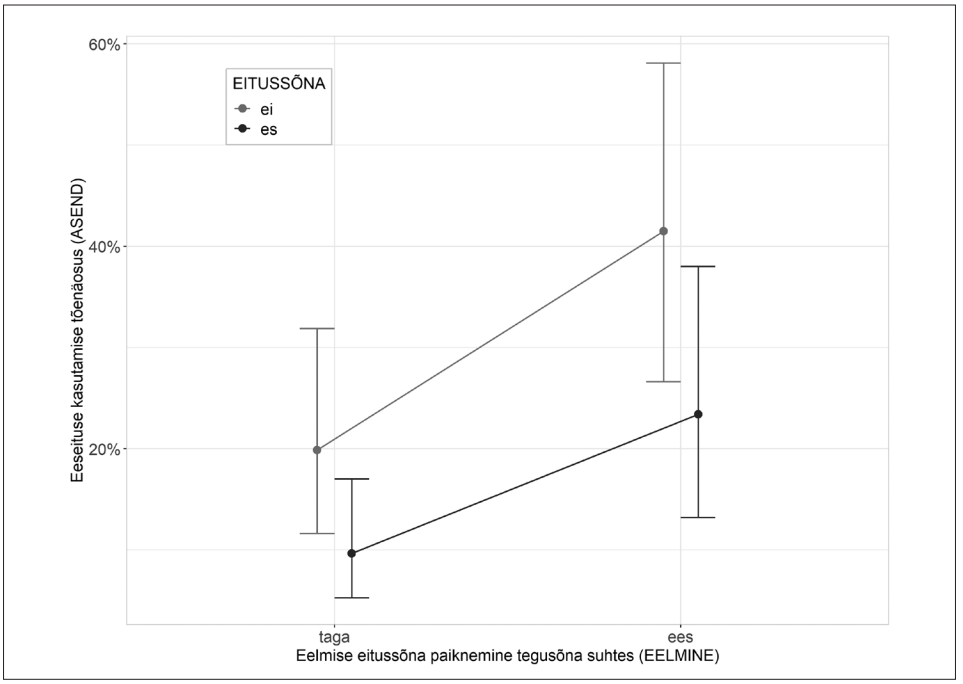
nõnda 922 vaatlust (nendest 25% ees- ja 75% tagaeitusega) ning võru andmestikku 569 vaatlust (nendest 70% ees- ja 30% tagaeitusega). Mõlemas andmestikus lasime mudelil ennustada vähem levinud eituskonstruktsiooni esinemise tõenäosust, st idaseto keeles eeseitust ja võru keeles tagaeitust.

5.2.2. Eitussõna paiknemist seletavad tunnused idaseto keeles

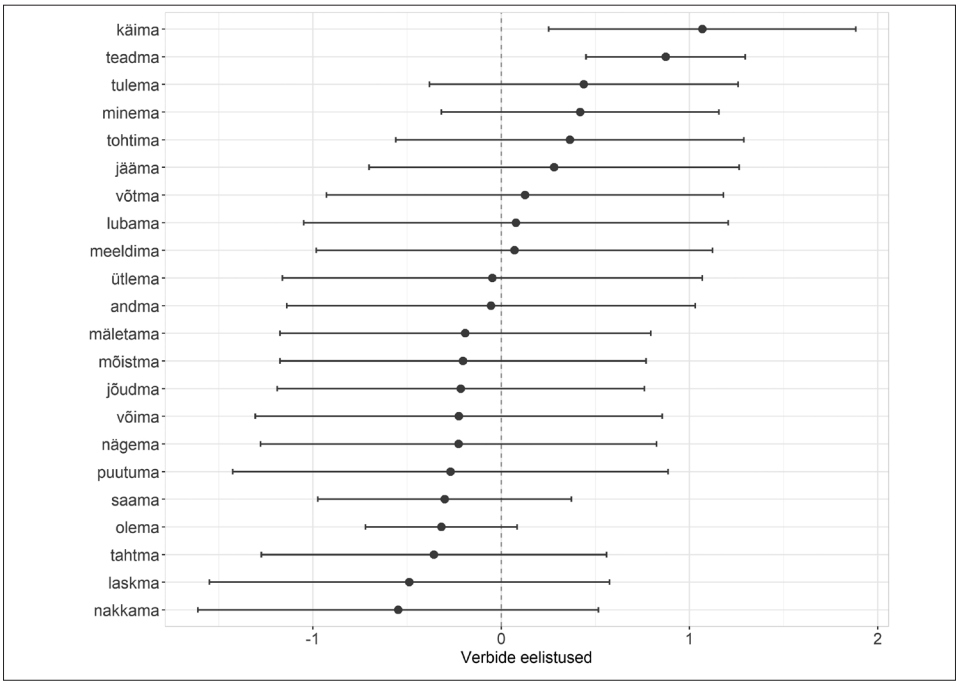
Selleks et selgitada välja, millised eeltoodud tunnustest eitussõna paiknemist tingivad, alustasime kõige lihtsamatest, üht seletavat tunnust sisaldavatest regressioonimudelitest ning liikusime järjest keerukamate kombinatsioonideni, lisades mudelisse ükshaaval seletavaid tunnuseid ning testides ka nendevahelisi koosmõjusid. Erinevate mudelite võrdlemiseks kasutasime tõepärasuhte testi (ingl *likelihood ratio test*, vt Pinheiro, Bates 2000: 83; Gries 2021). Idaseto andmestikus jõudsime optimaalse mudelini, kus eitussõna asendit seletab kõige paremini see, milline oli diskursuses viimati kasutatud eitussõna asend tegusõna suhtes, ning see, kas räägitakse olevikust või minevikust (oleviku eitussõnaga *ei* või mineviku eitussõnaga *es*). Lisaks arvestab mudel ka juhusliku kõneleja- ning verbispetsiifilise varieerumisega.[2] Muud, grammatilised tunnused oluliseks ei osutunud. Varieerumist seletavate tunnuste fikseeritud mõju kujutab joonis 3.

Eitussõna paiknemist mõjutavad tegurid idasetos on tihedalt seotud mälust, suhtlusolukorrast ja suhtlemise sagedusest tulenevate tendentsidega ning keelesisesed, grammatilised tegurid jäävad siinjuures tagaplaanile. Jooniselt 3 näeme, et üldiselt domineerib idasetos tagaeitus ning eeseituse kasutamise tõenäosus jääb igas vaadeldud kontekstis alla 50%. Kui kõnehetkele eelnenud eituskonstruktsioon (EELMINE), ükskõik kas öeldud kõneleja enda või intervjueerija poolt, on olnud samuti tagaeitusega, siis on tagaeituse uuesti kasutamise tõenäosus veelgi suurem ning eeseituse kasutamise tõenäosus seega väike (olenevalt eitussõnast kas 10% või 20%). Kui aga eelnev eitusvorm on olnud eeseitus, siis kasvab ka eeseituse uuesti kasutamise tõenäosus (olenevalt eitussõnast on see kas 23% või 42%). Oluline roll eitussõna asukoha ennustamisel on mudelis niisiis ka eitussõnal endal (*ei* või *es*). Eelneva eituskonstruktsiooni sõnajärjest olenemata on setole vähem omase eeseituse kasutamise tõenäosus suurem, kui eitussõnaks on *ei*. Seda võib seletada ühelt poolt sellega, et oleviku eitussõna moodustab sagedamini teatud verbidega suhtluses kinnistunud ühendeid (nt *ma=i tiijä*), mida kontaktsituatsioonis võidakse lähtekeele sõnajärjega üle võtta. Teisalt kasutataksegi teatud verbe suhtluses palju sagedamini olevikus kui teisi ning sel juhul peegelduvad eitussõna mõjus ka verbispetsiifilised eelistused. Verbide individuaalseid eelistusi (eitussõnast sõltumata) näitab joonis 4.

[2] Lõpliku, optimaalse mudeli kuju oli ASEND ~ EELMINE + EITUSSÕNA + (1|KÕNELEJA) + (1|LEMMA). Mudeli konkordantsiindeks *C* (vt Hosmer jt 2013: 173–182) oli 0,82, mis viitab mudeli heale võimele ees- ja tagaeituse kasutamise kontekste eristada (ideaalset eristusvõimet näitaks *C*-indeks väärtusega 1).



Joonis 3. Eeseituse kasutamise tõenäosus idaseto andmestikus vastavalt eelmise eitussõna paiknemisele ja eitussõna leksikaalsele kujule.



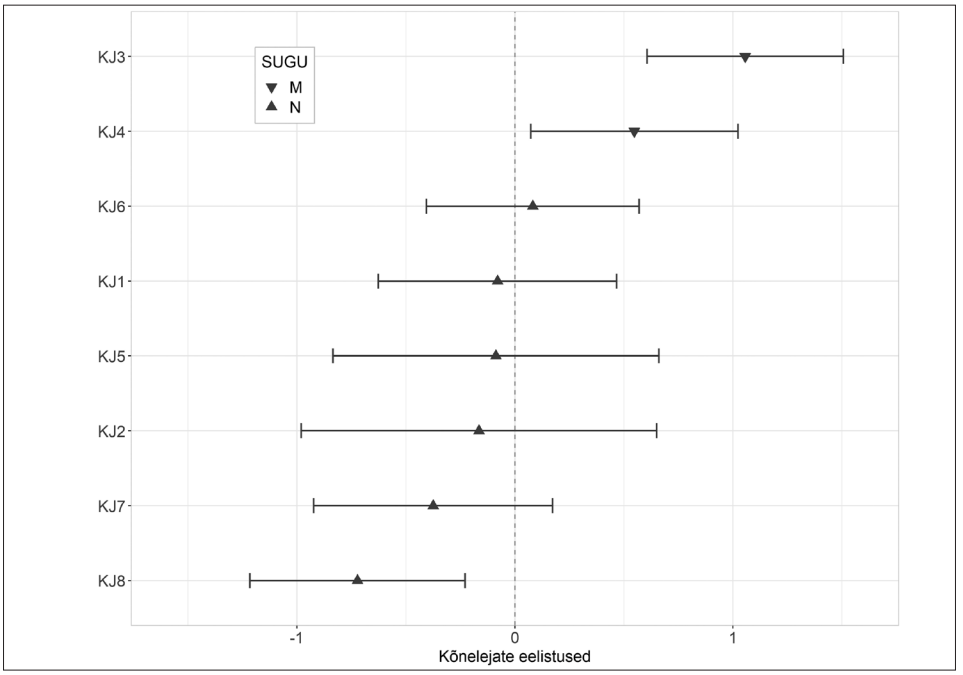
Joonis 4. Idaseto verbilemmade individuaalsed eelistused ees- ja tagaeituse kasutamisel.

Positiivsed väärtused joonisel 4 näitavad, et tegusõnal on tendents esineda eeseitusega keskmisest rohkem, negatiivsed väärtused jällegi, et tegusõna esineb keskmisest enam tagaeitusega.[3] Mida suurem on punkthinnangute 95% usaldusvahemik (joonisel 4 märgitud n-ö vurrudega), seda vähem kindlust mingi tegusõna eelistuse puhul on (üldjuhul seetõttu, et tegusõna esineb harva). Kui usaldusvahemik ei kata nulli, võib olla üsna kindel, et tegusõna käitub teistest erinevalt. Sellised tegusõnad on idasetos *käima* ja *teadma*, mis esinevad teistest tegusõnadest sagedamini eeseitusega (*ei/es käü*, *ei/es tiijäq*). Tegusõnad *nakkama*, *laskma*, *tahtma* ja *olema* aga kalduvad esinema keskmisest enam tagaeitusega (nt *olõ-õiq*, *taha-ass*), ehkki tegusõnade punkthinnangute usaldusvahemik on lai ja katab ka nulli ning seetõttu ei saa hinnangutes populatsiooni kohta väga kindel olla. Üksikuid tegusõnu eraldi vaadeldes näeme ka, et tegusõnad ei käitu eitussõna asukoha määramisel semantilise klassi poolest sugugi ühtselt. Näiteks kognitsiooniverbidel *teadma*, *meeldima*, *mäletama*, *mõistma*, *nägema* ja *tahtma* on väga erinevad eelistused ning kognitsiooniverbide klass tervikuna näib eelistavat eeseitust (vt Lindström jt 2021) pelgalt seetõttu, et eeseituselembene *teadma* on neist ülekaalukalt kõige sagedasem ja esineb eitussõnaga tõenäoliselt suhtluses võrdlemisi kinnistunud järjendis. Ühelt poolt võidakse suhtluses niisiis võtta kontaktkeeltest üle fikseeritud sõnajärjega ühendeid, teisalt aga võivad mingid sagedamad järjendid kinnistuda ka keelesiseselt (nt *es käü*, *olõ-õiq*), kusjuures siin võib olulist rolli mängida verbitüve morfofonoloogiline vorm.

Segamudel arvestab asjaoluga, et andmestikus esinevad eituse juhud võivad ka kõnelejati koonduda: kuigi kogukonna ühist, jagatud keelt võivad suunata mingid keelelised ja keelevälised tegurid, räägib iga kõneleja ikka pisut isemoodi. Nõnda võib mõni kõneleja kasutada rohkem eeseitust, teine tagaeitust. Neid eelistusi võib omakorda seostada näiteks kodukeele, meediatarbimise, hariduse, suhtlusringkonna jm sotsiaalsete teguritega. Seda eriti seto puhul, kus tihedad kontaktid vene keele kõnelejate, vene kultuuri ja religiooniga on võru keelega võrreldes põhiliseks fonoloogiliste, leksikaalsete ja ka süntaktiliste erinevuste allikaks (vt ptk 2). Samuti mõjutavad eriti Venemaa alal elavaid idaseto kõnelejaid üha kahanevad võimalused seto keelt igapäevaselt kasutada.

Kui vaatame idaseto kõnelejate eelistusi joonisel 5, siis näeme, et valimi kaks ainsat meessoost kõnelejat (KJ3 ja KJ4) kalduvad kasutama oma kõnes eeseitust teistest oluliselt enam. Siiski ei ole põhjust arvata, et meessoost kõnelejad üldiselt eelistavad eeseitust, vaid tõenäoliselt on tegemist sotsiaalsete tegurite mõjuga: neil on olnud elu jooksul tihedad kontaktid Eestiga (on töötanud Eestis või külastanud sageli Eestit) ning nad on abielus venelannadega (vt Lindström jt 2021). Võib seega oletada, et nende meeskõnelejate keeles on teiste keelte (vene, eesti) mõju.

[3] Kuna logistilise regressiooni mudelid ennustavad vaikimisi uuritava tunnuse ühe väärtuse esinemise (antud juhul eeseituse kasutamise) šansse logaritmitud skaalal, näitab *x*-telje skaala joonisel muutust šansi logaritmis. Kui väärtus on positiivne, siis suureneb konkreetse tegusõnaga eeseituse kasutamise logaritmitud šanss ning seega ka šanss ja tõenäosus. Kui väärtus on negatiivne, siis tuleb jällegi eeseituse kasutamise (logaritmitud) šansse ja tõenäosust kohandada tagaeituse kasuks.



Joonis 5. Idaseto kõnelejate ees- ja tagaeituse kasutamise eelistused.

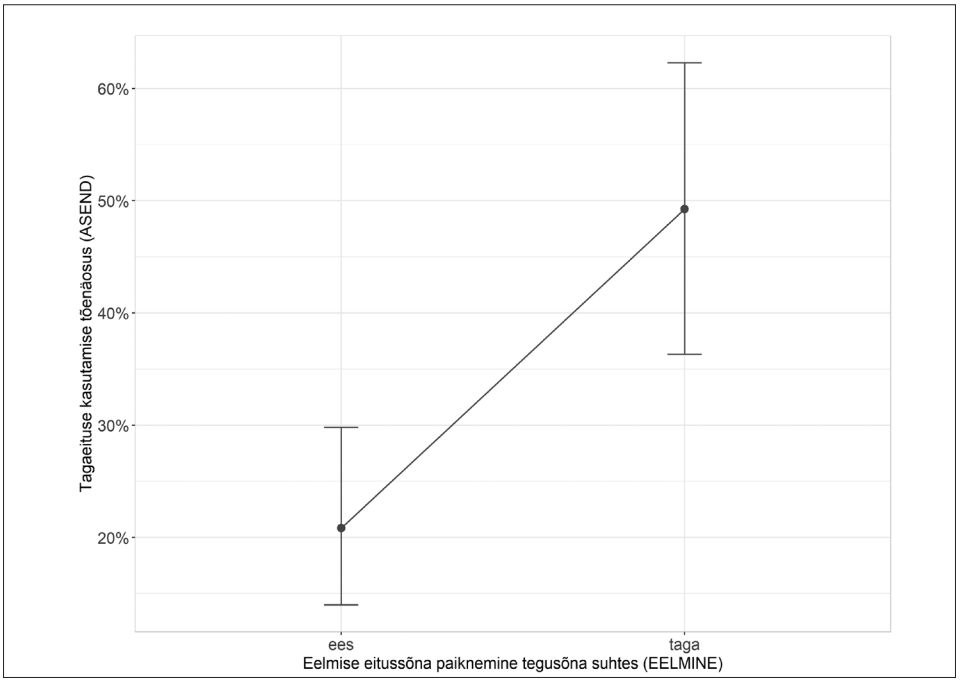
5.2.3. Eitussõna paiknemist seletavad tunnused võru keeles

Võru keel ja selle variandid näitasid üldjoontes väga tugevat eelistust just eeseituse kasutamisele (vt jooniseid 1 ja 2), mis ühtib standardeituse sõnajärjega ka eesti, vene ja läti keeles. Ainult kahes kihelkonnas – Rõuges ja Vastseliinas – esines arvestataval hulgal tagaeituse kasutamist. Võrreldes (ida)setoga on neis olukord vastupidine: üle kahe kolmandiku eituse kasutusjuhtudest esineb eeseitusega ja vaid alla ühe kolmandiku juhtudest tagaeitusega. Rõuge ja Vastseliina andmeid koondavas regressioonimudelis, mis ennustas harvema eituskonstruktsiooni ehk sedapuhku tagaeituse kasutamise tõenäosust, osutus aga ainsaks tugevaks varieerumist seletavaks tunnuseks eelmise eitussõna asend tegusõna suhtes (joonis 6), seega sama tunnus, mis oli oluline ka seto materjalis.

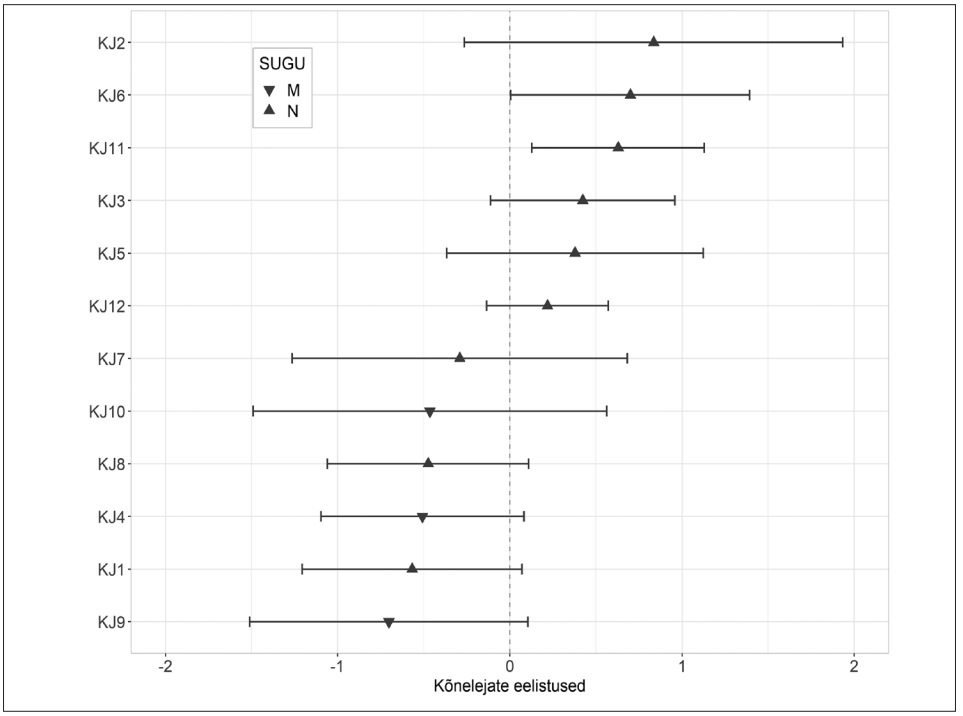
Jooniselt 6 näeme, et praimimise efekt võru keeles on idasetoga samapidine: kui kõnehetkele eelnenud eituskonstruktsioon esines eeseitusega, on palju tõenäolisem eeseituse uuesti kasutamine (tagaeituse kasutamise tõenäosus on sel juhul ainult 21%); kui aga eelmine eituskonstruktsioon on olnud tagaeitus, tõuseb märgatavalt ka tagaeituse uuesti kasutamise tõenäosus, jäädes siiski napilt alla 50%.

Ka kõnelejate juhuslik mõju aitab eitussõna positsiooni valiku varieerumist seletada,[4] tegusõna lemma võru keeles aga mitte. See, et tegusõna osutub oluliseks

[4] Lõpliku mudeli kuju oli ASEND ~ EELMINE + (1|KÕNELEJA). Mudeli *C*-indeks oli 0,74, mis näitab mudeli keskmist eristusvõimet.



Joonis 6. Tagaeituse kasutamise tõenäosus võru andmetes vastavalt eelmise eitussõna paiknemisele.



Joonis 7. Võru keele kõnelejate ees- ja tagaeituse kasutamise eelistused.

varieerumise seletajaks idasetos, aga mitte võrus, osutab vähemalt kaudselt piirkonna keelekontaktide mõju üldisele suunale: idaseto on saanud mõjutusi ümberkaudsetest keeltest, kus on valdav eeseitus (eesti, vene, võru) ning mis on mõjutanud eeseituse kasutust mõningate sagedaste verbidega (nt *ei tiijäq*). Vastupidi see aga ei näi toimivat: seto tagaeitusega keelena ei näi võru keelt samal viisil mõjutavat. See on ka tõenäoline, sest andmed on pärit 1960.–1970. aastatest, mil seto keele prestiiž oli võrukeste silmis madal ning laenamine ebatõenäoline. Samuti ei olnud sel ajal kasutuses võru kirjakeelt, kus tänapäeval on normiks kujundatud tagaeitus. Võru kõnelejate individuaalseid eelistusi kujutab joonis 7.

Jooniselt 7 näeme, et ka võru keele kõnelejad erinevad mõnevõrra individuaalsete eelistuste poolest ees- ja tagaeituse kasutamisel, kusjuures kõik meessoost kõnelejad näivad taas eelistavat keskmisest enam eeseitust. Paraku ei ole EMK-s võru murrakute kõnelejate kohta sama põhjalikku taustinfot kui idaseto kõnelejate kohta ning seetõttu on raske neid eelistusi sotsiolingvistiliste tegurite kaudu seletada. Küll aga näitab ka võru keele mudel, et kõnelejate idiosünkraatilise käitumisega arvestamine keele varieerumise uurimisel on oluline ning võimaldab arvesse võtta ka seda varjatud osa varieerumisest, mis fikseeritud lingvistiliste või ekstralingvistiliste kategooriatega ei pruugi ilmneda.

Statistilise mudeldamise strateegiaid on erinevaid (vt nt Hosmer jt 2013; Winter 2020; Gries 2021) ning nii keelekasutuse formaalsete mudelite koostamist kui ka tõlgendamist peab alati saatma dialoog keeleteooria ja varasemate uurimustega, eriti kui eesmärk on keelekasutuse võimalikult adekvaatne kirjeldamine, mitte mudelite parim ennustustäpsus. Mudelid, milleni siinse analüüsi käigus jõudsime, on meie hinnangul kasutada olevate andmete põhjal optimaalsed, kirjeldamaks olulisemaid tegureid, mis valimitesse sattunud kõnelejate ees- ja tagaeituse valikut suunavad. Tuleb aga rõhutada, et ehkki raporteerisime ruumipuudusel *p*-väärtuste alusel vaid lõplikud mudelid statistiliselt oluliste tunnustega, oli meie mudeldamisstrateegia pigem uuriv kui kinnitav (vt nt Winter 2020: 274–280). See tähendab, et meil puudusid tugevad eeldused eituskonstruktsiooni sõnajärge tingivate tegurite kohta, ning testisime kõiki võimalikke seoseid ja koosmõjusid, mida oleksime suutnud põhjendada ning mida andmestik lubas analüüsi kaasata. Selle uurimuse andmestikust koorunud eelistusmustrite paikapidavust tuleks seega võimalusel testida uute andmestikega. Hoolimata sellest, et mudelid ei ole kunagi ideaalsed esitused keeles esinevast varieerumisest, on mitmemõõtmelise statistika meetoditest abi kompleksete seoste tuvastamisel ja hindamisel suurtes andmehulkades, millest palja silmaga mustreid leida on keerukas.

6. Kokkuvõte

Artiklis näitlikustasime, kuidas korpusandmete kvantitatiivne analüüs võimaldab kirjeldada ja seletada keeles esinevat varieerumist, analüüsides juhtumiuuringus eitussõna paiknemist võru ja seto suulises keelekasutuses. Analüüsitud andmed pärinevad loodavast seto interdistsiplinaarsest korpusest, mis sisaldab muu hulgas

aastatel 2010–2013 Pihkva oblastis läbi viidud välitöödel tehtud intervjuusid, ning eesti murrete korpusest, kus on samalaadsed tekstid 1960.–1980. aastatest.

Võru ja seto on nii hääldus, vormimoodustuse kui ka sõnavara poolest sarnased, mistõttu on neid dialektoloogias käsitletud ühe Võru murdena. Ometi avaldub lausetasandil ka erinevusi, mida põhjendatakse sageli eesti ja vene üldkeele mõju erineva ulatusega keeleala eri piirkondades. Üks silmapaistvamaid võru ja seto erinevusi on eitussõna asukoht eituslauses. Ehkki mõlemas on võimalik paigutada eitussõna nii tegusõna ette kui ka taha, mis tüpoloogiliselt on küllalt haruldane, eelistavad võru keele kõnelejad kasutada ees- ja seto kõnelejad tagaeitust.

Artiklis vaatlesime esmalt eitusmustrite levikut võru-seto keelealal, et kaardistada ees- ja tagaeituse eelistusmustreid ning tuvastada üleminekuala ühe eelistamiselt teisele. Selgus, et setos on valdav tõepoolest tagaeitus, moodustades kolmveerandi kõikidest eituskonstruktsiooni kasutamise juhtudest. Võrus seevastu prevaleerib selgelt eeseitus ning ainult kahes kihelkonnas – Rõuges ja Vastseliinas – võis täheldada märgatavamat ees- ja tagaeituse varieerumist.

Seejärel uurisime mitmemõõtmelise statistilise analüüsi abil, kas kahe keelevariandi sees esinevat varieerumist suunavad ühesugused või erinevad tegurid. Selleks rakendasime segamõjudega logistilist regressiooni, mis võimaldab hinnata grammatiliste, semantiliste ja suhtluslike tegurite üldist rolli eituskonstruktsiooni valikul, võttes samal ajal arvesse ka kõneleja- ja tegusõnaspetsiifilisi eelistusmustreid. Tulemused näitasid nii võrus kui ka setos selget tendentsi korrata juba kõneleja või vestluspartneri aktiveeritud vorme: ehkki setos domineerib taga- ja võrus eeseitus, kasvab harvemini esineva variandi kasutamise tõenäosus mõlemas keeles märkimisväärselt siis, kui vastav eituskonstruktsioon on hiljuti juba suhtluses aktiveeritud.

Idasetos osutus oluliseks ka kasutatud eitussõna, mis vastab kas oleviku või mineviku ajavormile: oleviku eitussõnaga *ei* on eesti ja vene keele pärase eeseituse kasutamise tõenäosus oluliselt suurem kui mineviku eitussõnaga *es*. See on tõenäoliselt seotud kontaktkeeltest üle võetud suhtluses sagedaste kinnistunud sõnajärjega eituskonstruktsioonidega (nt *ei tiijäq*).

Sellistes võrdlemisi väikestes valimites ei saa mööda vaadata kõnelejate individuaalsest mõjust varieerumisele, ehkki selle mõju sisuline tõlgendamine võib kõneleja tausta teadmata osutuda keeruliseks. Kõneleja osutus oluliseks nii seto kui ka võru eitusmustrite eelistuste seletamisel. Setos saime siduda selle kõnelejate erineva kokkupuutega kontaktkeeltega (eesti, vene), ent võru kohta vastav info paraku puudus. Ka tegusõnade idiosünkraatiline käitumine võib heita valgust nii keelekontaktide rollile suhtluses sagedaste väljendite kasutamisel kui ka keelesisestele protsessidele, milles järjendid kasutussageduse ja/või morfofonoloogiliste tegurite mõjul võivad eituskonstruktsioonis kindla sõnajärjega kinnistuda ja automatiseeruda.

Artikli valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus) ning EKKD projekt „Interdistsiplinaarne seto korpus".

KIHELKONDADE LÜHENDID:

Hargla; **ISe** = Ida-Seto; **Kan**epi; **Krl** = Karula; **LSe** = Lõuna-Seto; **Plv** = Põlva; **PSe** = Põhja-Seto; **Rõu**ge; **Urv**aste; **Vas**tseliina.

VEEBIVARAD

EMK = Eesti murrete korpus. https://www.keel.ut.ee/et/keelekogud/murdekorpus

ggplot2. Create Elegant Data Visualisations Using the Grammar of Graphics. https://cran.r-project.org/package=ggplot2

Goldvarb Z. A multivariate analysis application for Macintosh. Toronto: Department of Linguistics, University of Toronto, 2018. http://individual.utoronto.ca/tagliamonte/goldvarb.html

Goldvarb X. A variable rule application for Macintosh and Windows. Toronto: Department of Linguistics, University of Toronto, 2005. http://individual.utoronto.ca/tagliamonte/goldvarb.html

lme4. Linear Mixed-Effects Models using 'Eigen' and S4. https://cran.r-project.org/package=lme4

Murdearhiiv = Tartu Ülikooli eesti murrete ja sugulaskeelte arhiiv. https://murdearhiiv.ut.ee

R Core Team 2020. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org

SetKo = Interdistsiplinaarne seto korpus. https://setko.ut.ee

sf. Simple Features for R. https://cran.r-project.org/package=sf

sjPlot. Data Visualization for Statistics in Social Science. R package version 2.8.7. https://CRAN.R-project.org/package=sjPlot

KIRJANDUS

Baayen, R. Harald 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. Harald; Davidson, Douglas J.; Bates, Douglas M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. – Journal of Memory and Language, kd 59, nr 4, lk 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, Douglas; Maechler, Martin; Bolker, Ben; Walker, Steve 2015. Fitting linear mixed-effects models using lme4. – Journal of Statistical Software, kd 67, nr 1, lk 1–48. https://doi.org/10.18637/jss.v067.i01

Bod, Rens; Hay, Jennifer; Jannedy, Stefanie (toim) 2003. Probabilistic Linguistics. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/5582.001.0001

Bresnan, Joan 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. – Roots: Linguistics in Search of its Evidential Base. Toim Sam Featherston, Wolfgang Sternefeld. Berlin: Mouton de Gruyter, lk 75–96.

Bresnan, Joan; Ford, Marilyn 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. – Language, kd 86, nr 1, lk 168–213. https://doi.org/10.1353/lan.0.0189

Cedergren, Henrietta J.; Sankoff, David 1974. Variable rules: Performance as a statistical reflection of competence. – Language, kd 50, nr 2, lk 333–355. https://doi.org/10.2307/412441

Diessel, Holger 2017. Usage-based linguistics. – Oxford Research Encyclopedia of Linguistics. Toim Mark Aronoff. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.363

Dryer, Matthew S. 2013. Order of negative morpheme and verb. – The World Atlas of Language Structures Online. Toim M. S. Dryer, Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/143 (11. VII 2019).

Eichenbaum, Külli; Pajusalu, Karl 2001. Setode ja võrokeste keelehoiakutest ja identiteedist. – Keel ja Kirjandus, nr 7, lk 483–489.

Grafmiller, Jason; Szmrecsanyi, Benedikt; Röthlisberger, Melanie; Heller, Benedikt 2018. General introduction: A comparative perspective on probabilistic variation in grammar. – Glossa: A Journal of General Linguistics, kd 3, nr 1, lk 94. https://doi.org/10.5334/gjgl.690

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. – Corpora, kd 10, nr 1, lk 95–125. https://doi.org/10.3366/cor.2015.0068

Gries, Stefan Th. 2021. (Generalized linear) mixed-effects modeling: A learner corpus example. – Language Learning, kd 71, nr 3, lk 757–798. https://doi.org/10.1111/lang.12448

Hagu, Paul; Pajusalu, Karl 2020. Lühkene seto keeleoppus. Lühike seto keeleõpetus. Seto Instituut. http://www.setoinstituut.ee/download/luhikene-seto-keeleopetus (2. VII 2021).

Hazen, Kirk 2011. Labov: Language variation and change. – The SAGE Handbook of Sociolinguistics. Toim Ruth Wodak, Barbara Johnstone, Paul Kerswill. SAGE Publications Ltd, lk 24–39. https://doi.org/10.4135/9781446200957.n3

Hinrichs, Lars; Szmrecsanyi, Benedikt 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. – English Language and Linguistics, kd 11, nr 3, lk 437–474. https://doi.org/10.1017/S1360674307002341

Hint, Helen; Taremaa, Piia; Reile, Maria; Pajusalu, Renate 2021. Demonstratiivpronoomenid ja -adverbid määratlejatena. Miks me oleme *siin ilmas*, *selles olukorras*? – Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics, kd 12, nr 1, lk 79–111.

Hosmer, David W. Jr.; Lemeshow, Stanley; Sturdivant, Rodney X. 2013. Applied Logistic Regression. Hoboken, NJ: John Wiley and Sons. https://doi.org/10.1002/9781118548387

Iva, Sulev 2002. Võro-eesti synaraamat. (Võro Instituudi toimõndusõq 12.) Võro: Võro Instituut.

Iva, Sulev 2007. Võru kirjakeele sõnamuutmissüsteem. (Dissertationes philologiae estonicae Universitatis Tartuensis 20.) Tartu: Tartu Ülikooli Kirjastus.

Iva, Sulev 2013. Võru ja seto kõrihäälikud *h* ja *q*. – Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastaraamat XI–XII. Tartu: Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskus, lk 102–116.

Janhunen, Juha 1982. On the structure of Proto-Uralic. – Finnisch-Ugrische Forschungen, nr 44, lk 23–42.

Jääts, Indrek 2000. Ethnic identity of the Setus and the Estonian-Russian border dispute. – Nationalities Papers, kd 28, nr 4, lk 651–670. https://doi.org/10.1080/00905990020009665

Kallio, Petri 2007. Kantasuomen konsonanttihistoriaa. – Sámit, sánit, sátnehámit: Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007. (Mémoires de la Société Finno-Ougrienne 253.) Toim Jussi Ylikoski, Ante Aikio. Helsinki: Société Finno-Ougrienne, lk 229–249.

Kallio, Petri 2014. The diversification of Proto-Finnic. – Fibula, Fabula, Fact: The Viking Age in Finland. (Studia Fennica Historica 18.) Toim Joonas Ahola, Frog, Clive Tolley. Helsinki: Suomalaisen Kirjallisuuden Seura, lk 155–168.

Keem, Hella; Käsi, Inge 2002. Võru murde tekstid. (Eesti murded VI.) Tallinn: Eesti Keele Instituut.

Klavan, Jane 2012. Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy. (Dissertationes linguisticae Universitatis Tartuensis 15.) Tartu: Tartu Ülikooli Kirjastus.

Klavan, Jane 2018. Kognitiivne keeleteadus arvude rägastikus. – Keel ja Kirjandus, nr 8–9, lk 697–712.

Klavan, Jane 2021 (ilmumas). The alternation between exterior locative cases and postpositions in Estonian web texts. – Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics, kd 12, nr 1.

Klavan, Jane; Divjak, Dagmar 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. – Folia Linguistica, kd 50, nr 2, lk 355–384. https://doi.org/10.1515/flin-2016-0014

Klavan, Jane; Pilvik, Maarja-Liisa; Uiboaed, Kristel 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. – SKY Journal of Linguistics, nr 28, lk 187–224.

Klavan, Jane; Veismann, Ann 2017. Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition *peal* 'on'. – Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics, kd 8, nr 2, lk 59–91. https://doi.org/10.12697/jeful.2017.8.2.03

Kortmann, Bernd 2010. Areal variation in syntax. – Language and Space: An International Handbook of Linguistic Variation. Theories and Methods. Toim Peter Auer, Jürgen E. Schmidt. Berlin–New York: De Gruyter, lk 837–864. https://doi.org/10.1515/9783110220278.837

Kortmann, Bernd 2021. Reflecting on the quantitative turn in linguistics. – Linguistics. https://doi.org/10.1515/ling-2019-0046

Krasnoukhova, Olga; Auwera, Johan van der; Crevels, Mily 2021 (ilmumas). Introduction: Postverbal negation: What, where, why. – Studies in Language.

Labov, William 1963. The social motivation of a sound change. – Word, kd 19, nr 3, lk 273–309. https://doi.org/10.1080/00437956.1963.11659799

Labov, William 1972. Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press.

Lindström, Liina 1997. Eitus Võru murde suulises kõnes. – Õdagumeresoomõ lõunapiir'. Toim Karl Pajusalu, Jüvä Sullõv. (Võro Instituudi Toimõtiseq 1.) Võro: Võro Instituut', lk 143–154.

Lindström, Liina 2001. Eesti murrete korpuse iseloomustus argivestlustega võrrelduna. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. (Tartu Ülikooli eesti keele õppetooli toimetised 17.) Toim Reet Kasik. Tartu: Tartu Ülikooli Kirjastus, lk 212–221.

Lindström, Liina 2015. Ülevaade eesti murrete korpusest seisuga 17.11.2015. https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf (2. VII 2021).

Lindström, Liina 2021 (ilmumas). Seto lause põhijooned. – Setomaa 3. Keel, rahvaluule ja tänapäeva kultuur. Toim Andreas Kalkun, Karl Pajusalu, Ergo-Hart Västrik. Tartu: Eesti Rahva Muuseum.

Lindström, Liina; Kalmus, Mervi; Klaus, Anneliis; Bakhoff, Liisi; Pajusalu, Karl 2009. Ainsuse 1. isikule viitamine eesti murretes. – Emakeele Seltsi aastaraamat 54 (2008). Tallinn: Teaduste Akadeemia Kirjastus, lk 159–185.

Lindström, Liina; Pilvik, Maarja-Liisa 2018. Korpuspõhine kvantitatiivne dialektoloogia. – Keel ja Kirjandus, nr 8–9, lk 643–662.

Lindström, Liina; Pilvik, Maarja-Liisa; Plado, Helen 2018. Nimetamiskonstruktsioonid eesti murretes: murdeerinevused või suuline süntaks? – Mäetagused, nr 70, lk 91–126. https://doi.org/10.7592/MT2018.70.lindstrom_pilvik_plado

Lindström, Liina; Pilvik, Maarja-Liisa; Plado, Helen 2021 (ilmumas). Variation in negation in Seto. – Studies in Language. https://doi.org/10.1075/sl.19063.lin

Lindström, Liina; Pilvik, Maarja-Liisa; Ruutma, Mirjam; Uiboaed, Kristel 2019. On the use of perfect and pluperfect in Estonian dialects: Frequency and language contacts. – Multilingual Finnic – Language Contact and Change. Toim Sofia Björklöf, Santra Jantunen. Helsinki: Finno-Ugrian Society, lk 155–193. https://doi.org/10.33341/uh.85035

Lindström, Liina; Uiboaed, Kristel 2017. Syntactic variation in 'need'-constructions in Estonian dialects. – Nordic Journal of Linguistics, kd 40, nr 3, lk 313–349. https://doi.org/10.1017/S0332586517000191

Lindström, Liina; Vihman, Virve-Anneli 2017. Who needs it? Variation in experiencer marking in Estonian 'need'-constructions. – Journal of Linguistics, kd 53, nr 4, lk 789–822. https://doi.org/10.1017/S0022226716000402

Lüdecke, Daniel 2018. ggeffects: Tidy data frames of marginal effects from regression models. – Journal of Open Source Software, nr 3 (26), lk 772. https://doi.org/10.21105/joss.00772

Mets, Mari 2010. Suhtlusvõrgustikud reaalajas: võru kõnekeele varieerumine kahes Võrumaa külas. (Dissertationes philologiae estonicae Universitatis Tartuensis 25.) Tartu: Tartu Ülikooli Kirjastus.

Miestamo, Matti; Tamm, Anne; Wagner-Nagy, Beáta 2015. Negation in Uralic languages. Introduction. – Negation in Uralic Languages. Toim M. Miestamo, A. Tamm, B. Wagner-Nagy. Amsterdam–Philadelphia: John Benjamins, lk 1–41. https://doi.org/10.1075/tsl.108.01int

Paas, Friedrich-Eugen 1927. Sega-abielud ja nende mõju rahvusele piiriäärsetes maakondades Eestis. Tartu: Eesti Vabariigi Tartu Ülikool.

Pajusalu, Karl 1996. Multiple Linguistic Contacts in South Estonian: Variation of Verb Inflection in Karksi. (Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja. Publications of the Department of Finnish and General Linguistics of the University of Turku 54.) Turku: Turku University.

Pajusalu, Karl; Hennoste, Tiit; Niit, Ellen; Päll, Peeter; Viikberg, Jüri 2002. Eesti murded ja kohanimed. Tallinn: Eesti Keele Sihtasutus.

Pajusalu, Karl; Velsker, Eva; Org, Ervin 1999. On recent changes in South Estonian: Dynamics in the formation of the inessive. – International Journal of the Sociology of Language, nr 139, lk 87–103. https://doi.org/10.1515/ijsl.1999.139.87

Paolillo, John C. 2002. Analyzing Linguistic Variation. (CSLI Lecture Notes 114.) Stanford: CSLI Publications.

Pebesma, Edzer 2018. Simple features for R: Standardized support for spatial vector data. – The R Journal, kd 10, nr 1, lk 439–446. https://doi.org/10.32614/RJ-2018-009

Pilvik, Maarja-Liisa; Lindström, Liina; Plado, Helen 2021. Murded, varieerumine ja korpusandmed: eitussõna paiknemine võru ja seto eituslausetes. Lisamaterjalid. https://doi.org/10.17605/OSF.IO/GCFT7

Pinheiro, José C.; Bates, Douglas M. 2000. Mixed-effects Models in S and S-PLUS. New York: Springer. https://doi.org/10.1007/978-1-4419-0318-1

Plado, Helen 2015. *des-* ja *mata*-konverbi kasutusest eesti murretes. – Emakeele Seltsi aastaraamat 60 (2014). Tallinn: Teaduste Akadeemia Kirjastus, lk 195–218. https://doi.org/10.3176/esa60.10

Reile, Maria; Plado, Helen; Gudde, Harmen B.; Coventry, Kenny R. 2020. Demonstratives as spatial deictics or something more? Evidence from Common Estonian and Võro. – Folia Linguistica, kd 54, nr 1, lk 167–195. https://doi.org/10.1515/flin-2020-2030

Saareste, Andrus 1955. Petit atlas des parlers estoniens. Väike eesti murdeatlas. Uppsala: Almqvist & Wiksells.

Siiman, Ann 2018. Variation of the Estonian singular long and short illative form: A multivariate analysis. – SKY Journal of Linguistics, nr 31, lk 139–167.

Siiman, Ann 2019. Vormikasutuse varieerumine ning selle põhjused osastava ja sisseütleva käände näitel. (Dissertationes philologiae estonicae Universitatis Tartuensis 45.) Tartu: Tartu Ülikooli Kirjastus.

Szmrecsanyi, Benedikt 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. – Corpus Linguistics and Linguistic Theory, kd 1, nr 1, lk 113–149. https://doi.org/10.1515/cllt.2005.1.1.113

Szmrecsanyi, Benedikt 2017. Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. – Canadian Journal of Linguistics / Revue canadienne de linguistique, kd 62, nr 4, lk 685–701. https://doi.org/10.1017/cnj.2017.34

Szmrecsanyi, Benedikt; Anderwald, Lieselotte 2018. Corpus-based approaches to dialect study. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Malden, MA: Wiley-Blackwell, lk 300–313. https://doi.org/10.1002/9781118827628.ch17

Tagliamonte, Sali A. 2012. Variationist Sociolinguistics: Change, Observation, Interpretation. Chichester: Wiley-Blackwell.

Tagliamonte, Sali A.; Baayen, R. Harald 2012. Models, forests, and trees of York English: *Was*/*were* variation as a case study for statistical practice. – Language Variation and Change, kd 24, nr 2, lk 135–178. https://doi.org/10.1017/S0954394512000129

Taremaa, Piia; Hint, Helen; Reile, Maria; Pajusalu, Renate 2021. Constructional variation in Estonian: Demonstrative pronouns and adverbs as determiners in noun phrases. – Lingua, kd 254, nr 103030. https://doi.org/10.1016/j.lingua.2021.103030

Torres Cacoullos, Rena; Travis, Catherine E. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. – Linguistics, kd 57, nr 3, lk 653–692. https://doi.org/10.1515/ling-2019-0011

Tulving, Endel 2002. Mälu. Tartu: Tartu Ülikooli Kirjastus.

Veismann, Ann; Klavan, Jane; Õim, Haldur 2018. Teoreetiline keeleteadus ja kvantitatiivsed meetodid. – Keel ja Kirjandus, nr 8–9, lk 609–621.

Viikberg, Jüri 2020. Eesti murrete grammatika. (Eesti keele varamu VIII.) Tartu: Tartu Ülikooli Kirjastus.

Viitso, Tiit-Rein 1998. Estonian. – The Uralic Languages. Toim Daniel Abondolo. London–New York: Routledge, lk 115–148.

Walker, James A. 2013. Variation analysis. – Research Methods in Linguistics. Toim Robert J. Podesva, Devyani Sharma. Cambridge: Cambridge University Press, lk 440–459.

Wickham, Hadley 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4_9

Winter, Bodo 2020. Statistics for Linguists: An Introduction Using R. New York: Routledge. https://doi.org/10.4324/9781315165547

Maarja-Liisa Pilvik (snd 1989), MA, Tartu Ülikooli doktorant ja rakendusliku dialektoloogia nooremteadur (Jakobi 2-430, 51005 Tartu), maarja-liisa.pilvik@ut.ee

Helen Plado (snd 1981), PhD, Tartu Ülikooli eesti keele teadur ja lektor ning Võru Instituudi teadur (Jakobi 2-426, 51005 Tartu), helen.plado@ut.ee

Liina Lindström (snd 1973), PhD, Tartu Ülikooli tänapäeva eesti keele professor ning digihumanitaaria ja infoühiskonna keskuse juhataja (Jakobi 2-443, 51005 Tartu), liina.lindstrom@ut.ee

Dialects, variation and corpus data: Position of the negation word in Võru and Seto

Keywords: dialects, language variation, multivariate analysis, corpus linguistics, negation, Võru, Seto

The article exemplifies contemporary language variation studies by examining the variation in the position of the negation word in the Võru and Seto varieties of South Estonian. Using frequency analysis of dialect corpus data, it is shown that despite being close both geographically and linguistically, Seto and Võru exhibit opposite tendencies with regard to their preferences towards pre- and postverbal negation. Võru speakers predominantly use preverbal negation patterns, showing more variation only in two southwestern parishes – Rõuge and Vastseliina. Seto, in turn, prefers the typologically rare postverbal negation construction, but occasionally also makes use of preverbal negation.

Mixed-effects logistic regression analysis was conducted to compare newer East Seto data with older data from Rõuge and Vastseliina with regard to the importance, strength and direction of the factors conditioning the variation in Seto and Võru. The results suggested that the variation in both language varieties is primarily affected by factors related to memory, processing, and communication. In particular, it was the word order of the previously activated negation construction that turned out to have the strongest effect: the use of both pre- and postverbal negation significantly raised the probability of the construction being repeated with the same respective word order. In East Seto, the lexical negation word (expressing grammatical tense) also had a significant role: the use of a preverbal negation construction became more likely with the present tense negation word *ei*. This had probably to do with copying the word order of frequent fixed expressions (*ma ei tiijäq* 'I don't know') from Estonian and Russian as contact languages. The individual effects of speakers and verb lemmas were also accounted for in the analyses as important sources of random variation.

Maarja-Liisa Pilvik (b. 1989), MA, University of Tartu, PhD Student, Junior Research Fellow (Jakobi 2-430, 51005 Tartu), maarja-liisa.pilvik@ut.ee

Helen Plado (b. 1981), PhD, University of Tartu, Research Fellow; Võro Institute, Researcher (Jakobi 2-426, 51005 Tartu), helen.plado@ut.ee

Liina Lindström (b. 1973), PhD, University of Tartu, Professor of Modern Estonian, Head of the Centre for Digital Humanities and Information Society (Jakobi 2-443, 51005 Tartu), liina.lindstrom@ut.ee

https://doi.org/10.54013/kk764a8

Kakasonist Kaljulaks

Perekonnanimede eestistamise andmebaasi analüüs

FRED PUSS

Nimi on oluline tunnus, mille põhjal tehakse isikut nägemata järeldus tema soo, rahvuse, sugulussuhete (patronüümid), sageli ka vanuse ning muude omaduste kohta. Kuna Eesti esimesel iseseisvusajal oli oluline rahvusriigi ülesehitamine, hakati rohkem tähelepanu pöörama ka nimede keelsusele. Tolleaegne nimede eestistamise kampaania puudutas sadu tuhandeid inimesi, kuid tegu on vaid osaliselt läbi uuritud teemaga. Selle taga on nähtud nii poliitilisi võtteid vaikival ajastul rahvale tegevuse pakkumise eesmärgil kui ka rahvuslikku algatust saamaks lahti „orjaöö" taagast. Nimede eestistamise eeskuju võeti Soomest, kuid õige hoog saavutati 1934. aasta riigipöörde järel tekkinud riikliku toetusega (Raun 2012).

Nimede rahvuskeelestamise kampaaniaid on tehtud näiteks nii XIX sajandil Ungaris, fašistlikus Itaalias (mõlemas küll vähemusrahvuste arvel), Teise maailmasõja järgses Poolas (Walkowiak 2016: 229, 235, 238) kui ka XX sajandi esimesel kolmandikul Soomes (Närhi 1987) ning vähesel määral Lätis (Staltmane 1977) jm. Kuigi nende kampaaniate käigus nimesid muutnud inimeste arv ulatub Eestiga sarnaselt sadadesse tuhandetesse (v.a Lätis), on Eesti kampaania unikaalne lühiajalisuse ja nimemuutjate osakaalu poolest elanikkonnast (iga kuues elanik!). Erinevalt teistest maadest on Eesti nimemuutmised onomastika andmebaasina (OA) juba mitukümmend aastat kõigile kättesaadavad: 1990. aastatel koostas üks digihumanitaaria pioneere Eesti ajalooteaduses (Tamm 2020: 42), Tartu Ülikooli arhiivinduse professor Aadu Must perekonnanimede paneku ja eestistamise andmebaasid. Aadu Mustalt 2000. aastate algul saadud algandmestikku (mis on väheste parandustega praegu veebis, vt OA) on siinkirjutaja sellest ajast alates parandanud ja täiendanud.[1]

Praeguseks on digitaalselt kättesaadavad ka mitmed lisaandmestikud. Esiteks andmebaasi aluseks olnud Riigi Teataja Lisa tärktuvastus DIGAR-is ning nimemuutmiste toimikud, mis seni olid teaduskäibest eemal, kuna need anti alles 2020. aasta mais Siseministeeriumist Rahvusarhiivi. Seal koostatud toimikute kirjeldusandmed arhiiviinfosüsteemis AIS annavad OA-le hea võrdlus- ning kontrollivõimaluse.[2] Samuti poleks soovitusnimestiku „Eesti nimi" tärktuvastuseta (DIGAR-is) olnud võimalik selle käsitelu.

[1] Tekstis viitab OA autori parandatud ja täiendatud versioonile. Kui on olnud vaja viidata algsele (avalikule) versioonile, on lisatud vastav selgitav sõna. Artiklis on kasutatud vaid OA eestistamise andmestikku.

[2] Autor tänab Rahvusarhiivi andmebaasi administraatorit Maiu Pevkurit toimikute andmete tabelkujul väljastamise eest ning Rahvusarhiivi arhivaari Andero Nimmerit toimikutes leiduvate andmete kontrollimise eest.

Artiklis antakse ülevaade perekonnanimede eestistamise senisest uurimisest ja eestistamise korraldusest, kirjeldatakse probleeme nimemuutmiste andmestikega töötamisel ning kasutatakse OA-d koos võrdlusandmestikega nimemuutjate üldarvu, soovitusnimestiku mõju ning võetud nimede statistiliste omaduste (esitäht, pikkus, ainukordsus) analüüsiks.

Historiograafia

Eestistamise kampaania kõrgajal 1935–1940 ilmus peaaegu iga päev ajakirjanduses teemakohaseid teateid ning kuni 1940. aastani avaldati mõningaid üldisi kokkuvõtteid (nt Elisto 1938), kuid protsess ja sellest ülevaadete andmine vaibusid 1940. aastal.

Välis-Eestis puudutati nimede eestistamise teemat vaid põgusalt. Keeleteadlane Hildegard Must (1957: 285) andis eestistamist tutvustades hinnangu, et „rohkem traditsioonidest ja loomulikkudest eeskujudest kinnipidamist oleks olnud ettevõttele kasuks". Puudustena mainis ta ebamääraste ja kunstlike nimede ning mitte-eestikeelsete sõnavormide kasutust, talunimede alakasutamist ning kohati teadmatust uue nime(vormi) kokkulangemisest võõrapärasega (nt *-ma* lõpp kattub tüüpiliste friisi nimedega) (Must 1957: 285). Teised Välis-Eesti (sh baltisaksa) kirjutised olid valdavalt referatiivsed ja mälestuslikud (Raag 1971: 189–191; Angelus 1971; ArA 1971; Grönberg 1985; Nimede eestistamine 1987).

Nõukogude ajal nimemuutmist kui sügavalt rahvuslikku üritust avalikult ei uuritud. Kogumik „Eestlasele eesti nimi" (EEN 1921) ja uute perekonnanimede nimestik „Eesti nimi" (EN 1935) olid kantud keelatud kirjanduse nimekirja (Glavlit 1948: 16; Glavlit 1963: 29), nii nagu kaasaegnegi perioodika. Oma mälestustes Emakeele Seltsist vaikib eestistamise peaaegu täielikult maha ka seltsi pikaaegne sekretär ja eestistamistoimkonna liige Elmar Elisto (1976: 89, viidatud tekst ilmus esmakordselt 1960). Arnold Kase (1960: 6) koostatud ülevaates Emakeele Seltsi 40 aasta tegevusest oli võimalik mainida, et seltsi propageeritud eestistamine ei leidnud poolehoidu ning hakkas „mõnevõrra levima" alles 1930. aastatel. Kümme aastat hiljem kirjutas Kask (1970: 8, 12) mõne rea jagu pikemalt ning objektiivsemalt (1934. aastal hakkas eestistamine „muutuma hoogsamaks"). Edaspidigi ilmus teemast vaid mälestuskatkeid (Roos 1980), isegi viljakaim nimeuurija ning Nimede Eestistamise Keskbüroo juhataja Edgar Rajandi (Uut hoogu 1940) ei kirjutanud teemast midagi peale mõne lõigu (Rajandi, Tarand 1976: 135), kuigi tal oli nimemuutmistest eraldi materjalikogu (Rajandi 1978: 96).

Taasiseseisvunud Eestis ilmunud uurimused käsitlevad eestistamise õiguslikku külge (Henno 2002; Mõistlik 2005), organisatsioonide tegevust (Henno 2001; Leivategija 2011), aktsiooni üldist käiku (Rajasalu 1993) või poliitilist tausta (Vaan 2005: 48–55; Raun 2012). Aadu Must on OA-d tutvustanud nii konverentsidel (2006, 2010), õpik-käsiraamatutes (2000, 2015) kui ka ajakirjanduses (1999). Neis töödes on käsitletud ajalooalaseid andmebaase kogu maailmas ja OA koostamist, kuid selle analüüs on piirdunud vaid näidete ja üldkokkuvõtetega.

Leidub mõningaid üliõpilastöid, kus OA-d on kasutatud piirkondliku nimemuutmise analüüsiks (Raidmaa 1999; Pärn 2013), ning selliste tööde panus on väiksema andmestiku detailne analüüs. Üldisem ülevaade aga puudub siiani.

Perekonnanimede muutmise korraldus

1919. aasta detsembris vastu võetud perekonnanime muutmise seadus (RT 1920, 2/3, 15)[3] andis r i i g i k o h t u l e õiguse nime muutmise otsustamiseks. Nime võisid muuta vaid Eesti kodanikud ning põhjendusteks võisid olla „halb tähendus, võõrakeelne või halb kõla või sugukonnanime alalhoidmise soov". Muutmisprotsess kestis umbes pool aastat. Protsessi keerukuse ning kalliduse tõttu läbisid selle vaid vähesed inimesed.

OA-s on 539 riigikohtu tehtud muutmisotsust. Tolle perioodi andmestikus on maapiirkonnad n-ö üleesindatud: umbes pooled tsaariajal maal hingekirjas olnud (ning sellisena muutmisotsuses ja seega ka OA-s näidatud) inimestest elasid vabariigi algusajal linnas. Näiteks *Bernhard* ja *Helmi Mühlberg* > *Mäelo* elasid Tartus, kuid olid hingekirjas Kuigatsi vallas ning see on muutmisotsuses nendega seostatav ainus kohanimi.[4] Samuti *Friedebert Mihkelson* > *Tuglas* elas koos abikaasaga Tartus, kuid muutmisotsuses kajastub hingekirjutuse koht Ahja vald. Sel põhjusel ei näitaks olemasolevate andmete põhjal tolle perioodi nimemuutmiste geograafiline uurimine tegelikkust objektiivselt, näiteks nimemuutmiste kese Tartu linn on andmete esitusviisi tõttu alaesindatud.

1919. aasta seadusest sõltumatult võidi oma võõrkeelse kõlaga perekonnanime muuta P e t s e r i m a a j a N a r v a t a g u s e elanikele perekonnanimede panekuks moodustatud k o m i s j o n i d e s (RT 1921, 26, 21). Aastail 1921–1923 kasutati seda võimalust 908 korral. Seda tuleks erinevalt senisest lugeda nimede eestistamise osaks, ent kuna tegu on eraldi andmekoguga, mille analüüs pole artikli eesmärk, on need välja jäetud. Muutmised hõlmasid ligikaudu 2000–3000 inimest, kes peamiselt vahetasid ühe vene nime teise vastu (*Šlopkin* > *Akatsiev*, *Ustinov* > *Baitsov*). Enamasti polnud tegu fikseeritud perekonnanime muutmisega, vaid patro- või paponüümi või mitteametliku perekonnanime (vn *уличная фамилия*) asemele ametliku perekonnanime võtmisega. Mõnesajal korral on võetud ka eesti nimi. Nendel puhkudel vahetati eesti (halvakõlaline) nimi teise eesti nime vastu (*Elevant* > *Ant*, *Lirtse* > *Seeder*) või ka vene (*Isvoštšik* > *Ehavald*) või saksa (*Stamberg* > *Haukanõmm*) nimi eestipärase vastu. (PENA 1921–1923) Osa selliselt muudetud perekonnanimedest (*Steinfeld* > *Kiviste*) jõudis August Rava koostatud teatmikku (MN 1935) ning selle kaudu ka OA-sse, kuid tegu on üksikjuhtumitega.

[3] Riigi Teataja (RT) puhul on viidatud aastale, numbrile ning artiklile, Riigi Teataja Lisa (RTL) puhul aastale, numbrile ja leheküljele. Sama kehtib nende jätkajate Eesti NSV Teataja (Lisa) ja Ametliku Teataja kohta.

[4] Teksti koormamise vältimiseks pole näidete viiteid esitatud. Need on OA-st hõlpsasti leitavad ning viite kaudu leiab muutmisotsuse Riigi Teataja Lisast, mis on digiteerituna olemas DIGAR-is.

Perekonnanimede muutmise seadus (RT 1930, 57, 388) andis otsustamise alates 1. augustist 1930 (VT 1934: 214) riigikohtult k o h t u- ja s i s e m i n i s t r i l e.[5] Peale eelmises seaduses mainitud muutmispõhjuste lisati loetellu ka nime lai levik. Kulud ning menetlusaeg jäid varasemaga sarnaseks.

Nimekorralduse seadus (RT 1934, 52, 438) kaotas kaheks aastaks (juunist 1934 juunini 1936) nime eestistamise kulud (need jäid kehtima halvakõlalise nime muutmisel). Seadus lõi võimaluse mitmest nimeteisendist ehk rööpnimest üks valida (varem loeti seda perekonnanime muutmiseks) ning ka nimi kaitseregistrisse kanda.

Soovides sisepoliitiliselt pingsal ajal veelgi hoogustada nimede eestistamist, kehtestati 1934. aastal eelmise seaduse asemel perekonnanimede korraldamise seadus (RT 1934, 91, 735), millega loodi kaheks aastaks (1. jaanuarist 1935 kuni 1. jaanuarini 1937) nimede muutmise lihtsustatud kord. Sellega nähti siseministrile ette õigus anda muutmise otsustamine p e r e k o n n a s e i s u a m e t n i k e l e (mida minister ka tegi: RT 1934, 107, 841), vastuvaidlemise tähtaega lühendati kolmelt kuult ühele ja lisati tasuta muudetavate hulka ka halvakõlalised nimed. Tasuta nimemuutmise tähtaega pikendati hiljem 1. jaanuarini 1939 (RT 1937, 101, 820), seejärel 1. jaanuarini 1940 (RT 1938, 106, 921) ning lõpuks 1. jaanuarini 1941 (RT 1939, 115, 893). Osal juhtudel (vaidluste, rööpnimede kaotamise soovi, laialt levinud nime muutmise soovi jm) jäi otsustajaks siseminister ning vastuvaidlemise tähtajaks kaks kuud.

Sellest tulenevad järgmised ebakõlad andmebaasi andmestikus. Nagu riigikohtu otsustest ei selgu alati isiku elukoha omavalitsus, ei selgu see ka siseministri tehtud otsustest. Näiteks muutis minister Valga-Võru raudteeliinil raudteemajas nr 513b elava *Artemon Muna* (ka *Jaitsov*) nime (> *Kallaste*) või lihtsalt ilma elukohta mainimata otsustati nimemuutus *Rudolf Kassuk* > *Talv*. Siseministri otsustatud nimemuutmised (andmebaasis 1366 kirjet) on siinses analüüsis arvestatud Tallinna alla, kuna selle omavalitsuse andmetes mõjutavad need tulemust kõige vähem (Tallinna kohta on 12 043 kirjet), ning arvestades üldise elanikkonna jaotust eestistamise tippaastal 1936 (Reiman 1936: 174), elaski nendest siseministri otsusega seotud 1366 inimesest 12,7% Tallinnas. Kuna tegu on üsna suure hulgaga, oleks edaspidi andmestike kättesaadavuse paranemisel (eelkõige tollaste perekonnaregistritena) vajalik elukoha omavalitsuste täpsustamine.

Alates muutmise otsustusõiguse üleminekust riigikohtult ministrile (1930) kuni muutmise lihtsustatud korra kehtima hakkamiseni 1. jaanuaril 1935 on andmebaasis 514 kirjet. Neile lisanduvad detsembris 1934 ja jaanuaris-veebruaris 1935 tehtud ja hiljem Riigi Teataja Lisas ilmunud otsused, kuna esimene perekonnaseisuametniku otsus nimemuutmise kohta ilmus alles 1. märtsil 1935 (RTL 1935, 17, 631). Kokku on seega siseministri ainuotsustuse ajal registreeritud 652 nimemuutust.

Nimede eestistamine ei lõppenud 1940. aastaga, nagu seda tavatsetakse arvata ja nagu OA avalik versioon seda kajastab. 1940. aastal alanud esimese Nõukogude okupatsiooni ajal nimede eestistamine esialgu jätkus ja viimased otsused ilmusid 3. detsembril 1940 (ENSV TL 1940, 28, 443). 1. jaanuarist 1941 hakkas Eestis kehtima Vene NFSV abielu-, perekonna- ja eestkosteseaduste koodeks (ENSV T 1940,

[5] 1. aprillil 1934 toimunud ministeeriumide lahutamise järel allus nimemuutmine Siseministeeriumile (RT 1934, 5, 36).

73, 1007) ning sellega katkes nimede muutmine seni kehtinud korra alusel. Saksa okupatsiooni ajal Eestis kehtiva õiguse määrusega (AT 1941, 2, 8) jõustusid taas Eesti Vabariigi seadused ja määrused ning perekonnanimede korraldamise seaduse alusel otsustas nimemuutmisi edaspidi sisedirektor. Perekonnaseisuametnikele enam seda õigust ei antud.

Ametliku Teataja III osas avaldati 8. juunist 1942 kuni 31. juulini 1944 kokku 138 sisedirektori nimemuutmisotsust, kusjuures väljaande viimases numbris ilmus veel seitse muutmise sooviavaldust (AT III 1944, 32, 546). Et otsused ilmusid väljaandes kuu aja pärast või hiljemgi, võidi muutmisotsuseid teha veel augustis ja septembris 1944, aga need jäid väljaande katkemise tõttu avaldamata. Näiteks võttis sisedirektoorium 16. augustil 1944 vastu riigilõivu nimemuutmise *Buschkov* > *Männits* ja 18. augustil 1944 nimemuutmise *Kommusaar* > *Poola* kuulutuse avaldamiseks (RA, ERA.5201.2.47594b; RA, ERA.5201.2.47594d). Sisedirektori muutmisotsused on lisatud OA-sse. Suurem täiendus on 1938. aasta sügise muutmisotsused (1479 kirjet), mis puuduvad siiani OA avalikust versioonist.

Onomastika andmebaasi probleemid

Erinevalt OA perekonnanimede paneku osast (mis oli relatsiooniline andmebaas) oli eestistamise andmestik algselt päringuvormiga tabel. Andmebaasi iga kirje sisaldab nelja välja: endine perekonnanimi, uus perekonnanimi, kuulutuse avaldanud Riigi Teataja Lisa kuupäev, muutmise koht ehk otsuse teinud perekonnaseisuametniku omavalitsus. Kirjeid on algversioonis 78 687, pärast siinkirjutaja täiendusi ja korduvkannete liitmist 76 364. Siinse analüüsi jaoks tehtud korrastamise käigus tehti tabelist relatsiooniline andmebaas, eraldades kohanimed eraldi tabelisse (vt ptk „Haldusjaotuse andmestik"). Samuti loodi lisaväljad, mida on andmete täpsustamise käigus täidetud vastavalt vajadusele. Suuremad muutused olid tegelemine mitme variandiga nimedega, haldusüksustega, samuti andmestiku analüüsil ilmnenud üksikute puudustega. Järgnevalt kirjeldatakse neid probleeme lähemalt.

Nimevariandid, korduvad kanded

Ühel isikul võis esineda mitu erinevat perekonnanime. Reeglina oli rööpnimede puhul üks kasutusel kirikuraamatutes, teine hingeloendites ja nende funktsiooni jätkavates valladokumentides. Perekonnaregistri sisseseadmisega 1926. aastal pidi ametlikuks kasutuseks valitama sünnisissekande ehk kirikuraamatu nimekuju ning seetõttu valladokumentides olnud rööpnimed valdavalt hääbusid. (Puss 2020)

OA algversioonis olid rööpnimedest ning ka ühe nime eri kirjapiltidest moodustatud eraldi kirjed (*Noks* > *Ehasaar* ja *Noksman* > *Ehasaar*). Ühe muutmise kohta oli kuni kuus eri kirjapilti, seega oli kunstlikult juurde tekitatud kuni viis kirjet. Need ühendati nüüd lisavälja abil üheks kirjeks (*Noks* ~ *Noksman* > *Ehasaar*).[6]

[6] Nimetatud töö tegi oma uurimistöö osana Hugo Treffneri gümnaasiumi õpilane Mihkel Paloots 2020. aasta kevadel.

Kaks või enam varianti või mitmiknimi esinevad OA-s 5148 kirjel. Tollal ei olnud mitmiknimed tavaliselt abielunaistel nagu tänapäeval, vaid tegu oli rööpnimedega ja nii on neid siin analüüsis ka käsitletud. Mitme nimevariandiga kirjetest 477 on juhud, kus üks eelmistest variantidest või mitmiknime komponentidest kattub uue nimega (*Pankrot-Pank* > *Pank*) ning seega pole tegu nime eestistamisega. Rööpnimede kaotamisena on siinkohal käsitletud ka neid, kus ükski eelmistest fikseeritud nimekujudest ei kattunud täpselt, kuid on näha soovi kahest nimest üks valida (*Tomson-Meus* > *Meos*), või tehti koos rööpnime kaotamisega ka ilma muutmisprotsessita lubatud õigekirjutuslik parandus (*Hanssen-Tagasar* > *Tagasaar*). Rööpnimede kaotamiseks pole loetud neid, kus peale kahe nime kaotamise nimi ka eestistati (*Käsper-Kesper* > *Käspre*). Üksikud juhtumid jäävad ortograafia tõttu arvutianalüüsil ja silmvaatlusel leidmata, kuigi tegu on samuti rööpnimede kaotamisega, nt *Miilberg* ~ *Miilperk* ~ *Uustallu* ~ *Uustallo* > *Uustalu* või *Jas* ~ *Jazi* > *Jaasi*. Rööpnimede kaotamistest 208 otsustati enne, 269 pärast 1. jaanuari 1935, seega on nende osakaal kõikidest muutmistest enne nimemuutmise lihtsustatud korda oluliselt suurem.

OA algversioonis oli rööpnimede puhul enamasti sisestatud vaid alles jäävaga mittekattuv nimi, st *Aavekukk* ~ *Aurvekukk* > *Aavekukk* sisestati kui *Aurvekukk* > *Aavekukk*. Siinkirjutaja tegi andmekontrollil selliste rööpnimede kohta mitusada parandust. Tõenäoliselt on selliseid kirjeid veel kuni pool tuhat, nendest mõnisada on rööpnimede kaotamise otsused ning seda arvestades peaks nimemuutjate üldarvust maha võtma veel kuni 300. See moodustab vaid 0,4% kirjete koguarvust ega mõjutaks lõpptulemust. Rööpnimede kaotamise kirjed (477) märgistati eraldi ning need on edaspidisest analüüsist välja jäetud.

Kui rööpnimede ja eri kirjapildiga kirjete liitmisel kahanes eestistamiskirjete hulk oluliselt (5148 kirje ehk 6% võrra), siis lahendamata jääb probleem, kus kirjeid peaks olema hoopis rohkem (vt lk 806).

Üksikjuhtumeid

OA töötlusel ning algandmestikuga võrdlemisel ilmnes huvitavaid üksikjuhte, mille tõttu andmebaasi on jõudnud kas ebatäpsused või tuleks neid edaspidi andmete täiendamisel eristada, sest tegu ei olnud endale vana nime asemele uue võtmisega. Nende hulk on minimaalne, kuid need on siinkohal esitatud illustreerimaks, et kogu toimunud protsess ning seda andmebaasina kajastav andmestik ei olnud ühetaoline.

1935. aasta veebruarini on nimed RTL-i kuulutuses käändes (*Koogile*, *Vaesele*, *Erelisele*) ja kuni jaanuarini 1935 mõnikord apostroofita (hiljem olenevalt perekonnaseisuametniku saadetud andmetest enamasti apostroofiga). Seetõttu pole alati õigesti tuletatud nime nominatiivne kuju: näiteks nimest *Sipsakale* tekkis OA algversiooni *Sipsak* pro *Sipsaka*.

Leidub üksikjuhtumeid, kui taastati oma esialgne nimi. 1921. aastal võttis *Johannes Muks* koos perega Narvas sealse nimepanekukomisjoni otsusega uueks nimeks *Roosilill*, kuid 1926. aastal võttis ta nime *Muks* tagasi. Hiljemalt 1935. aastal on

asutud seisukohale, et „perekonnanime saab muuta üks kord" (Siseministeeriumi ringkiri perekonnaseisuametnikele 24. V 1935, avaldatud Mõistlik 2005: 165), kuid siinkirjutajal pole õnnestunud leida sellisele seisukohale juriidilist alust.

Mitmekordset muutmist esines teistelgi põhjustel. Pärast eelmainitud rööpnimede kaotamist mõnikord ka see ainus nimi eestistati. Näiteks *Johannes Nebukat ~ Mihelson ~ Mihkelson* jättis endale 1934. aastal ainsaks nimeks *Mihelson*, kuid juba aasta hiljem eestistas ta selle nimeks *Mihklisoo*. Sellised nimemuutjad koos oma perekonnaliikmetega kajastuvad nimemuutjate hulgas kaks korda. Siiski on nende hulk väga väike ning praeguste andmestike põhjal pole neid ka võimalik leida.

Harvadel juhtudel võeti endale (tagasi) ka rööpnimed ja hiljem neist loobuti. Näiteks 1926. aastal võttis *Emilie-Rosalie Kallus* pärast lahutust tagasi oma eelmise, surnud mehe nime *Leu (Lõiv)* ja 1939. aastal jättis neist alles vaid nime *Lõiv*.

Mõned nimemuutmised viidi küll lõpule, kuid muudetud nime ei hakatud kasutama. Roela valla perekonnaseisuametnik määras 13. V 1940 perekond Voltenbergile uueks perekonnanimeks *Kaldoja*, kuid nad elasid edasi Voltenbergi nimega (Haudi, 10. V 2021). Palju sagedamini esines seda Eestist lahkunute puhul.

Senine nimi vahetati mõnikord oma varasema (mitte viimase) abikaasa nime vastu (*Tiits* > *Kõiv*), võeti endale poja nimi (*Grauding* > *Falk*), üsna sageli võeti tagasi enda neiupõlvenimi (*Kiiranen* > *Sepp*, *Huene* > *Aksenova*, *Majer* > *Alver*) või ühiselt naise neiupõlvenimi (*Titma* > *Toodu*), samuti ema neiupõlvenimi (*Unt* > *Aus*). Korduvalt muutis ema oma alaealise lapse nime enda nimega samaks (*Tillart* > *Endrikson*), samuti täisealine vallaslaps võttis ema abielunime (*Martinson* > *Sommer*). Muudeti enda nimega samaks ka oma kasulapse ehk mitte lapsendatud lapse nimi (*Poker* > *Juurikas*, *Juhanson* > *Riikson*), võeti ka lahutatud abikaasa nimi (*Tarkpea* > *Padalka*).

Ühelgi eelnimetatud juhul ei olnud võetud nimi uus, vaid perekonnas varem olnud või tolgi hetkel eksisteeriv nimi. Kui kasulaste või vallaslaste nime muutmine oli pigem erand ning seda tehti tõenäoliselt peamiselt 1920. aastatel, siis abielupaari puhul naise neiupõlvenime või emma-kumma abikaasa ema neiupõlvenime uueks nimeks võtmine on huvitav uurimisteema, milleks saab kasutada Siseministeeriumis juba sisestatud, kuid avalikkusele seni ilma juurdepääsuta 1926.–1949. aasta perekonnaregistri andmebaasi. Selgi juhul oleks tulemuslik kasutada laiendatud lähenemist, näiteks sarnaseid nimesid tuvastavate algoritmide abil. Nii võib leida ka juhtumi, kus *Voldemar-Gustav Korner* võttis endale uueks nimeks *Lillepott*, sest tema ema neiupõlvenimi oli *Lilleput*.

Haldusjaotuse andmestik

Nimede eestistamise ajal muudeti halduspiire ning toimus vallareform. Aastail 1922–1934 kaotati üheksa valda, 1. aprillil 1939 aga 117 valda. Kaotatud valdade perekonnaseisuraamatutes sündide, abielude ja surmade registrid lõpetati, kuid perekonnaregistreid polnud võimalik jaotada mitme valla vahel ning need läksid üle ühele uuele vallale (Teder 1939: 141–142). Perekonnaregistrite liitmine (andmestikku vt Teder 1939: 367–379) on hea alus nimemuutmiste analüüsiks enne ja

pärast vallareformi, sest samal põhimõttel on arhiivis korraldatud ka nimemuutmistoimikud.

Probleeme on aga samanimeliste omavalitsustega, millele on tähelepanu juhtinud OA koostaja (Must 2000: 78–79). Ühelt poolt on tegu eri kohas asuvate omavalitsustega (näiteks Roela vald Tartumaal ja Virumaal), kuid teiselt poolt on andmesisestusel jäänud eristamata samanimeliste valla- ja alevi- või linnaomavalitsuste andmed, näiteks on Võru linna ja Võru valla andmed kõik Võru nime all.

Ajaloolise haldusjaotuse geoinfona on maa-ameti veebilehe kaudu kättesaadavad muuhulgas ka 1938. aasta omavalitsuspiiride andmed.[7] Tegu on 1990. aastate teisel poolel koostatud andmetega, mis vajaksid praegu aga parandamist (*Väinvere* pro *Väinjärve*, *Kõrgessaare* pro *Kõrgesaare* jt) ning ka geoinfoandmete täpsustamist. Nii on seal kuus juhuslikku linna (nt Tallinn ja Kärdla) esitatud valdade kihis territooriumina, ülejäänud 12 linna vaid punktina, Otepää linn puudub andmestikust sootuks. OA ning olemasolevate ajalooliste geoinfoandmete ebaühtlus on üks põhjusi, miks artiklis pole kasutatud geograafilist visualiseerimist.

Andmebaasi korrastamise tulemus

Olles andmebaasi täiendanud 1938. aasta sügise ja 1942.–1944. aasta andmetega ning parandanud tuhandeid kirjeid, kontrollinud silmvaatlusega andmebaasi eri versioonides esinevaid mittekattuvaid kirjeid (mis osutusid enamasti sisestusvigadeks), ühendanud rööpnimedega ja eri kirjapildiga kirjed, kujunes kirjete koguhulgaks 76 364. Neist rööpnimede kaotamiseta ehk tegelikke nimede eestistamisi on 75 887 kirjet, mis on edaspidise analüüsi aluseks. Järgnevalt kasutatakse saadud andmeid, et korrigeerida nimemuutjate üldarvu ning saada ülevaade soovitusnimestiku kasutamisest ja nimevõtjate maitse-eelistustest.

Nimemuutjate üldarv

Nimemuutjate üldarvu kohta oli pikka aega käibel vaid osaline statistika ning kuigi suurusjärk on olnud sama, on põhjust seda uute võimaluste avanemise järel täpsustada.

1938. aastal avaldas Elmar Elisto (1938: 52) Nimede Eestistamise Keskbüroo 1937. aastani kogutud andmed. Ta tõdeb, et esiteks pole teada nime eestistajate arv (st eristatuna halvakõlaliste nimede jmt muutmisest) ning teada on vaid sooviavalduste, mitte muutmisotsuste arv. Tegu on üldjuhul kattuvate arvudega ning nagu esitatud statistikast näha, on aastate 1935–1937 andmetel saadud isikute arv sooviavalduste arvu korrutamisel 2,5-ga, kuna see oli võetud keskmiseks ühise avalduse esitanud isikute arvuks. Täpne tundub olevat vaid aastail 1920–1934 nime muutnute arv (2672).

[7] https://geoportaal.maaamet.ee/est/Ruumiandmed/Haldus-ja-asustusjaotus/Ajalooline-haldusjaotus-p827.html (10. V 2021).

Keskbüroo jõudis koostada ka statistika 1. jaanuarini 1940. Selles on esitatud nimede eestistamiste arv, seega jääb selgusetuks, kas see hõlmab kõiki nimemuutmisi. Nende andmete järgi eestistasid oma nime 1935.–1939. aastani 196 482 isikut (RA, ERA.14.2.700, l 288; sama RA, ERA.1068.1.1, l 4). Kuigi see ei jagu täpselt 2,5-ga, ei ole tõenäoline, et tolleks ajaks oli olemas täpne statistika isikute tasemel, vaid sooviavalduste arv korrutati 2,5-ga ning ümardati täisarvuni. 22. märtsil 1938 kinnitas Nimede Eestistamise Liit Siseministeeriumile, et „isikute arvutamisel on iga sooviavalduse kohta arvestatud 2,5 isikut (mis on üleriigiline keskmine)" (RA, ERA.14.2.369, l 8). Eeltoodud arvust 196 482 puuduvad ka enne 1935. ja pärast 1939. aastat muudetud nimede andmed. Ühe nimemuutmise keskmiseks isikute arvuks peeti varem isegi kolme ning selle põhjal tehti ka statistikat (Salurand 1935).

Siseministeerium hindas 1. aprilliks 1939 nime muutnute arvuks 181 000 (VT 1939: 101). Aprillist detsembrini 1939 ilmus 3378 nimemuutmisotsust (maksimaalselt 8500 inimest). Seega oli 1939. aasta lõpuks muutjaid kuni 190 000, mitte aga 199 154, nagu see oleks keskbüroo statistika põhjal (liites juurde kuni 1935. aastani nime muutnute arvu).

Arvestuse pooleli jäämine on mõjutanud ka hilisemaid hinnanguid. Inda Rajasalu on taasiseseisvunud Eestis ilmunud esimeses teemakohases kirjutises esitanud 1. detsembrini 1939 nime muutnute arvuks 195 000, toetudes teksti järgi keskbüroo tabelile, kuigi viited on kuni 1938. aastani ilmunud allikatele (Rajasalu 1993: 104, 106).

ENE ja EE esitavad aastail 1935–1940 nime eestistanute arvuks „umbes 211 000 isikut" (ENE 1974: 83; EE 1994: 255). Aadu Must (2010: 399, 2015: 66) on hiljemalt 2002. aastast toonud käibele nimemuutjate arvuna 210 000. Entsüklopeedia „Eesti. A–Ü" esitab arvatavasti Musta tsiteerides arvuks „ligi 210 000" (EE 2003: 409). Kõige täpsema arvu (213 670) toob Ilona Raidmaa (1999: 10), paraku pole ka koos temaga õnnestunud siinkirjutajal selle allikat tuvastada. Sageli kasutatakse viimast kaasaegset statistikat („ligi" või „umbes" 200 000), mis ulatus 1939. aasta lõpuni (Karjahärm, Sirk 2001: 288; Zetterberg 2010: 434; Raun 2012: 101–102).

Nagu eespool mainitud, on nüüdseks arhiiviinfosüsteemis AIS kättesaadavad Rahvusarhiivi (fondi ERA.5201) jõudnud nimemuutmistoimikute (n = 91 286) kirjeldusandmed. Need on sisestatud aga arhiivikirjelduse, mitte andmetöötluse eesmärke täites tekstina, kuid sisaldavad ka kõikide nimemuutjate eesnimesid. Säilikute tüüppealkirjade tõttu oli siinkirjutajal võimalik teha tekstist tabel, et seda statistiliselt analüüsida. Selleks kanti iga isiku eesnimi ja perekonnanimed eraldi lahtrisse.

Andmeid võrreldes selgus, et toimikuid on ka muutmistoimingute kohta, mis ei jõudnud otsuseni. Näiteks soovis 1935. aastal perekond Unujaan võtta uueks nimeks Ariste ning perekond Vänzel uueks nimeks Ventsel, kuid kummalgi seda ei võimaldatud (RA, ERA.5201.2.80202). Asjaajamistoimik on olemas, pealkirjast aga ei selgu, kas muutmisavaldus rahuldati. Selliseid toimikuid on hinnanguliselt 7500.

Kõiki nimemuutmistoimikuid selles arhiivifondis pole. Siinkirjutaja on toimikuid leidnud ka mujalt, sh üldtoimikutest (nt RA, ERA.14.2.700; RA, ERA.14.2.1262), kaitseregistrisse kandmise toimikute vahelt (RA, ERA.14.2.288) jm. Enim on puudu 1940. aastal lõpetatud toimikuid, kuna nähtavasti polnud neid veel Siseministeeriu-

misse saadetud ja need hävisid sõja ajal. Järelikult ei saa säilinud toimikute põhjal teha täpseid järeldusi nimemuutmiste arvu kohta.

Küll aga saab toimikute andmeid kasutada ühe muutmistoiminguga seotud keskmise isikute arvu teada saamiseks. Ühes toimikus on 1–22 isiku andmed. 22 isikuga on üks (*Wassikas* > *Heinlaid* (24. XI 1926 Pärsamaa) hõlmab nelja põlvkonna esindajaid), 14 isikuga kaks, 12 isikuga 13 toimikut. Kõige enam on ühe (39 487, 43,3% toimikutest) ning kahe (17 580, 19,3%) isikuga toimikuid. Kokku on toimikutes 215 580 isiku andmed. Kõige rohkem isikuid on kolme (41 826, 19,4% isikutest) ja nelja (41 268, 19,1%) isikuga toimikutes. Ühe toimiku keskmine isikute arv on 2,36.[8] Arv on oluliselt väiksem nimemuutmise ajal keskmiseks peetud 2,5-st. Näha on ka selge langustendents aastati, arvestades toimiku algdaatumit (vt tabelit 1).

OA kirjete arvu kasutamine muutjate üldarvu teada saamiseks ei annaks täpset tulemust, kuna ühe perekonnaseisuametniku poolt samas RTL-i numbris avaldatud sama vana ja uue nimega kuulutused on sisestatud ühe kirjena. Näiteks võisid korraga nime muuta täisealised õved (ka koos peredega) ning nende nimemuutmine kajastub eraldi otsustena, kuid andmebaasis ühe kirjena. Koosmuutmisi toimus lausa suguvõsati. Näiteks leidub andmebaasi ühe kirje *Vilberg* > *Vilbaste* (4. X 1935 Kodasoo) taga 15 kuulutust 39 inimesega, kirje *Laisk* > *Laaneväli* (13. VI 1922 Hellamaa) taga on viis kuulutust 27 inimesega.

Puuduolevate kirjete hulga arvutamiseks võrreldi aastate 1935–1940 igast kvartalist ühe RTL-i numbri otsuste ning sama numbri kirjete arvu OA-s (v.a 1938. aasta IV kvartal, kus otsuste ja kirjete arv on andmete hilisema sisestuse tõttu sama). 5445 otsuse kohta oli OA-s 4818 kirjet, puuduolevate kirjete hulk erines aastati (vt tabelit 1).

Nimemuutjatest tuleb eraldada rööpnimede kaotajad ehk 477 kirjet (olenemata sellest, et nad loeti omal ajal nimemuutjate arvestusse). Tuletades OA kirjetest otsuste arvu (ilma rööpnimede kaotamisteta) ning korrutades selle iga aasta keskmise isikute arvuga otsuse kohta, saab üldtulemuseks 197 153. Kuna praeguste andmestike põhjal puudub võimalus maha arvestada kõiki rööpnimede kaotamisi, samuti neid, kes võtsid tagasi oma neiupõlvenime, andsid oma kasu- või vallaslapsele enda perekonnanime jmt, on põhjust oma perekonnanime eestistanute arvuna kasutada ümardatult 195 000, mille kaasaegne statistika ületas juba 1. jaanuariks 1940. See on ka 15 000 võrra ehk 7,1% väiksem ligi pool sajandit kasutatud arvust 210 000 (nt EKK: 618).

Tabelis 1 on toodud eelnimetatud arvutused ning võrdlusandmed kaasaegse statistikaga. Autoril on õnnestunud leida vaid üks, Edgar Rajandi koostatud mustanddokument 1940. aasta kevadest, kus on toodud andmed 1939. aasta lõpuni ning seda nii sooviavalduste kui ka otsuste kohta.[9] Andmed pole aastati küll täpselt võrreldavad, sest otsused avaldati RTL-is viivitusega ning andmebaasis on avaldamise kuupäev. Andmeridu on seega kolm: sooviavaldused, otsused ning otsuste avaldamised.

[8] Täpne tulemus on 2,36161. Üldarvu arvestamisel tabelis 1 on lähtutud iga aasta keskmisest eraldi, mujal aga kõikide aastate keskmisest.

[9] Dokument „Perekonnanimede ja nimede eestistamise käik aastate järgi" kaustas „Nimede Eestistamise Keskbüroo mitmesugused materjalid", Edgar Rajandi perekonna valduses. Tänan siinkohal Rajandi järeltulijaid Merike Kallast ja Maiu Pevkurit materjalide kasutamise võimaluse eest.

Erinevus arvutusliku ja tegeliku tulemuse vahel perioodi 1920–1934 kohta on selgitatav sellega, et arvutuslikust tulemusest on välja jäetud rööpnimede kaotamise otsused, mis mõjutavad selle perioodi lõpptulemust oluliselt. Aastail 1935–1936 on näha sooviavalduste palju suurem hulk võrreldes muutmisotsustega ning on loomulik, et just sel põhjusel kasutati avalikkuses neist esimest arvu argumendina kõhklejate ärgitamiseks. Alles hiljem hakati avaldama tõepärasemat otsuste statistikat. Siiski jätkus isikute arvu saamiseks otsuste üldarvu korrutamine 2,5-ga.

Tabel 1. Nimemuutmiste arvutuslik üldarv ja võrdlus 1940. a-ni tehtud statistikaga.

Aasta(d)	Kirjete arv OA-s	Kirjeid ilma rööpnimede kaotamiseta	Hinnanguliselt puuduvaid kirjeid	Tõenäoline otsuste arv	Ühe muutmistoimiku keskmine isikute arv	Nimemuutjate arvutuslik hulk	1940. aastani peetud statistika	
							Sooviavaldusi x 2,5	Otsuseid x 2,5
1920–1934	1053	845	11,4%	954	2,16	2060	2672	2672
1935	10 544	10 487	10,8%	11 760	2,57	30 223	34 196	30 000[10]
1936	21 832	21 804	13,1%	25 086	2,48	62 214	78 164	62 753
1937	18 517	18 492	12,1%	21 046	2,24	47 144	43 069	50 518
1938	12 901	12 855	9,7%	13 963[11]	2,14	29 881	30 225	35 845
1939	5937	5877	11,9%	6668	2,02	13 470	15 313	15 740
1940	5433	5409	9,6%	5985	2,00	11 971	-	-
1942–1944	147	118	0,0%	118[11]	1,61	190	-	-
Kokku/ keskmine	**76 364**	**75 887**	**11,7%**	**85 581**	**2,36**	**197 153**	**203 639**	**197 528**

Nimemuutmiste analüüsil on võimalik lähtuda ühelt poolt nimevahetajate arvust, teiselt poolt vahetatud ja võetud nimede arvust. Keeleliseks analüüsiks on enamasti otstarbekam näiteks 52 muutmiskorra *Steinberg* > *Kivimäe* käsitlemine ühe muutusena. Seda on rakendatud siin töös soovitusnimestiku kasutuse analüüsil. Nimede esitähtede analüüsil on aga olulisem vaadelda muutmiste arvu ehk mitu korda muudeti *k*-algulised nimed samuti *k*-algulisteks, mitte aga seda, mitu *k*-algulist nime muudeti samuti *k*-alguliseks. Muutmiste arvust on allpool lähtutud esitähtede, nimepikkuste ja ainukordsuse analüüsil.

Soovitusnimestik „Eesti nimi" ja võetud nimed

Eestistamise algusest saadik üritasid algul keeleteadlased, hiljem ka riik suunata uue nime valikut. Selleks koostati soovitusnimestikke, loodi kaitseregister ning piirati laialt levinud nimede võtmist. Siinses peatükis on vaadeldud 1935. aasta soovitusnimestiku „Eesti nimi" sisu ning selle mõju võetud nimede valikule.

[10] Ligikaudne, täpne statistika puudus.

[11] 1938. aasta IV kvartali ja aastate 1942–1944 otsuste ja kirjete arv on vastavuses, seega pole nendele andmetele puuduvate kirjete hulka lisatud.

Üks seadustes loetletud nimevahetuse võimalikke põhjuseid oli alates 1930. aastast nime lai levik. Siseminister andis 14. detsembril 1934 välja laialt tarvitusel olevate perekonnanimede määruse (RT 1934, 107, 842) 188 eestipärase nimega, mida ei tohtinud endale uueks nimeks võtta. Mitte-eestipäraseid nimesid ei tohtinud eestlased nagunii võtta ning teiste rahvuste nimevaramuga tegelemist eesmärgiks ei seatud. Nimekirja kanti tol hetkel nimed, mida esines Tallinnas ja Tartus enam kui sajal kandjal (Lepp 1936: 265).

Loetelu täiendati 1935. aastal 19 nimega (RT 1935, 98, 806) ja 1938. aastal 82 nimega (RT 1938, 97, 840). Kokku oli nimekirjas 1938. aastaks 289 nime. Samas oli üle 100 kandjaga eestipäraseid nimesid 1934. aastal 1127 ja juba 1936. aastal leiti, et keelunimekirja tuleb täiendada, sest laialt levinud nimesid võeti agaralt endale uueks nimeks (Lepp 1936: 265). Põhjusena, miks nimekirja ei täiendatud, võib näha soovi mitte takistada kampaania käiku ja samuti lootmist soovitusnimestikule.

Esimene soovitusnimestik oli 1921. aastal ilmunud „15.000 uut sugunime" (US 1921), mida kasutati nimepanekul Petserimaal ja Narvatagusel ning mida tõenäoliselt sel põhjusel ning ka maitse muutumise tõttu peeti poolteist aastakümmet hiljem vananenuks. See sisaldab ka üsna palju ebaõnnestunud nimesid (*Karvastik*, *Pihkupere*, *Tusaste*, *Õngekalju*, *Õudusmaa*, *Üüritar*). 1935. aastal hinnati, et see oli „erakordse kiirusega koostatud ja on olnud ainult „hädalist tarvidust rahuldama ruttav katse", [---] sisaldades tunduval määral vähesoovitavat nimeainestikku" (EN 1935: 26). Paraku on väljaanne trükitehniliselt kesise kvaliteediga (mimeograafiline paljundus) ning seetõttu pole ka selle tärktuvastus arvutianalüüsiks kasutatav.

1935. aastal ilmus valimik „4000 uut nime. Kuidas leida uut nime" (UN 1935). Sellest avalik tärktuvastusversioon puudub, kuid ka selle olemasolul poleks see põhjaliku eeltöötluseta kasutatav, kuna seal on eri variandid esitatud liitekohta märkimata: *Avu(vur)*, *Paiga(gaste)*, *Palmik(vik)*. Raamat ilmus umbes 20. jaanuaril 1935, sest päev hiljem kirjutas sellest kirjastaja ajaleht Oma Maa. Eestistamist oli ajalehe arvates rahva väitel pidurdanud nimemoodustusjuhiste puudumine, sest „tahetakse ju, et nimi oleks ilusakõlaline ja sobiv", ning just neid uus brošüür pakkuvat (Põhjaklane 1935). Raamat ilmus Viljandis ning seal oli sellel kahtlemata oma mõju, mida on edaspidi põhjust uurida. 9. märtsil 1935 kirjutas Postimees: „Tartlased ja tallinlased ammu juba koostavad kahasse uute eesti perekonnanimede valikut: Tartlased teevad nimed, Tallinn maksab kinni kulud ja kirjad. Viljandlastel aga on oma „4000 uut nime" juba ammu valmis. [---] See raamat on igal viljandlasel taskus." (Andermann 1935)

Tartlaste ja tallinlaste nimetatud koostööna ilmus 15. märtsi paiku 1935 (RA, ERA.1093.1.219, l 7) Akadeemilise Emakeele Seltsi toimetiste sarjas Julius Mägiste ja Elmar Elisto toimetatud väljaanne „Eesti nimi. Valik uusi perekonnanimesid" (EN 1935). Nimestiku tärktuvastusversiooni kasutades avati analüüsi jaoks kõik välja pakutud variandid (*Saulep*, *-i* > *Saulep* ja *Saulepi*, *Salukan|d*, *-nu* > *Salukand* ja *Salukannu*), koos nendega selgus nimekirjas olevat 15 595 nime. Kõige pikem välja pakutud nimi on 12-täheline (*Haavistniidu*), 11-tähelisi on juba 21 (*Kloostremaa*, *Kaasikniidu*) ning 10-tähelisi 578. Soovitatud nimede keskmine pikkus on 7,3 tähte

(kvartiilid[12]: Q_1=6, Q_2=7, Q_3=8). Kuni seitsme tähe pikkuseid nimesid on 52,8% (n = 8239). Vähemalt 10-tähelised nimed (n = 600) moodustavad suhteliselt suure osa (3,9%). Esitähtedena on kasutatud 20 tähte, eesti omasõnade tähtedest on välja jäetud *b*, *d*, *g*. Kõige enam (17,8%) on välja pakutud *k*-algulisi nimesid (n = 2792), järgnevad *p*-algulised (n = 1500, 9,6%).

Analüüsiks kõrvutati soovitusnimestiku nimesid alates 16. aprillist 1935 ilmunud muutmisotsuste nimedega. Sooviavaldus pidi olema ilmunud vähemalt kuu aega varem ehk hiljemalt 15. märtsil 1935, seega enne soovitusnimestiku trükkimist. Tegelikult ilmusid ka muutmisotsused mõnevõrra hiljem. Näiteks 18. märtsil 1935 tehtud muutmisotsus *Ende* > *Sammalkivi* ilmus alles 12. aprillil. Nimi *Sammalkivi* on soovitusnimestikus, kuid kuna sooviavaldus ilmus juba 16. veebruaril, ei saanud nimi olla sealt võetud.

Kuni 15. aprillini 1935 võeti kokku 205 korral 151 nime, mis esinevad samuti soovitusnimestikus. Tegu on sageli kohanimedega (*Aimla, Maidla*, *Osula*), *ste*-lõpuliste nimedega (*Aunaste*, *Marguste*, *Tõniste*), loodussõnavarast tuletatud liht- (*Kalle*, *Lehis*, *Luide*) ja liitnimedega (*Mäepõld*, *Põldmaa*), mille väljamõtlemiseks ei olnud vaja soovitusnimestikku. Kõikidest sama perioodi muutmiskordadest moodustavad need 9%.

Alates 16. aprillist 1935 võeti kokku 73 591 korral 15 397 erinevat uut nime. Neist 30,5% (n = 4695) esinevad soovitusnimestikus. Kuna vaadeldud perioodil võetud nimede hulk on üsna lähedane soovitusnimekirja nimede hulgaga, on sarnane ka vastupidine näitaja: nimekirjas olevatest nimedest läks kasutusele 30,1%.

Nimestikust võetud nimede keskmine pikkus on 6,9 tähte ehk välja pakututest veidi vähem, kvartiilid aga kattuvad (Q_1=6, Q_2=7, Q_3=8). Siiski moodustavad kuni seitsme tähe pikkused võetud nimed (n = 3131) 66,2% ehk soovitatutest palju enam. Üle 10-tähelisi nimesid ei võetud, 10-tähelisi võeti 77 korral (1,6%). Seega eelistati soovitusnimestikust võtta veidi lühemaid nimesid (vt joonist 2).

Kõikide algustähtedega nimesid võeti soovitusnimestikust ka kasutusele, kõige enam välja pakutuga sarnaselt *k*-algulisi (n = 676, 14,3%), teisena aga *l*-algulisi (n = 449, 9,5%), *p*-algulised jäid kolmandaks (n = 411, 8,7%). Siintoodud analüüs põhineb võetud nimede koondnimekirjal, kus iga nimi esineb vaid korra. Muutmiskordi arvestades on pilt veelgi ilmekam.

Soovitusnimestikuga kattuvaid nimesid võeti 27 263 korral ehk 37%-l muutmistest. Üle 100 korra võetud nimed on tabelis 2. Üle 50 korra võeti 72 nime, ühel korral 1741 nime (37% nimede koguvalikust).

Muutmiskordi arvestades on nimede keskmine pikkus ikka 6,9, samad on ka kvartiilid. Nimede esitähtede järgi on esikohal endiselt *k*-algulised nimed (n = 3580, 13,1%), kuid teisel kohal on *r*-algulised (n = 2980, 10,9%), mis soovitustes on viiendal (8%) ning võetud nimede loetelus neljandal (8,5%) kohal. Kolmandal on muutmiskordi arvestades *l*-algulised (n = 2640, 9,7%). Seda selgitab esiteks üksikute *r*-alguliste nimede populaarsus (vt tabelit 2) ning teiseks populaarsete sõnadega *raud*, *rand*

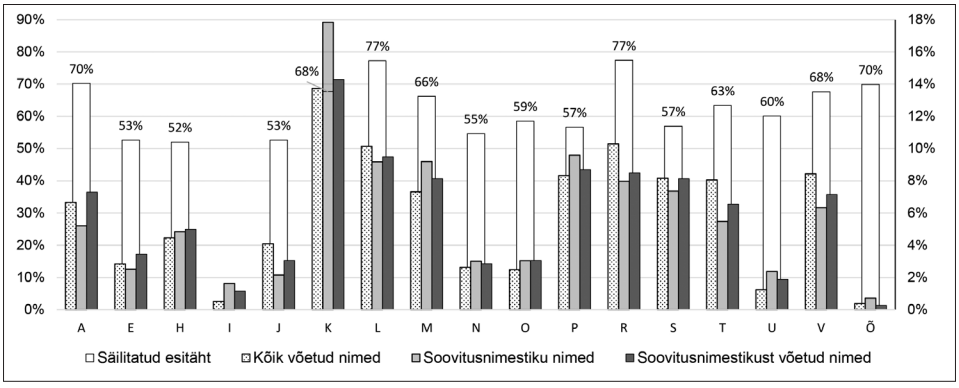
[12] Kvartiilid on tunnuse väärtused variatsioonireal, mis jagavad variatsioonirea neljaks ligikaudu võrdseks osaks. Näiteks Q_1 on tunnuse väärtus, millest väiksemaid või võrdseid tunnuseid on ligikaudu 25%.

ja *rehi ~ rei* algavate nimede suur hulk (*Ran-* 665, *Raud- ~ Raua-* 168, *Rei-* 528) ning *r*-alguliste nimede tugevana tajutav kõla. Võrdluse soovitusnimestiku, sealt kasutusele võetud nimede ning kõikide võetud nimede esitähtede kohta annab joonis 1.

Soovitusnimestiku kasutamist on põnev vaadelda ka omavalitsuseti. Nimestikuga kattuvate uute nimede osakaal neis on 0–83%. Ükski nimi ei kattu viies omavalitsuses, kuid kõigis nendes on nimemuutmiste koguarv ühest kuueni ning nimestikuga mittekattuvus võib olla juhus ega näita nimestiku teadlikku vältimist. Küll aga võib väita, et nimestikku peaaegu ei kasutatud suuremate nimemuutmiste arvuga Kärgula (1 nimi 12-st ehk 8%), Taheva (3/28 ehk 11%), Kuusalu (5/45 ehk 11%) ega Muhu (1/9 ehk 11%) vallas. Nimestiku kasutamist on näha selgemini, kuna tipus on muutmiste arvud suuremad: Põdrala (7/10 ehk 70%), Maasi (80/107 ehk 75%), Puhja (12/15 ehk 80%), Põide (13/16 ehk 81%) ja Kalda (19/23 ehk 83%) vald. Enamikus omavalitsustes kasutati nimestikku kolmandiku jagu (Q_1=32%, Q_2=37%, Q_3=43%).

Aastate lõikes on näha nimestiku mõju kerget suurenemist. Kui 1935. aastal võeti soovitusnimestikus esinev nimi 30,5%-l kõigist muutmiskordadest, siis 1939. aastal oli protsent juba 39,2 (vt joonist 3). Tõenäoliselt muutus vähelevinud nime soovijal sellise nime leidmine pidevalt raskemaks ning soovitusnimekirja abi tuli kasutada rohkem kui kampaania algusaastal.

Soovitusnimestiku „Eesti nimi" koostanud keeleteadlased Julius Mägiste ja Elmar Elisto olid seega pärast nimestiku ilmumist 37% ehk umbes 65 000 nimemuutja „ristiisad", nagu seda öeldi Villem Ernitsa ja Samuel Sommeri kohta Petserimaa ja Narvataguse nimepanekul. Viimati mainitud protsess oli küll pigem suunav ja eestistamine rohkem vabatahtlik, kuid nagu näha, oli soovitusnimestiku abi kasutamine massiline. Kindlasti ei võtnud kõik soovitusnimestikus leiduva nime saanud seda endale otse raamatust, vaid võisid seda teha sugulaste, tuttavate, kohanime või ajakirjandusest loetu eeskujul, kuid nimestiku tähtsus kogu protsessis on ilmselge.



Joonis 1. Esitähtede säilitamine nimemuutmisel (primaarteljel) ning soovitusnimestiku „Eesti nimi", sealt kasutusele võetud nimede ja kõikide võetud nimede esitähtede osakaal (sekundaarteljel). Välja on jäetud esitähed, millega algavaid nimesid võeti alla 100 korra (*b*, *d*, *f*, *g*, *š*, *z*, *w*, *ä*, *ö*, *ü*), ning need, mida säilitati nimemuutmisel alla 50% juhtudest.

Võetud nimede esitähed, pikkus ja ainukordsus

Siinses peatükis uuritakse, kui levinud oli nime esitähe säilitamine, samuti milline oli kõige „parema" pikkusega nimi ning kas eelistati nimesid, mida võtsid paljud, või neid, mida kellelgi teisel polnud.

Aadu Must (2000: 64) on oma loodud andmebaasi põhjal toonud üldistuse, et umbes pooltel kordadel oli uuel nimel vanaga sama esitäht. Selgub aga, et alliteratsiooni ja assonantsi kasutamine oli nimemuutmiste puhul täheti üsna erinev. 75 887 nimemuutmisest kattub uue nime esitäht vanaga 43 064 korral (56,8%). Üsna võrdne on jagunemine alliteratsiooni (56,4%) ja assonantsi (59,3%) vahel. Enim on sama esitähe säilitanud *r*- (77,4%), *l*- (77,3%), *a*- (70,2%) ja *õ*-algulised (69,9%) nimed, kõige vähem eesti omasõnade tähtedest aga *b*- (0,1%), *d*- (0,2%) ja *ö*-algulised (13,3%), nendega kaasneb ka võõrnimede täht *f* (0,2%). Alla 50% on kattuvus veel vaid esitähtedel *i*, *ä* ja *ü*.

Kronoloogiliselt analüüsides on näha selge suund esitähe säilitamisele: kui 1935. aastal kattus see 52,3%-l juhtudest, siis 1940. aastal 59,9% ning tõus on ühe väikese erandiga (1939. a) pidev (vt joonist 3). Veelgi ilmekam on populaarseima säilitatava *r*-tähe kasv samadel aastatel 74,4%-lt 82,8 protsendini, samuti ühe erandiga (1939. a) pidevas tõusus.

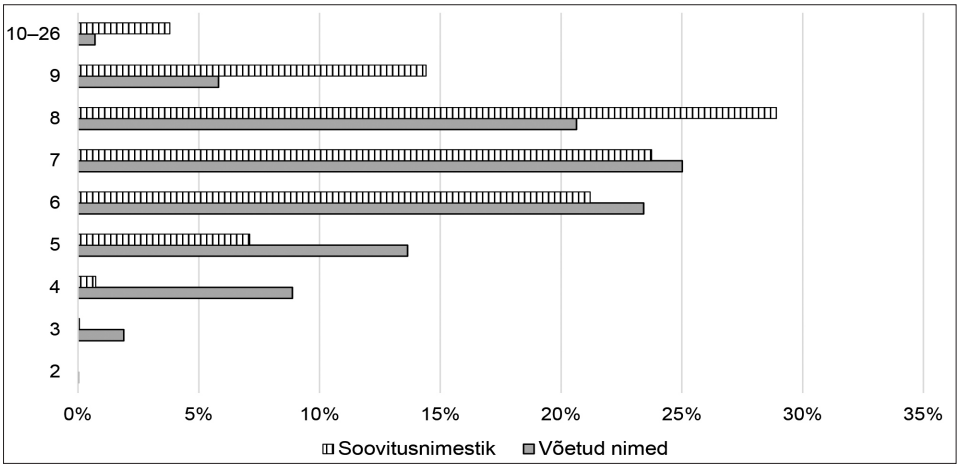
Piirkonniti on erinevaid tendentse. Olgu siinkohal toodud vaid Tallinna (60,9%) üle keskmise ning Tartu (54,5%) ja Narva (51,3%) alla keskmise näitajad esitähe säilitamisel.

Olenemata vanast nimest ja soovitusnimestiku eeskujuna kasutamisest, võeti uueks nimeks kõige enam kordi *k*- (13,7%), *r*- (10,3) ja *l*-algulisi (10,2%) nimesid. Kokku moodustavad need üle kolmandiku toimunud nimemuutmistest. Joonisel 1 on näidatud esitähed, mida säilitati üle 50%-l juhtudest, samuti enim kasutatud esitähtede osakaal.

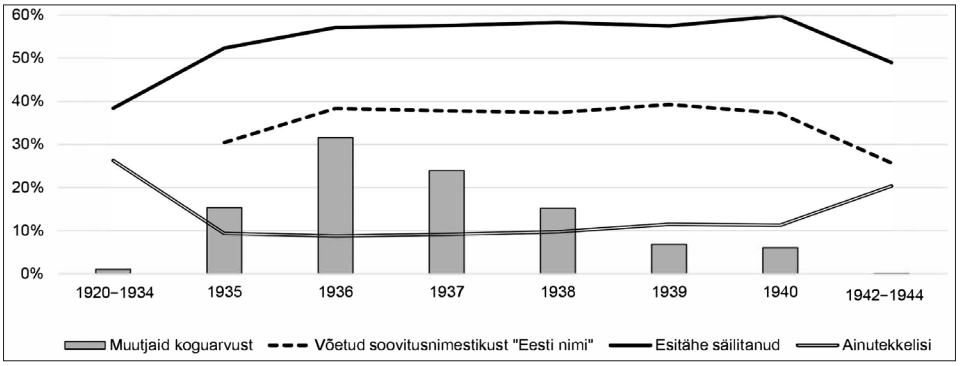
Pikkuse poolest eelistas 25% muutjatest endale võtta 7-tähelise, 23,4% 6-tähelise ja 20,6% 8-tähelise nime (vt joonist 2). Need kolm hõlmavad 69,1% kõikidest nimemuutmistest. Pikimad olid mitte-eesti nimed *Kessler Rossillon en Bugey* (26 tähte), *Popoff-Beboutoff* (16) ja *Bever-Bodenburg* (15), eesti nimedest *Tofer-Toriveri* (14). Kõiki neid ei tuleks lugeda mitte sisuliseks, vaid pelgalt vormiliseks muutmiseks, kuna osa nimest oli varemgi kasutusel. Pikim uus eesti nimi oli *Tammesaarväli* (13 tähte, võeti ühel korral), lühima nimega olid viis nimemuutmist (nimed *Ao* ja *Oa*).

Monofüleetilisi ehk ainutekkelisi nimesid on keeruline tuvastada, kuna võtmise kordade arv seda ei näita. Näiteks võeti kahel korral nimi *Kernumees* (Amblas ja Tallinnas), kuid et selle võtsid isa ja poeg perega (endine nimi *Kirnmann*), on nimi ainutekkeline. Nime *Surve* võeti samuti kahel korral (Kaiu ja Haljala vallas), kuid juba muudetavad nimed (*Klemmer* ja *Seks*) näitavad, et nimi *Surve* on polüfüleetiline ehk mitmiktekkeline. Seetõttu saab ilma mikroanalüüsita ainutekkelisteks pidada vaid ühel korral võetud nimesid, mida on 7361 (9,7% kõikidest muutmistest), nende osakaal aasta-aastalt kergelt tõusis (vt joonist 3). 42 omavalitsuses ei võetud ühtegi ainutekkelist nime. Ainutekkeliste nimede suure hulgaga paistavad silma

Petserimaa omavalitsused. Neist, kus on üldse võetud vähemalt kümme nime, ongi ainutekkeliste nimede poolest esikohal Roodva (ainutekkelisi 6 ehk 28,6%) ja Saatse vald (4 ehk 28,6%) Petserimaal ning Vasknarva vald (4 ehk 28,6%) Virumaal.



Joonis 2. Võetud nimede ja soovitusnimestiku „Eesti nimi" nimede pikkus tähtedes.



Joonis 3. Nimemuutjate koguarvu, soovitusnimestiku kasutamise, esitähe säilitamise ning ainutekkeliste nimede dünaamika.

Nimede ja nende võtjate kohta kehtib peaaegu Pareto reegel: 73% nimemuutjatest kasutas ära 22% võetud nimede valikust. Seetõttu olgu toodud ka populaarseimate võetud nimede edetabel (tabel 2). Osa neist esineb ka soovitusnimestikus, kuid neid on võetud mõnel korral ka enne nimestiku ilmumist. Nimede muutmise plahvatuslik kasv ja soovitusnimestikus esinemine ei ole alati seotud, kuna mõlemad langesid ajaliselt kokku. Eeldatavalt on aga just soovitusnimestikust alguse saanud osa populaarseimatest nimedest.

Muutmise tippajast 1935–1940 erinevad varasema ja hilisema perioodi nimevalik oluliselt. Mõlemal perioodil on ainutekkeliste nimede hulk suurem, esitähe säilitamine harvem ning Saksa ajal langes ka soovitusnimestiku kasutamine (vt joonist 3).

Tabel 2. Populaarseimad eestistamisel võetud nimed. Allajoonitud esinevad soovitusnimestikus, tärniga tähistatud laialt levinud nimede keelunimestikus. Esimene number näitab muutmiskordi, teine isikute arvutuslikku hulka, kolmas muutmiskordi, kui palju soovitusnimestikus olevat nime võeti enne nimestiku ilmumist.

1. Aas* 239/564	13. Raid* 148/350	25. Soosalu* 111/262/3
2. Tamme* 228/538	14. Aasa* 143/338	26. Raidma* 108/255/2
3. Kalda* 227/536	15. Merila* 143/338	27. Lillemets 108/255/0
4. Rohtla* 221/522	16. Kallaste* 129/305	28. Ojala 107/253/0
5. Kase* 194/458	17. Saare* 128/302	29. Ojaste 107/253
6. Tali* 191/451	18. Maidla* 127/300/1	30. Vahtra* 107/253
7. Järve* 185/437	19. Randmaa* 127/300/0	31. Heinla 104/246/1
8. Pärna* 182/430	20. Kivimäe* 126/298	32. Randma* 103/243
9. Laanemets* 169/399	21. Kaljula* 125/295	33. Männiste* 101/239/4
10. Toomsalu* 163/385/0	22. Randoja* 123/290/3	34. Kaldma* 100/236/0
11. Toome 153/361	23. Põldma* 122/288	
12. Hein* 149/352	24. Põldmaa* 116/274/2	

Kokkuvõtteks

Artiklis analüüsitakse 1990. aastate oskuse ja tollaste nimeuurimise teadmiste järgi koostatud nimede eestistamise andmebaasi ning antakse ülevaade selle kasutamise probleemidest. Kuigi andmebaas oleks analüüsitav ka oma algkujul, tuli parema tulemuse saamiseks selle andmestikku korrastada.

Täpsema tulemuse saamiseks kõrvutati erinevaid andmestikke. Nimemuutjate üldarvu arvutamisel olid aluseks nimemuutmistoimikute kirjeldusandmed Rahvusarhiivis, mille põhjal aastate kaupa arvutatud keskmine isikute arv ühes toimikus ja seega onomastika andmebaasi ühes kirjes võimaldas täpsustada nimemuutjate üldarvu.Viimane osutus seni kirjanduses käibel olevast üle 7% väiksemaks: 210 000 asemel umbes 195 000. Omaaegne lõpetamata statistika ning seal toodud muutumatu arvutusalus (2,5 isikut muutmise kohta) olid ülearu paisutanud muutjate hinnangulist arvu. Samuti arvestati siin artiklis esmakordselt rööpnimede kaotamist, mida ei saa pidada nimede muutmiseks, rääkimata eestistamisest.

1935. aastal ilmunud soovitusnimestikus „Eesti nimi" leiduvate nimede võrdlus sellest ajast alates kasutusele võetud nimedega näitab nimestiku massilist kasutust, mis aasta-aastalt suurenes, erinedes samas omavalitsuseti, ulatudes 83 protsendini Petserimaal Kalda vallas. Uueks nimeks võetud nimede koondvalikust 30,5% kattus soovitusnimestiku nimedega, olles sageli, kuid kindlasti mitte alati, ka sealt võetud. Muutmiskordi arvestades võeti nimi nimestikust lausa 37%-l kordadest. Nimestikus välja pakutud 15 595 nimest võeti kasutusele 30,2%, eelistades seal välja pakutud keskmisega (7,3 tähte) võrreldes lühemaid nimesid (6,9 tähte) ning välja pakutuga võrreldes suhteliselt rohkem *r*- ja *l*-algulisi. Absoluutarvu poolest olid esikohal *k*-algulised nii välja pakutute kui ka võetute hulgas.

Võetud nimedest kattub esitäht vana nimega 56,8%-l kordadest. Enim on esitähe säilitanud *r*- (77,4%) ja *l*-algulised (77,3%) nimed, alla poole kordadest on säilinud esitähed *b*, *d*, *f*, *i*, *ä, ö* ja *ü*. Esitähe säilitamise tendents kasvas aasta-aastalt.

69% võttis endale 6–8-tähelise nime, sh veerand muutjatest võttis endale seitsmetähelise nime. Iseloomulik näide esitähe säilitamisest ning uue, *k*-algulise ja seitsmetähelise nime võtmisest on muutus *Kakason* > *Kaljula*.

Ainutekkelisi ehk ühel korral võetud nimesid oli 9,7%, enamikus Petserimaa valdadest võeti aga neid eriti palju, kõige enam Roodva ja Saatse vallas (mõlemas 28,6%). 73% nimemuutjatest kasutas ära 22% võetud nimede valikust.

Edaspidi oleks võimalik analüüsida soovitusnimestikes välja pakutud ja tegelikult võetud nimede keelelist külge (siin tehtud automaatse morfoloogilise analüüsi katsed eesti keele sõnavara baasil[13] andsid selleks lootust, kuid jäid välja artikli mahu tõttu), esteetiliselt motiveeritud nimemuutusi (*Junn* > *Vetesina*), sugulussuhetel tuginevaid nimemuutusi (varem eksisteerinud nimede taaskasutamist), perekonna- ja talunimede seost (võrdluseks näiteks 1938. aasta katastriüksuste andmebaas), geograafilis-kronoloogilist dünaamikat (millised olid algul mahajäänud piirkonnad), võetud nimede geograafilist levikut (kus võeti enim kõige levinumaid nimesid, kas ülilevinud *kalju-* ja *ranna*-algulised jm merega seotud nimesid võtsid mereäärsed elanikud), nimede eeskuju (nime levima hakkamine raadioesinemiste järel, nt Johannes Silveti eeskuju), vanade ja uute nimede pikkuse seost, muutmiste sesoonsust, kampaaniate mõju jne.

Nimede eestistamise suure andmebaasi kasutusvõimalusi on palju ning olemasolevate naaberandmestike ning digihumanitaarsete oskuste ja võimaluste arenedes lisandub neid veelgi.

VEEBIVARAD

Haudi = Haudi. Kalmistute register. https://www.kalmistud.ee

OA = Onomastika andmebaas. Aadu Musta poolt 1995–1999 koostatud perekonnanimede paneku ja eestistamise andmed. http://www.ra.ee/apps/onomastika/. Täiendanud Fred Puss alates 2000. Fred Pussi valduses.

ARHIIVIALLIKAD

Rahvusarhiiv (RA)

ERA.14.2.288. Siseministeerium. Kirjavahetus perekonnanimede kaitseregistrisse kandmise kohta. 1936–1937.

ERA.14.2.369. Siseministeerium. Kirjavahetus Akadeemilise Emakeele Seltsiga ja Nimede Eestistamise Liiduga. Statistilisi andmeid nimede eestistamise kohta. 1937–1938.

ERA.14.2.700. Siseministeerium. Kirjavahetus asutuste ja isikutega perekonnanimede muutmise kohta. 1939–1940.

ERA.14.2.1262. Siseministeerium. Kirjavahetus perekonnaseisuametnike ja isikutega perekonnanimede ja nende muutmise küsimuses. 1940.

ERA.1068.1.1. Nimede Eestistamise Keskbüroo. Väljavõtted saabunud kirjadest ja statistilised tabelid perekonnanimede eestistamise kohta. 1939–1940.

[13] Nende eest tänab autor Eesti Keele Instituudi vanemarvutilingvisti Ahti Lohku.

ERA.1093.1.219. Informatsiooni keskus. Kirjavahetus nimede eestistamise ja Eesti lippude levitamise kohta. 1934–1936.

ERA.5201.2.47594b. Siseministeeriumi perekonnaarhiivi dokumentide kollektsioon. Julia-Pauline Buschkov perekonnanime muutmine Männits. 1944.

ERA.5201.2.47594d. Siseministeeriumi perekonnaarhiivi dokumentide kollektsioon. Gretchen-Sigrid Kommusaar perekonnanime muutmine Poola. 1944.

ERA.5201.2.80202. Siseministeeriumi perekonnaarhiivi dokumentide kollektsioon. Karl, Olga Ventsel (Vänzel) perekonnanime muutmine Ventsel; Elmar, Hilda-Marie Unujaan (Unojan, Unnojan) perekonnanime muutmine Ariste. 1935–1936.

KIRJANDUS

Andermann 1935. Bussiga Viljandisse. – Postimees 9. III, nr 67.

Angelus, Oskar 1971. Die Estisierung der Namen in Estland. – Jahrbuch des baltischen Deutschtums, kd XIX (1972), lk 90–97.

ArA 1971. Kui eestlane muutis nime. Fakte sellest aktsioonist. – Eesti Päevaleht = Estniska Dagbladet 13. XI, nr 90, lk 10.

AT = Ametlik Teataja. 1941–1944.

AT III = Ametlik Teataja. III osa. 1941–1944.

EE 1994 = Eesti entsüklopeedia. Kd 7 (Nõuk–rah). Tallinn: Eesti Entsüklopeediakirjastus.

EE 2003 = Eesti entsüklopeedia. Kd 12. Eesti (A–Ü). Tallinn: Eesti Entsüklopeediakirjastus.

EEN 1921 = Eestlasele eesti nimi. (Akadeemilise Emakeele Seltsi toimetused II.) Tartu: Akadeemiline Emakeele Selts.

EKK = Mati Erelt, Tiiu Erelt, Kristiina Ross, Eesti keele käsiraamat. Uuendatud väljaanne. Tallinn: Eesti Keele Instituut, Eesti Keele Sihtasutus, 2020.

Elisto, Elmar 1938. Nimede eestistamisest Eestis. – Omariikluse süvendamisel. Ü. S. Raimla koguteos. Peatoim Jüri Parijõgi. Tartu: Ü. S. Raimla kirjastus, lk 47–62.

Elisto, Elmar 1976. Õigest keelest, tervest meelest. Keelelisi küsimusi II. Tallinn: Eesti Raamat.

EN 1935 = Eesti nimi. Valik uusi perekonnanimesid. Toim Julius Mägiste ja Elmar Elisto. (Akadeemilise Emakeele Seltsi toimetised XXVII.) Tartu: Akadeemiline Emakeele Selts.

ENE 1974 = Eesti nõukogude entsüklopeedia. Kd 6 (Pais–rumo). Tallinn: Valgus.

ENSV T = Eesti NSV Teataja. 1940–1960.

ENSV TL = Eesti NSV Teataja Lisa. 1940–1957.

Glavlit 1948 = ENSV Kirjandus- ja Kirjastusasjade Peavalitsuse ülema käskkirjadega kõrvaldatavate eestikeelsete raamatute üldnimestik (1940–1947). Tallinn: RK Poliitiline kirjandus.

Glavlit 1963 = Raamatukogudest ja antikvariaatidest kõrvaldamisele kuuluvate kodanlikul ja okupatsiooni ajal Eestis ilmunud väljaannete koondnimekiri. II osa. Tallinn: *s.n.*

Grönberg, Artur 1985. Eesti Üliõpilaste Seltsi ajalugu. II: Iseseisvuse eelvõitluses (1906–1917). Omariikluse saavutustes (1918–1940). Montreal: Eesti Üliõpilaste Seltsi Vanematekogu kirjastus.

Henno, Kairit 2001. Emakeele Selts nimede eestistamise käivitajana. – Keel ja Kirjandus, nr 2, lk 73–79.

Henno, Kairit 2002. Riiklikust isikunimekorraldusest aastatel 1918–1940. – Õiguskeel, nr 4, lk 33–42.

Karjahärm, Toomas; Sirk, Väino 2001. Vaim ja võim. Eesti haritlaskond 1917–1940. Tallinn: Argo.

Kask, Arnold 1960. 40 aastat Emakeele Seltsi. – Emakeele Seltsi aastaraamat 6 (1960). Tallinn: Eesti Riiklik Kirjastus, lk 5–23.

Kask, Arnold 1970. Pool sajandit Emakeele Seltsi. – Emakeele Seltsi aastaraamat 16 (1970). Tallinn: Eesti Raamat, lk 7–30.

Leivategija, Jürgen 2011. Perekonnanimede eestistamine kui kultuuriajalooline protsess: Seltskondlike organisatsioonide roll kampaanias ja selle tausta kujundamisel aastatel 1921–1940. Bakalaureusetöö. Tallinna Ülikool, Ajaloo Instituut.

Lepp, A. 1936. Perekonnanimede statistika. – Eesti Statistika, nr 5, lk 264–267.

MN 1935 = Muudetud nimede register. I: 1919 kuni 1. juuli 1935. Toim August Juhan Rava. Tallinn: Aura.

Must, Aadu 1999. Perekonnanimede eestistamine. – Maakodu, nr 1, lk 6–7.

Must, Aadu 2000. Eestlaste perekonnaloo allikad. 2., parand ja täiend tr. Tartu: Ajaloo-kirjanduse Sihtasutus Kleio.

Must, Aadu 2006. Names as sources. The implementation of Estonian surnames in the historical research. – Proceedings of the 21st International Congress of Onomastic Sciences. Uppsala 19–24 August 2002. Kd 2. Toim Eva Brylla, Mats Wahlberg koostöös Lars-Erik Edlundiga. Uppsala: Språk- och folkminnesinstitutet, lk 214–221.

Must, Aadu 2010. Name formation, name change and name loss. An Estonian case (1826–1940). – Proceedings of the 21st International Congress of Onomastic Sciences. Uppsala 19–24 August 2002. Kd 5. Toim Eva Brylla, Maria Ohlsson, Mats Wahlberg koostöös Wolfgang Haubrichsi ja Tom Schmidtiga. Uppsala: Språk- och folkminnesinstitutet, lk 393–400.

Must, Aadu 2015. Perekonnaloo uurija käsiraamat. Tallinn: Hea Lugu.

Must, Hildegard 1957. Eestlaste perekonnanimedest. – Tulimuld, nr 6, lk 279–285.

Mõistlik, Piret 2005. Isikunime õiguslik korraldus. Nimeseaduse kommentaarid. Tallinn: Juura.

Nimede eestistamine 1987 = Nimede eestistamine poole sajandi eest. – Eesti Päevaleht = Estniska Dagbladet 3. VI, nr 43, lk 4.

Närhi, Eeva Maria 1987. The changing of surnames in Finland during the twentieth century, particularly in 1906 and 1935. – Studia Anthroponymica Scandinavica, kd 5, lk 109–121.

PENA 1921–1923 = Petserimaal ja Narvatagusel 1921–1923 pandud perekonnanimed. Andmebaas peamiselt kartoteegi (RA, EAA.5433.1.11 kuni 48) alusel. Koostanud Fred Puss, Eha Otsa, Mall Pendin. Fred Pussi valduses.

Puss, Fred 2020. Muuga või Kersna? Eesti rööpsete perekonnanimede teke ja kadu. – Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics, kd 11, nr 1, lk 43–85. https://doi.org/10.12697/jeful.2020.11.1.03

Põhjaklane 1935. Nimede muutmisel maal suur poolehoid. – Oma Maa 21. I, nr 9.

Pärn, Aveli 2013. Perekonnanimede eestistamine Tormas aastatel 1935–1940. Bakalaureusetöö. Tartu Ülikool, soome-ugri osakond. https://dspace.ut.ee/handle/10062/31614

Raag, Arno 1971. Kõuepilvede saatel. Mälestusi. Lund: Eesti Kirjanike Kooperatiiv.

Raidmaa, Ilona 1999. Perekonnanimede muutmise ja eestistamise liikumine Eestis 1918–1940. Muudetud ja eestistatud nimed Anija, Peningi, Raasiku ja Rae vallas. Seminaritöö. Tartu Ülikool, eesti keele õppetool. https://murdearhiiv.ut.ee/naita.php?t=kasikiri&id=5197

Rajandi, Edgar 1978. Mõtteid Eesti isikunimede uurimisest. – Sõna ja nimi. Emakeele Seltsi aastaraamat 23 (1977). Tallinn: Eesti Raamat, lk 93–102.

Rajandi, Edgar; Tarand, Helmut 1976. Eesti nimi läbi sajandite. – Keel, mida me uurime. Koost Mart Mäger. Tallinn: Valgus, lk 131–136.

Rajasalu, Inda 1993. Riiklik Propaganda Talitus ja eesti nimi. – Tundmatu Eesti Vabariik. Koost Jüri Ant, Eero Medijainen, Ago Pajur. Tallinn: Jaan Tõnissoni Instituut, lk 96–108.

Raun, Toivo U. 2012. The movement to Estonianize surnames in interwar Estonia. – Acta Historica Tallinnensia, kd 18, nr 1, lk 97–107. https://doi.org/10.3176/hist.2012.1.04

Reiman, H. 1936. Asustamistihedus ja rahvaarvu muutumine Eestis. – Eesti Statistika, nr 4, lk 169–176.

Roos 1980 = Sõnade ja nimede seltsis: Eduard Roos. [Intervjuu]. – Eesti sõnavara. Emakeele Seltsi aastaraamat 24 (1978). Tallinn: Eesti Raamat, lk 67–79.

RT = Riigi Teataja. 1918–1940.

RTL = Riigi Teataja Lisa. 1921–1940.

Salurand 1935 = 33 000 uuenimelist. [Vestlus hr Salurannaga Nimede Eestistamise Keskbüroost]. – Postimees 14. X, nr 280.

Staltmane 1977 = V. E. Staltmane, Iz istorii latyšskih familij. – Istoričeskaja onomastika. Otdel'nyi ottisk. Akademija nauk SSSR, Institut jazykoznanija. Moskva: Nauka, lk 159–180 [В. Э. Сталтмане, Из истории латышских фамилий. – Историческая ономастика. Отдельный оттиск. Академия наук СССР, Институт языкознания. Москва: Наука.]

Zetterberg, Seppo 2010. Eesti ajalugu. 4., parand ja täiend tr. Tallinn: Tänapäev.

Tamm, Marek 2020. Eesti ajalooteadus 21. sajandil: Sissejuhatuseks. – Acta Historica Tallinnensia, kd 26, nr 1, lk 3–45. https://doi.org/10.3176/hist.2020.1.01

Teder, Teofil 1939. Perekonnaseisuametniku käsiraamat. Tallinn: Siseministeeriumi Administratiiv-ala Kirjastus.

UN 1935 = 4000 uut nime. Kuidas leida uut perekonnanime. Viljandi: Oma Maa.

US 1921 = 15.000 uut sugunime. Lisa albumile „Eestlasele Eesti nimi". (Akadeemilise Emakeele Seltsi toimetused IV.) Tartu: Akadeemiline Emakeele Selts.

Uut hoogu 1940 = Uut hoogu nimede eestistamisel. – Sakala 1. IV, nr 37, lk 3.

Vaan, Laura 2005. Propagandatalitus Eesti Vabariigis autoritaarsel ajajärgul. Magistritöö. Tartu Ülikool, Eesti ajaloo õppetool. http://dspace.ut.ee/handle/10062/1208

Walkowiak, Justyna B. 2016. Personal Name Policy: From Theory to Practice. (Dysertacje Wydziału Neofilologii UAM w Poznaniu 4.) Poznań: Wydział Neofilologii UAM w Poznaniu. https://doi.org/10.14746/9788394601720

VT 1934 = Valitsusasutiste tegevus 1918–1934. Tallinn: Riigikantselei.

VT 1939 = Valitsusasutiste tegevus 1938/39. a. Tallinn: Riigikantselei.

Fred Puss (snd 1976), Eesti Keele Instituudi leksikograaf-nooremteadur; Tartu Ülikooli ajaloo ja arheoloogia instituudi doktorant (Jakobi 2-222, 51014 Tartu), fred.puss@eesti.ee

Changing Kakason to Kaljula: Analysing database of Estonianizing surnames

Keywords: anthroponymics, surnames, Estonian history, name changes, national campaigns

Surname changes have taken place throughout their existence, but not so often as a result of a nationwide campaign. In Estonia campaign of Estonianizing foreign (mostly German) sounding names took place in 1920–1944 with 99% of the name changes taking place in 1935–1940. It stands out for its intenseness which resulted a surname change of every sixth person. Also, public access to the complete database of name changes already since the 1990s (www.ra.ee/apps/onomastika/) is unique when compared to other countries. Author of this article has amended the public version of the database for the last two decades and explains the arisen problems with it.

This article uses new data analysis methods to correct the total number of surname changers during the campaign (see Table 1). The resulting number of approximately 195,000 is smaller than has been estimated so far (210,000). Estonian population at that time was 1.13 million, thus the amount of name changers was 17%.

New names were often taken from a 1935 book of suggested surnames titled *Eesti nimi* ("Estonian name"). The book contains 15,595 suggested names and 30% of them were actually taken as a new name. Also 30% of all taken names (37% when counting all individual name changes) come from the book, in some municipalities even up to 83%. The public preferred shorter names from the book and the use of the book slightly increased year by year (see Figures 2 and 3). Average length of all taken names was seven letters (69% took a name with 6 to 8 letters), the most popular first letters being *k* (13%), *r* (11%) and *l* (10%).

The first letter of the old and new name matches in 57% cases, most popular preserved first letters being *r* and *l* (both 77% , see Figure 1). Tendency of preserving the first letter also increased in time. 10% of the names were taken only once, with the highest (up to 29%) number in municipalities of Petseri county. Taking an unique name also slightly increased in time. Almost the Pareto rule applies to surname changers: 73% of name changers used 22% of the overall choice of taken names.

Many more analyses can be made with the huge database, for example analysis of the vocabulary used in taken names, aesthetically motivated name changes, reused names (for example taking a wife's maiden name as a new surname), geographical-chronological dynamics of name changes etc.

Fred Puss (b. 1976), Institute of the Estonian Language, Lexicographer-Junior Researcher; University of Tartu, Institute of History and Archaeology, PhD Student (Jakobi 2-222, 51014 Tartu), fred.puss@eesti.ee

https://doi.org/10.54013/kk764a9

Kuidas kodeerida kliimat?

Eesti ajaloolise kliimauurimise digitaalsest pöördest

ULRIKE PLATH, PRIIT RAUDKIVI, KAAREL VANAMÖLDER,
KRISTER KRUUSMAA, ANNA HELENA LIIV

Ajalugu saab ja peab tegema ka ilma matemaatiliste meetoditeta. Kuid mõnes mõttes tundub mulle ajaloolane, kes kaasajal ei taha kasutada neid meetodeid, umbes inimesena, kes tahab sõita Moskvasse. Kuid otsustab mitte sõita Estoniaga[1], vaid võtab kepi kätte ja hakkab jala minema... (Juhan Kahk 1988 [1980])

„Eile tabas meie linna selline rahetorm, millist ei mäleta ka kõige vanemad inimesed," kirjutati Rigasche Zeitungis 11. mail 1872. aastal (RiZ 1872a). Tõepoolest leidis 10.–11. mail (vkj) nii Eesti- kui ka Liivimaa kubermangus aset tugev torm, mis kohati paisus tornaadoks. Seda täheldati erinevates kohtades ning kummagi kubermangu ajalehtedes avaldati juhtunust kokku üle viiekümne teate. Neis ilmnevad tormi hävitustöö suured paikkondlikud erinevused. Näiteks Tallinnas oli torm väiksema jõuga ning suurtest kahjudest ei kirjutatud, seevastu Paides puhus tugev tuul pikali lossivaremete nurgatorni (ReZ 1872). Läti alal olid purustused veel ulatuslikumad: Allažis lennutas tuul hobuse koos vankriga üle viie jala kõrguse kirikaia müüri ning haaras kaasa kirikulised, kellest üks naine ja kaks tüdrukut hukkusid (RiZ 1872b). Võnnus (Cēsis) tõstis keeristorm kolm tüdrukut õhku ja kaks surid kukkumise tagajärjel, peale selle hukkus neli teenijat – ühest tüdrukust leiti vaid jalg ja käsi. Riias hinnati tormikahjude summa 60 000 rublale. (RiZ 1872c) Seesuguseid teateid ilma „trikkidest" leidub kirjalikes allikates omajagu. Kas ja kuidas saab ajalehti kasutada kliimaajaloo analüüsis ning milliseid võimalusi pakub tänapäevane digihumanitaaria, et nende värvikate kirjelduste abil ajaloonarratiivi uuendada?

Ajaloolase kireks on alati olnud andmete kogumine. Aastasadade ja -kümnete jooksul on vaid muutunud kogumise viisid ning eesmärgid: mida ja milleks koguda? Kogutud informatsiooni hulk on n-ö pikal XX sajandil (XIX sajandi lõpp kuni XXI sajandi algus) kasvanud aga nii suureks, et paremagi tahtmise juures ei suuda üksikisik seda enam hallata. Kvantitatiivse uurimismeetodi vaimustuses loodi eelmisel sajandil ka ajalooteaduses järjest rohkem kartoteeke ja mindi seejärel andmetöötluses üle statistiliste ja matemaatiliste meetodite rakendamisele (Mahoney 1988; Paju 2020).

Kuigi kliimaajalugu on interdistsiplinaarne uurimisvaldkond, kus publikatsioonide arv järjest kasvab, pole selle algusajast palju kirjutatud ei Eestis ega mujal Euroopas. Et täiendada üldpilti kliimaajaloo uurimisest XX sajandi teisel poolel Eestis ja mujal, oleme teinud intervjuu Andres Tarandiga (snd 1940), kes on siinse kliimaajaloo uurimise alusepanija (Tarand 2021). Samuti oli lahkesti nõus üldpilti täien-

[1] Tallinn-Moskva kiirrong.

dama Christian Pfister (snd 1944), Euroopa ühe suurema ja tuntuma kliimaajaloo andmebaasi Euro-Climhist[2] rajaja ja eestvedaja (Pfister 2021). Mõlema mälestuste ja pärandi – perfokaartide, kartoteekide, vanade andmebaaside – kaudu püüame selgitada, kuidas toimus digipööre kliimaajaloo uurimises, vaadeldes samal ajal laiemalt ajalooteaduse moderniseerumist XX sajandi teisel poolel. Käsitleme Andres Tarandi kliimaajaloo andmebaasi kui üht nähtust selles protsessis, mis on kestnud suuremate katkestusteta juba üle 40 aasta. Püüame näidata, kuidas kirjalikult ülestähendatud ilmanähtusest saab kodeerimisprotsessi käigus number ning kuidas on võimalik koodide alusel koostada uusi kliimaajalugusid.

Artikkel ei pretendeeri siiski Eesti mineviku kliima ülevaate andmisele. Sel teemal on oodata lähiajal omaette artikleid, mis valmivad koostöös loodusteadlastega, et valideerida ajaloolisi andmeid reaalteaduslike uuringutega (Huhtamaa jt ilmumas). Esmalt uurime aga küsimust, millal sai ajalooline klimatoloogia – geograafia ja ajaloo vaheline distsipliin – Eesti ajaloo uurimise osaks.

1872. aasta torm ja ajaloolise klimatoloogia sünd Eestis

Esimene põhjalik 1872. aasta maitormi käsitlus ilmus sündmusele järgnenud aastal Riia Loodusuurijate Seltsi väljaandes Gotthard Schwederi (1831–1913) sulest. See artikkel näitlikustab nii meedia huvi tõusu ilmanähtuste (näiteks ilmastikuvaatluste järjepidev avaldamine, teated ekstreemsetest ilmaoludest jne) vastu kui ka meteoroloogia teaduslikku taset Baltimaades XIX sajandi lõpus. Baltimaade ajaloos ainulaadseks peetud nähtust kirjeldati ajaleheteadete ja sündmuse tunnistajate intervjuudest saadud teabe põhjal. Kogutud materjali alusel ei kirjeldatud mitte ainult rahe ja trombi vormi ning hävitustööd, vaid juhtunut võrreldi ka viimase suure mäletatava rahetormiga 1795. aastast. Ühtlasi püüti selgitada tormi alguspunkti ja trajektoori. Lähtekohaks pakuti Kesk-Saksamaad, kus mais olevat samuti olnud ulatuslikud rahekahjud. (Schweder 1873: 59–60)

Meteoroloogiat uuriti ja õpetati Peterburis (Wild 1881) ja Tartus (Liidemaa 1992; Mürk 1992; Palm 1992), kus maailmakuulsaks trombide uurijaks sai Johannes Letzmann (1885–1971) (Peterson 1992). Kohalike meteoroloogide huvi ajalooliste ilmanähtuste vastu on olnud siiski piiratud. Nõnda kulus 1872. aasta tormist sada aastat, enne kui pandi alus Eesti ajaloolise klimatoloogia uurimisele, mis eeldas meteoroloogide ja ajaloolaste koostööd. Meteoroloog Heino Tooming (1930–2004) pööras samale tormile uuesti tähelepanu 1970. aastatel (Tarand, Kallis 2017: 30–32) ning samal ajal tekkis ka ajaloolaste seas huvi ajaloolise kliima vastu. Esimene, kes eeldas, et ilmal ja pikemas plaanis kliimal võib olla mõju Eesti rahvastikuprotsessidele, oli ajaloolane Sulev Vahtre (1926–2007), kes kasutas sel ajal uuenduslikku lähenemist keskkonna ja ühiskonna suhete mõtestamisel: oma doktoritöö (Vahtre 1970a) ettevalmistamisel kogus ta ilmaandmeid perioodist 1700–1870 (Vahtre 1970b). 1872. aasta tormi käsitlus ei mahtunud siiski tema andmekogu kronoloogilisse raamis-

[2] Andmebaasi haldab Šveitsi Berni ülikooli Oeschger Centre for Climate Change Research (vt Euro-Climhist).

tusse. 1970. aastate keskel alustas Eesti ajaloolise kliima andmebaasi loomisega ka Tarand. 1872. aasta tormi mainitakse veel kord, küll tagasihoidlikult, alles Tarandi, Jaak Jaaguse (snd 1956) ja Ain Kallise (snd 1942) raamatus „Eesti kliima minevikus ja tänapäeval" (2013: 296), detailsem kirjeldus on aga raamatus „Eesti tornaadod", mis vaatleb perioodi 1745–2015 (Tarand, Kallis 2017: 22–23, 99–105). Kliima- ja keskkonnaajaloo küsimustega on tegelenud ka Priit Raudkivi (2010, 2014, 2016a, 2016b).

Digihumanitaaria esiaeg Eesti ajalooteaduses

Personaalarvutite levik muutus Eestis üldiseks 1980. aastate lõpust alates. Selle eelset ajastut võiks tinglikult nimetada digihumanitaaria muinas- või esiajaks. Ajastuga seonduv mõiste on kliomeetria (ingl *cliometrics*), mis tähistab üldiselt matemaatiliste meetodite kasutamist ajaloouurimises. Perioodi alguseks võib pidada 1950. aastate lõppu, mil teadusasutustes hakati kasutama elektronarvuteid: 1959. aasta lõpus võeti Tartu Riiklikus Ülikoolis (TRÜ) kasutusele arvuti Ural 1 ja Tallinnas Küberneetika Instituudis (tegutses aastatel 1960–2017) käivitati peagi Eestis toodetud arvuti M3 Estonia jne (Tõugu 2018; Salum 1999). Suurte elektronarvutite soetamisvõimalused olid piiratud juba tulenevalt nõukogude teadusasutuste hierarhiast.[3] Lihtsamaks ja ilmselt levinuimaks suurte andmekogude haldamise ja analüüsi viisiks oli nende koondamine perfokaartidest kartoteekidesse, mida oli võimalik töödelda nii elektronarvutite kui ka kõikvõimalike mehaaniliste sorteerijate, tabulaatorite jms abil (Neithal 1963).

Ajalooteaduses mõisteti masinarvutusega kaasnevaid võimalusi üsna pea (Kahk 1964; Palli 1965a; Ruusalepp 1993; Krikmann 2006). Eesti NSV Teaduste Akadeemia Ajaloo Instituut (AI) ja TRÜ olid koos Moskva riikliku ülikooli ja Novosibirski teaduslinnakuga selles vallas teerajajad terves Nõukogude Liidus (Borodkin 1993: 8). 1962. aastal tegid ajaloolased Juhan Kahk (1928–1998) ja Herbert Ligi (1928–1990) AI-s esimese katse kasutada oma töös kõrgemat matemaatikat Küberneetika Instituudi arvutipargi ning täppis- ja majandusteadlaste abiga (Kahk, Ligi 1964; Jansen 1972: 335; Borodkin 1993: 8; Rosenberg 1998: 56; Tamm 2020: 42). Kahk tutvustas erialakirjanduses elektronarvutite kasutamise viise regressioonianalüüsi ja korrelatsioonimeetodi rakendamisel, rõhutades muu hulgas, kuidas masinad avavad võimaluse läbi töötada hiiglasuuri andmehulki. Samas toonitas ta jätkuvalt range allikakriitilise lähenemise vältimatust. (Kahk 1964)

Perfokaarte hakati Kahki initsiatiivil AI-s uurimistööks kasutama 1962. aastal. Laiemalt leidsid kasutust kaht liiki kaardid: sälkkaardid märkmete ja väljakirjutuste tegemiseks ning duaalkaardid masintöötluse tarbeks. Suurt tähelepanu pöörati kaartidel ühtsete koodide kasutamisele: see võimaldas AI töötajatel teha koostööd teiste asutustega, nagu etnograafiamuuseum (praegune Eesti Rahva Muuseum), või TRÜ õppejõudude ja üliõpilastega (Palli 1965a: 108). Kahk osales ühe esimese Eesti

[3] Nii oli näiteks 1970. aastatel Tallinna Botaanikaaial kui kolmanda järgu (ehk kõige vaesemal) Teaduste Akadeemia asutusel keeruline hankida elektronarvutit (Tarand 2013: L1).

„kvantifikaatorina" ehk matemaatiliste meetodite rakendamise üleliiduliselt tunnustatud autoriteedina ka 1981. ja 1987. aastal Tallinnas toimunud Nõukogude Liidu ja Ameerika Ühendriikide arvutiajaloolaste sümpoosionidel (Rosenberg 1998: 56). Prantsuse ajaloolase Emmanuel Le Roy Ladurie (snd 1929) kutsel sai Kahk 1970. aasta lõpul võimaluse töötada nn Annaalide koolkonna tugipunktis, Pariisis asuvas õppe- ja teaduskeskuses École des Hautes Études en Sciences Sociales (Rosenberg 1998: 54).

Kahki kõrval sai matemaatiliste meetodite ja perfokaartide agaraks propageerijaks ja juurutajaks AI-s Heldur Palli (1928–2003). Kahk ja Palli populariseerisid uusi meetodeid ja tutvustasid neid aktiivselt ka rahvusvahelisele teadusavalikkusele (Palli 1965b, 1967, 1969; Kahk 1971, 1974; Juhan Kahk 1988 [1980]). 1960. aastate keskpaigast kuni 1980. aastate lõpuni õpetati TRÜ-s ajaloolastele kursust matemaatilistest meetoditest ajalooteaduses, õppejõududeks alguses matemaatik Leo Võhandu (snd 1929), hiljem Ligi ja Kahk (Ruusalepp 1993: 24; Tartu Riikliku Ülikooli... 1983: 20–21).

Masinarvutust, kvantitatiivanalüüsi meetodeid, faktoranalüüsi jne kasutati peamiselt agraar- ning majandusajaloo, ajaloolise demograafia jms vallas statistilise ja sotsiaalmajandusliku andmestiku uurimiseks. Kahki, Ligi, Palli ja Vahtre kõrval tasub uurijatena nimetada ka Ants Ruusmanni (1935–2009), Enn Tarvelit (snd 1932), Raimo Pullatit (snd 1935) ja Helmut Piirimäed (1930–2017) (Jansen 1972: 335; Ruusalepp 1993: 23). Uurimine käis ajaloolaste ja matemaatikute koostöös: ajaloolased ise programme ei loonud, kuid andsid perfokaartidele ja muudele kandjatele kogutud andmestiku läbitöötamiseks arvutuskeskuste matemaatikutele, kes lõpptulemusena saatsid tagasi n-ö pikad paberirullid arvudega. Nii rakendas Piirimäe XVII sajandi Rootsi riigimajanduse uurimisel hoopis kolhooside majandustegevuse analüüsiks loodud programmi (Ruusalepp 1993: 23).

Digihumanitaaria muinasaja lõpp 1980. aastate teisel poolel ei ole Eesti ajalooteaduses seletatav ainuüksi personaalarvutite ilmumisega. Laulev revolutsioon tõstatas hoopis teised teemad ning uurimisküsimused. Eesti Vabariigi taasiseseisvumise järgsetel aastatel valitsenud üldises ainelises nappuses polnud lihtsalt võimalik soetada piisaval hulgal kaasaegseid arvuteid. Kirjanduses ja mälestustes on korduv motiiv erakordne sündmus, kui välismaalt saadi kingituse või abina üksik kaasaegne (sageli kasutatud) kompuuter (Krikmann 2006: 26; Ruusalepp 1993: 24). Ajaloolased, kes 1960. aastatel – oma parimas loomeeas – edendasid masinarvutust ja lõid andmekogusid, olid pensionile siirdumas. Aastakümneid väldanud kogumistöö käigus perfokaartidele ja muudele meediumidele talletatud andmebaaside digiteerimiseks polnud ressursse, tehnilist võimekust ega vahest ka huvi. Andmebaasid olid jäänud kinni eelmisse ajastusse, enamik neist unustati, paljud hävisid. On silmatorkav, et 1990. aastatel ajaloouurimise „komputeriseerimisest" kirjutanud autorid peatuvad selle eellool vähe või üldse mitte (Somelar 1993; Ruusalepp 1993, 1996; Must 1998).

Ajaloolise kliima andmebaasid Šveitsis ja Eestis

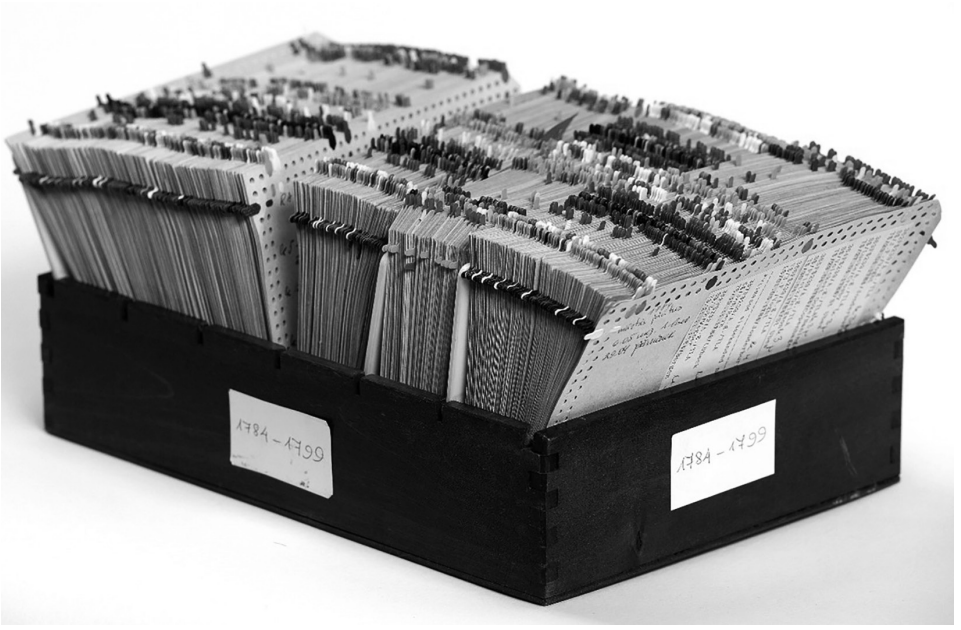
Kesk-Euroopas tõusis huvi ajaloolise klimatoloogia vastu 1970. aastatel. Selle protsessi eestvedaja oli Šveitsis Berni ülikoolis geograafia instituudis agraarajaloo uurimissuunal tegutsenud Christian Pfister, kes õppis aastail 1966–1970 Berni ülikoolis ajaloo kõrval geograafiat ning huvitus just põllumajanduse ja kliima vahelistest seostest (Rohr jt ilmumas).

Oluliseks Pfisteri mõjutajaks sai eespool mainitud prantsuse agraarajaloolase Le Roy Ladurie 1967. aastal ilmunud ja praeguseks kliimaajaloo klassikaks saanud teos „Kliima ajalugu alates aastast tuhat" („Histoire du climat depuis l'an mil"; Pfister 2021). Teoses näidatakse, et kombineerides eri distsipliinides materjali, nagu traditsioonilised ajalooallikad, fenoloogilised ning dendrokronoloogilised andmed, on võimalik rekonstrueerida kliimaajaloo pikki aegridu, mis sisaldavad sootuks täpsemat teavet, kui seda suudaksid pakkuda allikagrupid eraldi (Le Roy Ladurie 1967). Kindlasti mängis metoodika valikul rolli Le Roy Ladurie informaatikalembus – kuulsa lause „homne ajaloolane on kas programmeerija, või pole teda üldse" (Le Roy Ladurie 1968) autor oli ajalooteaduse tehnoloogilise uuenemise üks peamisi eestvedajaid XX sajandil (Lemny 2017). Samuti pidas ta vajalikuks andmebaaside loomist ning nende pidevat täiendamist ja kasutamist: „Arvutipõhine ajalugu ei vii meid ainult selgelt defineeritud uurimusteni. Selle tulemuseks on ka „arhiivi" loomine. Kord perforeeritud kaartidele või lintidele talletatud infot saab säilitada, pidades silmas tulevasi uurijaid, kes soovivad leida sealt uusi korrelatsioone." (Le Roy Ladurie 1973: 14) Erinevate andmetüüpide kombineerimine viis loomuldasa edasi kvantitatiivsete meetodite rakendamise juurde. Mälestustes toob Le Roy Ladurie selles valdkonnas eeskujuks oma lennukaima õpilase Pfisteri, kelle kogutud andmehulk olnuks tema sõnul varem kujuteldamatu (Le Roy Ladurie 2014: 44, 168, 177).

1976.–1977. aastal õppis Pfister Inglismaal Rochesteri ülikoolis elektronarvutite programmeerimiskeelt PL-1, mida ta kasutas kliimaajaloo andmebaasi loomisel: „Ma olin algusest peale veendunud, et üksnes kvalitatiivsete ja kvantitatiivsete meetodite ühildamisega on taolisel uurimistööl mõtet" (Pfister 2021). Andmebaas funktsioneeris algusest peale masinloetavas perfokaardisüsteemis. Pfisteri andmetel oli juba sel ajal kasutusel põhijoontes sama klassifikaatorite süsteem nagu tänapäeval, aga aja jooksul on seda täiendatud. Üleminek seni kasutatud masinloetavast perfokaardisüsteemist tänapäevasesse arvutisüsteemi toimus seega sujuvalt, kuigi suure tööjõukuluga. 1990. aastate alguses kujunes andmebaasist praegune kliimaandmebaas Euro-Climhist. 2011. aastal avati see veebilehena ja muudeti neli aastat hiljem uurijatele avalikult kättesaadavaks. Hetkel hõlmab Euro-Climhist üle 300 000 kliimateate keskajast tänapäevani valdavalt Kesk-Euroopast (Rohr jt ilmumas).

Nagu näha, mängivad isiklikud sidemed ja eeskujud vahel olulist rolli. Siinses käsitluses jooksevad niidid kokku sarnaste teadushuvidega (agraar- ja majandusajalugu) uurijate juures (Le Roy Ladurie, Kahk, Pfister). Nõnda on ootuspärane, et 1990. aastate alguses tutvusid omavahel Tarand ja Pfister, kes mõlemad haldasid kliimaajaloo andmebaase.

Tarandi andmebaas on üks väheseid edukaid näiteid sellest, kuidas omaaegsele andmekandjale talletatud info on läbinud formaadimuutuse ning on digiteerituna tänapäevani kasutatav.[4] Nagu ajaloolasedki, koondas Tarand kogutud kliimateated perfokaartidele, millest moodustus mahukas kataloog; andmestik hõlmab perioodi aastaist 1000–1923. Tuhanded perfokaardid mahuvad 21 kasti.[5] Andmehaldus toimus peamiselt spetsiaalsete värviliste plastikust reiterite süsteemi abil (vt pilt 1).



Pilt 1. Andres Tarandi kliimaandmebaasi kartoteegikast Tartu Ülikooli muuseumis (TÜM, ÜAM 1565:11 Ar, foto Kaarel Vanamölder).

Eesti kliima andmebaasi ajaloos eristab Tarand kolme faasi. Esimene, aastail 1975–1990, oli andmete kogumise organiseerimine ja perfokaart-kartoteegi koostamise faas. Sel ajal tegutseti Eesti-siseselt. Teises faasis (2004–2009) digiteeriti kanded, kirjutades info perfokaartidelt ümber Microsoft Wordi. Samal ajal täiendati andmestikku digikujul teadetega XIX sajandi ajakirjandusest. Umbes samal ajal tekkisid kontaktid Berni ülikooliga ja algas ka andmete kodeerimine Euro-Climhisti süsteemi järgi, eesmärgiga muuta andmebaas tulevikus sellega ühilduvaks. Kahjuks jäi protsess vahepeal seisma. Kui 2010. aastate keskpaigas asuti jälle tööle, oli Berni süsteem keerulisemaks läinud. Seega oli vaja siirata Wordi kanded Microsoft Excelisse vastavalt Euro-Climhisti uuele süsteemile. Pausid erinevate faaside vahel on seletatavad Tarandi muu tegevusega taasiseseisvunud Eestis. On selge, et niivõrd

[4] Peale Tarandi andmebaasi digiteeriti 1990. aastate alul näiteks ka Herbert Ligi nn erisurmade kartoteek u 10 000 kandega, kasutades selleks programmi dBase III (Rahi, Must 1993: 10).

[5] 2013. aastal annetas Tarand andmebaasi Tartu Ülikooli muuseumile ja see on leitav viite all TÜM, ÜAM 1565:1–21 Ar.

pikaajalise protsessi jooksul oli eripalgeliste ülesannete lahendamiseks töösse kaasatud palju abilisi, kelle tegevust Tarand koordineeris. Käimasolevas kolmandas faasis (alates 2017. aastast) valmistatakse andmeid (*ca* 40 000 sissekannet) ette lõplikuks üleviimiseks Euro-Climhisti andmebaasi. Aastatel 2017–2020 täiendati seni kogutud andmeid tormisuse osas perioodist 1650–1900 Eesti ja Liivimaa alal, eesmärgiga luua Euro-Climhistis Balti moodul tormidest. Lisaks tormidele on praeguseks Euro-Climhisti andmebaasis olemas ka Tarandi elutöö üks tähtsamaid tabeleid: Tallinna talvise keskmise õhutemperatuuri aegrida perioodist 1322–1915. (Tarand 2021)

Jagub palju näiteid sellest, kuidas informatsioonil on oht digiajastu tulekuga n-ö raputada saada. XXI sajandi alguse digiteerimisele on iseloomulik andmekandjate ja formaatide paljusus. Kontoritarkvara, nt Microsoft Office'i kättesaadavus on andnud küll võimaluse tuua digiteerimine kõikjale, kuid samal ajal ei ole see üldjuhul suuremate andmekogude haldamiseks kuigi hea tööriist ning võib kaasa tuua kadusid ja segadust. Õnneks on Tarandi kliimaandmestik need katsumused üle elanud ja avatud tänapäevase digimaailma võimalustele. Peamiste perspektiivide hulka kuulub kindlasti täiendavate andmete leidmine juba digiteeritud materjali hulgast. Näiteks digiteeritud ajalehtede andmebaasid, nagu Digar Eestis ja Periodika Lätis, annavad võimaluse kliimaajaloo seisukohalt tähtsa teabe kogumisel füüsiline andmekandja n-ö andmeringest üldse eemaldada. Lisaks on sellistele andmekogudele võimalik rakendada palju võimsamaid kvantitatiivseid meetodeid, andmekaevet ja -analüüsi.

Tormide kodeerimine ja andmebaasi kasutusvõimalused

Kuidas saavad 1872. aasta maitormi kirjeldused ehk kvalitatiivsed andmed kvantitatiivseteks andmeteks ehk arvudeks? Suurem osa ajaloolisest kliimateadusest on keskendunud homogeensetele aegridadele, näiteks ühe vaatleja või vaatlusjaama mitme või mitmekümne aasta jooksul kirja pandud vaatlusandmete analüüsile. Kui mõõtmiseks on kasutatud instrumente, on vastava ilmaelemendi muutumise dünaamika (temperatuuri, õhurõhu vms) rekonstrueerimine võrdlemisi lihtne, ehkki ka siis tuleb matemaatilise rekonstruktsiooni käigus tuvastada võimalikke väliseid mõjutajaid, näiteks termomeetri asukoha muutus (vt Moberg jt 2002). Kui vaatlused on järjepidevad, kuid sõnalised (näiteks ilmavaatluspäevikud), tuleb nende sisu numbriliseks teisendada. Selle jaoks on välja pakutud erinevaid indekseid ehk skaalasid. Näiteks Pfisteri loodud ja laialdast kasutust leidnud temperatuuri- ja sademeskaalad määravad igale kuule väärtuse seitsmepallisüsteemis: –3 (äärmiselt kuiv/külm) kuni +3 (äärmiselt märg/soe). Siiski on ka skaalade puhul vaja leida kattuvus teiste allikatega (instrumentaalsed jadad, loodusteaduslikud andmed), et fikseerida skaala nullpunkt, st ilmastiku n-ö normaalolek valitud vaatlusperioodil (Pfister jt 2018: 115–120; Brázdil jt 2005). Kirjeldustele numbriliste või diskreetsete väärtuste omistamise protsessi (olgu selle reeglid konkreetsel juhul millised tahes) kutsutakse kodeerimiseks ja see on kvantitatiivse analüüsi paratamatu eeltingimus. Selle analüüsi kvaliteet sõltub kodeerimise võimalustest, töö põhjalikkusest ja eelkõige ühtlusest.

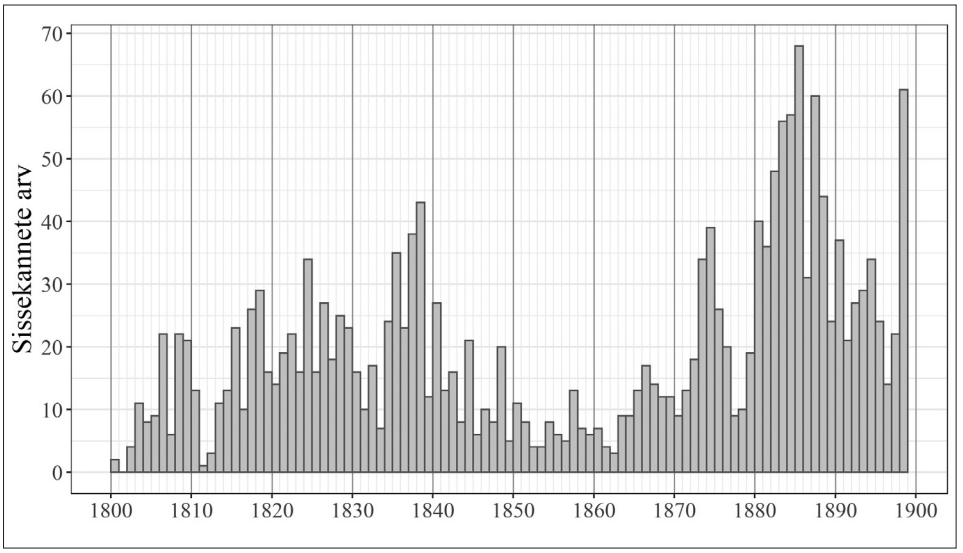
Tormid kui ekstreemsed ilmastikunähtused vajavad aga oma loomu tõttu teistsugust lähenemist kui eelkirjeldatud aegread, mis kujutavad näiteks sademete või temperatuuri pikaaegseid trende. Tormide puhul on matemaatilises mõttes tegu pigem ajas hajutatud punktidega kui kõverjoonega: see tähendab, et tänasest tormist ja üleeilsest tuulevaikusest pole kuigi hästi võimalik järeldada, et eile oleks tuul puhunud mõõduka kiirusega.[6] Üleüldse on ajaloolistes allikates tuule kiirusest haruharva juttu ning tormi tugevuse hindamiseks tuleb üldjuhul kasutada kirjeldust selle mõjust loodusele ja inimestele, mis on aga mitut moodi tõlgendatav. Probleemist, mis tekib kvalitatiivsete andmete numbriliseks teisendamisel või nende märksõnastamisel, kirjutas Heldur Palli juba 1965. aastal: „Üheks kõige enam segavaks asjaoluks on seejuures ajalooliste seoste väga suur mitmekesisus, eriti nende omavaheline põimumine. Kuna materjali statistiline uurimine näitas, et ajalooalasel uurimistööl on kõige otstarbekam kasutada hierarhilisel printsiibil koostatud märksõnastikku (ja selle põhjal üles ehitatud koodi), siis tekitas just mõistete põimumine kõige suuremaid raskusi." (Palli 1965a: 107)

Ka Euro-Climhistis on kasutusel hierarhiline kodeerimissüsteem. Iga sissekanne andmebaasis sisaldab peale kuupäeva, geograafilise asukoha koodi ja allika ilmastikunähtuse kategooriat (CAT). Näiteks kannab kood 1:1:5:3: tähendust „tugev tuul", 1:1:5:4: aga „torm" ja 1:1:5:5: juba „orkaan". Sarnaselt vastab rahesajule kood 1:1:2:7:, äikesele aga 1:1:2:8:. Neljakohaline kood 1:1:5:4: on osade kaupa loetav viisil: 1 – kirjeldus kui teabeliik, 1 – ilmastikukirjeldus, 5 – tuule tugevuse kirjeldus, 4 – tormi kirjeldus. Andmebaasi kodeerimissüsteem on seega hierarhiseeritud ja sisaldab mitusada erinevat koodi. Millal aga saab tugevast tuulest torm? See vajab iga sissekande korral eraldi kaalumist, samuti uurijatevahelisi kokkuleppeid. Kuna igal üksikjuhul on tarvis olla allikakriitiline ja vahel allikaid omavahel võrrelda, jääb tööprotsessi alati mingi veamäär, mis on ajaloouurimisel sageli paratamatu. Pidades silmas ajalooliste andmete komplekssust, allikakriitilist nõudlikkust ja kodeerimisprotsessist tulenevat tõlkekadu, on piisavalt suure hulga andmete põhjal siiski võimalik kvantitatiivseid järeldusi teha. Järgnevalt esitame mõne esmase järelduse Eesti ja Läti alade tormisuse kohta ajalooliste allikate põhjal. Nagu eespool mainitud, ei ole neid järeldusi veel võrreldud loodusteaduslike meetodite ja mõõdetud ilmavaatlustest saadud aegridadega.

Ehkki ajaliselt vanimad teadaolevad dateeritavad tuuleandmed Euro-Climhisti Balti moodulis pärinevad juba XVI sajandist, on siinseks analüüsiks võetud vaatlusperiood 1800–1900, kuna selles vahemikus on andmete tihedus piisavalt suur. Võttes andmebaasis vaatluse alla nimetatud ajavahemiku ja vaid tormid ning tugevad tuuled, saab tulemuseks 1925 sissekannet, mis on ajaloolaste töö tulemusel väga eriilmelistest allikatest numbriteks ja märksõnadeks saanud. Teave pärineb valda-

[6] Isegi tuule kiirust iseloomustavate pikkade aegridade olemasolu korral oleks tormide uurimine raskendatud. Nimelt toovad kliimaajaloos kasutusel olevad kalibreerimis- ja verifikatsioonimeetodid endaga matemaatilises plaanis kaasa ekstreemsuste n-ö mahalihvimise. See tähendab, et näiteks suurt kahju põhjustanud torm võib tuule kiirust koondava vaatlusjada kontekstis ekslikult tunduda statistilise veana. Sellist laadi probleemile on pakutud ka matemaatilisi lahendusi (vt Rodrigo 2008).

valt kahest allikagrupist. Esimese moodustab publitseerimata arhiiviaines, kust on olnud võimalik välja sorteerida tormisust kajastav info. Nii on Tarandi andmebaasi koondatud Eestimaa kubermanguvalitsuse kirjavahetust, magistraatide (Pärnu, Narva, Kuressaare) toimikuid laevanduse ja kaubanduse osas, Tallinna merekohtu materjale, kaubaettevõtete (nt Mayer & Co, Thomas Clayhills & Son) toimikuid jms. Publitseerimata materjali hulka võib arvata ka varajased ilmavaatlused (nt Palmse mõisa kalendrid ja Carl Kalki ilmavaatluse päevikud, mis asuvad Tallinna Ülikooli Akadeemilise Raamatukogu Baltika osakonnas). Nimetatud materjal katab vaatlusalust perioodi XIX sajandi algusest kuni 1840. aastateni, seejärel hakkab domineerima teine allikagrupp ehk omaaegne ajakirjandus. Balti mooduli andmetes hakkab ajakirjanduse osatähtsus muude allikate ees järk-järgult kasvama alates XVIII sajandi teisest poolest. See on mõistagi selgitatav ajakirjanduse üldise arengu ja levikuga Eesti- ja Liivimaal. Kui XVIII sajandi teisel poolel ilmusid üksikud saksakeelsed väljaanded vaid Riias, Tallinnas ja Tartus (Peegel jt 1994: 271–275), siis alates XIX sajandi teisest veerandist lisandus järjepidevalt saksa-, hiljem mittesaksakeelseid väljaandeid, ilmumine tihenes nädalalehest päevaleheks. Levinumad väljaanded valimis on ajakiri Das Inland ning ajalehed Rigasche Zeitung, Rigasche Stadtblätter, Libausche Zeitung, Revalsche Zeitung, Maarahva Näddala-leht, Perno Postimees, Eesti Postimees, Olevik, Saarlane. Nimetatud allikatest pärinevate sissekannete jaotust ajalises plaanis ja aastaaegade lõikes kujutab joonis 1.

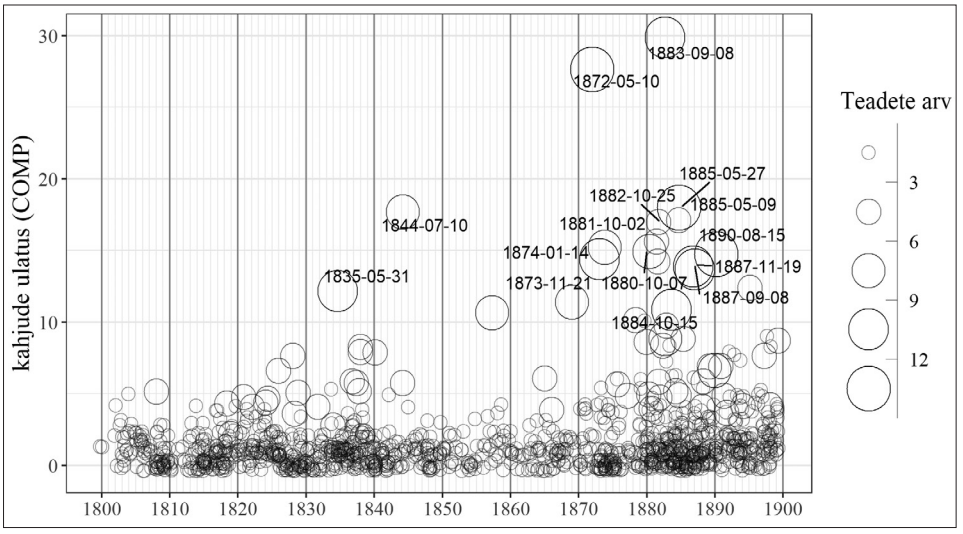


Joonis 1. XIX sajandi tormiteadete arv ja ajaline jaotus (joonis: Krister Kruusmaa).

Jooniselt 1 on näha, et ajalooliste tormiteadete hulk on tihedam kahel perioodil: esmalt 1820.–1830. aastate paiku ning siis 1880. aastatel. Siiski näitlikustab selline graafik vaid seda, kui palju on ühel või teisel aastal olnud allikates infot tormide kohta, ega ütle midagi näiteks tormide tugevuse kohta. Samuti tuleb meeles pidada, et tormide (ja teiste ilmastikunähtuste) kajastus ühiskonnas pole kunagi

täpselt võrdeline nende sageduse või ulatusega, vaid on mõjutatud vaatlejast: näiteks rannikualadel ja sadamalinnades ei äratanud talvised ja varakevadised tormid nii palju tähelepanu kui need, mis leidsid aset navigatsioonihooajal.

Peale eespool kirjeldatud nähtuse kategooria (CAT, nt torm, rahe jne) on kodeeritavad ka ühe või teise nähtuse üksikomadused, näiteks tajutud tugevus, materiaalsed kahjud ja nende ulatus, inimohvrite hulk jne, mis pärinevad eeldefineeritud iseloomustavate märksõnade nimistust (ingl *complementary*, COMP). Nende parameetrite abil on võimalik teadete ajas muutuvast kogumassist eristada tõeliselt tugevaid torme. Omistades igale tormi tagajärgi väljendavale koodsõnale arvulise väärtuse 1, on võimalik neid omavahel kokku liita. Näiteks annavad murtud puud ühe punkti, samuti nagu katkised katused, laastatud tänavad või häiritud laevaliiklus. Selline robustne süsteem on laiendatav kõigile allikatüüpidele ega sõltu oma ülevaatelisuses kuigivõrd tormi kirjeldanud autori subjektiivsusest (peale esmaseid järeldusi on võimalik erinevaid märksõnu ka kaaluda, näiteks omistada mõnele suurem, teisele väiksem väärtus). Võttes märksõnade kõrval abiks veel ühe muutuja, nimelt teadete arvu tormi kohta (suurest katastroofist kirjutavad paljud allikad), ning kombineerides selle tormi tagajärgi väljendavate punktide kogusummaga, võib XIX sajandi kõige märkimisväärsemaid torme kujutada nii, nagu näha joonisel 2.



Joonis 2. Tormid Balti kubermangudes XIX sajandil ja nende tugevus (joonis: Krister Kruusmaa).

Joonisel 2 saame nüansseerida varasemaid järeldusi: kahest suurema teadete arvuga perioodist tõuseb selgelt esile teine, 1880. aastad. Sellesse vahemikku paistab koonduvat suurem osa kõige tõsisemate tagajärgedega marutuultest. Sarnane pilt avaneb ka Põhjamere andmetest, kus tormisuses on 1883. aastal selge tipphetk (Feser jt 2014). Samuti on selle perioodi Põhjamere tormide kajastus ulatuslikum. Kvantitatiivne analüüs paigutab konteksti siinses artiklis näitena kasutatud 1872. aasta maitormi, mida on näha graafiku ülaosas. Arvulises mõttes on sellest üle vaid 1883.

aasta septembritorm, mis põhjustas allikate andmeil kogu Eesti-, Liivi- ja Kuramaal ulatuslikke kahjusid, sh mitmel pool surmasid.

Uute teadmiste valguses on võimalik kvalitatiivset lähenemist täpsustada, näiteks kõige tugevamaid torme võrreldes. Joonisel 2 saab ulatuse alusel kõige silmapaistvamad tormid jaotada kaheks: suuremat piirkonda raputanud tormid ning väiksema ulatusega tormid. Laia haardega oli 1883. aasta septembritorm, kui Tallinnas murdis „orkaanisarnane" tuul Kadriorus 600 puud ja uputas mitu laeva (EP 1883), painutas Vormsil lausa 2000 puud pikali (RiZ 1883a; Olevik 1883) ning möllas ka Riias (RiZ 1883b). 1885. aasta mai lõpupäevil teatati külmast tugevast tuulest, vihmast ja äikesest muu hulgas Valgas (ZfSL 1885), Kanepis (Olevik 1885), Liepajas (LiZ 1885b), Riias (RiZ 1885b) ja Bauskas (LiZ 1885a). Sama aasta 9. mail sadas Hallistes sõrmejämedusi raheterasid (Wirulane 1885a) ning tugev äikesetorm vihma ja rahega lammutas majasid Liivimaal Meņģeles (RiZ 1885a) ja surmas viis inimest Tallinnas (Wirulane 1885b). 15. augustil 1890 ajasid kanged tormituuled Hiiumaal randa mitu laeva (Saarlane 1890a), külvasid kahju Saaremaa kirikumõisas (Saarlane 1890b) ja Tallinnas „õnnetusse langenud laevu olla juba umbes 15 tükki teada" (Olevik 1890). Pingerea üks varasemaid torme, 10. juulist 1844, tekitas Võrumaal nii palju purustusi, et kõiki neid teateid ei hakatud tindi kokkuhoidmiseks trükkima (Das Inland 1844). Võrumaal hinnati kahjud mõnele tuhandele rublale, Oisus 3000 rublale ning Riias lausa 49 000 rublale (Das Inland 1845). Väiksema ulatusega, aga nähtavasti mitte nõrgemad tormid külvasid kahju Kuramaal Liepajas ja Ventspilsis 2. oktoobril 1881 (LiZ 1881a; LiZ 1881b). 7. oktoobril 1880 anti Riias tormile järgnenud hommikul sisse 68 tellimust purunenud akende parandamiseks (RiZ 1880). 25. oktoobril 1882 uppusid Saaremaa rannas umbes 30 kalameest (EP 1882). 14. jaanuaril 1874 paisutas edelatorm Haapsalus merd sedavõrd, et see koos jääga üle tammide linna tungis (ReZ 1874). Rannaäärsetel aladel sumasid inimesed lausa vööst saati vees (EP 1874).

Eelnevast on näha, millise iseloomu ja tugevusega on kvantitatiivse analüüsi tulemusel kõige olulisemateks liigitunud tormid. Järgmiseks võib andmestikule esitada uusi küsimusi ja otsida segavaid faktoreid, mis võiksid moonutada XIX sajandi tormisuse üldpilti. Olulisim küsimus on, millisel määral on allikate tihedus tõlgendatav suurenenud tormisusena: kas on võimalik, et 1880. aastate tormiteadete esiletõusu taga on ajakirjanduse mahu ja moe muutumine? Samuti võib oletada, et ajaliselt hilisemad tormid tekitasid rohkem purustusi, kuna infrastruktuur oli tihedam, peenem ja haavatavam: 1830. aastatel ei olnud näiteks kinnituisanud raudtee või murdunud telegraafipost Liivimaal veel probleem. Neile küsimustele võib esmased vastused anda ilma allikmaterjali suurendamata (mis on vähemalt tormide osas üsnagi ammendunud ja võiks vaid marginaalseid täiendusi pakkuda). Saab märkida, et tugevate, st enim kajastatud ja kahjusid tekitanud tormide jaotus ei paista korreleeruvat ajalehtede arvu ja ilmumissageduse ühtlase kasvuga XIX sajandi jooksul: 1840.–1860. aastad on allikates üpris tormivaesed, samuti nagu sajandi viimane kümnend. Infrastruktuuri kohta võib öelda, et mainitud tugevamate tormide ainelised kahjud puudutasid üldjuhul asju, mis olid juba ammusest ajast inimeste materiaalse elu osad, mitte XIX sajandi tehniliste murrangute viljad.

Kuidas on veel võimalik kontrollida kodeeritud andmete kõnekust ja kvantitatiivse analüüsi tulemusi? Esiteks võib kvantitatiivseid tulemusi võrrelda kvalitatiivsetega: Gotthard Schwederi uurimusest selgub, et tänapäevase andmeanalüüsi järgi äärmiselt tugevat 1872. aasta maitormi pidasid erakordseks ka tema kaasaegsed (Liiv 2021). Teiseks võib teavet võrrelda teiste andmeliikidega, näiteks kõrvutades arhiivi- ja ajakirjanduslikku materjali olemasolevate tuule kiiruse mõõtmisjadade või ökoloogiliste andmetega ja otsides neis kattuvusi. Kirjeldavatele allikatele võib esitada täiendavaid küsimusi kliimanarratiivide kohta: kas ja mil moel muutus teadlikkus ilmastikust ja kliimast Balti kubermangudes ning kas ilmast rääkimise viiside muutumine mõjutab seda, kuidas torme kajastati? Kui ühest küljest on selge, et Schwederi analüüs kinnitab 1872. aasta tormi erilisust, siis teisalt oli selliseks lähenemiseks vaja konkreetset ühiskondlikku, kultuurilist või teaduslikku huvi, mida varasematel perioodidel polnud. Laiemate hüpoteeside püstitamine ilmanarratiivide kohta ja nende kontrollimine vajab aga edasist tööd.

1872. aasta maitorm Euroopa ajaloolise klimatoloogia raames

Tormisuse rahvusvahelises võrdluses ei paista 1872. aasta torm siiski kõige enam silma. Samal aastal oli Euroopas rohkelt tugevaid tuuli. Nii möllas üks teine erakordse hävitusjõuga torm 12.–13. novembril Läänemere lõunaosas, tuues suurt kahju Taanis, Rootsi lõunaosas ja Saksamaa rannikualal ning jättes püsiva jälje kultuurimällu (vt Ostsee-Sturmflut 1872). Eesti ja Läti jäid sellest kas puutumata või pole siinmail nähtust peegeldavaid kirjalikke ülestähendusi. „Meie" hirmsast maikuu tormist pole aga omakorda jälgi mujal kui omaaegses Eesti- ja Liivimaa ajakirjanduses. Praeguseni puudub ka täpsem ülevaade, millist teed pidi maikuu raju siia üldse jõudis, ent on siiski teada, et Genova laht ja Põhja-Itaalia on piirkond, kus tekivad tugevad tsüklonid, mis jõuavad Eestisse just hiliskevadel ja suve alguses. 10–15% kõikidest siia jõudnud tormidest on vahemerelise päritoluga, ülejäänud kujunevad Atlandil. Vastav statistika on tehtud küll ajavahemiku 1948–2000 kohta, ent pakub 1872. aasta maitormi seletamiseks huvitavat mõtlemisainet. (Link, Post 2007) Ehkki puuduvad andmed ja analüüsivõimalused, et modelleerida 1872. aasta tormi laiemas kontekstis, on selge, et seda peab vaatlema üldise tormisuse tõusu ja erakordse purustusjõuga marutuulte sagenemise raames, mis XIX sajandi teisel poolel näib olevat üldine trend mitmel pool Euroopas. 1868. aasta detsembri alguses möllas Briti saartelt Austriani ulatuv maru, mille teele jäid Madalmaad, Belgia, Saksamaa, Poola ja Tšehhi. Mitmekülgselt on selle n-ö sajandi tormi mõju uuritud Kesk-Euroopa, eriti Tšehhi näitel (Brázdil jt 2017). 1870. aasta 26.–27. oktoobril tabas Tšehhit tugevuselt teine torm sel sajandil (Brázdil jt 2018). Praegu kasutatavatele andmetele toetudes pole aga neist Kesk-Euroopa XIX sajandi suurtest tormidest ükski Läänemere idakaldani ulatunud.

Siinse regiooni kliima eripära on andnud tõuke neid erijooni konkretiseerida ning täiendada Euroopa üldisemat pilti. Mainitud üksikjuhtumid on teaduslikus

mõttes õpetlikud ning annavad selget tunnistust tormiuuringute rahvusvahelistumisest, õieti selle vältimatusest. Tormide ja tormisuse uurimise hetkeseisu ei ole võimalik Euroopa kontekstis kokkuvõtvalt kirjeldada: hoolimata püüdlusest seda teha, on uurijate-kirjutajate koorekiht tunnistanud, et nende võimuses pole hõlmata tehtud ja tehtavat tööd kogumahus (Feser jt 2014). Samamoodi ettevaatlikud ollakse suuremate üldistuste tegemisel. Igal endast lugupidaval uurijal või uurimisgrupil on oma meelisprobleemid, lähenemisviisid ning enamasti ka geograafiline piirkond, mille kohta söandatakse järeldusi teha. Siinsed klimatoloogid on kasutanud oma töödes põhiliselt mõõtmisandmeid ning seetõttu on vaatluse all olnud maksimaalselt viimased 150 aastat, mille kohta leidub pikemaid vaatlusridu. Kuid enamasti on piirdutud viimase poolsajandi või isegi lühema perioodiga (Sepp jt 2018). Ajaloolises klimatoloogias kasutatavaid kirjalikke ajalooallikad, mida on samuti võimalik kvantifitseerida ja mis pakuvad võimalust ajas kaugemale minna, pole Tarandi ja Kallise pingutustest hoolimata tormiuuringutes seni kasutatud (Tarand, Kallis 2017).

Ajaloolisel klimatoloogial on mitu sihiseadet. Esmalt muidugi kliima kui füüsikaliste karakteristikute kogumi muutumise dünaamika selgitamine. Täpsem minevikupilt aitab nii mõista olevikku kui ka modelleerida tulevikku, ennustada kliima muutumise globaalsemaid ja regionaalsemaid stsenaariume. Teiseks suunaks on uurida kliimast tingitud ühiskondlikke reaktsioone. Kõige laiemalt võib neid nimetada ka kliima mõjuks kultuurile. Inimene on kohanemisvõimeline ning otsib kliimast tingitud kitsikusest väljapääsu. Milliseid strateegiaid ja taktikaid sealjuures rakendatakse, sõltub ühiskonna tundlikkusest keskkonnamuutuste suhtes ehk ühiskondliku haavatavuse määrast (vt Otto jt 2017). Nendegi probleemide lahendamine on seda edukam, mida ulatuslikumad on kasutada olevad andmebaasid (Brázdil jt 2005; Pfister 1999).

Globaalse kliimasüsteemi toimimise mõistmiseks pakuvad digipöördega tekkinud võimalused andmetöötluseks vähemalt teoreetiliselt suurepäraseid võimalusi. Ajalooline klimatoloogia on riigipiire ületava iseloomuga, sest ilm ja kliima ei tunnista inimese kehtestatud tõkkeid. Küll võib aga nende mõju olla erinev sõltuvalt ühiskondliku haavatavuse erijoontest. Viimase kümne aasta sees on hoogustunud initsiatiiv maailma eri paigus koostatud suuremate ja väiksemate kliimaandmebaaside ühendamiseks, mida näitlikustab ka Tarandi andmebaasi üleviimine EuroClimhisti. Kindla süsteemi järgi eelnevalt korrastatud (n-ö samasse keelde tõlgitud) andmete ühendamine annab võimaluse esitada uusi küsimusi ja saada kliimasüsteemi toimimise kohta täpsemaid vastuseid. Kokkulepitud printsiipide järgimine andmete sisestamisel hõlbustab ja kiirendab oluliselt täiendava informatsiooni ladestumist, mida saadakse uute allikate läbitöötamisel.

Ambitsioonikaid initsiatiive on mitmeid. Üheks seesuguseks on näiteks algatus nimega Tambora (Tambora; Riemann jt 2015). See loodi aastal, mil möödus 200 aastat samanimelise vulkaani purskest, mida peetakse millenniumi üheks võimsaimaks. Mitmel pool, näiteks Kanadas on aga tehtud algust vanade ilmavaatlusjaamade andmete digitaalsele alusele viimisega. Projekt DRAW (Data Rescue: Archives and Weather) on võtnud eesmärgiks McGilli observatooriumi (tegutses aastatel 1874–1953) andmestiku digiteerimise, rakendades vabatahtlike kaasamist ehk ühisloomet.

Andmete kogumine erialainimeste valvsa pilgu all on tõhus algatus, sest märkmetena on andmeid 10 000 leheküljel ja üksikuid sissekandeid u 4 miljonit (Siebera, Slonosky 2019). Ühendkuningriigis on käsil andmebaasi TEMPEST (Tracking Extremes of Meteorological Phenomena Experienced in Space and Time) arendamine, milles osalevad nii riiklikud institutsioonid (nt Ühendkuningriigi meteoroloogiaamet) kui ka asjast huvitatud seltsingud ja eraisikud (Veale jt 2017). Veelgi ambitsioonikam on kümmekond aastat tagasi käivitatud ulatuslik rahvusvaheline projekt ACRE 2010 (The Atmospheric Circulation Reconstructions over the Earth), millel on hulgaliselt alaprojekte üle kogu maailma.

Kokkuvõte

Järjest suuremaks muutuvad andmebaasid on ajaloouurimises tavalised nähtused juba pool sajandit. Digitehnoloogiline areng on kaasa aidanud aina kasvavate andmete kiiremale analüüsile, samal ajal võimaldades valdkondadevahelist ja -ülest koostööd. Andres Tarandi perfokaartidega alustatud Eesti kliima andmebaasist 1970. aastal on saanud XXI sajandil Euroopa suurima kliimaandmebaasi Euro-Climhisti osa Berni ülikoolis. Eesti andmetest on välja kasvamas selle Balti moodul, mis peaks pikemas perspektiivis hõlmama kolme Balti riiki ja Soomet. Nõnda kasvab võimalus võrrelda kohalikke teadmisi kliimanähtustest ning saada laiem ja täpsem pilt mineviku kliima dünaamikast laiemas regioonis.

Mida saab öelda 1872. aasta maitormi kohta ja kas digitaalne pööre on kaasa aidanud selle uurimisele? Konkreetse tormi kirjeldamise osas oleme Schwederi artiklis esitatust ainult mõnes detailis edasi jõudnud. Lisandunud on aga võimalus asetada nähtus teiste tormide konteksti nii kohalikul kui ka laiemal tasandil. Pole kahtlust, et 1872. aasta torm kuulub Liivimaa võimsamate tormisündmuste hulka, kuid talle ei kuulu esikoht, mille au osaliseks saab teadaolevate andmete järgi 1883. aasta sügistorm. Läänemere idaosa tormisusel ja kliimal tervikuna on Lääne-Euroopaga võrreldes erijooni. Ehkki kliimaekstreemsustel on regiooniti oma nägu, võivad need suures plaanis olla omavahel seotud. Suurt pilti mineviku kliima arengu üldistest trendidest ja kohalikest erinevustest on võimalik luua ainult andmebaasidesse kogutud materjali kaudu. Kodeerimine aitab selles kontekstis jõuda võrdluseks vajaliku materjalihulgani. Erinevates keeltes allikateated on muudetud numbrites väljendunud koodideks, olles seeläbi hõlpsamini lingitavad ka teiste maailma kliimaandmebaasidega. Ka siis, kui usk andmebaasidesse ja teadusse jääb vahepeal nõrgaks ja tekib küsimus, kas nendega kaasnev ressursikulu käib käsikäes teadmiste juurdekasvuga, pole kahtlust, et Balti kliimamooduli loomine Euroopa suurimasse ajaloolise kliima andmebaasi pakub suurepäraseid uurimisvõimalusi.

Kvantitatiivse ajaloouurimise üks suuremaid probleeme paistab olevat, kuidas kaasa minna kiiresti arendatavate digitaalsete võimalustega ning kuidas toime tulla süsteemidega, mis ei pruugi enam olla intuitiivselt rakendatavad ei andmebaasi täitja ega iga potentsiaalse kasutaja jaoks. Kui andmete kogumine muutub digiajastul uuesti vaid kitsa teadlasteringi hobiks ja andmed pole kasutatavad laiema publiku

jaoks, pole digitaalne andmebaas kartoteegist parem. Kas andmebaasidel on tulevikku, sõltub sellest, kas need jätkuvalt moderniseeruvad ja ühilduvad teiste andmebaasidega, ning sellest, kui operatiivselt on võimalik infot lisada ja kui kiiresti seda analüüsida.

Selleks et kliimaandmebaasidesse kogutud lugudest võiksid saada suuremad narratiivid, on ennekõike vaja interdistsiplinaarset koostööd mineviku kliima modelleerimisel või täiesti uute meetodite ja narratiivide loomist. Mineviku ilmaolukordade ja -sündmuste ning pikemate kliimatrendidega arvestamine peaks saama ajalookirjutuses ja -õpetamises sama normaalseks nagu poliitilise, majandusliku või kultuurilise taustsüsteemi arvesse võtmine ükskõik millise fakti tõlgendamisel.

Artikli valmimist on toetanud Tallinna Ülikooli uuringufondi projekt „Tormid – Eesti kliima ekstreemsused ja nende mõju ühiskonnale" (TF5516) ning Eesti Teadusagentuuri projekt „Eesti keskkonnaliikumine 20. sajandil: ideoloogia, diskursid, praktikad" (PRG908).

MUUSEUMIKOGUD

Tartu Ülikooli muuseum (TÜM)
ÜAM 1565:1–21 Ar – Kliimaandmed. Üleandja Andres Tarand.

PUBLITSEERIMATA INTERVJUUD

Pfister, Christian, veebruar 2021. Intervjueerinud Ulrike Plath.
Tarand, Andres, jaanuar 2021. Intervjueerinud Kaarel Vanamölder.

VEEBIVARAD

Euro-Climhist. https://www.euroclimhist.unibe.ch/en
Digar. Eesti artiklid. https://dea.digar.ee
Ostsee-Sturmflut 1872. Koost Thomas Sävert. http://www.naturgewalten.de/sturmflut1872.html (30. V 2021).
Periodika. Latvijas Nacionālā Digitālā Bibliotēka. http://periodika.lv
Tambora. https://www.tambora.org

KIRJANDUS

Borodkin, Leonid 1993. History and computing in the USSR and Russia: Retrospect, state of the art, perspectives. – History and Computing in Eastern Europe. (Halbgraue Reihe zur Historischen Fachinformatik A21.) Toim Leonid Borodkin, Wolfgang Levermann. St. Katharinen: Scripta Mercaturae Verlag, lk 7–20.
Brázdil, Rudolf; Pfister, Christian; Wanner, Heinz; von Storch, Hans 2005. Historical climatology in Europe – the state of the art. – Climatic Change, kd 70, nr 3, lk 363–430. https://doi.org/10.1007/s10584-005-5924-1

Brázdil, Rudolf; Szabó, Péter; Stucki, Peter; Dobrovolný, Ladislava; Řezníčková, Petr; Kotyza, Oldřich; Valášek, Hubert; Melo, Marián; Suchánková, Silvie; Dolák, Lukáš; Chromá, Kateřina 2017. The extraordinary windstorm of 7 December 1868 in the Czech Lands and its central European context. – International Journal of Climatology, kd 37, nr S1, lk 14–29. https://doi.org/10.1002/joc.4973

Brázdil, Rudolf; Szabó, Péter; Stucki, Peter; Dobrovolný, Ladislava; Řezníčková, Petr; Kotyza, Oldřich; Valášek, Hubert; Dolák, Lukáš; Zahradníček, Pavel; Suchánková, Silvie 2018. The second most disastrous windstorm of the nineteenth century in the Czech Lands, 26–27 October 1870. – Theoretical and Applied Climatology, kd 132, lk 1201–1216. https://doi.org/10.1007/s00704-017-2146-1

Das Inland 1844 = Werro, den 7. August. – Das Inland 15. (27.) VIII, nr 33, lk 523–524.

Das Inland 1845 = Riga. Ueber die Folgen des Unglücks... – Das Inland 9. (21.) I, nr 2, lk 29–30.

EP 1874 = Omalt maalt. Haapsalust. Kangest tormidest... – Eesti Postimees ehk Näddalaleht 6. II, nr 6, lk 32.

EP 1882 = Isamaalt. Saare maalt. Nagu juba minewal... – Eesti Postimees ehk Näddalaleht 17. XI, nr 47, lk 2.

EP 1883 = Uuemad sõnumid. Tallinnas olnud... – Eesti Postimees ehk Näddalaleht 14. IX, nr 37, lk 3.

Feser, Frauke; Barcikowska, Monika; Krueger, Oelof; Schenk, Frederik; Weisse, Ralf; Xia, Lili 2014. Storminess over the North Atlantic and northwestern Europe – A review. – Quarterly Journal of the Royal Meteorological Society, kd 141, lk 350–382. https://doi.org/10.1002/qj.2364

Huhtamaa, Heli; Jaagus, Jaak; Koff, Tiiu; Kruusmaa, Krister; Plath, Ulrike; Raudkivi, Priit; Rohr, Christian; Tarand, Andres; Vaasmaa, Tiit; Vanamölder, Kaarel; Vandel, Egert; Widmer, Tamara T. (ilmumas). Northern Baltic Sea storminess variability and trends in the 19th century based on early meteorological observations, descriptive documents, and paleogeographical evidence. – Climate of the Past.

Jansen, Ea 1972. 25 aastat Eesti NSV Teaduste Akadeemia Ajaloo Instituuti. – Eesti NSV Teaduste Akadeemia toimetised. Ühiskonnateadused, kd 21, nr 3, lk 331–338.

Juhan Kahk 1988 [1980]. – Akadeemikud, osa 47. Rež Mare Viljarand, saatejuht Andres Vihalem. Eesti Televisioon 1. I. https://arhiiv.err.ee/vaata/akadeemikud-juhan-kahk (18. I 2021).

Kahk, Juhan 1964. Elektronarvutite kasutamise võimalustest ajaloo uurimisel. – Eesti NSV Teaduste Akadeemia toimetised. Ühiskonnateaduste seeria, nr 1, lk 354–360.

Kahk, Juhan 1971. Computer analysis of socio-economic development in Estonia in the first half of the nineteenth century. – Mathematics in the Archaeological and Historical Sciences. Proceedings of the Anglo-Romanian Conference, Mamaia, 1970, organized by The Royal Society of London and the Academy of the Socialist Republic of Romania. Toim Frank Roy Hodson. Edinburgh: Edinburgh University Press, lk 500–504.

Kahk, Juhan 1974. Eine Zwischenbilanz der Anwendung mathematischer Methoden in der sowjetischen Geschichtswissenschaft. – Jahrbuch für Wirtschaftsgeschichte, kd 2, lk 223–230.

Kahk, Juhan; Ligi, Herbert 1964. K voprosu ob èkonomičeskom položenii i feodal'nyh povinnostjah krest'jan v Èstljandskoj gubernii v XVIII v. (Opyt primenenija èlektronno-sčetnyh mašin v istoričeskom issledovanii.) – Ežegodnik po agrarnoj istorii Vostočnoj Evropy 1962 g. Peatoim Viktor Kornelievič Jacunskij. Minsk: Nauka i tehnika, lk 43–58. [Юхан Кахк, Херберт Лиги, К вопросу об экономическом положении и феодальных повинностях крестьян в Эстляндской губернии в XVIII в. (Опыт применения электронно-счетных машин в историческом исследовании). – Ежегодник по аграрной истории Восточной Европы 1962 г. Ответственный редактор Виктор Корнелиевич Яцунский. Минск: Наука и техника.]

Krikmann, Arvo 2006. Infotehnoloogia minu elus. – Võim ja kultuur 2. Toim Mare Kõiva. Tartu: Eesti Kirjandusmuuseum, lk 11–26. https://doi.org/10.7592/VK2.2006.krikmann

Lemny, Stefan 2017. „L'historien de demain sera programmeur...": Emmanuel Le Roy Ladurie et les défis de la science. – L'histoire à la BnF. https://histoirebnf.hypotheses.org/1505 (13. II 2021).

Le Roy Ladurie, Emmanuel 1967. Histoire du climat depuis l'an mil. Paris: Editions Flammarion.

Le Roy Ladurie, Emmanuel 1968. La fin des érudits. L'historien de demain sera programmeur ou il ne sera pas. – Nouvel Observateur 8. V.

Le Roy Ladurie, Emmanuel 1973. Le territoire de l'historien. Paris: Gallimard.

Le Roy Ladurie, Emmanuel 2014. Une vie avec l'histoire. Paris: Tallandier. https://doi.org/10.3917/talla.ladur.2014.01

Liidemaa, Helene 1992. Tartu Ülikooli Meteoroloogia Observatooriumi rajamine ja meteoroloogia areng Eestis aastail 1865–1918. – Teaduse ajaloo lehekülgi Eestist. Kd VIII. Toim Hermann Mürk. Tallinn: Valgus, lk 63–79.

Liiv, Anna Helena 2021. 1872. aasta maikuu torm ja selle kajastamine Eesti-, Liivi- ja Kuramaa ajalehtedes. Bakalaureusetöö. Tallinn: Tallinna Ülikool.

Link, P.; Post, Piia 2007. On the spatial and temporal variability of cyclones in the Baltic Sea region. – Uurimusi Eesti kliimast. Studies on climate of Estonia. (Publicationes Instituti Geographici Universitatis Tartuensis 102.) Toim Jaan Jaagus. Tartu: Tartu Ülikooli Kirjastus, lk 7–17.

LiZ 1881a = Locales. Auch in Windau... – Libausche Zeitung 6. (28.) X, nr 232, lk 3.

LiZ 1881b = Locales. Der starke Sturm ... – Libausche Zeitung 5. (27.) X, nr 231, lk 3.

LiZ 1885a = Inland. Bauske. Ueber das Unwetter... – Libausche Zeitung 10. (22.) VI, nr 132, lk 2.

LiZ 1885b = Lokales. (Vom Hafen). Schon gestern... – Libausche Zeitung 29. V (10. VI), nr 122, lk 3.

Mahoney, Michael 1988. The history of computing in the history of technology. – Annals of the History of Computing, kd 10, lk 113–125. https://doi.org/10.1109/MAHC.1988.10011

Moberg, Anders; Bergström, Hans; Krigsman, Josefin Ruiz; Svanerd, Ola 2002. Daily air temperature and pressure series for Stockholm (1756–1998). – Climatic Change, kd 53, lk 171–212.

Must, Aadu 1998. Pilk Eesti komputeriseeritud ajaloolaste tööpõllule. – Ajalooline Ajakiri, kd 103, nr 4, lk 15–24.

Mürk, Herman 1992. Meteoroloogia- ja klimatoloogiaalasest uurimistööst Eestis 1920–1940. – Teaduse ajaloo lehekülgi Eestist. Kd VIII. Toim H. Mürk. Tallinn: Valgus, lk 80–98.

Neithal, Martin 1963. Sälkkaardid kirjanduseuurija abilistena. – Keel ja Kirjandus, nr 8, lk 504–506.

Olevik 1883 = Eestimaalt. Tallinnast. – Olevik 26. IX, nr 37, lk 3.

Olevik 1885 = Eestimaalt. Kanapääst. – Olevik 10. VI, nr 24, lk 1.

Olevik 1890 = Tallinnas, 23. aug. – Olevik 27. VIII, nr 35, lk 879.

Otto, Ilona M.; Reckien, Diana; Reyer, Christopher; Marcus, Rachel; Le Masson, Virginie; Jones, Lindsey; Norton, Andrew; Serdeczny, Olivia 2017. Social vulnerability to climate change: A review of concepts and evidence. – Regional Environmental Change, kd 17, nr 6, lk 1651–1662. https://doi.org/10.1007/s10113-017-1105-9

Paju, Petri 2020. The long road to 'digital history': History of computer-assisted research of the past in Finland since the 1960s. – Digital Histories. Emergent Approaches within the New Digital History. Toim Mats Fridlund, Mila Oiva, P. Paju. Helsinki: Helsinki University Press, lk 21–44. https://doi.org/10.33134/HUP-5-2

Palli, Heldur 1965a = Hel'dur Palli, Opyt ispol'zovanija perfokart v Institute Istorii Akademii Nauk Èstonskoj SSR. – Eesti NSV Teaduste Akadeemia toimetised. Izvestija Akademii Nauk Èstonskoj SSR. Ühiskonnateaduste Seeria. Serija obščestvennyh nauk, nr 1, lk 98–109. [Хельдур Палли, Опыт использования перфокарт в Институте Истории Академии Наук Эстонской ССР. – Eesti NSV Teaduste Akadeemia toimetised. Известия Академии наук Эстонской ССР. Ühiskonnateaduste seeria. Серия общественных наук.]

Palli, Heldur 1965b. Perfokaartide kasutamisest ajalooteaduses. Algteadmisi. Tartu: Tartu Riiklik Ülikool.

Palli, Heldur 1967. Perfokaart kodu-uurijatele: Sälkperfokaartide kasutamisest looduskaitseobjektide ja kultuurimälestiste kohta informatsiooni kogumisel. – Kodu-uurimise teateid. Eesti NSV Teaduste Akadeemia, kodu-uurimise komisjon, kd 7. Tallinn: Eesti NSV Teaduste Akadeemia, lk 34–47.

Palli, Heldur 1969. Lochkartenverfahren u. mathematische Methoden in der Geschichtswissenschaft. Erfahrungen in Institut für Geschichte der Akademie der Wissenschaften der Estnischen SSR. – Zeitschrift für Geschichtswissenschaften, kd 17, lk 504–514.

Palm, Ille 1992. Meteoroloogia õpetamisest Tartu ülikoolis aastail 1919–1940. – Teaduse ajaloo lehekülgi Eestist. Kd VIII. Toim Herman Mürk. Tallinn: Valgus, lk 99–122.

Peegel, Juhan; Aru, Krista; Issakov, Sergei; Jansen, Ea; Lauk, Epp 1994. Eesti ajakirjanduse teed ja ristteed. Eesti ajakirjanduse arengust (XVII sajandist XX sajandini). Tartu: Tartu Ülikool.

Peterson, Richard 1992. Johannes Letzmann: A pioneer in the study of tornadoes. – Weather and Forecasting, kd 7, nr 1, lk 166–184. https://doi.org/10.1175/1520-0434(1992)007<0166:JLAPIT>2.0.CO;2

Pfister, Christian 1999. Wetternachhersage: 500 Jahre Klimavariationen und Naturkatastrophen (1496–1995). Bern–Stuttgart–Wien: Verlag Paul Haupt.

Pfister, Christian; White, Sam; Mauelshagen, Franz (toim) 2018. The Palgrave Handbook of Climate History. London: Palgrave Macmillan.

Rahi, Aigi; Must, Aadu 1993. Katkenud töö. Herbert Ligi: Eesti rahvastiku kaotused Nõukogude Liidu vanglates ja sunnitöölaagrites. – Kleio. Ajaloo ajakiri, kd 8, lk 8–15.

Raudkivi, Priit 2010. Maa – meie ema, ilm – meie isa. Märkmeid looduse rollist Liivimaa 14. sajandi ajaloos. – Acta Historica Tallinnensia, kd 15, nr 1, lk 3–23. https://doi.org/10.3176/hist.2010.1.01

Raudkivi, Priit 2014. Islandi 1783. aasta vulkaanipurske võimalikust mõjust Eestis. Keskkonnaajalooline arutlus. – Acta Historica Tallinnensia, kd 20, nr 1, lk 51–73. https://doi.org/10.3176/hist.2014.1.02

Raudkivi, Priit 2016a. Der Vulkanismus als historischer Faktor. – Forschungen zur Baltischen Geschichte II. Toim Mati Laur, Karsten Brüggemann. Tartu: Akadeemiline Ajalooselts, lk 181–196.

Raudkivi, Priit 2016b. Ilm teeb ajalugu. Aspekte talurahva ühiskondlikust haavatavusest Eesti- ja Liivimaal 19. sajandi algul. – Acta Historica Tallinnensia, kd 22, nr 1, lk 18–40. https://doi.org/10.3176/hist.2016.1.02

ReZ 1872 = Inland. Reval, 13. Mai. Dem heftigen Sturme... – Revalsche Zeitung 13. (25.) V, nr 110, lk 1.

ReZ 1874 = Inland. Aus Hapsal, 16. Januar, wird... – Revalsche Zeitung 19. (31.) I, nr 16, lk 1.

Riemann, Dirk; Glaser, Rüdiger; Kahle, Michael; Vogt, Steffen 2015. The CRE tambora.org – new data and tools for collaborative research in climate and environmental history. – Geoscience Data Journal, kd 2, lk 63–77. https://doi.org/10.1002/gdj3.30

RiZ 1872a = Inland. Riga, 11. Mai. Gestern wurde... – Rigasche Zeitung 11. (23.) V, nr 108, lk 1.

RiZ 1872b = Inland. Riga, 19. Mai. – Rigasche Zeitung 19. (31.) V, nr 119, lk 2.

RiZ 1872c = Rigasche Zeitung 13. (25.) V, nr 110, lk 1–2, 6.

RiZ 1880 = Locales. (Sturmnotiz. – Eingesandt.). – Rigasche Zeitung 10. (22.) X, nr 236, lk 6.

RiZ 1883a = Inland. Estland. Privatnachrichten zufolge... – Rigasche Zeitung 22. IX (4. X), nr 218, lk 2.

RiZ 1883b = Witterungsbeobachtungen in Riga. – Rigasche Zeitung 10. (22.) IX, nr 208, lk 6.

RiZ 1885a = Inland. Der gewaltige Sturm... – Rigasche Zeitung 16. (28) V, nr 109, lk 1.

RiZ 1885b = Lokales. (Ein sehr heftiger Südweststurm). – Rigasche Zeitung 29. V (10. VI), nr 120, lk 6.

Rodrigo, Fernando 2008. A new method to reconstruct low-frequency climatic variability from documentary sources: application to winter rainfall series in Andalusia (Southern Spain) from 1501 to 2000. – Climatic Change, kd 87, nr 3, lk 471–487. https://doi.org/10.1007/s10584-007-9312-x

Rohr, Christian; Camenisch, Chantal; Widmer, Tamara (ilmumas). Dem Klima seit dem Mittelalter auf der Spur – projektorientiertes Arbeiten in Sekundarschule und Gymnasium mit der Datenbank Euro-Climhist.

Rosenberg, Tiit 1998. Juhan Kahk agraarajaloolasena. – Ajalooline Ajakiri, kd 103, nr 4, lk 51–64.

Ruusalepp, Raivo 1993. Computing and history in Estonia. Historical survey. – History and Computing in Eastern Europe. (Halbgraue Reihe zur Historischen Fachinformatik A21.) Toim Leonid Borodkin, Wolfgang Levermann. St. Katharinen: Scripta Mercaturae Verlag, lk 21–26.

Ruusalepp, Raivo 1996. Ajalooline informaatika – mis see veel on? – Kleio. Ajaloo ajakiri, kd 15, nr 1, lk 37–41.

Saarlane 1890a = Isamaalt. Hiiumaalt. 16. augustil möllas... – Saarlane 11. IX, nr 37, lk 2.

Saarlane 1890b = Isamaalt. Kaarmalt. Suur torm. – Saarlane 4. XI, nr 36, lk 2.

Salum, Henn 1999. Meenutusi Küberneetika Instituudi algajaloost. – Küberneetika Instituudi infoleht 2/99. https://www.ioc.ee/infoleht/1999_2/infoleht.html (18. I 2021).

Schweder, Gotthard 1873. Der Hagelsturm des 10. mai 1872. – Arbeiten des Naturforscher-Vereins zu Riga N.F.V. Riga: Häcker.

Sepp, Mait; Post, Piia; Mändla, Kaupo; Aunap, Raivo 2018. On cyclones entering the Baltic Sea region. – Boreal Environment Research, kd 23, lk 1–14.

Siebera, Renée; Slonosky, Victoria 2019. Developing a flexible platform for crowdsourcing historical weather records. – Historical Methods, kd 52, nr 3, lk 164–177. https://doi.org/10.1080/01615440.2018.1558138

Somelar, Erik 1993. Kastike, Kastike Laua Peal. – Kleio. Ajaloo ajakiri, nr 7, lk 59–60.

ZfSL 1885 = Inland. Walk. (Unwetter.) – Zeitung für Stadt und Land, 4. (16.) VI, lk 2.

Tamm, Marek 2020. Eesti ajalooteadus 21. sajandil: sissejuhatuseks. – Acta Historica Tallinnensia, kd 26, nr 1, lk 3–45. https://doi.org/10.3176/hist.2020.1.01

Tarand, Andres 2013. Ajaloolise kliima kartoteek. Tallinna Botaanikaaed 1975–1990. Käsikiri. Tartu Ülikooli muuseum, ÜAM 1565:25 Ar.

Tarand, Andres; Jaagus, Jaak; Kallis, Ain 2013. Eesti kliima minevikus ja tänapäeval. Tartu: Tartu Ülikooli Kirjastus.

Tarand, Andres; Kallis, Ain 2017. Eesti tornaadod. Tallinn: Varrak.

Tartu Riikliku Ülikooli struktuur ja isikkoosseis 83/84. Teatmik (andmed seisuga 1. oktoober 1983). Tartu, 1983.

Tõugu, Enn 2018. Arvutid, küberruum ja tehismõistus. Noppeid arvutite imepärasest eduloost. (Acta Universitatis Tallinnensis (Realia et Naturalia).) Tallinn: Tallinna Ülikooli Kirjastus.

Vahtre, Sulev 1970a. Hingeloendused Eestimaa kubermangus (1782–1858) ja nende andmed talurahva ajaloo allikatena. Tartu: Tartu Riiklik Ülikool.

Vahtre, Sulev 1970b. Ilmastikuoludest Eestis XVIII ja XIX sajandil (kuni 1870) ja nende mõjust põllumajandusele ning talurahva olukorrale. (Tartu Ülikooli toimetised, vihik 258.) Tartu: Tartu Riiklik Ülikool.

Veale, Lucy; Endfield, Georgina H.; Davies, Sarah J.; Macdonald, Neil; Naylor, Simon; Royer, Marie-Jeanne; Bowen, James; Tyler-Jones, Richard; Jones, Cerys 2017. Dealing with the deluge of historical weather data: The example of the TEMPEST database. – Geo. Geography and Environment, kd 4, nr 2, lk 1–15. https://doi.org/10.1002/geo2.39

Wild, Heinrich 1881. Die Temperatur-Verhältnisse des Russischen Reiches. St. Petersburg: Kaiserlichen Akademie der Wissenschaften.

Wirulane 1885a = Eesti maalt. Hallistest kirjutatakse: 9. mail... – Wirulane: Poliitika, kirjanduse, teaduse, põllu- ja käsitöö ajaleht 21. V, nr 21, lk 2.

Wirulane 1885b = Uuemad sönumed. Neljapäew 9. Mail tabas... – Wirulane: Poliitika, kirjanduse, teaduse, põllu- ja käsitöö ajaleht 11. V, nr 20, lk 3.

Ulrike Plath (snd 1972), PhD, Tallinna Ülikooli humanitaarteaduste instituudi baltisaksa uuringute ja keskkonnaajaloo professor (Narva mnt 25, 10120 Tallinn), Eesti Teaduste Akadeemia Underi ja Tuglase Kirjanduskeskuse vanemteadur, plath@tlu.ee

Priit Raudkivi (snd 1954), PhD, Tallinna Ülikooli humanitaarteaduste instituudi vanemteadur (Narva mnt 25, 10120 Tallinn), raudkivi@tlu.ee

Kaarel Vanamölder (snd 1981), PhD, Tallinna Ülikooli humanitaarteaduste instituudi kommunikatsiooni- ja praktilise ajaloo dotsent (Narva mnt 25, 10120 Tallinn), kaarel.vanamolder@tlu.ee

Krister Kruusmaa (snd 1996), BA, L'Université Paris Sciences & Lettres, Ecole nationale des chartes, digihumanitaaria magistrant (65, rue de Richelieu, 75002 Paris), krister.kruusmaa@hotmail.com

Anna Helena Liiv (snd 1997), BA, Tallinna Ülikooli humanitaarteaduste instituudi ajaloo magistrant (Narva mnt 25, 10120 Tallinn), annaliiv@tlu.ee

Kann man das Klima kodieren? Zur digitalen Wende in der historischen Klimaforschung in Estland

Stichwörter: Geschichte, Technik und Forschung, digitale Wende, Baltische Klimageschichte, Datenbank, Presse

In einigen Forschungsbereichen der estnischen Geschichtswissenschaft wie etwa der Agrargeschichte hielten digitale Methoden und Techniken bereits in den 1970er Jahren ihren Einzug. Dies gilt auch für die Klimageschichte, die sich in dieser Zeit wie im übrigen Europa im Übergangsgebiet zwischen Geographie und Geschichtswissenschaft entwickelte. Anhand eines der größten Sturmereignisse des Baltikums, dem Mai-Sturm von 1872, gibt der Artikel einen Überblick über die Phasen der digitalen Wende, die langjährige internationale Zusammenarbeit und die interdisziplinären Forschungsmöglichkeiten der baltischen Klimageschichte. Grosse Datenmengen, die für die Rekonstruktion und Bewertung des historisches Klimas notwendig sind, wurden in der ersten Phase der digitalen Wende (1970–1990) per Hand auf Lochkarten übertragen und in der zweiten Phase (1990–2017) digitalisiert. In der dritten Phase (ab 2017) werden digitale Methoden für die schnellere Datenerhebung und Kodierung eingesetzt, zudem wird an der interdisziplinären Auswertung der Datensätze gearbeitet. Die von Andres Tarand gesammelten Daten der Estnischen Klimadatenbank basieren für das 19. Jahrhundert zum überwiegenden Großteil auf Zeitungsberichten. Die Validierung dieser spezifischen Quellengattung anhand von gemessenen Daten der Wetterstationen oder naturwissenschaftlicher Messreihen steht noch aus. Daher vermag der Artikel keinerlei Aussagen über den Anstieg der

Stürme zu treffen, der für andere Länder Bestätigung gefunden hat. Auch ohne diese noch ausstehende Bestätigung muss dem Klima als einem Hauptfaktor historischer Entwicklung in Zukunft weitaus mehr Beachtung geschenkt werden.

Ulrike Plath (geb. 1972), Dr. phil., Professor für deutschbaltische Studien und Umweltgeschichte, Universität Tallinn, Institut für Geisteswissenschaften (Narva mnt 25, 10120 Tallinn), Wissenschaftliche Mitarbeiterin, Under-und-Tuglas-Literaturzentrum der Estnischen Akademie der Wissenschaften, plath@tlu.ee

Priit Raudkivi (geb. 1954), Dr. phil, Wissenschaftlicher Mitarbeiter, Universität Tallinn, Institut für Geisteswissenschaften (Narva mnt 25, 10120 Tallinn), raudkivi@tlu.ee

Kaarel Vanamölder (geb. 1981), Dr. phil., Dozent für Praktische und Kommunikationsgeschichte, Universität Tallinn, Institut für Geisteswissenschaften (Narva mnt 25, 10120 Tallinn), kaarel.vanamolder@tlu.ee

Krister Kruusmaa (geb. 1996), Magistrant im Fachbereich Digitale Geisteswissenschaften, L'Université Paris Sciences & Lettres, Ecole nationale des chartes (65, rue de Richelieu, 75002 Paris), krister.kruusmaa@hotmail.com

Anna Helena Liiv (geb. 1997), Magistrant im Fachbereich Geschichte an der Universität Tallinn, Institut für Geisteswissenschaften (Narva mnt 25, 10120 Tallinn), annaliiv@tlu.ee

ÕNNITLEME

AUGUST

13. **Ilmar Vene**, kirjanik – 70
15. **Eduard Tüür**, kirjanik – 60
29. **Anu Haak**, keeleteadlane – 80

SEPTEMBER

3. **Jaan Rannap**, kirjanik – 90
6. **Asko Künnap**, kirjanik – 50
9. **Vladislav Koržets**, kirjanik – 70
21. **Tiia Kriisa**, kirjanik, näitleja – 80
22. **Tõnu Tannberg**, ajaloolane – 60
30. **Tiit Maksim**, ES-i liige – 80

JÄRGMISTES NUMBRITES:

- Baltisaksa ossianismist
- Täiendust XVII sajandi eesti luulele
- Paronüümide probleem eesti keeles

9 770131 144034
TELLIMISE INDEKS 78167
ISSN 0131-1441
e-ISSN 2346-6014

Hind 2.80 €