https://doi.org/10.54013/kk795a2

Arvutuslikke vaateid läänemeresoome regilaulude varieeruvusele

„Harja otsimine" ja „Mõõk merest"

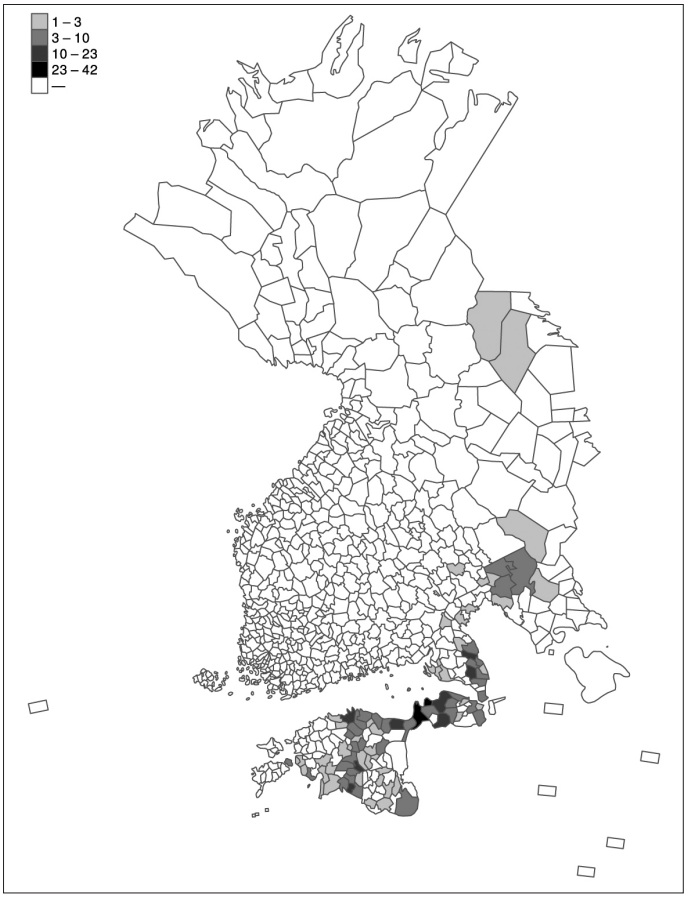
MARI SARV, KATI KALLIO, MACIEJ JANICKI

Regilaul, läänemeresoome kultuuride omapärasemaid nähtusi, on siinsetel aladel üks rikkalikumalt kogutud ja uuritud rahvaluuleżanre. Ajaloolis-geograafilise uurimismeetodi kasvav populaarsus rahvaluule suurkogumise aegadel on meile andnud erakordselt rikkalikud tekstikogud, kus – erinevalt varasemast kogumispõhimõttest jäädvustada teksti süżeest üks, võimalikult täiuslik versioon – pöörati tähelepanu kõigi variantide võimalikult täpsele dokumenteerimisele.

Suur teisendite hulk oli vajalik ajaloolis-geograafilise meetodi rakendamiseks folkloori ja sealhulgas regilaulude uurimisel, et välja selgitada laulude kujunemisteed ja võimalik algkodu võrdlev-ajaloolise keeleteaduse eeskujul. Folkloristika uurimisfookus pöördus XX sajandi jooksul tekstide võrdleva analüüsi juurest folkloorse kommunikatsiooni muude aspektide poole, nagu esitus, esitaja, loovus, laulmise kontekst, kultuurilised tähendused ja laulmise funktsioonid individuaalselt ja kogukondades. Teatud määral juba XX sajandi lõpukümnenditel, aga eriti käesoleval sajandil on arvutianalüüsi üha mitmekesisemad võimalused andnud põhjuse tagasi pöörduda suurte tekstihulkade analüüsi juurde. Regilauluainese arvutuslikud vaatlused võimaldavad meil paremini mõista folkloorse loomeviisi ja varieeruvuse olemust, saada teadmisi traditsiooni piirkondlikust jaotusest ja piirkondade eripärast üldisemas plaanis, aga mitmekesised otsingu- ja päringuvõimalused suures ainekogumis toetavad oluliselt ka detailisemate küsimuste analüüsi.

Soome ja Eesti teadlaste käimasoleva koostööprojekti „Vormellik intertekstuaalsus, teemavõrgustikud ja poeetiline varieeruvus läänemeresoome suulise luule piirkonnatraditsioonides" (2020–2024, lühend FILTER, rahastaja Soome Akadeemia) käigus on õnnestunud mõlemal maal paralleelselt ja samu andmestruktuure rakendades koostatud regilaulude andmebaasid liita ühisesse läänemeresoome regilaulude koondandmebaasi eesmärgiga saada parem ülevaade kogu regilaulukultuurist. Ehk küll regilaulu on uuritud põhjalikult eri aspektidest, on uurimistöö toimunud eri teadusruumides ning siiani puudub selge ülevaade, mis õigupoolest on sarnast ja mis erinevat eri rahvaste ja kultuurialade regilauludes, kui palju on sarnaseid süżeid, motiive, vormeleid ning milline on üldse regilaulutraditsiooni kujunemislugu. Läänemeresoome regilaulude koondandmebaasi ja tekstide arvutuslikku võrdlusmetoodikat arendades oleme töö kestel saanud parema ülevaate lauluvara żanrilise koostise piirkondlikest eripäradest, sõnavarakasutusest, tüpoloogilisest varieeruvusest, käimas on uurimistöö nii regilaulu ja kirjanduse vastasmõjude kui ka laulude, motiivide ja värsside stereotüüpsuse teemadel. Arvutuslik uurimistöö pakub rohkelt võimalusi ja väljakutseid.

Ühest küljest annab tekstide andmebaasi koondamine võimaluse teha statistilisi sõnapäringuid sadadest tuhandetest lauludest, kuid teisest küljest võib andmestikku ja selle eripärasid tundmata sooritatud otsing jätta osa huvipakkuvast ainesest kõrvale aga ka vastupidi: anda soovimatuid või kallutatud vasteid. Ainestiku kogumislugu ja -tihedust, žanrijaotust, keelelist varieeruvust arvestamata on oht tulemusi valesti tõlgendada. Andmestik on mitmeti ebaühtlane, mille põhjuseks on muuhulgas alamkorpuste koostamise erinevad põhimõtted, kogumislugu ja -põhimõtted (ajaline, geograafiline ja žanriline ebaühtlus) ning lõpuks mitmekesine sisuline, poeetiline ja keeleline (kirjaviis, murdelisus) varieeruvus.



Joonis 1. Laulutüüpidesse „Harja otsimine", „Suka mereen", „Mõõk merest", „Miekka merestä" liigitatud laulude ja nendega sarnaste laulude levikukaart.

Soome-Eesti uurimisprojekti raames on arvutiteadlane Maciej Janicki koostöös teiste projekti liikmetega ehitanud veebikeskkonna Runoregi, mis võimaldab leida ja vaadelda läänemeresoome regilaulude koondandmebaasis olevaid tekste, aga keskkonnas on rakendatud ka komplekt regilaulutekstide võrdlemiseks välja töötatud sarnasusarvutuste meetodeid, mis võimaldab võrrelda värsse, motiive või laule teiste

sarnaste üksustega ning vaadelda, millisesse konteksti huvipakkuv üksus asetub laulutekstide tervikkogumis. Veebikeskkonnaga liitub visualisatsioonide rakendus, kus Runoree abil leitud ainest on võimalik kujutada kaardil ja eri tüüpi graafikutel. Vaatamata enneolematult käepärastele võrdlusvõimalustele on Runoree abil siiski raske saada head ülevaadet suuremast tekstikogumist ja selle varieeruvusest. Ka on sarnasuste tuvastamisel (seni veel) takistuseks suuremad keelelised lahknevused.

Käesoleva artikli eesmärk on tutvustada läänemeresoome regilaulude koondandmebaasi koostist ning Runoree keskkonnas ja sellega liituvas visualisatsioonide rakenduses peituvaid võimalusi regilaulude vaatluseks ja võrdluseks, kombineerides lähilugemist arvutusliku analüüsi meetoditega. Lisaks oleme otsinud metoodilisi lahendusi laulutüüpidest, motiividest ja nende varieeruvusest parema ülevaate saamiseks. Selleks rakendame siinses artiklis värsside sarnasus- ja külgnevusarvutustel põhinevatel andmetel võrgustikuanalüüsi.

Vaatluse keskmes on Aado Lintropi (2024) selleteemalisest artiklist inspireerituna laulutüüp „Harja otsimine" (soome „Suka mereen"), mis on tuntud Eestis, Ingerimaal ja Karjalas ning liitub võrdlemisi püsivalt laulutüübiga „Mõõk merest" (soome „Miekka merestä", vt ka Salmela 1964; Lintrop 1999, 2000; Hiiemäe 2006). Laulutüüpide peamine sisu lühidalt kokku võetuna on järgmine: päike, neiu või poiss kammib oma pead, kamm sulpsatab merre, peategelane läheb seda otsima ning leiab merest mõõga. Vaatluse all on folkloristide poolt mõlemasse tüüpi kuuluvaks määratud tekstid (380 teksti) ning Runoree sarnasusarvutuste põhjal nendega sarnased tekstid (18 teksti, tekstide sarnasusindeks 0,5), kokku 398 teksti (neist 163 Eesti ja 235 Soome andmebaasist, joonis 1).

Läänemeresoome regilaulud: andmestik, andmebaas ja veebirakendused

Rahvaluuleuurimise algusaegadest saadik on andmete kogumine, haldamine, korraldamine ja süstematiseerimine olnud tööprotsessi oluline osa. Eestis ja Soomes lähtus vanemate rahvaluulekogude esmane korraldamine ajaloolis-geograafilise meetodi põhimõtetest. Ka Eesti ja Soome regilaulude tüpologiseerimine on alguse saanud samas teoreetilises kontekstis. Olgugi et tüpologiseerimispõhimõtteid ja tüübiliigitust on aegade jooksul kritiseeritud ja muudetud, pärinevad tekstide liigituspõhimõtted, paljud alajaotused ja tüübinimetused XX sajandi algusaegadest.

Läänemeresoome regilaulude koondandmebaasis on praegu üle 280 000 teksti. Andmebaas sisaldab peale regilaulude teatud hulgal ka muid tekste: riimilisi ja siirdevormilisi laule, regivärsiainelist omaloomingut, mitteregivärsilisi lastelaule, loodushäälendeid, teavet laulikute ja laulmise kohta ja muud. Lisaks sellele on Eesti, Soome ja Venemaa arhiivides käsikirju, heli- ja videosalvestusi ning trükiseid, milles leiduvad regilaulud ei ole andmebaasi (veel) jõudnud. Kavas on andmebaasi koostamine Petroskoi karjala uurimiskeskuse karjala kogudest (Kundozerova 2022).

Andmebaasi on koondatud kolm suuremat regilaulukogu, igaüks oma eripäraga. Lisaks arhiivimaterjalidele on andmebaasis eraldi täienev kogu trükis avaldatud

Tabel 1. Ülevaade läänemeresoome regilaulude koondandmebaasi alamkogudest: ERAB, SKVR ja JR (Kallio jt 2023: 62–63).[1]

Alamkogu	ERAB (2023)	SKVR	JR
Tekste	108 969	89 247	85 228
Värsse	2 162 948	1 417 090	893 288
Sõnu	7 346 075	4 259 398	2 599 158
Kogumisaeg	1644, 1804–1943	1564–1939	1653–1971
50% tekstidest kogutud aastatel	1889–1923	1884–1915	1930–1947
Maakondi	13	24	29
Kihelkondi (või linnu)	118	532	622
Kogujaid	4192	1634	2413
Peamised keeled	põhja- ja lõunaeesti	isuri, karjala, lüüdi, soome, vadja	isuri, karjala, soome
Muud ette tulevad keeled	ingerisoome, isuri, saksa, soome, vene	rootsi, ladina, vene, kreeka	eesti, vepsa, vadja, saami, roma, rootsi
Täielikkus	täieneb	valmis	valmis, võib täieneda
Tüpoloogiline korrastatus	töös (osa ainesest on tüpologiseerimata)	töös (vajab täpsustamist ja kontrollimist)	puudub (on karto-teegikaartidel)
Tüübinimetusi	2514 (masinakirjakoopiatelt saadud ühtlustamata tüübinimetusi 14 818)	7573	–
Tüübinimetusi, mille alla kuulub vaid üks tekst	828 (masinakirjakoopiatelt 10 181)	2827	–
Korrastatud andmebaasis	tekst (kirjaviisilt ühtlustatud versioon), koht, kogumisaeg, koguja	koht, kogumisaeg, koguja	koht, kogumisaeg, koguja

regilauludest ja regivärsivormis kirjandusloomingust, et saaks analüüsida regivärsivormis suulise ja kirjaliku kultuuri sarnasusi ja vastasmõjusid.

- ERAB on Eesti regilaulude andmebaasi alates 2003. aastast lisatud tekstide kogu, mis on kättesaadavad andmebaasi veebilehe kaudu https://www.folklore.ee/regilaul/andmebaas (praegu kokku 108 969 teksti). Andmebaasi täiendatakse pidevalt, lisaks regilauludele sisaldab see laulmist ja laulikuid puudutavaid tekste ning muudesse regilauluga liituvatesse või sellega seotud žanridesse kuuluvaid tekste, nagu tantsulaulud, lastelaulud, loitsud, kombestikukirjeldused, isegi muinasjutud. Andmebaasi esialgsed tüübinimed pärinevad Eesti Rahvaluule Arhiivi kogutud regilaulude aastakümnete jooksul tehtud masinakirjakoopiatelt, tüpoloogia korrastamine on pooleli. Andmebaasi on lisatud nii tekstide algupärane kui ka kirjaviisilt ühtlustatud versioon (viimast kasutatakse koondandmebaasi sarnasusarvutustes).

[1] Tekstide arv ERAB-i kogus on korrigeeritud.

- SKVR on Soome väljaandel „Suomen kansan vanhat runot" põhinev 89 247 tekstist koosnev kogu soome, karjala, isuri ja vadja regilaule ning osalt ka muid regilauludega seonduvaid tekste, mis on eraldi kättesaadav veebilehe www.skvr.fi kaudu. Andmekogu tüpoloogiline liigitus ühtlustati ja korrastati XX sajandi lõpul. Korpus sisaldab mõningaid XIX sajandil trükis ilmunud regilaule. Tekstide kirjaviis on ebaühtlane ja seda ei ole süstemaatiliselt korrigeeritud. Läänemeresoome regilaulu koondandmebaasi sarnasusarvutustes on kasutatud automaatselt ühtlustatud ja puhastatud versiooni, kust on näiteks eemaldatud kõik diakriitilised märgid.
- JR on Soome avaldamata laulude kogu (sm *julkaisemattomat runot* 'avaldamata laulud'), mis koosneb 85 228 tekstist. Neist suurem osa on regilaulud, kuid leidub ka riimilisi laule, lastelaule ja kirjandusliku algupäraga luulet. Kogu ei ole tüpologiseeritud ega kontrollitud, tekstid varieeruvad ka kirjaviisilt. Sarnaselt SKVR-i koguga kasutatakse andmebaasi sarnasusarvutustes automaatselt ühtlustatud tekstiversiooni.

FILTER-i projekti raames on need kolm kogu koondatud ühte SQL-andmebaasi, mida on võimalik uurida erinevates veebikeskkondades: teksti ja metaandmete päringukeskkond Octavo, sarnasuste tuvastamise keskkond Runoregi ja FILTER-i visualisatsioonide rakendus (FILTER visualizations), kus on võimalik kuvada andmeid kaartidel ja graafikutel.

Andmete ebaühtluse ja keelelise varieeruvuse probleem

Regilaule ja nendega seonduvaid tekste on erinevatel eesmärkidel ja eri põhimõtteid järgides kirja pandud juba alates XVII sajandist. Kogude tekkelugu on seotud erinevate inimeste ja institutsioonide huvide, valikute, eelistuste ja harjumustega ning andmekogu on mitmes mõttes ebaühtlane. Uurimistööd tehes tuleb neid aspekte arvesse võtta, et mitte teha ekslikke järeldusi.

Korpuse vanim tekst on kirja pandud aastal 1564, kuid laiaulatuslikum rahvaluule kogumine algas XIX sajandi alguses eeskätt Soome teadlaste algatusel, keda huvitasid peamiselt mütoloogia ja eepilised laulud, mis olid kõige rohkem levinud toonasesse Arhangelski kubermangu kuulunud Valge mere (ehk Viena) Karjalas, aga võrdlemisi laialdaselt ka teistes karjalakeelsetes piirkondades, Ingerimaal ja Ida-Soomes (karjala ja isuri keele rääkijaid käsitlesid toonased Soome teadlased soomlastena). Sõltuvalt kogujast pandi kirja ka muudesse žanridesse kuuluvaid laule. Eestis tekkis teaduslik huvi eestlaste ja nende rahvalaulude vastu baltisakslaste hulgas (nagu teame, oli eestlastel väga vähe võimalusi omandada kõrgemat haridust), kes laule teatud määral kogusid ja avaldasid. Aja jooksul huvi regilaulude vastu laienes nii žanriliselt kui ka geograafiliselt. Varased kogujad märkisid harva laulude juurde täpseid andmeid kogumise koha, aja ja esitaja kohta. Samuti ei pööranud nad tähelepanu teksti täpse kuju kirjapanemisele: detailide keeleline ühtlustamine arusaadavuse huvides oli tavaline. (Kallio jt 2023)

XIX sajandiks olid läänemeresoome keeleala eri piirkondade laulukultuurid kujundanud eriilmelisteks mitmesugused tegurid, nagu inimesi ümbritsev elu- ja looduskeskkond, kuulumine eri riikide, kirikute, õigusruumide ja kirjakultuuride mõjusfääri, mitmesugused ajaloosündmused ja moderniseerumisprotsessi erinev areng. Teadlased ja teised vanema kultuuri huvilised püüdsid kirja panna tekste, mis nende meelest olid kõige väärtuslikumad, terviklikumad ja arhailisemad. Eriti Soomes otsiti väärtuslikku ainest eelkõige oma kodukohast kaugemalt, oma aja kultuurilistest perifeeriatest (Kalkun 2015; Piela 2023; Tarkka 2005). Alles siis, kui folkloristika XIX sajandi lõpul hakkas kujunema omaette distsipliiniks, mille metoodiline kese oli ajaloolis-geograafiline meetod, seati eesmärgiks, et kirjapanekuid oleks ühtlase(ma)lt kogu traditsioonialalt. Eestis seostus Jakob Hurda algatus rahvuse konsolideerimisega, rahvaluulet sooviti kirja panna kõikjalt, kus elas eestlasi (Kikas 2014). Nii kogumisloo kui ka regilaulukultuuri erineva arengu tõttu eri aladel on andmekogus ebaühtlust laululiikide ja -tüüpide esindatuses nii ajaliselt kui ka ruumiliselt.

Eesti ja Soome regilaulukogudes on sarnaselt märgitud laulu päritolukoht, kirjapaneku aeg, koguja, arhiiviviide, laululiik ja -tüüp. Need kategooriad ei ole alati lihtsalt ja üheselt mõistetavad. Laulu päritolukoht võib tähistada kas esitaja sünnikohta või elukohta kogumise ajal, kirjapanemiskohta või laulu õppimiskohta ning olla esitatud kas talu, küla, kihelkonna või (eriti varasemas aineses) maakonna täpsusega. Mõnest üleskirjutusest on arhiivi jõudnud mitu koopiat või veidi erinevat versiooni, leidub ka tekste, millel päritoluteave puudub.

ERAB-i ja SKVR-i kogudel on oma tüübiindeksid, mis on loodud arhiiviainese haldamise ja süstematiseerimise käigus ning üldjoontes järgivad sama ideed, kuid tüpoloogiad on siiski loodud eraldi Eesti ja Soome arhiivikogude põhjal. Vaid mõned laulude eesti ja soome tüübinimetused on ühitatavad pelgalt tõlkimise teel. ERAB-i esialgsed tüübinimetused pärinevad ainese skaneerimise aluseks olnud umbes 60 aasta jooksul tehtud masinakirjakoopiatelt. Selle aja jooksul on tüübinimetused muutunud ja varieerunud. Tüpoloogia korrastamise pikaldane töö on käimas, praeguseks on ühtlustatud 54 453 teksti varieeruvad tüübinimed, kuid suur osa tüübimääratlustest vajab veel täpsustamist või täiendamist. Lisaks sellele on 16 976 teksti esialgse tüübinimega, ülejäänud tekstid on kas tüübimääratluseta või ei ole regivärsivormis. Soome ainesest on SKVR-i kogu tüpologiseeritud, JR-i kogu mitte – tüübikartoteegis on tüübinimed küll olemas, kuid kogu digiteeriti geograafiliselt korraldatud kartoteegi põhjal, mis oli täielikum. SKVR-i kogu trükiköidetes ebaühtlaselt kasutatud tüübinimetused korrastati XX sajandi lõpul ja mõne laululiigi osas revideeriti tüpoloogia täielikult. Tüpologiseerimise põhimõtted on laululiigiti erinevad: jutustavate laulude tüübinimetus on enamasti seotud süžeega (arvestatud on võtmemotiive ja -värsse); loitsud on indekseeritud funktsiooni järgi (laulutüübi alla määratud tekstide kogum võib olla väga kirev); lüüriliste laulude tüpologiseerimisel on aluseks väiksemad ühikud (motiivid); olukorra, kalendrikombestiku või muude tavanditega seotud laulud (laululiigid „Erilistes olukordades", „Pulmalaulud") on liigitatud nende kasutuskonteksti või rituaalse funktsiooni alusel (kuigi sageli on arvestatud ka tekstitüüpi ja võtmemotiive). See tähendab, et tüpoloogia eri alajaotused võivad põhineda

teksti süžeel, kinnismotiividel, kontekstil või funktsioonil ning need laulukogumid käituvad erinevalt, kui hindame tekstide sarnasust arvutuslike meetoditega.

Andmestiku keeleline varieeruvus on määratu. Läänemeresoome keelte sõnade ja morfoloogia murdeline varieeruvus on juba iseenesest rikkalik, kuid poeetiline keel sisaldab ka hulga arhailisi ja eripäraseid keelendeid. Kogujad on laule üles märkinud erineva keelelise täpsusega ja eri viisil: mõni kasutas üleskirjutamisel lühendeid, mõni ühtlustas tekste või lähendas neid kirja- või üldkeelele, teised seevastu on leiutanud eripäraseid kirjaviise või -märke, et kuuldu võimalikult täpselt üles kirjutada, või siis kasutanud laulude kirjutamisel foneetilist transkriptsiooni. Regilaule on kirja pandud eesti, vadja, isuri, karjala, lüüdi ja soome keelealade erinevates murretes, JR-i kogu sisaldab ka vepsa laule. Mitmekesise varieeruvuse tõttu võib mõnel olulisemal sõnal (nagu *Väinämöinen*) olla andmestikus sadu erinevaid esinemiskujusid. Lisaks sellele on paljud kujult sarnased sõnad erinevate tähendustega ning võivad pärineda hoopis eri tüvedest: sõnavormide homonüümia on tavaline. Regilaulutekstide keeleline mitmekesisus pakub regilaulude, nende sisu ja folkloorse varieeruvuse arvutuslikul analüüsil põnevaid väljakutseid. Esialgu veel jääb olemasolevatest standardsemate keelekujude analüüsiks loodud keeleanalüüsi vahenditest, eri keelte ja murrete sõnaraamatutest ja -korpustest väheks, et regilaulukeele sõnavaralist, häälikulist, morfoloogilist mitmekesisust täies mahus arvutuslikult analüüsida. Oma uurimistöös katsetame pidevalt eri meetodeid, et tekstide keeleline varieeruvus ei takistaks folkloristlikku analüüsi.

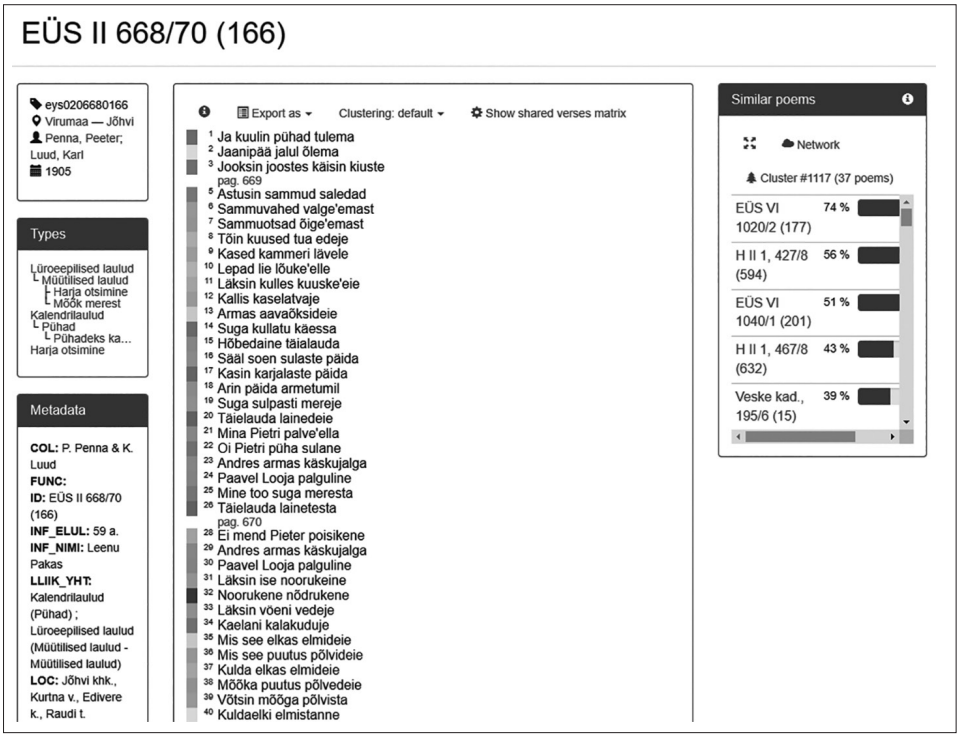
Värsside, värsijärjendite ja laulude sarnasus: Runoregi

Selleks et ületada laulutekstide keelelisest varieeruvusest tulenevaid probleeme ja oleks võimalik kokku viia sama sisuga värsid ning laulud on Janicki (2023; Janicki jt 2023) välja töötanud meetodid sarnaste värsside ja laulude tuvastamiseks. Neid meetodeid kasutades on arvutatud andmebaasi värsside ja laulude sarnasusnäitajad, mis on lisatud läänemeresoome regilaulude koondandmebaasi ning rakendatud neid ka Runoree veebikeskkonnas[2] eesmärgiga hõlbustada sarnaste üksuste võrdlust ja lähilugemist. Keskkonnas on kerge liikuda ühelt tekstilt, lõigult või värsilt teisele, (arvutuslikult) kõige sarnasematele üksustele või uurida mitut sarnast teksti kõrvuti sarnaste värsside järgi joondatult. Runoregi suudab tuvastada sarnaseid värsse ja laule, mis ei ole leitavad sõnaotsinguga või tüübimääratluste alusel.

Ühe laulu vaates (joonis 2) näitab Runoregi laulu metaandmeid, teksti ning kõige sarnasemate tekstide loendit, samuti pakub võimaluse uurida tuvastatud sarnaste värsside klastreid (klikkides ruudukesele värsi ees, ruudu tumedus näitab sarnaste värsside arvu), sarnaseid lõike (need ilmuvad, kui valida hiirega huvi pakkuv lõik laulust) ning saada kiire ülevaade, millised laulu värsid esinevad teistes tekstile kõige

[2] Runoree keskkond asub veebiaadressil: https://runoregi.rahtiapp.fi. Artikli kirjutamise ajal on uurimisprojekt alles käimas ning sellest tulenevalt kogu keskkond aktiivsel katsetuslikul arendamisel, niisiis on tõenäoline, et siin kirjeldatud tunnused ja võimalused tulevikus teatud määral muutuvad.

sarnasemates lauludes (vajutades kirjale: „Show shared verses matrix" 'näita ühiste värsside maatriksit').



Joonis 2. Laulu vaade ühe „Harja otsimise" tüüpi kuuluva teksti näitel keskkonnas Runoregi.

Värsside sarnasuse tuvastamiseks on leitud igas värsis esinevad kahe tähe järjendi (ehk bigrammi) paarid, arvutatud värsipaaride bigrammikoostise koosinussarnasus ning selle alusel värsid klasterdatud niinimetatud telefonimängu meetodil (ingl *Chinese whispers clustering method*, vt meetodi ja selle testimise kohta Janicki jt 2023). Meetodiga on võimalik tuvastada värsi sarnased variandid, näiteks:

(1a) Suga sulpsatas meresse (EKM ERA, E 7901/2 (31))[3]
(1b) Sulpsatas suga meresse, (EÜS VIII 1923/5 (169))
(1c) Suga läks sulpsates meresse. (H I 2, 86 (7))
(1d) Suga suissatas mereje (H II 37, 343/4 (11))

(2a) Ari läks sulpsates meresse (AES, MT 49, 21/2 (10))
(2b) Hari sumatas meresse, (ERA II 60, 47 (5))
(2c) Hari aga sulpsatas meresse, (H, Kolga-Jaani 3, 233 (215))

[3] Edaspidi viidatakse tekstis Eesti Kirjandusmuuseumi (EKM) Eesti Rahvaluule Arhiivi (ERA) rahvaluulekogudele kogu, köite ja lehekülje ning pala tähistega.

(3a) Hari kukus tal meresse (H, Ostrov 39/40 (38))
(3b) Suga kukkus käest meresse, (H II 41, 595/6 (2))
(3c) Hari kukkus käesta meresse (E A 946/8 (11)).

Oleme katsetanud sarnasusarvutuste eri versioone ja need on kasutatavad Runoree keskkonnas.[4] Mõni neist (*default* 'optimaalne sarnasusläve', 'tähepaaride lihtsagedus'; *normalized* 'normaliseeritud'; *loose* 'madal sarnasusläve') annab tulemuseks suuremad värsiklastrid, kus värsside varieeruvus on suurem ning mis sisaldavad rohkem valepositiivseid juhtumeid, mis tegelikult ei esinda sama värsitüüpi;[5] teised (*binary* 'binaarvektor'; *sqrt* 'ruutjuurväärtus'; *tight* 'kõrge sarnasusläve') annavad tulemuseks kitsama ulatusega klastrid, kuhu koondunud värsid esindavad enamasti sama värsitüüpi, kuid klastritest on välja jäänud hulk siiski samasse värsitüüpi kuuluvaid, kuid koostiselt vähem sarnaseid värsikujusid. Keskkonnas on võimalik uurida „naaberklastreid", näiteks ülaltoodud kolmest rühmast kaks esimest on naabrid, kuid kolmas on juba oma tähepaarikoostiselt eelmistest kaugemal.[6]

Tähepaarivektorite sarnasusel põhinev võrdlusmeetod suudab ainesest sageli tuvastada sama sisuga värsivariante, mille olemasolu peale uurija ise ei pruugi tullagi, samas aga paigutuvad värsi kaugemad variandid tihti teistesse klastritesse. Eriti kirjaviisilt ühtlustamata Soome alamkogude puhul on sellise klasterdamisega võimalik suuresti ületada kirjaviisi varieerumisest tulenevad erinevused, mis tihti piiravad oluliselt lihttekstiotsingu (sõnapäringute) tulemusi. Kaugemates murretes või lausa eri keeltes jääb värsside sarnasus enamasti tuvastamata, isegi kui värsside põhisisu on edasi antud samade tüvisõnadega, tähepaaride jadas mängivad statistiliselt olulist rolli peale sõnatüvede häälikulise varieeruvuse ka erinevused morfoloogias ja grammatilistes sõnades.

Tuginedes värsitasandi sarnasusarvutustele, on võimalik edasi liikuda suuremate üksuste, sarnaste lõikude ja terviklaulude vaatlusele (värsijadade ja pikemate tekstide sarnasuste tuvastamise meetodi kohta vt Janicki 2023). Kui üksikute värsside klasterdamisel esineb klastrites siiski ka valepositiivseid, häälikuliselt koostiselt sarnaseid,

[4] Sarnasusarvutused erinevad omavahel esiteks arvutusparameetrite väärtuste poolest: sarnaseks loetud värsside sarnasusläve saab seada kas (seniste kogemuste põhjal) optimaalseks (ingl *default*), kõrgemaks (*tight*) või madalamaks (*loose*), sõltuvalt sellest, kas konkreetsel kasutusjuhul peetakse olulisemaks korrektsust või täielikkust. Võrreldavad vektorid võivad olla kas tähepaaride lihtsagedused (*default*), ruutjuurväärtused nendest (*sqrt*) või lihtsalt binaarvektorid (*binary*), mis sedastavad, kas vastav tähepaar värsis esineb või mitte (sõltumata esinemiste arvust). Ruutjuur- ja binaarväärtusi oleme kasutanud, kuna tähepaaride lihtsageduste põhjal arvutatud koosinussarnasused tõid värsisarnasuste tuvastamisel ebaproportsionaalselt esile häälikukordused, mille puhul sarnaseid heakõlapaare või sõnakordusi sisaldavad värsid (nt *laula laula* ~ *lalala*) loeti sarnaseks vaatamata nende selgesti erinevale sisule. Viimaks oleme rakendanud väga robustset eesti ja soome kirjaviisi erinevuse ning vokaalharmoonia mõju silumist („normaliseerimist"), et proovida arvutuslikult lähendada põhjapoolse (soome, karjala, isuri, vadja) ja lõunapoolse (põhja- ja lõunaeesti) korpuse värsse (nt b → p, ä → a jne; sarnasuse tuvastamise meetodid *norm* ja *norm-sqrt*).

[5] Värsitüübina käsitame oma lähenemises värsside kogumit, mille sisusõnad kattuvad, kuid grammatilised sõnad, sõnade vormid ning sõnajärg võivad varieeruda.

[6] Kaks värsiklastrit loetakse „naabriteks", kui klastri mis tahes värsi koosinussarnasus teise klastri mis tahes värsiga on suurem kui 0,75 (see on väikseim oluliseks peetav väärtus).

kuid sisult erinevaid värsse, siis pikemate tekstide sarnasuse tuvastamine toimib keskkonnas üldiselt usaldusväärselt. Kui tekstid on murdeliselt või kirjaviisilt teineteisest kauged (ja isegi nende sisult sarnased värsid on häälikukoostiselt niivõrd erinevad, et ei klasterdu kokku), siis ei suuda ka arvuti nende sarnasust tuvastada, isegi kui lugeja teatavat sisulist sarnasust märkab. Nii näiteks ei loe Runoregi omavahel sarnaseks[7] laulutüübi „Sai paha peigmehe" („Pahan sulhon saanut") eri piirkondadest kogutud variante: 1) Tveri karjala küladest, kuhu karjalased rändasid XVII sajandil Laadoga Karjalast, 2) 19 muust karjalakeelsest piirkonnast (sealhulgas Laadoga Karjalast ja soome karjalapärastelt aladelt) ning 3) Karjala kannaselt ja Ingerimaalt kirja pandud tekste.[8] Küll aga on nähtavad juhuslikumad sarnasused värsi ja lõigu tasandil eri piirkondade karjala lauludes.

Enamasti on sama lauliku, sama küla, sama piirkonnatraditsiooni ala, sama keeleala lauluvariandid omavahel sarnasemad nii värsitüüpide kasutuse, sisu kui ka poeetilise ülesehituse osas ning nende sarnasust on kergem tuvastada, samuti on hästi tuvastatavad samast kirjalikust allikast pärinevad tekstid. Laias laastus tähendab sarnasusnäitaja 80–100% sisuliselt sama teksti (näiteks sama laulu kordusüleskirjutused samalt laulikult, käsikiri ja selle kergelt toimetatud koopia või trükiallikas ja sealt õpitud laul, väga lühike püsiva kujuga loits või lastelaul); sarnasusnäitaja 50–80% osutab väga sarnastele variantidele samast piirkonnast (tihti samalt laulikult, samast perest või külast), toimetatud koopiale, kirjandusallikale või väga püsiva kujuga lühikesele rahvalaulule. Alla 50% sarnane variant viitab tihti tavapärasele suulisele varieeruvusele, kuid võib siiski esineda ka sama lauliku lauldud lauluvariantide, osaliselt kopeeritud ja toimetatud käsikirjade või kirjanduslike mõjude puhul. Laulutekstide sarnasusnäitajaid kasutame praegu ka ERAB-i tüübimääratluseta laulude tüpologiseerimisel ning edaspidi on seda kavas teha Soome JR-i korpusega. (Janicki 2023)

Runoree laulu vaate paremas ülanurgas sarnaste laulude kastis on võimalik valida üks või rohkem lauluteksti (rohkem laulutekste on võimalik valida, kui vajutada nelja noolega ikoonile) ja võrrelda neid vaates, kus sarnased värsid on kohastikku paigutatud. Seda, mil määral kohastikku paigutatavad värsid sarnased peavad olema, saab seadistada (nupust *Threshold* 'lävi'). Värsile klõpsates värvuvad teised sarnased värsid sõltumata nende asukohast laulus (sarnasusmäär kuvatakse hiirega sarnase värsi kohale libisedes).

Runoree avalehel on toodud Soome ja Eesti tüpoloogilise klassifikatsiooni peakategooriad, millele klõpsates avaneb vastava alaliigi laulude tüübinimetuste loend. Teine võimalus konkreetsesse tüüpi kuuluvate laulude leidmiseks on kasutada esilehe otsinguakent, mille eri sakkidel on otsitulemused tüübinimedest ja -kirjeldustest, laulude metaandmetest, värssidest ja tekstidega liituvatest proosaosadest.

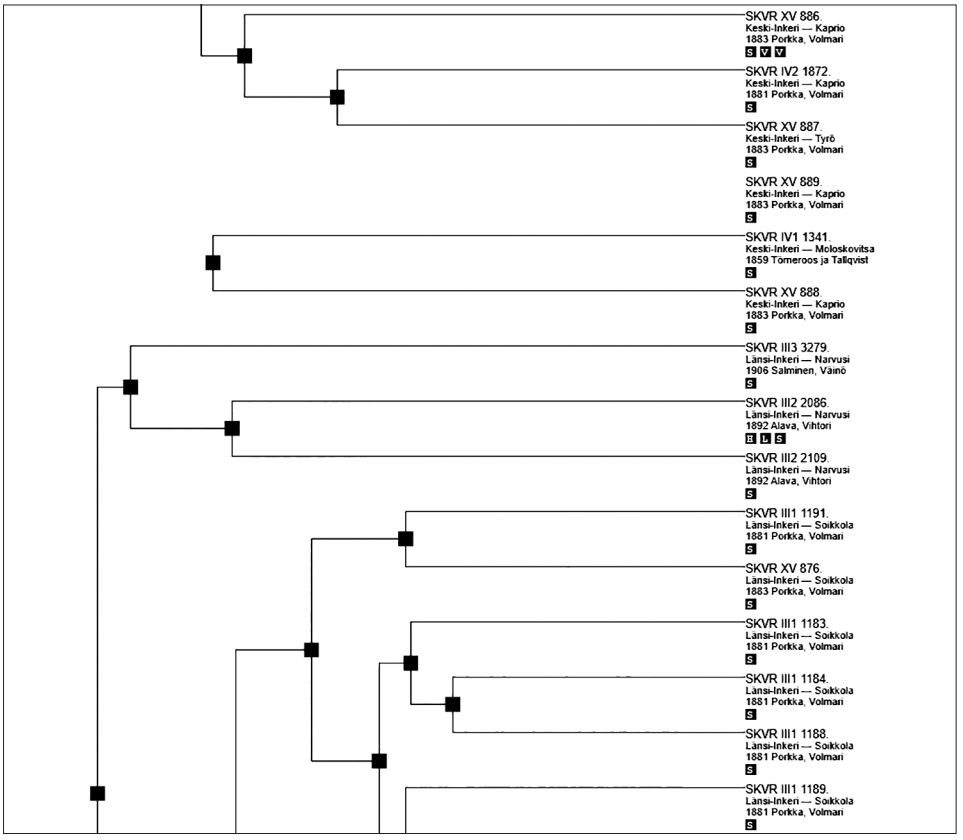
Kui klõpsata laulutüübi nimetusele, siis avaneb lehekülg, kus on tüübikirjeldus (juhul kui see on andmebaasis olemas) ning loend kõigi selle tüübi alla liigitatud

[6] Selleks peab laulude sarnasusnäitaja olema suurem kui 10%.

[8] Eranditeks on Tveri karjala teisend SKVR II: 1133, mis klasterdub kokku teiste Karjala alade tekstidega, ja Kainuu karjalakeelsest külast kirja pandud teisend SKVR I 2: 1161, mis klasterdub kokku Tveri lauludega.

laulude andmetega. Loend on järjestatud alfabeetiliselt viite alusel, mis SKVR-i alamkogu puhul tähendab ka algset avaldamise järjekorda. Laulutüübi kohta on võimalik kuvada ka muid ülevaateid: laulude esinemust võib kuvada kaardil, vaadata ülevaatlikku visualisatsiooni uuritava tüübiga kontamineeruvatest laulutüüpidest ning vaadata puu (dendrogrammi, joonis 3) vaates, kuidas tüüpi kuuluvad laulud sarnasuse alusel omavahel rühmituvad. Puu vaadet võib täiendada tüüpi kuuluvate sarnaste lauludega. Nii on võimalik vaatlusse kaasata laulud, millel tüübimääratlus puudub või mis on määratud muudesse tüüpidesse. Puu vaates rühmituvad kokku enamasti lähedastest kohtadest pärit laulud, tuues esile sarnasuse piirkonnatraditsioonide sees. Puu vaate sõlmedele klõpsates on võimalik vaadata kõrvuti lähedasi lauluvariante, mis aitab kiiremini läbi töötada ja lihtsamalt võrrelda sarnaseid tekste.

Keerukamaid teksti- ja metaandmepäringuid saab teha kas SQL-andmebaasi kasutades või Octavo päringukeskkonnas, mille internetiviide leidub Runoree päringuakna infotekstis. Octavo keskkonnas on eraldi vaade sõnavormi variantide otsimiseks ja valimiseks ning otsitulemused on omakorda seostatud Runoree laulu vaatega. Ka Octavo päringutulemused saab kuvada kaardil või žanrijaotuse skeemil, mis võimaldab saada kiiremini ülevaate päringutulemuste kogumist.



Joonis 3. Osa „Suka mereen" laulutüüpi kuuluvate tekstide dendrogrammist.

Ülevaade laulutüübi varieeruvusest: võrgustikuanalüüs

Runoregi ja Octavo päringukeskkond võimaldavad küll hästi tuvastada suures tekstikogumis sarnaseid värsse, lõike ja laule ning neid omavahel lähilugemise teel mugavalt võrrelda, samuti koostada mitmesuguseid visuaalseid ülevaateid otsitulemustest erinevate metaandmete põhjal, kuid oleme oma töös põrkunud probleemiga, et raske on saada ülevaadet suuremast laulukogumist, selle varieeruvusest, sisulistest ja piirkondlikest eriarendustest. Siin vaadeldava(te) laulutüübi (laulutüüpide) ligi 400 teksti läbitöötamine on ajamahukas töö isegi Runoree paindlikult kasutusmugavas keskkonnas ning kokkuvõtete tegemine ainesest keerukas.

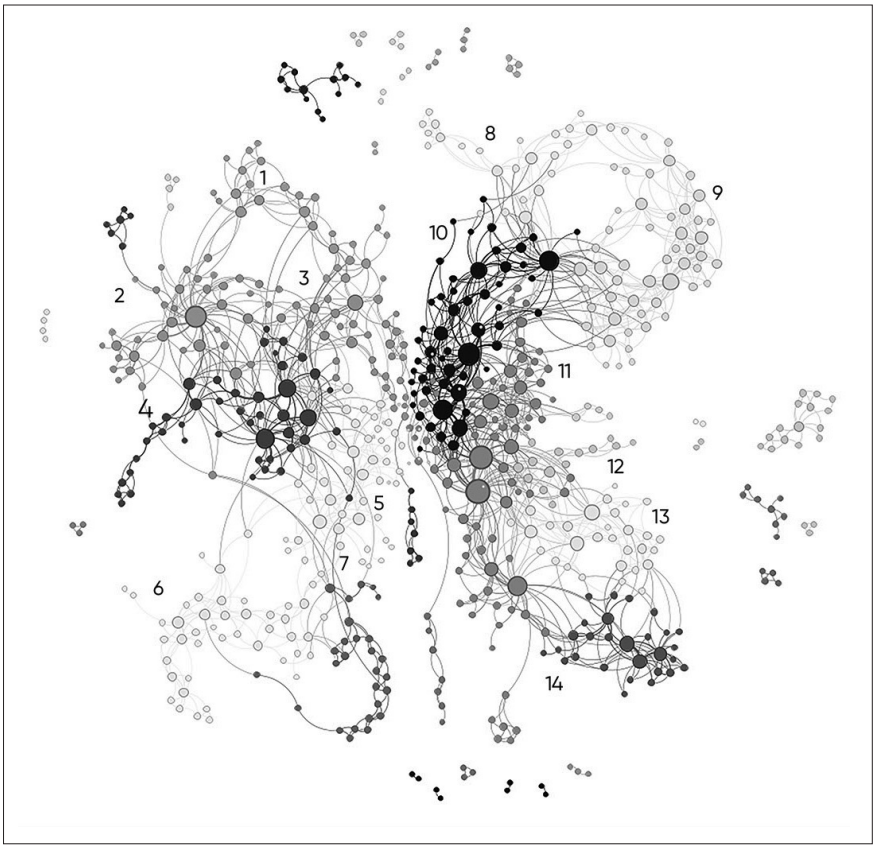
Üheks FILTER-i projekti uurimistöö eesmärgiks oleme seadnud tekstimassiivist korduvate üksuste tuvastamise ehk teisisõnu arvutusliku (või naturaalse) liigituse koostamise, arvestades regilaulule (ja suulisele luulele üldisemalt) iseloomulikku tekstiloomeviisi, kus stereotüüpsemad (ehk eri kontekstides korduvad) üksused vahelduvad tüübikindlamate või unikaalsemate üksustega. Oleme püüdnud tuvastada tekstides korduvaid osiseid, uurides, kui regulaarselt eelkirjeldatud sarnaste värsside klastrid üksteisega külgnevad, selgitamaks, millised värsitüübid tekstides tavaliselt lähestikku paiknevad.

Kuna värsiklastrite külgnevusandmete põhjal joonistatud võrgustikugraafil tundusid üksikute laulutüüpide graafid hästi ilmestavat laulude sisulisi hargnevusi (kui neid oli), siis otsustasime siinse artikli kontekstis rakendada värsiklastrite külgnevusanalüüsi piiratud tekstikogumil, et uurida, kuidas ühe laulusüžeega seotud värsid külgnevuse alusel rühmituvad, lootuses tuvastada laulude püsivamad koostisosad (motiivid) ja piirkondlikud eriarendused.

Joonisel 4 on esitatud kammi kaotamisega seotud laulude värsiklastrite külgnevusseoste graaf.[9] Graafi iga sõlm esindab ühte värsiklastrit ning graafile on valitud kõik värsiklastrite paarid, mis tekstis esinevad üle ühe korra kas kõrvuti või teineteisest ühe värsi kaugusel. Seose olulisust näitab sõlmedevahelise joone jämedus, mille aluseks on seosetihedushinnang.[10] Võrgustikus ilmnevad omavahel tihedamalt seotud värsiklastrite alagrupid, suuremad alagrupid on joonisel tähistatud numbritega 1–14 ning need viitavad kas piirkondlikele eriarendustele või väiksematele sisuüksustele (motiividele). Alagrupid 1–7 esindavad laulu eesti tekste ning grupid 8–14 Ingerimaa, soome ja karjala tekste. Suuremad sõlmed võrgustikus viitavad laulu kesksetele värssidele, mis esinevad rohkemates lauluvariantides ja külgnevad rohkemate värssidega.

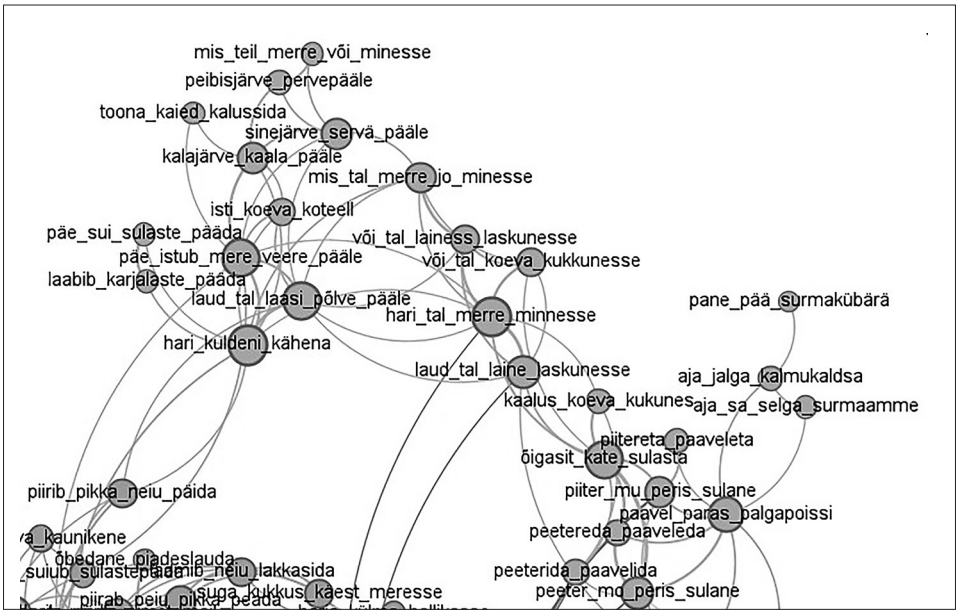
[9] Graaf on koostatud võrgustikuanalüüsi ja võrgustike visualiseerimise programmi Gephi (Bastian jt 2009) abil. Võrgustiku gruppide tuvastamiseks on kasutatud Gephisse sisse ehitatud võrgustiku modulaarsusarvutuse meetodit (Blondel jt 2008).

[10] Seosetihedushinnangu arvutamiseks on kasutatud leksikograafi vastastikuse informatsiooni (ingl *lexicographers mutual information*) näitajat (Kilgarriff jt 2004).



Joonis 4. Laulutüüpide „Harja otsimine", „Suka mereen", „Mõõk merest" ja „Miekka merestä" ning neile sarnaste laulude värsiklastrite külgnevusseoste võrgustik (graafil on kuvatud kõik rohkem kui üks kord esinevad seosed rohkem kui kolm korda ette tulevates värsitüüpides). Võrgustiku Gephi modulaarsusanalüüsiga saadud alagrupid on nummerdatud ja võimaluste piires ka eristatud tooniga, sõlme suurus näitab temaga külgnevate teiste värsiklastrite arvu.

Nagu eespool öeldud, suudab Runorees rakendatud värsside sarnasusarvutusmeetod küll kokku viia häälikuliselt veidi erinevad värsivariandid lähedase keele- või murdeala piirides, kuid ei suuda alati kokku viia sama sõnalise koostisega värsivariante eri keeltes ja eri murretes. Eriti järsk lõhe on Soome ja Eesti korpuse tekstide vahel, mis on vaid harva automaatselt kokku viidavad. See piir langeb üldjoontes (v.a vadja laulude osas) kokku läänemeresoome keelte jaotusega põhja- ja lõunarühmaks. Ka joonisel 4 toodud graafil põhineb vasakpoolne osa Eesti korpuse tekstidel ning parempoolne Soome korpuse tekstidel. Lisaks sellele on võrgustiku alagrupid tihti seotud kitsamate piirkondadega. Alagrupi leviala selgitamiseks kaardistasime iga grupi värsiklastritesse kuuluvate värsside summaarse päritolu. Materjali üldjoontes tundes oli üksikute alarühmade olemust lähivaate põhjal (joonis 5) võrdlemisi lihtne tõlgendada, eelkõige tuginedes alarühma kesksetele värssidele, mis esindavad süžee stabiilsemaid elemente.



Joonis 5. Värsiklastrite võrgustiku alagrupp 1: lõunaeestilised värsitüübid.

- Grupp 1 esindab selgelt laulu lõunaeestilisi värsivariante (värsiklastrite levikuala piirdub Setu, Võru ja Mulgi keelealadega), mis üksnes üksikutel juhtudel külgnevad põhjaeestilisematega. Alagrupi värsitüüpide jada kuvab laulu algusosa süžee: päev soeb sulaste päid, kamm kukub merre või Koiva jõkke ning päev läheb paluma Peetrit ja Paavlit. Samas on teiste, põhjaeestilisemate alagruppide värsid siiski tihti tuntud ka lõunaeesti keelealal, eriti sagedasti Mulgimaal.
- Grupp 2 esindab laulu algusmotiivi: päev või minategelane soeb sulaste päid ja suga sulpsatab merre. Alagrupi keskne värss, mis seostub erinevate algusosa arendustega, on *kammib karjalaste päida*.
- Grupp 3 on taas selgelt lõunaeestiline alagrupp, lauluosa või -motiiv, kus pühakud keelduvad kaotatud kammi järele minemast. Kesksemad värsid on: *mine too hari meresta*, *kuldakammi kaldaasta*.
- Grupp 4 on tüüpiline põhjaeesti ja Mulgi pühakute palumise ja nende keeldumise motiiv, kesksed värsid on: *oo, Pieter, püha sulane*, *mine too mu suga meresta*, *päidelauda lainetesta*. Selle alagrupiga liitub laulualguse eriarendus (tüüpvärsiga *Jaanikene, kaanikene*), kus erinevalt grupi 2 algusmotiivist on juuste kammijaks vend (nii nagu enamikus karjala variantides).
- Grupp 5 esindab lauluosa, kus kammija läheb ise merre kammi otsima ning leiab sealt mõõga. Alagrupp koondab nii põhja- kui ka lõunaeestilisi (eriti Mulgi) värsse, mis on hästi näha kahest paralleelsest keskse värsi kujust: *kaelani kalakuduje ja kaalani kalakuduje*.
- Grupp 6 esindab laulu lõpumotiivi „Mõõk merest", kus leitud mõõk viiakse ilmarahva või sakste näha ja arvatakse, et tegu on sõjamõõgaga, mis on sõjas

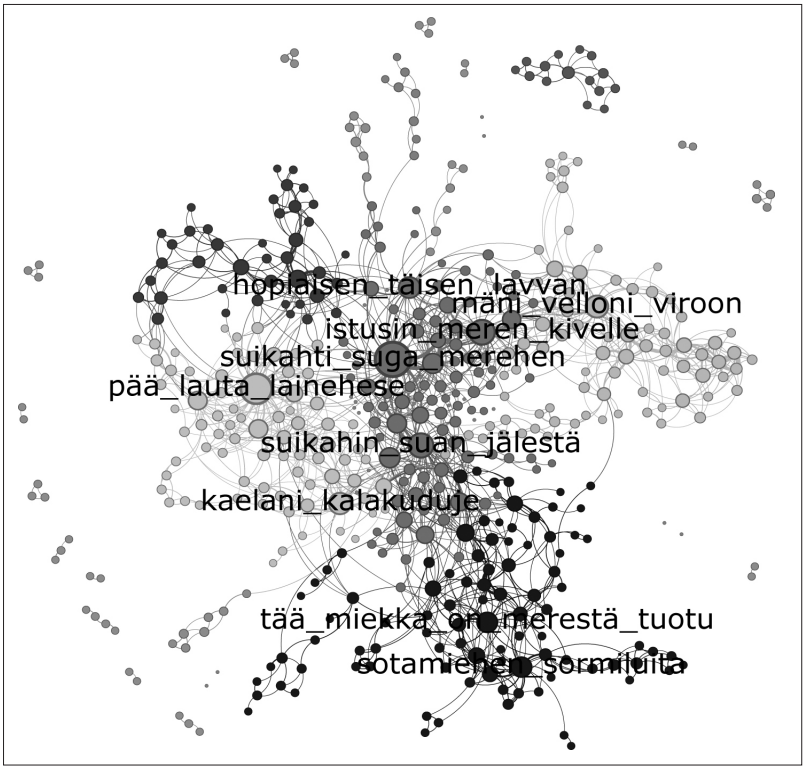
käinud ning seal päid ja luid raiunud. Alagrupi kesksemaks värsiks ongi *sõjameeste sõrmeluist*.
- Grupp 7 esindab terve süžee regionaalset arendust (nii nagu grupp 1) levikukeskmega Pärnu- ja Mulgimaal, kus minategelase vend toob pühadeks kasedkuused toa ette ning minategelane läheb sinna otsa pead kammima, hari kukub merre, minategelane läheb *kaelani kalakudusse* ja leiab mõõga.

Ülejäänud grupid esindavad laulutüübi variante, mis pärinevad Soomest, Karjalast ja Ingerimaalt. Kammi kaotamise laulu ongi kõige rohkem kirja pandud Ingerimaalt ja Karjala kannaselt ning sealsed arendused on võrgustikus nähtavamad.

- Grupp 8 esindab Ingerimaa tüüpilist laulualgusmotiivi, kus lauljal kasvavad nii pikad ja tihedad juuksed, et vajab paremat kammi, vend läheb Virumaale maarahasid maksma (ja kammi järele).
- Grupp 9 on pikem Ingerimaa ja idapoolse kannase laulualgusmotiiv juuste kasvatamisest emakodus, kesksemaks värsiks *kasvo miulle kassa pitkä* 'mulle kasvas pikk pats'.
- Grupis 10 on laulu keskne motiiv, mis on levinud Ingerimaal ja lõunakarjala aladel: vend läheb Virumaale, kammimäele vm, toob kammi, laulik-neiu istub kivile pead kammima ja suga sulpsatab merre. Kesksed värsid on: *mäni vellooni viroon* 'vend läks Virumaale', *istusin meren kivelle* 'istusin mere kivile', *harjaella hiuksiani* 'harjama oma juukseid', *sulpahti suka mereen* 'sulpsatas suga merre'.
- Grupp 11 esindab kogu peamise süžee paralleelarendust (vend tõi Virumaalt soa, istusin juukseid kammima, suga suisatas merre, tõin mõõga veest välja), kus keskne soa järele sukeldumise motiiv on ühine kogu põhjapoolsele laululalale. Grupi kesksed värsid: *suikahti suga merehen* 'suga suigatas merre', *painettaisin katsomaan* 'painutasin end vaatama', *suikahin sua jälestä* 'suigatasin soa järele', *mie miekan maalle* kannoin 'mina kandsin mõõga maale'.
- Grupp 12 on eelmise grupi vette sulpsatamise teemaga liituv Lääne-Ingeri eriarendus, kus mõõk puutub põlvedesse või helmeisse, tundis mehe meele, millele järgneb sõjamotiiv paljude irrutatud kehaosadega.
- Grupp 13 on eelmiste paralleelne (idapoolse Ingerimaa ja Karjala kannase) eriarendus, kus laulja leiab mõõga ja see viiakse inimestele näha (koju, vennale, mõisa), keksetes värssideks: *mie miekan merestä löysin* 'ma leidsin mõõga merest', *mie miekan venollein* 'viisin mõõga vennale'.
- Grupp 14 on kogu Ingerimaal levinud laulu lõpumotiiv, kus vaatajad (ja mõõk ise) arutlevad mõõga päritolu üle: see kas on või pole sõjas olnud, verd joonud ning luud ja liha söönud. Kesksed värsid: *tämä miekka on miestä syönyt* 'see mõõk on meest söönud', *miestä syönyt, verta juonut* 'meest söönud, verd joonud', *sotamiehen sormiluita* 'sõjamehe sõrmeluid'.

Kesksete alagruppide kõrval on veel mitmeid eriarendusi ja -motiive, osa neist on graafil näha väiksemate gruppidena, mis otsapidi liituvad kesksemate motiividega,

osa on aga nii haruldased (näiteks põhjakarjala versioon, kus pead kammib päeva poeg), et ei ole läbinud meetodi kasutamisel rakendatud kerget statistilist sõela. Põhjakarjala, Aunuse karjala ja soome keele karjalapärastes murretes teisendeid, mis on süžee poolest mõneti eripärased, on laulukogumis nii vähe, et need ei ole rikkaliku Ingerimaa ja Karjala kannase ainese kõrval üldpildis mõjule pääsenud.



Joonis 6. Kammi kaotamise ja mõõga leidmisega seostuvate laulutüüpide värsitüüpide graaf (samasse värsitüüpi kuuluvad värsiklastrid on kokku viidud käsitsi). Tekst on esitatud nende värsitüüpide puhul, millega laulutekstides külgneb kõige rohkem (>24) erinevaid värsitüüpe. Need on kogumis sõnastuselt kõige stabiilsemad ja ühtlasi tundub, et ka süžee kulu seisukohalt enamasti olulised värsid. Värsitüübi näidistekst esindab tegelikkuses varieeruvat värsikogumit (nii nagu eelpool kirjeldatud).

Seesugusel lähenemisel saame kiire ülevaate laulu süžee komponentidest, võrgustikus kesksemad värsid on olulised süžee arengu seisukohalt ja üksnes neile tuginedes on laulu põhisisu üldjoontes selge. Samas aga on näha, et võrgustiku ülesehituses on lahutamatult põimunud keelelise väljenduse ja sisu tasand. Erineva keelelise väljenduse tõttu on samasisulised arendused ja värsid sattunud eri gruppidesse, rääkimata sellest, et kogu võrgustik jaguneb kaheks (Eesti ja Soome korpuse osaks), kuigi laulu süžee ja tunnusvärsid suuresti kattuvad. Samuti eristuvad võrdlemisi selgelt lõunaeestilised alagrupid. Tulemus peegeldab lisaks tekstide keelelisele

ja folkloorsele varieeruvusele meie kasutatud sarnasusarvutuse meetodi piire: see on suuteline kokku tooma pigem värsid sama keele piires, kuid ei küüni enamasti ühendama värsse läänemeresoome eri murretes.

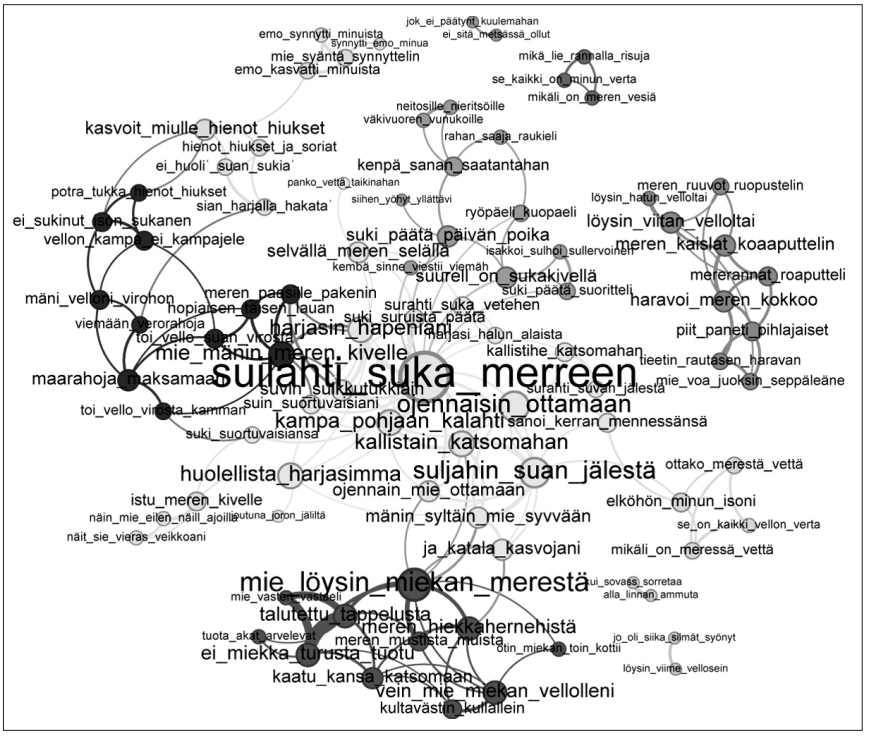
Selleks et vaadata motiivistiku hargnemist läänemeresoome traditsiooni üleselt, viisime käsitsi kokku värsid, mis kuulusid samadesse värsitüüpidesse. Samaks värsitüübiks lugesime värsid, kus sisusõnad olid samade tüvedega (onomatopoeetiliste sõnade puhul ka lähedaste tüvedega: *suisahti ~ sulpsatas ~ vulksatas*), erineda võisid sõnajärg, grammatilised sõnad, tuletusliited ja morfoloogilised lõpud. Pärast samasse värsitüüpi kuuluvate värsiklastrite kokkuviimist arvutasime uuesti külgnevusnäitajad ja joonistasime uue graafi (joonis 6).

Näeme, et keelelise varieerumise ühtlustamisel sulandub tekstikogum varasema kahe eraldi suure grupi asemel viiest suuremast alagrupist koosnevaks tervikuks. Eesti ainesele iseloomulik on algusmotiiv, kus päike või minategelane kammib kuldse harjaga karjalaste päid (kesksed värsid: *ari tal kuldane käessa*, *kammib karjalaste päidä*). Ingerimaa ja Lõuna-Karjala tavapärane algus seostub neiu pikkade juuste kasvatamisega, mille ohjeldamiseks toob vend Virumaalt kammi (keskne värss: *mäni vellooni Viroon*). Järgnevad eesti ja Ingerimaa paralleelsed, kuid tükati omavahel põimunud arendused. Eesti variantides: kammi kaotamine merre, pühakute palumine, kammi järele minek ja merest mõõga leidmine (kesksed värsid: *päälauta lainehese*, *kaelani kalakuduje*). Ingerimaa ja Karjala kannase variantides mererannal kammimine, kammi vette kukkumine ja merest mõõga leidmine (kesksed värsid: *istusin meren kivelle*, *suikahti suka merehen*, *suikahin suan jälestä*). Viimane suurem sisuüksus on ühine mõõga arvustamine.

Lõpuks rakendasime külgnevate värsside arvutust ja võrgustikuanalüüsi kitsama piirkonna lauludel, et testida meetodit keeleliselt ühtlasemal ainesel. Valisime selleks karjala laulud, mille puhul karjala traditsioonile iseloomulikud päeva poja ning meres hukkumise motiivid tundusid suures võrgustikus kaduma minevat. Kaasasime sellesse valimisse 70 eri karjala aladelt pärit teksti. Tulemus on kuvatud joonisel 7.

Näeme, et laulu keskse sisu moodustab ikkagi kammi kaotamise ja sellele järele sukeldumise motiiv, millega liitub mõõga leidmise motiiv, kuid selgelt eristuvad kaks algusmotiivi: ka Ingerimaal laialt levinud juuste kasvamise ja venna Virus käimisega laulualgus, kuid teiseks algusmotiiv, kus pead soeb päeva poeg. Karjala teisendites saab kammija – päeva poeg või vend – tavaliselt merre minnes surma, mõnikord saadab ta sugulastele hoiatusi, mõnes variandis leiab õde surnud venna.

Kasutatud meetod (bigrammikoostise sarnasuse alusel leitud sarnaste värsside klastrite kuvamine võrgustikuna klastritesse kuuluvate värsside külgnevuse alusel) võimaldab saada kergesti ülevaate laulutüübi kõige tavalisemast sisust, peamistest motiividest ja võtmevärssidest. Tulemuste mõistmiseks ja tõlgendamiseks ei ole vaja läbi lugeda ega omavahel kõrvutada kõiki tekste, vaid piisab tutvumisest olulisemate süžeearendustega. Keeleline varieeruvus aga teeb siingi ülevaate saamise keerukaks ning kergem on tulemusi tõlgendada keeleliselt ühtlasema ainese puhul. Lisaks tulevad seesugusel vaatlusel paremini esile sagedasemad motiivid ja arendused, kuid haruldasemad eriarendused võivad kergesti pildilt välja jääda.



Joonis 7. Kammi kaotamise ja mõõga leidmisega seonduvad värsid karjala lauludes.

Kokkuvõtteks

Sõltumata uurimuse eesmärkidest ja teoreetilisest raamistusest on arvutuslikud meetodid abiks tekstikogumite analüüsimisel ja tõlgendamisel, kui need on nii mahukad, et nende läbilugemine ja inimjõul võrdlemine on liiga aeganõudev (nt Cohen 1999: 23; Moretti 2000). On tõeliselt keeruline saada head ülevaadet sadadest või tuhandetest tekstidest lihtsalt lugemise teel ilma arvuti abita (vt Mäkelä jt 2020). Samas on arvutuslikul teel saadud tulemuste kontrollimiseks, hindamiseks ja adekvaatseks tõlgendamiseks siiski tekstide või väiksemate tekstiosade lähilugemine ja uuritava ainese hea tundmine vältimatu.

Läänemeresoome regilaulukorpus koos hästi korrastatud metaandmetega võimaldab teha mitmesuguseid teksti- ja andmepäringuid ning tulemusi eri viisil visualiseerida. Regilaulude jaoks välja töötatud ning Runoree keskkonnas rakendatud värsside ja tekstide sarnasuse arvutamise meetodid hõlbustavad oluliselt sarnaste üksuste tuvastamist ja nende võrdlemist. Viimaks, täiendav värsside külgnevusseoste võrgustiku vaatlus võimaldab tuvastada aineses tihedamalt seotud sisuüksusi (stereotüüpseid värsse ja värsipaare, motiive ja laulutüüpe), esile tuua ainestikus kesksed värsitüübid ning sedakaudu saada kiiresti ülevaade näiteks ühte tüüpi kuuluvate laulude ainesest ning sisuhargnemistest. Samas on arvutusliku analüüsi

meetodite abil laulude sisutasandini jõudmisel jätkuvalt probleemiks laulutekstide suur keeleline varieeruvus ning ebaühtlus aineses.

Artikli valmimist on toetanud Soome Akadeemia (projektid 333138 „Vormellik intertekstuaalsus, teemavõrgustikud ja poeetiline varieeruvus läänemeresoome suulise luule piirkonnatraditsioonides" (FILTER) ja 346342 „Läänemeresoome suulise luule piirkondlikud kultuurid: võrdlev vaatenurk"), Eesti Teadusagentuur (projekt PRG1288 „Folkloorse varieeruvuse korpuspõhine käsitlus: regilaulutraditsiooni piirkondlikud stiilid, teemavõrgustikud ja suhtlusviisid", tippkeskus TK215 „Eesti juured: rahvastiku ja kultuuri kujunemise transdistsiplinaarsete uuringute tippkeskus") ning Eesti haridus- ja teadusministeerium (projekti EKKD126 „Kuidas allikatest saab kultuur: eesti aines Eesti Kirjandusmuuseumi kogudes ja andmebaasides II").

ARHIIVIALLIKAD

Viited digiteeritud näitetekstide käsikirjalistele allikatele Eesti Kirjandusmuuseumi (EKM) Eesti Rahvaluule Arhiivis (ERA):

AES – Akadeemilise Emakeele Seltsi kogu
E – Matthias Johann Eiseni rahvaluulekogu
ERA – Eesti Rahvaluule Arhiivi rahvaluulekogu
ERM – Eesti Rahva Muuseumi rahvaluulekogu
EÜS – Eesti Üliõpilaste Seltsi rahvaluulekogu
H – Jakob Hurda rahvaluulekogu

VEEBIVARAD JA ELEKTROONILISED TÖÖRIISTAD

ERAB = Eesti regilaulude andmebaas. Koost Janika Oras, Liina Saarlo, Mari Sarv, Kanni Labi, Merli Uus, Reda Šmitaite. Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiiv. ttps://www.folklore.ee/regilaul/andmebaas

FILTER andmebaas. Koost Maciej Janicki, Eetu Mäkelä ja FILTER projekti töörühm. Helsingi Ülikool, Soome Kirjanduse Selts, Eesti Kirjandusmuuseum.

FILTER visualizations. Loonud Maciej Janicki, Kati Kallio, Eetu Mäkelä, Jukka Saarinen, Mari Sarv. Tööversioon. Helsingi Ülikool, Soome Kirjanduse Selts, Eesti Kirjandusmuuseum. https://filter-visualizations.rahtiapp.fi

JR = Julkaisemattomat runot, digiteeritud versioon. Soome Kirjanduse Selts.

Octavo UI. Loonud Eetu Mäkelä. Helsingi Ülikool. https://jiemakel.github.io/octavo-nui

Runoregi. Loonud Maciej Janicki, Kati Kallio, Mari Sarv, Eetu Mäkelä. Helsingi Ülikool (HELDIG), Soome Kirjanduse Selts, Eesti Kirjandusmuuseum (versioon kuupäevast 28.11.2023). https://runoregi.rahtiapp.fi

Harja otsimine. https://runoregi.rahtiapp.fi/type?id=erab_001001003
https://runoregi.rahtiapp.fi/type?id=erab_orig2312;
https://runoregi.rahtiapp.fi/type?id=erab_orig10132

Miekka merestä.
https://runoregi.rahtiapp.fi/dendrogram?source=type&type_id=skvr_t010100_2270

Mõõk merest. https://runoregi.rahtiapp.fi/type?id=erab_001001013 https://runoregi.rahtiapp.fi/type?id=erab_orig958; https://runoregi.rahtiapp.fi/type?id=erab_orig1619)

Suka mereen. https://runoregi.rahtiapp.fi/type?id=skvr_t010100_3380

SKVR = SKVR-tietokanta – kalevalaisten runojen verkkopalvelu. Suomalaisen Kirjallisuuden Seura. https://skvr.fi

KIRJANDUS

Bastian, Mathieu; Heymann, Sébastien; Jacomy, Mathieu 2009. Gephi: an open source software for exploring and manipulating networks. Proceedings of the International AAAI Conference on Weblogs and Social Media, kd 3, nr 1, lk 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Blondel, Vincent D.; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne 2008. Fast unfolding of communities in large networks. – Journal of Statistical Mechanics: Theory and Experiment, nr 10, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Cohen, Margaret 1999. The Sentimental Education of the Novel. Princeton: Princeton University Press.

Hiiemäe, Mall 2006. Kosmogoonilise harja otsimine. – Regilaul – esitus ja tõlgendus. (Eesti Rahvaluule Arhiivi toimetused 23.) Toim Aado Lintrop. Tartu: Eesti Kirjandusmuuseum, lk 21–48.

Janicki, Maciej 2023. Large-scale weighted sequence alignment for the study of intertextuality in Finnic oral folk poetry. – Journal of Data Mining and Digital Humanities, nr NLP4DH. https://doi.org/10.46298/jdmdh.11390

Janicki, Maciej; Kallio, Kati; Sarv, Mari 2023. Exploring Finnic written oral folk poetry through string similarity. – Digital Scholarship in the Humanities, kd 38, nr 1, lk 180–194. https://doi.org/10.1093/llc/fqac034

Kalkun, Andreas 2015. Seto laul eesti folkloristika ajaloos. Lisandusi representatsiooniloole. (Eesti Rahvaluule Arhiivi toimetused 33.) Tartu: Eesti Kirjandusmuuseum.

Kallio, Kati; Janicki, Maciej; Mäkelä, Eetu; Saarinen, Jukka; Sarv, Mari; Saarlo, Liina 2023. Eteneminen omalla vastuulla. Lähdekriittinen laskennallinen näkökulma sähköisiin kansanrunoaineistoihin. – Elore, kd 30, nr 1, lk 59–90. https://doi.org/10.30666/elore.126008

Kikas, Katre 2014. Folklore collecting as vernacular literacy: Establishing a social position for writing in the 1890s Estonia. – Vernacular literacies – Past, present and future. Toim Ann-Catrine Edlund, Lars-Eric Edlund, Susanne Haugen. (Northern Studies Monographs 3. Vardagligt skriftbruk 3.) Umeå: Umeå University, Royal Skyttean Society, lk 309–323.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel; Tugwell, David 2004. The sketch engine. – Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient, France, July 6-10, 2004. Lorient: Université de Bretagne-Sud, lk 105–115.

Kundozerova 2022 = Мария Кундозерова, База данных «Карельские руны»: идея создания, концепция, перспективы. – Альманах североевропейских и балтийских исследований, вып. 7, lk 233–240. https:// doi.org/10.15393/j103.art.2022.2386

Lintrop, Aado 1999. Suur tamm, kuduvad neiud ja punane paat, kadunud harjast rääkimata. – Mäetagused, nr 10, lk 7–23. https://doi.org/10.7592/MT1999.10.tamm

Lintrop, Aado 2000. Suur tamm ja õde-venda. – Mäetagused, nr 13, lk 24–42. https://doi.org/10.7592/MT2000.13.suurtamm

Lintrop, Aado 2024. Kosmogooniline hari ja selestiline kiik. – Keel ja Kirjandus, nr 3, lk 219–237. https://doi.org/10.54013/kk795a1

Moretti, Franco 2000. The Slaughterhouse of Literature. – Modern Language Quarterly, kd 61, nr 1, lk 207–227. http://muse.jhu.edu/journals/mlq/summary/v061/61.1moretti.html

Mäkelä, Eetu; Koivunen, Anu; Kanner, Antti; Janicki, Maciej; Harju, Auli; Hokkanen, Julius; Seuri, Olli 2020. An approach for agile interdisciplinary digital humanities research – a case study in journalism. – TwinTalks 2020: Understanding and Facilitating Collaboration in Digital Humanities 2020. Proceedings of the Twin Talks 2 and 3 Workshops at DHN 2020 and DH 2020. (CEUR Workshop Proceedings 2717.) Toim Steven Krauwer, Darja Fišer. Aachen: RWTH Aachen University, lk 4–14. http://ceur-ws.org/Vol-2717/paper01.pdf

Piela, Ulla 2023. Toiveiden maa. Ylioppilaiden matkakertomuksia autonomian ajalta. (Tietolipas 282.) Helsinki: Suomalaisen Kirjallisuuden Seura.

Salmela, Alfred 1964. Päivän suka. – Kalevalaseuran vuosikirja, kd 44, lk 100–116.

Tarkka, Lotte 2005. Rajarahvaan laulu. Tutkimus Vuokkiniemen kalevalamittaisesta runokulttuurista 1821–1921. (Suomalaisen Kirjallisuuden Seuran Toimituksia 1033.) Helsinki: Suomalaisen Kirjallisuuden Seura.

Mari Sarv (snd 1972), PhD, Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiivi juhtivteadur (Vanemuise 42, 51003 Tartu), mari@haldjas.folklore.ee

Kati Kallio (snd 1977), PhD, Helsingi Ülikooli folkloristika dotsent ja Soome Akadeemia stipendiaatteadur Soome Kirjanduse Seltsis (Hallituskatu 1, 00171 Helsingi, Soome), kati.kallio@finlit.fi

Maciej Michał Janicki (snd 1989), PhD, Helsingi Ülikooli digihumanitaaria keskuse teadur (järeldoktor), maciej.janicki@helsinki.fi

Computational insights into the variation of Finnic folk songs: "Searching for the Comb" and "Sword from the Sea"

Keywords: folklore, oral poetry, runosong, digital humanities, Finnic languages, variation

The article introduces the joint Finnic runosong database and associated web environments and applications developed collaboratively by computer scientists and folklorists from Finland and Estonia. These tools facilitate new approaches to analyzing the extensive dataset. Within the research framework, various computational solutions have been devised in order to identify and associate with one another simi-

lar verses and texts that differ in orthography, language, and content. These methods have also been implemented in the web environment Runoregi (runoregi.rahtiapp.fi), allowing researchers and enthusiasts interested in traditional oral poetry to easily navigate the network of variant verses, motifs and texts, and to compare various texts and their elements. Additionally, there is a web application for maps and other visualizations integrated with the database and Runoregi environment.

While Runoregi serves as a valuable tool for the close reading and comparison of texts, obtaining an overview of large amounts of texts (the database currently contains over 280,000 texts) remains a challenge. We address this issue through an examination of the frequently contaminated song types "Searching for the Comb" and "Sword from the Sea". Given that not all texts in the database are consistently typologized by folklorists, our sample includes texts identified by means of similarity calculations as similar to those sorted under the types under consideration. We computed adjacency scores for verse clusters obtained as a result of clustering verses by their similarity scores using the Chinese whispers method, presenting the results as a network graph (with verse clusters as nodes and adjacency scores as edges). The groups appearing in the network reflect regional plot developments and elements. Despite sharing plotlines and even poetic formulas, a clear divide emerged between Northern and Southern Finnic texts. As our verse similarity calculations may not capture linguistically distant variants, we manually consolidated variants of the same verse (with identical root composition of content words) across different dialects and languages. By applying adjacency computation and network visualisation, the graph now represents the general Finnic plot with the main alternative developments. The graph also highlights the stabler cross-Finnic verse types associated with significant plot turns.

Mari Sarv (b. 1972), PhD, Lead Research Fellow, Estonian Folklore Archives of the Estonian Literary Museum (Vanemuise 42, 51003 Tartu), mari@haldjas.folklore.ee

Kati Kallio (b. 1977), PhD, Academy Research Fellow at the University of Helsinki and the Finnish Literature Society (Hallituskatu 1, 00171 Helsinki, Finland), kati.kallio@finlit.fi

Maciej Michał Janicki (b. 1989), PhD, Postdoctoral Researcher, Department of Digital Humanities, University of Helsinki, maciej.janicki@helsinki.fi