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 komplekssuse 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?

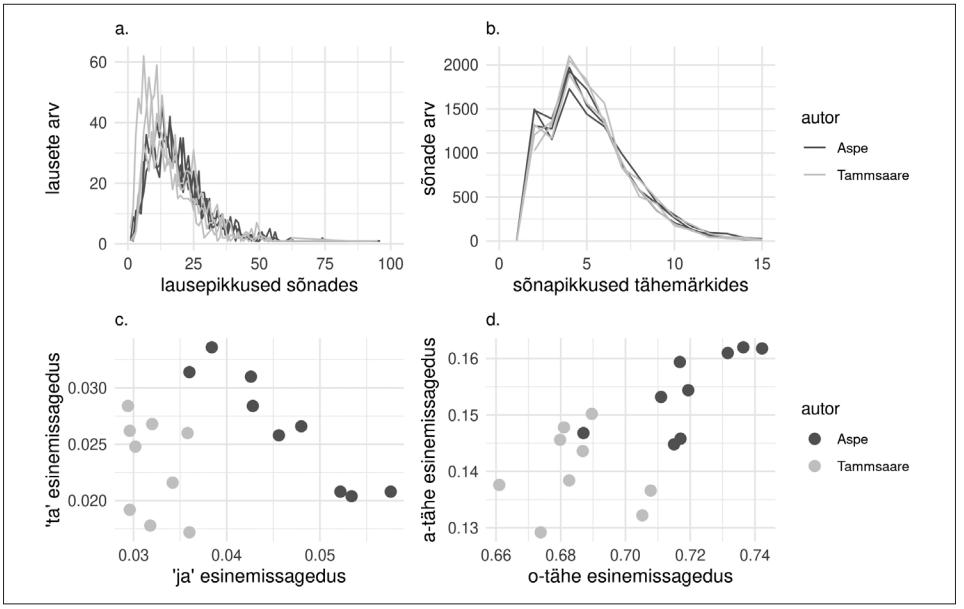
Mitmemõõtmelise stiilikäsituse poole

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).

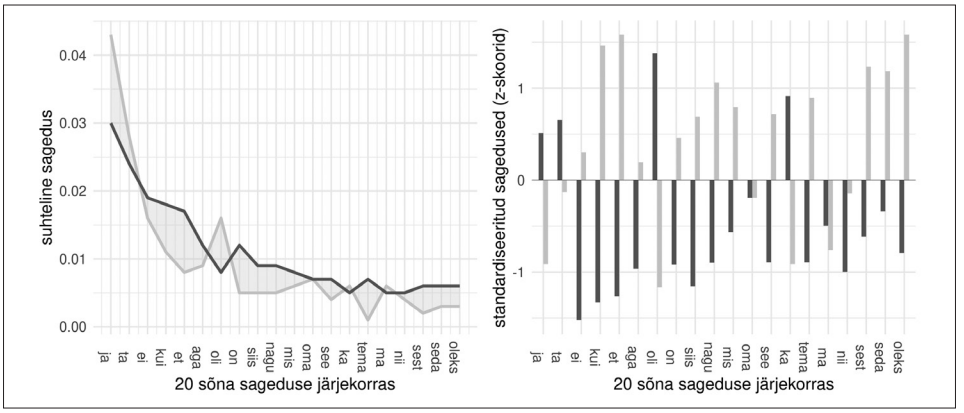
Tekst	Suhtelised sagedused					Standardiseeritud sagedused (*z*-skoorid)				
	ja	**ta**	**ei**	**kui**	**et**	**ja**	**ta**	**ei**	**kui**	**et**
Aspe „Aastate pärast”	0,054	0,021	0,019	0,012	0,011	1,715	–0,721	0,304	**–0,932**	**–0,316**
Aspe „Ennosaare Ain”	0,043	0,028	0,016	0,011	0,008	0,511	0,655	–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.

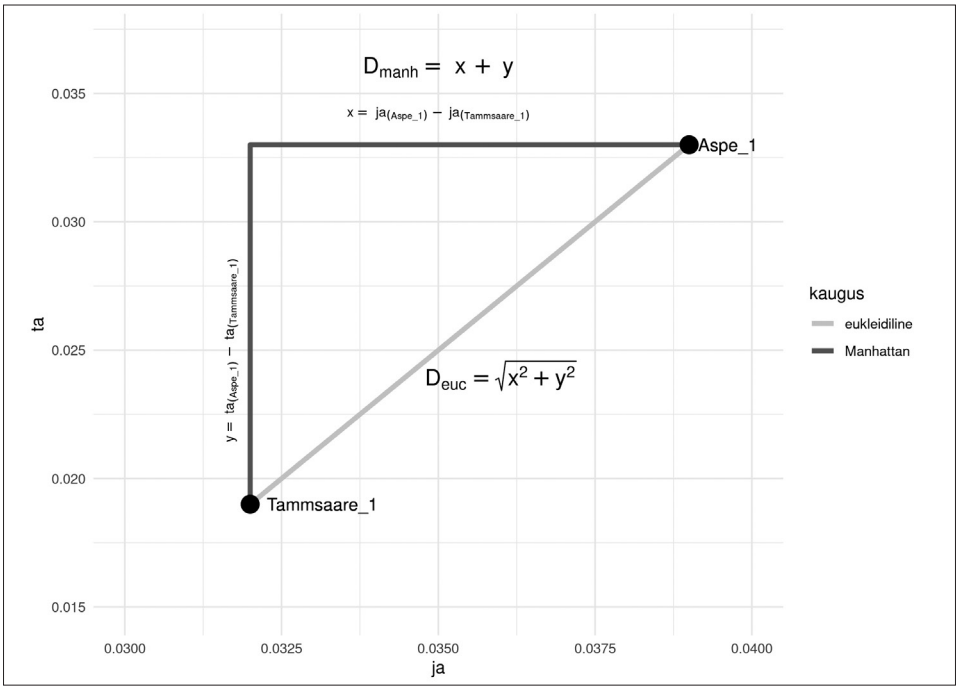
Joonise 2 parempoolsel osal võib näha *z*-skoori kujul esitatud sageduste mõju: kuna nüüd on iga tunnust skaleeritud standardhälbega, on tulemuseks juba paljude tunnuste lõikes võrreldavad erinevusmäärad. Isegi kui sõna esinemissagedus on madal, võib see delta arvutamisse panustada samal (matemaatilisel)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

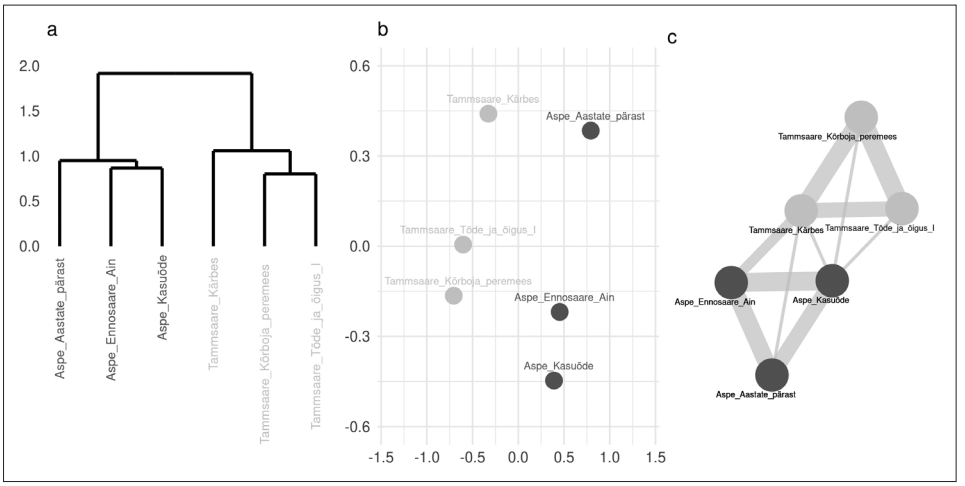
Tabel 3. Kuue teksti vastastikuste kauguste (Burrowsi delta) ruutmaatriks.

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. A g l o m e r a t i i v n e h i e r a r h i l i n e k l a s t e r d a m i n e 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 m i t m e m õ õ t m e l i s t s k a l e e r i m i s t (*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

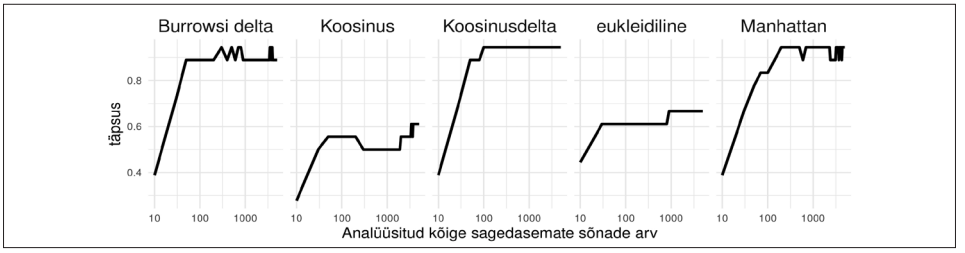
Selleks et demonstreerida kirjeldatud meetodit ning anda aimu selle toimivusest erinevate eesti kirjanduse tekstiliste omaduste puhul, viisin läbi väikese autorituvastuskatse varase eesti ilukirjanduse põhjal. Kasutatud korpus on avatud juurdepääsuga (Uiboaed 2018) ning seda on varem rakendatud stilomeetria 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.

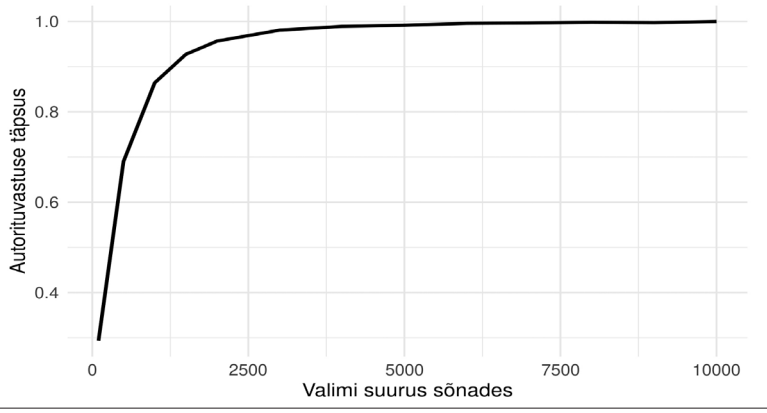
Üldiselt kipub tuvastuse täpsus kõigi kaugusmõõdikute korral tipnema ja stabiliseeruma umbes saja kõige sagedasema sõna kaasamisel. Edasine sõnavektorite mõõtmelisuse kasv ei näi tuvastamise jaoks enam lisainformatsiooni andvat. See kinnitab ootuspäraselt veel kord, et kõige sagedasematel funktsioonisõnadel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