Tartu Ülikool

Humanitaarteaduste ja kunstide valdkond

Eesti ja üldkeeleteaduse instituut

Leena Karin Toots

LGBT+ kogukonnaga seotud sõnavara

Eesti LGBT+ aktivistide taskuhäälingutes

Bakalaureusetöö

Juhendajad Liina Lindström ja Maarja-Liisa Pilvik

Tartu 2023

Sisukord

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud lähtudes Tartu Ülikooli eesti ja üldkeeleteaduse instituudi lõputöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Sissejuhatus

LGBT+ kogukonnaga seonduvad teemad muutuvad tänapäeva ühiskonnas järjest aktuaalsemaks. Kuigi suur osa kogukonda puudutavat temaatikat jääb pigem ühiskonnateaduste uurimisvaldkonda, on LGBT+ kogukonnal ja teistel vähemusgruppidel tihe seos ka keelega. Keelekasutus määrab muuhulgas ära selle, kuidas kajastatakse kogukonda ühiskondlikus diskussioonis ning meedias ja millist suhtumist keeleliste vahenditega edasi antakse. Varasemalt on New York Timesi ajalehes LGBT+ kogukonna ja selle liikmete dehumaniseerimise uurimiseks rakendatud arvutilingvistika meetodeid (Mendelsohn, Tsvetkov, Jurafsky 2020). Korpuslingvistika meetodite rakendamisest ja probleemidest LGBT+ sõnavara analüüsimisel on põhjaliku ülevaate teinud Heiko Motschenbacher (2022: 41–63).

Kogukonna liikmetele on kogukonnaspetsiifiline sõnavara oluline, kuna see aitab neil oma identiteeti määratleda, ülejäänud maailmaga suhestuda ning tugevdab kogukonnasisest ühtsustunnet. LGBT+ kogukonnaga seotud sõnavara päritolu ja selle muutumist on käsitlenud Paul Baker (2012) ja Heiko Motschenbacher (2022). Baker keskendus oma uurimuses LGBT+ kogukonnaga seotud keelekasutuse muutumisele kogukonda puudutavates konverentsiettekannetes ning Motschenbacher kirjeldas USA gei meeste sõnavara enne ja peale LGBT+ aktivismi murrangupunktiks olnud Stonewalli mässu. Eraldi on uuritud ajalooliselt stigmatiseerivate ja negatiivsete konnotatsioonidega sõnade nagu *queer* (ee *kväär*) või *faggot* (ee *pede*) omaks võtmist ning enesesildistamise protsessi (Galinsky jt 2013).

Siinne töö keskendub LGBT+ kogukonnaga seotud sõnavarale kahes Eesti LGBT+ aktivistide taskuhäälingus, Lillas Agendas ja Homokringlis. Taskuhäälingud on aktivistide sõnavara uurimiseks hea allikas nende lihtsa kättesaadavuse ja loomuliku keele kasutuse tõttu. Kuna tegu on suhteliselt vabas vormis vestlusega, annavad need hea ülevaate inimeste tavapärasest keelekasutusest. Lisaks on taskuhäälingud eesti keele uurimise kontekstis suhteliselt uus andmeformaat, mis pakub võimalust testida, kui hästi tulevad andmetega toime eesti keele automaattöötlusvahendid nagu automaatne kõnetuvastus ning morfoloogiline analüsaator.

Bakalaureusetöös otsin vastuseid järgmistele uurimisküsimustele:

1. Millist LGBT+ kogukonnaga seotud sõnavara kasutavad Eesti LGBT+ aktivistid oma taskuhäälingutes?
2. Kuidas tuleb automaatne kõnetuvastus toime LGBT+ kogukonnaga seotud suhteliselt uue ja inglise keelest mõjutatud sõnavaraga?
3. Milliste meetodite ja vahendite abil on võimalik taskuhäälingutest LGBT+ sõnavara tuvastada?

Bakalaureusetöö esimene etapp on taskuhäälingute transkribeerimine. Selleks kasutan Tallinna Tehnikaülikoolis loodud eesti keele automaatset kõnetuvastustarkvara (Olev, Alumäe 2022). Töö käigus kaardistan taskuhäälingute transkribeerimisel esinevaid probleeme ja hindan lõpptulemuse kvaliteeti. Peale transkribeerimist lemmatiseerin transkriptsioonid EstNLTK[1] teekide kogumikku kuuluva Vabamorfi[2] morfoloogilise analüsaatoriga. Morfoloogilise analüüsi kvaliteedi hindamine töö eesmärkide hulka ei kuulu, kuid pööran siiski tähelepanu võimalikele läbivamatele probleemidele.

LGBT+ kogukonnaga seotud sõnavara kasutust pole eesti keeles varem uuritud, seega käsitleb bakalaureusetöö suhteliselt uut valdkonda. Ülemaailmselt juhib kogukonnaga seotud sõnavara arengut inglise keel, mistõttu on enim uurimusi ilmunud nimelt selle keele kontekstis. See jätab kõrvale teisi keeli emakeelena kõnelevate LGBT+ inimeste kogemused ning raskused uute sõnade loomisel, kasutusele võtmisel ja levitamisel. Ivana Buntak (2020) leidis inglise ja horvaatia keele LGBT+ kogukonnaga seotud neologisme võrreldes, et suurem osa LGBT+ sõnavarast jõuab teistesse keeltesse esmalt laensõnadena. Üks bakalaureusetöö eesmärke on välja selgitada, kas ja kui suurel määral kasutavad Eesti aktivistid inglise tsitaat- või laensõnu ning kui laialt on levinud uuemate sõnade eesti keelde kohandatud versioonid.

Töö koosneb kolmest peatükist. Esimeses peatükis annan ülevaate LGBT+ kogukonnaga seotud sõnavara päritolust ja ajaloost, selle seostest normatiivsuse ja sildistamisteooriaga, olulisusest LGBT+ kogukonna jaoks ning vajadusest sõnavara kaasajastada. Lisaks toon

[1] https://github.com/estnltk/estnltk
[2] https://github.com/Filosoft/vabamorf

välja korpuslingvistika meetodite rakendamise kitsaskohad LGBT+ sõnavara uurimisel ning kirjeldan lühidalt eesti keele automaatset kõnetuvastussüsteemi ja morfoloogilist analüsaatorit. Teises peatükis annan ülevaate kasutatud andmetest ja töö etappidest. Kolmandas peatükis toon üksikasjalikult välja töö igas etapis kasutatud meetodid, nende rakendamisel esinenud raskused ja etappide tulemused.

1. Taust

1.1. LGBT+ sõnavara ja aktivism

1.1.1. Normatiivsus ja LGBT+ sõnavara teke

LGBT+ kogukonnaga seotud sõnavara on väga tihedalt seotud (lingvistilise) normatiivsusega. Normatiivsuse all mõistetakse seda, et teatud sotsiaalseid praktikaid nähakse ühiskonnas normaalsetena ning teisi ebanormaalsetena (Motschenbacher 2019). See tähendab, et normidel on võime määrata, milliseid identiteete ja tegevusi ühiskonnas marginaliseeritakse ning milliseid mitte. Kuna normid eeldavad, et neid järgitakse, on normidega vastuollu sattumise tagajärjeks tihti halvakspanu või sotsiaalsed sanktsioonid. Normatiivsus toetab seega eelkõige normidele alluvaid ja seeläbi ühiskonnas privilegeeritud inimgruppe ja surub alla normidele mittealluvaid marginaliseeritud gruppe. (Motschenbacher 2022: 12–13)

Läänemaailmas algas seksuaalidentiteetide normatiividele allutamine 19. sajandi lõpus. Sellele pani aluse meditsiiniterminite kasutuselevõtt inimeste seksuaalsuse alusel kategoriseerimiseks, millest arenesid hiljem välja seksuaalidentiteetide sildid nagu *hetero*, *gei*, *lesbi* ja *biseksuaal*. Seksuaalidentiteetide normatiivsel kategoriseerimisel oli nii positiivseid kui negatiivseid tagajärgi. Ühest küljest surus seesugune kategoriseerimine identiteedid normatiivide „kastidesse“, teisest küljest andis identiteetidele nime andmine neile teatud legitiimsuse, muutes need nähtavamaks ja aktuaalsemaks. Seksuaalvähemuste hulka kuuluvatel inimestel oli seega esmakordselt võimalus end oma identiteedi sildi all tõsiseltvõetavana esitada. (*ibid*., lk 13)

Ühiskond jaguneb normatiivsuse silmis laias laastus kaheks – normidele vastavaks ja neist hälbivaks. LGBT+ kogukonnast rääkides on üks olulisimaid normatiive heteronormatiivsus, mille kohaselt on normiks heteroseksuaalsed suhted. Mitteheteroseksuaalsed suhted hälbivad heteronormatiivsuse vaatepunktist automaatselt normist. (*ibid*., lk 14) Kuna normatiivsuse rakendamiseks on vaja nii normidele vastavaid kui neist hälbivaid nähtuseid kirjeldada, avaldub selle mõju tugevalt ka keeles. LGBT+ kogukonnast rääkides keskendun edaspidi normidest hälbimise keelelisele avaldumisele.

1.1.2. Sildistamine ja LGBT+ aktivism keeles

Normidest hälbimise kirjeldamisel on oluline roll sildistamisteoorial (ingl *labeling theory*) (Berk 2015: 150). Sildistamisteooria jaotab normidest hälbimise omakorda kaheks – primaarseks hälbimiseks, mis ei ole laiemas ühiskonnas veel tajutav, ning sekundaarseks hälbimiseks, mille põhjustavad primaarse hälbimise avalikustumisel tekkinud negatiivsed ühiskondlikud tagajärjed, eelkõige inimese halvustavalt sildistamine. (Lemert 1951: 75–77) LGBT+ inimeste puhul on inglise keeles selliste halvustavate siltidena kasutatud näiteks sõnu *faggot*, *queer* ja *homosexual*. (Motschenbacher 2022: 34) Eesti keeles on mõned sarnased näited *pede*, *homo* ja *lilla* (Ostrat 2023).

Kuigi seesugused sildid mõjuvad normist hälbivale inimesele esialgu negatiivselt, põhjustades diskrimineerimisest tingitud enesehinnangu langust, häbi ja vaimse tervise probleeme, annavad need talle ühtlasi võimaluse silte omaks võtta ja end seeläbi võimestada. Sildi omaks võtmine annab inimesele võimaluse leida sama identiteediga inimesi, luua nendega sidemeid ja tunda end taaskord aktsepteerituna. Sekundaarse hälbimise viimane samm on omaks võetud sildi all poliitiliselt aktiivseks muutumine. (Motschenbacher 2022: 34) Halvustavate ning stigmatiseerivate siltide omaks võtmine kirjeldab väga hästi LGBT+ kogukonna kogemust ning on üks LGBT+ aktivismi alustalasid.

Inglise keeles on stigmatiseerivate siltide omaks võtmist uurinud näiteks Adam D. Galinsky jt (2013). Nende eesmärk oli selgitada välja, kuidas on omavahel seotud üksikisikute ja ühiskonnagruppide võimestatus ning siltide omaks võtmine. Uurimuses püstitati hüpotees, et enesesildistamine (ingl *self-labeling*) oma kogukonda alandava terminiga muudab kogukonna teiste silmis võimsamaks ning et võim ja enesesildistamine on omavahel vastastikku seotud. See tähendab, et inimesed võtavad stigmatiseerivaid silte suurema tõenäosusega omaks siis, kui nad tunnevad, et nende grupil on selleks piisavalt võimu. Samuti oletati, et kuna enesesildistamine seab kasutatud sildi negatiivse varjundi kahtluse alla, võib see aidata kaasa stigma vähendamisele. (*ibid.*)

LGBT+ kogukonnaga seotud siltidest uuriti ingliskeelset sõna *queer*, mis on inglise keelt kõnelevate maade LGBT+ kogukondades praeguseks laialt levinud. Väljaspool LGBT+ kogukonda on sõnal *queer* aga endiselt negatiivne tähendusvarjund. Et selgitada välja, kas enesesildistamise puhul on sildi seos stigmatiseerimisega nõrgem kui kogukonnavälisel sildistamisel, katsetati seda võrdluses mittestigmatiseeriva sildiga *LGBT* ja enamusgruppi kirjeldava sildiga *straight* (ee *sirge, hetero*). Katse tulemusena selgus, et sõna *queer* tajuti enesesildistamise puhul tunduvalt vähem negatiivsena kui juhul, kus sõna kasutas teine inimene. Teiste katses kasutatud sõnadega seesugust tendentsi esile ei tulnud. (*ibid.*)

1.1.3. LGBT+ sõnavara päritolu, olulisus ja muutumine

Nagu varasemalt mainitud, on LGBT+ sõnavara LGBT+ identiteetide ajaloolise patologiseerimise tõttu suuresti pärit kunagisest meditsiiniterminoloogiast. Viimase sajandi jooksul on LGBT+ kogukond aga iseseisvalt uue, kõiki erinevate soo- ja seksuaalidentiteetidega inimesi kaasavama ning positiivsema sõnavara loomise ja propageerimisega tegelenud (Baucom 2018). LGBT+ kogukonda kirjeldava sõnavara areng on oluline eelkõige seetõttu, et sõnavaral, mida kogukonna ja selle liikmete kirjeldamiseks kasutatakse, on võim LGBT+ inimeste olemasolu kas jaatada või eitada (Buntak 2020). LGBT+ kogukonna kirjeldamisel kasutatava sõnavara arengut ilmestavad Paul Bakeri (2012) ja Heiko Motschenbacheri (2022) diakroonilised uurimused LGBT+ sõnavara ja selle kasutuse muutumisest.

Baker (2012) keskendus oma uurimuses sellele, kuidas on 18 aasta jooksul muutunud soo- ja seksuaalvähemuste keelekasutusele pühendatud konverentsi „Lavender Languages and Linguistics Conference“ LGBT+ kogukonda puudutav keelekasutus. Tema uurimusest ilmnes, et keelekasutus on eristavamatelt ja spetsiifilisematelt terminitelt nagu *gay* (ee *gei*) ja *lesbian* (ee *lesbi*) liikunud laiemate terminite nagu *queer* ja *LGBT* poole. Samuti välditakse rohkem seksuaalkäitumisele viitavaid meditsiinilise alatooniga sõnu nagu *homosexual* (ee *homoseksuaal*) ning *transsexual* (ee *transseksuaal*). (*ibid*.)

Motschenbacher (2022) uuris lähemalt seda, kuidas erineb LGBT+ kogukonnaga seotud keelekasutus USA gei meeste hulgas enne ja peale 1969. aastal New Yorgis toimunud Stonewalli mässu. Stonewalli mässu peetakse LGBT+ aktivismi murrangupunktiks, mistõttu tähendas see soo- ja seksuaalvähemuste hulka kuuluvate inimeste jaoks suuri muutusi. Üks olulisim neist muutustest oli samasooliste suhete nähtavamaks tegemine ja seeläbi normaliseerimine. Motschenbacheri uurimuses ilmnes, et muutuste tulemusena oli enne Stonewalli mässu kasutatud vastandus *homosexual* (ee *homoseksuaalne*) – *normal* (ee *normaalne*) peale mässu suuresti asendunud vastandusega *gay* (ee *gei*) – *hetero* (ee *hetero*). Ühtlasi järeldus uurimuse tulemustest, et Stonewalli mäss pani aluse seksuaalsuse nägemisele inimese identiteedi osana laiemalt kui ainult meditsiinidiskursuses. (*ibid*.)

1.1.4. LGBT+ sõnavara ajakohastamine

LGBT+ terminoloogia korrektsuse ja ajakohasuse küsimus on samuti esile tõusnud infoteaduse valdkonnas, kus põhiprobleemiks on marginaliseeritud ühiskonnagruppide liikmetele neile olulise informatsiooni kättesaadavaks tegemine. Selles valdkonnas on ilmnenud, et terminid, mida kasutatakse soo, rassi või seksuaalse orientatsiooni põhjal eristuvate kogukondade esindamiseks on tihti aegunud või sisaldavad endas valeinfot. Ühtlasi ei ole kasutatavad terminid sageli kogukonnaliikmete infovajaduse katmiseks piisavalt detailsed. (Campbell 2000: 122)

Eestis võib sarnasel teemal näiteks tuua transsooliste inimeste soolise üleminekuga seotud meditsiiniteenuseid reguleeriva määruse, mis võeti vastu aastal 1999. Nii määruses kui muudes soolise üleminekuga seotud avaldustes kasutatakse aegunud termineid nagu *soo muutmine*, *soovahetus*, *transseksualism* ja *transseksuaalsus*. (Soolise üleminekuga seotud...) Seesugune terminikasutus paneb transsoolised inimesed riigiga suheldes ebamugavasse positsiooni, kuna aegunud terminid on kogukonnaliikmete jaoks alandavad ning dehumaniseerivad.

Juba 2000. aastal tõi Lääne-Ontario Ülikooli infoteaduse professor Grant Campbell välja, et gei- ja lesbikogukondade üha suurem nähtavus on infoteadlaste hulgas tekitanud diskussiooni vananenud sõnavara ja info klassifikaatorite muutmise vajadusest

(Campbell 2000: 123). Ka Eestis on olnud mitmeid diskussioone eelkõige transsoolisusega seotud aegunud määruste ja sõnavara ajakohastamise ümber (Läbipaistmatu süsteem ja... 2021; Sõnastik: transfoobia, cis-seksism... 2019; Vikerkaar arstikabinetis 2018), kuid muutusi pole need veel toonud.

1.1.5. LGBT+ neologismid ja nende ülekanne inglise keelest

Kuigi suur osa LGBT+ kogukonda kirjeldavast sõnavarast on aja jooksul sama kuju säilitanud ning vaid oma tähendusvarjundit muutnud, on seoses soo ja seksuaalsuse mõistete avardumisega tekkinud vajadus ka uue, kõigi erinevate LGBT+ identiteetide kirjeldamiseks sobiva sõnavara järele. See vajadus pani aluse LGBT+ neologismide ehk uudissõnade tekkele. Inglise ja horvaatia keele näitel on LGBT+ neologismide teket ja kasutust uurinud Ivana Buntak (2020). Kuna inglise keel juhib hetkel maailmas soo ja seksuaalsusega seotud sõnavara muutusi, jõuavad neologismid teistesse keeltesse tihti inglise laensõnadena ning võetakse esmalt kasutusele LGBT+ kogukonna liikmete ja aktivistide hulgas. (*ibid*.)

Buntak käsitles oma uurimuses viit inglise LGBT+ neologismi ja nende horvaatiakeelseid vasteid eelkõige sõnamoodustuse ja seejärel sõnade kasutuse vaatepunktist. Käsitletud sõnad olid *pansexuality* (ee *panseksuaalsus*), *demisexual* (ee *demiseksuaalne*), *cisgender* (ee *cis-sooline*, *paiksooline*), *non-binary* (ee *mittebinaarne*) ja *transphobia* (ee *transfoobia*) ning nende horvaatiakeelsed vasted. Ta leidis, et inglise keeles moodustatakse neologisme eelkõige ees- või järelliidete lisamise kaudu, kusjuures liited pärinevad peamiselt kreeka ja ladina keelest. Horvaatia keeles esinevad neologismid olid laenatud inglise keelest ning mugandatud sihtkeelele kohasteks. Neologismide kasutuse osas leidis Buntak, et inglise keeles on sõnad levinud laiemalt kui horvaatia keeles. Inglise keeles kasutatakse neid kõigis keeleregistrites ning sõnad ei vaja erilisi lisaselgitusi, kuid horvaatia keeles kasutavad uudissõnu peamiselt LGBT+ kogukonna liikmed. (*ibid*.)

Sarnane inglise keelest mugandamise protsess leiab aset ka paljudes teistes, sealhulgas eesti keeles, kuhu eelmainitud sõnad on põhimõtteliselt analoogselt üle võetud. Eesti keeles kasutusel olevast LGBT+ sõnavarast annavad hea ülevaate LGBT+ ühingu ja Feministeeriumi sõnastikud (Sõnastik; Sõnastik 2023). Sõnavara kasutuskontekste ega

levikut pole eesti keeles veel uuritud, kuid on tõenäoline, et see on horvaatia keelega sarnases arengujärgus. Seetõttu on just LGBT+ aktivistidel oluline roll korrektse ja kaasava keelekasutuse arendamisel ja levitamisel.

1.1.6. LGBT+ sõnavara ja korpuslingvistika

Korpuslingvistika meetodite kasutamist ja sellega seotud kitsaskohti LGBT+ sõnavara analüüsimisel on uurinud John Paul Baker (2018) ja Heiko Motschenbacher (2022). Alates 2000. aastatest on korpuslingvistika meetodid seksuaalsuse uurimisel üha populaarsemaks muutunud ning seoses sellega on kerkinud esile vajadus hinnata, kuivõrd need soo ja seksuaalsusega seotud sõnavara uurimiseks sobivad. Motschenbacher (*ibid*.) on välja toonud viis korpuslingvistika metodoloogia aspekti, mida soo ja seksuaalsuse uurimisega seoses kritiseeritud on.

Korpuslingvistika esimese probleemina on välja toodud selle fookus kvantifitseerimisele ja objektistamisele. Seesugune lähenemine jätab tagaplaanile soo- ja seksuaalsusuuringutes olulise kontekstianalüüsi, mis mittenormatiivseid diskursuseid paremini mõista võimaldab. Teiseks tõusevad korpuslingvistikas esiplaanile korpustes sageli esinevad sõnad, mille tagajärjel võivad harvemini esinevad ja rohkem marginaliseeritud mustrid märkamatuks jääda. Ühtlasi toetub korpuslingvistika suuresti konkreetsetele sõnavormidele, tuvastades mustreid leksikaalsel tasandil. Teised soo- ja seksuaalsusuuringutes olulised keeletasandid nagu morfoloogia, süntaks, semantika ja pragmaatika jäävad korpuslingvistikas tagaplaanile. (*ibid*., lk 45–48)

Korpuslingvistilises analüüsis on kesksel kohal kategoriseerimine, mis on soo- ja seksuaalsuse kirjeldamisel problemaatiline, kuna kategooriad taanduvad tihti lihtsatele kaksikjaotustele nagu „mees – naine“ või „homoseksuaalne – heteroseksuaalne“. See süvendab kaksikjaotustest välja jäävate identiteetide stigmatiseerimist ja marginaliseerimist. Viimane kirjeldatud probleem on korpuslingvistika keskendumine tekstidevaheliste erinevuste leidmisele. Säärane lähenemisviis süvendab soo- ja seksuaalvähemuste eristamist normatiivsest ühiskonnast, kujutades neid millegi teistsuguse ja võõrana. (*ibid*., lk 50–52)

Kriitikast hoolimata on praeguseks leitud, et läbimõeldult rakendades võivad korpuslingvistika meetodid soo- ja seksuaalsusega seotud sõnavara uurimisel anda olulist infot eelkõige selle kohta, kuidas vähemusgruppe avalikus diskursuses kujutatakse. Selle info saamiseks on võimalik kasutada soo- ja seksuaalidentiteete kirjeldavate sõnade *n*-gramme ning sagedusi. Samuti on kasulik vaadata nii üksiksõnade kui *n*-grammide konkordantse, mis annavad sõnakasutusele laiema konteksti. (*ibid.*, lk 62–63) Ühtlasi võimaldab korpuslingvistika vaadelda spetsiifiliste sõnade koosesinemissagedusi ehk kollokatsioone, mida saab kasutada soo- ja seksuaalidentiteetide gruppide uurimiseks. (*ibid*., lk 67) Nende meetodite abil püüan ka oma töös Eesti LGBT+ kogukonna keelekasutusest ülevaate saada.

1.1.7. LGBT+ aktivism Eestis

Eesti organiseeritud LGBT+ aktivismi alguseks võib pidada 1990. aastate algust, kui loodi Eesti Lesbiliit (1990), Eesti Gayliit (1992) ja dekriminaliseeriti homoseksuaalsed suhted meeste vahel (1992). Katrin Tiidenberg ja Airi-Alina Allaste (2020) on uurinud, kuidas mõtestavad Eesti LGBT+ aktivistid oma aktivismi. Nad leidsid, et aktivistide arvates on Eesti LGBT+ kogukonna liikmete ühtsustunne nõrk. Selle põhjusena toodi välja sisemist homofoobiat, suhteliselt aktsepteeritavaid inimõigusi ja eestlastele loomuomast passiivsust. Aastatel 2007–2008 sai LGBT+ aktivism aga hoogu juurde. Tolleaegses ühiskonnas muutusid oma õiguste eest seismine ja poliitikute tegevuse kritiseerimine järjest populaarsemaks ning see andis tuge ka LGBT+ aktivismile. (*ibid.*)

Praegu on Eestis LGBT+ aktivismi suurim eestvedaja ja organiseerija Eesti LGBT Ühing, mis asutati aastal 2008. Asutamisel seati ühingu põhieesmärgiks „suurendada LGBT-noorte ühiskondlikku aktiivsust ja suurendada ühiskonna teadlikkust LGBT-teemadest“. (Eesti LGBT Ühing 2008) LGBT ühingu 2022–2024 aasta strateegiasse kuuluvad muuhulgas abieluvõrdsuse ja uue soo tunnustamise regulatsiooni vastuvõtmine ning seksuaal- ja soovähemuste osas teadlikkuse parandamine ja diskrimineerimise vähendamine (Eesti LGBT Ühing).

Eesti LGBT Ühing pole aga ainus LGBT+ kogukonda kaasav ja toetav organisatsioon. Näiteks teeb ühing koostööd Eesti Inimõiguste Keskuse, Geikristlaste Kogu, Tallinn

Bearty ja LGBT+ filmifestivali Festheartiga (Koostöö). Lisaks eelmainitud organisatsioonidele kasvas Lilla Agenda taskuhäälingust 2022. aasta lõpus välja Eesti Transinimeste ühing. Samuti tegutseb Tartus projekti „Peemoti Raamatud ehk Tartu LGBT+ kogukonna isemajandava keskuse loomine koosloome ning teenusedisaini meetodite abil“ raames LGBT+ kogukonnakeskus (Projektist 2022).

1.2. Eesti keele automaattöötluse vahendid

1.2.1. Automaatne kõnetuvastus

Minu töö materjal pärineb kahest LGBT+ aktivistide taskuhäälingust, Lillast Agendast ja Homokringlist, milles püüan tuvastada aktivistide kasutatavat LGBT+ kogukonnaga seotud sõnavara. Taskuhäälingutes kasutatava sõnavara väljaselgitamiseks ja analüüsimiseks on vaja see esmalt kirjalikule kujule viia. Selleks kasutan eesti keele automaatset kõnetuvastust. Siinses peatükis annan ülevaate kõnetuvastussüsteemide arengust ja eesti keele kõnetuvastuse tööpõhimõtetest.

Inimeste käitumist jäljendavate masinate arendamine on teadlastele huvi pakkunud juba mitu sajandit. Kuna algselt oli arendustöö peamine fookus nn intelligentsete masinate loomine, keskenduti esmajärjekorras automaatsele kõnesünteesile ning alles seejärel kõnetuvastusele. Esimesed katsed inimkõnet automaatselt luua pärinevad oletatavasti 18. sajandi teisest poolest ning automaatselt kõnet tuvastada 20. sajandi keskpaigast. Algsed kõnetuvastussüsteemid toetusid akustilisele foneetikale, mis kirjeldab kõne foneetilisi elemente ja selgitab, kuidas neid kõnes realiseeritakse. (Juang, Rabiner 2005)

Esimesed 1960. aastatel loodud kõnetuvastussüsteemid suutsid nende akustilis-foneetiliste tunnuste põhjal tuvastada ligikaudu 10–100 erinevat sõna. 1970. aastateks tuvastasid süsteemid sõnades esinevate mustrite abil juba 100–1000 sõna. 1980ndatel hakati statistilisi meetodeid kasutades katsetama üle 1000 sõna tuvastamist. Sellel ajal võeti laiemalt kasutusse peidetud Markovi mudelid ja tehisnärvivõrgud, mis võimaldasid suulise teksti varieeruvust paremini hallata. 1990. aastatel keskenduti suurema sõnavaraga keelemudelite loomisele ning statistiliste treeningmeetodite kasutusele

võtmisele. Edasine areng on peamiselt toetunud keelemudelite laiendamisele ja masinõppe rakendamisele. (*ibid*.)

1.2.2. Eesti keele automaatse kõnetuvastuse loomine

Eesti keele automaatse kõnetuvastuse esimene prototüüp valmis Tanel Alumäe 2002. aastal kaitstud magistritöö osana (CV: Tanel Alumäe; Alumäe 2003: 34). 2003. aastal ilmunud artiklis „Eestikeelse kõne tuvastus: prototüübi loomine“ kirjeldatakse väikse kuni keskmise suurusega sõnavara (2–1000 sõna) tuvastamiseks mõeldud prototüübi arendamist ja selle peamisi komponente. Prototüübi arendamisel ei keskendutud veel valmis kõnetuvastussüsteemi loomisele vaid pigem süsteemi realiseerimiseks vajalike komponentide täiustamisele. (*ibid.*, lk 34–35)

Tüüpiline kõnetuvastussüsteem järgib kõne teksti kujule viimisel kindlat malli. Sisendkõnest väljundteksti saamiseks digitaliseerib kõnetuvastussüsteem esmalt sisendi ning jaotab selle võrdse pikkusega segmentideks. Seejärel arvutatakse iga segmendi tunnusvektor ehk segmendis oleva info kompaktne kujutis. Viimaks dekodeeritakse vektorite jada akustiliste mudelite ja keelemudeli abil tekstiks. Dekodeerimisel kasutatakse valikuliselt ka keele morfoloogial ja süntaksil põhinevat otsingut. (*ibid*., lk 36)

Eesti keele kõnetuvastuse esialgses prototüübis kasutati akustika modelleerimiseks peidetud Markovi mudeleid, mis võimaldasid arvutada sõnadele ja lausetele vastavaid akustilisi tõenäosusi. Selleks, et võtta arvesse häälikute akustika sõltuvust nende naabrusest, kasutati prototüübis kontekstist sõltuvaid häälikumudeleid. Häälikumudelite treenimise aluseks oli eesti keele foneetiline andmebaas, mis sisaldas 12 tundi kõnematerjali. Prototüübi treenimiseks kasutati andmebaasis olevaid lausungeid ja tekstilõike. (*ibid.*, lk 38–41)

Kõnetuvastuse keelmudel sisaldab endas sisendkeele grammatikat ja sõnastikku. Grammatikat kasutatakse tuvastatavate sõnajadade defineerimiseks ning sõnastikku selleks, et näidata, millistest põhiühikutest (siin foneemidest) sõnad koosnevad. Eesti keele puhul oli prototüübi üks suurimaid probleeme hea keelemudeli koostamine, kuna

keeles on erinevate sõnavormide arv on väga suur. Prototüübi loomisel oletati, et paremaid tulemusi võiks anda morfeemidel või teistel sõnaosadel põhinev grammatika ja sõnastik. (*ibid*., lk 46)

1.2.3. Eesti keele automaatse kõnetuvastuse hetkeseis

Eesti keele kõnetuvastussüsteemi viimaseid uuendusi on kirjeldatud artiklis „Estonian Speech Recognition and Transcription Editing Service“ (Olev, Alumäe 2022). Praegune süsteem on disainitud poolspontaanse kõne, sealhulgas vestlussaadete, loengusalvestiste ja intervjuude tuvastamiseks. Lisaks kõne tekstiks transkribeerimisele tuvastab süsteem kõnelemiskeele, et vältida võõrkeelsete fraaside eesti keele mudeli järgi transkribeerimist, lisab kirjavahemärgistuse, normaliseerib teksti ja tuvastab kõnelejaid, kui tegemist on tuntud avaliku elu tegelastega. Kõnetuvastus on avalikkusele kättesaadav nii veebiliidese kui avatud lähtekoodiga tarkvara kujul. (*ibid*.)

Kõnetuvastussüsteemide viimaste uuenduste hulka kuuluvad tehisnärvivõrkude ja süvaõppemeetodite rakendamine nii akustiliste kui keelemudelite treenimisel ning treeningandmete mahu suurendamine. Praeguse eesti keele kõnetuvastussüsteemi akustilise mudeli treeningandmed pärinevad enamjaolt vestlussaadetest. Lisaks vestlussaadetele kuulub akustilise mudeli treeningandmete hulka Eesti keele spontaanse kõne foneetilise korpuse (Lippus jt 2021) andmeid, Seenioride kõnekorpuse (Meister 2021) andmeid ning kõnede, loengute ja parlamendikõnede salvestisi. Keelemudeli treeningandmetest moodustab suurima osa 2019. aasta Eesti keele ühendkorpuse (Koppel, Kallas 2020) veebilausete allkorpus. (Olev, Alumäe 2022)

Automaatselt tuvastatud kõne ortograafilise korrektsuse saavutamisel on kõige olulisem roll keelemudelil. Kuigi keelemudeli treeningmeetodid ja -andmed on viimastel aastatel kiiresti arenenud, leidub poolspontaanses kõnes paratamatult sõnu, mida keelemudeli sõnastikus ei esine. Nende sõnastikuväliste sõnade oletuslikuks tuvastamiseks kasutatakse praeguses kõnetuvastussüsteemis spetsiaalset foneemipõhist lähenemist, kus sõnu esitatakse esmalt kõigi keelemudelis esinevate sõnade häälduste põhineva foneemimudelina. Seejärel rekonstrueeritakse sõnad väljundis neile kõige tõenäolisemalt

vastava foneemijärjendina. Seesugune lähenemine parandab enamikul juhtudel sõnastikuväliste sõnade tuvastamise kvaliteeti ja muudab väljundi kognitiivselt lihtsamini mõistetavaks. (Alumäe, Tilk, Asadullah 2019) Kuna suur osa siinses töös uuritavast LGBT+ kogukonnaga seotud sõnavarast on uudissõnavara, mida tõenäoliselt keelemudeli treeningandmetes ei esine, on kõnetuvastussüsteemi võimekus sõnastikuväliseid sõnu tuvastada töö kontekstis väga tähtis.

1.2.4. Automaatse kõnetuvastuse hindamine

Automaatse kõnetuvastuse kvaliteedi hindamise olulisust on varasemalt välja toodud poliitikatekstide näitel. Kuigi palju poliitilisi tekste on transkribeeritud kujul kättesaadavad, jääb suur osa neist transkriptsioonide puudumise tõttu teaduslikult analüüsimata. Transkribeerimata tekstide hulka kuuluvad näiteks poliitilised vestlussaated ning poliitikute raadio- ja teleintervjuud. Kuna seesugust materjali on väga palju, oleks selle käsitsi transkribeerimine mõttetult kulukas ja aeganõudev. Materjali transkribeerimisel võiks potentsiaalselt kasu olla automaatsetest kõnetuvastussüsteemidest, kuid automaatsete transkriptsioonide kasutamiseks tuleks nende kvaliteeti esmalt hinnata. (Proksch, Wratil, Wäckerle 2019: 339–340)

Automaatsete transkriptsioonide hindamisel on peamine fookus nende täpsusel. Transkriptsiooni täpsust hinnatakse enamasti sõnavigade määra (ingl *word error rate*) alusel. Sõnavigade määra arvutamiseks võrreldakse omavahel automaatse kõnetuvastuse koostatud oletuslikku teksti ja süsteemile sisendiks antud referentsteksti. Vigade määra arvutamise valem on järgmine:

$$WER = \frac{S + D + I}{N}$$

Valemis tähistab N referentsteksti sõnade arvu, S oletuslikus tekstis asendatud sõnade arvu ehk sõnu, mis olid valesti tuvastatud, D oletuslikus tekstis kustutatud sõnade arvu ehk sõnu, mis olid automaatses transkriptsioonis puudu ning I oletuslikku teksti lisatud sõnade arvu ehk sõnu, mis puudusid referentstekstist. (*ibid*., lk 340) Eesti keele automaatse kõnetuvastuse sõnavigade määr teleuudiste transkribeerimisel on ligikaudu

8,5%, vestlussaadete transkribeerimisel ligikaudu 13,4% ning pressikonverentside transkribeerimisel ligikaudu 8,1% (Alumäe jt 2023: 495).

1.2.5. Morfoloogiline analüüs ja lemmatiseerimine

Eesti keele rikkaliku morfoloogia tõttu on tekstide automaattöötluse ja -analüüsi üks esimesi etappe morfoloogiline analüüs. Morfoloogiline analüsaator võtab sisendiks tekstis esineva sõnavormi ja väljastab selle märksõnavormi ehk lemma (nimisõna puhul ainsuse nimetav kääne ja tegusõna puhul *ma*-tegevusnimi) ning sellele liidetud grammatilised morfeemid (Viks 2000: 11–13). Esmalt analüüsitakse sõnavorme sõnastiku põhjal, kasutades erinevate morfeemide loendeid ja nende kombineerimise eeskirju. Kuna sõnastikud ei sisalda aga kõiki tekstides esinevaid sõnu, kasutatakse tundmatute sõnade analüüsimiseks reeglipõhist morfoloogiat, mis määrab sõnadele analüüsid grammatikareeglite järgi. (Muischnek jt 2012: 69–70)

Eesti keele automaattöötluseks mõeldud teekide kogu EstNLTK[3] koosseisu kuulub vabavaraline morfoloogiline analüsaator Vabamorf[4]. Vabamorfi analüüsiprotsess koosneb kahest peamisest osast, morfoloogilisest analüüsist ja ühestamisest, kuid võimaldab neile sätestada erinevaid parameetreid, mida on täpsemalt kirjeldatud EstNLTK morfoloogilise analüsaatori kasutusjuhendis[5]. Morfoloogilise analüüsi käigus lisatakse sõnavormidele kõik võimalikud analüüsivariandid ja sellele järgneval ühestamisel proovitakse lausekonteksti põhjal analüüsivariantidest õige variant või variandid välja selgitada. Seetõttu eeldab morfoloogiline analüüs ja ühestamine, et tekstis on lausepiirid eelnevalt märgendatud. (Kaalep, Vaino 2001) Eesti keele ajakirjandus-, ilukirjandus-, teadus- ja veebitekstide korpuste põhjal jäi 2022. aasta seisuga morfoloogilisel analüüsil õige lemma analüüsita 0,87% sõnadest (Saul 2022: 28).

[3] https://github.com/estnltk/estnltk
[4] https://github.com/Filosoft/vabamorf
[5] estnltk/01_morphological_analysis.ipynb at main · estnltk/estnltk · GitHub

2. Andmed ja meetod

2.1. Andmed

Oma bakalaureusetöös kasutan andmetena kahe Eesti LGBT+ aktivistide taskuhäälingu, Lilla Agenda ja Homokringli episoode ajavahemikust 28.04.2020–05.10.2022. Lilla Agenda autorid on LGBT+ aktivistid ning Eesti Transinimeste Ühingu asutajad Mel Zelmin ja Paul Vahtra ning Homokringli autorid LGBT+ aktivistid Anette Mäletjärv ja Heinrich Sepp, keda teatakse ka *drag*-artist Helgi Saldona. Lisaks taskuhäälingute autoritele astuvad neis üles erinevad LGBT+ kogukonda kuuluvad külalised.

Kõnetuvastus tehti kahe taskuhäälingu peale 31st episoodist. Neist episoodidest 17 kuulusid Lillasse Agendasse ning 14 Homokringlisse. 31 episoodi ajaline kogumaht oli ligikaudu 29 tundi. Kõnetuvastuse väljundiks olid iga analüüsitud faili transkriptsioonid CTM, SRT, TRS ja JSON formaadis. Edasises töös võtsin aluseks TRS ja JSON formaadis failid, mis sisaldasid kokku umbes 60 000 sõne transkriptsiooni.

2.2. Töö etapid

Kuna siinne töö koosnes paljudest omavahel kronoloogiliselt seotud etappidest, on selles peatükis välja toodud etappide üldine järjestus. Töö esimene etapp oli taskuhäälingute transkribeerimine automaatse kõnetuvastuse abil. Peale transkriptsioonifailide kätte saamist valisin mõlemast taskuhäälingust välja viis faili, mille transkriptsioonid käsitsi üle kontrollida. Kontrollitud failidest leidsin LGBT+ kogukonda puudutava sõnavara levinuimad tuvastusvead ja kasutasin neid ülejäänud failides esinevate LGBT+ sõnavara tuvastusvigade parandamiseks. Seejärel lemmatiseerisin kõik parandatud failid. Lemmatiseeritud failide põhjal koostasin nii üldised kui LGBT+ sõnavara sagedusloendid, bi- ja trigrammid ning kahe ja kolme sõna kollokatsioonid. Transkriptsioonide kontrollimiseks kasutasin foneetikatarkvara Praat (Boersma, Weenink 2023), andmete töötlemiseks ja analüüsiks rakendustarkvara R versiooni 4.2.2 (R Core Team 2023) ning morfoloogilise analüüsi rakendamiseks programmeerimiskeele Python

versiooni 3.10 (Van Rossum, Drake 2009). Kõigis etappides kasutatud meetodite kirjeldused, etappide tulemused ja nende analüüs esitatud järgmistes peatükkides. Töö käigus koostatud ja kasutatud kood on leitav GitHubis lingil https://github.com/leenakt/LGBT_sonavara.

3. Analüüs ja tulemused

3.1. Tuvastusvigade arv, veatüübid ja sõnavigade määr

Kõnetuvastuse kvaliteedi hindamiseks kontrollisin käsitsi üle kummastki taskuhäälingust viie, kokku kümne episoodi transkriptsioonid. Episoodid valisin Lilla Agenda puhul mugavusvalimi ja Homokringli puhul juhuvalimi alusel. Seda põhjusel, et erinevalt Homokringlist on Lilla Agenda episoodide pealkirjades kirjas, mis teemadel käesolevas episoodis räägitakse. Nii sain välja valida episoodid, mis minu töö kontekstis kõige informatiivsemad oleksid. Kuna Homokringlis episoodide pealkirju sellisel kujul ei ole, valisin analüüsitavad episoodid juhuslikult. Valitud episoodide pealkirjad, kuupäevad, kõnelejad ning pikkused on näha tabelis 1.

Tabel 1. Kontrollimiseks valitud episoodid.

Taskuhääling	**Episoodi pealkiri**	**Kuupäev**	**Kõnelejad**	**Sõnade arv**	**Episoodi pikkus**
Lilla Agenda	Tulla või mitte tulla (kapist välja) – see on küsimus!	28.04.2020	Mel Zelmin, Paul Vahtra	2857	00:22:38
Lilla Agenda	Miks me marsime?	27.06.2020	Mel Zelmin, Paul Vahtra, Eva Marta Sokk	4550	00:37:27
Lilla Agenda	Trans mees, trans naine ja mittebinaarne inimene astuvad baari…	20.01.2021	Sara Arumetsa, Mel Zelmin, Paul Vahtra	6456	00:48:33
Lilla Agenda	Mõnda kanni sa tahad, mõni on kunstiteos ehk aseksuaalsuse ABC	03.03.2021	Mel Zelmin, Paul Vahtra	4267	00:34:13
Lilla Agenda	Intiimse sõbratari otsinguil ehk kväär ajaloo eri II	21.07.2021	Mel Zelmin, Paul Vahtra, Taavi Koppel	5818	00:44:11
Homokringel	-	07.07.2022	Anette Mäletjärv, Helgi Saldo (Heinrich	5881	00:58:18

			Sepp), Rene Köster		
Homokringel	-	09.06.2022	Anette Mäletjärv, Helgi Saldo (Heinrich Sepp), Emma Kuppart	5424	00:58:50
Homokringel	-	02.09.2021	Anette Mäletjärv, Helgi Saldo (Heinrich Sepp), Krõõt Juurak	7757	00:58:43
Homokringel	-	05.08.2021	Mia Moore, Sebastian Freiberg, Anette Mäletjärv, Helgi Saldo (Heinrich Sepp)	9238	01:00:17
Homokringel	-	08.07.2021	Ann Vaida, Barbara Oja, Anette Mäletjärv, Helgi Saldo (Heinrich Sepp)	8204	00:58:21

Transkriptsioonide kontrollimiseks kasutasin foneetikatarkvara Praat, mis võimaldab heli- ja tekstifailidega paralleelselt töötada. Kuna Praatis analüüsitavad tekstifailid peavad olema TextGrid formaadis, teisendasin esmalt valitud episoodide TRS formaadis väljundid Oregoni ülikooli veebipõhise tööriista[6] abil TextGrid formaati. Peale failide teisendamist laadisin alla valitud episoodide audiofailid. Kuna varasemat infot eesti keele kõnetuvastuse kontrollimise kohta ei õnnestunud mul leida, töötasin transkriptsioonide kontrollimiseks ise välja süsteemi, milles liigitasin taskuhäälingutes esinevad vead 15 kategooriasse:

1. **Voorus vale kõneleja** – ühe kõneleja vooru vahele on tuvastatud teise kõneleja öeldu; ka vead, kus ühe kõneleja tekst on jäänud tuvastamata, kuid see oleks teise kõneleja vooru keskel.

[6] http://lingtools.uoregon.edu/tools/trans_to_praat.php

2. **Voor ülearu** – transkriptsioonis on erinevaid voore tuvastatud rohkem, kui konkreetses episoodis kõnelejaid on.
3. **Ingliskeelne lause/lauseosa puudu / vigaselt tuvastatud –** kahest või rohkemast ingliskeelsest sõnast koosnevad fraasid, mis on transkriptsioonist puudu või valesti tuvastatud.
4. **Eestikeelne lause/lauseosa puudu / vigaselt tuvastatud –** kahest või rohkemast eestikeelsest sõnast koosnevad fraasid, mis on transkriptsioonist puudu või valesti tuvastatud.
5. **Terve sõna puudu –** terve eesti- või ingliskeelne sõna, mis on transkriptsioonist puudu.
6. **Poolik sõna puudu –** kõneleja pooleli jäetud sõna, mis on transkriptsioonist puudu.
7. **Poolik sõna valesti tuvastatud –** kõneleja pooleli jäetud sõna, mis on valesti tuvastatud.
8. **Sõna üle –** transkriptsioonis esineb sõna, mida kõnelejad öelnud pole.
9. **Nime ortograafiaviga –** nimi, mis on valesti tuvastatud; kuna kõnetuvastusel on teatud võimekus nimesid tuvastada, on siin arvestatud ka esisuurtähevigu.
10. **Eestikeelse sõna ortograafiaviga –** eestikeelne sõna tuvastatud valesti viisil, mida ei saa pidada ühekski eesti keeles olemasolevaks sõnaks.
11. **Ingliskeelse sõna ortograafiaviga –** ingliskeelne sõna tuvastatud valesti viisil, mida ei saa pidada ühekski inglise keeles olemasolevaks sõnaks, sh ingliskeelsed sõnad, mis on tuvastatud eestikeelsete sõnadena.
12. **Vale eestikeelne sõna –** korrektse eestikeelse sõna asemel tuvastatud mõni teine eesti keeles olemasolev sõna.
13. **Vale ingliskeelne sõna –** korrektse ingliskeelse sõna asemel tuvastatud mõni teine inglise keeles olemasolev sõna.
14. **Muu võõrkeelne sõna valesti tuvastatud –** valesti tuvastatud sõnad mõnes muus keeles peale eesti ja inglise keele.
15. **Akronüüm valesti tuvastatud –** valesti tuvastatud lühendid; siia kuuluvad igat tüüpi vead, mis lühendite tuvastamisel tehtud on, sh tähevead, tuvastamine mõne muu sõnana, ainult osaliselt õigesti tuvastamine jne.

Kategooriad kujunesid välja esimese episoodi kontrollimise käigus, kus vaatasin jooksvalt, milliseid veatüüpe transkriptsioonis esineb. Kui olin esimese episoodi põhjal suurema osa kategooriaid välja töötanud, hakkasin episoode järjest kuulama ja vigu paralleelselt nii TextGrid failis parandama kui Exceli tabelisse üles märkima. TextGrid failides tegin iga kõneleja vooru jaoks uue duplikaatvooru, kus märkisin tehtud parandused kantsulgude vahele. Töö tulemuseks olid parandatud TextGrid failid ning koondtabel kõigi kontrollitud episoodides esinenud vigadega, mis on saadaval GitHubis[7]. Vigade hulk ja tüübid taskuhäälingute lõikes on näha tabelis 2.

Tabel 2. Kõnetuvastuse veatüübid ja vigade arvud taskuhäälingute lõikes.

	Lilla Agenda	**Homokringel**	**Kokku**
Voorus vale kõneleja	780	1337	2117
Voor ülearu	45	5	50
Ingliskeelne lause/lauseosa puudu/vigaselt tuvastatud	37	438	475
Eestikeelne lause/lauseosa puudu/vigaselt tuvastatud	383	1192	1575
Terve sõna puudu	843	1479	2322
Poolik sõna puudu	276	435	711
Poolik sõna valesti tuvastatud	35	45	80
Sõna üle	24	48	72
Nime ortograafiaviga	62	224	286
Eestikeelse sõna ortograafiaviga	214	287	501
Ingliskeelse sõna ortograafiaviga	128	490	618
Vale eestikeelne sõna/sõnavorm	475	846	1321
Vale ingliskeelne sõna/sõnavorm	17	56	73
Muu võõrkeelne sõna valesti tuvastatud	2	3	5
Akronüüm valesti tuvastatud	39	57	96
Kokku	3360	6942	10302

[7] https://github.com/leenakt/LGBT_sonavara/blob/main/Failide eelt%C3%B6%C3%B6tlus/Vigade koondtabel.xlsx

Kokku märkisin kõigi kümne episoodi peale üles 10 302 viga, neist 3360 Lillas Agendas ja 6942 Homokringlis. Saadud vigade arv ei pruugi olla täiesti täpne, kuna mitmeid vigu oli keeruline ühe kindla tüübi alla kategoriseerida. Kindlalt mitmesse tüüpi kuuluvad vead, nt vale kõneleja eestikeelse sõna valetuvastus läksid vigadena arvesse mõlemas kategoorias. Siiski võib vigade märkimises ja kategoriseerimises esineda väheldast ebaühtlust. Kõigi episoodide lõikes oli kõige levinum veatüüp terve sõna puudumine transkriptsioonist. Sellele järgnesid voorus vale kõneleja tuvastamine ning eestikeelse fraasi või lauseosaga eksimine.

Lisaks sagedasimatele veatüüpidele sain transkriptsioonide kontrollimise käigus aimu sellest, millised kohad automaatsele kõnetuvastusele enim probleeme valmistavad. Suurim probleem on minu hinnangul ebaselge hääldus, sealhulgas aktsent või erinevate sõnade sarnaselt hääldamine. Sarnasest häälduses tuleneb suure tõenäosusega ka üks levinumaid veatüüpe, vale eestikeelse sõna tuvastamine. Loomuliku kõne puhul on tavaks hääldada sõnu kiiresti, jättes sõnalõpud tihti selgelt välja hääldamata. Seetõttu ongi mitmes kohas korrektse sõna asemel tuvastatud vale, enamasti lühem sarnaselt kõlav sõna. Mõned näited valesti tuvastatud sõnadest on näha joonisel 1, kus vasakpoolses sõnade tulbas on kujutatud õigeid sõnu ja parempoolses nende tuvastusvigu.

```
"3" "05082021_result.TextGrid" 1 "vahvaid" "vaid"
"553" "05082021_result.TextGrid" 14 "veits" "võti"
"628" "05082021_result.TextGrid" 17 "kellelegi" "kellelgi"
"129" "ajalooeri_result.TextGrid" 11 "habet" "habe"
"275" "ajalooeri_result.TextGrid" 18 "tagasi" "tagas"
"152" "transmbi_result.TextGrid" 4 "avalikkuses" "avalikkus"
"154" "transmbi_result.TextGrid" 4 "meedias" "meeldis"
"769" "09062022_result.TextGrid" 42 "üldse" "sa"
"807" "09062022_result.TextGrid" 44 "naeruga" "arvu"
```

Joonis 1. Tuvastusvead transkriptsioonides.

Teine levinud probleem oli kõnelejate sarnane hääletoon, mis põhjustas voorus vale kõneleja tuvastamist. Kuna see viga pole minu töö seisukohast aga kuigi oluline, otsustasin selle edasisest analüüsist välja jätta. Järgmine probleemkoht, mis tihti esile kerkis, oli läbi naeru rääkimine, mis on samuti loomuliku kõne puhul üsna tavapärane. Mulle tuli üllatusena, et isegi kui inimkõrvaga on läbi naeru öeldud tekst täiesti selgelt

tuvastatav, on kõnetuvastus sellistel juhtudel tihti pikemaid lausejuppe lihtsalt tuvastamata jätnud.

Vigade koondtabelist järeldasin, et levinuimad vead olid mõlema taskuhäälingu ning kõigi episoodide lõikes enamjaolt samadest tüüpidest. Väiksed erinevused tulenesid kõnelejate erinevusest ning konkreetse taskuhäälingu tavadest või episoodi teemast ja pikkusest. Näiteks on tabelist näha, et Homokringli episoodides on eranditult rohkem vigu põhjusel, et Homokringli episoodid on keskmiselt Lilla Agenda episoodidest ligikaudu 15 minutit pikemad. Samuti on Homokringlis märgatavalt suurem osakaal ingliskeelsete sõnade puhul tehtud vigu, kuna Lilla Agenda saatejuhid kasutavad süsteemselt palju vähem inglise keelt.

Taskuhäälingute sõnavigade määra leidmiseks kasutasin rakendustarkvara R funktsiooni *wersim* (Proksch, Wratil, Wäckerle 2019) ja failide parandamise tulemusena saadud TextGrid faile. *Wersim* funktsioon võtab sisendiks oletusliku korpuse ehk siinses kontekstis kõnetuvastuse väljundi ning referentskorpuse ehk parandatud teksti. Seejärel joondab funktsioon kaks korpust sõnade alusel, leiab tekstis tehtud sõnaasenduste ja lisatud ning kustutatud sõnade arvud ning arvutab nende põhjal sõnavigade määra. (*ibid.*) Joonisel 2 on näha osa joondatud failist, kus esimeses sõnade tulbas on parandatud ja teises parandamata faili sisu. Kõik joondatud failid on leitavad GitHubis[8]. Funktsiooni abil leidsin keskmise sõnavigade määra nii kahe taskuhäälingu koondarvestuses kui kummagi taskuhäälingu kohta eraldi. Kogu selleks kasutatud kood on leitav siin[9].

8 https://github.com/leenakt/LGBT_sonavara/tree/main/Joondus%20ja%20s%C3%B5navigade%20m%C3%A4%C3%A4r/Joondused

9 https://github.com/leenakt/LGBT_sonavara/blob/main/Joondus%20ja%20s%C3%B5navigade%20m%C3%A4%C3%A4r/sonavigade_maar.rmd

```
348  "347" "kapistvalja_result.TextGrid" 16 "aga" "aga"
349  "348" "kapistvalja_result.TextGrid" 16 "nad" "nad"
350  "349" "kapistvalja_result.TextGrid" 16 "panevad" "panevad"
351  "350" "kapistvalja_result.TextGrid" 16 "selle" "selle"
352  "351" "kapistvalja_result.TextGrid" 16 "kahtluse" "kahtluse"
353  "352" "kapistvalja_result.TextGrid" 16 "alla" "alla"
354  "353" "kapistvalja_result.TextGrid" 16 "jaa" "####"
355  "354" "kapistvalja_result.TextGrid" 16 "ja" "ja"
356  "355" "kapistvalja_result.TextGrid" 16 "ja" "ja"
357  "356" "kapistvalja_result.TextGrid" 16 "see" "see"
358  "357" "kapistvalja_result.TextGrid" 16 "on" "on"
359  "358" "kapistvalja_result.TextGrid" 16 "raske" "raske"
360  "359" "kapistvalja_result.TextGrid" 17 "jaa" "ja"
361  "360" "kapistvalja_result.TextGrid" 17 "sa" "sa"
362  "361" "kapistvalja_result.TextGrid" 17 "paned" "paned"
363  "362" "kapistvalja_result.TextGrid" 17 "mis" "####"
364  "363" "kapistvalja_result.TextGrid" 18 "ennast" "ennast"
365  "364" "kapistvalja_result.TextGrid" 18 "väga" "väga"
366  "365" "kapistvalja_result.TextGrid" 18 "nagu" "nagu"
367  "366" "kapistvalja_result.TextGrid" 18 "haavatavasse" "haavatavasse"
368  "367" "kapistvalja_result.TextGrid" 18 "olukorda" "olukorda"
369  "368" "kapistvalja_result.TextGrid" 18 "sellega" "sellega"
```

Joonis 2. Lilla Agenda episoodi „Tulla või mitte tulla (kapist välja) – see on küsimus!“ joonduse näide.

Kahe taskuhäälingu keskmine sõnavigade määr oli 23,7%. Vigade tabeli põhjal oletasin, et Lillas Agendas oli vähem vigu peamiselt seetõttu, et kontrollitud episoodid olid lühemad. Sõnavigade määr näitas aga samuti Lilla Agenda paremat tuvastuskvaliteeti. Lilla Agenda keskmine sõnavigade määr oli 16,3% ning Homokringli oma 12% suurem ehk 28,3%. Lilla Agenda parem tuvastuskvaliteet tuleneb tõenäoliselt autorite ja külaliste aeglasemast kõnetempost, vähesemast inglise keele kasutusest ning selgemast häädusest.

Taskuhäälingute sõnavigade määrad olid oluliselt suuremad varem dokumenteeritud sõnavigade määradest. Selle põhjuseks võib pidada andmeformaatide erinevust. Nimelt on töös käsitletud taskuhäälingute näol tegemist eelnevalt kontrollitud andmetest tunduvalt spontaansema kõnega, mida iseloomustab kiirem kõnetempo, sagedasem teistest kõnelejatest üle rääkimine ning suurem ingliskeelsete sõnade hulk kõnes. Kõik need faktorid raskendavad automaatse kõnetuvastuse tööd. Siiski ei ole töös leitud sõnavigade määr kõnetuvastuse kvaliteedi hindamiseks ja varasemate tulemustega võrdlemiseks täiesti pädev alus, kuna selle arvutamise aluseks olnud sõnapõhine joondus ei pidanud sajaprotsendiliselt paika. Joondus eksis kõige rohkem kohtades, kus parandatud teksti oli lisatud või sellest kustutatud palju järjestikuseid sõnu. Eksliku

joonduse tõttu tõusis ka leitud sõnavigade määr, kuid probleemi siinse töö raames lahendada ei õnnestunud.

3.2. LGBT+ sõnavara tuvastusvead

Kõnetuvastuse väljundis pöörasin erilist tähelepanu LGBT+ kogukonnaga seotud sõnavara tuvastuskvaliteedile, kuna see mõjutab otseselt töö tulemusi. Töös vaadeldavat LGBT+ kogukonnaga seotud sõnavara esines parandatud failidest kõige rohkem mõlema taskuhäälingu *pride*-kuule keskenduvates episoodides: Lilla Agenda „Miks me marsime“ ja Homokringli 09.06.2021 episoodis. Nendes episoodides oli ühtlasi suurim hulk lühendite tuvastamisel tehtud vigu, millest enamuse moodustasid LGBT+ kogukonnaga seotud lühendite vead.

LGBT+ sõnavara tuvastamisel tehtud vigade täpsemaks uurimiseks joondasin parandatud ja parandamata failid sõnade järgi. Selleks kasutasin osa sõnavigade määra leidmise funktsiooni *wersim* koodist. Kuna sõnavigade määra leidmiseks peab funktsioon sõnad esmalt joondama, võtsin koodist selle osa välja ja salvestasin selle väljundi iga parandatud ja parandamata failide paari kohta eraldi tekstifaili.

Joonduse tegemisel võtsin referentstekstiks parandatud faili ja oletuslikuks tekstiks parandamata faili. Parandamata failidest leidsin esmalt kõnelejad ning seejärel puhastasin failid kirjavahemärkidest, topelttühikutest ja tühjadest ridadest (NA). Parandatud failidest eemaldasin samuti kirjavahemärgid, topelttühikud ja tühjad read ning lisaks sellele ka teksti juurde lisatud kõnelejanimed. Joonduse salvestamiseks lõin tühja andmetabeli, kuhu lisasin tulbad rea numbri ning referentsfaili ehk parandatud faili ja algse faili väljundi jaoks. Kui parandatud või parandamata failis mõne sõna puhul väärtus puudus, märkisin sellele kohale vastavas tulbas „####“. Selleks kasutatud kood on leitav siin[10]. Seejärel otsisin failidest manuaalselt LGBT+ sõnu sisaldavaid ridu ja märkisin eraldi

[10] https://github.com/leenakt/LGBT_sonavara/blob/main/Joondus%20ja%20s%C3%B5navigade%20m%C3%A4%C3%A4r/sonade_joondus.Rmd

dokumenti kõik erinevad valetuvastuste variandid. LGBT+ sõnade kindlakstegemisel toetusin Eesti LGBT Ühingu ja Feministeeriumi sõnastikele ning isiklikule kogemusele.

LGBT+ sõnadest esines kõige rohkem tuvastusvigu sageli kasutatud sõnades *drag* (15), *LGBT+* (23), *LGBT* (33), *pride* (20) ja *transsooline* (11). Üllatavalt hästi oli tuvastatud sõnu *trans*, *gei*, *lesbi* ja *homo*, mis võib olla seotud nende laialdasema kasutusega kõnetuvastuse treeningandmetes. Joonisel 3 on kujutatud osa tuvastusvigade loendist, terve loend on olemas GitHubis[11]. Vigade üles märkimisel eraldasin need püstkriipsudega, et loendit hiljem hõlpsalt regulaaravaldiste koostamisel kasutada. Valetuvastused, mis langesid osaliselt või täielikult kokku mõne eesti keeles esineva sõnavormiga, märkisin loendis tärnidega, kuna neid ei oleks olnud võimalik regulaaravaldistes kasutada. Trellid lisasin nende ridade ette, millest oli parandustes võimalik kasutada vähemalt ühte sõnavormi.

```
#kväär  *väär|käär|kvar|var
#heteronormatiivse  tarnatiivse|heteronormalatiivse
#drag king  tracking|trancking|traking
#queer  quer|*kui|quir|que|quire|keer|*kuur|koir|kuer
#hetero hetega|hedra|hedro|*häda
#gei    *keegi|*key|kei|käi|*kee
#straight   *sheet|*strate|
#lesbian    les|lespen|lespin
#cishet siset
#transfoob  transfob
#trans education    transagucation
#top surgery    top serchie
trans   *ran|*prants
masc    *mask
transfoobide    *transformatsioonide
```

Joonis 3. Näide LGBT+ sõnavara tuvastusvigadest.

LGBT+ kogukonnaga seotud sõnavigade tekkepõhjused kattusid suurel määral ülejäänud sõnades esinenud vigade tekkepõhjustega, kuid lisaks neile mõjutas LGBT+ sõnade tuvastuskvaliteeti asjaolu, et sõnavarast moodustavad suure osa inglise laensõnad või eesti uudissõnad, mis puuduvad tõenäoliselt kõnetuvastuse treeningandmetest täielikult. Kuigi sõnastikuväliste sõnade tuvastamiseks kasutatav lähenemine on võimaldanud

11 https://github.com/leenakt/LGBT_sonavara/blob/main/Failide%20eelt%C3%B6%C3%B6tlus/LGBT_vead.txt

nende tuvastuskvaliteeti parandada, jääb see siiski alla sõnastikus esinevate sõnade tuvastuskvaliteedile.

3.3. Kõnetuvastuse väljundi korrastamine

Andmete analüüsimiseks viisin need esmalt ühtsele kujule, kus kõigi taskuhäälingute episoodide transkriptsioonid olid võimaluste piires parandatud ja tekstifailideks teisendatud. Kuna peale kümne episoodi käsitsi parandamist jaotusid transkriptsioonid kaheks – parandatud TextGrid failid ja parandamata CTM, SRT, TRS ja JSON failid, ühtlustasin neid kahes osas.

3.3.1. Parandatud failid

Kõnetuvastuse väljundina saadud ja eelnevalt parandatud failide korrastamiseks kasutasin samuti rakendustarkvara R. Kuna Praati TextGrid formaadis failid ei olnud minu edasiseks andmeanalüüsiks sobivad, teisendasin need esmalt R-i paketi *textgRid* (Reidy 2016) abil TXT failideks. Teisenduse tulemusena sain eraldi faili kõigi episoodide iga kõneleja kõnevooru väljundist (kokku 71 faili), millele lisasin kõnelejate eristamiseks kõigi ridade ette vastava kõneleja koodi (K01 – K05). Seejärel kleepisin iga episoodi kõigi kõnelejate failid kokku üheks tekstifailiks. Teisenduse lõpptulemusena sain iga episoodi kohta selle parandatud ja parandamata versiooni väljundit sisaldavad tekstifailid. Kogu teisendusteks kasutatud kood on leitav siin[12].

3.3.2. Parandamata failid

Parandamata 21 faili edasises analüüsis kasutamiseks viisin need esmalt JSON formaadist lihtteksti kujule. Selleks kasutatud kood on leitav siin[13]. Seejärel parandasin neis

[12] https://github.com/leenakt/LGBT_sonavara/blob/main/Failide%20eelt%C3%B6%C3%B6tlus/failide_teisendused.Rmd

[13] https://github.com/leenakt/LGBT_sonavara/blob/main/Failide%20eelt%C3%B6%C3%B6tlus/json_tekstifailiks.rmd

tõenäoliselt esinevaid LGBT+ sõnavara tuvastusvigu regulaaravaldiste abil. Selleks kasutasin eelnevalt failide joondamisel kindlaks tehtud sagedasi valetuvastusi. Kuna valetuvastuste hulgas oli palju selliseid juhtumeid, kus LGBT+ sõnad olid tuvastatud mõne muu eesti või inglise keeles esineva sõna või fraasina, näiteks *homo – oma*, *LGBT – ega te*, *gei – key*, jätsin need regulaaravaldistest välja.

Kokku sain parandustes kasutada 198 valesti tuvastatud sõnavormi. Otsisin neid vorme regulaaravaldiste abil parandamata failidest ja asendasin need korrektsete LGBT+ sõnadega. Kuna parandatud failide maht oli võrdlemisi suur, ei ole need peale parandamist kindlasti täielikult korrektsel kujul, kuid annavad LGBT+ sõnavara kasutusest siiski üsna hea ülevaate. Failide parandamiseks kasutatud kood on leitav GitHubis[14].

3.4. Morfoloogiline analüüs

Korrastatud failidele lisasin morfoloogilise analüüsi EstNLTK koosseisu kuuluva Vabamorfi morfoloogilise analüsaatori abil. Kasutasin selleks Vabamorfi standardvarianti vaikesätetega, lisades ainult slängileksikoni. Kuna töö kontekstis on oluline ainult lemmatiseerimine, jätsin muu morfoloogilise analüüsi väljundis oleva info välja. Analüüsiks kasutatud kood on leitav siin[15]. Morfoloogilise analüüsi piiranguna pidin arvesse võtma seda, et suulise kõne transkriptsiooni puhul pole lausete piirid korrektselt märgistatud. Kuna morfoloogiline analüsaator kasutab sõnadele õige analüüsi määramiseks lausekonteksti, võib transkriptsiooni analüüside täpsus varasemalt raporteeritud täpsusest erineda. Samuti esineb käsitsi kontrollimata transkriptsioonides valetuvastusi, mida pole võimalik korrektselt analüüsida. Siiski võimaldas

14 https://github.com/leenakt/LGBT_sonavara/blob/main/Failide%20eelt%C3%B6%C3%B6tlus/LGBT_parandused.Rmd

15 https://github.com/leenakt/LGBT_sonavara/blob/main/Failide%20eelt%C3%B6%C3%B6tlus/morf_analyys.py

lemmatiseerimine mitmeid sõnavorme kokku koondada ja lihtsustas seeläbi edasist analüüsi.

3.5. LGBT+ sõnavara ja selle kasutus

LGBT+ sõnavara analüüsimisel kasutasin R-i paketti *quanteda* (Benoit jt 2018), mis on mõeldud tekstikorpuste töötlemiseks. Pakett võimaldab olemasolevatest tekstifailidest korpuseid koostada ning neile nii lihtsamaid kui keerukamaid statistilise analüüsi meetodeid rakendada. Leidsin, et minu töö kontekstis oleks kõige informatiivsem vaadelda LGBT+ kogukonnaga seotud sõnade sagedusi, neid sisaldavaid bi- ja trigramme ning sõnade kollokatsioone. Sel viisil saab võrrelda spetsiifilisema ja üldistavama, kaasavama sõnavara kasutust ning uurida terminoloogia kaasajastamise mõju kogukonnaliikmete keelekasutusele. Samuti võimaldab see hinnata eesti- ja ingliskeelsete paralleelvormide levikut ning selgitada välja, millises kontekstis LGBT+ kogukonnaga seotud sõnad esinevad ja kuidas kajastub kogukonnaliikmete kõnes enesesildistamine.

3.5.1. Sagedused

LGBT+ sõnavara kasutuses uurimiseks koostasin tekstide põhjal esmalt temaatilise sagedusloendi. Kogukonnaga seotud sõnade valimiks võtsin tekstide parandamise käigus koostatud sõnaloendi, mille 74st sõnast ja fraasist sain sagedusloendi koostamiseks kasutada 38 sõnavormi (Lisa 1). Vähene kasutatud sõnavormide hulk tulenes sellest, et paljud sõnad kattusid omavahel osaliselt, seega sain sagedusloendi koostamisel sellesse kaasata ka aluseks võetud 38 sõnavormi sisaldavad sõnad. Lisaks oli isegi peale morfoloogilise analüüsi rakendamist tekstides palju erinevaid sama või sarnast tähendust kandvaid vorme. Kuna lisasin algselt sagedusloendisse kõik aluseks võetud sõnavorme sisaldavad sõnad, sattus sellese märgatav hulk soovimatuid sõnu, mille uue sõnaloendi (Lisa 2) abil välja filtreerisin. Peale loendi puhastamist jäi sellesse 322 erinevat LGBT+

kogukonnaga seotud sõna. Kogu sagedusloendi tegemiseks kasutatud kood on leitav siin[16].

Tabelis 3 on näha taskuhäälingute 50 kõige sagedasemat LGBT+ kogukonnaga seotud sõna ja nende sagedused. Kogu tabel on leitav GitHubis[17]. Kõige rohkem esines tekstides lühendit *LGBT*, mille märkimisväärselt suur esinemissagedus on osaliselt seotud sellega, et morfoloogiliselt analüüsitud tekstides polnud võimalik teha vahet LGBT ja LGBT+ kirjapiltide vahel. Ootuspäraselt kasutati taskuhäälingutes sageli erinevaid identiteedisilte, sealhulgas *gei*, *trans*, *lesbi*, *homo*, *mittebinaarne*, *bi* ja *aseksuaalne*.

Keelelisest vaatepunktist leidub sagedaste sõnade hulgas mitmeid eesti-inglise sõnapaare nagu *gei – gay*, *pride – praid*, *kväär – queer* ning *lesbi – lesbian*. Nende sõnapaaride sagedustest järeldub, et kõigil juhtudel peale *pride – praid* vastanduse kasutatakse eestikeelset sõnavormi sagedasemini. See näitab üldiselt, et Eesti LGBT+ aktivistid on huvitatud ingliskeelsetele terminitele eestikeelsete vastete leidmisest ja valmis neid igapäevaelus kasutama. Siiski tuleb eesti- ja ingliskeelsete sõnade kasutuse hindamisel arvesse võtta kõne- ja keeletuvastuse kvaliteeti. Näiteks sõna *pride* puhul oli failide parandamisel võimalik keeli eristada vaid konteksti põhjal ning kuna ka see jättis võimaluse mitmeti tõlgendamiseks, kasutasin parandustes eranditult sagedasemat *pride* vormi. Samas ei lugenud ma vigadeks juhtumeid, kus *praid* vorm juba transkriptsioonis esines. Teiste sõnade puhul on häälduserinevused veidi selgemad ja alust loota, et transkriptsiooni kvaliteet seetõttu samuti parem.

Lisaks eelmainitutele tõusid sagedaste sõnade hulgas esile abielu ja abieluvõrdsuse temaatika ning transsoolisusega seotud sõnad, mis annab aimu sellest, et need teemad on Eestis LGBT+ kogukonna jaoks olulised ja aktivistid soovivad neile tähelepanu tõmmata. Üsna palju esineb ka sõnu *hetero*, *heteroseksuaalne* ja *cis*, mida harilikult kõnes eraldi välja ei tooda. See näitab, et erinevalt laiemast ühiskondlikust vaatest nähta

16 https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Sagedusloend/sagedusloend.Rmd

17 https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Sagedusloend/sagedusloend.xlsx

heteroseksuaalsust ja paiksoolisust LGBT+ kogukonnas vaikimisi normina, mis eraldi markeerimist ei vaja.

Tabel 3. LGBT+ kogukonnaga seotud sõnavara sagedusloend.

Rida	**Sõna**	**Sagedus**
1	lgbt	202
2	drag	81
3	gei	80
4	abielu	76
5	pride	76
6	trans	69
7	lesbi	69
8	kväär	67
9	transinimene	66
10	hetero	61
11	praid	61
12	homo	58
13	queer	52
14	mittebinaarne	48
15	abielluma	41
16	transsooline	35
17	vikerkaar	29
18	bi	25
19	gay	22
20	transnaine	21
21	aseksuaal	21
22	lesbiline	20
23	lesbiliit	20
24	transmees	20
25	homokringel	18
26	aseksuaalne	18
27	transsoolisus	18
28	heteronormatiivne	17
29	race	17
30	abieluvõrdsus	17
31	aseksuaalsus	16
32	heterokringel	15
33	homoseksuaalsus	14
34	homofoobia	14
35	gender	14
36	heteroseksuaalne	14
37	transinimesi	13

38	homofoob	12
39	they	12
40	homoseksuaalne	12
41	lesbian	11
42	binaarne	11
43	heteromees	10
44	stonewall	9
45	transnormatiivsus	9
46	homofoobne	8
47	cis	8
48	biseksuaalne	8
49	homoseksuaal	7
50	praidi	7

3.5.2. Bi- ja trigrammid

Peale sagedusloendi koostamist otsustasin vaadata LGBT+ sõnavara sisaldavaid bi- ja trigramme. Bi- ja trigrammide tegemisel võtsin aluseks juba eelmises sammus kasutatud 38 sõnavormi. Kuna *quanteda* bi- ja trigrammide koostamise funktsioon ei lisa neile automaatselt esinemissagedusi, koostasin mõlema loendi kohta ise sagedustabeli ja salvestasin selle eraldi andmetabelisse. Seejärel eemaldasin uue sõnaloendi (Lisa 3) abil bi- ja trigrammide hulgast soovimatud tulemused. Kogu selleks kasutatud kood on leitav siin[18].

Tabelis 4 on näha 50 kõige sagedasemat LGBT+ kogukonnaga seotud sõnu sisaldavat bigrammi. Kogu tabel on leitav GitHubis[19]. Tabelist tuleb selgelt esile, et sagedastes bigrammides mainitakse palju transsoolisuse ja eriti mittebinaarsusega seotud sõnu ja väljendeid nagu *mittebinaarne inimene*, *olema mittebinaarne*, *olema trans* ja *transsooline inimene*. Kuna transsoolised inimesed on üks LGBT+ kogukonna enim marginaliseeritud

[18] https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Bi-%20ja%20trigrammid/bi_trigrammid.Rmd

[19] https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Bi-%20ja%20trigrammid/bigramm.xlsx

gruppe, on loomulik, et taskuhäälingutes soovivad aktivistid transteemadele rohkem tähelepanu pöörata.

Samuti on bigrammides levinud akronüüm *LGBT*, mida kasutatakse eelkõige LGBT inimestest, kogukonnast ja Eesti LGBT Ühingust rääkides. LGBT kogukonna sage mainimine näitab, et kogukonnaliikmete jaoks on kogukonna olemasolu tähtis ning seda rõhutatakse palju. Eesti LGBT Ühingust rääkimisega näitavad aktivistid, kust Eesti LGBT+ inimestel on võimalik vajadusel tuge ning abi leida. Bigrammides esineb ka mitmeid sagedusloendis levinud identiteedisilte nagu *gei* ja *lesbi* ning grupisilte *LGBT* ja *kväär*, mida kasutatakse enamasti koos tegusõnaga *olema*. See näitab, et aktivistid ja teised LGBT+ kogukonna liikmed toovad tihti kogukonnaga seotud teemadest rääkides esile erinevate identiteetide olemasolu. Grupisiltide sage kasutus viitab aga ka kaasavama sõnavara kogukonnasisesele populaarsusele.

Tabel 4. LGBT+ kogukonnaga seotud sõnavara bigrammid.

Rida	**Bigramm**	**Sagedus**
1	mittebinaarne inimene	42
2	lgbt inimene	32
3	lgbt kogukond	32
4	lgbt ühing	27
5	olema mittebinaarne	26
6	eesti lgbt	23
7	olema trans	22
8	olema gei	21
9	olema lgbt	21
10	olema lesbi	19
11	rääkima lgbt	18
12	olema kväär	17
13	lgbt teema	16
14	ja queer	15
15	lgbt sõnasupp	15
16	queer elu	15
17	transsooline inimene	15
18	gei mees	14
19	mittebinaarne ja	14
20	nagu mittebinaarne	14
21	see lgbt	14
22	trans inimene	14

23	olema hetero	12
24	olema homo	12
25	biseksuaalne naine	10
26	drag queen	10
27	ja mittebinaarne	10
28	lesbi ja	9
29	see praid	9
30	bi ja	8
31	et lgbt	8
32	nagu kväär	8
33	olema transsooline	8
34	siis lgbt	8
35	trans ja	8
36	drag king	7
37	nagu gei	7
38	olema aseksuaal	7
39	olema pride	7
40	pride olema	7
41	see drag	7
42	see kväär	7
43	transnaine ja	7
44	baltic pride	6
45	et pride	6
46	heteroseksuaalne naine	6
47	homo olema	6
48	homo või	6
49	lgbt ajalugu	6
50	lgbt organisatsioon	6

Tabelis 5 on näha 50 kõige sagedasemat LGBT+ kogukonnaga seotud sõnavara sisaldavat trigrammi. Kogu tabel on leitav GitHubis[20]. Sagedasemate trigrammide hulgas paistavad esmalt silma Lilla Agenda iga episoodi sissejuhatusest pärinevad sõnakolmikud ridadel 1–5. Samuti on Lilla Agenda sissejuhatusest pärit trigramm *queer elu eestima*, mille tegelik kuju enne lemmatiseerimist oli *queer elust Eestis*. Muus osas on trigrammid bigrammidega üsna sarnased, kuid annavad sõnade kasutuskontekstist veidi laiema ülevaate. Trigrammidest avaldub näiteks sõnade *mittebinaarne* ja *trans* sage kasutus koos

[20] https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Bi-%20ja%20trigrammid/trigramm.xlsx

isikuliste asesõnadega, mis viitab sellele, et transsoolisusest rääkides toetuvad taskuhäälingute autorid ja külalised suuresti isiklikule kogemusele.

Isikuliste asesõnadega koos esinevad tihti ka teised identiteedisildid, näiteks *homo, gei*, *lesbi* ja *aseksuaal*. Neist trigrammidest on kõige silmapaistvam kolmiku *mina olema homo* suhteliselt sage kasutus. Kuna varasemalt on sõna *homo* olnud pigem negatiivse konnotatsiooniga, väljendab selle kasutus endale viitamisel selgelt negatiivsete sõnade kogukonnasisest omaksvõtmist.

Tabel 5. LGBT+ kogukonnaga seotud trigrammid.

Rida	**Trigramm**	**Sagedus**
1	ja queer elu	15
2	lgbt sõnasupp ja	15
3	mis rääkima lgbt	15
4	rääkima lgbt sõnasupp	15
5	sõnasupp ja queer	15
6	mina olema mittebinaarne	12
7	eesti lgbt ühing	10
8	ja mittebinaarne inimene	10
9	queer elu eestima	10
10	siin olema gei	8
11	mina olema trans	7
12	mina olema homo	6
13	mittebinaarne inimene aga	6
14	nagu mittebinaarne inimene	6
15	olema mittebinaarne inimene	6
16	olema mittebinaarne ja	6
17	olema mittebinaarne siis	6
18	mina olema gei	5
19	olema lgbt inimene	5
20	olema see lgbt	5
21	olema trans siis	5
22	eesti lgbt kogukond	4
23	ei olema hetero	4
24	lgbt kogukond olema	4
25	mina olema lesbi	4
26	mittebinaarne inimene ei	4
27	mittebinaarne ja see	4
28	mittebinaarne siis mina	4
29	nagu mittebinaarne et	4
30	nagu see lesbian	4

31	olema bi ja	4
32	olema gei mina	4
33	olema lesbi ja	4
34	olema nagu mittebinaarne	4
35	olema trans ja	4
36	sina olema trans	4
37	ei olema lgbt	3
38	et lgbt inimene	3
39	et pride olema	3
40	et transinimene ei	3
41	gei mees kes	3
42	lgbt inimene olema	3
43	lgbt kogukond inimene	3
44	lgbt ühing ja	3
45	mina olema aseksuaal	3
46	mina olema queer	3
47	mina rääkima lgbt	3
48	mittebinaarne inimene transmees	3
49	mittebinaarne ja transinimene	3
50	mittebinaarne või transinimene	3

3.5.3. Kollokatsioonid

Viimaks koostasin andmete põhjal kahe- ja kolmesõnalised kollokatsioonid, et tuua esile sõnapaarid või -kolmikud, mille koosseisus olevatel sõnadel on tekstides suurem tõenäosus esineda üksteise naabruses kui eraldi. Kollokatsioonide leidmiseks kasutasin samuti *quanteda* paketti ja varasemat sõnavormide loendit. Kogu kasutatud kood on leitav GitHubis[21]. Kuna pakett on eelkõige mõeldud kinnistunud väljendite leidmiseks, on selle peamine piirang kollokatsiooniakna suurus, mis on vaikimisi sama suur kui kollokatsiooni sõnade arv. See tähendab, et kollokatsioonid leitakse ainult otsitavast ja sellega vahetult kõrvuti olevatest sõnadest.

Kollokatsiooni seoste tugevuse hindamiseks kasutatakse paketis lambda ja z-standardiseeritud lambda statistikuid. Lambda statistiku arvutamisel jaotatakse tekst otsitava kollokatsiooniga võrdse pikkusega *n*-grammideks ehk *n* sõnast koosnevateks

21 https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Kollokatsioonid/kollokatsioonid.Rmd

sõnajärjenditeks ning võrreldakse iga otsitavat *n*-grammi kõigi tekstis esinevate *n*-grammidega. Võrdlemise tulemusena loetakse kokku kõigi võimalike sõnakattuvuste-mittekattuvuste kombinatsioonide sagedused. Näiteks võivad kahe bigrammi võrdlemisel kattuda mõlemad sõnad (1, 1), ainult esimene sõna (1, 0), ainult teine sõna (0, 1) või mitte kumbki sõnadest (0, 0). Lambda statistik on võrdne Poissoni log-lineaarse mudeli kõrgeima järgu interaktsiooni koefitsiendi väärtusega, kui saadud kombinatsioonide sagedusi ennustada vastavate kombinatsioonide *n* elemendi kattuvate-mittekattuvate väärtuste ja nende koosmõjude põhjal. (Blaheta, Johnson 2001)

Tabelis 6 on näha LGBT+ kogukonnaga seotud sõnavara 50 suurima lambda väärtusega kahesõnalist kollokatsiooni. Kogu tabel on leitav GitHubis[22]. Tabelist tuleb esile, et kõige suurema koosesinemissagedusega on ootuspäraselt kinnistunud väljendid ja nimetused nagu *vikerkaar kangelane* (vikerkaarekangelane), *queer planeet* (LGBT+ ürituste sari), *top surgerycer / top surgerycher* (*top surgery*, ee *ülakehaoperatsioon*) ning *they them* (sooneutraalsed asesõnad). Samuti tuleb esile Lilla Agenda taskuhäälingu sissejuhatusest pärinev *LGBT sõnasupp*.

Lisaks kinnistunud väljenditele esineb suurema lambda väärtusega kollokatsioonides sõnu, mida tekstides üldiselt nii tihti ei kasutata ja mis seega sagedusloendis ega bigrammide hulgas esile ei tõuse. Kuna harvemini kasutatud sõnadel on suurem tõenäosus kindlates väljendite koosseisus esineda, annab see parema sissevaate LGBT+ kogukonnaga seotud sõnavara mitmekesisusse. Näiteks tähistab kõige suurema lambda väärtusega kollokatsioon *gray ace* väga spetsiifilist aseksuaalsete inimeste gruppi. Esile tõusevad ka kogukonnasiseses diskursuses tähenduse omandavad väljendid nagu *paiksooline heteroseksuaalne*, *queer code*, *heteronormatiivne trash* ja *spicy hetero*.

Tabel 6. LGBT+ kogukonnaga seotud sõnavara kahesõnalised kollokatsioonid.

Rida	Kollokatsioon	Sagedus	Lambda	Z
1	gray ace	2	11.91808608	9.84102078
2	gender happene	2	11.40722045	7.243585734

22 https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Kollokatsioonid/2s_kollokatsioon.xlsx

3	gender trable	2	11.40722045	7.243585734
4	top surgerycer	2	10.86485614	6.946242542
5	top surgerycher	2	10.86485614	6.946242542
6	geiliit arhiiv	2	10.73430265	10.89099116
7	biology is	2	10.69421284	6.847362583
8	vikerkaar kangelane	2	10.61869639	6.802988754
9	they them	5	10.30859482	14.62286846
10	paiksooline heteroseksuaalne	3	10.21762749	11.62389193
11	queer code	2	10.01080679	6.435447722
12	gei hümn	2	9.569556988	6.160788133
13	lgbt sõnasupp	15	9.424266605	10.9813345
14	geiliit üheksakümmend	2	9.246172241	10.67488879
15	heteronormatiivne trash	2	9.246172241	10.67488879
16	queer planet	2	8.912190057	8.550173976
17	them they	2	8.748333812	11.17465748
18	gay people	2	8.715246273	10.6393743
19	drag queen	10	8.700120731	16.12664073
20	ukraina lgbt	2	8.63129865	5.565694029
21	spicy hetero	3	8.590738783	10.24732943
22	drag king	7	8.523221962	14.03139093
23	vikerkaar värv	2	8.421454031	10.37718157
24	hetero gringe	2	8.237316988	9.114334112
25	vikerkaar värviline	2	8.220778889	10.45570428
26	queer skeene	3	8.170037331	11.24212183
27	vikerkaar perekond	3	8.158960987	12.2847273
28	baltic pride	6	8.067596096	14.21989447
29	transnormatiivsus mõiste	2	7.713228955	10.27801681
30	bioloogiline pere	2	7.701270445	9.627146265
31	jätma binaarsus	2	7.632894399	9.837440473
32	cis hetero	2	7.281787755	9.639407022
33	riia praid	4	7.249812635	12.82363611
34	kväär id	3	7.158041435	11.227202
35	valge heteromees	2	7.157944955	9.767272833
36	lgbt ühing	27	7.11919326	25.63020557
37	biseksuaalne naine	5	7.113023341	10.36422379
38	queer dating	2	7.066327792	9.751812635
39	rohkem queere	2	7.008357817	6.776421749
40	rupaul drag	2	6.991918102	9.290166747
41	mittebinaarne karakter	2	6.874416417	9.707247749
42	vikerkaar pere	5	6.849602126	13.9430954
43	pride marss	2	6.788625354	9.315457502
44	vikerkaar lipp	2	6.768446522	9.780407208
45	praid traditsioon	3	6.766131679	11.16382214

46	teine geipaar	2	6.610563273	6.394686446
47	aseksuaal grupp	2	6.543480587	9.490455298
48	trans gender	2	6.501647742	9.242727247
49	drag artist	2	6.421351006	9.085345669
50	pride korraldamine	2	6.402945083	9.115242033

Tabelis 7 on näha LGBT+ kogukonnaga seotud sõnavara 50 suurima lambda väärtusega kolmesõnalist kollokatsiooni. Terve tabel on leitav GitHubis[23]. Kolmesõnalisi kinnistunud väljendeid leidub tekstides üsna vähe ning tulemused kattuvad suuresti trigrammidega. Kõige suurema lambda väärtusega kollokatsioon *ja queer elu* ning kolmandal real olev *mis rääkima LGBT* pärinevad Lilla Agenda sissejuhatusest. Huvitav on aga järgmine sõnakolmik *homofoob olema mina*, mida on tõenäoliselt kasutatud homofoobiaga kokku puutumise kirjeldamisel. See näitab taaskord, et ka negatiivsete teemade puhul räägivad aktivistid ja teised saatekülalised eelkõige oma isiklikest kogemustest.

Tabel 7. LGBT+ kogukonnaga seotud sõnavara kolmesõnalised kollokatsioonid.

Rida	**Kollokatsioon**	**Sagedus**	**Lambda**	**Z**
1	ja queer elu	15	7.288470573	3.555206274
2	siis lesbi suhe	2	5.373089585	2.770276005
3	mis rääkima lgbt	15	5.327969143	6.266130996
4	homofoob olema mina	2	5.277210853	2.488444937
5	et rupaul drag	2	5.223021544	2.283138423
6	kui olema pride	2	5.074816263	3.141170452
7	ja hetero kringel	2	5.035859137	2.589378502
8	lgbt fänn mina	2	4.78941591	2.577492848
9	transsooline või mis	2	4.643160171	2.202843341
10	nägema nagu homo	2	4.486302811	2.692886098
11	et transnaine ei	2	4.301757205	2.038461991
12	sina olema transmees	2	4.229825009	2.003728687
13	aga aga kväär	2	4.096262359	1.948777623
14	spicy hetero et	2	3.973507062	1.696801937
15	see kuidas transinimesi	2	3.966019727	1.869689725
16	olema gei mina	4	3.9456503	2.602771922
17	see kväär i	2	3.809221231	2.943063297
18	pride olema just	2	3.769117912	2.330821334

[23] https://github.com/leenakt/LGBT_sonavara/blob/main/LGBT%20s%C3%B5navara%20anal%C3%BC%C3%BCs/Kollokatsioonid/3s_kollokatsioon.xlsx

19	sõna lesbi mina	2	3.652156949	3.081167194
20	mittebinaarne siis mina	2	3.627522463	2.063540116
21	olema trans siis	5	3.597232918	2.390549438
22	mina olema homo	6	3.515755712	2.303324055
23	queer naine kogukond	2	3.457408492	1.592121733
24	transsoolisus ja see	2	3.448370716	1.948273345
25	transsooline mees olema	2	3.44067313	2.073418434
26	hea pride hea	2	3.385085299	1.843291244
27	sõnasupp ja queer	15	3.369484827	1.296149649
28	olema mittebinaarne siis	3	3.357282624	2.161084833
29	olema hetero siis	2	3.349350396	2.110010115
30	mina olema queer	3	3.284558783	2.564224579
31	olema lesbi siis	3	3.257426612	2.115739889
32	olema pride siis	3	3.237779198	3.362765602
33	sina olema transinimene	2	3.217415191	1.979836976
34	ja mittebinaarne inimene	5	3.206847684	2.109702533
35	olema gei iga	2	3.069927893	1.915523799
36	olema transsooline ja	2	3.06682852	1.904834007
37	heteronormatiivne trash olema	2	2.97569406	1.265250768
38	transsoolisus olema olema	2	2.966465709	2.476228009
39	homokringel mina olema	2	2.926242024	1.653526317
40	näide trans ja	2	2.902282999	2.475284901
41	gei mina ei	2	2.850631035	1.699296338
42	lesbi pidama olema	2	2.84164338	1.78050531
43	pride hea pride	2	2.796170915	1.44125687
44	siin olema gei	8	2.786299683	1.87035684
45	olema ka kväär	2	2.758347552	1.69738268
46	et transinimene ei	3	2.710375805	2.417502934
47	olema kväär aga	2	2.710221106	1.718013918
48	mina olema gei	5	2.662527394	1.771134083
49	et gei olema	2	2.637194333	2.255952747
50	nagu queer naine	2	2.63637702	2.191925152

Kokkuvõte

Bakalaureusetöös uurisin LGBT+ kogukonnaga seotud sõnavara kahes Eesti LGBT+ aktivistide taskuhäälingus, Lillas Agendas ja Homokringlis. Teemavaliku aluseks oli isiklik huvi LGBT+ kogukonda puudutava sõnavara vastu ja asjaolu, et teemat ei ole eesti keeles varasemalt uuritud. Töö peamine eesmärk oli selgitada välja, millist sõnavara kasutavad LGBT+ teemadel rääkides Eesti aktivistid. Kogukonnaga seotud sõnavara kaardistamine on muuhulgas oluline sisend edasistele uurimustele. Saadud tulemuste põhjal on võimalik võrrelda kogukonnasisest keelekasutust kogukonnavälisega või käsitleda LGBT+ sõnavara arengut erinevate ajaperioodide lõikes.

Töö käigus uurisin ka, kui hästi töötab eesti keele automaatne kõnetuvastussüsteem taskuhäälingutest pärineval materjalil. Taskuhäälingud on eesti keele uurimisel uus andmeformaat ning kuna materjali on palju, võimaldab see keelekasutust täiesti uue nurga alt vaadelda. Kvantitatiivsete meetodite rakendamiseks tuleb suuline kõne aga esmalt kirjalikule kujule viia, mistõttu on oluline teada, milline on transkriptsioonide kvaliteet eesti keele praeguse kõnetuvastussüsteemi kasutamisel. Informatsioon tüüpiliste vigade ja probleemide kohta on oluline sisend kõnetuvastuse edasisele arendamisele.

Bakalaureusetöö esimene etapp oli taskuhäälingute automaatne transkribeerimine, valitud transkriptsioonide parandamine ning kõnetuvastuse kvaliteedi hindamine. Kvaliteedi hindamiseks kontrollisin käsitsi mõlema taskuhäälingu viie episoodi transkriptsioone. Transkriptsioonides esinenud vead märkisin üles Exceli tabelisse ning arvutasin kontrollitud episoodide põhjal keskmise sõnavigade määra. Kahe taskuhäälingu keskmine sõnavigade määr oli 23,7%, Lilla Agenda sõnavigade määr 16,3% ning Homokringli sõnavigade määr 28,3%.

Tulemused illustreerivad seda, kui suurel määral varieerub kõnetuvastuse kvaliteet erinevate sisendite puhul ja toovad esile kõnetuvastuse treeningandmete laiendamise olulisuse. Praegusel juhul jäi transkriptsioonidesse suur hulk vigaselt tuvastatud sõnavorme, sh LGBT+ kogukonnaga seotud sõnavara valetuvastusi. Osa vigaseid vorme oli võimalik regulaaravaldiste abil parandada, kuid kuna parandused koostati 31st

kasutatud episoodist kümne põhjal, ei katnud neis esinenud vead kõiki vigu ja arvestatav osa uuritavast sõnavarast jäi endiselt valele kujule.

Taskuhäälingute transkriptsioonide analüüsimiseks lemmatiseerisin need EstNLTK teekide kogumikku kuuluva Vabamorfi morfoloogilise analüsaatori abil. Morfoloogilise analüsaatori kvaliteeti ma töö raames ei hinnanud, kuid tõin ka selles etapis välja mõned peamised probleemkohad, mis andmeanalüüsi raskendasid. Lemmatiseerimisega kaasnenud raskused tulenesid suuresti morfoloogilise analüsaatori sisendi kvaliteedist. Oli igati oodatav, et morfoloogiline analüsaator ei suuda analüüsida transkriptsioonides leiduvaid vigaseid sõnavorme ega taskuhäälingutes sageli esinevaid ingliskeelseid sõnu. Osad lemmatiseerimata jäänud sõnad olid aga üsna ootamatud, näiteks ei suudetud ühist algvormi leida sõnale *transinimene*, mille vormimoodustus ühtib sagedase sõnaga *inimene*. Lisaks märgiti tõenäoliselt kirjavahemärgistuse ning suurtähtede puudumise tõttu sõnavormi *Eestis* lemmaks *eestima*. Kuna Vabamorf on eelkõige mõeldud kirjakeelse teksti analüüsimiseks, tasuks edaspidistes spontaanse suulise keele uurimustes katsetada analüsaatori erinevate komponentide modifitseerimist.

LGBT+ kogukonnaga seotud sõnavara uurimiseks koostasin andmetest sagedusloendi, vaatasin bi- ja trigramme ning kahe- ja kolmesõnalisi kollokatsioone. Peale taskuhäälingute transkribeerimist ja lemmatiseerimist jäi andmete põhjal koostatud tabelitesse mitmeid segaseid vorme, millega tuli tabelite tõlgendamisel arvestada. Sellest hoolimata aitasid kasutatud kvantitatiivsed meetodid andmetest suhteliselt kiiresti ja hõlpsalt otsitavat materjali leida ning tõstsid esile suulise kõne automaattöötlusega kaasnevaid kitsaskohti.

Sagedusloendi alusel hindasin, millised teemad on Eesti aktivistide arvates olulisimad ning mille osas soovitakse ühiskonna teadlikkust tõsta. Kõige enam esines taskuhäälingutes transsoolisuse, spetsiifilisemalt mittebinaarsusega seotud sõnu, mis näitab, et transsoolised inimesed on LGBT+ kogukonna liikmetest kõige enam marginaliseeritud. Tihti kasutati ka erinevaid LGBT+ identiteetide silte, sealhulgas grupisilte nagu *LGBT* ja *kväär*. Selle põhjal võib järeldada, et kaasava sõnavara kasutus on LGBT+ aktivistide hulgas populaarne. Lisaks uurisin töös seda, kui suure osa

aktivistide sõnavarast moodustavad inglise tsitaatsõnad ja kui palju kasutatakse eesti keelde kohandatud sõnu. Eesti keelde kohandatud sõnu kasutati taskuhäälingutes peaaegu eranditult rohkem kui ingliskeelseid tsitaatsõnu, seega võib öelda, et Eesti LGBT+ aktivistid ja kogukonna liikmed on eestikeelse sõnavara arendamisest ja levitamisest huvitatud.

Bi- ja trigrammides esinevad sõnad kattusid suuresti sagedusloendis esinevatega, kuid andsid sõnade kasutusest veidi laiema ülevaate. LGBT lühendi kasutust vaadates tuli esile, et LGBT+ kogukonna liikmete jaoks on oluline kogukonnatunne. LGBT+ sõnavara kollokatsioone vaadates ilmnes, et kogukonnaga on seotud ka mitmeid kinnistunud väljendeid. Samuti esinesid LGBT+ identiteedisildid tihti koos isikuliste asesõnadega, mis näitab, et aktivistid toetusid LGBT+ teemadest rääkides suurel määral isiklikele või oma tuttavate kogemustele. Sellest võib järeldada, et enesesildistamine on LGBT+ kogukonnas eelkõige oluline viis oma kogemusi jagada ning nende alusel kogukonnasisest ühtsustunnet kasvatada.

Vocabulary relating to the LGBT+ community in Estonian LGBT+ activists' podcasts

Discussions surrounding the LGBT+ community are becoming increasingly prominent in contemporary society and the vocabulary used in those discussions plays a significant role in constructing the discourse around LGBT+ matters. This highlights the importance of understanding how different groups of people portray the LGBT+ community and its members through their choice of words. The aim of this bachelor's thesis was to give an overview of the use of words regarding the LGBT+ community in two Estonian LGBT+ activists' podcasts, Lilla Agenda and Homokringel.

The textual data used in this thesis was collected by transcribing a total of 31 podcast episodes, 14 from Homokringel and 17 from Lilla Agenda, using the Estonian speech recognition system (Olev, Alumäe 2022). Since this was the first attempt to use the automatic transcriptions of Estonian podcast data in a computational linguistic analysis, another aim of the thesis was to evaluate the quality of the podcast transcriptions. For the evaluation, five transcriptions from both podcasts were manually corrected and used to calculate the average word error rate. The average word error rate based on both podcasts was found to be 23.7%. The word error rates of the individual podcasts were 16.3% for Lilla Agenda and 28.3% for Homokringel. All the word error rates were higher than the previously reported rates based on the transcriptions of TV news, talkshows and press conferences.

In order to analyse the words related to the LGBT+ community, tables of word frequencies , bi- and trigrams and two or three word collocations were compiled from the texts using the R programming language (R Core Team 2023) and *quanteda* natural language processing package (Benoit jt 2018). The tables were used to qualitatively analyse the use of various words and phrases describing or connected to the LGBT+ community or its members.

The results show that Estonian LGBT+ activists prefer using Estonian LGBT+ terms instead of their English counterparts. The most frequent terms relate to transgender and

non-binary people, which highlights the activists' desire to raise public awareness on those topics. Group terms such as *kväär* and *LGBT* were also prevalent, conveying the importance of inclusive terminology and a sense of community for LGBT+ people. LGBT+ identity labels were most often paired with personal pronouns such as I or you. This indicates that when talking about matters concerning LGBT+ people, the activists often rely on their or their acquaintances' experiences and use self-labeling as a a tool for strengthening the community.

Kirjandus

Alumäe, Tanel 2003. Eestikeelse kõne tuvastus: prototüübi loomine. – Toimiv keel I: Töid rakenduslingvistika alalt. (Eesti keele instituudi toimetised 12.) Tallinn: Eesti Keele Sihtasutus, lk 34–49.

Alumäe, Tanel, Kalda, Joonas, Bode, Külliki, Kaitsa, Martin 2023. Automatic Closed Captioning for Estonian Live Broadcasts. – Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). Tórshavn, Faroe Islands: University of Tartu Library, lk 492–499.

Alumäe, Tanel, Tilk, Ottokar, Asadullah 2019. Advanced Rich Transcription System for Estonian Speech. – Human Language Technologies – The Baltic Perspective, lk 1–8.

Baker, Paul 2012. 'From gay language to normative discourse: a diachronic corpus analysis of Lavender Linguistics conference abstracts 1994-2012. – Journal of Language and Sexuality, kd 2, nr 2, lk 179–205.

Baker, Paul 2018. Language, sexuality and corpus linguistics: Concerns and future directions. – Journal of Language and Sexuality, kd 7, nr 2, lk 263–279.

Baucom, Erin 2018. An Exploration into Archival Descriptions of LGBTQ Materials. – The American Archivist, kd 81, nr 1, lk 65–83.

Benoit, Kenneth, Watanabe, Kohei, Wang, Haiyan, Nulty, Paul, Obeng, Adam, Müller, Stefan, Matsuo, Akitaka 2018. quanteda: An R package for the quantitative analysis of textual data. – Journal of Open Source Software, kd 3, nr 30, lk 774.

Berk, Bernard 2015. Labeling Theory, History of. – International Encyclopedia of the Social & Behavioral Sciences, lk 150–155.

Blaheta, Don, Johnson, Mark 2001. Unsupervised learning of multi-word verbs. Presented at the ACLEACL Workshop on the Computational Extraction, Analysis and Exploitation of Collocations.

Boersma, Paul, Weenink, David 2023. Praat: doing phonetics by computer [Computer program]. Veebiaadress: *http://www.praat.org/*. [Vaadatud: 17.12.2022].

Buntak, Ivana 2020. A corpus analysis of five neologisms from the area of gender and sexuality studies. Zagreb: University of Zagreb, Faculty of Humanities and Social Sciences. Department of English language and literature.

Campbell, D. Grant 2000. Queer theory and the creation of contextual subject access tools for gay and lesbian communities. – Knowledge Organization, kd 27, lk 122–131.

CV: Tanel Alumäe. – ETIS. Veebiaadress: *https://www.etis.ee/CV/Tanel_Alum%C3%A4e/eng*. [Vaadatud: 20.04.2023].

Eesti LGBT Ühing 2008. Eesti LGBT Ühingu põhikiri. Veebiaadress: *https://docs.wixstatic.com/ugd/5a1900_69378f4988e84029a70110ac7dc2622d.pdf*. [Vaadatud: 05.05.2023].

Eesti LGBT Ühing. Eesti LGBT Ühingu strateegia 2022-2024. Veebiaadress: *https://www.lgbt.ee/_files/ugd/5a1900_51b447751e304d5ab8d6837a119e526a.pdf*. [Vaadatud: 05.05.2023].

Galinsky, Adam D., Wang, Cynthia S., Whitson, Jennifer A., Anicich, Eric M., Hugenberg, Kurt, Bodenhausen, Galen V. 2013. The Reappropriation of Stigmatizing Labels: The Reciprocal Relationship Between Power and Self-Labeling. – Psychological Science, kd 24, nr 10, lk 2020–2029.

Juang, B., Rabiner, Lawrence 2005. Automatic Speech Recognition – A Brief History of the Technology Development.

Kaalep, Heiki-Jaan, Vaino, Tarmo 2001. Complete Morphological Analysis in the Linguist's Toolbox. Veebiaadress: *https://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf*. [Vaadatud: 06.05.2023].

Koostöö. – LGBT Ühing. Veebiaadress: *https://www.lgbt.ee/koostoo*. [Vaadatud: 18.05.2023].

Koppel, Kristina, Kallas, Jelena 2020. Eesti keele ühendkorpus 2019. DOI: *https://doi.org/10.15155/3-00-0000-0000-0000-08565L*. [Vaadatud: 22.04.2023].

Lippus, Pärtel, Aare, Kätlin, Malmi, Anton, Tuisk, Tuuli, Teras, Pire 2021. Phonetic Corpus of Estonian Spontaneous Speech v1.2. Institute of Estonian and General Linguistics, University of Tartu. Veebiaadress: *https://datadoi.ee/handle/33/351*. [Vaadatud: 22.04.2023].

Läbipaistmatu süsteem ja hinnanguid andvad spetsialistid ehk Kuidas kohtleb riik trans-inimesi 2021. – Feministeerium. Veebiaadress: *https://feministeerium.ee/labipaistmatu-susteem-ja-hinnanguid-andvad-spetsialistid-ehk-kuidas-kohtleb-riik-trans-inimesi/*. [Vaadatud: 22.04.2023].

Meister, Einar 2021. A corpus of elderly Estonian speech (under development). DOI: *https://doi.org/10.15155/9-00-0000-0000-0000-00220L*. [Vaadatud: 22.04.2023]

Mendelsohn, J., Tsvetkov, Y., Jurafsky, D. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. – Frontiers in Artificial Intelligence, kd 3.

Motschenbacher, Heiko 2019. "Language and sexual normativity.". – Oxford Handbook of Language and Sexuality. Eds. Rusty Barrett, Kira Hall.

Motschenbacher, Heiko 2022. Linguistic Dimensions of Sexual Normativity: Corpus-Based Evidence. New York: Routledge.

Muischnek, Kadri, Fišel, Mark, Kaalep, Heiki-Jaan, Koit, Mare, Müürisep, Kaili, Orav, Heili, Vare, Kadri, Õim, Haldur 2012. Arvutilingvistika ja keeletehnoloogia Tartu Ülikoolis. – Emakeele Seltsi aastaraamat, kd 57, nr 1, lk 66–102.

Olev, Aivo, Alumäe, Tanel 2022. Estonian Speech Recognition and Transcription Editing Service. – Baltic Journal of Modern Computing, kd 10, nr 3, lk 409–421.

Projektist 2022. – Peemoti keskus. Veebiaadress: *https://peemot.ee/projektist/*. [Vaadatud: 18.05.2023].

Proksch, Sven-Oliver, Wratil, Christopher, Wäckerle, Jens 2019. Testing the Validity of Automatic Speech Recognition for Political Text Analysis. – Political Analysis, kd 27, nr 3, lk 339–359.

R Core Team 2023. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Reidy, Patrick 2016. textgRid: Praat TextGrid Objects in R.

Saul, Kertu 2022. ESTNLTK morfoloogilise analüsaatori ja ühestaja kvaliteedi hindamine. Tartu: Tartu Ülikool.

Soolise üleminekuga seotud meditsiiniteenused ja juriidiline soo tunnustamine. – LGBT Ühing. Veebiaadress: *https://www.lgbt.ee/soo-tunnustamine*. [Vaadatud: 15.04.2023].

Sõnastik. – LGBT Ühing. Veebiaadress: *https://www.lgbt.ee/sonastik*. [Vaadatud: 17.04.2023].

Sõnastik 2023. – Feministeerium. Veebiaadress: *https://feministeerium.ee/sonastik/*. [Vaadatud: 17.04.2023].

Sõnastik: transfoobia, cis-seksism ja transvaenulik soopimedus 2019. – Feministeerium. Veebiaadress: *https://feministeerium.ee/dictionary/transfoobia-cis-seksism-ja-susteemne-cis-privilegeeritus/*. [Vaadatud: 20.04.2023].

Tiidenberg, Katrin, Allaste, Airi-Alina 2020. LGBT activism in Estonia: Identities, enactment and perceptions of LGBT people. – Sexualities, kd 23, nr 3, lk 307–324.

Van Rossum, Guido, Drake, Fred L. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vikerkaar arstikabinetis 2018. – Feministeerium. Veebiaadress: *https://feministeerium.ee/vikerkaar-arstikabinetis/*. [Vaadatud: 20.04.2023].

Viks, Ülle 2000. Eesti keele avatud morfoloogiamudel. – Arvutuslingvistikalt inimesele. Toim Tiit Hennoste. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikooli kirjastus, lk 9–36.

Lisa 1. Sagedusloendi, bi- ja trigrammide ning kollokatsioonide tegemisel kasutatud LGBT+ sõnad.

drag
bi
lgbt
kväär
hetero
queer
gei
straight
lesbian
mittebinaar*
soodüsfooria
pride
praid
hatecrime
gender
gay
stonewall
homo
they
trans
cis
viker
aseksuaal
alloseksuaal
aromanti*
biseksuaal
asexual
lgbtqia
ally
nonbinary
pronoun
panseksuaal
androgüün*
deadname
ace
mikrolabel
top

Lisa 2. Sagedusloendi puhastamiseks kasutatud sõnad.

läbi	racist	totally	bioloogiliselt
abi	sebima	debiilik	bik
klubi	space	really	lemmikleiklubi
sobima	seksiabi	sobilik	babie
facebook	rubina	araabia	especially
häbi	veebipood	tarbimine	inimbarbienukk
läbima	destabiliseerima	bin	face
veebileht	ööbima	bio	bigge
facebookis	arstiabi	stabiilne	racism
läbipõlemine	ambitsioon	sobiv	inimbarbie
sobituma	veebinõustamine	abielutseremoonia	desirability
l*ibid*o	beebi	läbimõeldult	enesehäbistamine
literally	libiido	läbirääkimine	häbistama
bioloogia	valehäbi	hobi	transformation
bioloogiline	translate	noortekabinet	embrace
abil	actually	häbitunne	bit
transport	bitch	jobistama	rabbit
viibima	läbimõeldud	pulbitsev	surface
kiirabi	biology	robima	mobile
tarbima	place	kitkatklubin	sobivam
l*ibid*ot	racers	abielumees	versace

syncis	responsibility	kompatibilitt	ambivalentne
bid	ambisies	biker	sööbima
facebooki	bilan	geiti	häbiväärne
eventually	abitoo	plastopring	keskeltläbi
autobiograafiline	stabilisaator	horoskoobir	abistama
firebirdi	stabiilselt	abiellunud	biopsia
süübimine	nilbitsema	häbiväärselt	hüübima
prooviabielu	changeing	rambivalgus	rebimine
rebima	läbinisti	abielupühadus	ühistransport
läbimõtlemine	columbia	abieluväline	tuubike
veebiarendus	hobiseltskond	abitu	abiellumata
kombinatsioon	bluusiklubi	abielluga	sobitama
biit	püsieksibitionine	abieluriik	transformatiivsus
humanitaarabi	kabiin	mobiiltelefon	kombineerima
beebitoit	birgi	abiline	läbimine
tarbija	klubivetsus	kabis	transporditsus
bipolaarsus	elsibith	tace	eneseabi
bipolaarne	transform	häbitult	abiotsing
kriisiabi	bionair	krõbisev	esmaabi
transkulpturel	challengeing	franciscu	ohvriabi
viibimine	stabiliseerima	birt	abitelefon
competbility	läbiv	kiibitsema	abiotsimine

kabinet	libima	tops	kristoper
sobi	l*ibid*omõiste	toppima	topik
abiseadus	*ibid*ot	topelt	ortopeedia
abieluteemaline	l*ibid*os	christopher	topnoh
transpordiküsimus	h*ibid*o	autopiloot	topeltelu
libisema	läbikukkumine	abikaasa	
autobiograafia	big	christoper	
straightegia	häbiasi	offtop	
libi	top	stoper	

Lisa 3. Bi- ja trigrammide ning kollokatsioonide puhastamiseks kasutatud sõnad ja sõnavormid.

äbi
abi
klubi
obi
face
ibi
rally
transpor
bim
rac
space
bin
eebi
bit
translate
ually
place
totally
really
bik
especially
transfor
bil
versace
ncis
bid
rbi
bip
transkulpturel
mbi
changeing
birgi
challengeing
geiti
plastopring
bii
tace
krõbisev
birt
big
tops
toppima
topelt
christopher
autopiloot
christoper
offtop
stoper
topik
ortopeedia
topnoh
topeltelu
rebi
rubi

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Leena Karin Toots,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „LGBT+ kogukonnaga seotud sõnavara Eesti LGBT+ aktivistide taskuhäälingutes“, mille juhendajad on Liina Lindström ja Maarja-Liisa Pilvik, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Leena Karin Toots

30.05.2023