

Tartu Ülikool
Humanitaarteaduste ja kunstide valdkond
Eesti ja üldkeeleteaduse instituut

**Twitteri korpus ajas muutuva keelekasutuse
uurimiseks**

Bakalaureusetöö

Martin Vainikko
Juhendajad Heiki-Jaan Kaalep ja
Siim Orasmaa

Tartu 2022

Sisukord

Sissejuhatus	5
1. Taust	7
1.1. Netikeel	7
1.2. Mis on Twitter?	8
1.3 Twitter kui uurimismaterjal	8
1.4. Uurimisküsimuse püstitamine	9
1.5. Valdkonnad, mida saab uurida	10
1.5.1. Metagraafia varieerumine	11
1.5.2. Ebastandardne morfoloogia	11
2. Twitter	13
2.1. Twitteri kasutamine	13
2.2. Ajalugu	14
2.3 Programmatiline ligipääs Twitteri säutsudele	16
3. Korpus ja teisendamine	17
3.1. Lähtematerjal	17
3.2. Korpuse iseloomustus	18
3.2.1. Säutsude jaotus	18
3.2.2. Kasutajad	19
3.2.3. Aktiivsus	21
3.2.4. Kellaajalisus	22
3.3. Korpuse korrastamine	23
4. Korpuse ettevalmistus ja netikeelsuse nähtuste ilmingud	25
4.1. Ortograafia ja metagraafia tunnuste märgendamine	25
4.1.1. Olemasolev netikeele tunnuste tuvastaja	25
4.1.2. Übermärgendajast märgendajaks	26

4.2. Leksika	28
4.3. Ilmingud	29
4.3.1. Metagraafia tunnused ajas	29
4.3.2. Leksikaalsed tunnused ajas	31
5. Arutelu	33
5.1. Põhilised tähelepanekud	33
5.2. Materjali poolt tõstatatud küsimused	34
Kokkuvõte	35
Kirjandus	37
Twitter corpus for analyzing language changing in time	41
Summary	41
Lisad	44
Lisa 1. Netikeele tunnuseid säutsu kohta ajas – isikud	44
Lisa 2. Netikeele tunnuseid säutsu kohta ajas – naised	45
Lisa 3. Netikeele tunnuseid säutsu kohta ajas – mehed	46
Lisa 4. Netikeele tunnuseid säutsu kohta ajas – organisatsioonid	47
Lisa 5. Tunnuseid keskmiselt säutsus – isikud vs organisatsioonid	48
Lisa 6. Tunnuseid keskmiselt säutsus – mehed vs naised	49
Lisa 7. Tunnuseid keskmiselt säutsus (enne säutsude korduste eemaldamist)	50
Lisa 8. Tunnuste kombinatsioonide kasutussagedused (enne säutsude korduste eemaldamist)	51
Lisa 9. (Kemp 2019)	52
Lisa 10. (Kemp 2022)	53
Lisa 11. Partikleid ja lühendeid keskmiselt säutsus	54

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

Internetis suhtlemine on saanud igapäevaseks lahutamatuks osaks elust. Keel, mida selleks kasutame, on oma iseärasustega ning võib erineda kirjakeelest – netikeel. Üks võimalus netikeelt uurida on Twitter. Twitter kui meedium lühikeste teadete jagamiseks võib olla soodne paik netikeele vohamisele. Seda, kas ja kui palju on Twitteri eesti keeles netikeelele omaseid tunnuseid ajalises dimensioonis, proovitaksegi siinses bakalaureusetöös välja selgitada.

Netikeel hoomab mõistena väga suurt osa keelest, sest interneti keelekeskkonnad varieeruvad suuresti (Kaalep 2017; Hennoste 2013; Oja 2010). Laiema pildi nägemiseks heidetakse pilk varasematele töödele. Vaatamata sellele, et eelnevad tööd ei pruugi uurida sama küsimust samas netikeele keskkonnas, aitavad need täpsustada uurimisküsimust. Ühtlasi saab aimu, millised on netikeele tunnused, mida peaks uurima.

Siinse bakalaureusetöö eesmärk on ette valmistada Twitteri korpus edasiseks uurimiseks, tuua välja esmased ilmingud ning proovida näha kuivõrd sobivad Twitteri säutsud keelekorpuseks. Töö praktiline osa on arvutiprogrammi kirjutamine netikeele tunnuste märgendamiseks. Põhirõhk on materjali ettevalmistamisel, et tulevikus oleks parem uurida. Joondudes varasematest netikeele uurimustest, jaotatakse siinse töö programmeeriline lähenemine kahte ossa: orograafia ja metagraafia ning leksikoloogia osaks.

Bakalaureusetöö on jaotatud viite peatükki. 1. peatükk jaguneb omakorda viite alapeatükki. Esimeses alapeatükis kirjeldatakse netikeele avaldumisvorme. Teises alapeatükis tehakse kiire ülevaade Twitterist ja saadaolevast Twitteri statistikast. Kolmandas alapeatükis on teemaks varasemad Twitteri korpusuuringud, neljandas on varasemad netikeele uurimused ning oletused, et kuidas seonduksid tunnused Twitteri korpusega. Viiendas alapeatükis on vaadeldavate tunnuste täpsustamine, jagunedes kaheks alamosaks: metagraafia varieerumine ja ebastandardne morfoloogia. 2. peatükis tuleb kolmes alapeatükis täpsemalt juttu Twitterist, Twitteri ajaloost tehnilisest vaatepunktist ning Twitteri korpuse kogumisest ja levitamisest. 3. peatükis on kolm

alapeatükki, esimeses alapeatükis on kirjeldatud seda, kuidas korpus koguti. Teises alapeatükis kirjeldatakse korpust: kuidas jaotuvad kasutajad ja säutsud ajas. Kolmandas alapeatükis tehakse ülevaade korpuse eeltöötlemisest. 4. peatükk on jaotatud kolmeks alapeatükiks. Alapeatükkides tuleb juttu praktilisest osast: programmi loomisest, ortograafia ja metagraafia ning leksika tunnuste märgendamisest. Viimases alapeatükis tuuakse välja ilmingud märgendatud korpuse põhjal. 5. peatükis tuuakse välja põhilised tähelepanekud ning materjali poolt tõstatatud küsimused.

1. Taust

1.1. Netikeel

Mati Erelt jt on öelnud kokkuvõtvalt, et “Keel on märgisüsteem, mida kasutatakse suhtlemiseks” (M. Erelt jt 2020: 38). Keele kui märgisüsteemi saab jagada erinevateks tasanditeks, mille väikseim üksus on foneem ning suurim üksus on lause. Foneeme uurib fonoloogia ning lauseid süntaks. Nende keeleteaduse harude vahele jääb veel morfoloogia, moodustades kokku grammatika. Grammatikareegleid kirjakeeles kirjeldab ortograafia. Keele uurimise seisukohalt võivad veel huvi pakkuda leksikoloogia – mis sõnu kasutatakse, ning süntaks – mida sõnad tähendavad. (M. Erelt jt 2020)

Selle kompleksse märgisüsteemi üheks kasutusruumiks on internet. Anni Oja järgi on internetis omakorda väga palju keelekeskondi, igaüks oma erisustega. Ta jagab internetisuhtluse viite rühma: üldmeedia, personaalmeedia, otsesuhtlus, grupisuhtlus, ja veebipõhine suhtlusvõrgustik. Twitteri liigitab ta personaalmeedia alla. (Oja 2010) Tänapäeval on piirid internetivestluse rühmade vahel hägusamad. Sotsiaalmeedia platvormid sobituvad oma funktsionaalsuselt mitmesse rühma. Nt on Twitteris võimalik postitada veebipõhisele suhtlusvõrgustikule omaselt suletumas ringis ning teha otsesuhtlust. Samuti on olemas platvormid nagu Discord, kus on nii otsesuhtlus kui grupisuhtlus (Discordi kanalid) mõlemad kasutuses.

Internetis kasutatava eesti keele ehk eesti netikeele¹ tunnuseid on mitmel keeletasandil. Kasutatakse mitte-kirjakeelseid lühendeid, (nt *mt* - *mis teed*, *prst* - *pärast*); emotikone: tähemärkidest, kirjavahemärkidest ja numbritest kombineeritud pildid, (nt *:O*); ebastandardset kirjavahemärke ja tühikuid, (nt on lause lõpus kirjavahemärgi asemel reavahetus); kõnekeelseid võid toorlaenulisi sõnu, (nt *ellou* või *davai*); lühendatud lauseid, (nt tegusõna ärajätmine); puuduvat või üleliigset läbivat suurtähte. (Kaalep 2017; Hennoste 2013) Tunnuste nimekirja saab veel lisada *emoji*'de kasutamise.

¹ Kaalep (2017) arutleb, et “netikeel” on terminina väga laialdane, sest keel internetis varieerub suurelt, mistõttu võiks vahetule internetivestlusele välja mõelda eraldi termini. Hennoste (2013) nimetas otsesuhtluse stiilis vestluse internetis netikeeleks, internetis kasutatava keele allkeeleks.

Emoji'd on väiksed pildid, mis võivad sümboliseerida emotsioone või igasugu muid objekte. *Emoji*'del on oma Unicode'i koodid. (Bai jt 2019) Nt 😊, 🍔 (vt lisaks Emojipedia²).

Kõige enam netikeele tunnuseid on vahetus vestluses. (Oja 2010; Hennoste 2013)

1.2. Mis on Twitter?

Twitter on 2006. aastal loodud mikroblogimise keskkond, kus kasutajad saavad jagada omavahel kuni 280 tähemärgi pikkuseid teateid ehk säutse (ing *tweet*) (Britannica 2022). Lisaks tekstile saavad kasutajad postitada pilte, veebilinke ja videoid. Kasutajad saavad säutsudele vastata, neid edasi postitada (ing *retweet* - edasipostitus, edasisäuts), säutse meeldivaks märkida, säutsude alla kommenteerida, teisi kasutajaid oma säutsu märkida jpm. Säutse seovad omavahel teemaviidad (ing *hashtag*): teemaviida alla koonduvad teemaviita sisaldavad säutsud. (Twitter Help Center) Nt #EiJaksa (TWT)

2019. aastal kasutas Twitterit Eestis ligikaudu 108 200 inimest, kellest 33% olid naised ja 67% mehed (vt Lisa 12) (Kemp 2019). 2022. aastal kasutas Eestis Twitterit ligikaudu 124 500 inimest (vt Lisa 11) (Kemp 2022). Varasemate aastate kohta kui 2019 Eesti Twitteri kohta statistikat eriti ei leidu. Täpset kasutajate arvu on raske hinnata, sest muud allikad annavad vastuolulisi andmeid.

1.3 Twitter kui uurimismaterjal

Twitter on uurimismaterjalina äratanud huvi paljudes uurijates erinevatest valdkondadest.

Näiteks uuris Kimmo Elo, kuidas kajastuvad poliitilised sündmused Twitteris teemaviidetena: kas Twitteris peegelduvad teemad, mis on aktuaalsed muus meedias või on seal omaette arutelud. Uuriti kahte ligikaudu miljon säutsu sisaldavat korpust, esimeses olid säutsud kuu aega enne 2017. a Saksa parlamendivalimisi ja teises kuu aega enne 2019. a Euroopa Parlamendi valimisi. Teemaviidetel rakendati klasterdamist

² <https://emojipedia.org>

ja tf-idf meetodit, et näha, milliseid teemaviitasid kasutatakse poliitilises arutelus. Üldistusi kogu Twitteri kohta ei tehtud, sest uurimismaterjal kattis vaid ühte osa Twitterist, kuid leiti, et Twitteris peegeldus muu meedia. (Elo 2019)

José Santos uuris oma magistritöös seost edasisäutsumise ja kasutaja eelmiste postituste vahel: kas kasutaja säutsub edasi selliseid säutse, millega tal on rohkem ühist. Kasutaja ja säutsu vahelise ühisosa leidmiseks rakendas ta masinõppe meetodit ning tõdes, et kaugus kasutaja ja edasisäutsu teemade vahel on väiksem kui kaugus kasutaja ja edasiäutsumata säutsu vahel. (Santos 2011)

Käle, Rikters uurisid 2021. aastal kuidas kirjeldatakse toitu 2,4 miljonit lätikeelset säutsu sisaldavas korpuses. Toitu kirjeldava kasutajasõnastiku abil leiti, et viimasel ajal olid lätikeelses Twitteris ülekaalus neutraalsed toidukirjeldused, positiivseid ja negatiivseid oli umbes sama palju. Ühtlasi leiti, et tervisliku toiduga seostati märgatavalt vähem toitu kirjeldavaid sõnu kui maitstva toiduga. (Käle, Rikters 2021)

1.4. Uurimisküsimuse püstitamine

Keele uurimine Twitteris toob esile hulga küsimusi. Esiteks, saab arutleda, kas keel Twitteris on muutunud. Planet Foorumite tekstides täheldati netikeelsuse vähenemist perioodil 2001–2008: järgiti rohkem ortograafia reegleid, kasutati vähem emotikone (Haiba 2016). Siinses töös märgendatav Twitteri säutsude korpus pärineb hilisemast ajast: 2009–2018 (vt ptk 3.), teisalt on Twitteri postituste formaat erinev foorumi omast. Tuvastamaks, kas keel on muutunud või mitte, saab uurida tunnuste kasutamise sagedust ajas (Kaalep 2017).

Teiseks, tuleks kindlaks määrata tunnused, mida uurida. Eelnevalt toodi välja hulk netikeelele omaseid tunnuseid. Lisaks eelnimetatutele võib uurimismaterjalis leida seniteadmata või ülejäänud netikeelele unikaalseid tunnuseid. Saab valida, kas keskenduda ortograafia, punktuatsioon, morfoloogia, süntaksi või võõrkeelte kasutuse tuvastamisele. Iga suunda võiks esindada omaette tunnuste komplekt. Nt kui tahta punktuatsiooniga seonduvat uurida, võiks tunnuste komplekti kuuluda kirjavahemärkidega seonduv, emotikonide ning *emoji*'de kasutus.

Netikeele muutumise puhul tekib küsimus, et kas muutub üksiku kasutaja keelekasutus või muutus tuleneb uute kasutajate pealetulekust (Kaalep 2017). Planet Foorumi staažikate kasutajate keelekasutuses leiti, et muutused toimusid ka üksikisiku keelekasutuses (Haiba 2016). Kuna Twitteri kasutajaskond on laiem ning vähem kogukondlik kui foorumis, võib oletada, et ühtset muutust staažikate kasutaja keelekasutuses on vähem.

Muutused võivad olla seotud tehniliste vahenditega. Twitterit saab kasutada nii nutitefonis kui arvutis, kummalgi omad eripärad. Näiteks võib arvuti klaviatuuri keele paigutusest sõltuda täpitähtede kasutus. Nutitelefoni kaasaskantavuse tõttu võib aga olulist rolli mängida kiirus, mis võib mõjutada teksti pikkust, lühendite ja täpitähtede kasutust (Hennoste 2013). Nutitelefoni klaviatuurid pakuvad üldiselt ka teksti automaatparanduse funktsionaalsust, mis võib samuti mõjutada netikeelsust. Samuti võib üksikisiku keelekasutust mõjutada see, mis säutse ta oma ajajoonel näeb. Ajajoon võib olla kronoloogilises järjestuses või selles järjekorras, mis suurema tõenäosusega paelub kasutajat – ajajoone algoritm (Casper 2017).

Saab ka spekuloida, et missugused netikeele tunnuste muutused ja mõjutavad tegurid on omavahel seoses. Nt klaviatuuri tüüp võib mõjutada ortograafiat täpitähtede mitte kasutamise näol, nutitefoniga kiires olukorras säutsumine aga süntaksi: kiiruga jäetakse nt verb lausesse panemata (Hennoste 2013). Ajajoonel nähtav võib olla seoses kasutaja sõnavaraga, nt slängisõnade kasutamisega.

Ka piisava koguse andmete põhjal ei saa teha terve eestikeelse Twitteri keele muutumise kohta lõplikke järeldusi. Kogutud korpus on vaid väike alamhulk kõikide eestikeelsete säutsude hulgast.

1.5. Valdkonnad, mida saab uurida

Selleks, et täpsustada, mis netikeele tunnustele siinse töös ülevaade anda, tehakse alapeatükkides detailsemad ülevaated kahest eesti netikeele uurimusest.

1.5.1. Metagraafia varieerumine

Kristiina Toots (2014) uuris eesti netikeelt metagraafia vaatenurgast. Erelt jt on metagraafia kohta kirjutanud "On olemas kirjamärkide liigitus, kus kõik muud märgid ja tähistusviisid peale tähestiku tähtede arvatakse METAGRAAFIA kategooriasse. Metagraafia piirid ei ole selged. Kindlalt kuuluvad siia sõnasümbolid, tehnilised osundid, esiletõstetud trükikirjadega, värvusesiletõstetud, samuti kirjavahemärgid. Küsitav on, kas on vajalik tuua metagraafia mõiste alla osa tähestikuseiku, nagu suur- ja väiketähe opositsioon .." (Erelt jt 1993: 387).

Kristiina Tootsi uurimismaterjal koosnes jututubade, kommentaariumide ning foorumite tekstidest. Korpuse tekstid pärinevad aastatest 2002–2004. Töös uuris ta suure algustähe kasutust, punktuatsioonile eelneva ning järgneva tühiku olemasolu ja emotikonide kasutust. Leidis, et jututubades on suure algustähe kasutus harv võrreldes kommentaariumide ja foorumite tekstidega. Põhjuseks võib olla jututubade infovahetuse sünkroonsus. Emotikone kasutati teksti lõpus rohkem kui teksti alguses. Punktuatsiooni kasutuse kohta Toots lõplikke järeldusi ei teinud. (Toots 2014)

1.5.2. Ebastandardne morfoloogia

Muischnek jt tegid eesti internetikeele korpusele automaatse morfoloogilise analüüsi ning täiendasid morfoloogilist analüsaatorit kasutajasõnastikuga. Kasutati OÜ Filosoofi morfoloogilist analüsaatorit *etmrf* ilma oletamise ja ühestamiseta. Analüsaatorile tundmatuks jäänud sõnede hulk näitas potentsiaalset internetikeelsust. Leiti, et kõige rohkem tundmatuid sõnesid sõnede koguarvust on korpuse jututubade tekstides: 27,2%. Foorumite, kommentaaride ja uudisgruppide tekstides jäi see osakaal 5 ja 12% vahele. (Muischnek jt 2011)

Tundmatud sõned jagunesid seitsmesse klassi:

- partiklid, nt *tre, irw, ok, kle, we* (KM jt, 115);
- emotikonid, nt :) või :))) (KM jt, 120), ühe märgi enam kui kolmekordsed kordused asendati tekstis kolmekordse kordusega;

- täheasendused, nt *raffas*, *küll* (KM jt, 117);
- pärisnimed, k.a kasutajanimed;
- võõrkeelsed toorlaenud ja sõnavormid, nt *help ma nii vesinud* (KM jt, 118);
- kirjakeelega vastuolus olevad vormid, nt *loogish*, *minuteada*, *lissalt* (KM jt, 119);
- kirjavead. (Muischnek jt 2011)

Tähe-, märgi- ja silbikordused viidi tekstis ühtsemale kujule. Kui kordusi oli rohkem kui kolm, asendati originaalne sõne variandiga, kus kordus on kolmekordne, kui oli vähem kui kolm, siis asendati ühekordsega. Kasutajasõnastikku lisati eelnimetatud klassidest erinevaid vorme osaliselt automaatselt. Korpusele tehti teist korda automaatne morfoloogiline analüüs. Seekord märgendas analüsaator korpust kasutajasõnastikuga, kusjuures kasutajasõnastik oli prioriteetsem kui analüsaatori sõnastik. Tundmatute sõnede osakaal kõikidest sõnedest langes jututubade tekstides 27,2%-lt 10,5%-le, ülejäänud tekstides langes osakaal kuni 1,5% võrra. Täheledatai, et kasutajasõnastikupõhist meetodit saab edukalt rakendada antud korpuse analüüsimisel, kuid teise korpuse analüüsimiseks tuleks luua uus kasutajasõnastik. (Muischnek jt 2011)

2. Twitter

Siin peatükis on kirjeldatud detailsemalt Twitteri funktsioone ning ajalugu tehnilisest vaatepunktist.

2.1. Twitteri kasutamine

Twitteri tuum on säuts ehk postitus, mis võib sisaldada kuni 280 tähemärgi pikkust teksti või muud meediat (kuni nelja pilti või ühte videot/gif'i, veebilinki) (Twitter Help Center). Säutsu saab meeldivaks märkida, seda edasisäutsuda ja sellele vastata. Edasisäutsule saab kasutaja lisada omalt poolt teksti või muud meediat, originaalsäuts jääb omaette tsitaadiks. Säutsu vastused kuvatakse säutsu all, vastaja ajajoonel kuvatakse kellele nad vastavad. Säutsus saavad kasutajad märkida @-märgiga teisi kasutajaid #-märgiga lisada teemaviitasid. Võib paikneda säutsus ükskõik kus, aga tekst peab vahetult järgnema märgile, nt @kasutaja või #teemaviit. Kasutajamärkele vajutamine viib kasutaja profiilile. Teemaviidale vajutamine, kuvab teisi säutse sama teemaviidaga ning teemaviida järgi saab säutse otsida. (Twitter Help Center)

Näiteid säutsudest:

nii hea meel et sobrannadel hasti laheb janii 😊 (TWT)

Miks ei võiks saada snapi best friende ise valida??? (TWT)

Twitteri keskne koht on ajajoon: siin näeb kasutaja jälgitavate kasutajate säutse ning kasutajale valitud säutse. Muid säutse valib Twitteri algoritm kasutajale selle järgi, mis kasutajat potentsiaalselt huvitab. Soovitusi otsitakse algoritmiga, mille sisendiks on kasutaja Twitteri ringkonnas populaarsed säutsud ja ühised teemad. Nõnda soovitatakse ajajoonel ka kasutajaid, keda jälgida. Säutsud paiknevad ajajoonel voona, kasutaja saab valida, kas järjestus on kronoloogiline või on talle potentsiaalselt huvipakkuvad säutsud eespool. Huvipakkuvad säutsud valitakse jällegi erinevate tegurite järgi, fookuses on kasutajad ja säutsud, kellega kasutaja on rohkem interakteerunud. (Twitter Help Center)

Kasutajad saavad igal hetkel muuta oma säutsud avalikust privaatseks või vastupidi. Muutes oma säutsud privaatseks, näevad kasutaja säutse ainult tema jälgijad ning uued jälgijad vajavad kasutaja jälgimiseks temalt luba. Kui säutsud on avalikud, siis pole kasutaja luba jälgimiseks vaja. Privaatseid säutse ei saa edasisäutsuda ning mitte-jälgija ei näe neid ka siis, kui on säutsutud vastusena tema säutsule. Avalike säutsude puhul saab valida, kes saavad säutsule vastata: kõik, ainult jälgitavad kasutajad või säutsus (@-märgiga) märgitud kasutajad. (Twitter Help Center)

2.2. Ajalugu

Twitter on 2006. aastal loodud sotsiaalmeedia platvorm.

Chris Messina eestjuhtimisel hakkas Twitteri kogukond 2007. aastal #-märki teemaviidana kasutama. #-märgi kasutamine teemaviidana oli inspireeritud IRC-st³, kus kasutati märki kanaliga ühinemiseks. 2009. aasta juulis implementeeris Twitter teemaviitade linkimise: teemaviidale vajutades sai näha teisi säutse sama teemaviidaga. (Anić 2017)

Sarnaselt teemaviidale, tuli kasutusele säutsule vastamine. Esimese teadaoleva @-märgiga vastuse Twitteris tegi Robert S. Anderson formaadis “@ kasutajanimi” 2006. aasta novembris. (Seward 2013) Twitteri kogukonnas kogus nõndaviisi kasutajatele vastamine populaarsust kuni 2007. mais lisati vastamise funktsionaalsus Twitterisse (Stone 2007).

2009. aasta novembris implementeerisid Twitteri arendajad edasisäutsumise funktsionaalsuse. Edasisäutsuminst praktiseeris Twitteri kogukond juba varem erinevates formaatides kuni lõpuks kinnistusid vormid “RT @kasutajanimi” ja “ReTweet @kasutajanimi”, millele järgnes kopeeritud säuts. Edasisäutsumise nupp oli kolmanda osapoole Twitteri äpis Twhirl implementeeritud juba 2007. aasta lõpus. Kasutaja sai siis edasisäutsumise formaadi ise sätestada. Kolmanda osapoole äpp TweetDeck eelsätestas edasisäutsumise formaadi, mis aitas kaasa lühema edasisäutsumise variandi “RT” kinnistumisele. Edasisäutsude loendamine ja

³ Internet Relay Chat

sündmused, kus Twitterit kasutati meedia alternatiivina, tõid edasisäutsumisele rohkem tähelepanu kuni see lõpuks lisati Twitterisse. (Paßmann 2019)

2010. aasta aprillis lasi Twitter käiku oma esimese mobiilirakenduse, senimaani olid kasutuses kolmandate osapoolte rakendused (Carlson 2010). Kuu aega hiljem tuli kasutusele Twitteri Androidi äpp. Enne seda kasutati kolmandate osapoolte äppe. (Scott 2010) Sama aasta mai keskel saabus Twitteri äpp ka IOS-i turule. Twitter ostis ära Tweetie, kolmanda osapoole Twitteri IOS-i äpi, ning arendas seda edasi. (Whitney 2010)

Alates 2011. aasta augustist saavad kasutajad lisaks tekstile säutsuda pilte. Enne seda kasutati selleks kolmanda osapoole pildite jagamise võimalusi. (Independent 2011).

2014. aasta aprillis lisati *emoji*'de toetus Twitteri veebilehele. Enne seda kuvati *emoji*'sid ainult IOS-i ja Androidi Twitteri äpis, arvutis mitte. (Moscaritolo 2014) 2014. aasta juunis lisati võimalus postitada GIF-e (Summers 2014).

2016. aasta veebruaris muutis Twitter ajajoont: ajajoone tipus ilmusid säutsud, mis on kasutajale potentsiaalselt huvipakkumad ning nende all ülejäänud säutsud kahanevas kronoloogilises järjestuses. Ajajoont uuendades nägid kasutajad säutse jälle reaajas, nagu enne uuendust. Algoritmi põhine ajajoon oli valikuline, kuid tekitas Twitteri kogukonnas poleemikat. (Casper 2017)

Märgati, et jaapani, hiina ja korea keeles mahutab üks tähemärk umbes kaks korda rohkem informatsiooni kui muudes keeltes. Jaapani- ja ingliskeelseid säutse võrreldes nähti, et jaapanikeelne säuts on keskmiselt 15 tähemärki pikk, ingliskeelne säuts aga 34 tähemärki pikk. Tähemärgi limiit tuli ette 0,4%-l säutsudest jaapani keeles, inglise keeles 9%-l säutsudest. 2017. aasta septembris testiti maksimaalse säutsu tähemärkide arvu suurendamist. Testgrupil tõsteti säutsu tähemärkide limiit 140 tähemärgi pealt 280 peale. (Rosen, Ihara 2017) Sama aasta novembris tõsteti limiit tervele kasutajaskonnale, 280 tähemärgi peale, eeldusel, et keel on toetatud keelte hulgas (Perez 2017).

2.3 Programmatiline ligipääs Twitteri säutsudele

Twitter võimaldab massiliselt säutse koguda Twitteri otsingu programmiliidese⁴ (Twitter Search API) kaudu. Säutsude kogumisele on seatud omad limiidid, nt Twitteri Developer Platformi dokumentatsiooni järgi saab akadeemilisel eesmärgil kuus kuni 10 miljonit säutsu koguda. (Twitter Developer Platform 2022)

Piirangud on seatud ka Twitterist kogutud korpuse levitamisele. Levitada tohib ainult kindla arvu säutsude või kasutajate ID-sid kindlas ajaraamis. Ka siin tehakse akadeemilisel eesmärgil kogutud korpustele erandeid, kuid levitada tohib siiski ainult ID-sid. (Twitter Developer Policy 2022)

⁴ <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview>

3. Korpus ja teisendamine

3.1. Lähtematerjal

Korpus koosneb eestikeelsetest Twitteri säutsudest (ing *tweet*). Säutsud on kogunud Dage Särg 2018. aasta juunis TweetCat⁵ tööriista abil. Siim Orasmaa on säutsud viinud EstNLTK⁶ v1.6 andmekujule ning lisanud ka märgendused. Ühtlasi märgendati kasutajad profiilide põhjal meesteks ja naisteks ning isikuteks ja organisatsioonideks. Korpus on JSON-formaadis, säutsude Text-objektid on ridade kaupa. Iga säutsuga tuleb kaasa hulganisti meta-andmeid. Nt on meta-andmetes säutsu ja säutsuva kasutaja ID-d, säutsumise kuupäev ja kellaaeg, info edasisäutsumise kohta (kui on edasisäuts) jpm.

TweetCat on 2017. aastal tehtud avatud lähtekoodiga tööriist säutsude kogumiseks, kus tööriista kasutaja valib keele, milles ta säutse kogub. TweetCat loodigi selleks, et säutse saaks mugavalt koguda ka väiksemates keeltes. Senised tööriistad säutsude kogumiseks olevat rohkem suunatud ingliskeelsete säutsude kogumiseks. (Ljubešić jt 2014)

TweetCat sisestab Twitteri otsingu programmiliidesesse paarkümmend otsingusõna: TweetCati kasutaja määratud sagedased sõnad, mis kuuluvad ainult uuritava keele sõnavarasse. Kui Twitteri otsingu programmiliides leiab vaste, algab kasutajapõhine keeletuvastus tingimusel, et kasutaja ajajoonel on vähemalt 100 säutsu (alla selle olevat keele tuvastamine keeruline). Järgmisena tehakse keeletuvastus sajakonna levinud sõnaga otsitavast keelest – ka need määrab TweetCati kasutaja. Kui keel on tuvastatud otsitava keelena, salvestatakse kasutaja järjendisse ning TweetCat võtab ette kasutaja jälgijad ja jälgitavad ning kogu protsess algab uuesti. Viimaks võetakse salvestatud kasutajate järjend ette ning kogutakse nende säutsud korpusesse. (Ljubešić jt 2014)

⁵ <https://github.com/clarinsi/tweetcat>

⁶ https://github.com/estnltk/estnltk/tree/version_1.6

3.2. Korpuse iseloomustus

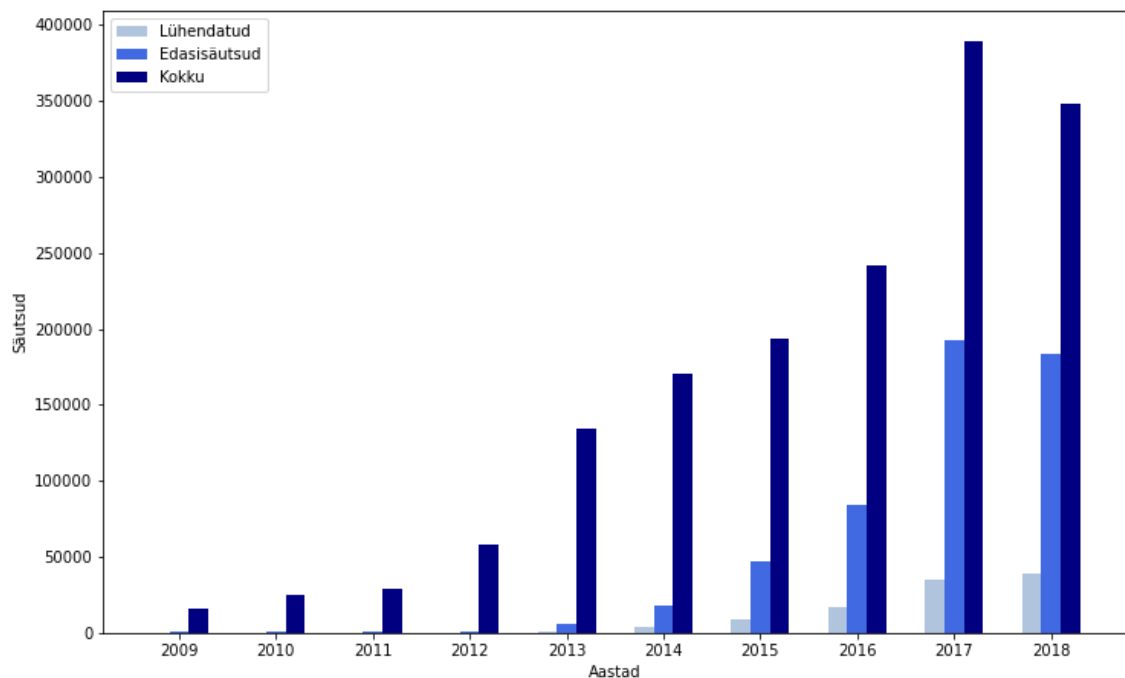
3.2.1. Säutsude jaotus

Enne säutsude jaotuse tegemist tuli täiendada meta-andmeid. Reeglina on edasisäutsude tekstiosa formaadis *RT @kasutajanimi: säuts* kuid umbes 1300 edasiäutsul puudus meta-andmetest edasisäutsumise kohta info. Selle reeglipärasuse põhjal sai statistika tegemiseks märke edasisäutsumise kohta lisatud. Ühtlasi polnud meta-andmetes infot lühendatud (ing *truncated*) säutsude kohta. Need on säutsud, mis on Twitteri otsingu programmilidies lühendanud ning lisanud lõppu ...-sümboli (pole kolm tavalist punkti vaid üks sümbol). Lühendatud säutsude kohta sai meta-andmetesse samuti märke lisatud.

Joonisel 1 on näha, et korpuses on ligikaudu 1,6 miljonit säutsu ajavahemikust 2009–2018. Kõige rohkem säutse on aastast 2017. Kui korpus oleks kogutud 2018. aasta lõpus, oleks kõige rohkem säutse kogumise aastal. Mida aasta vanem, seda vähem on korpuses säutse: 2009. aastal on kõige vähem säutse. Põhjus võib olla see, et Twitteri otsingu programmiliides võib anda tulemuseks hiljutisemad säutsud, mis võib tähendada ka uuemaid kasutajaid. Et sõelale satuvad pika staažiga kasutajad, on vähetõenäolisem. TweetCat hakkab siis säutse koguma arvatavasti värskemalt kasutajaskonnalt.

Samuti on iga aastaga kasvanud edasisäutsude arv nii summarselt kui protsentuaalselt. Kui 2013. aastal moodustavad kõikidest säutsudest vaid 10% edasisäutsud, siis 2018. aastal on juba üle 50% säutsudest edasisäutsud.

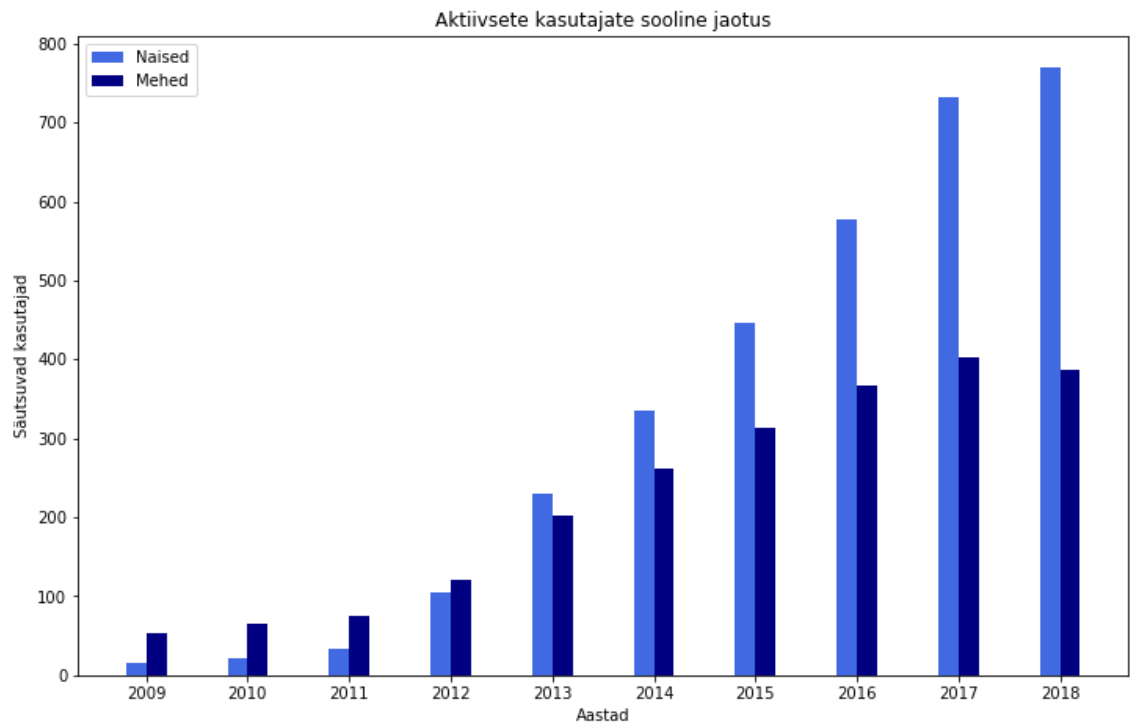
Lühendatud säutsud on ajas ühtlaselt kasvanud kuni 2018. aastani, mil nende osakaal on protsentuaalselt tõusnud.



Joonis 1. Lühendatud, edasisäutusde ning kõikide säutsude koguarv aastate kaupa

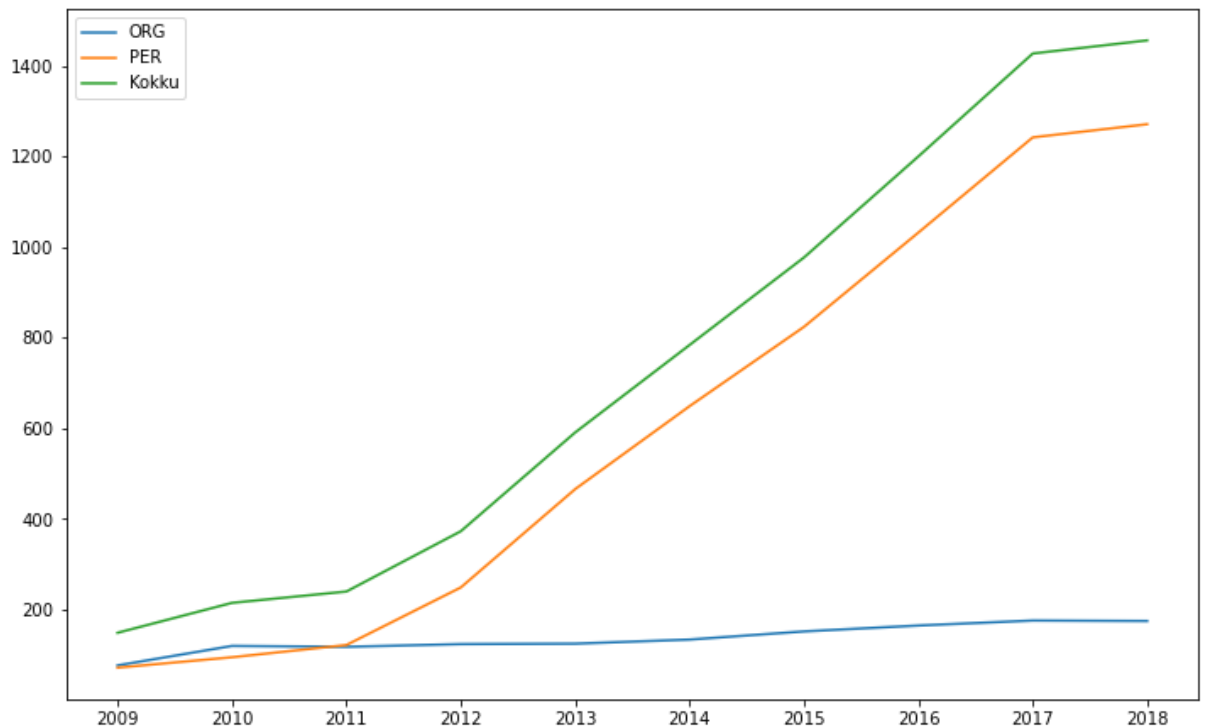
3.2.2. Kasutajad

Kasutajate sooline jaotus aastate lõikes näitab, et korpuse algusaastatel oli Twitteris rohkem meessoost kasutajaid (joonis 2). Kasutajate osakaalus on näha naissoost kasutajate ühtlast tõusutrendi. 2012. ja 2013. aasta vahel oli kasutajate sooline jaotus võrdne, kuid 2013. aastast alates on naissoost kasutajate osakaal järjest suurem. 2018. aastal on meessoost kasutajaid umbes poole vähem kui naissoost kasutajaid, kusjuures 2009. aastal oli naissoost kasutajaid umbes kolm korda vähem kui meessoost kasutajaid. 2018. aastal võib aktiivseid meessoost kasutajaid olla sellepärast vähem, et sel aastal on poole aasta jagu andmeid.



Joonis 2. Kasutajate sooline jaotus aastate kaupa: kasutaja on vaadeldaval aastal teinud vähemalt ühe säutsu

Järgmisel joonisel on muutusi näha ka isikute ja organisatsioonide kasutajate jaotuses. Kuni 2011. aastani on kas isikuid umbes sama palju kui organisatsioone. Peale 2011. aastat toimub isikute kasutajate arvus järsk ning ühtlane kasv. Organisatsioonide arvus toimub samuti kasv, kuid palju tasasem, mistõttu moodustavad organisatsioonid 2018. aastal kõikidest kasutajatest vaid u 12%. Inimesi on arvatavasti rohkem kui organisatsioone, mis tagab ka suurema kasvu isikute kasutajate arvus.



Joonis 3. Vaadeldaval aastal säutsuvate kasutajate jaotus isikute ja organisatsioonide vahel

Naise säuts:

ole vait jobu :D :D nkn isutd, kõnnid koooli minuga jne (TWT)

Mehe säuts:

Kõikide toolide kriuksuvad nagu lämbuvad kajakad (TWT)

Organisatsiooni säuts:

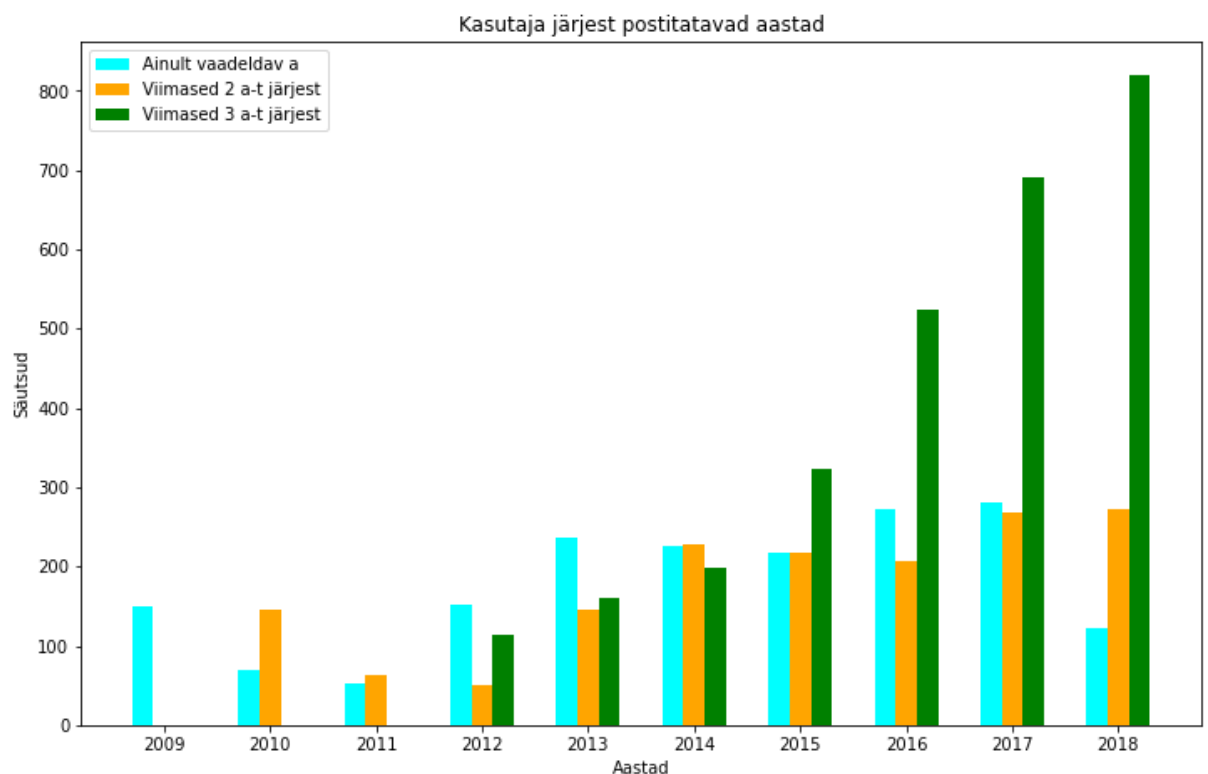
Upal on meil nüüdsest suurem ja palju nutikam CNC-pink. #väikelaevaehitus #kompetentsikeskus (TWT)

3.2.3. Aktiivsus

Joonisel 4 on vaatluse all kasutajate aktiivsust 1–3 järjestikuse postitatava aasta põhjal. 2009. aastal saavad olla ainult vaadeldaval aastal postitavad kasutajad, sest varasemaid

andmeid korpuses pole. Jooniselt on näha, et mida hilisem aasta, seda rohkem on 3 aastat järjest postitavaid kasutajaid. Peale 2014. aastat on ühtlaselt rohkem neid kasutajaid, kes on postitanud 3 aastat järjest.

Joonis on tagasivaatav, seega arvesse pole võetud kasutajate postitamised tulevatel aastatel.



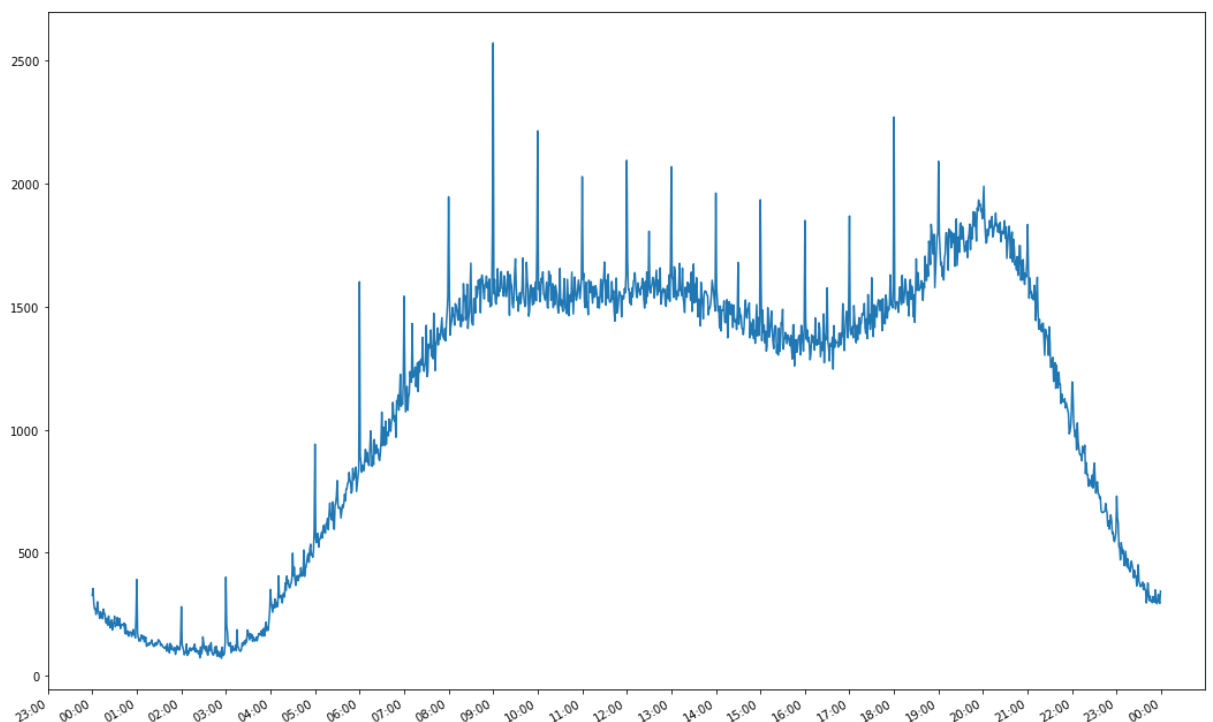
Joonis 4. Kasutajate säutsumise järjepidavus tagasivaatavalt: eelduseks on, et kasutaja on teinud aastas vähemalt ühe korra säutsunud

3.2.4. Kellaajalisus

Joonisel 5 on säutsud jaotatud 24-tunnilisele ajateljele. Jooniselt on näha, et kõige rohkem postitatakse ajavahemikus 19.00–20.30. Arvata on, et siis on inimestel kõige rohkem vaba aega sotsiaalmeedia kasutamiseks, sest päevased kohustused on täidetud. Üsna ühtlast trendi on näha kella 8 ja 13 vahel, samuti on see teine kõige tihedam postitamise ajavahemik. Arvatavasti on need postitused töölt ja koolist. Peale kella 13

on märgata langustrendi, mis jõuab madalpunkti vahemikus 15–16. Selle põhjuseks võib olla töö- ja koolipäeva lõpp: inimesed on teel koju ning postitamiseks pole mahti.

Õhtul peale kella 20 algab postitamise tiheduses järsk langustrend, mis muutub tasasemaks peale kella 00. Inimesed lähevad magama see kell. Kõige vähem postitatakse öösel kell 2–3, postitajateks ilmselt öösel üleval olijad. Peale seda hakkab postitamise tihedus ühtlaselt tõusma, ilmselt on need postitajad varajased ärkajad.



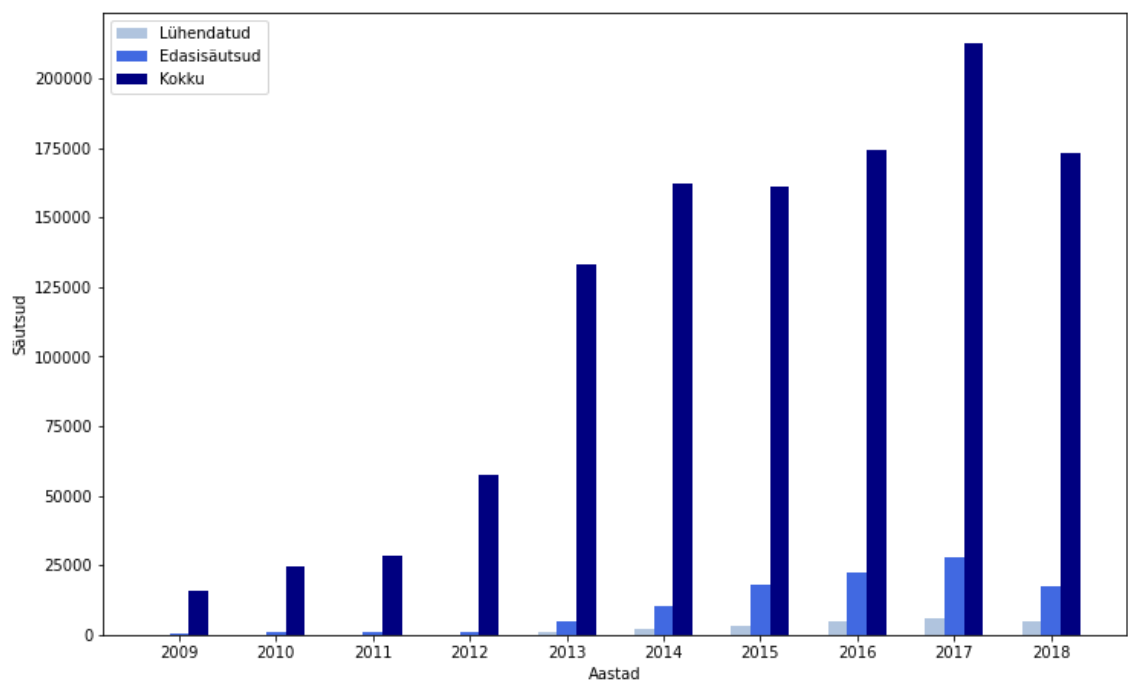
Joonis 5. Säutsumine kellaajaliselt, y-teljel on säutsude koguarv ning x-teljel on kellaajad minuti täpsusega

3.3. Korpuse korrastamine

Korpusesse kogutud originaalsäutsude ja edasisäutsudega tekivad andmekordused. Nt on korpusesse sattunud originaalsäuts, mida on mitu kasutajat edasisäutsunud. Valdav osa kordustest on siiski ilma originaalsäutsuta, vaid mitmest edasisäutsust koosnevad. Kuna kordused mõjutavad keeletunnuste sagedusi, said need eemaldatud. Eemaldatud said need korduvad edasisäutsud, millel oli meta-andmetes Twitteri otsingu

programmiliidese lisatud info edasisäutsumise kohta. Edasisäutsumise metainfo üks märgendeid oli originaalsäutsu ID, mille abil sai kindlaks teha, et säuts on korpuses üks kord. Terve edasine töö on tehtud korpuse peal, millest on kordused eemaldatud. Mõned kordused võisid sisse jääda eelmainitud edasisäutsumise metainfo puudulikkuse tõttu.

Joonisel 6 on kuvatud säutsude jaotus pärast korduste eemaldamist. Edasisäutsude arvud langesid märkimisväärselt. Säutsude koguarv arv langes korpuses 1,6 miljoni pealt 1,1 peale.



Joonis 6. Lühendatud, edasisäutsude ning kõikide säutsude koguarv aastate kaupa pärast korduste eemaldamist

4. Korpuse ettevalmistus ja netikeelsuse nähtuste ilmingud

Programmide peamine tööriist on EstNLTK versioon 1.6.9b0, mis on installeeritud Conda paketihalduri⁷ keskkonda. Conda keskkonda on installeeritud Python⁸ 3.7.11 versioon. Programmi kirjutasin Jupyter Notebook'i⁹ programmeerimiskeskonnas.

4.1. Ortograafia ja metagraafia tunnuste märgendamine

4.1.1. Olemasolev netikeele tunnuste tuvastaja

Programmi aluseks on Helen Kaljumäe 2019. aastal EstNLTK projektis kirjutatud lõigu põhine netikeele tunnuste loendur ParagraphWebLanguageScoreRetagger¹⁰. Übermärgendaja võtab sisendiks EstNLTK lõigumärgenduse, leiab sealt tunnused ning lisab lõigu atribuutidena tunnuste loendused.

Kaljumäe tunnuste loendurist võtsin kasutusle järgnevad tunnused.

Kirjavahemärkide kordused on igasuguste kirjavahemärkide (v.a punkti) kordused: kaks või enam kirjavahemärki järjest, nt *tere!!*. Kui tegu on punktiga, siis läheb korduse alla see, kui punkt on tekstis kaks või rohkem kui kolm korda järjest, sest kolm punkti on kirjakeelne tunnus. Nt *tore..* või *tore....*

Tähemärkide kordus on järjestikune tähemärk rohkem kui kaks korda. Nt *jaaa*. Paraku tuvastatakse ka haruldased liitsõnad, kus on kaks või enam järjestikust tähte ning sõnade vahel pole sidekriipsu, nt *jäääär*.

Puuduvad tühikud tuvastatakse siis, kui peale lauset, osalauset või loendi elementi puudub kirjavahemärgi ja sellele järgneva sõna vahel tühik. Nt *Tere!Nägemist*.

⁷ <https://docs.conda.io/en/latest/>

⁸ <https://www.python.org/>

⁹ <https://jupyter.org/>

¹⁰ https://github.com/estnltk/ettenten-experiments/tree/master/weblang_detection

Läbiv suurtäht saab märgenduse siis, kui tekstis on vähemalt kaks järjestikust läbiva suure tähega vähemalt kahe tähemärgi pikkuselist sõna. Nt *see oli NII LAHE*.

Võõrtähe tuvastuse alla lähevad sõnad, kus on kasutatud võõrtähti *c, q, w, x, ja y*. Võõrtähte sisaldav sõna peab olema läbiva väikese tähega, et vältida pärisnimede tuvastust. Nt *kax kassi*.

Puuduv lausealguline suurtäht on lause alustamine väikese tähega. Nt *tšau! kuidas läheb?* Samuti märgendatakse lausetevahelised valed tühikud. Nt *Tere ! Nägemist!*

Märgendatakse emotikonid, mis viimane ei hooma täielikku emotikonide nimekirja emotikonide paljususe ja valetuvastuste vältimise tõttu. Emotikonid tuvastatakse tekstist asukohast sõltumatult. Nt :) või :))))).

ParagraphWebLanguageScoreRetagger'i sisendiks on EstNLTK CompoundTokenTagger'i¹¹ väljund. CompoundTokenTagger teeb juba osa tööd ära, tuvastab netikeele üksuseid: emotikone, veebilinke, teemaviitasid, kasutajanimed jm. Täiendasin märgendaja emotikonide tuvastust Haiba bakalaureusetööst pärineva regulaaravaldise mustriga (Haiba 2016). Proovisin testhulgal emotikonide tuvastamist ning muutsin mustrit vastavalt. Muster pole täielik – ei tuvasta kõiki laialdalt kasutatavaid emotikone. Mõned emotikonid pidi eemaldama valetuvastuste vältimiseks, mõned tuli lisada.

4.1.2. Ümbermärgendajast märgendajaks

Märgendaja ja ümbermärgendaja vahe seisneb selles, et märgendaja loob uusi kihte ning ümbermärgendaja muudab juba olemasolevaid kihte (EstNLP). Tunnuste loendamiseks oleks piisanud ka ümbermärgendaja funktsionaalsusest: tunnuseid saab loendada metakihti (EstNLP), kuid edasise kasutatavuse mõttes oli parem, kui tunnused on ära märgendatud. Nii sai hiljem teha sagedusloendeid, näha tunnuse asukohta tekstis jm.

¹¹

https://github.com/estnltk/estnltk/blob/version_1.6/tutorials/nlp_pipeline/B_02_segmentation_compound_tokens.ipynb

WebLangTagger¹² on EstNLTK põhine märgendaja, mille kirjutasin kasutaja Kaljumäe loodud ParagraphWebLanguageScoreRetagger'i¹³ põhjal. ParagraphWebLanguageScoreRetagger'i struktuuri ja põhilisi koodiplokke sai suures osas taaskasutada, kuid toon suuremad muudatused välja. Uus märgendaja ei vaata teksti enam lõigu kaupa, vaid terve teksti kaupa, sest säutsud ei ole pikad. Ümbermärgendaja loomine tuli viia märgendaja loomise kujule, kuhu juurde tuli kihi loomine. Lisasin märgendajale kaks uut tunnust: *emoji* ja z-tähte sisaldavad sõnad, mis tõenäoliselt on võõrsõnad. *Emoji*'d (v.a lipu-*emoji*'d) tuvastasin Pythoni *Emoji* mooduli¹⁴ abil. Lipu-*emoji*'sid moodul ei leia, selle jaoks tuli need tuvastada Unicode'i koodide vahemikust¹⁵. Võõra z- ja ž-tähe tuvastamiseks lisasin regexi sõnastikku uue mustri ning validaatori, et välistada eestikeelseid z- või ž-tähte sisaldavaid sõnu. Kogusin z- ning ž-tähti sisaldavad sõnad neljast sõnavormiloendist: tasakaalukorpuse, teadustekstide, ajakirjade ja ilukirjanduse sõnavormiloenditest¹⁶. Ümbermärgendajaga tuli kaasa ka tundmatute sõnade ja puuduvate komade tuvastamine, kuid praegu aja kokkuhoiu mõttes nende tunnuste otsimist ei kohandanud ning märgendamisel lülitasin nende tunnsute lipud välja.

Olemasolevaid regexi mustreid tuli kohandada Twitteri säutsudele vältimaks valetuvastusi. Täiendasin ka mõnda mustrit, sest mõned äärehujud olid jäänud tuvastamata. Muudetud sai tähekordeste, puuduva tühiku, läbiva suurtähe ja võõrtähtede mustrit. Lisasin ka validaatorid selleks, et veebilinkides (mida CompoundTokenTagger ei tuvastanud) tunnuseid mitte tuvastada ja, et järjestikuseid korduvaid *emoji*'sid ei tuvastataks punktuatsioonikordustena.

Viimaks märgendasin korpuse vastloodud märgendajaga.

¹² <https://github.com/nitram-v/web-lang-analysis>

¹³ https://github.com/estnlk/ettenten-experiments/tree/master/weblang_detection

¹⁴ <https://pypi.org/project/emoji/>

¹⁵ <https://stackoverflow.com/questions/43146528/how-to-extract-all-the-emojis-from-text>

¹⁶ <https://www.cl.ut.ee/ressursid/sagedused1/>

4.2. Leksika

Korpuses metagraafia-ortograafia tunnuseid märgendades, lisasin Text-objektile ka morfoloogilise analüüsi koos oletamise ja ühestamisega. Nüüd tegin morfoloogilise analüüsi uuesti, aga ilma oletamise ja ühestamiseta. Uus morfoloogiline analüüs läks eraldi kihti nimega *morph_analysis_without_guessing*. Korpuses oli nüüd hulganisti tundmatuid sõnu (märgendusesks *None*). Tegin kaks uut kihti: tundmatute sõnade analüüsimiseks *none_string_analysis* (vt Joonis 7) ja tundmatute fraaside analüüsimiseks *none_phrase_lang_analysis* (vt Joonis 8). *none_string_analysis* kiht on *morph_analysis_without_guessing* kihi alamkihiks. ja *none_phrase_lang_analysis* kiht on *none_string_analysis* kihi ümbriskiht.

Mdea, ma pidevalt vaatan inimeste tweete ja mõtlen, et kirjutaks neile: "You, sa tundud nii lahe inimene, let's be friends fo... (TWT)

layer name	attributes			parent	enveloping	ambiguous	span count
none_string_analysis	capital_letter	particle_abbrev	en	morph_analysis_without_guessing	None	True	7
text	capital_letter	particle_abbrev	en				
Mdea	True	[None]	False				
tweete	False	[None]	False				
You	True	[None]	True				
let's	False	[None]	False				
be	False	[None]	True				
friends	False	[None]	True				
fo	False	[None]	True				

Joonis 7. *none_string_analysis* kihi näide: ülalpool on toodud märgendatud säutsu tekst

layer name	attributes	parent	enveloping	ambiguous	span count
none_phrase_lang_analysis	lang, probability, phrase_length	None	none_string_analysis	False	1

text	lang	probability	phrase_length
['let's', 'be', 'friends', 'fo']	en	-3.0746937	4

Joonis 8. *none_phrase_lang_analysis* kihi näide: ülalpool on toodud märgendatud säutsu tekst

Tundmatute sõnade analüüsi kihti lisasin atribuutidena info selle kohta, kas sõna on suure algustähega, partikkel, lühend või ingliskeelne. Mitu atribuuti võivad korraga ka tõesed olla. Partikleid, lühendeid ja ingliskeelseid sõnu otsin kasutajasõnastikest. Partiklite ja lühendite kasutajasõnastikud on pärit Dage Säre magistritöö korpusest¹⁷. Ingliskeelsed sõnad pärinevad Githubi projektist¹⁸, kuhu on kogutud 466 000 ingliskeelset sõna. Pärinevad kirjandustekstidest. Kuna tundus, et kõige levinumad sõnad jäid veel katmata, lisasin veel ühe sagdussõnastiku¹⁹ 10 000 kõige sagedasemat sõna kasutajasõnastikku.

Järjestikused tundmatud sõned märgendasin tundmatuteks fraasideks. Tundmatud fraasid lasin läbi *langid*²⁰ keeletuvastuse. *langid* annab analüüsi tulemuseks keelekoodi ja tõenäosuse. Üldiselt oli näha, et lühemate fraaside ja üksikute sõnade peal on valeanalüüse üpriski palju. Inglise keelt tuvastab kõige ühtlasemalt, sest ilmselt on *langid* treeningmaterjalides kõige rohkem inglise keelt näinud.

4.3. Ilmingud

4.3.1. Metagraafia tunnused ajas

Tunnuste madalseis kuni 2012. aastani on tingitud algusaastate andmete vähesusest. Langustrend vahemikus 2017–2018 on tingitud selle tõttu, et 2018. aastal on ainult poole aasta jagu andmeid. Edaspidiseid trende saab juba tõlgendada. 2014. a keskpaigas

¹⁷ <https://www.cl.ut.ee/korpused/jutumorf/>

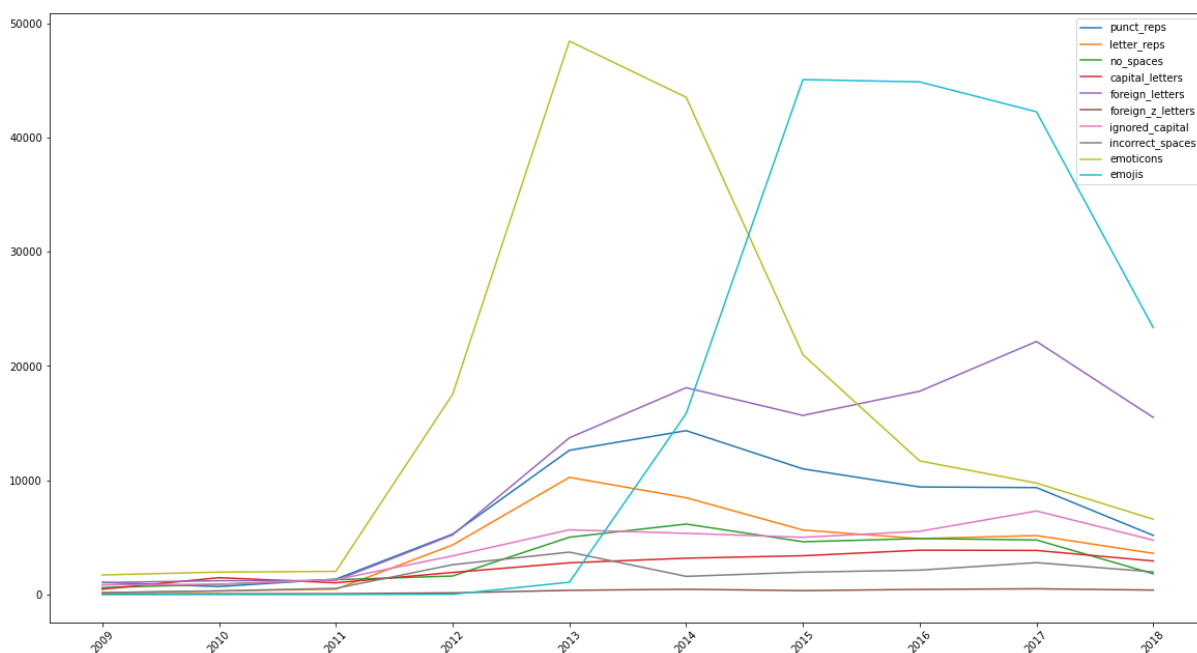
¹⁸ <https://github.com/dwyl/english-words>

¹⁹ <https://github.com/first20hours/google-10000-english>

²⁰ <https://pypi.org/project/py3langid/>

asenduvad emotikonid *emojidega*, kusjuures suurem tõus *emoji*de kasutuses algab 2013. aastast.

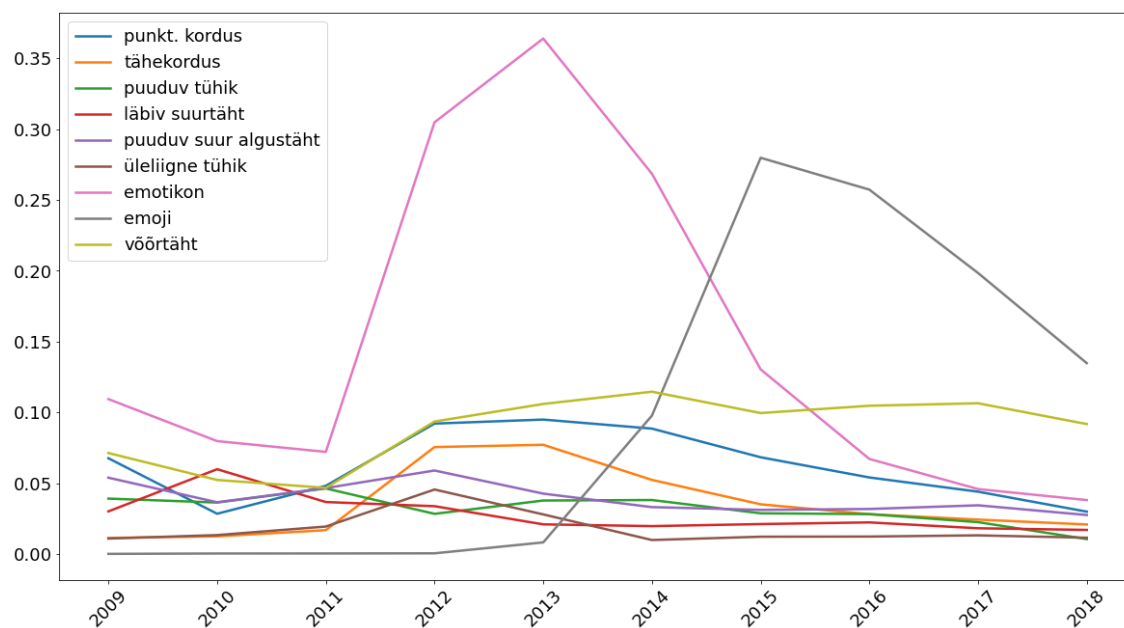
Kuna joonisel 7 on ajajoonel tunnuste koguarv, siis on tunnuste trendid mõjutatud säutsude koguarvust. Täpsema ülevaate saaks tunnuste suhtarvu uurides.



Joonis 7. Metagraafia tunnuste koguarvud aastate kaupa

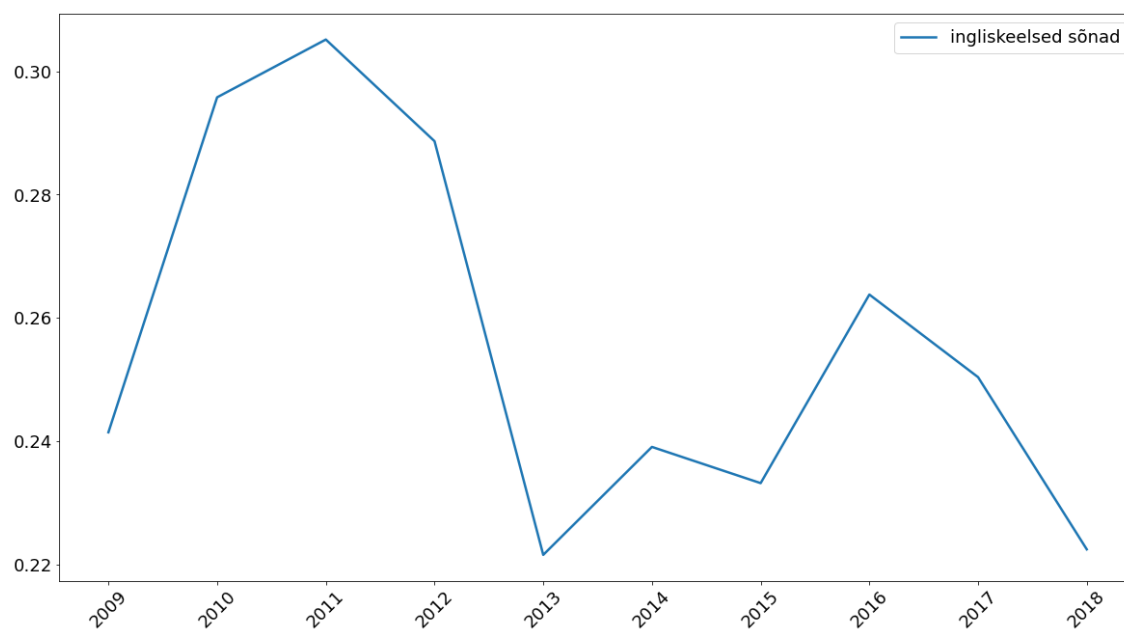
Joonisel 8 on välja toodud keskmine tunnuste arv säutsus aasta kaupa. Joonistub välja emotikonide kasutuse suur tõus, mis algab peale 2011. aastat, jõuab tippu 2013. aastal ning hakkab peale seda langema, kusjuures samal aastal algab peegelpildis *emoji*-de kasutuse tõus. 2014. ja 2015. aasta vahel on *emoj*-de ja emotikonide kasutus võrdne. 2015. aastast alates aga on *emoji*-d sagedasemad kui emotikonid, olles samal aastal ka oma sageduse tipus.

Vahemikus 2012–2014 on märgata internetikeelsuses tõusu: kõikide tunnuste sagedus kasvab, v.a läbiva suurtähe ning puuduvate tühikute sagedus. Peale 2014. aastat tunnuste sagedus kas vähenb või stabiliseerub.

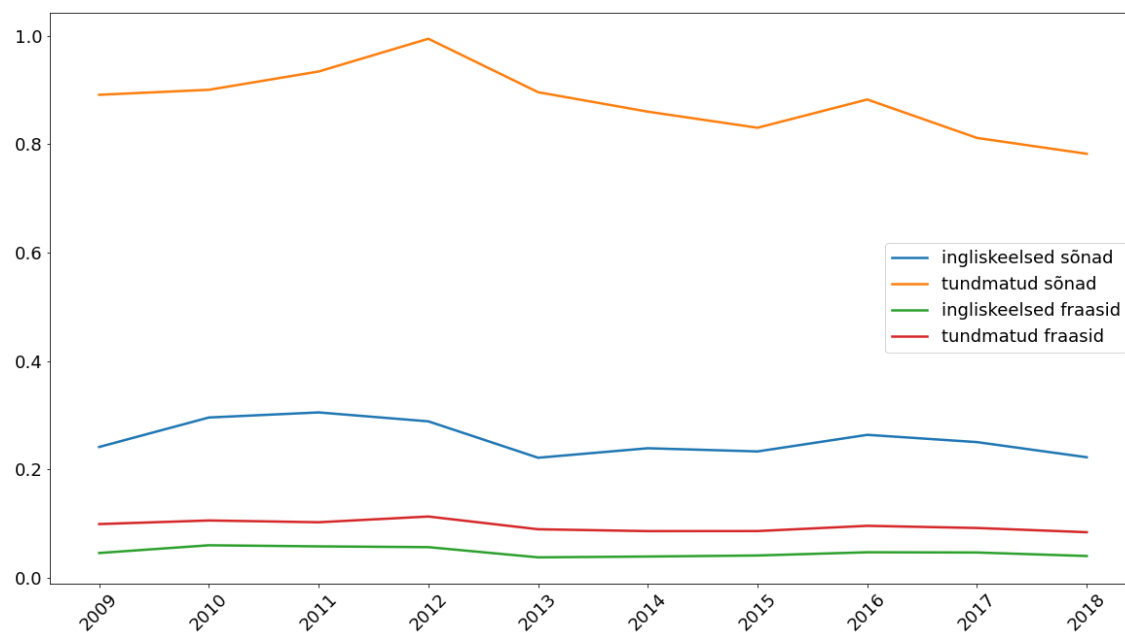


Joonis 8. Keskmise metagraafia tunnuste arv säutsus aastate kaupa

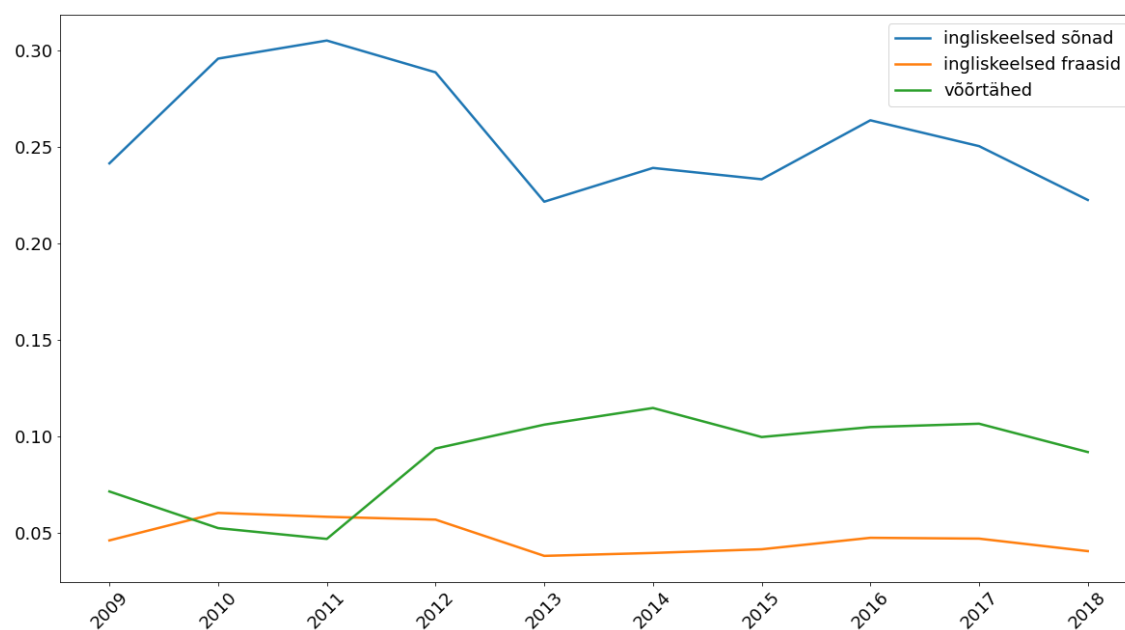
4.3.2. Leksikaalsed tunnused ajas



Joonis 9. Ingliskeelseid sõnu keskmiselt säutsus



Joonis 10. Inglisekeelseid sõnu ja ingliskeelseid fraase ning tundmatuid sõnu ja tundmatuid fraase keskmiselt säutsus



Joonis 11. Inglisekeelseid sõnu ja ingliskeelseid fraase ning võõrtähti keskmiselt säutsus

5. Arutelu

5.1. Põhilised tähelepanekud

Säutsude arv korpuse pole ajas ühtlane: mida aasta tagasi, seda vähem on säutse. See sõltub ilmselt säutsude kogumise metoodikast ning võib oletada, et Twitteri otsingu programmiliides annab päringu vastuseks hiljutisemad säutsud eespool.

Korpuses oli edasisäutsude kujul võrdlemisi palju kordusi. Pärast korduste eemaldamist langes edasisäutsude arv märkimisväärselt.

Kasutajate soolises jaotuses oli märgata naiste ja meeste koguarvudes kasvu, kusjuures korpuse algusaastatel oli meessoost kasutajaid rohkem, viimastel aastatel oli aga vastupidi. See aga läheb üldise Twitteri kasutuse statistikaga vastuollu, ka üldine statistika pole allikati ühtlane.

Metagraafia ja ortograafia tunnuste puhul on näha langustrendi alates 2015. aastast (vt ka Lisa 4). Joonistus välja emotkonide suureosaline vahetumine *emoji*'de vastu. Saab oletada, et *emoji*'de toetuse lisamine Twitteri veebilehele soodustas *emoji*'de kasutust.

Esile tuli ka erinev keelekasutus meeste ja naiste vahel. Järelduste tegemiseks pole piisavalt infot kasutajaskonna kohta.

Ingliskeelsete sõnade kasutussageduses on näha kõikumisi ning kuni 2013. aastani pole võõrtähtede ja ingliskeelsete sõnade kasutussagedused korrelatsioonis. Kõikumised on üsnagi väiksed.

Senised tulemused viitavad sellele, et Twitteri korpus pole ühtlane kogum / homogeenne. Tulemusi ei saa vaadata ühtselt, sest äärmused on üpriski erinevad (nt erinevus naiste ja meeste keelekasutuses).

5.2. Materjali poolt tõstatatud küsimused

Seoses tunnuste langustrendidega alates 2015. aastast tekib küsimus, kas netikeelsus on vähenenud peale 2015. aastat. Muutused tunnuste kasutussaguduses võivad olla keelest sõltumatud statistilised kõikumised, mistõttu tuleks kindlaks teha, kas uurimismaterjali on piisavalt.

Korpuse homogeensuse küsimuse tõttu võiks korpust võrrelda mõne teise eestikeelse Twitteri korpusega, mis eeldaks küll korpuse kogumist. Huvitav oleks Twitteri eesti netikeelsuse uurimine ajavahemikus 2018–2022. Kas pandeemia mõjutas Twitteri kasutuse dünaamikat ning keelt? Kas tunnuste näiline langustrend jätkub või on tunnuste trendid uues korpuses sootuks erinevad siinse töö korpusest?

Omaette küsimus on ka keele muutumine kasutaja tasandil (Kaalep 2017). Haiba (2016) järgi saaks defineerida aktiivse kasutaja, kellel on ühtlasi piisavalt postitamise staaži, ning uurida keelemuutust aktiivsete kasutajate alamhulga peal. Ehk muutuks korpus nii ka ühtlasemaks?

Kokkuvõte

Bakalaureusetöös oli töötlemiseks üks eestikeelseid säutse sisaldav Twitteri korpus ajavahemikust 2009–2018. Töö teoreetilises osas püstitati uurimisküsimust eelnevate netikeele uurimuste põhjal ning teoretiseeriti, kuidas ja miks võivad sarnased jooned avalduda Twitteri keelekasutuses. Heideti pilk Twitteri kasutamisesse, ajalukku tehnilise poole pealt ning Twitteri kui korpus uurimisse. Varasemate uurimuste põhjal selgitati välja siinses bakalaureusetöös vaadeldavad netikeele tunnused, mis koondusid netikeele metagraafia ja ortograafia ning leksikaalsete tunnuste alla.

Bakalaureusetöös anti ülevaade korpus eeltöötlemisest, säutsumise dünaamikast ning kasutajate jaotusest. Esitati joonised edasisäutsude, lühendatud säutsude ning kõikide säutsude kohta ajas. Leiti, et korpus oli võrdlemisi palju kordusi edasisäutsude kujul, mis eemaldati edasiseks tööks.

Töö praktilises osas esimeses pooles arendati netikeele tunnuste märgendajat, programmi aluseks võeti Kaljumäe 2019. aastal kirjutatud netikeele tunnuste märgendaja. Märgendatavateks tunnusteks osutusid punktuatsioonikordused, tähekordused, puuduvad tühikud, läbivad suurtähed, puuduvad suured algustähed, üleliigsed tühikud, emotikonid, *emoji*'d ja võõrtähed. Praktilise osa esimese poole tulemusena valmis netikeele tunnuste märgendaja ning vastvalminud märgendajaga märgendatud korpus, kus on tunnused eraldi kihis.

Praktilise osa teises pooles keskenduti leksikaalsele analüüsile. Korpuseteostati EstNLTK morfoloogiline analüüs ilma oletamise ja ühestamiseta. Kirjutati programm tundmatuteks jäänud sõnade ingliskeele tuvastamiseks kasutajasõnastikust ja fraaside keeletuvastuseks *langid* programmi abil. Dage Säre 2015. aasta magistr töö korpus põhjal koostati ka üks partiklitest ja lühenditest koosnev kasutajasõnastik ning märgendati ka need. Praktilise osa teise poole tulemusena valmis tundmatuks jäänud sõnade märgendatud korpus, tundmatud sõnad märgendati kahte uude kihti, fraasid ja sõnad eraldi.

Bakalaureusetöö praktilises osas märgendatud tunnused visualiseeriti ajajoonel. Märgati emotikonide kasutuse suurosalist vahetust *emoji*'de vastu ning oletati, et see on Twitteri veebiliidese uuendusest tingitud. Ortograafia-metagraafia joonisel täheldati tunnuste vähenemist, kuid järeltusi ei tehtud, sest korpus pole ühtlane. Selgus, et äärmused kõikusid seinast seinast: meeste ja naiste keelekasutuse erinevus. Kuvati leksikaalsed tunnused joonistel, kuid ka siin ei hakatud järeltusi tegema.

Viimaks tehti ettepanekuid edasiseks uurimiseks, nt kas netikeelsus on langenud alates 2015. aastast. Lisaks uurida Haiba (2016) järgi aktiivse kasutajagrupi keelekasutust.

Kirjandus

Anić, Ines 2017. The history of hashtags. – Wallery;
<https://walleryapp.com/2017/08/08/the-history-of-hashtags/>. Vaadatud 29.05.2022;

Bai, Qiyu; Dan, Qi; Mu, Zhe; Yang, Maokun 2019. A Systematic Review of Emoji: Current Research and Future Perspectives. – Frontiers in Psychology;
<https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02221/full>. Vaadatud 02.06.2022;

Carlson, Nicholas 2010. Twitter Announces "Official BlackBerry App," Kills First Third-Party App. – Insider;
<https://www.businessinsider.com/twitter-announces-official-blackberry-app-2010-4>. Vaadatud 30.05.2022;

Casper, Heather 2017. Everything You Need To Know About Twitter's Timeline Algorithm. – ClickToTweet;
<https://clicktotweet.com/blog/everything-you-need-to-know-about-twitter-timeline-algorithm/>. Vaadatud 30.05.2022;

Elo, Kimmo 2019. Is Twitter just a Media Amplifier? Evidence from a data-driven, comparative Analysis of German Federal Elections 2017 and European Elections 2019; <https://halshs.archives-ouvertes.fr/halshs-02425856>. Vaadatud 29.05.2022;

Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2020. Eesti keele käsiraamat. Tallinn: Eesti Keele Instituut, EKSA

Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina, Saari, Henn; Tael, Kaja; Vare, Silvi 1993. Eesti keele grammatika II. Süntaks. Lisa: kiri. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut;

EstNLP.
https://nbviewer.org/github/d009/EstNLP/blob/master/estnltk_baastarkused_sisukorrage.ipynb. Vaadatud 02.06.2022

- Haiba, Sabiina 2016.** Kuidas on netikeel muutunud aastatel 2001–2008. Bakalaureusetöö. Tartu
- Hennoste, Tiit 2013.** Tiit Hennoste: kuule ma eemale nüüd. – Postimees. Arvamus 7. dets.;
- Kaalep, Heiki-Jaan 2017.** Netikeele kirjakeelestumine. – Sirp 15. dets.;
- Käle, Maija; Rikters, Matīss 2021.** Fragmented and Valuable: Following Sentiment Changes in Food Tweets; <https://arxiv.org/abs/2106.04903>. Vaadatud 29.05.2022;
- Kemp, Simon 2019.** Digital 2019: Estonia. – DataReportal; <https://datareportal.com/reports/digital-2019-estonia>. Vaadatud 28.05.2022;
- Kemp, Simon 2022.** Digital 2022: Estonia. – DataReportal; <https://datareportal.com/reports/digital-2022-estonia>. Vaadatud 28.05.2022;
- Ljubešić, Nikola; Fišer, Darja; Erjavec, Tomaž 2014.** TweetCaT: a tool for building Twitter corpora of smaller languages – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Ed. by -. Reykjavik, Iceland: European Language Resources Association (ELRA), 2279–2283;
- Moscaritolo, Angela 2014.** Twitter Adds Emoji Support for the Web. – PC Magazine; <https://www.pcmag.com/news/twitter-adds-emoji-support-for-the-web>. Vaadatud 29.05.2022;
- KM jt = Muischnek, Kadri; Kaalep, Heiki-Jaan; Sirel, Raul 2011.** Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. – Eesti Rakenduslingvistika Ühingu aastaraamat 7. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Rakenduslingvistika Ühing, 111–127;
- Oja, Anni 2010.** Sissevaateid internetisuhtluse. – Oma Keel 20, 11–18;
- Paßmann, Johannes 2019.** Less Mutable, More Mobile: The Role of Twitter Apps in the History of the Retweet Button. – Computational Culture 7;

<http://computationalculture.net/less-mutable-more-mobile-the-role-of-twitter-apps-in-the-history-of-the-retweet-button/>. Vaadatud 29.05.2022;

Perez, Sarah 2017. Twitter officially expands its character count to 280 starting today. – TechCrunch;
<https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/>. Vaadatud 02.06.2022;

Rosen, Aliza; Ihara, Ikuhiro 2017. Giving you more characters to express yourself. – Twitter Blog;
https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself. Vaadatud 29.05.2022;

Santos, José 2011. Säutsude taaspostitamise käitumise analüüs kasutades teemamudeleid; <https://dspace.ut.ee/handle/10062/33036>. Vaadatud 29.05.2022;

Scott, Jennifer 2010. Twitter launches official Android app. – IT Pro;
<https://www.itpro.com/622936/twitter-launches-official-android-app>. Vaadatud 30.05.2022;

Seward, Zachary M. 2013. The first-ever hashtag, @-reply and retweet, as Twitter users invented them. – Quartz;
<https://qz.com/135149/the-first-ever-hashtag-reply-and-retweet-as-twitter-users-invented-them/>. Vaadatud 29.05.2022;

Stone, Christopher Isaac 2007. Are You Twittering @ Me? – Twitter Blog;
https://blog.twitter.com/official/en_us/a/2007/are-you-twittering-me.html. Vaadatud 29.05.2022;

Summers, Nick 2014. Twitter adds native animated GIF support to Twitter.com, Android and iPhone. – The Next Web;
<https://thenextweb.com/news/twitter-adds-native-animated-gif-support-twitter-com-android-iphone>. Vaadatud 31.05.2022;

Toots, Kristiina 2014. Netikeele metagraafia. Bakalaureusetöö. Tartu;

Independent 2011. Twitter introduces native photo uploading, plans on adding Twitter user photo galleries. – Independent; <https://www.independent.co.uk/tech/twitter-introduces-native-photo-uploading-plans-on-adding-twitter-user-photo-galleries-2335293.html>. Vaadatud 29.05.2022;

Twitter Help Center; <https://help.twitter.com/en>. Vaadatud 29.05.2022.

Whitney, Lance 2010. Twitter debuts official iPhone app. – CNET; <https://www.cnet.com/tech/mobile/twitter-debuts-official-iphone-app/>. Vaadatud 30.05.2022;

Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2020. Eesti keele käsiraamat. Tallinn: Eesti Keele Instituut EKSA

<https://digi.geenius.ee/rubriik/uudis/150-000-eesti-inimest-kaib-iga-kuu-twitteris/>

Twitter Developer Platform 2022. <https://developer.twitter.com/en>. Vaadatud 02.06.2022;

Twitter Developer Policy 2022. <https://developer.twitter.com/en/developer-terms/policy#id8>. Vaadatud 02.06.2022;

Britannica, T. Editors of Encyclopaedia 2022. Twitter. – Encyclopedia Britannica; <https://www.britannica.com/topic/Twitter>. Vaadatud 02.06.2022;

TWT - Dage säre kogutud säutsude korpus, mida siinses töös kasutan.

Twitter corpus for analyzing language changing in time

Summary

This bachelor's thesis tries to find out if a Twitter text corpus (1.6 million tweets from years 2009 to 2018) in Estonian can be used for corpus-based Estonian internet language analysis. The main objectives of this thesis are the following: to find out, what kind of attributes to look for; to write a programming tool for tagging said attributes, and to prepare the corpus for future analysis.

In the theoretical part of the thesis, overall Estonian internet language is described, past internet language analyses are described, and attributes are chosen based on them. Also, there is a description of using Twitter and a brief history of Twitter from a technical point of view. There is also an overview of tweeting dynamics in the corpus, distribution of tweets (retweets, truncated tweets, and total tweets), and users over the corpus time period. Some preprocessing is done, including removing repetitions in the form of the same retweets from the corpus, leaving the size of the corpus to 1.1 million tweets.

In the first section of the practical part of the thesis, a program for tagging orthography and punctuation attributes is developed based on a past internet language tagging project²¹. For example, the attributes are punctuation repetitions, use of uppercase and lowercase, use of emoticons, use of emojis, etc. The program is written in Python using the EstNLTK toolkit²². The attributes are tagged into layers. In the second section of the practical part of the thesis, a morphology analysis without disambiguation and guessing is done. Unknown words and phrases' language is analyzed with the use of a user dictionary and *langid* tool²³. Also, based on a user dictionary, words are tagged into particles and abbreviations. User dictionary is based on Dage Särg's (2015) master thesis' corpora.

²¹ https://github.com/estnltk/ettenten-experiments/tree/master/weblang_detection

²² <https://github.com/estnltk/estnltk>

²³ <https://github.com/saffsd/langid.py>

The attributes' frequencies are visualized and starting from 2015, a decrease in the frequencies is noticed. Also, the frequencies of emojis and emoticons switch places, which is assumed to be because of a Twitter web page update supporting emojis. Any general conclusions aren't made because the corpus isn't homogenous. Finally, questions for future work are formulated. For example, has the internet language changed starting from year 2015?

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, _____ Martin Vainikko _____,

(autori nimi)

1. _____ annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

_____ Twitteri korpus ajas muutuva keele uurimiseks _____,

(lõputöö pealkiri)

mille juhendajad on _____ Siim Orasmaa ja _____ Heiki-Jaan Kaalep _____

(juhendaja nimi)

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 Commons'i 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

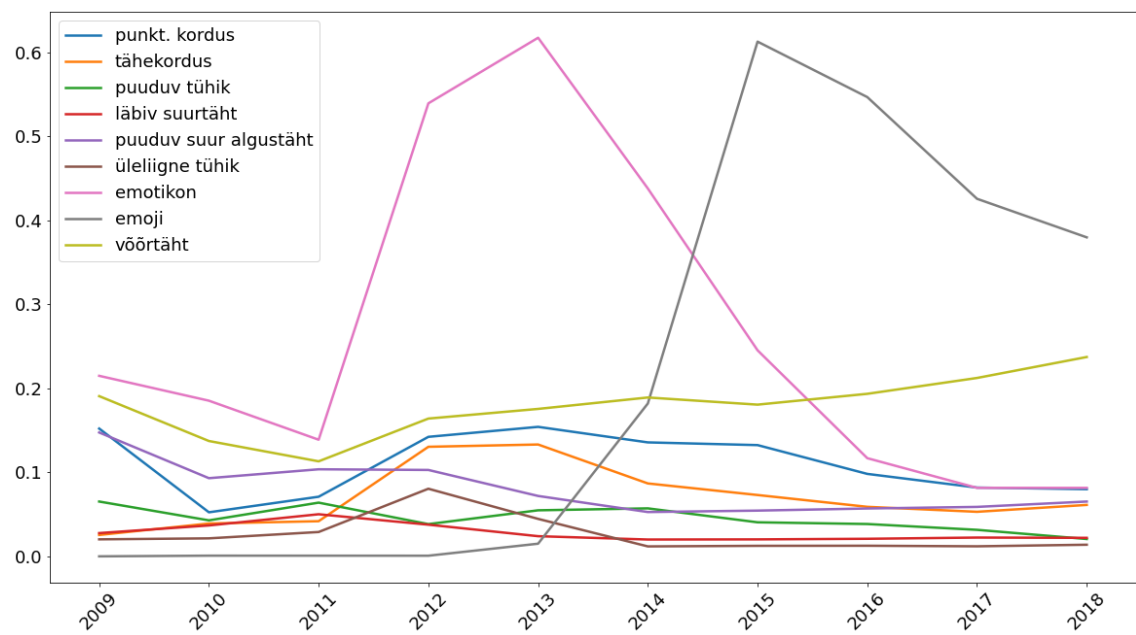
intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Martin Vainikko

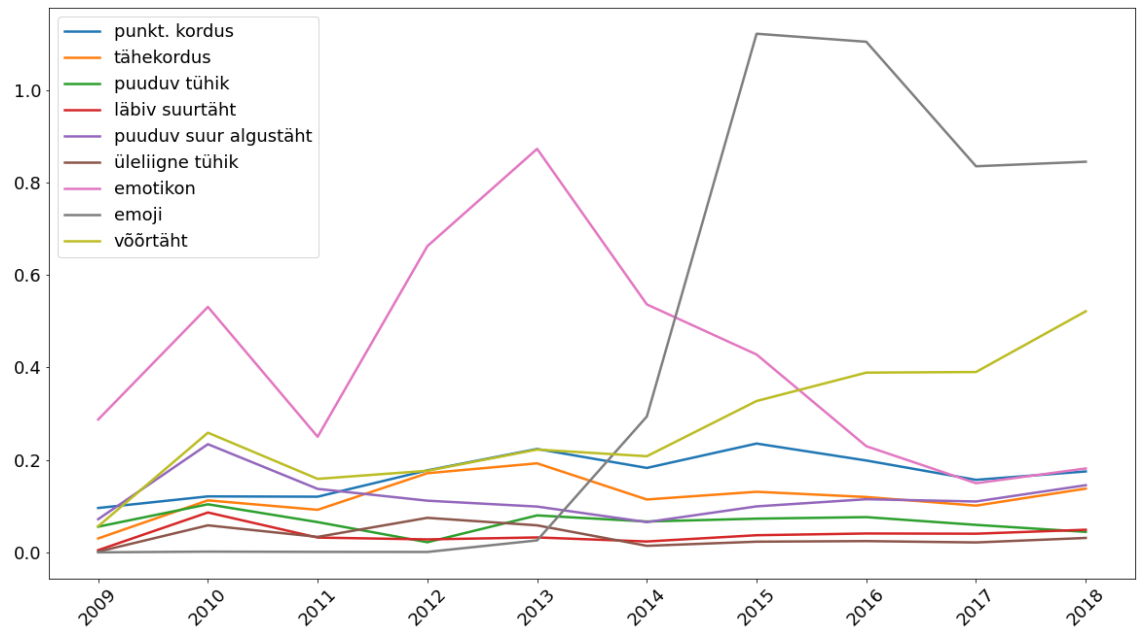
02.06.2022

Lisad

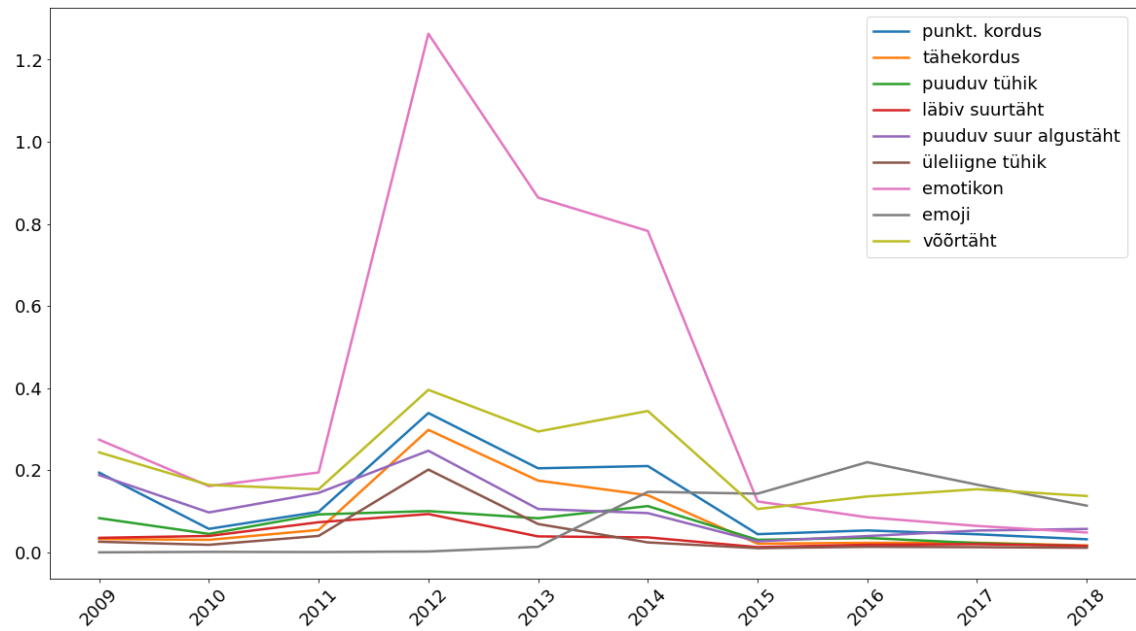
Lisa 1. Netikeele tunnuseid säutsu kohta ajas – isikud



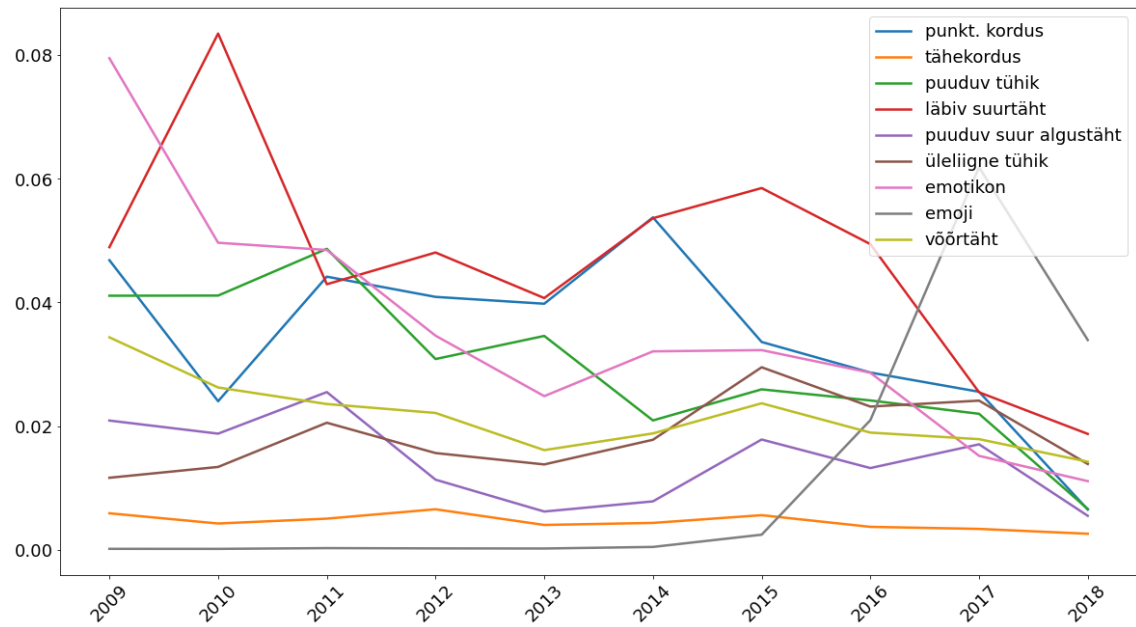
Lisa 2. Netikeele tunnuseid säutsu kohta ajas – naised



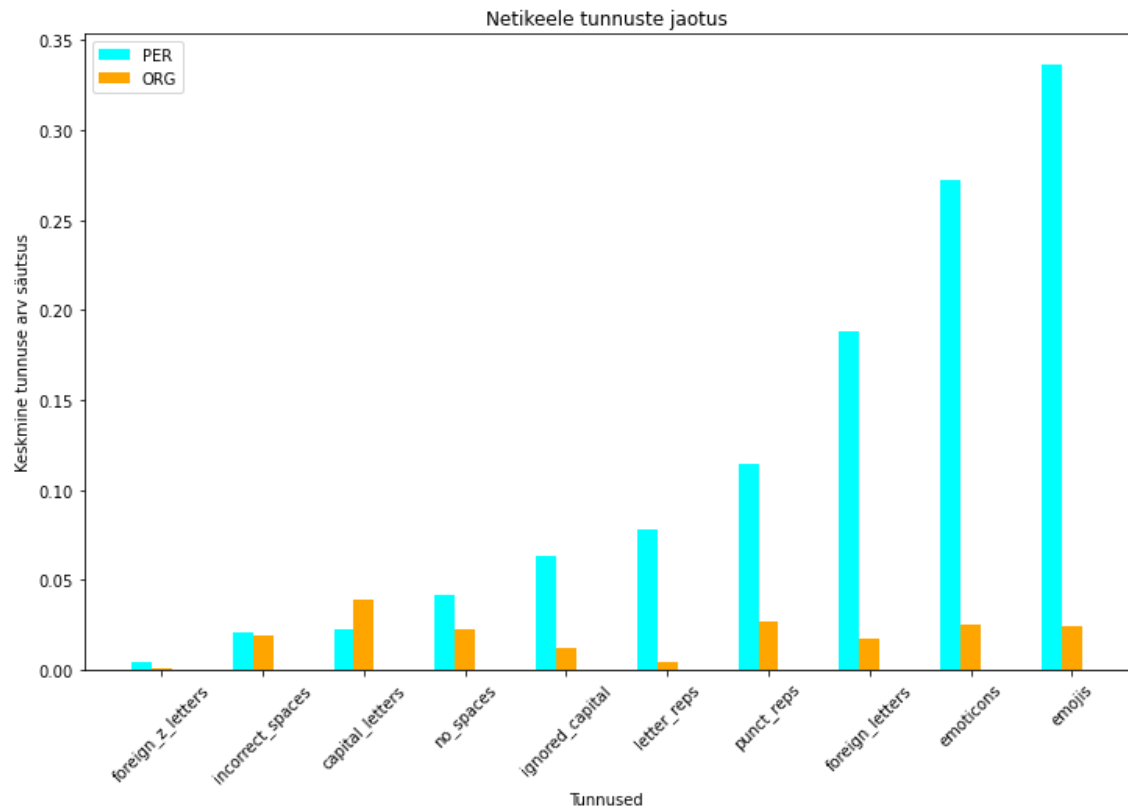
Lisa 3. Netikeele tunnuseid säutsu kohta ajas – mehed



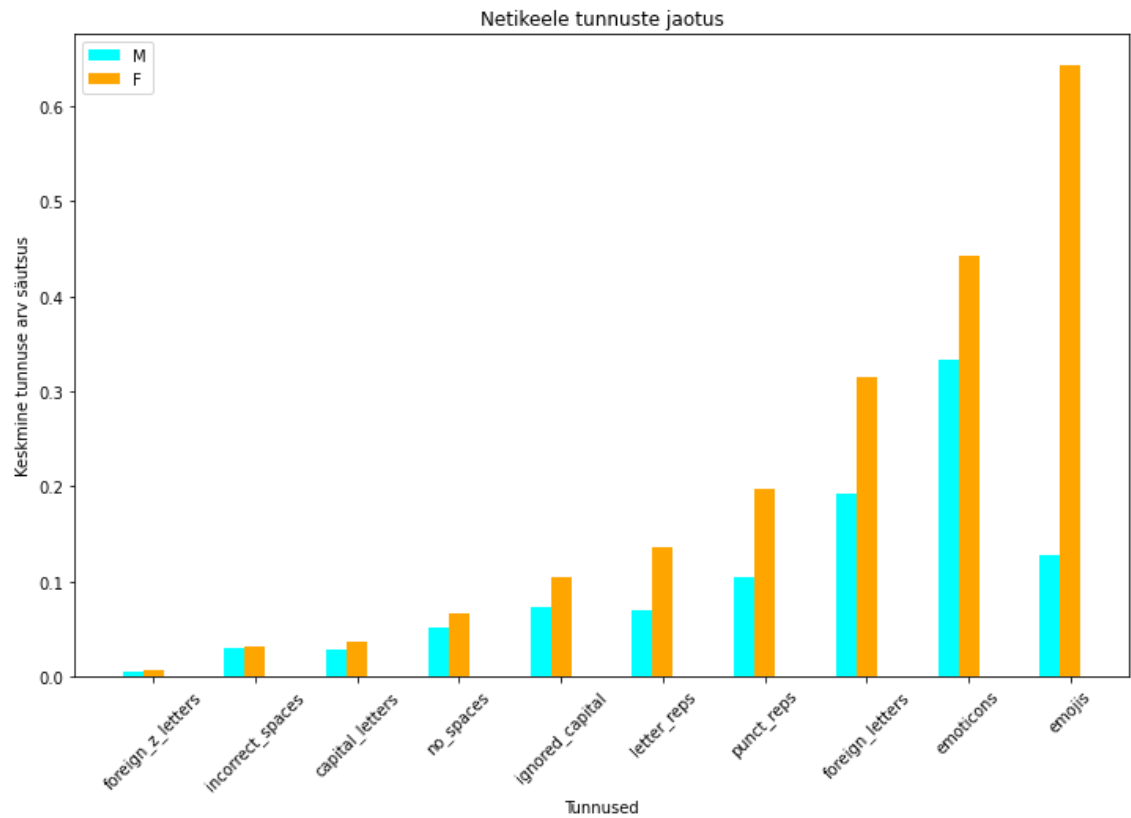
Lisa 4. Netikeele tunnuseid säutsu kohta ajas – organisatsioonid



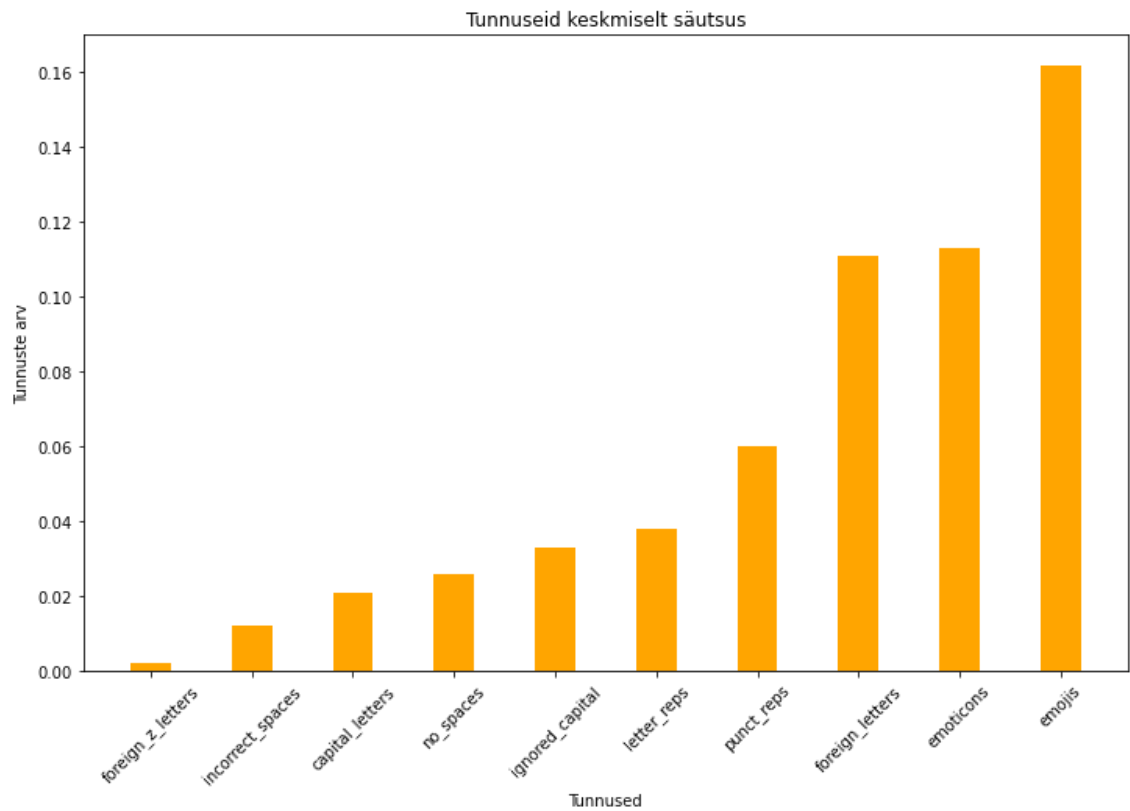
Lisa 5. Tunnuseid keskmiselt säutsus – isikud vs organisatsioonid



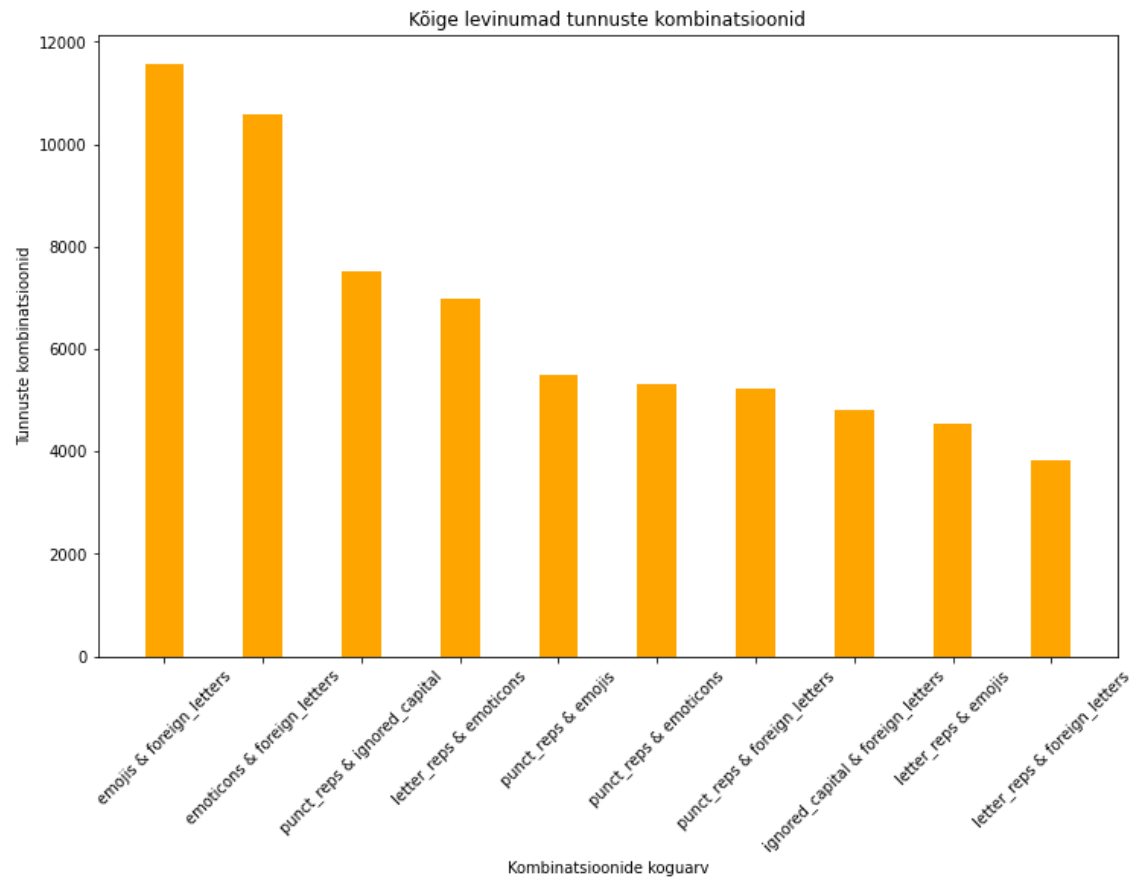
Lisa 6. Tunnuseid keskmiselt säutsus – mehed vs naised



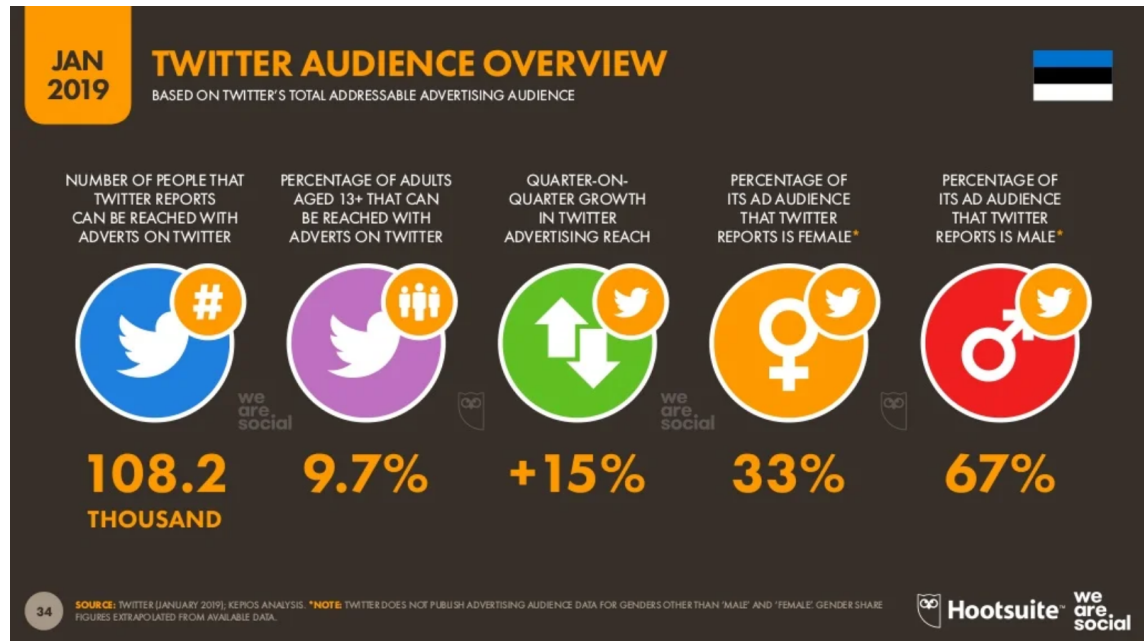
Lisa 7. Tunnuseid keskmiselt säutsus (enne säutsude korduste eemaldamist)



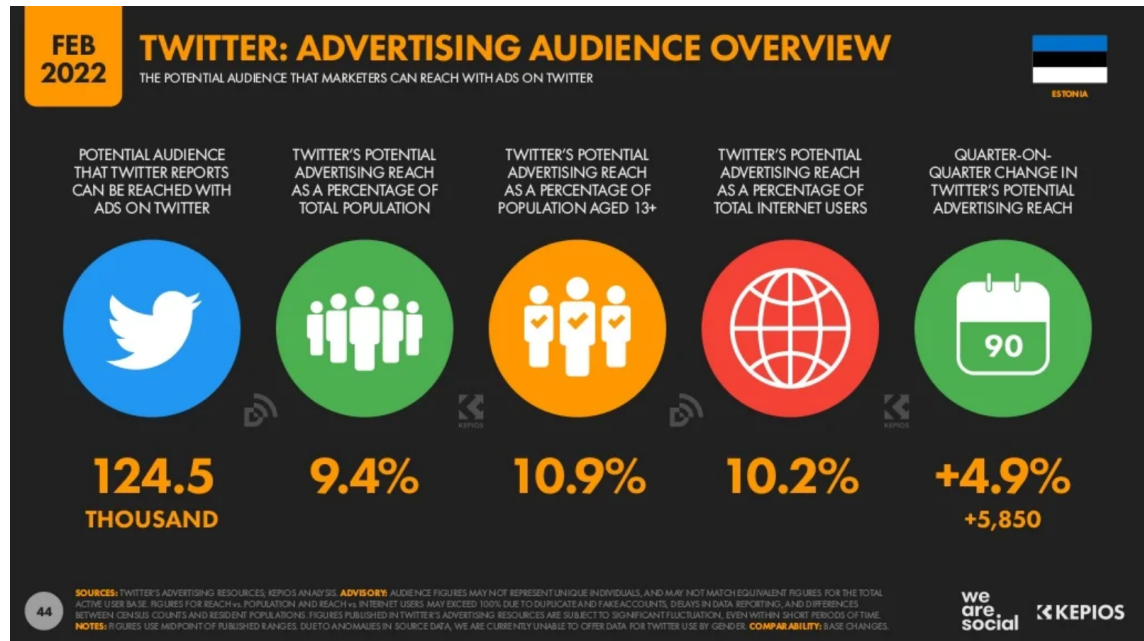
Lisa 8. Tunnuste kombinatsioonide kasutussagedused (enne säutsude korduste eemaldamist)



Lisa 9. (Kemp 2019)



Lisa 10. (Kemp 2022)



Lisa 11. Partikleid ja lühendeid keskmiselt säutsus

