Tartu Ülikool
Filosoofiateaduskond

Kristjan Link

UUE MEEDIA TEKSTIDE SAGEDUSSÕNASTIK JA SELLE VÕRDLUS KIRJAKEELE SAGEDUSSÕNASTIKUGA

Bakalaureusetöö

Juhendaja: Kadri Muischnek

Tartu
2014

Sisukord

Sissejuhatus

Uue meedia keel ehk internetikeel ehk netikeel on internetis kasutatav keelevariant ja ta ei ole ühtne, vaid koosneb allkeeltest. Ühtset äratuntavat netikeelt ei ole olemas, küll on aga lugematul hulgal keelevariante, mis sõltuvad žanrist, olukorrast, tehnilistest kitsendustest, eesmärgist ja veel hulgast teguritest. (Oja 2010) Netikeele uurimine ja töötlemine on arvutilingvistikas populaarne teema. Uuritakse ka tehnoloogiat, mille abil omavahel suheldakse. (Baron 2008)

Sõnade sageduse uurimisel on mitmeid põhjusi. Näiteks on sellest kasu keele õppimisel. Seda on vaja ka keeletehnoloogias (keeletuvastus ja masintõlge). Samuti võib sagedusloendites sisalduv info olla ka niisama huvitav. Näiteks Briti rahvuslikus korpuses on sõna *man* kaks korda sagedam kui sõna *woman,* aga sõna *woman* mitmus *women* on sagedam kui sõna *man* mitmus *men*. (Leech jt 2001)

Selle töö eesmärk on koostada uue meedia keele sõnavormide sagedusloendid ja võrrelda neid sagedusloendeid normeeritud kirjakeele baasil koostatud sarnaste loenditega.

Töö on üles ehitatud järgmiselt. Osas 1 tutvustatakse töö teoreetilisi lähtekohti: osas 1.1 käsitletakse Zipfi seadust, osas 1.2 antakse ülevaade eesti keele varasematest sagedussõnastikest, osas 1.3 tutvustatakse uue meedia erijooni, osas 1.4 vaadatakse varasemaid eesti internetikeele uurimusi. Osas 2 kirjeldatakse materjali. Osas 3 seletatakse töö käiku. Osast 4 alustab tulemuste kvantitatiivne analüüs: kirjeldatakse tabeleid ning osas 4.1 on *Venni* diagrammid korpuste sagedusloendite kattuvuse kohta. Osast 5 alustab tulemuste kvalitatiivne analüüs: osas 5.1 uuritakse uue meedia ja kirjakeele sagedusloendite ühise sõnavara sagedaimat osa, mis on järjestatud uue meedia põhjal ja analüüsitakse neid sageduse järjekorranumbri alusel, osas 5.2 uuritakse uue meedia ja kirjakeele sagedusloendite ühise sõnavara sagedaimat osa, mis on järjestatud uue meedia põhjal ja vaadatakse sõnavorme sõnaliikide kaupa ning uue meedia korpuses sagedamini esinevaid sõnavorme, osas 5.3 analüüsitakse vaid uue meedia tekstides esinenud sõnavorme.

1. Töö teoreetilised lähtekohad

Selles osas antakse ülevaade erinevatest uurimustest, mis on mingil määral seotud antud tööga. Zipfi seadus näitab seost sageduste vahel, eesti keele varasemate sagedussõnastike osas antakse ülevaade sagedussõnastike uurimise ajaloost, uue meedia keele erijoonte osa iseloomustab internetikeelt.

1.1 Zipfi seadus

Üks peamisi statistilisi seaduspärasusi, mis on kindlaks tehtud loomulike keelte tekstide põhjal koostatud sagedussõnastike analüüsimisel, on funktsionaalne seos sõnavormi sageduse ja astaku vahel, kusjuures astaku all mõeldakse sõna järjekorranumbrit sageduste kahanevas reas. Nime sai seadus Harvardi keeleteaduste professori George Kingsley Zipfi järgi, kes formuleeris lõplikult funktsionaalse seose sõnade esinemissageduse ja järjekorranumbri vahel. Ta leidis, et sõna absoluut- või suhteline sagedus ning vastav astak sagedussõnastikus on seotud küllalt pika teksti puhul sõltuvusega, mida on hakatud nimetama Zipfi seaduseks.

Joonis 1. Zipfi seadus.

$$F_r = \frac{C}{r^{\gamma}}$$

Nagu näha jooniselt 1, võrdub sõna esinemissagedus F_r konstandi C ja astaku r jagatisega, kusjuures astakut on astendatud konstandiga ɣ, mille väärtus on 1 ümber. Esinemissageduse astak r peab olema sama, mis jagajal. (Tuldava 1977)

Zipfi seadus ei pruugi alati tõele vastata. Mõnes sagedusloendis võib olla rohkem sageli korduvaid sõnu kui teises. See tuleb esile eriti siis, kui sagedusloendist on parema analüüsi huvides mingid sõnavormid välja jäetud. Väike variatsioon on alati, kuna mitte kunagi ei kirjutata identseid tekste. Igal inimesel on omamoodi keel, mis võib mõjutada sagedussõnastike tulemusi. Üldjoontes peab Zipfi seadus siiski paika, kuna kui teha jooniseid, on võimalik näha, et tulemused ongi sellised, nagu seadus kirjeldab. Mida rohkem on sagedusloendis sõnu, seda rohkem peab Zipfi seadus paika ja seda täpsemad on ka joonised.

1.2 Eesti keele varasemad sagedussõnastikud

Eestis hakati sõnavarastatistikaga tegelema 1960. aastatel, kui Juhan Tuldava juhtimisel moodustati keelestatistika uurimisrühm (Kasik 2011). 2001. aastal koostati Eesti Keele Instituudis terviklikku tekstikorpust kattev grammatiline sagedussõnastik „Seadusetekstide grammatiline sagedussõnastik“. Sagedussõnastiku lähtematerjaliks oli kümnest seadusest koosnev tekstikorpus: kõik sõnavormid analüüsiti kõigepealt morfoloogiliselt, seejärel ühestati mitmesed tulemused. (Viks jt 2001) Aastal 2002 koostati eesti keele sagedussõnaraamat „Eesti kirjakeele sagedussõnastik“ (Kaalep jt 2002), mis on 1990. aastate ilukirjanduse ja ajakirjanduse 1 miljoni sõna suuruse korpuse põhjal peaaegu täisautomaatselt

arvutatud lemmatiseeritud sõnade sagedusloend. Mõlema korpuses kasutatud tekstiklassi maht oli ümmarguselt pool miljonit sõna. Ilukirjanduse tekstidena kasutati eesti keele 90. aastate ilukirjanduse allkorpuse tekste.

1.3 Uue meedia keele erijooni

Selles bakalaureusetöös uuritakse uue meedia keele ehk internetikeele sõnavara. Internet võimaldab peaaegu igasugust kommunikatsiooni: vahetada saab nii kirjalikke tekste kui ka suulist informatsiooni; suhelda saavad kaks inimest, üks inimene mitme teise inimesega, grupp inimesi mingi teise inimgrupiga; suhtlus võib toimuda kas reaalajas või nii, et adressaat loeb edastatut hiljem; saata võib dokumente jne. (Cantoni 2006)

Võrgusuhtlusel on palju jooni, mis sarnanevad suulise ja kirjaliku keelega ja ka jooni, mis erinevad nendest (Cantoni 2006). *Kõne* ja *kiri*, mida inimesed senini on harjunud lahus hoidma, sulanduvad küberruumis üheks (Gurak 2001), paljud ootused ja reeglid, mis inimestele suulise ja kirjaliku keelekasutusega seostuvad (nagu kõnesituatsiooni vahetus *vs* kirjutaja ja lugeja eraldatus ajas, kõnelemisega kaasnevad keelevälised vihjed (miimika, žestid jm) ja tagasiside kuulajalt *vs* keeleväliste vihjete ja lugeja vahetu tagasiside puudumine kirjas, kõne planeerimatus *vs* teksti redigeeritus jt), ei kehti enam, sest internetisuhtlus kasutab mõlemale suhtlusviisile iseloomulikke jooni (Crystal 2001).

Nii arvataksegi, et internetikeele mõju üksikkeeltele on pigem halb kui hea. Kõige rohkem kardetakse, et keeled kaotavad mitmekesisuse ja segunevad inglise keelega (Crystal 2001). Internetikeel muudab ka suhtlust ebaviisakamaks, kuna lühikesed sõnumid ei saagi sisaldada viisakust. Probleemne on ka õigekiri, mida internetikeeles ignoreeritakse. Uuringud näitavad seesuguseid tendentse lisaks inglise keelele ka teiste keelte, nt kreeka keele puhul. (Soodla 2010)

1.4 Varasemad eesti internetikeele uurimused

1.4.1 Eesti keel internetis ja sissevaateid internetisuhtlusesse

Eestikeelset internetisuhtlust on analüüsinud Anni Oja (2006, 2010).

Arvutisuhtluses on keeleuurija jaoks olulised aeg, vestlejate arv ja nende omavaheline suhe. Aja poolest jaguneb suhtlus kaheks: sünkroonne ja asünkroonne.

Sünkroonse puhul suhtlevad osapooled ühel ajal. Sünkroonsed suhtluskeskkonnad on jututoad, kus suhtlevad paljud inimesed korraga reaalajas. Rollimängudes kasutatakse jututubade võimalusi ja fantaasiamaailma seiklusi. Paarisuhtlusprogrammid on tehtud privaatse suhtluse tarvis. Nendega on võimalik ka vahetada faile ja kutsuda vestlusesse rohkem osalejaid. Internetitelefon ja videokonverentsid täidavad telefoni funktsiooni. (Oja 2006)

Asünkroonse puhul ei pruugi partnerid olla samal ajal arvutis ja vastatakse siis kui võimalik. Sünkroonse suhtluse puhul on pigem oluline kiirus, asünkroonse puhul õigekiri. Tavaliselt lühendatakse levinumaid sõnu ja käibefraase, suurtähed ja kirjavahemärgid jäetakse ära. Trükivigu ignoreeritakse. Asünkroonsed suhtluskeskkonnad on portaalid, kus pakutakse kasutajale mitmeid suhtlusvõimalusi. Foorumites pakutakse võimalust arutleda teatud teemadel. Tekst on kirjakeelsem ja suhtlus asjalik. Postiloendid ja uudisgrupid sarnanevad foorumitele, aga suhtlus käib posti teel. Postiloend koondab huviliste meiliaadresse ja kõik postiloendi liikmed saavad loendisse saadetud kirjad. Anonüümsust on siin vähem. Kommentaariumides on teksti juurde võimalik lisada oma arvamus. See on anonüümne ja seetõttu kommenteeritakse jällegi julgemalt. Vikipeediad on kollektiivselt koostatavad vabad võrguentsüklopeediad. E-post on asendus traditsioonilisele kirjasaatmisele. Ajaveebid ehk blogid on avalikult peetavad päevikud. Nad võivad koosneda fotodest ja videotest. Isiklikud veebilehed on monoloogilised ja staatilisema sisuga. Seal ei oodata tagasisidet. (Oja 2006)

Internetis kasutatakse emotikone ja sümbolpilte, mis koosnevad mitmest reast. Internetisuhtluses on esindatud monoloog, dialoog ja ka multiloog. Monoloogis kirjutab inimene ja ei oota vastust, dialoogis suhtlevad kaks inimest omavahel ning multiloogis suhtleb mitu inimest omavahel korraga. Need kolm kattuvad tihti omavahel. Kuna internetis on võimalik jääda anonüümseks, on kommentaarid sageli avameelsemad ja solvavamad kui kohtades, kus nõutakse identifitseerimist. Samas võivad tekkida eristatavad kasutajad, kes suudavad oma identiteeti esile tuua. (Oja 2006)

Keel on internetis saanud omapärase vormi. Interneti teel on lihtsam suhelda kui teiste meediakanalite kaudu. Internet ei moonuta keelt, vaid on lihtsalt tema vahendamise kanal. Internetisuhtlus jaguneb viieks:

1) Üldmeedia – ajakirjandus internetis
2) Personaalmeedia – blogid ja Twitter
3) Otsesuhtlus – Skype jms

4) Grupisuhtlus – foorumid, moodle jms

5) Veebipõhised suhtlusvõrgustikud – Facebook jms

Piirid nende vahel on kohati hägused. (Oja 2010)

Netisuhtlust mõjutavad paljud tegurid, millest tähtsaimad on sünkroonsus, asünkroonsus, anonüümsuse määr, tehnilised võimalused, suhe avalikkuse ja privaatsuse vahel ning sotsiaalne taust ja suhtlusvõimekus. Sünkroonse suhtluse puhul on inimesed ühel ajal internetis, asünkroonse puhul mitte (e-post jms). Sünkroonse suhtluse puhul on vähe aega ja kasutatakse lühendeid, asünkroonse suhtluse puhul on rohkem aega ja ta meenutab rohkem kirjakeelt. Anonüümsuse määr ja avalikkuse/privaatsuse suhe mõjutavad seda, kui vabalt suhtleja end suheldes tunneb. Kui kõik on anonüümsed, räägitakse tabuteemadel vabamalt. Tehnilised võimalused on seotud suhtluseks kasutatud tehnika, interneti kiiruse ja suhtluskeskkonnaga. Tehniline varustatus arvutil määrab selle, kas saab pidada audio- ja videokõnesid. Võimalik on ka see, et klaviatuuril on mõned tähed puudu ja need tuleb asendada. Suhtluskeskkonnast sõltub mida seal kasutada saab, nagu emotikonid, videokõned jms. (Oja 2010)

Internetis on esindatud teemad, mida päris elus pole. Näiteks mingis kultuuris ei räägita mingist asjast, aga internet võimaldab seda. Veel võib olla takistus geograafiline, st inimesed asuvad üksteisest kaugel. Ainus eeltingimus on see, et suhtlevad inimesed peavad rääkima ühist keelt. Internetis saab rääkida teemadel, mida päris elus ei julgeks. Sellele aitab kaasa anonüümseks jäämise võimalus. Interneti murekohaks on piiride hägustumine (inimesed ei tea enam kus missugust keelt kasutada sobib) ja selle tagajärjel keelepiiride hägustumine. (Oja 2010)

Internetis on murekohaks see, et inimesed arvavad, et nad ei pea vastutama oma tegude eest internetis. Arvatakse, et internetti pandud tekste, pilte ja videoid niikuinii keegi ei vaata. Postitaja isiku saab enamasti lihtsalt kindlaks teha. (Oja 2010)

1.4.2 Morfoloogilisi, morfosüntaktilisi ja sõnamoodustuslikke nähtusi eesti internetikeeles

See lõputöö on tehtud kogu internetikeele kohta, seepärast ei saa teda võrrelda eraldi uue meedia tekstiklassidega.

Karin Soodla (2010) on uurinud eesti internetikeelt eelkõige keelenormist kõrvalekaldumise seisukohalt. Tema andmetel paistab internetikeeles silma mitmesuguste ühendite kokkukirjutamine. Kõige sagedamini kirjutatakse kokku kaassõnaühendeid. Kaassõnade kujunemise allikaks on tihti nimisõnad ja tegusõnad. Kõige sagedamini kirjutati kokku nimisõna ja kaassõna. Kõige enam kirjutati eelnev sõna kokku kaassõnaga *peal*. (Soodla 2010)

Üksikühendid kirjutatakse kokku tähenduse muutumise tõttu. Kui tähendus on metafoorne, siis esineb ka kirjapildis kokkukirjutamine. Näiteks *elusees* on omandanud määruselise tähenduse *iialgi*. Omadussõna ja nimisõna kokkukirjutamist illustreerivad ühendid *uuekooli* ja *vanakooli*, mis kirjutati 92,8% kasutusjuhtudest kokku. Kõige sagedamini kokkukirjutatav asesõna ja nimisõna ühend on *koguaeg*. Üksikjuhtumid on kahe nimisõna kokkukirjutamisel tekkinud ühendid ja määrsõnadega liituv nimisõna *kord*. Kordagi ei kirjutatud lahku liitnimisõna *suvalaks*. Ka asesõnad kirjutatakse kõige sagedamini kokku kaassõnaga. Asesõnadega liituvad kõige sagedamini kaassõnad *arust, meelest, pool* (*poole, poolt*), *jaoks, teada* ja *peale*. Kõige rohkem kirjutati asesõna kokku kaassõnaga *arust*. Määrsõna ja asesõna esinesid eitust väljendavates ühendites nagu *mittemidagi*. Omadussõnaühendeid esines suhteliselt vähe. Kõige rohkem kirjutati omadussõnu kokku kaassõnadega, üksikutel juhtudel määrsõnadega. Verbidest moodustavad kõige sagedamini ühendeid *teadma* ja *saama*, mis seostuvad oleviku eitusvormidega *ei tea* ja *ei saa*. Hüüdsõna ja sellega sageli koos esinevat määr- või asesõna on hakatud tajuma tervikkeelendina ning sellest tulenevalt ka kokku kirjutama. Kokkukirjutamisega võidakse taotleda ka suuremat ekspressiivsust. Sagedaimad ühendid on *einoh* ja *võinoh*, ka *novot* (*novott, noovot*). Kokkukirjutuse üksikjuhtumid on kokkukirjutatud määrsõnad, mida tajutakse liitena, nagu näiteks *niiväga* ja *justnimelt*. Esile kerkis veel kaks liitsidendit – *kasvõi* ja *niiet*. Esimene on grammatiseerunud keelend. Erinevalt esimesest, on teisel ühendil ülekaalus veel lahku kirjutamine. Teine ühend esineb lauses järeldava modaalpartiklina. Internetikeele iseloomulikuks jooneks on mitmesuguste sageli kõrvuti esinevate ühendite kokkukirjutamine. Kõige rohkem kirjutatakse käändsõnaga kokku kaassõnu. Kokkukirjutamise tingib kaassõnaühendi muutunud tähendus ja kasutus uutes kontekstides. (Soodla 2010)

2. Materjali kirjeldus

Selles ja järgmises peatükis esitatakse töö materjal ja käik, st kirjeldatakse, millisest allikmaterjalist ja millisel viisil on koostatud siin töös esitatavad sõnavormide sagedusloendid ja millisel viisil on neid võrreldud normeeritud kirjakeele vastavate loenditega.

Käesoleva bakalaureusetöö materjalina on kasutatud Tartu Ülikooli arvutilingvistika uurimisrühma koostatud uue meedia korpuse, foorumite, uudisgruppide ja kommentaaride allkorpusi. Uue meedia korpus sisaldab ka jututubade allkorpust, kuid see on siinkohal kõrvale jäetud, sest on märgendatud teistmoodi ja seega oleks vajanud teiste programmide (skriptide) kirjutamist. Korpused on kättesaadav aadressilt http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/foorumid_uudisgrupid_kommentaarid. Kasutatud on ilma kordusteta varianti. Korduste all mõeldakse seda, kui keegi tsiteerib kedagi teist selleks, et oleks näha, kellele postitaja vastab või mis teemal räägib. Tsitaadid võivad olla omakorda tsiteeritud, näiteks kui tsiteeritakse kedagi, kes midagi tsiteeris. Kui kedagi tsiteeritakse, on see <*quote*> märgendite vahel. (Uue meedia korpus):

(1) <quote type="postitas_Zorru"> <s> irw väga hea . </s> <s> aga see on nagu päriselt ka juhtunud v ? </s> </quote>

Korpustes on xml-märgendus. Iga fail on jagatud alamosadeks <*divX*> märgendi abil. Foorumite korpuses tähistab <*div3*> postitust. <*div3*> järel on autor <*name*> ja <*p*> märgendite vahel. Seejärel on järgmiste <*p*> märgendite vahel olevate <*time*> märgendite vahel postitamise aeg minuti täpsusega. Kolmandate <*div3*> märgendite vahel olevate <*p*> märgendite vahel on lõigud ning <*s*> märgendite vahel on laused. Uudisgruppide korpuses tähistab postitust <*div2. <head*>-märgendite vahel on postituse pealkiri, <*name*>-märgendite vahel postitaja nimi. Nagu foorumite korpuseski on lõigud <*p*> ja <*/p*> vahel ja laused <*s*> ja <*/s*> vahel. Kommentaaride korpuses tähistab postitust samuti <*div2* märgend, lõike <*p*> märgend ning lauseid <*s*> märgend. Kõikides korpustes moodustab ühe lõigu tavaliselt üks postitus. (Uue meedia korpus):

(2) </div3>
<div3 type="postitus"> <name type="postitaja">joosep12</name>

<time> 20-07-2004 , 13:06 </time>
 <s> cool . </s> <s> tädi oli ikka puhta loll </s> </div3>

(3) </div2>
<div2 type="kommentaar"> <head> 1001 , Mihhail Lendur </head>
<time> 24.12.2003 08:29 </time>
 <s> politruk </s> </div2>
<div2 type="kommentaar"> <head> Patser </head>
<time> 24.12.2003 08:34 </time>
 <s> Ega see , et ma piiblit muinasjutuks pean ei tähenda veel seda , et ma jumalat ei usu . </s> </div2>

Korpuste suurus on järgmine: foorumite korpuses on 8,7 miljonit sõna, kommentaaride korpuses 1,9 miljonit ning uudisgruppide korpuses 4,9 miljonit sõna. Kõigis kolmes korpuses on kokku 15,5 miljonit sõna, millest on käesoleva analüüsi jaoks tehtud ühine fail iga üksiku allkorpusega võrdlemiseks. Foorumite korpus on koostatud Planet Foorumite (endise nimega Zone foorumid) salvestuste lehelt http://forum.planet.ee/ saadud tekstidest. Tekstid pärinevad erinevatest aastatest vahemikus 2000 kuni 2008. Uudisgruppide korpus sisaldab eesti uudisgruppide salvestusi aastatest 2000 kuni 2004. Kommentaaride korpuse aluseks on Delfi kommentaarid ajavahemikust 26.01.2004 – 31.03.2004.

Bakalaureusetöö raames koostati uue meedia korpuste sõnavara sagedusloendid ja võrreldi neid normeeritud kirjakeelt sisaldavate korpuste sõnavara sagedusloenditega. Kasutatud on sõnavorme sageduse järjekorras. Kirjakeele sagedusloendid on koostatud tasakaalus korpuse põhjal, ajakirjanduse sagedusloendi allikateks on 5 miljonit sõna ajalehetekste, ilukirjanduse sagedusloendil 5 miljonit sõna ilukirjandust ja teadustekstide sagedusloendil 5 miljonit sõna teadustekste, tasakaalus korpus põhineb 15 miljonil sõnal, aga sagedusloenditest on välja jäetud sõnad, mille sagedus oli väiksem kui 10. Samuti on välja jäetud kirjavahemärgid, pärisnimed, lühendid, genitiivatribuudid, numbriga kirja pandud arvud ja ka rooma numbrid. Eemaldatud on ka osaliselt numbriga kirjutatud sõnad, näiteks *mp3*, *3D* ja *6-aastane*. Täpsemad andmed saab Riin Kirt'i magistritööst (Kirt 2013). Selles töös kasutatud kirjakeele sagedusloendid pärinevad aadressilt http://www.cl.ut.ee/ressursid/sagedused1/.

3. Töö käik

Sagedusloendite koostamiseks ja võrdlemiseks kirjutasin rea käsureaskripte (*shelli* skriptid). Skript, mis teeb uue meedia korpustest sagedusloendid, töötab järgmiselt. Kõigepealt kustutatakse failidest ära märgid ja numbrid *(«»_="".,\?!():\-0123456789*), seejärel eraldatakse <*s*>-märgendiga tähistatud tekst (st laused) muust infost ja iga lause pannakse eraldi reale. Siis võetakse välja <*s*>-märgendiga read, st sagedusloendid on tehtud ainult lausemärgendite vahel oleva teksti põhjal. Seejärel kustutatakse *hi rendpealkiri*-märgendiga (algses failis <*hi rend="pealkiri"*>, aga märgid said eelnevalt eemaldatud) read, st eemaldatakse vahepealkirjad, millel tehnilistel põhjustel oli lisaks pealkirja märgendile ka lausemärgend. Vahepealkirja märgendiga tähistatakse pealkirju, kirja stiili ja emotikone. Kuna pealkirjad on ka laused, pannakse nad mõnikord vahepealkirja märgendi vahele. Peale seda eemaldatatakse kõik märgendid, mis on ümbritsetud kolmnurksulgudega (<, >). Siis tehakse kõik tühikud ühekordseteks. Järgmisena teisendatakse kõik lause alguses olevad suurtähed väiketähtedeks, et lause alguses olev algselt suure algustähega sõne poleks sagedussõnastikus eraldi sõnena. Seejärel teisendatakse kõik tühikud reavahetusteks. Peale seda sorteeritakse kõik read, loetakse üle kõik kordused ja lõpuks järjestatakse sagedusloend sageduse järgi ja suunatakse väljundfaili.

Järgmises etapis ühendatakse uue meedia korpuste põhjal loodud sagedusloendid eelnevalt Riin Kirdi poolt tasakaalus korpuse põhjal loodud sagedusloenditega.

Siiamaani töödeldi vaid uue meedia faile, edasi töödeldakse juba nii uue meedia faile kui ka kirjakeele faile. Järgmise skriptiga pannakse igale reale järjekorranumbrid. Siis võetakse ära kõik tühikud järjekorranumbri ja sageduse vahel. Järgnes failide ühendamine. Failid ühendati teise välja alusel (sõnavormide järgi) käsu *join* abil. Kohtadesse kus ühist sõnavormi ei olnud, pannakse sõna *TÜHI*. Peale seda oli suhteliselt lihtne välistada seda sõna sisaldavad read, et alles jääks vaid need read, kus oli ühine sõnavorm mõlemal failil. Tulemus kirjutati väljundfaili:

(4) aaalfa 481338:1 TÜHI TÜHI

(5) TÜHI TÜHI aaamatukogu 338655:1

Näites 4 on näha, et sõna *aaalfa* leidus esimeses failis, aga puudus teises. Näites 5 leidus sõna *aaamatukogu* teises failis, aga puudus esimeses. Kooloni ees olev arv on sõnavormi

järjekorranumber sageduse kahanemise alusel järjestatud loendis ja arv peale koolonit on esinemissagedus.

Ühendatud failidest esimene oli kas foorumite, kommentaaride, uudisgruppide või nendest kõigist kokku pandud sagedusloend, teine oli ajakirjanduse, ilukirjanduse, teadustekstide või tasakaalus korpuse sagedusloend.

Ühendamise skriptiga loodi 16 faili, igale ühendatud failile vastab üks lahter Tabelis 2. Üks käsurea skriptide näide on olemas ka aadressil http://kodu.ut.ee/~lkristja/. Täielikud sagedusloendid on samuti saadaval aadressil http://kodu.ut.ee/~lkristja/.

4. Tulemuste kvantitatiivne analüüs

Tabel 1. Üldinfo.

Tabel 1.	Tekstisõnu	Sõnavara	Sõnavara ilma ühekordseteta	Üks kord esinenud sõnade protsent	Üks kord esinenud sõnade arv
Foorumid	8.7 mln	481454	184134	61.8	297320
Kommentaarid	1.9 mln	192365	75463	60.8	116902
Uudisgrupid	4.9 mln	326348	140368	57.0	185980
Uus meedia kokku	15.5 mln	766746	302944	60.5	463802
Ajakirjandus	5 mln	338658	141714	58.2	196944
Ilukirjandus	5 mln	314573	129816	58.7	184757
Teadustekstid	5 mln	340859	153630	54.9	187229
Tasakaalus korpus kokku	15 mln	82458	82458	0	0

Tabelist 1 võime näha, kui palju on tekstisõnu sagedusloendite aluseks olnud korpustes. Kõige vähem sõnu on kommentaaride korpuses, kõige rohkem aga foorumite korpuses. Sõnavara tulp näitab meile kui palju erinevaid sõnavorme igas sagedusloendis on. Kui ühekordsed sõnade esinemised ära võtta, jääb järele tunduvalt vähem sõnu. Üks kord esinenud sõnade protsent on iga sagedusloendi korral üle 50%. Tüüpiliselt moodustavadki üks kord esinevad sõnad umbes 50% korpuse sõnavarast (vt lähemalt osa 1.1). Viimasest tulbast on näha kui palju neid arvuliselt on. Viimase rea kahes viimases lahtris on väärtusteks 0, kuna tasakaalus korpuse sagedusloendist olid juba loendi koostamisel välja jäetud kuni üheksa korda esinenud sõnad.

Tabel 2. Sõnavara kattuvus sagedusloendites.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	116724	111027	92491	64028
Kommentaarid	80507	78221	68157	51880
Uudisgrupid	92771	33799	80736	56739
Uus meedia	150147	138715	122068	71861

Tabelist 2 saame teada, et kõige rohkem ühiseid sõnu on ajakirjanduse ja uue meedia (foorumid, kommentaarid, uudisgrupid) sõnaloendites. Kui arvestada kõiki uue meedia sagedusloendeid eraldi, siis kõige rohkem ühiseid sõnu on foorumite ja ajakirjanduse sõnaloendites, aga väga palju ei jää maha foorumite ja ilukirjanduse sagedusloendite ühiste

sõnade arv. Kõige vähem ühiseid sõnu on uudisgruppide ja ilukirjanduse sagedusloendites. Viimases tulbas on kattuvus väiksem kui teistes tulpades, kuna tasakaalus korpuse sagedusloendist on osad sõnad välja jäetud (vt. „Materjali kirjeldus“). Järelikult on vähem sõnavorme, mis võivad tasakaalus korpuse sõnavaraga kattuda.

Tabel 3. Sõnavara kattuvus sagedusloendites üks kord esinenud sõnu arvestamata.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	66104	63089	52194	54720
Kommentaarid	43491	42050	37136	39065
Uudisgrupid	52661	47034	46277	45848
Uus meedia	85503	78727	69876	66011

Tabel 3 näitab, kui suur on korpuste ühine sõnavara, kui üks kord esinenud sõnu ei arvestata. Üks kord esinenud sõnavormide kõrvalejätmist võib pidada õigustatuks, sest interneti nn kasutaja loodud sisus on palju trükivigadega sõnu, mis tüüpiliselt esinevad vaid ühe korra ja millele ei leidu vastet normeeritud kirjakeele korpuste baasil koostatud sagedusloendites. Tabelist 3 näeme ka seda, et endiselt on kõige rohkem ühiseid sõnu ajakirjanduse ja uue meedia korpustes ning kui mitte arvestada kogu uue meedia ja tasakaalus korpuse sagedusloendeid, siis on kõige rohkem ühised sõnu foorumite ja ajakirjanduse sagedusloendites. Kõige vähem ühiseid sõnu on kommentaaride ja teadustekstide sagedusloendites.

Tabel 4. Ühiste sõnade protsent sõnavarast ühekordseid sõnu arvestades.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	24.2/34.5	23.1/35.3	19.2/27.1	13.3/77.6
Kommentaarid	41.9/23.8	40.7/24.9	35.4/20.0	27.0/62.9
Uudisgrupid	28.4/27.4	10.4/10.7	24.7/23.7	17.4/68.8

Uus meedia	19.6/44.3	18.1/44.1	15.9/35.8	9.4/87.1

Tabelis 4 on igas lahtris vasakul pool olev väärtus esimeses tulbas oleva korpuse sagedusloendi ja talle vastava esimeses reas oleva korpuse sagedusloendi ühiste sõnade protsent esimeses tulbas oleva korpuse sagedusloendi sõnavarast. Paremal pool olev väärtus igas lahtris on samade korpuste ühiste sõnade protsent kirjakeele sagedusloendist (st esimeses reas olevast korpusest). Näiteks moodustab foorumite ja ajakirjanduse ühine sõnavara 24,2 % foorumite sõnavarast ja 34,5% ajakirjanduse sõnavarast. Võime näha, et kõige väiksema osa vastavate korpuste sagedusloenditest moodustavad ilukirjanduse ja uudisgruppide sagedusloendite ühised sõnad, mida on veidi üle 10%. Suurima osa moodustab uue meedia ja tasakaalus korpuse sõnaloendite ühine osa tasakaalus korpuse sagedusloendist, mida on üle 87%. Kuna tasakaalus korpuse sagedusloendist on osad sõnad välja jäetud, siis on uue meedia korpuste sagedusloenditega ühiste sõnade protsent ka selle võrra väiksem. Kui aga mitte arvestada koondkorpusi, siis moodustab suurima osa kommentaaride ja ajakirjanduse sagedusloendite ühine osa kommentaaride sagedusloendist (41,9%).

Tabel 5. Ühiste sõnade protsent sõnavarast ühekordseid sõnu arvestamata.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	35.9/46.6	34.3/48.6	28.3/34.0	29.7/66.4
Kommentaarid	57.6/30.7	55.7/32.4	49.2/24.2	51.8/47.4
Uudisgrupid	37.5/37.2	33.5/36.2	33.0/30.1	32.7/55.6
Uus meedia	28.2/60.3	26.0/60.6	23.1/45.5	21.8/80.1

Tabelist 5 võime näha, et kõige suurema osa moodustab kogu uue meedia sagedusloendi ja tasakaalus korpuse sagedusloendi ühine osa tasakaalus korpuse sagedusloendist (80,1%). Kui aga kombineeritud korpusi mitte arvestada, moodustab kõige suurema osa kommentaaride ja ajakirjanduse sagedusloendite ühine osa kommentaaride sagedusloendist (57,6%). Kõige väiksema osa moodustab kogu uue meedia sagedusloendi ja tasakaalus korpuse sagedusloendi ühine osa uue meedia korpuse sagedusloendist (21,8%). Kui aga siingi kombineeritud korpusi

mitte arvestada, moodustab väikseima osa kommentaaride ja teadustekstide sagedusloendite ühine osa teadustekstide sagedusloendist (24,2%).

Tabel 6. Mitte-ühise sõnavara suurus protsentides ühekordseid sõnu arvestades.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	75.8/65.5	76.9/64.7	80.8/72.9	86.7/22.4
Kommentaarid	58.1/76.2	59.3/75.1	64.6/80.0	73.0/37.1
Uudisgrupid	71.6/72.6	89.6/89.3	75.3/76.3	82.6/31.2
Uus meedia	80.4/55.7	81.9/55.9	84.1/64.2	90.6/12.9

Tabelist 6 võime näha, et kõige rohkem mitte-ühiseid sõnu on ilukirjanduse ja uudisgruppide sagedusloendites. Eripäraste sõnade protsent on kummastki sagedusloendist peaaegu 90%. Kõige vähem mitte-ühiseid sõnu on uue meedia ja tasakaalus korpuse sagedusloendites. Eripärased sõnad moodustavad tasakaalus korpuse sagedusloendist peaaegu 13%. Kuna üks kord esinevate sõnade hulgas on uue meedia korpuses palju trükivigu, väikese algustähega kirjutatud pärisnimed jms sõnavormid, millel ei saagi olla vastet kirjakeele korpuste sagedusloendites, siis on uue meedia sagedusloendid tasakaalus korpuse sagedusloenditest selle võrra rohkem erinevad.

Tabel 7. Mitte-ühise sõnavara suurus üks kord esinenud sõnu arvestamata.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	118030/75610	121045/66727	131940/101436	129414/27738
Kommentaarid	31972/98223	33413/87766	38327/116494	36398/43393
Uudisgrupid	87707/89053	93334/82782	94091/107353	94520/36610
Uus meedia	217441/56211	224217/51089	233068/83754	236933/16447

Tabelist 7 võime näha, et üks kord esinevaid sõnu mitte arvestades on kõige rohkem eripäraseid sõnu uue meedia ja tasakaalus korpuse sõnaloendite ühisest osast uue meedia tekstides (ligi 237000), kuna võrreldud on kogu uue meedia ühendi sagedusloendit rohkete väljajättudega tasakaalus korpuse sagedusloendiga. Kui need aga välja jätta, on kõige suurem

erinevus foorumite sagedusloendi ja teadustekstide sagedusloendi ühisosa ja foorumite sagedusloendite vahel, ligi 132000 sõna. Kõige väiksem on erinevus kommentaaride ja ajakirjanduse sagedusloendite ühiste sõnavormide ja kommentaaride sagedusloendil, alla 32000 sõna.

Tabel 8. Mitte-ühise sõnavara suurus protsentides üks kord esinenud sõnu arvestamata.

	Ajakirjandus	**Ilukirjandus**	**Teadustekstid**	**Tasakaalus korpus**
Foorumid	64.1/53.4	65.7/51.4	71.7/66.0	70.3/33.6
Kommentaarid	42.4/69.3	44.3/67.6	50.8/75.8	48.2/52.6
Uudisgrupid	62.5/62.8	66.5/63.8	67.0/69.9	67.3/44.4
Uus meedia	71.8/39.7	74.0/39.4	76.9/54.5	78.2/19.9

Tabelist 8 näeme, et uue meedia ja tasakaalus korpuse sõnaloendite ühiste sõnavormide ja uue meedia sagedusloendite vahel on kõige suurem erinevus (78,2%). Kui jätta välja mitmest sagedusloendist koosnevad loendid, siis on suurim erinevus foorumite ja teadustekstide sagedusloendite ühise osa ja foorumite sagedusloendi vahel, mida on 71,7%. Väikseim erinevus on ajakirjanduse ja kommentaaride sõnavormide sagedusloendi ning kommentaaride sagedusloendi vahel (42,4%) mis näitab, et ühekordseid sõnu arvestamata on sagedusloendite ühisosad sagedusloenditest endist mitte nii erinevad kui sel juhul kui me polnud üks kord esinenud sõnu välja jätnud.

Sellest osast järeldub, et kui sagedusloendites arvestada ühekordseid sõnu, on suurim uue meedia ja tasakaalus korpuse sõnaloendite ühiste sõnavormide protsent tasakaalus korpuse sagedusloendist (87,1%) ja väikseim ilukirjanduse ja uudisgruppide sagedusloendite ühiste sõnavormide protsent uudisgruppide sagedusloendist (10,4%). Kui sagedusloendites mitte arvestada ühekordseid sõnu, on suurim kogu uue meedia sagedusloendi ja tasakaalus korpuse sagedusloendi ühiste sõnavormide protsent tasakaalus korpuse sagedusloendist (80,1%) ning väikseim kogu uue meedia sagedusloendi ja tasakaalus korpuse sagedusloendi ühiste sõnavormide protsent uue meedia korpuse sagedusloendist (21,8%).

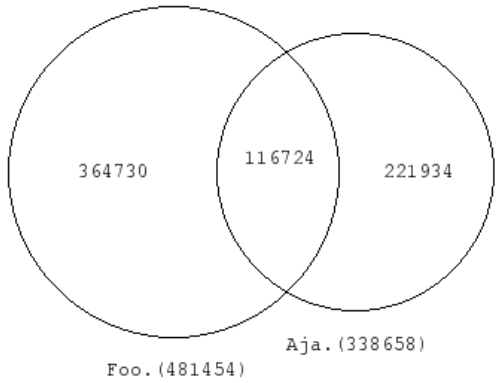
4.1 Venni diagrammid

Uue meedia ja normeeritud kirjakeele sõnavormide kattuvuse ja erinevuse visualiseerimiseks on selles töös kasutatud *Venni* diagramme. *Venni* diagramm on hulkadevaheliste loogiliste suhete illustreerimiseks kasutatav diagramm.

Diagrammide loomiseks on kasutatud tarkvara bioinforx kodulehelt BxToolBoxist nimega Venn Diagram, mis on kättesaadav lehelt http://apps.bioinforx.com/bxaf6/tools/app_overlap.php. Tarkvara kasutamiseks on vaja eelnevalt registreerida kasutajaks.

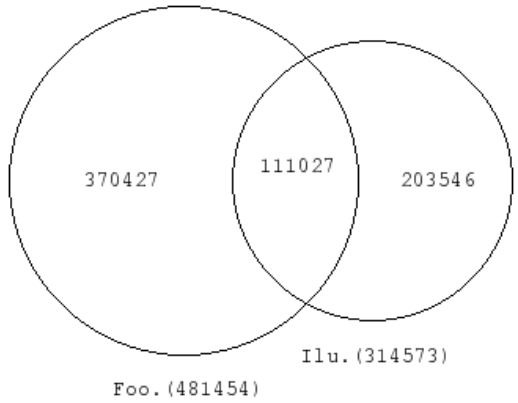
4.1.1 Koos ühekordsete sõnavormidega

Joonis 2. Foorumite (Foo.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



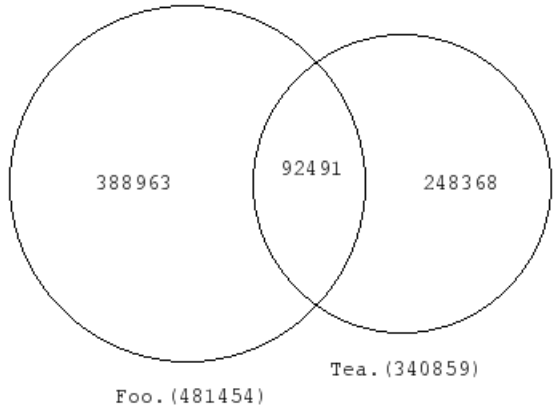
Joonisel 2 on kujutatud foorumite korpuse (*Foo.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et foorumite korpuse 481454 sõnavormist ei kattu 364730 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 338658 sõnavormist ei kattu 221934 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 116724.

Joonis 3. Foorumite (Foo.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



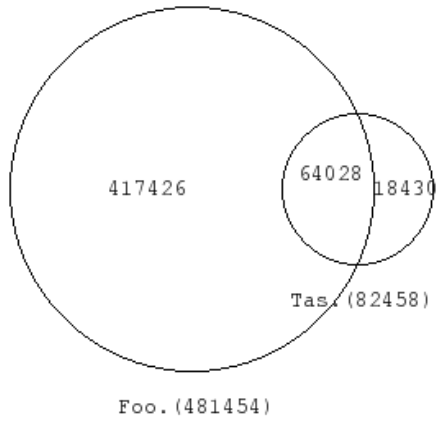
Joonisel 3 on kujutatud foorumite korpuse (*Foo.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et foorumite korpuse 481454 sõnavormist ei kattu 370427 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 314573 sõnavormist ei kattu 203546 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 111027.

Joonis 4. Foorumite (Foo.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



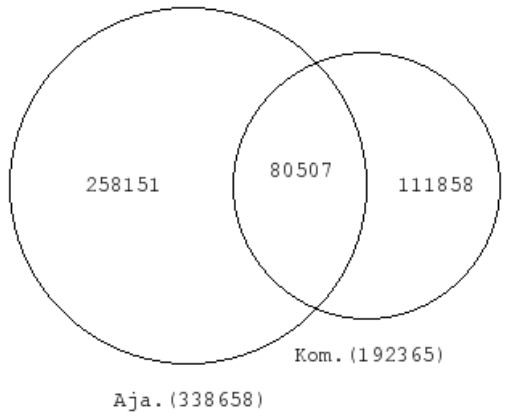
Joonisel 4 on kujutatud foorumite korpuse (*Foo.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et foorumite korpuse 481454 sõnavormist ei kattu 388963 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 340859 sõnavormist ei kattu 248368 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 92491.

Joonis 5. Foorumite (Foo.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



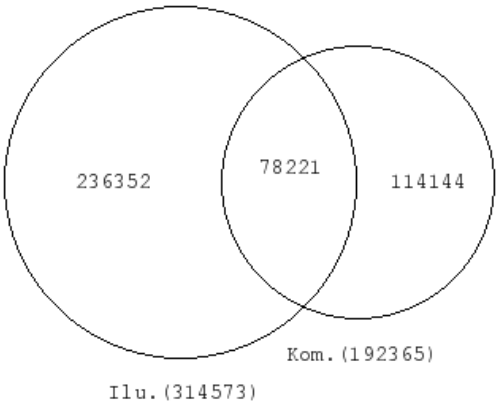
Joonisel 5 on kujutatud foorumite korpuse (*Foo.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et foorumite korpuse 481454 sõnavormist ei kattu 417426 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 18430 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 64028.

Joonis 6. Kommentaaride (Kom.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 6 on kujutatud kommentaaride korpuse (*Kom.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kommentaaride korpuse 192365 sõnavormist ei kattu 111858 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 338658 sõnavormist ei kattu 258151 sõnavormi kommentaaride korpusega. Näeme ka seda, et kommentaaride korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 80507.

Joonis 7. Kommentaaride (Kom.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



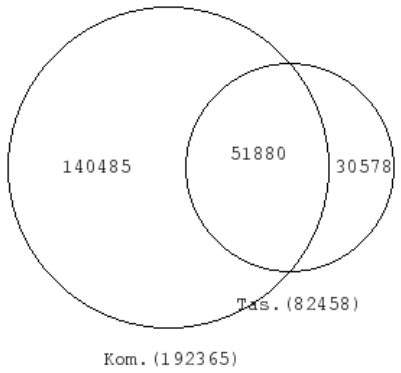
Joonisel 7 on kujutatud kommentaaride korpuse (*Kom.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kommentaaride korpuse 192365 sõnavormist ei kattu 114144 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 314573 sõnavormist ei kattu 236352 sõnavormi kommentaaride korpusega. Näeme ka seda, et kommentaaride korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 78221.

Joonis 8. Kommentaaride (Kom.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus koos ühekordsete sõnavormidega.

Joonisel 8 on kujutatud kommentaaride korpuse (*Kom.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kommentaaride korpuse 192365 sõnavormist ei kattu 124208 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 340859 sõnavormist ei kattu 272702 sõnavormi kommentaaride

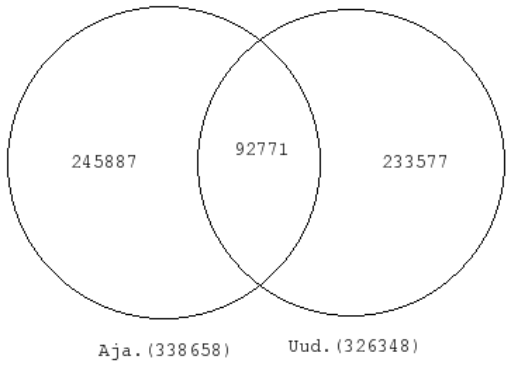
korpusega. Näeme ka seda, et kommentaaride korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 68157.

Joonis 9. Kommentaaride (Kom.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 9 on kujutatud kommentaaride korpuse (*Kom.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kommentaaride korpuse 192365 sõnavormist ei kattu 140485 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 30578 sõnavormi kommentaaride korpusega. Näeme ka seda, et kommentaaride korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 51880.

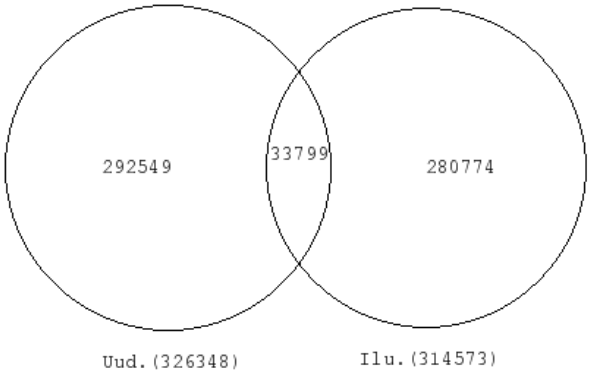
Joonis 10. Uudisgruppide (Uud.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 10 on kujutatud uudisgruppide korpuse (*Uud.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et uudisgruppide korpuse 326348 sõnavormist ei kattu 233577 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 338658 sõnavormist ei kattu 245887 sõnavormi uudisgruppide

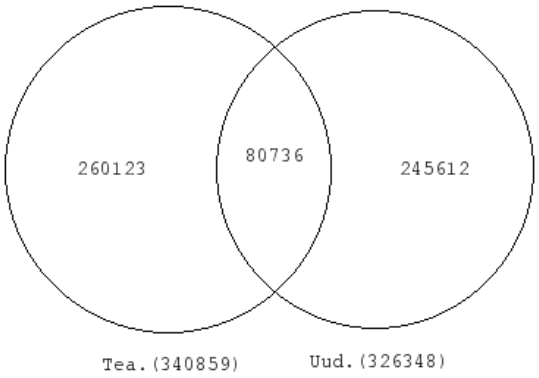
korpusega. Näeme ka seda, et uudisgruppide korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 92771.

Joonis 11. Uudisgruppide (Uud.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 11 on kujutatud uudisgruppide korpuse (*Uud.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et uudisgruppide korpuse 326348 sõnavormist ei kattu 292549 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 314573 sõnavormist ei kattu 280774 sõnavormi uudisgruppide korpusega. Näeme ka seda, et uudisgruppide korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 33799.

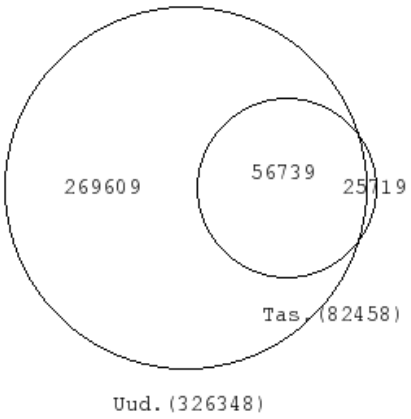
Joonis 12. Uudisgruppide (Uud.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 12 on kujutatud uudisgruppide korpuse (*Uud.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et uudisgruppide korpuse 326348 sõnavormist ei kattu 245612 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 340859 sõnavormist ei kattu 260123 sõnavormi uudisgruppide

korpusega. Näeme ka seda, et uudisgruppide korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 80736.

Joonis 13. Uudisgruppide (Uud.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



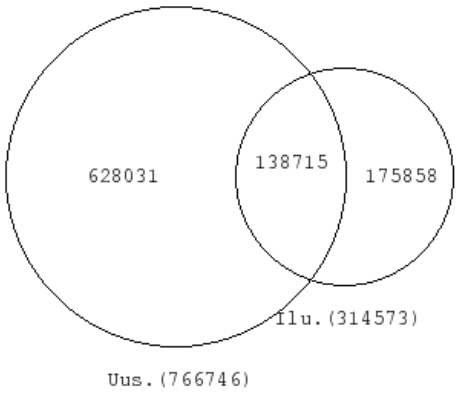
Joonisel 13 on kujutatud uudisgruppide korpuse (*Uud.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et uudisgruppide korpuse 326348 sõnavormist ei kattu 269609 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 25719 sõnavormi uudisgruppide korpusega. Näeme ka seda, et uudisgruppide korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 56739.

Joonis 14. Kogu uue meedia (Uus.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus koos ühekordsete sõnavormidega.

Joonisel 14 on kujutatud kogu uue meedia korpuse (*Uus.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kogu uue meedia korpuse 766746 sõnavormist ei kattu 616599 sõnavormi ajakirjanduskorpusega ja

ajakirjanduskorpuse 338658 sõnavormist ei kattu 188511 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 150147.

Joonis 15. Kogu uue meedia (Uus.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



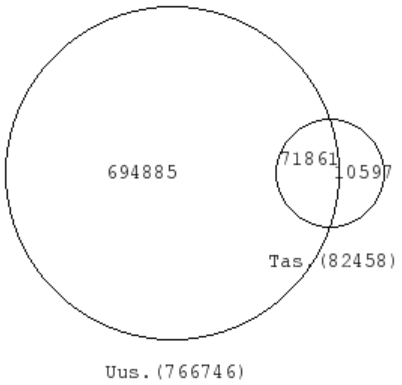
Joonisel 15 on kujutatud kogu uue meedia korpuse (*Uus.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kogu uue meedia korpuse 766746 sõnavormist ei kattu 628031 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 314573 sõnavormist ei kattu 175858 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 138715.

Joonis 16. Kogu uue meedia (Uus.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus koos ühekordsete sõnavormidega.

Joonisel 16 on kujutatud kogu uue meedia korpuse (*Uus.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kogu uue meedia korpuse 766746 sõnavormist ei kattu 644678 sõnavormi teaduskirjanduskorpusega ja

teaduskirjanduskorpuse 340859 sõnavormist ei kattu 218791 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 122068.

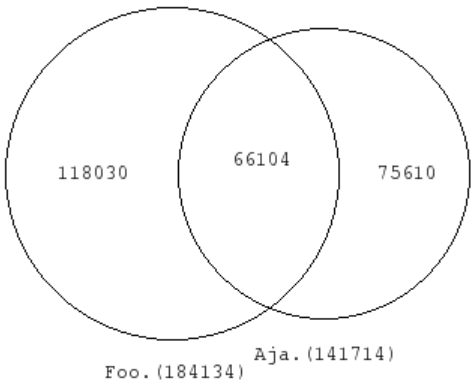
Joonis 17. Kogu uue meedia (Uus.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus koos ühekordsete sõnavormidega.



Joonisel 17 on kujutatud kogu uue meedia korpuse (*Uus.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus arvestades üks kord esinenud sõnavorme. Näeme, et kogu uue meedia korpuse 766746 sõnavormist ei kattu 694885 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 10597 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 71861.

4.1.2 Ilma ühekordsete sõnavormideta

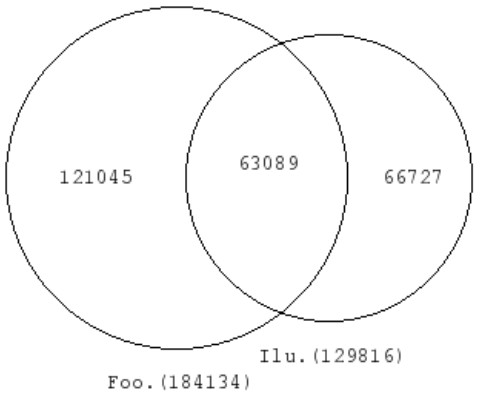
Joonis 18. Foorumite (Foo.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 18 on kujutatud foorumite korpuse (*Foo.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et foorumite korpuse 184134

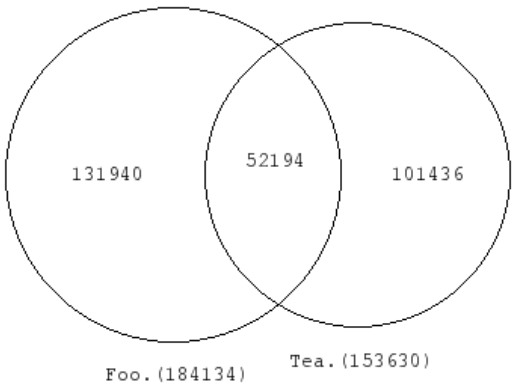
sõnavormist ei kattu 118030 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 141714 sõnavormist ei kattu 75610 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 66104.

Joonis 19. Foorumite (Foo.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 19 on kujutatud foorumite korpuse (*Foo.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et foorumite korpuse 184134 sõnavormist ei kattu 121045 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 129816 sõnavormist ei kattu 66727 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 63089.

Joonis 20. Foorumite (Foo.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 20 on kujutatud foorumite korpuse (*Foo.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et foorumite korpuse 184134 sõnavormist ei kattu 131940 sõnavormi teaduskirjanduskorpusega ja

teaduskirjanduskorpuse 153630 sõnavormist ei kattu 101436 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 52194.

Joonis 21. Foorumite (Foo.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.

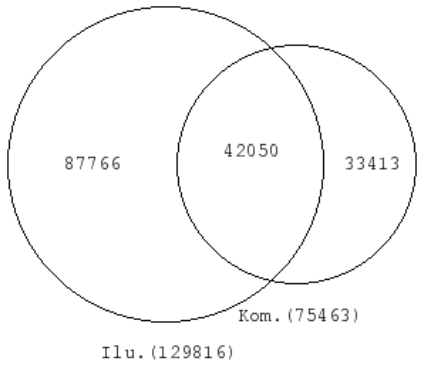
Joonisel 21 on kujutatud foorumite korpuse (*Foo.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et foorumite korpuse 184134 sõnavormist ei kattu 129414 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 27738 sõnavormi foorumite korpusega. Näeme ka seda, et foorumite korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 54720.

Joonis 22. Kommentaaride (Kom.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.

Joonisel 22 on kujutatud kommentaaride korpuse (*Kom.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kommentaaride korpuse 75463 sõnavormist ei kattu 31972 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 141714 sõnavormist ei kattu 98223 sõnavormi kommentaaride

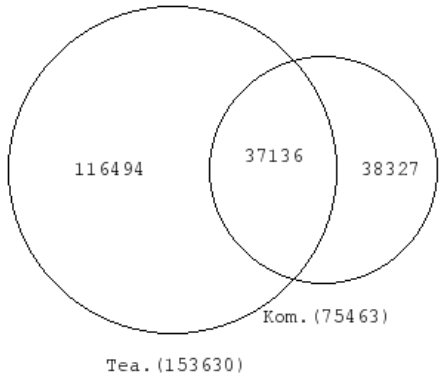
korpusega. Näeme ka seda, et kommentaaride korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 43491.

Joonis 23. Kommentaaride (Kom.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 23 on kujutatud kommentaaride korpuse (*Kom.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kommentaaride korpuse 75463 sõnavormist ei kattu 33413 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 129816 sõnavormist ei kattu 87766 sõnavormi kommentaaride korpusega. Näeme ka seda, et kommentaaride korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 42050.

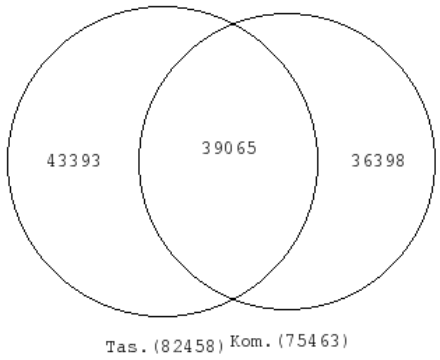
Joonis 24. Kommentaaride (Kom.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 24 on kujutatud kommentaaride korpuse (*Kom.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kommentaaride korpuse 75463 sõnavormist ei kattu 38327 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 153630 sõnavormist ei kattu 116494 sõnavormi kommentaaride

korpusega. Näeme ka seda, et kommentaaride korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 37136.

Joonis 25. Kommentaaride (Kom.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



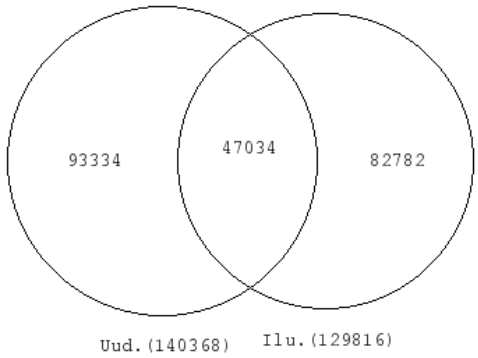
Joonisel 25 on kujutatud kommentaaride korpuse (*Kom.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kommentaaride korpuse 75463 sõnavormist ei kattu 36398 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 43393 sõnavormi kommentaaride korpusega. Näeme ka seda, et kommentaaride korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 39065.

Joonis 26. Uudisgruppide (Uud.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.

Joonisel 26 on kujutatud uudisgruppide korpuse (*Uud.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et uudisgruppide korpuse 140368 sõnavormist ei kattu 87707 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 141714 sõnavormist ei kattu 89053 sõnavormi uudisgruppide korpusega.

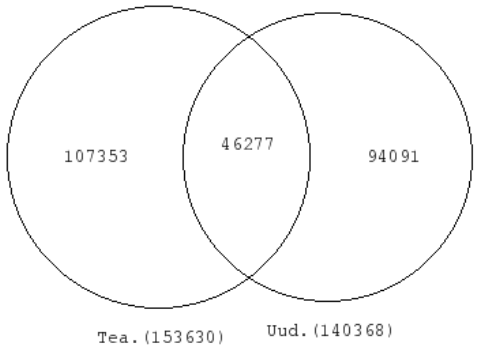
Näeme ka seda, et uudisgruppide korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 52661.

Joonis 27. Uudisgruppide (Uud.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 27 on kujutatud uudisgruppide korpuse (*Uud.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et uudisgruppide korpuse 140368 sõnavormist ei kattu 93334 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 129816 sõnavormist ei kattu 82782 sõnavormi uudisgruppide korpusega. Näeme ka seda, et uudisgruppide korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 47034.

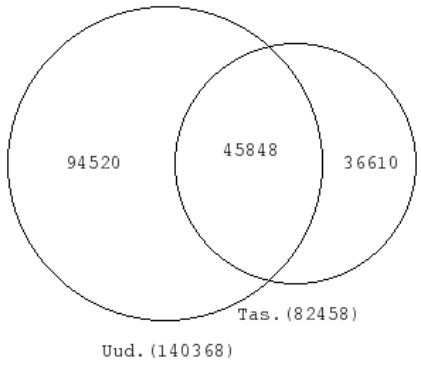
Joonis 28. Uudisgruppide (Uud.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 28 on kujutatud uudisgruppide korpuse (*Uud.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et uudisgruppide korpuse 140368 sõnavormist ei kattu 94091 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 153630 sõnavormist ei kattu 107353 sõnavormi uudisgruppide

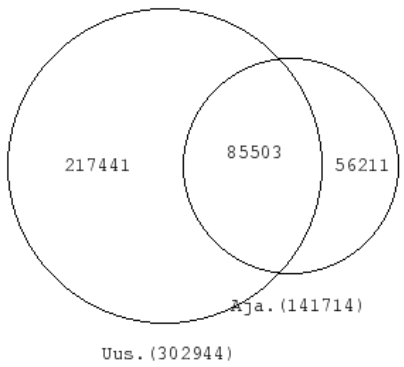
korpusega. Näeme ka seda, et uudisgruppide korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 46277.

Joonis 29. Uudisgruppide (Uud.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 29 on kujutatud uudisgruppide korpuse (*Uud.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et uudisgruppide korpuse 140368 sõnavormist ei kattu 94520 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 36610 sõnavormi uudisgruppide korpusega. Näeme ka seda, et uudisgruppide korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 45848.

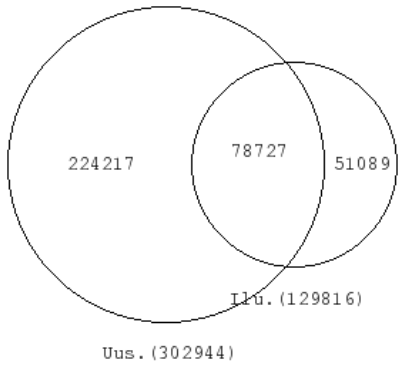
Joonis 30. Kogu uue meedia (Uus.) ja ajakirjanduskeele (Aja.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 30 on kujutatud kogu uue meedia korpuse (*Uus.*) ja ajakirjanduskorpuse (*Aja.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kogu uue meedia korpuse 302944 sõnavormist ei kattu 217441 sõnavormi ajakirjanduskorpusega ja ajakirjanduskorpuse 141714 sõnavormist ei kattu 56211 sõnavormi kogu uue meedia

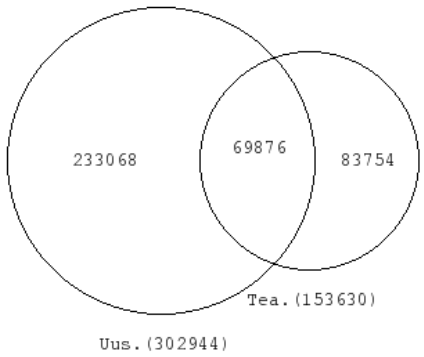
korpusega. Näeme ka seda, et kogu uue meedia korpuses ja ajakirjanduskorpuses on ühiseid sõnavorme 85503.

Joonis 31. Kogu uue meedia (Uus.) ja ilukirjanduskeele (Ilu.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 31 on kujutatud kogu uue meedia korpuse (*Uus.*) ja ilukirjanduskorpuse (*Ilu.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kogu uue meedia korpuse 302944 sõnavormist ei kattu 224217 sõnavormi ilukirjanduskorpusega ja ilukirjanduskorpuse 129816 sõnavormist ei kattu 51089 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja ilukirjanduskorpuses on ühiseid sõnavorme 78727.

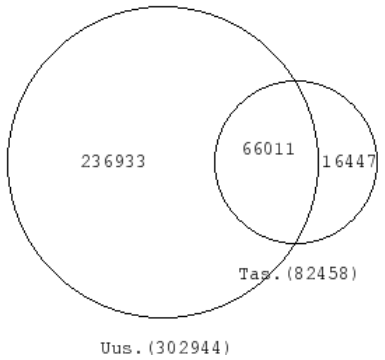
Joonis 32. Kogu uue meedia (Uus.) ja teaduskirjanduskeele (Tea.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 32 on kujutatud kogu uue meedia korpuse (*Uus.*) ja teaduskirjanduskorpuse (*Tea.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kogu uue meedia korpuse 302944 sõnavormist ei kattu 233068 sõnavormi teaduskirjanduskorpusega ja teaduskirjanduskorpuse 153630 sõnavormist ei kattu 83754 sõnavormi kogu uue meedia

korpusega. Näeme ka seda, et kogu uue meedia korpuses ja teaduskirjanduskorpuses on ühiseid sõnavorme 69876.

Joonis 33. Kogu uue meedia (Uus.) ja kogu tasakaalus korpuse (Tas.) sõnavormide kattuvus ilma ühekordsete sõnavormideta.



Joonisel 33 on kujutatud kogu uue meedia korpuse (*Uus.*) ja kogu tasakaalus korpuse (*Tas.*) sõnavormide kattuvus üks kord esinenud sõnavorme arvestamata. Näeme, et kogu uue meedia korpuse 302944 sõnavormist ei kattu 236933 sõnavormi kogu tasakaalus korpusega ja kogu tasakaalus korpuse 82458 sõnavormist ei kattu 16447 sõnavormi kogu uue meedia korpusega. Näeme ka seda, et kogu uue meedia korpuses ja kogu tasakaalus korpuses on ühiseid sõnavorme 66011.

5. Tulemuste kvalitatiivne analüüs

5.1 Kahe korpuse ühise sõnavara sagedaim osa uue meedia põhjal

Selles osas on kirjeldatud mõlema sagedusloendi viit sagedaimat sõnavormi, mis olid nii uue meedia sagedusloendis kui ka temaga võrreldud kirjakeele sagedusloendis. Võrdlemise aluseks on sõnavormi järjekorranumber sageduse kahanemise alusel järjestatud loendis.

Edasi analüüsitakse ka uue meedia sagedusloendi esimest sadat sõnavormi, millele leiduvad vasted tasakaalus korpuses. Mitmeti määratavad sõnavormid on liigitatud sagedasema kasutusviisi alusel (näiteks *aru* omaette nimisõnana on palju harvem kui ühendi *aru saama* koosseisus). Sagedased sõnavormid on liigitatud sõnaliikide kaupa. Täispikkuses sagedusloendid on aadressil http://kodu.ut.ee/~lkristja/.

5.1.1 Kogu uue meedia kõige sagedasemad tasakaalus korpusega kattuvad sõnavormid

Uue meedia ja ajakirjanduse ühises sõnavaras on esimestel ja teistel kohtadel mõlemal vastavalt *on* ja *ja*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ei* ning viiendatel kohtadel mõlemal *kui*.

Uue meedia ja ilukirjanduse ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *ei*, kolmandatel *ei* ja *on*, neljandatel *et* ja *ta* ning viiendatel kohtadel *kui* ja *et*.

Uue meedia ja teadustekstide ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *on*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ka* ning viiendatel kohtadel mõlemal *kui*.

Uue meedia ja tasakaalus korpuse ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *on* ning kolmandatel, neljandatel ja viiendatel kohtadel mõlemal vastavalt *ei*, *et* ja *kui*.

Kogu uue meedia sagedusloendi esimese saja sõnavormi hulgas on 3 nimisõnavormi, 1 omadussõnavorm, 27 asesõnavormi, 17 tegusõnavormi, 36 määrsõna, 12 sidesõna ja 4 kaassõna.

Nimisõnavormid on *asi*, *näiteks*, *aru*. Nimetavakäändeline laia ja ebamäärase tähendusega nimisõnavorm *asi* on uue meedia tekstides palju kasutuses mitmesugustes püsiväljendites (*asi on selles, selline asi, et, asi käib nii*) või siis mingi üldise olukorra tähistajana:

(6) asi täitsa vaadatav

Saavakäändeline parasiitsõna *näiteks* esineb foorumitekstides mitmel pool. Selle sõna võib ära jätta ilma et tähendus oluliselt muutuks:

(7) Samas aitab näiteks foorumi KKK lugemine... ja sealt näite 1:1 kopeerimine

Nimetavakäändeline laia kasutusega nimisõnavorm *aru* esineb peamiselt väljendverbi kooseisus:

(8) püüan aru saada

Samuti ka iseseisva nimisõnana:

(9) aga varsa aru!!

Ainus omadussõnavorm on *hea*; Omadussõnavormi *hea* kasutatakse atribuudina:

(10) Selle raha eest on suht hea tase...

Asesõnavormid on isikuliste asesõnade vormid *ma, sa, ta, mul, minu, mina, mulle*; küsiv-siduvad asesõnad *mis, kes, mida*; Näitavad asesõnad *sama, see, seda, selle, need, neid, nad, selline, sellest, nende*; umbmäärased asesõnad *üks, midagi, mingi, kõik, keegi*; enesekohased asesõnad *oma, ise*.

Tegusõnavormid on *olema* vormid *on, pole, oli, ole, oleks, olen, olla, olema, olnud*; *saama* vormid *saab, saa*; *võima* vormid *võib, võiks* ja muud tegusõnavormid *teha, tuleb, peaks, tea*.

Määrsõnad on asemäärsõnad *siin, kas, kus, miks, seal, kuidas, siis, nii*; rõhumäärsõnad *ka, veel, juba, palju, ainult, lihtsalt, rohkem, enam, isegi, ikka, küll, ju, ei, väga, vist, eriti, just, kõige, päris, nüüd, no, ära, nagu, mitte, jah* ja abimäärsõnad *välja, üle, vaja*.

Sidesõnad on *ja, et, kui, aga, või, ning, kuid, sest, ehk, vaid, kuna, ega*.

Kaassõnad on *peale, eest, kohta, pärast*.

Kõigi uue meedia korpuste sagedusloendites kokku on sageduse järjekorras kirjakeele korpusega võrrelduna kõrgemal kohal sõnad *on, siis, see, aga, ma, seda, või, nii, pole, kas, sa, ära, veel, ole, midagi, mitte, oleks, ikka, küll, ise, ju, mingi* (suhe uue meedia ja kirjakeele korpuse sageduse vahel 36:182), *väga, olen, mul, saab, peale, minu, keegi, seal, palju, need, teha, siin, hea, asi* (suhe 54:242), *lihtsalt, saa, jah* (suhe 61:249), *vist* (suhe 62:257), *mina,*

peaks, ehk, selline (suhe 73:214), *rohkem, olema, mulle, vaja, kuna, eriti, just, sama, miks, tea, aru, no* (suhe 96:537), *päris* (suhe 97:221) ja *võiks.*

Tasakaalus korpuse sagedusloendites on sageduse järjekorras kõrgemal kohal sõnad *ja, oli, oma, ta, nagu, kes, ning, mida, juba, kõik, välja, kuid, nüüd, ainult, üks, olla, tuleb, neid, kuidas, võib, kus, sest, nad, vaid, üle, sellest, ega, näiteks, enam, eest, isegi, kohta, kõige, pärast, olnud* ja *nende*.

Sageduse järjekorras samal kohal on sõnad *ei, et, kui, ka, mis* ja *selle.*

5.1.2 Foorumite kõige sagedasemad tasakaalus korpusega kattuvad sõnavormid

Foorumite ja ajakirjanduse ühises sõnavaras on sõnad *on* ja *ja* on sageduse poolest samal kohal, kuid foorumite sagedusloendis on järgmised *ei, et* ja *kui*, ajakirjanduse sagedusloendis *et, ei* ja *kui*.

Foorumite ja ilukirjanduse ühises sõnavaras on foorumite sagedusloendis esikohal sõna *on*, ilukirjanduse vastavas loendis sõna *ja*. Teistel kohtadel on vastavalt *ja* ja *ei*, kolmandatel *ei* ja *on*, neljandatel *et* ja *ta* ning viiendatel kohtadel *kui* ja *et*.

Foorumite ja teadustekstide ühises sõnavaras on esimestel kohtadel vastavalt *on* ja *ja*, teistel *ja* ja *on*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ka* ning viiendatel kohtadel mõlemal *kui*.

Foorumite ja tasakaalus korpuse ühises sõnavaras on esimestel kohtadel vastavalt *on* ja *ja* nagu eelmistel, teistel kohtadel ka nagu eelmisel sagedusloendipaaril *ja* ja *on*, kolmandatel mõlemal *ei*, neljandatel mõlemal *et* ja viiendatel kohtadel ka ühtemoodi sõna *kui*.

Foorumite sagedusloendi esimese saja sõnavormi hulgas on 3 nimisõnavormi, 1 omadussõnavorm, 27 asesõnavormi, 16 tegusõnavormi, 40 määrsõna, 11 sidesõna ja 2 kaassõna.

Nimisõnavormid on *asi, näiteks, aru*. Nimetavakäändeline laia ja ebamäärase tähendusega nimisõnavorm *asi* on foorumitekstides palju kasutuses mitmesugustes püsiväljendites (*asi on selles, selline asi, et, asi käib nii*) või siis mingi üldise olukorra tähistajana:

(11) asi täitsa vaadatav

Saavakäändeline parasiitsõna *näiteks* esineb foorumitekstides mitmel pool. Selle sõna võib ära jätta ilma et tähendus oluliselt muutuks:

(12) Samas aitab näiteks foorumi KKK lugemine... ja sealt näite 1:1 kopeerimine

Ainus omadussõnavorm on *hea*. Omadussõnavormi *hea* kasutatakse nagu teisteski tekstides atribuudina:

(13) Selle raha eest on suht hea tase...

Asesõnavormid on isikuliste asesõnade vormid *ma, sa, ta, mul, minu, mina, mulle, sul*; küsiv-siduvad asesõnad *mis, kes, mida*; näitavad asesõnad *sama, see, seda, need, selle, selline, sellest, neid, nad*; umbmäärased asesõnad *üks, midagi, mingi, kõik, keegi* ja enesekohased asesõnad *oma, ise*.

Tegusõnavormid on *olema* vormid *on, oli, pole, ole, oleks, olen, olla, olema, olnud*; *saama* vormid *saab, saa* ja muud tegusõnavormid *teha, tuleb, peaks, võib, tea*.

Määrsõnad on asemäärsõnad *miks, kus, seal, siin, kas, kuidas, siis, nii, kunagi*; rõhumäärsõnad *ka, veel, juba, palju, lihtsalt, ainult, rohkem, kah, enam, isegi, üldse, kohe, ikka, küll, ju, vist, väga, mitte, ei, eriti, päris, kõige, just, no, nagu, ära, jah, nüüd* ja abimäärsõnad *välja, üle, vaja*. Sõnavorm *üle* võib tekstis olla nii abimäärsõna kui kaassõna:

(14) Siinkohal pakud sa küll üle.

(15) Kõik tuleb üle võrgu NAS seadmest või striimin internetist.

Afiksaaladverb *vaja* on ühendverbide *vaja olema* ja *vaja minema* koosseisus, *olema* vorm võib olla ka välja jäetud:

(16) Hea et nägin teemat kuna endal ka vaja just teha üks võõrkeelne töö-Igatahes Tänan

Sidesõnad on *ja, et, kui, aga, või, ning, kuid, sest, kuna, ehk, ega*.

Kaassõnad on *peale, pärast*.

Järgnevalt analüüsitakse foorumite sõnavormide sagedusi võrdluses tasakaalus korpuse sõnavormide sagedusega. Sõna *on* on foorumite sagedusloendis sageduselt kõrgemal kohal kui tasakaalus korpuse sagedusloendis. Sõna *ja* on tasakaalus korpuses sageduse poolest järjekorras kõrgemal kohal. Sõnad *ei, et, kui, mis, või* ja *isegi* on mõlemas loendis sageduse järjekorra poolest samal kohal.

Sõna *siis* on foorumite sagedusloendis tunduvalt sagedam, olles 6. kohal, tasakaalus korpuses on see sõna 15. kohal. Sõna *see* on foorumites sagedam, samuti sõnad *ma*, *aga*, *seda*, *nii*, *pole*, *sa*, *selle*, *ära*, *kas*, *midagi*, *veel*, *ole*, *mingi* (palju sagedam, foorumite korpuses 27. kohal, tasakaalus korpuses 182. kohal), *ikka*, *mul*, *küll*, *oleks*, *väga*, *olen*, *ju*, *ise*, *hea*, *seal*, *palju*, *peale*, *saab*, *minu*, *jah* (palju sagedam, sageduse järjekordade suhe 48:249), *teha*, *need*, *lihtsalt*, *keegi*, *asi*, *vist*, *saa*, *siin*, *mina*, *mulle*, *peaks*, *rohkem*, *eriti*, *selline*, *päris*, *kuna*, *tea*, *ehk*, *just*, *olema*, *no* (suhe 84:537), *vaja*, *kah* (suhe 86:1055), *sul*, *üldse*, *miks*, *sama*, *kohe*, *aru* ja *kunagi*.

Tasakaalus korpuses on sagedamad sõnad *ka*, *oli*, *ta*, *oma*, *nagu*, *ning*, *mitte*, *kes*, *kõik*, *juba*, *mida*, *välja*, *nüüd*, *kuid*, *üks*, *olla*, *ainult*, *neid*, *sest*, *kuidas*, *tuleb*, *nad*, *kus*, *võib*, *üle*, *kõige*, *enam*, *ega*, *sellest*, *näiteks*, *pärast* ja *olnud*.

5.1.3 Kommentaaride kõige sagedasemad tasakaalus korpusega kattuvad sõnavormid

Kommentaaride ja ajakirjanduse ühises sõnavaras on esimestel kohtadel jällegi *ja* ja *on*, teistel *on* ja *ja*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ei* ning viiendatel kohtadel mõlemal sagedusloendil *kui*.

Kommentaaride ja ilukirjanduse ühises sõnavaras on esimestel kohtadel mõlemal *ja*, teistel *on* ja *ei*, kolmandatel *ei* ja *on* nagu ka foorumite ja ilukirjanduse ühises sõnavaras, neljandatel *et* ja *ta* jälle nagu foorumite ja ilukirjanduse ühises sõnavaras, viiendatel kohtadel taas nagu eelmainitud kahe korpuse ühises sõnavaras *kui* ja *et*.

Kommentaaride ja teadustekstide ühises sõnavaras on esimestel kohtadel mõlemal *ja*, teistel mõlemal *on*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ka* ning viiendatel kohtadel mõlemal *kui*.

Kommentaaride ja tasakaalus korpuse ühises sõnavaras on viis sagedaimat sõna mõlemal samad: *ja*, *on*, *ei*, *et* ja *kui*.

Kommentaaride sagedusloendi esimese saja sõnavormi hulgas on 3 nimisõnavormi, 1 omadussõnavorm, 31 asesõnavormi, 17 tegusõnavormi, 31 määrsõna, 11 sidesõna ja 6 kaassõna.

Nimisõnavormid on *aru*, *inimene*, *elu*. Nimetavakäändeline laia kasutusega nimisõnavorm *aru* esineb väljendverbi *aru saama* koosseisus:

(17) püüan aru saada

Väljendverbi koosseisus samuti nimisõnana:

(18) aga varsa aru!!

Nimetavakäändeline sõna *inimene* on kasutuses inimesest üldiselt rääkimise korral:

(19) Kahjuks ei saa inimene oma sünniaega valida.

Sõna *inimene* on kasutusel ka üttena koos omadussõnaga *hea*:

(20) Ütle hea inimene kes mind tööle tahab võtta?

Nimetava-, omastava- ja osastavakäändeline nimisõnavorm *elu* on kasutuses atribuudina:

(21) Elu realiteedid on aga tundetud.

Ainus saja sagedasema sõna hulka mahtuv omadussõnavorm on *hea*. Omadussõnavormi *hea* kasutatakse nagu foorumitekstides atribuudina:

(22) hea mõte, kõik diilerid ja narkarid tuleks kinni saata.

Asesõnavormid on isikuliste asesõnade vormid *ma*, *sa*, *ta*, *nad*, *minu*, *meie*, *tema*, *mina*, *me*, *sinu*, *mul*, *mulle*; küsiv-siduvad asesõnad *mis*, *kes*, *mida*; näitavad asesõnad *see*, *seda*, *selle*, *need*, *nende*, *neid*, *sellest*, *selles*; umbmäärased asesõnad *üks*, *midagi*, *kõik*, *keegi*, *mingi* ja enesekohased asesõnad *oma*, *ise*.

Tegusõnavormid on *olema* vormid *on*, *pole*, *oli*, *ole*, *oleks*, *olen*, *olla*, *oled*, *olema*, *olnud*; *saama* vormid *saa*, *saab*; *pidama* vormid *peaks*, *peab* ja muud tegusõnavormid *tuleb*, *võib*, *teha*.

Määrsõnad on asemäärsõnad *miks*, *kus*, *siin*, *seal*, *kas*, *siis*, *nii*, *kuidas*; rõhumäärsõnad *ka*, *veel*, *juba*, *ainult*, *palju*, *enam*, *lihtsalt*, *isegi*, *rohkem*, *tõesti*, *ikka*, *küll*, *ju*, *ei*, *väga*, *just*, *mitte*, *nüüd*, *no*, *nagu*, *ära* ja abimäärsõnad *välja*, *üle*, *vaja*. Sõnavorm *üle* võib tekstis olla nii abimäärsõna kui kaassõna:

(23) Jääb üle oodata, millal paavst sõna võtab.

(24) Mina olen Bushi üle uhke.

Nagu ka foorumites, on ka siin afiksaaladverb *vaja* ühendverbide *vaja olema* ja *vaja minema* koosseisus, *olema* vorm võib olla ka välja jäetud:

(25) Kui kulutad rohkem, kui mõistlikult vaja, läheb asi luksuseks.

Sidesõnad on *ja, et, kui, aga, või, vaid, sest, ning, kuid, ega, ehk.*

Kaassõnad on *vastu, peale, eest, pärast, kohta, poolt*.

Nii kommentaaride kui ka tasakaalus korpuse sagedusloendites on sageduselt samal kohal sõnad *ja, on, ei, et, kui, mida* ja *kõik.*

Kommentaaride sagedusloendis on sageduse poolest kõrgemal kohal sõnad *see, ka, siis, aga, oma, ma, pole, seda, kes, nii, kas, sa, siin, mitte, ole, ju, vaid, ära, midagi, oleks, ikka, küll, ise, need, minu, meie, olen, ega, ainult, neid, saa* (suhe 54:102), *olla, oled* (suhe 57:226), *palju, miks, mina, me, sellest, peale, keegi, aru, inimene* (suhe 73:220), *sinu* (suhe 76:329), *ehk, teha, lihtsalt, peaks, isegi, elu, olema, vaja, hea, selles, mul, poolt, no* (suhe 100:537), *tõesti* ja *mingi*.

Tasakaalus korpuse sagedusloendis on sageduse poolest kõrgemal kohal sõnad *mis, või, ta, selle, oli, nagu, veel, nad, sest, ning, juba, kuid, nende, välja, tema, üks, väga, kus, vastu, nüüd, tuleb, kuidas, võib, saab, eest, üle, enam, seal, rohkem, pärast, olnud, just, peab, kohta* ja *mulle*.

5.1.4 Uudisgruppide kõige sagedasemad tasakaalus korpusega kattuvad sõnavormid

Uudisgruppide ja ajakirjanduse ühises sõnavaras on esimestel kohtadel mõlemal *on*, teisel mõlemal *ja*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ei* ning viiendatel kohtadel mõlemal *kui*. Uudisgruppide ja ilukirjanduse ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *ei*, kolmandatel *ei* ja *on*, neljandatel *et* ja *ta* ning viiendatel *kui* ja *et*. Uudisgruppide ja teadustekstide ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *on*, kolmandatel *ei* ja *et*, neljandatel *et* ja *ka* ning viiendatel kohtadel *et* ja *kui*. Uudisgruppide ja tasakaalus korpuse ühises sõnavaras on esimestel kohtadel *on* ja *ja*, teistel *ja* ja *on*, kolmandatel mõlemal *ei*, neljandatel mõlemal *et* ja ka viiendatel mõlemal *kui*.

Uudisgruppide sagedusloendi esimese saja sõnavormi hulgas on 3 nimisõnavormi, 1 omadussõnavorm, 26 asesõnavormi, 19 tegusõnavormi, 33 määrsõna, 1 hüüdsõna, 12 sidesõna ja 5 kaassõna.

Nimisõnavormid on *asi*, *näiteks*, *asja*. Nimetavakäändeline laia ja ebamäärase tähendusega nimisõnavorm *asi* on uudisgruppide tekstides, nagu ka foorumitekstides, palju kasutuses mitmesugustes püsiväljendites (asi on selles, selline asi, et, asi käib nii) või siis mingi üldise olukorra tähistajana:

(26) Asi ei vasta lepingutingimustele

Siingi oli ainus saja sagedasema sõna hulka mahtuv omadussõnavorm *hea*. Omadussõnavormi *hea* kasutatakse nagu kommentaaride tekstides atribuudina:

(27) Eriti hea koostöö WIN XP-ga.

Asesõnavormid on isikuliste asesõnade vormid *ma*, *ta*, *sa*, *minu*, *mul*, *mina*; küsiv-siduvad asesõnad *mis*, *mida*, *kes*; näitavad asesõnad *sama*, *see*, *seda*, *selle*, *need*, *selline*, *neid*, *sellest*, *nad*, *nende*; umbmäärased asesõnad *üks*, *keegi*, *midagi*, *mingi*, *kõik* ja enesekohased asesõnad *oma*, *ise*.

Tegusõnavormid on *olema* vormid *on*, *pole*, *oli*, *ole*, *oleks*, *olen*, *olla*, *olema*, *olemas*; *saama* vormid *saab*, *saa*, *saada*, *sai*; *pidama* vormid *peaks*, *peab*; *võima* vormid *võib*, *võiks* ja muud tegusõnavormid *teha*, *tuleb*.

Määrsõnad on asemäärsõnad *seal*, *siin*, *kus*, *kas*, *siis*, *nii*, *kuidas*; rõhumäärsõnad *ka*, *veel*, *juba*, *ainult*, *palju*, *lihtsalt*, *rohkem*, *enam*, *isegi*, *küll*, *ikka*, *ju*, *ei*, *mitte*, *väga*, *vist*, *just*, *eriti*, *nüüd*, *ära*, *nagu* ja abimäärsõnad *välja*, *vaja*, *üle*, *läbi*, *ette*.

Ainus hüüdsõna on *tere*.

Sidesõnad on *ja*, *et*, *kui*, *aga*, *ning*, *ehk*, *kuid*, *kuna*, *vaid*, *sest*, *ega*, *või*.

Kaassõnad on *peale*, *kohta*, *eest*, *jaoks*, *sisse*.

Uudisgruppide sagedusloendis on sageduse poolest kõrgemal kohal sõnad *on*, *siis*, *ka*, *see*, *aga*, *mis*, *või*, *kas*, *seda*, *nii*, *pole*, *selle*, *ole*, *veel*, *oleks*, *mitte*, *küll*, *ära*, *saab*, *ise*, *ikka*, *keegi* (suhe uudisgruppide ja tasakaalus korpuse sageduse vahel 32:126), *midagi*, *peale*, *asi* (suhe 38:242), *ju*, *väga*, *teha*, *peaks*, *mingi*, *ainult*, *olen*, *minu*, *seal*, *ehk*, *tere* (suhe 55:1723), *mul*,

vaja, kuidas, palju, olema, saa, selline, vist, kuna, sama, näiteks, lihtsalt, siin, mina, hea, olemas (suhe 81:296), *just, asja* (suhe 86:315), *võiks, saada, eriti, jaoks, sisse* ja *sai.*

Tasakaalus korpuse sagedusloendis on sageduse poolest kõrgemal kohal sõnad *ja, ma, oli, oma, nagu, mida, ning, välja, juba, ta, kes, võib, kõik, kuid, tuleb, üks, kus, olla, need, üle, neid, kohta, nüüd, vaid, sellest, eest, rohkem, nad, enam, sest, ega, läbi, ette, peab, nende* ja *isegi*.

Sageduse poolest samal kohal on sõnad *ei, et, kui* ja *sa*.

Nagu näha on korpusest sõltumata esimesed sada sõna suures osas kattuvad. Uudisgruppide korpuses esineb hüüdsõna *tere*, teistes korpustes esimese saja hulgas seda ei esine. Põhjuseks on see, et uudisgruppides on kombeks teretada. Uue meedia korpustes on sõnad eripärasemad, mis tähendab, et uuel meedial on kirjakeelega võrreldes rohkem mitte-ühiseid sõnu.

5.2 Ainult uue meedia tekstides esinevad sõnavormid

Alustuseks kirjeldan esimest kümmet sõnavormi, mis midagi võivad tähendada (st võtan uuele meediale iseloomulike sõnavormide tabelist välja need read, kus sõnavormi tulbas pole märke ega suvalisi tähtede jadasid) ja on uue meedia sagedusloendites, kuid puuduvad kirjakeele sagedusloendites.

Seejärel analüüsin sageduse järjekorras neid uue meedia korpuse sagedusloendis esinenud sõnavorme, mida ei leidunud tasakaalus korpuse põhjal koostatud sagedusloendites ning jagan iga uue meedia tekstiklassi iseloomulikud sõnad funktsioonide järgi gruppidesse ja toon kasutusnäited.

Kõikide ainult uue meedia sagedusloendites esinenud sõnavormide loendid koos välja filtreerimata märgijadadega ja aluseks olnud tabelid uuele meediale iseloomulike sõnade kohta on olemas aadressil http://kodu.ut.ee/~lkristja/.

5.2.1 Foorumite tekstides esinevad sõnavormid koos välja filtreerimata märgijadadega

Tabel 9. Foorumitele iseloomulikud sõnavormid koos välja filtreerimata märgijadadega.

Sõnavorm	**Järjekorranumber**	**Sagedus**
jne	104	7962
a	109	7556
vms	189	4654
‹	195	4572
%	201	4378
eesti	214	4149
/	216	4124
+	218	4096
#	232	3880
the	250	3528

x	264	3303
Eesti	300	2793
ok	307	2714
>	323	2586
says	370	2229
krt	374	2222

Sõnavara, mis on foorumite sagedusloendis, aga mida pole normeeritud kirjakeelt esindavas korpuses on lühend *jne*, täht *a*, mis on kasutusel kas *aga* või *ahhaa* tähenduses, lühend *vms*, *eesti* väikese algustähega, inglise keele definiitne artikkel *the*, täht *x*, mis on peamiselt kasutusel mõõtmete sõnastamiseks, korrutusmärgina või tundmatu muutujana:

(28) 64 x 50 pixels foorumi defauldid on osad suuremad, kui 50x50

(29) Aga see muidugi kordab filme 5 x nädalas.

(30) Samas ei tohiks ka X olla võrdne X-ga vaid peaks olema määramatus.

Veel on foorumitele iseloomulik *Eesti* suure algustähega, *ok*, mis väljendab nõusolekut, *says*, mis on automaatselt genereeritud ja partikkel *krt*, mis on lühend sõnast *kurat*. Lühendid ja pärisnimed sisalduvad uue meedia tekstides, kuna neid ei olnud kõrvale jäetud, erinevalt kirjakeele korpuste sagedusloenditest, millest on juttu osa "Materjali kirjeldus" viimases lõigus. Nagu öeldud osas 2, on tasakaalus korpuse põhjal sagedusloendi koostamisel välja jäetud lühendid, pärisnimed ja genitiivatribuudid.

Nüüd tuuakse näiteid foorumitele iseloomulike sõnavormide kasutuse kohta.

Esimeses grupis on üksikud tähed. Foorumites tulevad üksikud tähed sellest, et foorumites on kohati kasutatud sõrendust ja lühendeid:

(31) nii et aitäh k õ i k i d e l e sest matude oli nii kena

(32) R. I. P.

Üksikud a tähed võivad esineda tähenduses *aga* ja *ahhaa* ning ka inglise keele indefiniitse artikli funktsioonis:

(33) A mis sa arvad , mitu punkti tulex juurde,

(34) a no eks õhtul chekin ära kamh.

(35) a dawn of fantasy on tõeliselt lame nimi

Üksik *i* täht võib olla ka inglise keele ainsuse esimese isiku asesõna:

(36) I have a message for you.

Teine grupp sõnu on lihtsustatud, mugandatud või lühendatud sõnad. Näiteks sõna *aint* mis on lühend sõnast *ainult*:

(37) See teema aint uutele tegijatele.

Veel kuuluvad siia alla normeeritud kirjakeeles kasutatavad lühendid:

(38) st. need nö kriitilised?

(39) see nn "ulmefoorum" toetub eelkõige filmidele ja siis paar mingit väikest muud teemat.

Kolmandas grupis on inglisekeelsed sõnad, kuna kirjutatakse ka inglise keeles:

(40) i like snow!

(41) Edit: päris palju aastaid juba.

Näites (13) on kasutatud sõna *edit*, mis lisatakse siis kui oma postitust muudetakse. See aitab märgata vanal postitusel uut infot.

Veel esineb ka sõna *says*, mis viitab tsitaadile ja on selle alguses:

(42) batoonike says:

Neljas grupp on keelemäng, oletatavasti tahtlik kirjakeele normi rikkumine, enamasti sõnavormide kokku kirjutamine, mille eesmärgiks on sagedasti koos esineva tervikliku mõistesisuga üksuse kokkukuuluvuse rõhutamine ka vormis (Soodla 2010):

(43) niiet mina ei ole süüdi

(44) minuarust ei ole nii, et kõik ühel ajal hüppavad.

Viiendas grupis on arvutialane sõnavara:

(45) ise m6tlesin et vist tuleks windows uuesti peale panna...

(46) See on nagu diskett... või kõvaketas.

Kuuendas grupis on sõnad, mis on teistmoodi kirjutatud tehnilistel põhjustel, erinevused normeeritud kirjapildist puudutavad peamiselt täpitähti (ameerika klaviatuur):

(47) Nyyd on siin menyy kaa.

(48) et ei ole liiga v2lja venitatud, m6ni nurk liiga lai jne.

Seitsmendas grupis on pärisnimed:

(49) Ja Eesti keelt kah minu teada seal ei olnud.

Kaheksandas grupis on partiklid (*novot*, *irw*, *lol*, *krt*) ja lauselõpulised küsipartiklid (*vä*):

(50) novot sellest ma ju rääkisingi

(51) irw, panin mööda vist

(52) kurihiir? Lol

(53) üleeile vist vä

(54) Miks sa krt. üldse eksisteerid?

5.2.2 Kommentaaride tekstides esinevad sõnavormid koos välja filtreerimata märgijadadega

Tabel 10. Kommentaaridele iseloomulikud sõnavormid koos välja filtreerimata märgijadadega.

Sõnavorm	**Järjekorranumber**	**Sagedus**
eesti	75	2297
Eesti	80	2161
jne	102	1686

to	114	1505
Sa	116	1484
Kingo	140	1279
;	144	1241
vene	160	1152
Jumala	182	991
Eestis	212	866
a	219	843

Kommentaaride keelele iseloomulik sõnavara on tasakaalus korpuse keelega võrreldes sõna *eesti* väikese ja suure tähega (kuna uuest meediast ei välistatud pärisnimesid) ning jällegi lühend *jne*. *To* tähistab seda, kui keegi suunab oma ütluse kellelegi:

(55) to Aivar,

Sa suure algustähega on tavaline asesõna. *Kingo* on pärisnimi, mis näitab, et nime *Kingo* on kommentaarides palju mainitud. *vene* väikese algustähega on genitiivatribuut, mida on kommentaarides palju mainitud. *Jumala* on kasutusel peamiselt afektiivsetes püsiühendites:

(56) Ärge Jumala eest laske Keskerakonda enam võimule!!

(57) Ei tohi alluda tänavapoliiti(kale)kutele, jumala pärast!!!

Sõnavorm *Eestis* on kasutatud Eesti riigile viitamisel ja on sõna *Eesti* inessiivi vorm. Ka kommentaarides on üksik a täht *aga* tähenduses ja ka ladinakeelsetes väljendites nagu *a la* ja *a priori*.

Järgmisena esitatakse näiteid kommentaarides esinevate iseloomulike sõnavormide kohta. Esimeses grupis on taas üksikud tähed. Foorumitega võrreldes on kommentaarides üksikuid tähti vähem. Üksikut a tähte kasutatakse sarnaselt foorumitele:

(58) A me seome talle vahel tugitoolile rihma ette.

Teises grupis on pärisnimed, eriti *Jumal* ja *Jeesus*, kuna tasakaalus korpuse põhjal koostatud sagedusloenditest on pärisnimed välja jäänud:

(59) Usk Jumalasse ei tähenda alati ristiusku..

(60) isegi Kingo on magama lainud va

Kolmandas grupis on lühendid nagu *nn* ja *nö*:

(61) Mida rohkem nn "establishment"-i vastane seda parem.

Palju sagedasi eripäraseid sõnu on põhjustatud asjaolust, et tasakaalus korpuse sagedusloendist on pärisnimed, genitiivatribuudid ja lühendid välja jäetud ja ka asjaolust, et kommentaarides on mitmed sõnavormid keset lauset olnud kirjutatud suure tähega, näiteks *Sina* ja *Jumal*.

Neljandas grupis on partiklid (*novot*, *irw*) ja lauselõpulised küsipartiklid (*vä*):

(62) Novot ja selle lausega suutsidki mind viia segadusse.

(63) Joel mind ei tahtnudki... irw

(64) Mängid lolli vä

Kommentaaride tekstides on vähem inglise keelt, kuna kommenteeritakse eesti keeles, arvutialast sõnavara pole ka, kuna selle kohta antud tekstides kommentaarid puuduvad. Klaviatuuri või operatsioonisüsteemi piirangute tõttu on ka siin kohati eesti keelele omaste tähtede asemel kasutatud numbreid või muid tähti. Leidub ka keelemängu. Ühikuid on vähesel määral, kuna seda ei kommenteerita, see on pigem teiste uue meedia korpustele omane.

5.2.3 Uudisgruppide tekstides esinevad sõnavormid koos välja filtreerimata märgijadadega

Tabel 11. Uudisgruppidele iseloomulikud sõnavormid koos välja filtreerimata märgijadadega.

Sõnavorm	**Järjekorranumber**	**Sagedus**
a	82	4415
;	94	3850

>	96	3804
jne	109	3406
andres	139	2813
x	155	2511
eesti	157	2491
+	159	2430
Eesti	160	2429
Soolo	221	1899
s	229	1817
t	244	1702
terv	246	1690

Uudisgruppide keelele iseloomulik sõnavara on tasakaalus korpuse keelega võrreldes täht *a*, mis on siingi kasutusel *aga* ja *ahhaa* tähenduses, lühend *jne* ja pärisnimi *andres* väikese algustähega. *X* täht on kasutusel peamiselt tundmatu muutujana. *eesti* väikese ja suure tähega on kasutusel vastavalt keelest ja riigist rääkides. *Soolo* suure algustähega on uudisgruppides sagedasti postitanud inimese perekonnanimi. *S* täht on kasutusel lühendites *P.S.*, *S.O.* ja *S.T.* ja ka nimede lühendamisel. *T* täht samamoodi lühendis *S.T.* ja nimede lühendamisel.

Lõpuks tuuakse näiteid ka uudisgruppidele iseloomulike sõnavormide kasutuse kohta. Nagu ikka on esimeses grupis üksikud tähed. Uudisgruppides tähistavad üksikud tähed sõrendust:

(65) E - M A I L M A R K E T I N G ! !

Üksik v täht on lühend sõnast *või*:

(66) usun et k6ik suudavad kirjad l2bi lugeda siis kui on ylevalt alla v alt yles postitatud.

Teises grupis on inglisekeelsed sõnad, näiteks suhtluskeskkonna poolt tekitatud sõna *wrote*, mis tähistab tsitaati:

(67) Arvi Pruuli wrote:

Kolmandas grupis on lühendatud sõnad (*terv*, *ntx*, *tel*, *nr*):

(68) Terv, Tarmo

(69) Ntx tööl sai kolleeg pangast pakkumise summale 0.- krooni.

(70) Kuna tel. liini ei ole, siis on valik suht piiratud.

(71) V.t. menüü nr. 15.

Neljandas grupis on jällegi pärisnimed, millest sagedaim on *andres*:

(72) Andres Soolo wrote:

Siit järeldub, miks korpuses on esimese 20 sõna hulgas just need kolm sõna. Härra Andres Soolo oli uudisgruppides sage postitaja ja seetõttu oli just see automaatselt tekitatud rida uudisgruppides ka kõige sagedasem.

Viies grupp koosneb lühenditest:

(73) Ja niiviisi ca 4 aastat.

(74) Abiks oleks ka lingid, viited jms.

Kuues grupp on ühikud, peamiselt mõõtühikud (*km*, *MB*, *mm*, *cm*), kuna neid oli uudisgruppide tekstides rohkelt:

(75) 15000 km on UTOOPIA!!

(76) Kas kirjutaski 900 MB ära??

(77) Rusikareegliga kuskil 45-55 mm vahemikus

(78) All oli 30 cm vedrusid ja siis 20 cm ja siis veel 5.

Seitsmes grupp on partiklid (*novot*, *irw*, *lol*, *krt*) ja lauselõpulised küsipartiklid (*vä*):

(79) Novot, seda ma kardangi.

(80) irw.. tundub, et see ei toimi

(81) midagi aru ei saa... lol.

(82) Kaubamärk maksab vä?

(83) krt vale pakkija oli

5.2.4 Kogu uue meedia tekstides esinevad sõnavormid koos välja filtreerimata märgijadadega

Tabel 12. Kogu uuele meediale iseloomulikud sõnavormid koos välja filtreerimata märgijadadega.

Sõnavorm	**Järjekorranumber**	**Sagedus**
jne	103	13054
a	107	12814
eesti	153	8937
Eesti	198	7383
+	212	6926
%	215	6895
>	224	6737
;	239	6147
/	248	5953
x	252	5916
vms	253	5870
‹	295	4657
#	332	4087
to	343	3946

the	360	3783
vene	376	3668
ok	378	3657

Kogu uue meedia keelele iseloomulik sõnavara on tasakaalus korpuse keelega võrreldes lühend *jne*, täht *a*, pärisnimi *eesti* väikese ja suure tähega, täht *x*, lühend *vms*, ingliskeelne eessõna *to*, definiitne artikkel *the*, sõna *vene* ja sõna *ok*. Kokkuvõtteks võib öelda, et palju on lühendeid ja pärisnimesid ning inglise keelest pärinevaid sõnu. Rohkesti on ka üksikuid tähti, mille allikaks on enamasti sõrendatult trükitud sõnad, mille tähed olid üksteisest tühikutega eraldatud.

Selle osa kokkuvõtteks võib öelda, et uues meedias on palju inglise keelt ja lühendeid, mistõttu üksikuid sõnu ja tähti ei saa ilma kontekstita üheselt tõlgendada. Inglise keelt leidub kõige rohkem foorumitekstides, kuna seal suheldakse peamiselt arvutialastel teemadel ja arvutialane terminoloogia on põhiliselt inglisekeelne (kuigi on hakatud kasutama juba ka eestikeelseid väljendeid nagu *RAM* asemel *mälu*, *Hard Drive*asemel kõvaketas jne. Partiklitest sagedaimad, kuigi sageduse positsioonidelt varieeruvad, olid erinevates uue meedia korpustes enam-vähem samad (*novot*, *irw*, *lol*, *krt*). Kommentaaride foorumites kasutati kõige vähem ühikuid, kuna kommentaarid annavad enamasti hinnangu või avaldavad arvamust. Kõigis uue meedia tekstides domineerisid üksikud tähed, mis kirjakeeles on harvem nähtus (kui sõrendus välja arvata).

Kokkuvõte

Käesolevas bakalaureusetöös käsitleti eesti uue meedia keele sõnavarasagedusi võrdluses normeeritud kirjakeelega ja analüüsiti sõnavorme järjekorranumbrite järgi sageduse kahanemise alusel järjestatud loendis ning tõi korpuse tekstist näiteid sõnavormide kasutuse kohta.

Esmalt andis töö ülevaate eelnevatest uurimustest, mis on mingil määral seotud selle tööga. Siis esitati töö materjal ja käik, st kirjeldati, millisest allikmaterjalist ja millisel viisil olid koostatud siin töös esitatavad sõnavormide sagedusloendid ja millisel viisil oli neid võrreldud normeeritud kirjakeele vastavate loenditega. Järgmisena esitati tabelid sõnavormide hulga, kattuvuse ja selle kohta, kui suure osa moodustas kattunud osa vastavatest korpustest. Viimased kaks esitati nii arvuliselt kui protsentuaalselt. Vaadati sagedusloendeid nii koos ühekordsete sõnavormidega kui ka ilma. Sõnavormide kattuvuste kohta esitati ka *Venni* diagrammid. Viimases osas analüüsiti uue meedia sagedusloendi esimest sadat sõnavormi, millele leidusid vasted tasakaalus korpuses, seejärel analüüsiti sageduse järjekorras neid uue meedia korpuse sagedusloendis esinenud sõnavorme, mida ei leidunud tasakaalus korpuse põhjal koostatud sagedusloendites. Iga uue meedia korpuse kohta esitati ka näiteid ka talle iseloomulike sõnavormide kohta.

Selle töö eesmärk oli koostada uue meedia keele sõnavormide sagedusloendid ja võrrelda neid sagedusloendeid normeeritud kirjakeele baasil koostatud sarnaste loenditega. Sagedusloendid koostati *shelli* skrpitide abil. Võrdlused tehti tabelite põhjal ja illustreeriti *Venni* diagrammidega. Uue meedia korpustele iseloomulike sõnade sagedusloendite tabelitesse jäid küll mõned märgid sisse, kuid see ei mõjutanud oluliselt analüüsi kulgu, sest analüüsiti ainult neid märgijadasid, mis omasid tähendust. Käesoleva töö tulemuseks oli palju uut informatsiooni ja järeldusi uuritavate sagedusloendite kohta.

Kirjandus

Baron, Naomi 2008. Always On. Oxford: University Press.

Cantoni, Lorenzo jt 2006. Internet. London and New York: Routledge.

Crystal, David 2001. Language and the Internet. Cambridge University Press.

Gurak, Laura 2001. Cyberliteracy: Navigating the Internet with Awareness. London: Yale University Press.

Kaalep, Heiki-Jaan jt 2002. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastuse trükikoda.

Kasik, Reet 2011. Stahli mantlipärijad. Tartu: Tartu Ülikooli Kirjastus.

Kirt, Riin 2013. Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning „Eesti keele seletava sõnaraamatu“ märksõnaloendi võrdlus; http://www.murre.ut.ee/arhiiv/naita.php?t=kasikiri&id=5095. Vaadatud 13.06.2014.

Leech, Geoffrey jt 2001. Word Frequencies in Written and Spoken English. Harlow: Pearson Education Limited.

Oja, Anni 2006. Eesti keel internetis. – Keel ja Arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Tartu: Tartu Ülikooli kirjastus, 259–267.

Oja, Anni 2010. Sissevaateid internetisuhtlusesse. – Oma Keel, nr 20, lk 11-18; http://www.emakeeleselts.ee/omakeel/2010_1/OK_2010-1_02.pdf. Vaadatud 13.06.2014.

Soodla, Karin 2010. Morfoloogilisi, morfosüntaktilisi ja sõnamoodustuslikke nähtusi eesti internetikeeles; http://www.murre.ut.ee/arhiiv/naita.php?t=kasikiri&id=2669. Vaadatud 13.06.2014.

Tuldava, Juhan 1977. Sagedussõnastik leksikostatistilise uurimise objektina. Tartu: Tartu Riikliku Ülikooli trükikoda.

Viks, Ülle jt 2001. Seadusetekstide grammatiline sagedussõnastik; http://www.eki.ee/teemad/seadused_dic/sonastikud.pdf. Vaadatud 13.06.2014.

Zipfi seadus; http://kodu.ut.ee/~hkaalep/arvutimorf_12/loeng7.htm. Vaadatud 13.06.2014.

Uue meedia korpus; http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/foorumid_uudisgrupid_kommentaarid.php?lang=et/. Vaadatud 13.06.2014.

Summary

The word frequency list of Estonian Internet media and its comparison with the word frequency list of Estonian written language

The purpose of this thesis is to create word frequency lists of Estonian Internet media (corpora which include data from forums, comments and newsgroups) and compare them with similar word frequency lists of the written language of Estonian (corpora which include data from both journalism, fiction and scientific literature separately and as part of the "Balanced Corpus of Estonian").

To form the word frequency lists of the Estonian Internet media (the word frequency lists of the written language of Estonian were already readily available), the subcorpora of Internet media, forums, comments and newsgroups of the University of Tartu's Research Group of Computational Linguistics were used. There were two versions of the corpora – one with repetitions (for example quoting on the forums), the other without repetitions. In this thesis, the latter was used. The files of the corpora are in xml format.

In this thesis, word frequency lists of Estonian Internet media were formed and compared with the word frequency list of the Estonian written language. The version of the word frequency list of the Estonian written language that is arranged by frequency was used.

To form the word frequency lists, command line scripts were used (*shell* scripts). The word frequency lists were joined with the join command after having been created and numbered. In total 16 files were made with the join command. This allowed to compare the frequencies of the respective word frequency lists. The numbering also provided extra information about at what position the word was in the word frequency list.

In the fourth chapter, a quantitative analysis of the corpora was carried out. Tables were formed to show relations between the corpora both as values and percentages. The relations were also depicted with Venn diagrams.

In the fifth and final chapter, a qualitative analysis was performed. It contains analyses of the 100 most frequent words from the Estonian Internet media word frequency list which had counterparts in the word frequency list of Estonian written language. Examples of the use of some words from the Estonian Internet media have also been provided.

Although some symbols were not completely excluded from the word frequency lists of the Estonain Internet media, it had little impact on the process of the analyses carried out because only those sequences of characters were analysed, which had a meaning. The results of this thesis were a lot of new information and conclusions about the analysed word frequency lists.

Mina Kristjan Link

(autori nimi)

(sünnikuupäev: 26.07.1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
UUE MEEDIA TEKSTIDE SAGEDUSSÕNASTIK JA SELLE VÕRDLUS KIRJAKEELE SAGEDUSSÕNASTIKUGA,

(*lõputöö pealkiri*)

mille juhendaja on Kadri Muischnek,

(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 12.06.2014 (*kuupäev*)