doi:10.5128/ERYa7.07

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 7, 111–127

KORPUSLINGVISTILINE LÄHENEMINE EESTI INTERNETIKEELE AUTOMAATSELE MORFOLOOGILISELE ANALÜÜSILE

Kadri Muischnek, Heiki-Jaan Kaalep, Raul Sirel

Ülevaade. Artiklis analüüsitakse eesti uue meedia keelekasutuse e internetikeele leksikaalseid ja ortograafilisi eripärasusi ning nendest tulenevaid automaatsel morfoloogilisel analüüsil kerkivaid raskusi. Esitatakse meetodid nende probleemide lahendamiseks: sagedased kõrvalekalded normeeritud kirjakeelest lahendatakse kasutajasõnastiku abil ning harvem esinevad, kuid regulaarsed kõrvalekalded automaatseid teisendusreegleid rakendades. Korpusespetsiifilise leksika analüüsiks pakutakse välja kasutajasõnastiku automaatse täiendamise meetod. Artikli autorid on seisukohal, et internetisuhtluses kasutatava keelevariandi erinevused normeeritud kirjakeelest on valdavalt kirjutajate teadliku keelemängu tulemus, mitte kehva kirjaoskuse väljendus.*

Võtmesõnad: arvutilingvistika, korpuslingvistika, morfoloogia, morfosüntaks, ortograafia, sõnaliigid, eesti keel

1. Sissejuhatus

Morfoloogiline analüüs on eestikeelsete tekstide korpuslingvistilisel ja/või automaatsel analüüsil oluline etapp. Selle käigus lisatakse igale tekstisõnale tema algvorm e lemma ja info tema grammatiliste kategooriate kohta. Lemma kaudu saab tekstisõna ühendada sõnastikuga, grammatiliste kategooriate kaudu võrrelda eri tekste ja tekstiliike omavahel. Morfoloogiliselt analüüsitud tekst on sisendiks ka automaatse lingvistilise analüüsi järgmistele etappidele: süntaktilisele ja semantilisele analüüsile. Samuti on teadmine lemmade kohta abiks automaatsel infootsingul.

Käesoleva artikli eesmärgiks on analüüsida uue meedia keele erijooni, mis muudavad ta ebamugavaks kirjakeele tarvis loodud morfoloogilise analüsaatori jaoks. Uue meedia keele all on siin mõeldud interaktiivset internetikeelt ehk võrgusuhtluse keelt. Enam-vähem samas tähenduses on eesti keeles kasutatud ka

* Artikli valmimist on toetanud Euroopa Regionaalarengute Fond Eesti Arvutiteaduse Tippkeskuse kaudu ning Haridus- ja Teadusministeerium (sihtfinantseeritav teema sF0180078s08 "Loomulike keelte arvutitöötluse formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine" ja riiklik programm "Eesti keele keeletehnoloogiline tugi").

terminit internetikeel (nt Soodla 2010) ja me kasutame selles artiklis neid väljendeid sünonüümsetena.

Analüüsime morfoloogiaanalüsaatorile tundmatuks jäävaid sõnavorme uue meedia korpuses ja esitame oma meetodi analüsaatori kohandamiseks uuele tekstitüübile.

Internetikeelt on Eesti keeleteaduses käsitlenud Anni Oja (2006, 2010), kelle põhitähelepanu on pööratud selle keelevariandi sotsiolingvistilistele aspektidele. Uue meedia allkeeltest on jututubade keelt ja selle taustaks olevaid suhtlusreegleid ning -tavasid analüüsinud Sigrid Salla (2002) ning veebikommentaare tekstilingvistika vaatepunktist Krista Kerge (2004). Foorumitekstide morfoloogilisi, morfosüntaktilisi ja sõnamoodustuslikke erijooni on uurinud Karin Soodla (2010).

Artikli ülesehitus on järgmine. Esmalt (2. osas) kirjeldame uue meedia korpust, selle koostist ja märgendust ning selgitame lühidalt morfoloogilise analüüsi protsessi ja sellega seotud mõisteid. Kolmandas osas analüüsime morfoloogia-analüsaatorile uue meedia tekstides tundmatuks jäänud sõnavorme. Artikli neljandas osas käsitleme tundmatute sõnade analüüsi võimalikke meetodeid ja viiendas osas uurime täiustatud morfoloogilise analüsaatori töö tulemust.

2. Materjal: uue meedia korpus ja morfoloogiline analüsaator

2.1. Uue meedia korpus

Tartu Ülikooli koondkorpus[1] on üle 200 miljoni sõna suurune tänapäeva kirjaliku eesti keele elektrooniline tekstikogu. Ühe osa sellest moodustab 22 miljoni sõna suurune uue meedia keelekasutuse allkorpus[2]. Uue meedia korpus sisaldab omakorda nelja allkorpust: jututubade tekste u 7 miljonit sõna, uudisgruppide tekste u 8 miljonit sõna, foorumitekste u 5 miljonit sõna ja kommentaaride tekste u 2 miljonit sõna.

Nimetatud allkorpuste märgendus on erinev. Muudes koondkorpuse allkorpustes (peamiselt ajalehtede, ilukirjanduse ja teaduse tekstid) on tekstiüksustena märgendatud lõigud, nende allosadena laused ja omaette üksustena veel pealkirjad ja autorikirjed.

Kommentaaride allkorpus on märgendatud enam-vähem samade põhimõtete järgi, kuid jututubade, foorumite ja uudisgruppide allkorpuste märgendamine lähtus tõdemusest, et jututoasalvestus või uudisgrupi ja foorumi arhiiv on nagu näidendi üleskirjutus: tegelased tulevad lavale, esitavad oma repliigid ja lahkuvad sealt. Kõik kõnelejad, st kasutajanimed ehk pseudonüümid on tähistatud märgendiga <speaker> ja kasutaja kirjutatud tekst on ümbritsetud lõigumärgenditega <p>. Kõneleja kasutajanimi ja tema toodetud tekst on ümbritsetud veel märgendiga <sp> (1). Jututoad sisaldavad ka automaatteateid jututoas osalejate liitumise või lahkumise kohta (2) ja jututoas osalejate kommentaare oma tegevuse kohta (3). Nii jututoas osaleja lausung kui ka automaatsete ja kasutajate loodud kommentaaride juures on märgendiga <time> märgendatud lausungi või kommentaari lisamise aeg.

[1] http://www.cl.ut.ee/korpused/segakorpus/ (26.02.2011).
[2] http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/ (26.02.2011).

(1) <time> 16:26 </time> <sp> <speaker> rohi </speaker> ok aga nyyd minek </sp>

(2) <time> 16:32 </time> <stage> * tessa has joined #Kreisiraadio </stage>

(3) <time> 13:20 </time> <stage> * Camille muigab </stage>

Kommentaaride, foorumite ja uudisgruppide allkorpused on automaatselt lausestatud (s.t lisatud on lausepiiride märgendid) kirjakeele normidest lähtuvalt, mis tähendab seda, et kui lause ei alga suure algustähega, siis ei ole seda eraldi lausena märgendatud.

Omaette probleemiks on laused jututubade tekstides. Automaatselt lausestada pole neid tekste õnnestunud, sest jututubade tekstides ei kasutata lause alguses suurtähte ja ka lauselõpumärgid võivad puududa. Morfoloogiline ühestaja vajab aga teadmist lausepiiride kohta. Otsustasime sel otstarbel võrdsustada lõigu lausega. See lihtsustus vastab sageli tõele, jututubade lõik, s.t ühe kasutaja postitus, on tüüpiliselt lühike lause või (nimisõna)fraas. Kuna lausega võrdsustatud üksus jututubade tekstides ei vasta tüüpilise lause tunnustele (vt EKK: 429–430), kasutame selle kohta siin artiklis nimetust lausung.

Kõigis uue meedia allkorpustes on nendes esinenud meili- ja internetiaadressid asendatud märgendiga <gap desc='hüperlink' />. Korpust tehes üritati ka võõrkeelne tekst korpusest välja jätta ja asendada märgendiga <gap desc='võõrkeelne_tekst' />, kuid nagu edaspidises analüüsis selgub, on võõrkeelset materjali tekstidesse ikkagi hulgaliselt jäänud.

Automaatse morfoloogilise analüüsi sisendiks on nendes korpustes ainult kasutajate tekst, mitte kasutajanimed või automaatteated (2). Samas kuuluvad kasutajate kommentaarid oma tegevuse kohta (3) samuti kasutaja loodud teksti hulka, kuid need on käesoleval etapil morfoloogilisest analüüsist välja jäänud.

Jututubade korpuses oli morfoloogilise analüüsi tarvis vaja teha veel eeltööd, et sisend oleks mõistlik. Sõnade eraldamiseks on jututubades kasutusel tühikud, aga ka kirjavahemärgid. Viimased on sageli järgneva ja emotikonid eelneva sõnaga kokku kleepunud; emotikonid ja kirjavahemärgid on samuti sageli koos. Seetõttu tuli enne morfoloogilist analüüsi sõnad kirjavahemärkidest ja emotikonidest eraldada.

Üks jututubades esinev nähtus on läbustajad. Need on osalejad, kelle eesmärgiks näib olevat teiste osalejate segamine ja vihastamine, milleks nad sisestavad võimalikult väikese vahega mõttetuid märgijadasid ja/või fraaside mitmekordseid kordusi. Tavaliselt õnnestub läbustajal oma postitust paar korda korrata, enne kui ta jututoast välja visatakse. Automaatselt on võimalik selliseid teateid ära tunda selle järgi, et nad on tavalisest postitusest pikemad, sisaldavad pikki korduvaid fraase ja/või korduvad ebatavaliselt palju (*lõika ja kleebi* -meetodit kasutavad oma postitustes küll ka jututubades osalejad ise, aga läbustajad kordavad end oluliselt rohkem). Jätsime sellised postitused morfoloogilisest analüüsist kõrvale ning ei arvestanud neid ka korpuse mahu arvutamisel.

2.2. Morfoloogiline analüsaator

Tekstide morfoloogiliseks analüüsiks kasutasime OÜ Filosoft morfoloogilist analüsaatorit *etmrf*.[3] Tegemist on sama firma programmi ESTMORF (Kaalep, Vaino 2000) edasiarendusega. Võrreldes ESTMORF-iga on *etmrf*-il suurem leksikon – 71 000 sõna – ja ka mõneti kvaliteetsem liitsõnade ja produktiivsete tuletiste analüüsialgoritm.

Morfoloogiline analüsaator "näeb" tekstis korraga ainult ühte sõnavormi ja lisab sellele kõik võimalikud analüüsid konteksti arvestamata. Konkreetses kontekstis ainuõige tõlgenduse väljavalimist nimetatakse morfoloogiliseks ühestamiseks ja sellest käesolevas artiklis juttu ei tule.

Meie jaoks oli oluline, et *etmrf*-i käitumist morfoloogilise analüsaatorina saab muuta, kui anda talle sobiv kasutajasõnastik. Kasutajasõnastik on tekstifail, milles igal real on nii analüüsitav sõnavorm kui ka väljund, mille *etmrf* meie arvates peaks antud sõnale andma. Iga analüüsimist vajava sõna korral kontrollib *etmrf* kõigepealt kasutajasõnastikust, kas sobiva analüüsi saaks võtta otse sealt. Alles juhul, kui sealt vastust ei saa, minnakse tegema morfoloogilist analüüsi. Seega saab kasutajasõnastikku panna sõnu, mida *etmrf* muidu analüüsida ei suudaks, aga ka sõnu, mis meie arvates peaks konkreetses tekstis saama teistsuguse analüüsi kui tavaliselt.

3. Esimene katse: lihtne morfoloogiline analüüs

Esimesel katsel analüüsisime uue meedia keelt ilma analüsaatorit kohandamata. Teostasime morfoloogilise analüüsi ilma oletamise ja ühestamiseta. Tundmatuks jäänud sõnade osakaal allkorpuste kaupa on esitatud tabelis 1.

Tabel 1. Tundmatu sõna analüüsi saanud märgijadade hulk ja osakaal allkorpuste kaupa

	Maht sõnades	**T%**
Jututoad	7 017 000	27,2
Foorumid	4 981 000	10,3
Kommentaarid	1 987 000	5,6
Uudisgrupid	6 851 000	11,7

Järgnevalt analüüsime tundmatu sõna analüüsi saanud märgijadasid ja arutleme võimalike analüüsimeetodite ning morfoloogiliste märgendite üle. Kuna kõige rohkem jäi sõnavorme tundmatuks jututubade tekstides, siis moodustab nende analüüs ka suurima osa järgnevast käsitlusest, kuid tähelepanu on pööratud ka teiste allkorpuste – foorumite, uudisgruppide ja kommentaaride – tundmatuks jäänud sõnavormidele.

Morfoloogilisel analüüsil tundmatuks jäänud märgijadade põhiosa jaguneb seitsme rühma vahel: partiklid, emotikonid, täheasenduste jms sihiliku kirjapildi muutustega sõnavormid, pärisnimed, võõrkeelsed sõnavormid ja toorlaenud, normeeritud kirjakeele seisukohalt valed (kuid kõnekeeles/murretes esinevad) sõnavormid ja lihtsalt trükivigadega sõnavormid. Järgnevalt analüüsime neid rühmi lähemalt.

[3] Programmi demoversioon on tasuta kasutatav veebiaadressil http://www.filosoft.ee/html_morf_et/ (01.03.2011).

3.1. Partiklid

Eelkõige jututubade tekstides moodustavad tundmatuks jäänud sõnade sagedusloendi tipu lühikesed, tihti lühenenud sõnavormid, mis sageli moodustavad üksi lausungi ja mis lausungi osana toimivad üldlaiendina, nt *tre*, *irw*, *ok*, *kle*, *we* jt. Tegemist on suulise kõne sõnaliikide süsteemist tuttava sõnaliigi (või sõnaliikide komplekti) – partikliga.

Tiit Hennoste (2002: 63) on suulises keelekasutuses esinevat partiklit defineerinud kui süntaktiliselt sõltumatut sõna, mis on semantiliselt sisutühi, paikneb lause propositsioonilisest sisust väljaspool ja millel on eelkõige suhtluslik ja emotiivne funktsioon.

Partiklit omaette sõnaliigina on varem eristatud suulise kõne korpuse (Hennoste jt 2002) ja eesti murrete korpuse (Lindström jt 2006) morfoloogilisel märgendamisel.

Suulises eesti keeles esinevad partiklid liigitab Hennoste (2002: 71–72) üksiesinevateks partikliteks, tekstipartikliteks ja toimetamispartikliteks. Üksiesinevad partiklid jagunevad Hennoste järgi omakorda dialoogipartikliteks, afektiivseteks partikliteks ja aktiivseteks suhtluspartikliteks (nn tähelepanupüüdjad). Tekstipartiklid ehk üksi mitteesinevad partiklid jaotab Hennoste veel edasi piiripartikliteks ja vabalt liikuvateks partikliteks.

Uue meedia keeles esinevad kõik Hennoste esitatud partiklite alaliigid, kuid tekstisõnade sõnaliikidesse jaotamisel jätsime nende seast välja vabalt liikuvad partiklid, mille alla Hennostel on liigitatud traditsioonilised lausemodaalid (nt *ikka*, *alles*, *vist*, *kindlasti*, *veel*, *muidugi*, *väga*, *täitsa*). Sõnaliigiliselt kuuluvad need adverbide või modaaladverbide hulka. Sellesse rühma kuuluvad sõnad saavad morfoloogiliselt analüsaatorilt analüüsi adverbiks ja me ei pidanud otstarbekaks seda muuta, vähemalt mitte ilma eelneva põhjalikuma uurimiseta.

Partikli sõnaliik uue meedia keeles hõlmab seega järgmisi alaliike:

- **dialoogipartiklid**, mis osutavad kuuldel olemist või sõnumi vastuvõtmist, nt *aa*, *asoo*, *jaja*, *jep*, *nunuh*, *ok*;
- **afektiivsed partiklid** – keelekasutaja reaktsioonid, väljendavad kasutaja tundeid ja meeleolusid, nt *auts*, *irw*, *icc*, *krt*, *oih*, *wau*;
- **aktiivsed suhtluspartiklid** (tähelepanupüüdjad), nt *tre*, *tsau* (ka *sau*, *saux* jms), *kle* (ka *kule* ja *kuule*), *vata* (*vaata*), *ota* (*oota*).

Tähelepanupüüdjad on Hennoste järgi üksiesinevad partiklid, kuid jututubades kasutatakse neid peamiselt koos üttega (nt *kle* + kasutajanimi) ja tihti pikema lausungi sees (kuigi partikkel pole neis süntaktiliselt lauseliige), nt (4).

(4) AL_Capone **kle** ma tulen Pärnu mul jube nälg ;D

Meie korpuse andmetel ei saa nõustuda ka Sigrid Salla (2002: 134) väitega, et jututoavestluses puudub sageli tervitus: jututoakorpuses on lisaks partiklile *tre* väga sagedased ka tervituspartiklid *tsau*, *sau*, *tsauks* jms. Meie ja Salla uurimused põhinevad erinevate jututubade tekstidel ja on võimalik, et erinevates jututubades käibivad erinevad suhtlustavad;

- **piiripartiklid** – sõnad lausungi piiri lähedal, mis osutavad, kuidas nendega algav üksus on seotud käsiloleva tekstiga, nt küsipartikkel *ve* (*we*);

- **toimetamispartiklid,** nt *ee*, *hmm*, näitavad kirjutaja mõtte takerdumist ja mõttepöördeid (5); põhiliselt kasutatakse jututubades takerduste edasiandmiseks küll ainult mõttepunkte (6).

(5) viimane mäng mus mu arvutis oli.... **hmm** see oli 3 kuud tagasi
(6) ära seleta nii palju .. ehk siis kriba aeglasemini

Tüüpiliselt on partikli funktsioonis olev sõnavorm morfoloogiliselt ühene, s.t tal on võimalik ainult üks morfoloogiline analüüs. Kuid korpuses esines ka selliseid sõnavorme, nii mitte-kirjakeelseid kui ka kirjakeelseid, mida kasutati paralleelselt nii partikli kui ka mingi muu sõnaliigi liikmena, näiteks *ütleme* (*ytleme*) või *ommik*.

(7) **ytleme** mina teen kodus lan game
(8) **ytleme** siis kenasti suure pika lause kokku
(9) **ommik** kõigile
(10) ma tavaliselt **ommik** vara läen ja öösel tulen

Sellistel nii partikli kui muu sõnaliigi esindajana kasutatavatel sõnavormidel võib olla mitu kirjavarianti, tavaliselt kirjakeelne vorm ja selle baasil tekkinud lühivorm või -vormid. Näiteks sõnavormid *vata* ja *vaata* võivad esineda nii "tähelepanupüüdjana" kui ka täistähendusliku verbivormina. Võiks oletada, et lühenenud vormi *vata* kasutatakse pigem partiklina ja kirjakeelset vormi *vaata* verbina, kuid esialgne korpuseuuring seda oletust ei kinnitanud: nii *vata* kui *vaata* võivad esineda nii partikli kui ka verbivormina (11)–(14). Selliseid kirjakeelse vormi ja tema lühenenud variantide komplekte esines tekstides veelgi: *ota* ja *oota*; *kle*, *kule* ja *kuule* jms.

(11) **vata** siin põlvkondade konfliktiks kisup
(12) ei suuda veel nii et ei **vata** seda klaviatuuri
(13) nuh **vaata** mul on siin mitu inimest
(14) ära minu otsa **vaata**

Partikleid esineb vähemal määral ka teiste uue meedia allkorpuste tekstides. Jututubade tekstides on 5,8% tekstisõnadest partiklid, foorumitekstides on see protsent 0,6, kommentaariumide tekstides 0,24 ja uudisgruppide tekstides 0,25.

3.2. Emotikonid

Emotikonid on kirjavahemärkidest kombineeritud ikoonilised märgid, mida kasutatakse internetikeeles emotsioonide väljendamiseks ja ka üneemide ja hüüundite asendajatena, nt : P , :) , :D (vt ka Salla 2002: 138). Kuna emotikonid panustavad internetikeele teksti oma tähendusnüansi ja võivad moodustada üksi lausungi, siis oleks mõistlik neid analüüsida ja märgendada nagu sõnu, luues ka nende jaoks uue sõnaliigi. Emotikonid moodustavad jututubade tekstides 3,2% tekstisõnadest, foorumites 0,16% tekstisõnadest, kommentaaride tekstides 0,28% ja uudisgruppides 0,37% tekstisõnadest.

3.3. Täheasendused jm sihilikud kirjapildi muutmised

Uue meedia keeles esinevad sümboliasendused, mille puhul ühe tähe või tähekombinatsiooni asemel kasutatakse teisi tähti või numbreid. Levinumad asendused on *hv* asemel *ff* (nt *raffas*), *ks* asemel *x* (nt *näitex*), *ü* asemel *y* (nt *kyll*), *ts* asemel *c* (nt *täica*), *ä* asemel *2* (nt *h2sti*), *õ* asemel *6* (nt *h6be*), *ö* asemel *8* (nt *t88*). Vähemal määral esineb tekstides ka *z* kasutamist *s* asemel (nt *meez*).

Probleemi lahendamiseks on hea teada, kas sellised asendused piirduvad ainult kindla väikese hulga sagedaste sõnavormidega või on produktiivsed, s.t selliste asendustega on kirjutatud suur hulk väikese sagedusega sõnavorme. Esimesel juhul võime need sõnavormid lihtsalt lisada kasutajasõnastikku, teisel juhul on vaja mingit eeltöötlust, mis tundmatu sõna analüüsi saanud märgijadas asendaks ühe sümboli teisega ja prooviks uuesti automaatselt morfoloogiliselt analüüsida.

hv asendamine *ff*-ga näib olevat kinnistunud ainult teatud sagedaste sõnavormide eripäraks: *ff*-ga kirjutatakse peamiselt sõnavorme *raffas* (mis on jututubade tekstides sage populaarse tervituse *tere raffas* tõttu), *vaffa* (*waffa*) ja *aff*, harvem ka *koff* (omastav *kofi*).

Seevastu *ts* asendamine *c*-ga ei toimu mitte ainult kindlates, sagedastes sõnavormides (nt *ei viici*, *täica*), vaid ka kogu korpuses vaid üks kord esinevates sõnavormides (nt *ülbicema*, *veremaice*). *ks* asendatakse *x*-ga samuti nii sagedastes sõnavormides *mix*, *olex*, *ex*, *plix* 'plika' kui ka korpuses 1–2 korda esinevates sõnavormides, eriti käändsõna translatiivi (*kirurgix*, *ajatäitex*) ja verbi konditsionaali muutelõpuna (*ärkax*, *tõestax*).

Täheasenduste hulka võib lugeda ka *h* ärajätmise sõna algusest (*ommik*, *uvitav*, *ullem* jpt).

Veel oleme siia rühma lugenud sõnumi emotsionaalse sisu rõhutamiseks kasutatud sihilikud tähe või silbi kordused, nt *hahahahaha*, *ehhhhh*.

Kirjeldatud täheasendused ja silbi- või tähekordused esinevad vähemal määral ka foorumite, kommentaaride ja uudisgruppide tekstides.

3.4. Pärisnimed

Kuigi morfoloogilise analüüsi sisendiks on ainult kasutajate loodud tekst, mitte kõnelejateks või autoriteks märgendatud sõnavormid, sisaldavad tekstid ikkagi palju pärisnimesid. Jututubade tekstides on tüüpiline tundmatu sõna analüüsi saanud pärisnimi väikese algustähega kirjutatud kasutajanimi, mida on kasutatud ütte funktsioonis (15). Väikese algustähega kirjutatud pärisnimesid kohtab sageli ka kommentaaride tekstides (16). Uudisgruppides, nagu ka foorumites esineb eriti palju võõrpärisnimesid (17).

(15) **krizzy** tule privva
(16) tead **sokrates** sinuga on ükskõik kellel täiesti mõtetu vaielda
(17) Ma saan aru, et minu **Sierra** kapoti alt võib leida mitu ...

Suur- ja väiketähe eristusel ei järgita jututubade tekstides normeeritud kirjakeele reegleid, selle funktsiooniks on hoopis rõhutamine ja emotsioonide väljendamine; nii kirjutatakse pärisnimed reeglina väikese algustähega. Väiketähelisi kasutaja-

nimesid ei suudaks pärisnimedeks määrata ka mitte oletaja (programm, mis annab analüüsi morfoloogiaanalüsaatorile tundmatuks jäänud sõnavormile, lähtudes sõnavormi kirjapildist ja sõnaliikide ning muutevormide sagedusest tekstides). Isegi kui väikesetäheline pärisnimi sisaldub morfoloogiaanalüsaatori sõnastikus (suurtähelisena), jääb ta ikkagi ära tundmata (18). Ka uudisgruppides jääb palju pärisnimesid ära tundmata väikese algustähe tõttu (19).

(18) ma kuulsin et **abja paluoja** kultuurimaja on kuum koht
(19) Küsimus, et kas keegi **eestis** ka müüb (vahendab)?

3.5. Võõrkeelsed sõnavormid ja toorlaenud

Foorumid ja uudisgrupid on peamiselt mingi kindla huvi- või erialaga tegelevate keelekasutajate info- ja arvamusvahetuse paikadeks. Nii kasutatakse seal palju erialakeelt ja -slängi, mis on tugevalt mõjutatud vastava ala ingliskeelsetest tekstidest ja terminoloogiast.

Foorumite ja uudisgruppide tekstides moodustavadki automaatsel morfoloogilisel analüüsil tundmatuks jäänud sõnavormidest põhiosa inglise keele sõnavormid ning toorlaenulised erialaslängi sõnad ja erialaterminid. Ingliskeelne võib olla terve lausung, tavalisemad on siiski ingliskeelsed sõnavormid või -fraasid toorlaenudena ning tsitaatsõnade või -fraasidena eestikeelses lauses (20–22).

(20) kui sa **diskilt bootida** tahad

(21) saab ka nii kui panna **SpyBotil advanced mode**- > **ignore products**- > linnuke ette **DSO Exploit** ja **exit**

(22) osta originaalid ehk MITTE autojupp , vaid **The Real Thing**

Jututubade ja kommentaaride tekstides esineb samuti võõrkeelseid, peamiselt ingliskeelseid sõnavorme, kuid nende osakaal tundmatuks jäänud sõnavormide hulgas on väiksem kui foorumites ja uudisgruppides.

Võõrkeelsed sõnavormid võivad tekstides esineda nii terve lausungi ulatuses (23), üksiksõnadena eestikeelse teksti sees (24); esineb ka lausungeid, kus poole peal keel vahetub (25). Kasutatakse, eriti jututubades, ka inglise keelest pärinevaid partikleid, nt *hello*, *fuck*, samuti on eriti jututubades kirjutatud ingliskeelset teksti häälduspäraselt (26).

(23) Ezkimo no offence but that is not true
(24) help ma nii vesinud
(25) Metaxa ma juba lootsin et sa teistsugune ja puha ,, but no !! :(
(26) luk huus tooking :D

Eesti keele morfoloogiaanalüsaator ei ole mõeldudki võõrkeelsete sõnade analüüsiks, nii et lahendus oleks võõrkeelse teksti eelnev äratundmine ja sellisena märgendamine, et morfoloogiaanalüsaator saaks selle vahele jätta.

3.6. Normeeritud kirjakeele seisukohalt valed sõnavormid: kõnekeelsused, murdevormid, slängisõnad, lühenenud sõnad ja sõnakatked, kirjakeele vormimoodustusnõuetele mittevastavad sõnavormid

Sellesse rühma võib lugeda järgmised keelendid:

- uue meedia allkeelele **eripärased (uudis)sõnad**, nt *loogish* (ka *logish*), *friik, tydo*, verbide *privama* ja *ruulima* vormid;
- iseseisvate sõnadena käibivad **sõnakatked** (vt ka Salla 2002: 133 ja Soodla 2010: 116 jj), nt *suht*, *tegelt*, *norm*, *aint*;
- ***gi/ki*-liite** kirjakeele normist erinev asetus sõna muutelõppude seas, nt *kellegil*, *kellegile*, *kellegiga*, *millegist*, *millegiga*;
- ***nud*-partitsiibi** kõnekeelne *nd*-lõpuline variant, nt *läind*, *surnd* jpt (vt ka Soodla 2010: 59–61);
- õigekeelenormidele mittevastavad **kokkukirjutised**, nt *eiole*, *midaiganes*, *niiet*, *eksole*, *minuteada* (vt ka Soodla 2010: 87–114 ja Salla 2002: 145);
- muud **õigekeelenormidele mittevastavad keelendid**, nt *midagist*, *kudagi*, *mudu*, *ikkagist*, *kussa*, *mingine*, *lissalt* 'lihtsalt', *ikki*, *prääga*, *õhtast*, *jummala*.

3.7. Trükivigadega sõnad

Uue meedia keelekasutus on huvitav just oma spontaansuse tõttu. Spontaansuse ja kiirustamisega kaasnevad rohked trükivead, mida on eriti palju jututubade kui sünkroonse keskkonna tekstides, kus kiirus on oluline. Vead on aga juhuslikud ja ebasüstemaatilised ja seetõttu on neid raske automaatselt tuvastada või parandada.

3.8. Vahekokkuvõte

Eeltoodud rühmi üldistades võib öelda, et uue meedia keelekasutus erineb normeeritud kirjakeelest nii oma leksika kui ka ortograafia poolest. Leksikaalsete omapärade hulka kuuluvad partiklid, emotikonid, allkeelespetsiifilised uudissõnad, lühendid ja toorlaenud ning nn kõnekeelsused. Leksikoni koostise seisukohalt on eripärane pärisnimede suur hulk tekstides.

Ortograafiliste eripärade hulka kuuluvad nn ortograafiamängud – ühe tähe või tähejärjendi asendamine teise või teistega ning suur- ja väiketähtede kasutamine mitte ortograafiareeglitest lähtuvalt, nagu pärisnime või lause alguse tähistamiseks, vaid emotsioonide väljendamiseks.

Vincent Ooi (2002: 96–98) kirjeldab ingliskeelsete jututubade tekstide automaatse morfoloogilise analüüsi katset ja rühmitab probleemseid sõnavorme, mis annab meile hea võimaluse võrrelda eesti ja inglise vastavat keelekasutust. Ingliskeelsete jututubade automaatse morfoloogilise analüüsi jaoks probleemsed sõnade rühmad sarnanevad siinsete jaotistes 3.1–3.7 kirjeldatud rühmadega. Nii nimetab Ooi sagedasemate probleemidena emotikone, diskursuspartikleid, väikese algustähega kirjutatud pärisnimesid, žanrispetsiifilisi lühendeid, mittestandardset ortograafiat ning tähekordusi.

Analüüsides jututubade korpuses kasutatud sõnavara ja selle erinevust kirjakeelsest, jääb mulje, et valdav osa erinevustest on teadliku keelemängu tulemus. Kasutajad justkui ütleksid endale, et "siin me kirjutame nii, nagu räägime" või "siin me kirjutame oma reeglite järgi, mitte nii, nagu koolis õpetatakse", ja kirjapildist häälduse väljalugemine on osa mängu võlust. Sellest siis vormid *näed* asemel *nääd*, *vaata* asemel *vata*, aga ka *enivei anyway* ja *jumala* asemel *jummala*. Ka täheasendused *ks* asemel *x*, *hv* asemel *ff*, *ts* asemel *c*, *ü* või *j* asemel *y* ei tundu olevat põhjustatud kiire trükkimise vajadusest, vaid pigem keelemängu lustist.

4. Kasutajasõnastik ja selle automaatne täiendamine

Nagu eespool (jaotises 2.2) öeldud, on morfoloogilist analüsaatorit *etmrf* võimalik allkeelespetsiifiliseks kohandada kasutajasõnastiku abil, mis annab analüüsi muidu tundmatuks jäävatele sõnavormidele ja mille abil saab anda üldkeelest erineva tõlgenduse allkeelespetsiifilise kasutuse ja funktsiooniga sõnavormidele.

Lisaks kasutajasõnastikule kasutasime uue meedia korpuste analüüsi hõlbustamiseks ka teksti eeltöötlust. Näiteks kasutatakse jututubade tekstides millegi rõhutamiseks sageli tähe või silbi kordamist, nt *eieieieiei*, *teeereeeee*. Ehkki kordus ise on tahtlik ja tal on kommunikatiivne funktsioon, ei ole korduvate silpide või tähtede täpne arv arvatavasti oluline. Seetõttu jätsime enne morfoloogilist analüüsi teksti ühtlustamisel korduvuse küll alles, kuid teisendasime kõik pikemad kui kolm kordust kolmeks, saades seega *eieiei* ja *teeereee*.

Sarnane probleem oli emotikonide kumuleerimine, s.t nende mõne koostisosa mitmekordistamine väljendamaks emotsiooni tugevust, nt :))))) , kusjuures korduste arv võis olla isegi suurem kui 100. Need kordused normaliseeriti samuti eeltöötluse käigus.

Oleme seisukohal, et kasutajasõnastiku loomisel tuleb arvestada tekstikorpuse ja tema sõnavara statistiliste karakteristikutega. On teada, et sõnade sagedused korpuses järgivad Zipfi seadust, mille kohaselt (lihtsustatult ja ligikaudselt) on sõna sagedus pöördvõrdelises seoses selle sagedusega sõnade arvuga. Ehk teiste sõnadega, väga väike hulk sõnu on väga sagedased ja väga paljud sõnad esinevad väga harva. Näiteks seitsme miljoni sõnalise jututubade korpuse erinevate sõnavormide arv on ümmarguselt 350 000 ja neist esineb korpuses üks kord 220 000. Ka jututubade korpuses morfoloogilise analüsaatori jaoks tundmatute sõnavormide sagedusjaotus on sarnane: erinevaid tundmatuid sõnavorme on kokku 240 000, neist üks kord esineb 160 000. Võrdluseks: eesti kirjakeele sagedussõnastiku (Kaalep, Muischnek 2002) aluseks olevas ühe miljoni sõnalises korpuses on ümmarguselt 155 000 erinevat sõnavormi, millest üks kord esineb 95 000.

Võib arvata, et internetikeele kui suhtlusvahendi sõnavara peaks järgima põhimõtet, et kui sõnavorm erineb kirjakeelsest ja on samas haruldane, siis on ta kirjakeelsest vormist mingi regulaarse teisenduse abil tuletatud. Teiselt poolt kui sõnavorm ei ole kirjakeelsest vormist tuletatav regulaarse teisenduse abil, peaks ta olema sageli kasutatav, et tema tähendus ja funktsioon kasutajatel meeles püsiks. Morfoloogilise analüsaatori kohandamine seisneb siis selles, et sagedased, ebaregulaarsed sõnavormid tuleb lisada kasutajasõnastikku käsitsi, haruldasemad ja regulaarselt kirjakeelest tuletatavad aga automaatselt.

Tundmatute sõnade sagedusloendi tipus on partiklid (vt jaotis 3.1). Nende jaoks lisati morfoloogiaanalüsaatori sõnaliikide süsteemi uus sõnaliik, partikkel märgendiga _B_. Partikkel on muutumatu sõna, tema algvorm on tema tekstis esinemise kuju, nt *jap*, *jep* ja *jup* on kolm erineva algvormiga tagasisidepartiklit. Sellest reeglist on ka erand: partiklites esineb sageli tähe- ja silbikordusi, need on algvormile viimisel eemaldatud, nt tekstis esines partikkel kujul *irwwwwww*, eeltöötluse käigus sai sellest *irwww* ja selle partikli algvorm on *irw*.

Tuleb rõhutada, et uue meedia tekstide morfoloogilisel analüüsil saavad kasutajasõnastikust analüüsi partikliks peamiselt need sõnavormid, mis ilma oletamiseta said tundmatu sõna analüüsi ja paiknesid tundmatute sõnade sagedusloendis piisavalt kõrgel kohal. Lähtuvalt analoogiast suulise keelega kontrolliti veel mõne sõnavormi kasutust ja vajadusel lisati partikli märgend (nt *kuule*, *vaata*, *oota*, *ütleme*, *ütleks*), kuid süstemaatiline jututubade (ja teiste internetiregistrite) partiklite uurimine on tegemata.

Kasutajasõnastiku abil anti morfoloogilise analüüsi märgend ka emotikonidele, mille jaoks samuti lisati morfoloogilise analüsaatori sõnaliikide süsteemi uus märgend, emotikon _E_. Kuna emotikone on lõplik hulk ja nad on alati ühesed, on lihtne neile kasutajasõnastiku abil morfoloogiline tõlgendus anda. Kasutajasõnastikus on 100 erinevat emotikoni. Selline suhteliselt suur arv on tingitud sellest, et eeltöötluse käigus jäeti alles kuni kolme sümbolikordusega emotikonid ja need lisati kasutajasõnastikku eraldi, nii on kasutajasõnastikus eraldi kirjed :) , :)) ja :))) jaoks.

Kasutajasõnastiku abil said analüüsi ka sagedasemad toorlaenud (nt adjektiiv *cool*, partiklid *ok* ja *bye*), kuid võõrkeelsete sõnade probleemi ei pidanud me õigeks kasutajasõnastiku abil lahendada, vaid tulevikus luua või leida parem keele tuvastaja, mis mitte-eestikeelsed sõnad tekstis ära tunneks ja vastavalt märgendaks.

Normeeritud kirjakeele seisukohalt valedele sõnavormidele püüti kasutajasõnastiku abil anda kirjakeelne algvorm, nt sõnavormi *mudu* algvorm on *muidu* ja vormi *kellegile* analüüsitakse nagu vormi *kellelegi*. Probleemiks on siin piiri tõmbamine ühelt poolt allkeelespetsiifiliste sõnade ja teiselt poolt nn valede sõnavormide vahele, nt kas sõnavorm *plix* on kirjakeele normile mittevastav variant sõnast *plika* (millele kasutajasõnastik peaks andma algvormiks *plika*) või selle allkeele sõna (mille algvorm on *pliks*).

Ülejäänud tundmatute sõnade rühmad – pärisnimed ja muudetud kirjapildiga sõnavormid – on küll olulised, kuid iga rühm koosneb hulgast keskmise või madala sagedusega sõnavormidest, mille kasutajasõnastikku käsitsi lisamine on töömahukas. Nii katsetasime kasutajasõnastiku automaatset täiendamist.

Kasutajasõnastiku automaatne täiendamine toimus järgmiselt. Algul leidsime sõnad, mida *etmrf* ei suutnud analüüsida, nt *kick* ja *viici*. Seejärel teisendasime need sõnad mingil (jaotises 3.3 kirjeldatud) regulaarsel moel, nt asendades kõigis sõnas *c ts*-iga (saime *kitsk* ja *viitsi*) ja lasime *etmrf*-il neid sõnu uuesti analüüsida. Kui *etmrf* sai analüüsiga hakkama, lisasime kasutajasõnastikku esialgse sõnavormi ja uue analüüsi, seega *viici viitsi+0 //_V_ 0, //*

Teine hea näide kasutajasõnastiku automaatsest täiendamisest on *nud*-partitsiibi kõnekeelse *nd*-lõpulise variandi automaatne äratundmine ja kasutajasõnastikku lisamine. Esimesel katsel tundmatuks jäänud sõnavormide hulgast eraldati *nd*-lõpulised, nt *istund* ja *around*. *nd*-lõpp teisendati *nud*-lõpuks ja tulemust (*istunud* ja *arounud*) lasti jälle uuesti analüüsida. Kui analüüs õnnestus, lisati

sõnavorm ja tema tõlgendus kasutajasõnastikku. *nd*-lõpulisi sõnavorme, millele sai kasutajasõnastiku abil anda *nud*-partitsiibi analüüsi, oli üle viiesaja.

Keerulisemaks tegi kasutajasõnastiku automaatse täiendamise asjaolu, et samas sõnavormis võib olla kasutatud mitut teisendust, nt sõnavormides *n2ind* või *viicix*.

Regulaarsed teisendused, mida rakendasime, olid korduvate tähtede asendamine (kolme asemel kaks või üks, kahe asemel üks) ja konkreetsed täheasendused (nt *2* asemel *ä*, *x* asemel *ks*), nagu on kirjeldatud eespool jaotises 3.3. Sageli võimaldas selline teisendus tunda ära ka modifitseeritud versiooni korpusespetsiifilisest sõnast, mis oli juba varem kasutajasõnastikku lisatud.

Omaette rühma tundmatuks jäävate sõnavormide hulgas moodustavad pärisnimed, mida jututubades reeglina ja teistes allkorpustes juhuslikumalt kirjutatakse väikese algustähega. Üksjagu pärisnimesid tähistab isikuid, kes vestluses osalevad ja kelle poole pöördutakse. Eriti kehtib see jututubades, kus nimeline pöördumine on sageli ainus viis oma postituse adressaati näidata: 58% pärisnimedest ehk 3,8% kõigist tekstisõnadest selles korpuses on kasutajanimed. Õnneks on jututubade korpuses osalejate nimed (õigemini pseudonüümid) olemas ja märgendatud (27), (28). Seega saab automaatselt teha osalejate nimesid sisaldava kasutajasõnastiku, kus nimedele antakse pärisnime analüüs. Juhul, kui kasutajanimi langeb kokku mõne tavalise sõnaga (27), tuleb pärisnime analüüs tavaanalüüsile lisada, et ei juhtuks nii, et sama sõna kasutus tavapärases tähenduses jääks ilma õige analüüsita. Seejuures tuleb kasutajasõnastikku panna suurtähelised nimed ka väiketähelistena, sest jututubades on kombeks suurtähti vestluses mitte kasutada.

(27) <speaker> **kaabakas** </speaker> suusatamine on mingi gei-sport

(28) <speaker> **Frode_Estil** </speaker> kaabakas milline ei ole gei sport ?

Ülaltoodud näite puhul lisanduvad kasutajasõnastikku kolm kirjet:

```
Frode_Estil Frode_Estil+0 //_H_ sg n, //
frode_estil frode_estil+0 //_H_ sg n, //
kaabakas kaabakas+0 //_S_ sg n, // kaabakas+s //_S_ sg in, //
kaabakas+0 //_H_ sg n, //
```

Automaatne kasutajasõnastiku täiendamine suurendas jututubade sõnastiku 30 000 sõnani, kuid kasutajanimed on sellest arvust välja jäetud. Ühe jututoa kasutajate nimed lisati ajutiselt kasutajasõnastikku sellesama jututoa töötlemise käigus ja neid järgmise jututoa analüüsil ei kasutatud. Põhjuseks oli see, et kasutajanimed on väga muutuv klass, mõned neist langevad kokku tavaliste sõnadega (nt *keegi*) ja seega võiksid nad kaasa tuua mitmeste analüüside mõttetu kasvu seal, kus vastavate nimedega kasutajaid ei olegi.

Automaatselt loodud kasutajasõnastiku väärtus omaette, uute tekstide analüüsiks, ei ole arvatavasti kuigi suur: selles on palju sõnavorme, mis esinesid ainult selles korpuses, mille põhjal ta tehti, ja uues korpuses olevaid sõnavorme ta ei hõlma. Pigem võib väärtuseks pidada tema loomise metoodikat: regulaarsete teisenduste kasutamist sõnavormide modifitseerimisel ja sellele järgnevat sõnastikupõhist

analüüsi, mille õnnestumise korral oletatakse, et teisenduse tulemuseks saadigi just algsele sõnavormile vastav kirjakeelne vorm. Uue korpuse analüüsil on mõtet rakendada sama metoodikat ja luua uus korpusepõhine kasutajasõnastik.

5. Teine katse: morfoloogiline analüüs eeltöötluse ja kasutajasõnastikuga

Teisel katsel analüüsisime uue meedia keelt, rakendades eelmises osas kirjeldatud kasutajasõnastikku ja eeltöötlust. Teostasime jälle morfoloogilise analüüsi ilma oletamise ja ühestamiseta. Tundmatuks jäänud sõnade hulk ja osakaal allkorpuste kaupa on esitatud tabelis 2. Tabeli veerg "1. katse T%" näitab ilma kasutajasõnastikuta morfoloogilisel analüüsil tundmatuks jäänud tekstisõnade protsenti ja tabeli veerg "2. katse T%" eeltöötluse ja kasutajasõnastikuga morfoloogilisel analüüsil tundmatuks jäänud tekstisõnade protsenti.

Tabel 2. Kasutajasõnastikuta (1. katse) ja eeltöötluse ning kasutajasõnastikuga (2. katse) morfoloogilisel analüüsil tundmatuks jäänud tekstisõnade osakaal

	Maht sõnades	**1. katse T%**	**2. katse T%**
Jututoad	7 017 000	27,2	10,5
Foorumid	4 981 000	10,3	8,8
Kommentaarid	1 987 000	5,6	4,8
Uudisgrupid	6 851 000	11,7	10,5

Paranemine on suurim jututubade korpuses, kus algne tulemus oli halvim. Tulemuse väikest paranemist foorumite ja uudisgruppide allkorpustes saab vähemalt osaliselt seletada ingliskeelsete sõnade rohkusega nendes tekstides.

Enamik nüüd veel tundmatuks jäävatest sõnavormidest on inglise, vähemal määral ka vene või mõne muu võõrkeele sõnad. Nüüd on ka jututubade tundmatute sõnade sagedusloendi tipus inglise keele sagedased sõnad: *the*, *is*, *to*, *in*, *my*, *it*, *what*, *or*, *of*, *am*, *go* jpt.

Muude tundmatuks jäävate sõnade analüüsil kooruvad välja veel mõned regulaarsed kirjapildi muutused, täpsemalt täheasendused, millega kasutajasõnastiku koostamisel ja tekstide eeltöötlusel polnud arvestatud. Vokaalidest on vastastikku asendatud *e*-d ja *ä*-d, rohkem esineb siiski *ä* kasutamist *e* asemel (*tärä* 'tere', *eksolä*, *lähän*, *polä*, *okäi* jt) kui *e* kasutamist *ä* asemel (*jergmine*, *jelle*, *verk*, *reegiks*). Vähesel määral esineb ka *ä* kasutamist *a* asemel (*ärä*) ja *y* kasutamist *ö* asemel (*yelda* jt verbi *ütlema* vormid; *yyse*l) ning *y* kasutamist *i* asemel (*yrw*).

Regulaarsetest kirjapildi muutustest on veel suhteliselt sagedased *ia* kasutamine *ea* asemel verbide *pidama* (nt *pian*, *piab*, *piaks*) ja *teadma* (nt *tian*, *tiate*) vormides ning muutumatutes sõnades *pial* ja *piale*, *vahepial*, *piaaegu* ning *sialt*.

Keerulisem on lugu tähejärjendiga *ää*, mis võib asendada nii tähejärjendit *ea* (*hääd*, *pääle*, *pääst*, *päält*, *säält*), tähejärjendit *äe* (*pääv*, *nään*, *nääb* jt verbi *nägema* vormid, *kääs* jt), tähejärjendit *ähe* verbi *minema* vormides (nt *ei lää*, *lääme*, *läävad*) kui ka tähejärjendit *äi* (*nääta*).

Sõna alguses puuduva *h*-ga sõnavormid on kasutajasõnastiku automaatse täiendamise tulemusena analüüsi saanud, nüüd jäävad tundmatuks veel mõned

verbi *minema* vormid, kus *h* on ära jäetud sõna keskelt (*läed*, *lääme*, *ei läe*, *läeks* jt).

Suhteliselt sagedase ja regulaarse teisendusena saab välja tuua *p* kasutamise *b* asemel verbi oleviku ainsuse 3. isiku vormi tunnusena (*käip*, *näep*, *saap*, *tulep*, *akkap*, *tahap*, *vaatap*, *jääp* jpt) ja *t* kasutamise *d* asemel verbi oleviku ainsuse 3. isiku vormi tunnusena (*teet*, *olet*, *näet*, *võtat* jt). Ka sõna algul esineb *g/b/d* asendamist *k/p/t*-ga (*tiivan*, *tüsgraafid*, *karaas*, *karderoob*, *kümnaasium*, *panaan*, *peib*, *pakter*), kuid need jäävad pigem nn juhuslike asenduste tasemele.

Esineb ka sõnavorme, milles kirjapildi muutused on kombineerunud ortograafiareeglitele mittevastava liitsõnamoodustusega, nt vormid *maitia*, ka *maidia*, *eitia*, *ääküll*.

6. Kokkuvõte

Ka internetis kasutatav eesti keel on eesti keel ning väärib uurimist ning automaatset töötlemist. Käesolev artikkel vaatles võimalusi internetikeele e uue meedia keele nelja tekstiklassi – jututubade, foorumite, kommentaaride ja uudisgruppide – tekstide automaatseks morfoloogiliseks analüüsiks. Uue meedia keelekasutus erineb normeeritud kirjakeelest nii oma leksika kui ka ortograafia poolest. Leksikaalse omapära alla kuuluvad partiklid, emotikonid, allkeelespetsiifilised uudissõnad, lühendid ja toorlaenud ning nn kõnekeelsused. Leksikoni koostise seisukohalt on eripärane pärisnimede suur hulk tekstides.

Ortograafiliste eripärade hulka kuuluvad nn ortograafiamängud: ühe tähe või tähejärjendi asendamine teise või teistega, sõnaalgulise *h* ärajätmine ning suur- ja väiketähtede kasutamine mitte ortograafiareeglitest lähtuvalt, nagu pärisnime või lause alguse tähistamiseks, vaid emotsioonide väljendamiseks.

Käesolev artikkel keskendus aspektidele, mis eristavad uue meedia keelekasutust normeeritud ja toimetatud kirjakeelest: ortograafiale ja sõnavormide erinevusele kirjakeelsetest. Analüüsiti erinevuste liike ja pakuti välja viis, kuidas olemasolevat tarkvara kohandada, et ta internetikeele töötlemisega hakkama saaks: luua (osaliselt automaatselt) morfoloogilisele analüsaatorile uus korpusepõhine kasutajasõnastik ning kasutada regulaarsete ortograafiliste teisenduste lahendamiseks automaattöötlust.

Kasutajasõnastikku lisati korpuses sagedased kirjakeelest erinevad sõnavormid – juba nimetatud partiklid, emotikonid, allkeelespetsiifilised uudissõnad, lühendid ja toorlaenud ning nn kõnekeelsused. Jututubade tekstides sageli esinevad väikesetähelised pärisnimed lisati iga jututoa jaoks kasutajasõnastikku automaatselt, kasutades selleks korpuses juba eelnevalt märgendatud “kõnelejate” loendit. Regulaarsed ortograafilised asendused püüti lahendada automaatsete teisenduste abil.

Automaattöötluse ja kasutajasõnastiku rakendamisel vähenes tundmatute sõnade osakaal kõige rohkem jututubade tekstides, 27,2 protsendilt 10,5 protsendile. Teistes tekstiklassides oli tulemuse paranemine väiksem (vt tabel 2). Pärast kasutajasõnastiku rakendamist moodustavad enamiku tundmatuks jäävatest sõnavormidest võõrkeele, põhiliselt inglise keele sõnad. Nende osakaal on suurem foorumite ja uudisgruppide tekstides, kus tulemuse paranemine oligi seetõttu

suhteliselt väike. Järelikult peaks üks edaspidiseid töösuundi olema parema keeletuvastaja (programm, mis suudab otsustada, millises keeles on mingi tekstilõik) loomine või mõne olemasoleva statistilise tuvastaja treenimine.

Interneti keelekasutust ei saa mingil juhul vaadelda ühtse allkeelena, lisaks on ta ajas kiiresti muutuv nähtus. Seetõttu ei ole ühe korpuse põhjal loodud sõnastiku väärtus sellisena kuigi suur: temas on palju sõnavorme, mis esinesid ainult selles korpuses, mille põhjal ta tehti, ja uues korpuses olevaid sõnavorme ta ei hõlma. Pigem võib väärtuseks pidada tema loomise metoodikat: regulaarsete teisenduste kasutamist sõnavormide modifitseerimisel ja sellele järgnevat sõnastikupõhist analüüsi, mille õnnestumise korral oletatakse, et teisenduse tulemuseks saadigi just algsele sõnavormile vastav kirjakeelne vorm. Uue korpuse analüüsil on mõtet rakendada sama metoodikat ja luua uus korpusepõhine sõnastik.

Analüüsides uue meedia keele korpuses kasutatud sõnavara ja selle erinevust kirjakeelsest, jääb mulje, et valdav osa erinevustest on teadliku keelemängu tulemus.

Lühendid ja sümbolid

<p>	kõneleja kirjutatu
<sp>	kõneleja ja kirjutatu
<speaker>	kõneleja
<stage>	kommentaar osaleja tegevuse kohta
<time>	lausungi või kommentaari lisamise aeg
H	pärisnimi
in	inessiiv
n	nominatiiv
S	substantiiv
sg	singular
V	verb

Viidatud kirjandus

EKK = Erelt, Mati; Erelt, Tiiu ja Ross, Kristiina 2007. Eesti keele käsiraamat. Kolmas, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – Renate Pajusalu, Ilona Tragel, Tiit Hennoste, Haldur Õim (Toim.). Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 4. Tartu: Tartu Ülikooli Kirjastus, 56–73.

Hennoste, Tiit; Lindström, Liina; Gerassimenko, Olga; Jansons, Airi; Rääbis, Andriela; Strandson, Krista; Toomet, Piret; Vellerind, Riina 2002. Suuline kõne ja morfoloogia-analüsaator. – Renate Pajusalu, Tiit Hennoste (Toim.). Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 3. Tartu: Tartu Ülikooli Kirjastus, 161–171.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: TÜ kirjastus.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Tiit Hennoste (Toim.). Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 1. Tartu: Tartu Ülikooli Kirjastus, 73–100.

Kerge, Krista 2004. Veebikommentaariumi mitmetahuline maailm. – Reet Kasik (Toim.). Tekstid ja taustad III. Lingvistiline tekstianalüüs. Tartu: Tartu Ülikooli Kirjastus, 51–73.
Lindström, Liina; Bakhoff, Liisi; Kalvik, Mari-Liis; Klaus, Anneliis; Läänemets, Rutt; Mets, Mari; Niit, Ellen; Pajusalu, Karl; Teras, Pire; Uiboaed, Kristel; Veismann, Ann; Velsker, Eva 2006. Sõnaliigituse küsimusi eesti murrete korpuse põhjal. – Ellen Niit (Toim.). Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised, 30. Tartu: Tartu Ülikooli Kirjastus, 154–167.
Oja, Anni 2006. Eesti keel internetis. – Mare Koit, Renate Pajusalu, Haldur Õim (Toim.). Keel ja arvuti. Tartu ülikooli üldkeeleteaduse õppetooli toimetised, 6. Tartu: Tartu Ülikooli Kirjastus, 259–267.
Oja, Anni 2010. Sissevaateid internetisuhtlusse. – Oma Keel, 1, 11–18.
Ooi, Vincent 2002. Aspects of computer-mediated communication for research in corpus linguistics. – P. Peters, P. Collins, A. Smith (Eds.). New Frontiers in Corpus Research. Amsterdam: Rodopi, 91–104.
Salla, Sigrid 2002. Jututuba kui võrgusuhtlusvorm. – R. Kasik (Toim.). Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu: Tartu Ülikooli Kirjastus, 128–156
Soodla, Karin 2010. Morfoloogilisi, morfosüntaktilisi ja sõnamoodustuslikke erijooni eesti internetikeeles. Teadusmagistritöö. Tartu Ülikool, filosoofiateaduskond, eesti keele osakond. http://hdl.handle.net/10062/15263 (10.09.2010).

Võrgumaterjalid

Eesti kirjakeele kooondkorpus. http://www.cl.ut.ee/korpused/segakorpus/ (26.02.2011).
etmrf. Morfoloogiline analüsaator. Demoversioon http://www.filosoft.ee/html_morf_et/ (01.03.2011).
Uue meedia korpus. http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/ (26.02.2011).

Kadri Muischneki (Tartu Ülikool) teaduslikeks huvialadeks on korpuslingvistika ning eesti keele (automaatne) morfosüntaktiline ja süntaktiline analüüs.
kadri.muischnek@ut.ee

Heiki-Jaan Kaalepi (Tartu Ülikool) põhilised uurimisvaldkonnad on korpuslingvistika ja eesti keele morfoloogia.
heiki-jaan.kaalep@ut.ee

Raul Sirel (Tartu Ülikool) tegeleb dialoogsüsteemidega ning eesti keele automaatse morfoloogilise analüüsiga.
rsirel@ut.ee

A CORPUS-BASED APPROACH TO THE AUTOMATIC MORPHOLOGICAL ANALYSIS OF ESTONIAN COMPUTER-MEDIATED COMMUNICATION

Kadri Muischnek, Heiki-Jaan Kaalep, Raul Sirel
University of Tartu

This article concentrates on aspects of Estonian that are different in computer-mediated communication and the standard written language: orthography and the divergence of word-forms. The authors present an analysis of these differences and propose a way to adapt an existing morphological analyser for analysing computer-mediated communication. The method entails the creation of a user lexicon for the morphological analyser, deployed largely in an automated manner, and the automatic pre-processing of texts.

While analysing the word-forms used in the texts of new media and comparing them with those of standard written language, one gets the feeling that most of the differences are the result of conscious language play. The lexical traits of Internet language include particles, emoticons, genre-specific neologisms, acronyms, borrowings from foreign languages and colloquial words. There is a great deal of play with orthography: substituting letters, omitting letters, lengthening and shortening letter sequences, and non-standard use of capitalization.

As a result of pre-processing and the user lexicon, the percentage of unrecognized tokens decreases from 27.2 to 10.5 for chatroom texts, from 10.3 to 8.8 for texts of Internet forums, from 5.6 to 4.8 for comments, and from 11.7 to 10.5 for newsgroup texts. The main source of errors while analyzing texts with the customized morphological analyzer are non-Estonian words, phrases, and sentences that the analyzer cannot handle.

Keywords: computational linguistics, corpus linguistics, morphology, morpho-syntax, wordclass, orthography, Estonian