doi:10.5128/ERYa12.15

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 12, 253–267

INTERNETIKEELE AUTOMAATNE SÜNTAKTILINE ANALÜÜS KITSENDUSTE GRAMMATIKAGA

Dage Särg

Ülevaade. Artikkel kirjeldab eesti keele kitsenduste grammatika kohandamist internetikeelele. Selleks parsiti 19 809 sõne suurune jututubade korpus eesti kirjakeele jaoks väljatöötatud reeglistikuga. Korpuse märgenduse käsitsi kontrollimisel leitud vigade põhjal tehti reeglistikku muudatusi neljas etapis: osalausepiiride tuvastamine, ühendverbide tuvastamine, pindsüntaktiline analüüs ning sõltuvussüntaktiline analüüs. Töö käigus leiti, et internetikeele süntaksi olulisemateks erijoonteks on laialdane partiklite ja ütete kasutus, väiksem täiendite osakaal, lausete lühidus ja väljajätteliste lausete sage esinemine. Reeglistiku kohandamise tulemusel paranesid nii pindkui ka sõltuvussüntaktilise analüüsi näitajad. Kõige enam vigu tekkis subjektide, predikatiivide ja adverbiaalide funktsioonide märgendamisel. Sõltuvussüntaktilisel analüüsil esines enim vigu adverbiaalide sõltuvusmärgendites.*

Võtmesõnad: arvutilingvistika, keeletöötlus, süntaks, sõltuvussüntaks, keele varieerumine, eesti keel

1. Sissejuhatus

Süntaktiline analüüs on sisendteksti süntaktilise kirjelduse kindlaksmääramine ilma semantilise interpretatsioonita (Roosmaa jt 2001: 11). Selleks võib määrata lauses olevatele sõnadele süntaktilised funktsioonid või leida lause puukujulise struktuuri: kas sõltuvus- või fraasistruktuuri. Süntaktiline analüüs on üks automaatse keeletöötluse põhietappe ning seda saab kasutada mitme edasise ülesande lahendamisel, näiteks masintõlkesüsteemides või informatsiooni eraldamisel tekstist.

Eesti kirjakeele jaoks loodi süntaktiline analüsaator juba 2000. aastal (Müürisep 2000), mida on sellest ajast saadik pidevalt edasi arendatud. Suur osa teksti, mis inimeste igapäevaelu käigus tekib – näiteks kirjalikud interneti teel peetavad

* Artikkel põhineb autori magistritööl (Särg 2015).

“vestlused”, aga ka mitmesugused märkmed, mis enda või teiste jaoks üles tähendatakse, – ei vasta kirjakeele normile. Seetõttu on vaja ka automaatseid keeletöötlusvahendeid kohandada seda sorti tekstidega toime tulema.

Käesolevas artiklis kirjeldan eesti kirjakeele süntaktiliseks analüüsiks mõeldud reeglistiku kohandamist internetikeelele. Reeglistiku kohandamisel ning tulemuste hindamisel võtsin aluseks erinevate jututubade tekstid, mille morfoloogiline analüüs ja ühestamine olid eelnevalt käsitsi üle kontrollitud.

Artikkel on üles ehitatud järgmiselt. Kõigepealt tutvustan kitsenduste grammatika formalismi ja annan ülevaate sagedasematest vigadest, mis tekivad kirjakeele reeglistiku rakendamisel internetikeelele, samuti kirjeldan internetikeele süntaksi eripärasid. Seejärel käsitlen reeglistikku tehtud muudatusi nende eripäradega toimetulemiseks. Lõpuks toon välja kohandatud reeglistikuga saavutatud tulemused.

2. Süntaktiline analüüs, kitsenduste grammatika

Lause süntaktilist struktuuri on võimalik esitada erinevalt: kas sõltuvusstruktuurina või fraasistruktuurina. Esimesel juhul leitakse sõnadevahelised sõltuvussuhted, teisel juhul fraasid (moodustajad) ja nendevahelised suhted. (Roosmaa jt 2001: 11–12) Mõlema variandiga võib kaasneda ka süntaktiliste funktsioonide kindlaksmääramine. Eesti keele puhul – nagu üldse vaba sõnajärjega keelte puhul – näib olevat sobivam sõltuvusstruktuuri kasutamine (Nivre 2005: 24, Muischnek jt 2014a: 111).

Süntaktilise märgendamise grammatikat on võimalik koostada kahel viisil: lingvistiliselt või statistiliselt. Esimesel juhul koostatakse käsitsi reeglid süntaktilise struktuuri määramiseks, teisel juhul aga kasutatakse masinõpet: analüsaatorit treenitakse etteantud näidete abil. (Roosmaa jt 2001: 14–15)

Eesti kirjakeele süntaktiline analüsaator põhineb Fred Karlssoni loodud kitsenduste grammatikal (Karlsson 1990, Karlsson jt 1995). Kitsenduste grammatika on reeglipõhine formalism tekstide süntaktiliseks parsimiseks ning seda luues seadis Karlsson eesmärgiks, et see oleks keelest sõltumatu ja suudaks toime tulla ka reaalse keelekasutusega, mitte vaid keeleteadlaste väljamõeldud lausetega. Samuti pidas ta oluliseks selgust ja lihtsust: et kitsendused (e märgendusreeglid) oleksid eraldatud programmikoodist, mis neid tekstile rakendab, ning et rakendamine oleks efektiivne ja mitte ülemäära keeruline. (Karlsson 1990: 168)

Kitsenduste loomise aluseks on ulatuslikud korpuspõhised uuringud ning kitsendused võivad olla kas absoluutsed reeglid või kajastada teatava tõenäosusega kehtivaid suundumusi. Eelistatavamad on absoluutsed reeglid ning reeglid järjestatakse nii, et absoluutseid rakendataks esmajärjekorras. Kitsenduste loomisel kasutatakse võimalikult palju ära morfoloogilist informatsiooni, aga ka leksikaalset infot. (Karlsson 1990: 168)

Kitsenduste grammatika rakendamisel lisatakse esmalt tekstis sõnadele kõik relevantsed märgendid – igale sõnale vähemalt üks –, seejärel eemaldatakse nii palju märgendeid kui võimalik ning ideaaljuhul jääb igale sõnale alles üks ja õige märgend. Seejuures ei määrata grammatika algse versiooni kohaselt ära kogu lause struktuuri, vaid tuvastatakse ainult see, kas fraasi põhisõna jääb laiendist lauses paremale või vasakule. (Karlsson 1990: 168, 171)

Sõltuvusstruktuuri määramine lisati kitsenduste grammatikasse hiljem juurde (Bick 2009). Selle puhul ei lisata sõnale kõiki võimalikke sõltuvusmärgendeid, vaid iga järgneva reegli rakendumine asendab eelnevalt määratud sõltuvusmärgendi.

Pindsüntaktiliste märgendite lisamiseks, asendamiseks, eemaldamiseks või mitme märgendi hulgast ühe väljavalimiseks võib grammatikas olla väga palju erinevat tüüpi reegleid[1]. Reeglid põhinevad lisaks morfoloogilisele ja leksikaalsele infole ka eelnevalt juba lisatud pindsüntaktilistel märgenditel: näites (1) toodud reegel lisab adverbile adverbilise eestäiendi märgendi (@DN>), juhul kui adverbile järgneb lauses substantiiv või pronoomen enne finiitset verbivormi.

Karlsson (1990: 171) mainib, et kitsendused võivad esmapilgul näidagi üsna lihtsad ja isegi triviaalsed, võrreldes sellega, mida tavaliselt n-ö lingvistiliselt olulisteks üldistusteks peetakse. See on seotud kitsenduste grammatika olemusega, mis nõuab, et kitsendused oleksid lihtsal viisil väljendatavad ning parsimise seisukohast efektiivsed.

(1) ADD (@DN>) TARGET Adv (*1 SbPr BARRIER FinV);

Sõltuvussüntaktilise info jaoks on kasutusel kahte liiki reeglid: sõnale ülemuse (ingl *governor*) määramiseks ning sõnale alluva (ingl *governee*) määramiseks. Näites (2) on toodud üks ülemuse määramise reegel, millega seatakse kaassõna ülemuseks temast lauses paremal paiknev infiniitne öeldis, mis kuulub samasse osalausesse.

(2) SETPARENT (K) TO (*1 (@IMV) BARRIER CLB);

Eesti kirjakeele jaoks mõeldud reeglistikku on Müürisep ja Uibo (2005) kohandanud transkribeeritud suulise keele pindsüntaktiliseks analüüsiks ning Lindström ja Müürisep (2009) transkribeeritud murdekeele analüüsimiseks. Sõltuvussüntaktilise analüüsiga on eesti keele puhul tegeletud vaid kirjakeelega (Muischnek jt 2014a).

3. Süntaktiline analüüs kohandamata analüsaatoriga

3.1. Andmestik ja tarkvara

Andmestikuna kasutasin käesolevas töös jututubade tekste, mis on pärit TÜ Uue Meedia korpusest,[2] sealt on võetud ka kõik toodud näitelaused. Kuigi suhtlemine jututubades ei ole enam eriti levinud, esindavad sealt kogutud tekstid siiski hästi spontaanset kirjalikku keelekasutust ning jututubade keele töötlemise kogemus peaks olema edukalt laiendatav ka teistele spontaanse kirjaliku keelekasutuse žanritele.

Kuna süntaktilise analüüsi teostamisel kitsenduste grammatikaga kasutatakse morfoloogilise analüüsi märgendeid, siis otsustasin käesolevas töös võtta aluseks jututubade tekstid, mille morfoloogiline analüüs on eelnevalt käsitsi parandatud. Nimelt leidsin oma bakalaureusetöös, et internetikeele automaatse morfoloogilise ühestamise täpsuseks on vaid 74,7% ning saagiseks 80,6% (Särg 2012: 16). Seega põhjustaksid ebakorrektsed morfoloogilised märgendid süntaktilisel analüüsil niivõrd palju vigu, et süntaksi eripäradest tulenevad vead jääksid sel juhul tõenäoliselt tagaplaanile.

[1] Constraint Grammar Manual http://beta.visl.sdu.dk/cg3/single/ (28.9.2015).
[2] http://www.cl.ut.ee/korpused/segakorpus/uusmeedia (28.9.2015).

Jututubade teksti lausestamisel on ühikuks võetud kasutaja korraga esitatud repliik, mis ei pruugi tegelikult kattuda kirjakeele lausega. Suulise keele uurimisel on Hennoste (2000: 2226) kasutanud selle tähistamiseks terminit "lausung", mida Muischnek jt (2011: 113) on kasutanud ka internetikeele puhul.

Jututoa lausung võib sisaldada kirjakeele seisukohast mitut lauset, nagu näites (3), kus kasutaja korraga edastatud lausungi võiks jagada kolmeks lauseks: esimest lauset lõpetaks küsimärk, teine oleks *m22ratud parentiks ja k6ik* ning kolmas oleks *ot ma l2hen otsin selle faili ylesse*. Samas võib üks kirjakeele lause olla jagunenud mitme lausungi vahel. Kuna aga kirjavahemärkide kasutamine jututubades on juhuslik ja vähene, siis pole võimalik ega otstarbekas kasutajate repliike kirjakeele reeglite järgi lauseteks jagama hakata.

(3) mis millise configa? m22ratud parentiks ja k6ik ot ma l2hen otsin selle faili ylesse

Süntaktilise analüüsi teostamiseks kasutati VISL-CG3 parserit[3] ning eesti kirjakeele jaoks väljatöötatud kitsenduste grammatika reegleid.

3.2. Tulemused kohandamata reeglitega

Esmalt, et saada algtaseme tulemus ning näha, missugused on suurimad probleemid internetikeele parsimisel, parsiti käsitsi kontrollitud morfoloogilise analüüsiga jututubade tekstid, kasutades kirjakeelele väljatöötatud reegleid, millele olid lisatud märgend @B partiklite märgendamiseks – seda kasutasid ka Müürisep ja Uibo (2005) suulise keele analüüsil – ning @E emotikonide määramiseks; samuti kõige elementaarsemad reeglid nende lauseliikmete sõltuvussuhete märgendamiseks.

Tulemuste hindamiseks kontrollisin käsitsi üle korpuse, mis sisaldas tekste üheksast eri jututoast ning mille suuruseks ilma kirjavahemärkideta oli 19 809 sõnet. Jututubades sai väära pindsüntaktilise analüüsi 8,01% sõnedest. Kõige enam vigu tekkis subjekti (32,07% kõigist vigadest), öeldistäite (21,55%) ning adverbiaali (10,08%) tuvastamisel. Pindsüntaktilise märgenduse täpsuseks oli 83,97% ning saagiseks 91,99%.

84,39% jututubade sõnedest said kirjakeele reeglistiku rakendamise järel ühese analüüsi. See osakaal on väiksem kui Lindströmi ja Müürisepa (2009) poolt saadud tulemus murdekeele puhul: kirjakeele analüsaatoriga jäi murdekeeles ühese analüüsi saanud sõnede osakaal 89–92% vahele. Ka kirjakeele tekstide puhul saab ühese analüüsi 88–90% sõnedest (Muischnek jt 2014a: 115). Mitmeseks jäänud sõnadele anti enamasti (1223 juhul ehk 82% juhtudest) kaks erinevat märgendit. Kõige enam mitmesusi tekkis sellest, et ei suudetud valida määruse ja järeltäiendi tõlgenduse vahel. Niisugune mitmesus jääb alles nii määrsõnade kui ka nimi-, ase- ja arvsõnade puhul. Näites (4) on toodud juhtum, kus mitmeseks jääb määrsõna *isegi*, saades märgendid @ADVL ja @<DN, kuna talle eelneb nimisõnu, millele ta võiks olla järeltäiend, samas võib ta esineda ka määrusena.

(4) tegelt winil oli subpixeli värk isegi (@ADVL @<DN) enne olemas vist, kui X'il

[3] http://beta.visl.sdu.dk/cg3.html (28.9.2015).

UAS (ingl *unlabeled attachment score*) ehk sõltuvusmärgendite saagis kirjakeele analüüsireegleid kohandamata kogu käsitsi üle kontrollitud korpusel oli 74,95%, st ligi kolmveerand sõnadest said õige ülemuse märgendi. Erinevate jututubade lõikes varieerus saagis 72,80%-st 80,48%-ni.

LAS (ingl *labeled attachment score*) ehk sõnade saagis, millel on nii õige pindsüntaktiline kui ka sõltuvusmärgend, oli korpusel 71,86%, seejuures lugesin õigeks ka need juhud, kui pindsüntaktilisi märgendeid oli antud mitu ja üks nendest oli õige.

Sõnadest, mis ei saanud õiget ülemuse märgendit ja mis loeti seega vigade hulka, jäid 30% ilma sõltuvusmärgendita ning 70% said vale märgendi. Ligi poole kõigist vigadest moodustasid partiklid ja adverbiaalid, esimesed 24,69% ning teised 24,63%. Lisaks oli muudest lauseliikmetest suurem vigade osakaal veel subjektide (13,62%) ja finiitsete öeldiste (8,26%) märgendite puhul. Kuna tegu on ka korpuses kõige enam esindatud lauseliikmetega, siis on see ootuspärane.

4. Internetikeele süntaktilised erijooned

Eesti internetikeele süntaksit on siiani uuritud äärmiselt vähe: Hennoste on küll oma töödes (Hennoste 2012; Hennoste 2013) toonud esile mõningaid selle keelevariandi süntaktilisi erijooni, kuid ühtegi põhjalikumat käsitlust ei ole. Seega, olles kirjakeele parsimisreeglite abil märgendatud korpuse käsitsi üle kontrollinud, analüüsisin edasise töö huvides, mille poolest internetikeele süntaks kirjakeele omast erineb.

4.1. Lühikesed ja väljajättelised lausungid

Üheks põhiliseks internetikeele süntaksile omaseks jooneks on lühikesed lausungid: jututubade korpuses on keskmine lausungi pikkus koos kirjavahemärkidega 4 sõna. Eesti keele sõltuvuspuude pangas seevastu on lause keskmine pikkus 14 sõna[4]. 66% internetikeele lausungitest on 1–4 sõna pikkused ning 90% kuni 9 sõna pikkused. Siiski on korpuses ka 1% lauseid, mille pikkus on 20 või enam sõna.

Suur osa internetikeele lausungitest ei vasta tüüpilisele kirjakeele lausele, mis sisaldab finiitset verbivormi ja vähemalt ühte käänd- või määrsõna (EKK). Ühesõnalised lausungid on enamasti partiklid või emotikonid. Ka kahesõnaliste lausungite üheks (või mõlemaks) elemendiks on sageli partikkel või emotikon, kombineerituna näiteks üttega. Seda sorti lausungeid luuakse palju tervitamisel: jututubadesse lisandub pidevalt uusi vestlejaid, keda tervitatakse ja kes kohalolijaid vastu tervitama hakkavad.

Teiseks põhjustab väljajättelisi ja väga lühikesi lausungeid see, et sageli sarnaneb niisugune spontaanne kirjalik suhtlus suulise dialoogiga. Hennoste (2012: 37) on MSN-i[5] vestluste kohta märkinud, et nendes vahetatakse sarnaselt suulise kõnega spontaanselt lühikesi kõnevoore. Vaatamata sellele, et jututoas viibib enamasti rohkem kui kaks kasutajat korraga, on seal näha lühikestest voorudest koosnevaid dialooge, milles lausungist puuduv info selgub ümbritsevast kontekstist.

Internetikeele lausungist võib olla välja jäetud lisaks muudele lauseliikmetele ka öeldis – seda on Hennoste (2013) nimetanud "eesti netikeele kõige radikaalsemaks

[4] https://www.keeletehnoloogia.ee/et/ekt-projektid/vahendid-teksti-mitmekihiliseks-margendamiseks-rakendatuna-koondkorpusele/soltuvussuntaktiliselt-analuusitud-korpus (28.9.2015).

[5] MSN Messenger – endine Microsofti kiirsuhtluse klientprogramm.

erijooneks". Niisugune väljajätt toimub enamasti *olema*-verbiga, kuid võib esineda ka liikumisverbide puhul (5).

(5) neitsik... ma homme õhtul vist tallinna

4.2. Kirjavahemärgistus ja emotikonid

Kirjavahemärgistust kasutatakse internetikeeles vähe ja juhuslikult: jututubade korpuses leidub üks kirjavahemärk iga 10 sõna kohta, samas kui sõltuvuspuude pangas olevates ilu-, aja- ja teaduskirjandustekstides on üks kirjavahemärk iga 5 sõna kohta. Lauselõpumärk jututubades enamasti puudub, kõige sagedamini on ta olemas küsilausete puhul, nagu on nentinud ka Hennoste (2013) MSN-i vestluste tekste uurides.

Lausesisene kirjavahemärgistus sõltub nii konkreetsest kasutajast kui ka lause pikkusest. On kasutajaid, kes panevad ka lühematesse lausetesse õigekeelsusreegleid järgides vajalikud kirjavahemärgid, kuid paljud seda ei tee. Pikematesse lausetesse pannakse enamasti vähemalt osa kirjavahemärke, pigem kaldutakse komadega eraldama pikemaid osalauseid.

Erinevusena kirjakeelest võib välja tuua selle, et näitamaks, kellele lausung suunatud on, kasutatakse koolonit ning @-märki. Koolonit on kasutatud näites (6), kus lausung on suunatud jututoa kasutajale *konn*, @-märk näites (7) enne sõna *Adi* annab teada, et repliik on mõeldud sellenimelisele kasutajale. Näidetes on toodud märkide tüüpiline kasutus: koolonit kasutatakse reeglina siis, kui adressaat asetseb lausungi alguses, @-märki siis, kui adressaat on lausungi lõpus.

Lisaks tavapärastele kirjavahemärkidele kasutatakse internetikeeles emotikone. Emotikone nimetab Hennoste (2013) "kõige tuntumaks netikeelsuseks", kirjakeeles (ega suulises kõnes) neid reeglina ei esine. Emotikon paikneb enamasti lausungi lõpus ning meenutab seetõttu lauselõpumärki. Lisaks lausungi lõpu tähistamisele omab emotikon aga veel funktsioone, võimaldades näiteks väljendada emotsiooni lausungis, kust see teisiti välja ei tuleks: kui näide (8) ei lõppeks emotikoniga :), siis ei pruugiks lausungi lugeja aru saada, et ütleja rõõmus või rahul on.

(6) konn: kui seda ei kasuta, siis pole mõtet uluda ka, et aeglane on

(7) ma tean @ Adi

(8) Bloodhound ma alles jõudsin koju ja söön ja siis magama :)

4.3. Süntaktilised funktsioonid

Internetikeele lausungile iseloomulike elementidena võib välja tuua partiklid ning ütte. Need mõlemad käituvad üldlaienditena, olemata lausungiga grammatiliselt seotud. Partiklid on internetikeele üheks ühisjooneks suulise keelega. Selliste partiklite esinemine, mis lausungi sisu otseselt ei muuda (vt näide 9 – *noh*), kinnitab, et lausungeid ei genereerita siiski aja ja tähemärkide osas niivõrd ratsionaalselt ja kokkuhoidlikult, nagu võiks arvata paljude väljajättude ja lauselõpumärkide mittekasutamise põhjal.

(9) noh nunnu sibelgas kuidas jääb

Ütte kasutamine on samuti internetikeele ja suulise keele ühiseks jooneks, kirjakeeles esineb ütet vähe. Tõenäoliselt kasutatakse ütet jututubades oluliselt rohkem kui mõnes muus internetikeele alaliigis (näiteks Skype'i vestlustes, mis on reeglina dialoogid) ning ka rohkem kui suulises kõnes, kuna jututubades ei ole võimalik teisiti märku anda, missugusele kasutajale lausung suunatud on.

Suurim erinevus internetikeele ja kirjakeele süntaktiliste funktsioonide jaotuse vahel seisneb täiendite osakaaludes. Jututubades moodustavad kõik täiendiliigid kokku 7,39% korpuse mahust, sõltuvuspuude pangas olevate tekstide põhjal on ilukirjanduses täiendeid 13,52%, ajakirjanduses 25,70% ning teaduskirjanduses 34,05% korpuse mahust. Teistest keelevariantidest rohkem esineb internetikeeles öeldistäiteid: jututubades oli neid 3,36%, ilukirjanduses 2,52%, ajakirjanduses 2,13% ning teaduskirjanduses 1,79%.

Internetikeele süntaksi eripärade kokkuvõttena võib öelda, et tegu on kirjalikult esitatud keelega, mille muudavad muust kirjakeelest erinevaks mitmed sarnasused suulise keelega: partikli ja ütte sage kasutus, lühikesed kõnevoorud ning kirjakeele tüüpilise lause tingimustele mittevastavad lausungid. Samas esineb jooni, mida suulises keeles ei ole, näiteks öeldise väljajätt lausungist.

5. Reeglite kohandamine internetikeelele

Reeglite kohandamisel kasutasin arenduskorpusena 5909 sõnast koosnevat alamkorpust, mis sisaldas tekste kolmest jututoast ning mille esialgsed täpsus ja saagis olid sarnased kõigi 9 jututoa põhjal saadutega. Reeglite kohandamiseks võrdlesin kirjakeele reeglistikuga parsitud korpust käsitsi kontrollitud versiooniga ning tegin leitud vigade põhjal reeglistikku muudatusi. Muudatuste tegemine käis nelja etapina, mida on kirjeldatud järgnevates alaosades: osalausepiiride määramine, ühendverbide tuvastamine, pindsüntaktiline analüüs ning sõltuvussüntaktiline analüüs.

5.1. Osalausepiiride määramine

Osalausepiirid on info, mida kasutavad paljud süntaktiliste funktsioonide määramise kitsendused, seetõttu on nende eelnev kindlaksmääramine vajalik. Kuna kirjakeele osalausestamisreeglid põhinevad suuresti kirjavahemärgistusel, siis oli vaja lisada juurde reegleid, mis aitaksid paika panna osalausepiiri juhul, kui jututoa kasutaja kirjavahemärke kasutanud ei ole. Selleks lisasin reeglistikku olemasolevale 67 reeglile, millest paljud on mõeldud väga spetsiifiliste juhtude jaoks ning millest jututubade arenduskorpusele rakendus tegelikult vaid 28, kümmekond täiendavat osalausepiiride määramise reeglit.

Uute reeglite lisamisega püüdsin lahendada peamiselt juhtumeid, kus korrektselt kirjavahemärgistamata osalausepiiri oli võimalik tuvastada kindlate sõnade põhjal. Niisugusteks olid eeskätt alistavad sidesõnad, aga ka muud sageli uue osalause algust tähistavad sõnad, nagu *kes*, *mis*, *seega*, *järelikult* jne, samuti verbid, kui kõrvuti oli kaks finiitses vormis põhiverbi. Näites (10) on toodud üks niisugune juhtum: sõna *et* ees oleks kirjakeeles koma, mille külge pandaks osalausepiiri märgend CLB. Näites esitatud jututoa lausungis koma ei ole ja seetõttu on vaja osalausepiir

märkida osalause esimese sõna ehk sõna *et* külge, nagu uue reegli rakendumisel tehtigi. Näites (11) lisati vastava reegli rakendumisel osalausepiir sõna *ütlege* külge, kuna talle eelnev sõna *on* on samuti finiitne põhiverb.

(10) mis rahvas teeb et (CLB) ei räägi? (CLB)
(11) palju rahvast siin on ütlege (CLB) JAA

Kuna internetikeeles on liitlauseid vähe, siis suurt paranemist täpsuses ja saagises osalausepiiride reeglite kohandamine ei andnud. Testkorpuse algseks pindsüntaktiliste märgendite täpsuseks reegleid muutmata oli 83,46%, saagiseks 92,15%. Osalausepiiride reeglite kohandamise järel tõusis testkorpuse täpsus 83,72%-le, saagiseks sai 92,17%.

5.2. Ühendverbide tuvastamine

Ühendverb on verbist ja afiksaaladverbist koosnev ühend, mis käitub lauses liitöeldisena (EKG II: 20). Nende tuvastamine on pindsüntaktilise analüüsi seisukohast vajalik: esiteks on analüüs vastasel juhul ebatäpne, kuna ei märgita ära, et ka afiksaaladverb on osa predikaadist, teiseks aga mängivad ühendverbid olulist rolli objekti tuvastamisel: perfektiivsusest sõltub sageli objekti kääne (Muischnek jt 2014b: 227–228).

Ühendverbid internetikeeles on üldjoontes samad, mis kirjakeeles, kuid kaks väikest muudatust reeglistikus siiski tegin. Ühendverbide tuvastamine põhineb nii leksikonil kui ka kombineerimisreeglitel: on koostatud loendid tegusõnadest, mis saavad ühendverbe moodustada, ning abimäärsõnadest, mis koos tegusõnaga kindlates konstruktsioonides ühendverbi moodustama peaksid (Muischnek jt 2014b). Muuhulgas on eraldi loenditena tegusõnad, mis saavad moodustada ühendverbi sõnadega *järgi* ja *järele*. Internetikeeles (ja üldse kõnekeeles) neil aga enamasti vahet ei tehta, kasutatakse näiteks nii vorme *järgi proovima* kui *järele proovima*. Seetõttu otsustasin need kaks loendit ühendada.

Teiseks lisasin abimäärsõnade loenditesse, kus esines *üles*, ka vormi *ülesse*, kuna internetikeeles (ja kõnekeeles) on see väga sageli kasutatav kuju, ehkki kirjakeeles ebakorrektne.

5.3. Pindsüntaktiline analüüs

Nagu öeldud osas 3, tekkis pindsüntaktilisel analüüsil kõige enam vigu subjektide tuvastamisel, seejuures määratakse kirjakeele reeglistikuga subjektideks ka ütted. Kuna internetikeeles esineb ütteid palju, otsustasin ütte subjektist eristada ning võtta kasutusele märgendi @VOC, nagu on tehtud näiteks saami keele puhul[6].

Kirjakeeles eristatakse üte muust tekstist komadega, internetikeeles aga ainult sellest lähtuda ei saa. Kuna enamasti kasutatakse üttena kas pärisnime või mingisse rühma kuulumist märkivaid nimisõnu, nagu *mehed, rahvas, inimesed, sõbrad* jne, siis otsustasingi üttena kaaluda vaid nimetavas käändes pärisnimesid, sealhulgas jututoa kasutajanimed, mis moodustavadki suurema osa ütetest, ning kindlaid nimisõnu.

[6] http://giellatekno.uit.no/doc/lang/common/docu-sme-syntaxtags.html (28.9.2015).

Ütte märgendamiseks lisati reeglitega esmalt võimalikele kandidaatidele ütte märgend, seejärel hakati kontrollima muid tingimusi: kas lauses on subjekt olemas, kas ja mis vormis verb järgneb sõnale, mis on saanud nii ütte- kui ka subjektimärgendi jms. Samuti asendati subjektimärgend ütte omaga internetikeelespetsiifilisel @-märgiga pöördumisel (vt näide 7). Kui nimetavas käändes pärisnimi oli lisaks ütte märgendile saanud täiendi, objekti või adverbiaali märgendi, siis need eemaldati, kuna vastav kasutus on üttega võrreldes ebatõenäoline.

Lisaks subjektile esines palju vigu predikatiivi tuvastamisel, sest selle märgendamine kirjakeeles põhineb kindlate verbide (*olema*, *näima*, *tunduma* jm) olemasolul lauses, mis internetikeeles sageli lausest välja jäetakse. Niisugustest lausetest predikatiivi tuvastamisel lähtusin eeldusest, et predikatiiv on nimetavas käändes omadussõnafraas. Predikatiiv saab küll olla ka nimisõnafraas, kuid verbita lauses on nimetavas käändes nimisõnafraas sagedamini subjekt – milleks kirjakeele reeglistiku järgi üksi esinev nimisõnafraas märgendataksegi – või üte ning sellistel juhtudel ei osutunud otstarbekaks predikatiivimärgendi lisamisega mitmesusi juurde tekitada.

Palju probleeme esines täiendite tuvastamisel, kuid kuna täiendite jaoks on kasutusel kümme eri märgendit (nimisõnaline, omadussõnaline, määrsõnaline, infinitiivtäiend ja kaassõnafraas, kõik nii ees- kui ka järeltäienditena), siis eraldi vaadatuna olid vead juhuslikku laadi. Täiendite vead puudutasid pigem konkreetseid leksikoniüksusi, näiteks sõnadega *täna*, *eile*, *homme* tuleb valida alati adverbilise eestäiendi märgend, kui järgneb mingit päevaaega tähistav sõna (näide 12), või sõna *mõni* puhul tuleks valida täiendi tõlgendus, kui järgneb samas käändes nimisõna.

(12) ma täna (@DN>) päeval peaks saama

Väikseid kohandusi tegin ka muude märgendite osas, näiteks lisasin reegli, mis annab prepositsiooni *ilma* järel olevale abessiivis nimi- või asesõnale kaassõnafraasi laiendi märgendi – see reegel ei ole internetikeelega seotud, vaid tõenäoliselt on kirjakeele reeglistikus kogemata käsitletud sõna *ilma* postpositsioonina. Samuti täiendasin mõningaid juba olemasolevaid reegleid, lisades uute märgendite – ütte ja partikliga – seotud kitsendusi.

5.4. Sõltuvussüntaktiline analüüs

Sõltuvussüntaktiliste reeglite kohandamisel pöörasin tähelepanu eeskätt internetikeele jaoks märgendussüsteemi lisatud süntaktiliste funktsioonidega sõnadele sõltuvussuhete leidmisele, mille jaoks kirjakeele reeglistikus reegleid ei olnud. Samuti oli vaja käsitleda väljajättelisi lauseid, sest nende kohta käivaid reegleid oli vähe ja seetõttu jäid sõltuvussuhted nendes lausetes sageli määramata või määrati valesti.

Kuna pindsüntaktiliste reeglite kohandamise käigus lisati üttele eraldi märgend @VOC, siis oli vaja sellele luua vastavad sõltuvuste märgendamise reeglid. Olemuselt on üte üldlaiend ehk sõna või fraas, mis kuulub lause või mingi selle osa juurde, olemata sellega grammatiliselt seotud, ja annab sellele mingi lisatähenduse (EKK: 393, 434). Seega otsustasin määrata ütte ülemuseks selle (osa)lause kõige kõrgema ülemuse, mille juurde ta kuulub: näites (13) on ütte *rallikas* ülemuseks määratud esimese osalause öeldis *tuled*.

Mõnel puhul võib küll tunduda, et üte sarnaneb järellisandiga ja võiks olla märgendatud pigem subjektile alluvaks. Ütte ja järellisandi vahe tuleb sellest, et järellisand käändub koos oma ülemusega, üte jääb nimetavasse käändesse (EKK: 434). Kuna lisand on nimisõnalise täiendi erijuhtum (EKK: 434), siis kasutatakse süntaktilisel märgendamisel täiendi märgendeid @NN> ja @<NN.

Näites (14) on toodud üks niisugune mitmeti analüüsitav juhtum: sõnale *Tegija* on antud ütte märgend @VOC ja märgitud tema ülemuseks öeldis *oled*. Kuna näites on subjekt *Sina* nimetavas käändes, siis pole võimalik öelda, kummaga tegu on: toodud märgenduse asemel võiks sõnal *Tegija* olla ka nimisõnalise järeltäiendi märgend @<NN, sel juhul oleks vaja ülemuseks määrata subjekt *Sina*. Teiste käänete puhul sellist mitmetitõlgendamise võimalust pole: kui lause oleks näiteks *Sinul Tegijal on alati midagi minu vastu* või *Sinul Tegija on alati midagi minu vastu*, siis oleks selge, et esimesel juhul on tegu järellisandiga, kuna kääne ühildub, teisel juhul üttega, kuna ei ühildu.

(13) rallikas (@VOC #1->3) äkki tuled (@FMV #3->0) ise siia ja koristad?

(14) Sina Tegija (@VOC #2->3) oled (@FMV #3->0) alati minu vastu

Kui öeldis puudub, nagu jututubades sageli juhtub, siis märgiti ütte ülemuseks (osa)lause kõige kõrgem ülemus, näite (15) puhul sõna *Inetu-tujutu* ülemuseks subjekt *nimi*. Erandiks lugesin siin juhtumid, kus lausung koosnes partiklist ja üttest (näide 16): sellistel puhkudel otsustasin määrata partikli ütte ülemuseks, mitte vastupidi, kuna lausungi sisulist tähendust – nii palju kui seda niisugustes lausungites on – annab edasi partikkel.

Üte sai lause kõrgeimaks ülemuseks vaid siis, kui see esineb üksi või koos kirjavahemärkide või emotikonidega, või seda laiendavad täiendid (nagu näites 15 üttele *Inetu-tujutu* allub järeltäiend *naiksa*, millele omakorda allub eestäiend *kallis*).

Samuti kehtivad koordineeritud ütete puhul tavapärased koordinatsioonireeglid, mille kohta on toodud näide (17): esimene koordineeritud üksus ehk *Caspar* on põhi ja allub öeldisele *tehke*, teine koordineeritud üksus *operaator-k6ps* allub esimesele ning sidesõna *ja* allub teisele koordineeritud üksusele *operaator-k6ps*.

(15) Inetu-tujutu (@VOC #1->7) kallis naiksa miks taas selline nimi (@SUBJ #7->0)?

(16) hendric (@VOC #1->2) omik (@B #2->0)

(17) tehke (@FMV #1->0) siis Caspar (@VOC #3->1) ja (@J #4->5) operaator-k6ps (@VOC #5->3)

Ka partiklite ülemuste määramise reeglid vajasid täiendamist. Esialgses katses oli partikli ülemuseks määratud alati lause kõrgeim ülemus: üldiselt predikaat, erandjuhtudel subjekt. See lähenemine on õige, kuna sarnaselt üttega on ka partikkel enamasti muu lausega grammatiliselt sidumata ega kuulu ühegi lauseliikme juurde. Sellele leidub siiski erandeid, nt näites (18) on partikkel *vot* subjekti rõhutaja, mitte ei kuulu kogu lause juurde.

Ainult partiklist või partiklist ja kirjavahemärkidest ja/või emotikonidest koosnevates lausetes määrati partikkel lause kõrgeimaks ülemuseks. Mitmest partiklist koosnevates lausetes sai ülemuseks viimane partikkel (19).

(18) vot (@B #1->2) see (@SUBJ #2->4) mind vihastabki
(19) no (@B #1->2) kle (@B #2->0)

Esialgses katses kirjakeele reeglitega oli kasutusel reegel, mis määras emotikonide ülemuseks lause kõrgeima ülemuse nagu partiklite puhul, kuid sellest otsustasin reeglite kohandamise käigus loobuda. Nimelt nii minu enda tähelepanekute kui ka varasemate käsitluste põhjal (nt Hennoste 2013) tundub, et emotikonide kasutus sarnaneb kirjavahemärkide kasutusega, kirjavahemärkidele aga sõltuvussuhteid ei määrata, seega näib õigustatud jätta ka emotikonide sõltuvussuhted muude lauseliikmetega määramata.

Vähesel määral lisasin sõltuvusreegleid ka teiste lauseliikmete sõltuvusmärgenduse parandamiseks. Predikatiivide puhul tuli lisada võimalus, et predikatiiv ise ongi lause kõige kõrgem ülemus – kas siis esinedes täiesti üksinda või pikema lausungi põhilise elemendina, nagu näites (20), kus lisaks predikatiivile *head* on veel kolm määrust, millest esimene ehk *üsna* on selgelt predikatiivi laiend ning teised kaks on üldlaiendid, mis osalause kõrgeimale ülemusele alluma peavad. Samuti oli vaja käsitleda juhte, kui predikatiiv esineb öeldiseta lauses ja tuleb märkida subjektile alluvaks (21).

Mitmel puhul (nt objekti, adverbiaali, alistava sidendi ülemuse määramine, vt näites (22) sõna *mida*) oli vaja täiendada reegleid selliste juhtumite osas, kus osalausepiir on märgitud uue osalause esimese sõna külge, kuna ülemust otsitakse üldiselt lausest sõnast nii ette- kui tahapoole vaadates kuni osalausepiirini. Niisugustel juhtudel kirjakeeles eelneks osalausepiir sõnale, kuna oleks märgitud koma külge, koma puudumise tõttu tuleb internetikeeles märkida ta aga sõna enda külge.

(20) üsna (@ADVL #1->2) head (@PRD #2->0) juba (@ADVL #3->2) tegelt (@ADVL #4->2)

(21) ma (@SUBJ #1->0) täna varane (@PRD #3->1)

(22) mhmh, laps peab saama mida (CLB @OBJ #6->8) laps tahab (@FMV #8->5) :)

6. Tulemused kohandatud reeglitega

Tulemuste hindamiseks kasutasin testkorpusena 5821 sõne suurust alamkorpust, mis sisaldas samuti tekste 3 jututoast ja mille esialgsed täpsus ja saagis olid lähedased arenduskorpuse ja kogu korpuse näitajatega. Kuna arendus- ja testkorpus sisaldasid erinevate jututubade tekste, siis omavahel need ei kattunud.

Nagu näha tabelist 1, paranesid reeglite kohandamise tulemusel nii pind- kui ka sõltuvussüntaktilise analüüsi näitajad. Seejuures tuleb arvesse võtta, et algsete näitajate arvutamisel ei eristatud ütet subjektist. Kui kohandatud reeglitega märgendatud korpuse puhul mitte lugeda vigadeks juhte, kus subjekt on märgendatud ütteks või vastupidi, siis oleks saagis ja täpsus veel pisut kõrgemad: vastavalt 94,43% ning 86,54%.

Tabel 1. Süntaktilise analüüsi näitajad (%) enne ja pärast reeglistiku kohandamist. UAS ja LAS on sõltuvussüntaktilise analüüsi näitajad (vt ptk 3.2)

Näitaja	Kohandamata	Kohandatud
Täpsus	83,97	85,16
Saagis	91,99	93,55
UAS	75,03	84,60
LAS	72,21	82,19

Ka pärast reeglite kohandamist esines kõige enam vigu subjekti ja predikatiivi tuvastamisel, kuid mõlema puhul vähenes tehtavate vigade arv tänu reeglite kohandamisele märgatavalt. Subjekti puhul oli algne saagis 83,07%, kohandatud reeglitega aga ütte puhul 91,18% ning subjekti puhul (mille hulka ütteid loetud ei ole) 85,04%. Predikatiivi saagis tõusis 49,49%-lt 70,92%-le.

Teiste suurema esindatusega märgendite veaprotsent vähenes kuni paari protsendipunkti võrra, välja arvatud adverbiaalide puhul, millel see kasvas poole protsendipunkti võrra. Põhjuseks on asjaolu, et adverbiaalide klassi käsitletakse n-ö ülejäägina: kui sõna vastab määruse tingimustele ning talle pole muid süntaktilisi funktsioone määratud, siis loetakse ta määruseks (Müürisep 2000: 66).

Mitmesuste osakaal testkorpuses jäi reeglistiku kohandamise tulemusel praktiliselt samaks: kirjakeele reeglitega sai mitmese analüüsi ehk lisaks korrektsele märgendile veel mõne(d) märgendi(d) 480 sõna, kohandatud reeglitega 484 sõna. Ka sagedasemad mitmesuse liigid jäid samaks, kuid nende osakaalud vähenesid pisut. Uusi mitmesusi lisandus ütte märgendi lisamise tõttu märgendussüsteemi, enim niisuguseid, kus sõnale antud märgenditeks olid subjekt ja üte.

Sõltuvussüntaktilise analüüsi näitajad UAS ja LAS kasvasid reeglite kohandamise tulemusel mõlemad pea 10 protsendipunkti võrra ning on seega pisut kõrgemad Muischneki jt (2014a) leitud eesti kirjakeele vastavatest näitajatest UAS 83,4% ning LAS 80,3%.

Esialgne sõltuvusmärgendite saagis oli madal peamiselt seetõttu, et paljudele sõnadele jäid ülemused määramata, kuna internetikeele lausungitena esines niisugustest lauseliikmetest koosnevaid konstruktsioone, mille võimalikku olemasolu kirjakeele sõltuvusreeglid ei käsitlenud. Seega oli peamiselt vaja lisada reegleid selliste juhtumite käsitlemiseks, samuti internetikeele jaoks märgendisüsteemi lisatud partiklite ja ütete sõltuvussuhete määramiseks. Samas vähenes reeglite kohandamise käigus ka vale ülemuse märgendi saanud sõnade hulk (arvestamata esialgseteks vigadeks sõnu, millele ülemust ei määratud) 17%-lt 15%-le.

Kõige enam aitas saagise tõusule kaasa partiklitele põhjalikumate ülemuse määramise reeglite kirjutamine: partiklite sõltuvusmärgendite saagis kasvas testkorpuses 44,71%-lt 88,15%-ni. Samas tõusid ka edasiste rakenduste jaoks olulisemate lauseliikmete sõltuvusmärgendite saagised: predikatiivide puhul näiteks 10,71 protsendipunkti võrra ehk 72,96%-lt 83,67%-le. Objektide, adverbiaalide ning öeldiste sõltuvusmärgendite saagised tõusid kõik 4–5 protsendipunkti võrra, samuti subjektide ja ütete oma, kui neid algtulemusega võrdlemise huvides koos vaadelda. Sidendite sõltuvusmärgendite saagis kasvas 2,95 protsendipunkti ning täienditel 1,44 protsendipunkti võrra. Väike langus toimus vaid omadus- ja määrsõnaliste täiendite ning kaassõna laiendite puhul, mida esineb korpuses kokku suhteliselt vähe.

Võrreldes teiste eesti keele variantidega, millele reeglistikku kohandatud on – nimelt transkribeeritud suulise keele ja murdekeelega –, on internetikeelel saadud tulemused pisut madalamad. Üks põhjus on see, et jututubade keelekasutus on väga mitmekesine: mõned kasutajad eelistavad jääda võimalikult truuks kirjakeele reeglitele, samas kui teised püüavad kirjutamisel imiteerida rääkimist. Samuti on vead sageli väga juhuslikku laadi, tulenedes näiteks trükivigadest, sellest, et kasutaja saadab oma lausungi jututuppa mitme osana, et kiire tempoga vestlusega kaasas püsida jms.

7. Kokkuvõte

Artiklis kirjeldasin katset kohandada kirjakeele jaoks mõeldud kitsenduste grammatika reeglistikku internetikeele süntaktilise analüüsi teostamiseks. Reeglistiku kohandamiseks oli vaja tutvuda vastava keelevariandi süntaksi eripäradega ning seega andsin artiklis ka lühikese ülevaate internetikeele süntaksist.

Kirjakeele reeglistiku rakendamisel jututubade korpusele tekkis enim pindsüntaktilise analüüsi vigu ja seega vajasid enim parandamist subjekti (sealhulgas ka ütte), öeldistäite ning adverbiaali tuvastamise reeglid, sõltuvussüntaktilisel analüüsil oli kõige rohkem vigu adverbiaalide, partiklite ja subjektide ülemuste määramisel. Kitsenduste grammatika reeglite kohandamise tulemusel paranesid nii pind- kui ka sõltuvussüntaktilise analüüsi tulemused. Pindsüntaktilise analüüsi puhul tõusis testkorpuse saagis 92,15%-lt 93,35%-le ning täpsus 83,46%-lt 85,16%-le. Sõltuvussüntaksi näitajad LAS ja UAS tõusid vastavalt 72,21% ja 75,03%-lt 82,19% ja 84,60%-ni. Kõige veaohtlikumad lauseliikmed olid ka reeglite kohandamise järel üldjoontes samad.

Kuna süntaktiline analüüs on üks automaatse keeletöötluse põhietappidest ning selle korrektsest väljundist on kasu paljude edasiste ülesannete lahendamisel, siis on oluline teemaga edasi tegeleda. Internetikeel on kirjakeelest selgelt eristuv keelevariant ning vägagi laialdaselt kasutusel. Arvestades internetikeele suurt mitmekesisust, oleks ehk tulevikus kasulik lisaks reeglipõhisele lähenemisele katsetada ka andmepõhist lähenemist.

Viidatud kirjandus

Bick, Eckhard 2009. A Dependency Constraint Grammar for Esperanto. – Constraint Grammar Workshop at NODALIDA 2009, Odense. NEALT Proceedings Series, Vol. 8, 8–12. http://visl.sdu.dk/~eckhard/pdf/cg-workshop2009_dep.pdf (28.3.2016).

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

EKK = Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2000. Eesti keele käsiraamat. Tallinn: Eesti Keele Sihtasutus.

Hennoste, Tiit 2000. Sissejuhatus suulisesse eesti keelde: lausung suulises kõnes I. – Akadeemia, 10, 2223–2254.

Hennoste, Tiit 2012. Enda algatatud eneseparandus eestikeelsetes MSN-i dialoogides. – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 37–54. http://dx.doi.org/10.5128/ERYa8.03

Hennoste, Tiit 2013. kuule ma eemale nüüd. – Sirp, nr 46. http://www.sirp.ee/s1-artiklid/c9-sotsiaalia/2013-12-05-17-05-50 (6.5.2015).
Karlsson, Fred 1990. Constraint Grammar as a Framework for Parsing Running Text. – Proceedings of Coling-90. Vol. 3, 168–173. http://www.aclweb.org/anthology /C90-3030.
Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha; Anttila, Arto 1995. Constraint Grammar: A Language Independent System for Parsing Unrestricted Text. Berlin–New York: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110882629
Lindström, Liina; Müürisep, Kaili 2009. Parsing Corpus of Estonian Dialects. – Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing. http://dspace.utlib.ee/dspace/bitstream/handle/10062/14288/ lindstrommuurisep2.pdf?sequence=1 (28.3.2016).
Muischnek, Kadri; Kaalep, Heiki-Jaan; Sirel, Raul 2011. Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 111–127. http://dx.doi.org/10.5128/ERYa7.07
Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina 2014. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. – Andrius Utka, Gintarė Grigonytė, Jurgita Kapočiūtė-Dzikienė, Jurgita Vaičenonienė (Eds.), Human Language Technologies – The Baltic Perspective. Frontiers in Artificial Intelligence and Applications 268. IOS Press, 111–118. http://dx.doi.org/10.3233/978-1-61499-442-8-111
Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina 2014. Ühendverbid eesti keele pindsüntaktilises analüüsis. – Eesti Rakenduslingvistika Ühingu aastaraamat, 10, 227–240. http://dx.doi.org/10.5128/ERYa10.14
Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu: Tartu Ülikooli kirjastus.
Müürisep, Kaili; Uibo, Heli 2005. Shallow Parsing of Spoken Estonian Using Constraint Grammar. Treebanking for Discourse and Speech. – Peter Juel Henrichsen, Peter Rossen Skadhauge (Eds.), Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse. Copenhagen Studies in Language 32. Samfundslitteratur, 105–118.
Nivre, Joakim 2005. Dependency Grammar and Dependency Parsing. Växjö University.
Roosmaa, Tiit; Koit, Mare; Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina; Uibo, Heli 2001. Eesti keele formaalne grammatika. Tartu: Tartu Ülikooli kirjastus.
Särg, Dage 2012. Internetikeele automaatse morfoloogilise ühestamise kvaliteedi uuring. Bakalaureusetöö. Käsikiri Tartu Ülikooli üldkeeleteaduse osakonnas.
Särg, Dage 2015. Internetikeele süntaktiline analüüs kitsenduste grammatikaga. Magistritöö. Käsikiri Tartu Ülikooli üldkeeleteaduse osakonnas. http://dspace.ut.ee/bitstream/handle/10062/47666/Sarg_2015.pdf (28.3.2016).

Dage Särg on Tartu Ülikooli keeleteaduse doktorant, kelle peamisteks uurimisvaldkondadeks on automaatne süntaktiline analüüs ning kirjakeele normile mittevastavate tekstide töötlus.
Jakobi 2, 50090 Tartu, Estonia
dage@ut.ee

SYNTACTIC ANALYSIS OF ESTONIAN NETSPEAK USING CONSTRAINT GRAMMAR

Dage Särg
University of Tartu

The paper provides an overview of an attempt to adapt the Estonian Constraint Grammar rule set for netspeak. The rule set has been developed by Kaili Müürisep and Tiina Puolakainen for shallow and dependency parsing of standard written Estonian, and it has previously been adapted for shallow parsing of spoken Estonian by Kaili Müürisep and Heli Uibo.

First, in order to adapt the rules, a chatroom corpus was parsed with the existing rule set. The corpus was manually revised and based on the errors that were found, changes were made to the rule set. The changes regarded detection of clause boundaries and particle verbs, as well as assignment of syntactic tags and dependency relations. Extensive use of discourse particles and direct addresses, short sentence length, and small percentage of attributes among the syntactic functions used in text appeared to be the most distinctive features of netspeak, as well as the large amount of elliptical sentences from which, in addition to other syntactic functions, a predicate can be left out.

As a result of adapting the rule set, the results of both shallow and dependency parsing improved. The most error-prone syntactic functions were subjects, predicatives, and adverbials. In dependency parsing, the largest number of errors was made in determining the governors of adverbials.

Keywords: computational linguistics, natural language processing, syntax, dependency parsing, language variation, Estonian