TARTU ÜLIKOOL

FILOSOOFIATEADUSKOND

EESTI JA ÜLDKEELETEADUSE INSTITUUT

Dage Särg

INTERNETIKEELE SÜNTAKTILINE ANALÜÜS KITSENDUSTE GRAMMATIKAGA

Magistritöö

Juhendaja dotsent Kadri Muischnek

TARTU 2015

Sisukord

Sisukord 2
Sissejuhatus 4
1. Internetikeel 6
2. Süntaktiline analüüs 9
2.1. Fraasi- ja sõltuvusstruktuur 9
2.2. Kitsenduste grammatika 10
2.2.1. ESTKG märgendussüsteem 11
2.2.2. Süntaktilise analüüsi tulemuste hindamine 15
2.3. Varasemaid teemakohaseid uurimusi 15
3. Pindsüntaktiline analüüs kohandamata analüsaatoriga 19
3.1. Kasutatud andmestik ja tarkvara 19
3.2. Pindsüntaktilise analüüsi esialgsed tulemused 21
3.2.1. Subjekt 25
3.2.2. Predikatiiv 26
3.2.3. Täiend 27
3.2.4. Muu 28
3.2.5. Mitmesused 29
4. Sõltuvussüntaktiline analüüs kohandamata reeglistikuga 32
4.1. Vigade jaotus 32
4.2. Vigade põhjused 34
4.2.1. Ülemuse märgendita sõnad 34
4.2.2. Vale ülemuse märgendiga sõnad 35
5. Pindsüntaksi reeglistiku kohandamine internetikeelele 38
5.1. Osalausepiiride reeglite kohandamine 39
5.2. Ühendverbide tuvastamise reeglite kohandamine 41
5.3. Pindsüntaktilise analüüsi reeglite kohandamine 42
5.3.1. Üte 42

5.3.2. Predikatiiv 44
5.3.3. Täiend 44
5.3.4. Muu 46
5.4. Tulemused kohandatud reeglitega 46
6. Sõltuvussüntaktilise analüüsi reeglite kohandamine 50
6.1. Üte 50
6.2. Partikkel 54
6.3. Emotikon 56
6.4. Predikatiiv 56
6.5. Adverbiaal 57
6.6. Muu 57
6.7. Tulemused kohandatud sõltuvussüntaksi reeglistikuga 59
7. Internetikeele süntaksi erijooned 62
7.1. Lausungi pikkus 62
7.2. Kirjavahemärgistus 63
7.3. Mittetäielik lausung 66
7.4. Internetikeelele iseloomulikud elemendid lausungis 69
7.5. Süntaktiliste funktsioonide osakaalude võrdlus kirjakeelega 71
Kokkuvõte 74
Kirjandus 76
Lisa. Internetikeelele kohandatud kitsenduste grammatika reeglistik 82
Syntactic Analysis of Estonian Web Language Using Constraint Grammar. Summary 83

Sissejuhatus

Käesolevas magistritöös käsitletakse internetikeele automaatset süntaktilist analüüsi. Valitud teema on aktuaalne, kuna internet on saanud paljudele inimestele igapäevaselt ülimalt oluliseks ja seega puututakse kokku ka seal kasutatava keelega; samas on see keel piisavalt erinev nii (toimetatud) kirjakeelest kui ka suulisest kõnest, et väärida eraldi käsitlemist. Automaatse süntaktilise analüüsi korrektsest väljundist oleks kasu teiste automaatsete keeletöötlusülesannete lahendamisel, näiteks võib tuua meelestatuse analüüsi või masintõlke.

Töö eesmärgiks on kohandada eesti kirjakeele süntaktiliseks analüüsiks mõeldud kitsenduste grammatika reeglistik internetikeelele ning anda ülevaade internetikeele süntaksi eripäradest. Reeglistiku kohandamisel ning tulemuste hindamisel võeti aluseks erinevate jututubade tekstid, mille morfoloogiline analüüs ja ühestamine olid eelnevalt käsitsi üle kontrollitud.

Töö koosneb sissejuhatusest, seitsmest peatükist ning kokkuvõttest. Esimeses peatükis iseloomustatakse internetikeelt üleüldiselt, tuuakse välja selle liigid ning põhjused, miks internetikeel kirjakeelest erineb. Teises peatükis käsitletakse süntaktilist analüüsi, kirjeldatakse käesoleva töö aluseks olevat kitsenduste grammatika formalismi ning antakse ülevaade varasematest teemakohastest uurimustest.

Kolmas kuni kuues peatükk kajastavad autori praktilist tööd kirjakeele süntaktilise analüüsi reeglistiku kohandamisel internetikeelele. Kolmandas ja neljandas peatükis esitatakse tulemused, mis saadakse kirjakeele pind- ja sõltuvussüntaksi reeglistiku rakendamisel internetikeelele. Viiendas peatükis kirjeldatakse pindsüntaktilise analüüsi reeglistiku kohandamist ning kuuendas peatükis sõltuvussüntaksi reeglistiku kohandamist internetikeelele.

Seitsmendas peatükis esitatakse kokkuvõtlikult praktilise töö käigus välja tulnud internetikeele eripärad ning tuuakse välja mõningaid võrdlusi kirjakeelega.

1. Internetikeel

Internetikeele (ka uue meedia keele, netikeele) all mõistetakse keelt, mida kasutatakse arvutisuhtluses. See keelevariant tekkis 1990-ndatel seoses arvutite ja interneti kiire levikuga ning on tänu oma eripärasusele pälvinud mitmete uurijate tähelepanu, näiteks on David Crystal (2001) pühendanud selle kirjeldamisele ja analüüsimisele terve raamatu „Language and the Internet“. Tänapäeval toimub arvuti vahendusel väga erinevat laadi suhtlus, nii suuline (näiteks Skype'i vahendusel) kui kirjalik, nii isiklik kui ametlik. On selge, et keel, mida kasutatakse ametlikes e-kirjades, on võrdlemisi erinev sellest, mida kasutatakse sõpradega suheldes, vaatamata sellele, et mõlema puhul võib vahendajaks olla arvuti. Tiit Hennoste (2013) on eristanud n-ö päris internetikeelt kogu arvutisuhtluses kasutusel olevast keelest suulise keele omaduste ning internetikeele eripärade olemasolu alusel kirjalikult esitatud keeles.

Suulise keelega sarnaseid jooni esineb rohkem sünkroonse arvutisuhtluse keeles ehk näiteks jututubades, Skype'i suhtluses jm, mis toimub reaalajas ja kus suhtlemise kohta kasutatakse enamasti sõna „rääkima“, mitte „kirjutama“. Asünkroonne arvutisuhtlus, näiteks ajalehekommentaarid, foorumipostitused jms, sarnaneb üldiselt rohkem traditsioonilise kirjaliku keelega, kuid sisaldab sellegipoolest sageli just internetikeelele omaseid jooni. (Ooi 2002: 93)

Millised need internetikeelele eripärased jooned on? Esimesena hakkavad silma tõenäoliselt emotikonid, muus kirjakeeles mitte kasutusel olevad lühendid (nt *lol – laughing out loud*), kirjakeele reeglitele mittevastav kirjavahemärkide ja suurtähtede kasutus, tähtede kordamine (nt *paluuuun*) ja asendamine (nt *tydrik*) sõnades jne. Samuti on kasutusel palju kõnekeelseid sõnu ja väljendeid.

Ent mis on põhjuseks, miks internetikeel kirjakeelest erineb? Jacob Eisenstein on püüdnud sellele küsimusele vastata Twitteri tekstide kohta. Ta toob välja, et üheks võib-olla seni alahinnatud põhjuseks on erinevad teksti sisestamise võimalused

arvutisuhtluses – näiteks ennustava tekstisisestuse kasutamine mobiiltelefonides ja sisestamine puutetundlikult ekraanilt. Samuti on Eisensteini järgi olulised sotsiaalsed tunnused – teksti kirjutaja vanus, sugu, päritolu jne. Seega võib internetikeele korpus erineda näiteks ajakirjanduskeele korpusest juba seetõttu, et ajakirjanikena töötavad enamasti üsna kindlate sotsiaalsete tunnustega inimesed, samas kui internetikasutajate profiil on ilmselt laiem. (2013: 361-362)

Samas lükkab Eisenstein ümber oletuse, et vähemasti Twitteri tekstide puhul võiks erinevust kirjakeelest põhjustada säutsudele kehtestatud 140-tähemärgiline pikkusepiirang – miljoni ingliskeelse säutsu põhjal leiab ta, et ebastandardseid lühenenud vorme kasutatakse pigem lühemate, alla 90 tähemärgi pikkuste säutsude puhul, ja kuna lühenenud vormi kasutamine säästab vaid paar tähemärki, siis järelikult ei ole selle kasutamise põhjuseks pikkusepiirang. (2013: 360-361)

Michelle Drouin ja Claire Davies (2009) on lükanud SMS-sõnumite puhul ümber oletuse, et seal kasutatava keele erinevust kirjakeelest võiks põhjustada inimeste kirjaoskamatus – nende uuringu kohaselt ei erinenud SMS-sõnumites ebastandardseid vorme kasutavate inimeste kirjaoskus tegelikult nende omast, kes ebastandardseid vorme ei kasuta, samas olid huvitaval kombel mõlemast grupist rohkem kui pooled katsealused arvamusel, et ebastandardsete vormide kasutamine mõjub halvasti nende võimele luua korrektset kirjakeelset teksti (2009: 46).

Erinevalt kirjakeelest kasutatakse internetikeeles emotikone, põhjusena toob Vincent B Y Ooi (2009: 105) intonatsiooni ja kehakeele kasutamise võimaluse puudumise internetikeeles. Eli Dresner ja Susan C. Herring (2010) on jõudnud oma emotikone käsitlevas artiklis järeldusele, et emotikonidel on kolm funktsiooni: nad võivad kas näidata emotsiooni väljendavat näoilmet (nt rõõmus, kurb nägu), näoilmet, mis ei väljenda otseselt emotsiooni (nt silmapilgutus näitamaks, et tehakse nalja), või olla illokutiivse jõu indikaatoriks (nt kaebusega kaasnev naeratus, mis leevendab selle mõju). Eisenstein mainib, et emotikonidega sarnaseid funktsioone täidavad sageli ka fraaside lühendid nagu *smh – 'shake my head'* jne, kuid erinevalt emotikonidest saavad need käituda lauses ka moodustajatena (2013: 361).

Jennifer Foster *et al.* (2011a) toovad välja, et internetikeele töötlemise muudab

keeruliseks see, et internet ei ole tegelikult nii-öelda üks žanr. Nii sünkroonsel kui asünkroonsel arvutisuhtlusel on hulk erinevaid alamliike ja ka ühe ja sama liigi sees võib erinevate kasutajate keelekasutus olla väga erinev – näiteks postitavad mõned kasutajad Twitteris täiesti kirjakeelseid säutse, samas kui teiste puhul on säutsud kirjakeelest üsna kaugel.

Eisenstein (2013: 366) rõhutab ka seda, et internetikeel ei ole püsiv nähtus, vaid muutub pidevalt, seega staatilised süsteemid võivad küll nende rajamise hetkel olla efektiivsed, aga aja möödudes muutuvad nad järjest kasutumaks, kuna ei vasta enam parasjagu kasutusel olevale keelele.

Internetikeele süntaksi erinevust kirjakeelest illustreerib see, et inglise keele puhul saavutavad praegused parserid ajalehetekstide parsimisel suurema kui 90%-lise tulemuse, rakendatuna samal kujul internetikeelele aga vaid 75-80%-lise tulemuse (Petrov, McDonald 2012). Prantsuse keele puhul on saadud kirjakeele parseri rakendamisel internetikeelele sõltuvalt internetikeele konkreetse valdkonna teksti mürarikkusest tulemusi 39,11-71%, mis on võrrelduna kirjakeele puhul saavutatava 86-89%-ga tunduvalt madalam. (Seddah *et al.* 2012: 2443)

Eesti internetikeelest on üldisi ülevaateid kirjutanud näiteks Sigrid Salla (2002) Anni Oja (2006) ja Tiit Hennoste (2013), viimane on vestlusanalüüsi põhimõtetest lähtudes uurinud ka eneseparandusi internetikeeles (2012). Internetikeele automaattöötlust – nimelt morfoloogilist analüüsi – on käsitlenud Heiki-Jaan Kaalep ja Kadri Muischnek (2011) ning Muischnek jt (2011), internetikeele morfoloogilisest analüüsist tegi oma bakalaureusetöö ka käesoleva töö autor (Särg 2012). Eesti internetikeele süntaksi kohta on Tiit Hennoste (2011) pidanud küll ühe ettekande, kuid autorile teadaolevalt ühtegi käsitlust ilmunud ei ole.

2. Süntaktiline analüüs

Süntaktiline analüüs on sisendteksti süntaktilise kirjelduse kindlaksmääramine ilma semantilise interpretatsioonita (Roosmaa jt 2001: 11). Selleks võib määrata lauses olevatele sõnadele nende süntaktilised funktsioonid või leida lause struktuuri. Süntaktiline analüüs on vajalik samm mitmete keelega seotud ülesannete lahendamisel: Kaili Müürisep on näiteks oma doktoriväitekirjas (2000: 8), hakates eesti keele jaoks süntaktilise analüüsi reeglistikku looma, välja toonud nimisõnafraaside tuvastamise, tõlkijate abivahendite, masintõlke, grammatika- ja stiilikontrollijate loomise, automaatsete sisukokkuvõtete tegemise jms. Kõik nimetatud ülesanded on jätkuvalt aktuaalsed ka praegu ning enamal või vähemal määral ka internetikeele puhul.

2.1. Fraasi- ja sõltuvusstruktuur

Lause struktuuri on võimalik esitada erinevalt: kas sõltuvusstruktuuri 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.

Nii fraasi- kui sõltuvusstruktuuri jaoks on võimalik kasutada erinevaid grammatikaid. Fraasistruktuuri jaoks on tuntumad HPSG (*Head-Driven Phrase Structure Grammar* e peajuhitav fraasistruktuurigrammatika), TAG (*Tree Adjoining Grammar*, puid ühendav grammatika), kategoriaalne grammatika (*Categorial Grammar*) ning ka LFG (*Lexical Functional Grammar*, leksikaalne funktsionaalne grammatika). Sõltuvusstruktuuri grammatikatest on üks tuntuim Fred Karlssoni loodud kitsenduste grammatika (*Constraint Grammar*). Kuna kitsenduste grammatikat on kasutatud ka käesolevas töös, siis on seda lähemalt kirjeldatud järgmises alapeatükis.

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. Samuti on võimalik kahe meetodi kombineerimine. (Roosmaa jt 2001: 14–15)

2.2. Kitsenduste grammatika

Kitsenduste grammatika on formalism tekstide parsimiseks, mille töötas 1990-ndatel välja Fred Karlsson (vt Karlsson 1990, Karlsson *et al*. 1995). Grammatikat luues oli Karlssoni eesmärgiks, et see oleks keelest sõltumatu ning 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 muidugi absoluutsed reeglid ning reeglid järjestatakse nii, et neid rakendataks esmajärjekorras. Kitsenduste loomisel kasutatakse võimalikult palju ära morfoloogilist informatsiooni, samuti 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 jääks igale sõnale alles üks ja õige märgend. Seejuures ei määrata grammatika algse versiooni kohaselt ära kogu lause struktuuri, vaid tuvastatakse vaid, 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.

Kitsenduste grammatikat on rakendatud edukalt mitmetele keeltele, sealhulgas lõi Kaili Müürisep oma doktoritöö raames vastava grammatika eesti keele jaoks (Müürisep 2000), nimetades seda ESTKG-ks. Sellest ajast on grammatikat pidevalt edasi arendatud, sealhulgas näiteks on lisatud ühendverbide märgendamisreeglid (Muischnek jt 2013). ESTKG-d on rakendatud ja kohandatud lisaks kirjakeelele ka transkribeeritud suulisele keelele (Müürisep, Uibo 2005; Müürisep, Nigol 2007) ning murdekeelele (Lindström, Müürisep 2009).

2.2.1. ESTKG märgendussüsteem

Eesti keele kitsenduste grammatika parserina kasutatakse VISL-CG3 parserit (VISL-CG3) ning märgendussüsteemi, mille väljundist on toodud järgnev näide 1 (kõik töös kasutatud näitelaused pärinevad Uue Meedia korpuse jututubade alaosast, vt täpsemalt peatükk 3.1 ja internetiviide Jututubade korpus): lause algust ja lõppu tähistavad *"<s>"* ja *"</s>"*, tekstisõna on toodud jutumärkides ja noolsulgudes. Järgneval taandega algaval real on sõna algvorm ja morfoloogiline analüüs (morfoloogiliste märgendite tähendusi vt viidetest Morfosüntaktilised kategooriad), noolsulgude vahel paikneb süntaksianalüsaatori abiinfo, näiteks *<Tr>* sõna *nimetatakse* juures tähendab, et selle verbi laiend peab olema translatiivis. @-märgi järel on pindsüntaktiline märgend, mida võib olla sõnale alles jäänud ka mitu, nagu näites sõnade *asja*, *mida* ja *DNS* puhul. Märgend *CLB* enne pindsüntaksi märgendit tähistab osalausepiiri.

Pindsüntaktiliste märgendite tähendused on esitatud tabelis 1, täpsemat infot märgendussüsteemi kohta leiab näiteks eesti keele kitsenduste grammatika süntaksianalüsaatori märgendite lehelt (vt viidetest Märgendusjuhend).

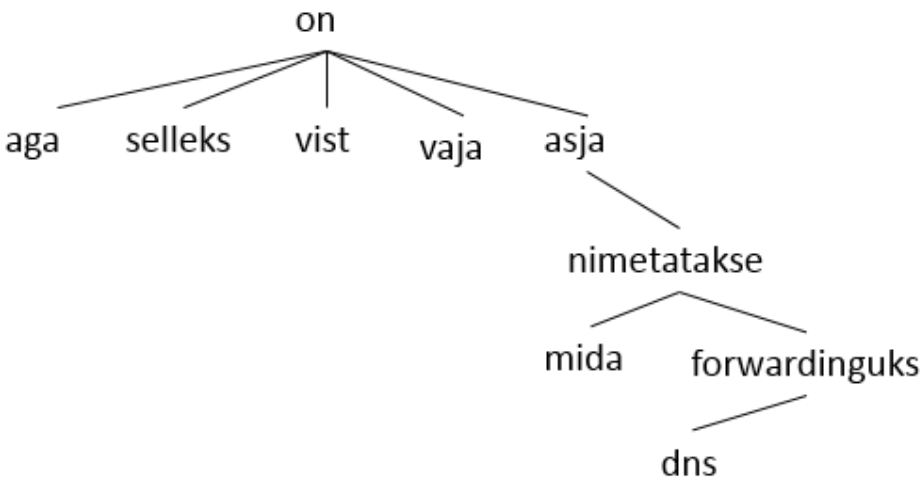
Nn trellidele (#) järgnevad numbrid tähistavad sõltuvussüntaktilist infot: esimene number on sõna järjekorranumber lauses, teine on sõna ülemuse (*head*) järjekorranumber ehk näites 1 esimese sõna *aga* ülemuseks kolmas sõna *on* – nende põhjal on võimalik esitada lause struktuur sõltuvuspuuna nagu joonisel 1. Sõltuvussüntaktilise märgendamise kohta leiab rohkem infot näiteks sõltuvuspuude

panga juurest (vt viidetest EKT).

(1)

```
"<s>"

"<aga>"
        "aga" L0 J crd @J #1->3
"<selleks>"
        "see" Lks P dem sg tr @ADVL #2->3
"<on>"
        "ole" L0 V main indic pres ps3 sg ps af <FinV> <Intr> <PhVerb> <0> <PhVerb> @FMV #3->0
"<vist>"
        "vist" L0 D @ADVL #4->3
"<vaja>"
        "vaja" L0 D @Vpart #5->3
"<asja>"
        "asi" L0 S com sg part @NN> @SUBJ #6->3
"<mida>"
        "mis" Lda P inter rel sg part CLB @OBJ @ADVL #7->8
"<nimetatakse>"
        "nimeta" Ltakse V main indic pres imps af <FinV> <NGP-P> <Tr> @FMV #8->6
"<DNS>"
        "DNS" L0 Y nominal <ccap> @NN> @OBJ @ADVL #9->10
"<forwardinguks>"
        "forwarding" Lks S com sg tr @ADVL #10->8
"</s>"
```



Joonis 1. **Sõltuvuspuu.** Üleval paikneb lause kõige kõrge ülemus *on,* talle allub viis sõna ning nendest viimasele, sõnale *asja*, allub omakorda kõrvallause öeldis *nimetatakse*.

Tabel 1. **Pindsüntaktilised märgendid ja nende tähendused.** < ja > mõnede märgendite juures tähistavad selle märgendi saanud sõna ülemuse asukoha suunda lauses

Märgend	**Tähendus**
@FMV	Finiitne öeldis
@FCV	*olema* liitaegades ning modaalverbid ahelverbides, finiitne vorm
@IMV	Infiniitne öeldis
@ICV	*olema* liitaegades ning modaalverbid ahelverbides, infiniitne vorm
@NEG	Verbi eitus
@SUBJ	Subjekt
@OBJ	Objekt
@PRD	Predikatiiv
@ADVL	Adverbiaal
@NN>, @<NN	Nimi-, ase- või arvsõnaline täiend
@AN>, @<AN	Omadussõnaline täiend
@DN>, @<DN	Määrsõnaline täiend
@KN>, @<KN	Kaassõnafraas täiendina
@INFN>, @<INFN	Infinitiivne täiend
@P>, @<P	Kaassõnafraasi laiend
@Q>, @<Q	Kvantorifraasi laiend
@J	Sidend
@??	Tundmatu, nt võõrkeelne sõna

Grammatikas võib olla väga palju erinevat tüüpi reegleid pindsüntaktiliste märgendite lisamiseks, asendamiseks, eemaldamiseks, mitme märgendi hulgast ühe väljavalimiseks jne (vt CG Manual). Reeglid põhinevad morfoloogilisel ja leksikaalsel infol ning eelnevalt lisatud pindsüntaktilistel märgenditel. Näiteks näites 2 toodud reegel lisab adverbile adverbilise eestäiendi märgendi (@DN>), juhul kui adverbile järgneb lauses

substantiiv või pronoomen enne finiitset verbivormi, reegel näites 3 aga eemaldab sama märgendi juhul, kui adverbile vahetult järgnev sõna pole substantiiv ega pronoomen või adverb ei kuulu eelnevalt defineeritud sobivatesse hulkadesse („Täna“, „KohamäärIlAl“ jne). Karlsson (1990: 171) mainib, et kitsendused võivad esmapilgul näidagi üsna lihtsad ja isegi triviaalsed, võrreldes sellega, mida üldiselt n-ö lingvistiliselt olulisteks üldistusteks peetakse, ent see on seotud kitsenduste grammatika olemusega: et kitsendused oleks lihtsal viisil väljendatavad ning parsimise seisukohast efektiivsed.

(2)

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

(3)

```
REMOVE (@DN>) (NOT 1 SbPr)
        (NOT 0 Täna OR KohamäärIlAl OR KohamäärInAd OR KohamäärElAb OR Ainult);
```

Reeglid võivad olla kas kindlad (nt kui lause ainuke verb on ainsuse kolmandas pöördes, siis *mina* ei saa alus olla) või heuristilised (nt kui lauses on vaid üks nimetavas käändes nimisõna ja verb on isikulises tegumoes, siis on nimisõna lauses aluseks).

Pindsüntaktiliste märgenditega küll kogu lause struktuuri kindlaks ei määrata, kuid näiteks täiendite märgendites on siiski kajastatud see, kummas suunas paikneb peasõna: näites 1 on sõnal *DNS* märgendiks muuhulgas *@NN>*, mis tähendab, et see oleks nimisõnaline eestäiend.

Sõltuvussüntaktilise info jaoks on kasutusel kahte liiki reeglid: sõnale ülemuse määramiseks ning sõnale alluva määramiseks. Näites 4 on toodud üks ülemuse määramise reegel, millega seatakse kaassõna ülemuseks temast lauses paremal paiknev infiniitne öeldis, mis kuulub samasse osalausesse.

(4)

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

2.2.2. Süntaktilise analüüsi tulemuste hindamine

Süntaktilise analüüsi headuse hindamiseks saab kasutada täpsust (leitud korrektsete märgendite arv jagatuna kõigi leitud märgendite arvuga) ja saagist (leitud korrektsete märgendite arv jagatuna kõigi õigete märgendite arvuga). Samuti on kasutusel saagist ja täpsust kombineeriv F1-skoor, mis on saagise ja täpsuse harmooniline keskmine ehk $\mathrm{F1}=2*\left(\frac{\text{saagis}*\text{täpsus}}{\text{saagis}+\text{täpsus}}\right)$.

Sõltuvussüntaksi puhul saab kasutada mõõdikuid LAS (*labeled attachment score*), UAS (*unlabeled attachment score*) ja LA (*label accuracy*). LAS märgib nende sõnede osakaalu, millele süsteem määrab nii õige ülemuse (*head*) kui ka funktsiooni (*dependency label*), LA arvetab ainult seda, kas on määratud õige funktsioon, UAS aga ainult seda, kas on määratud õige ülemus (Buchholz, Marsi 2006: 150).

2.3. Varasemaid teemakohaseid uurimusi

Internetikeelest ja selle käsitlustest anti ülevaade juba töö esimeses peatükis, seega keskendutakse selles osas spetsiifilisemalt süntaksile. Nagu esimeses peatükis juba mainiti, on eesti internetikeele süntaksit on siiani uuritud äärmiselt vähe. Tiit Hennoste on küll oma töödes (Hennoste 2012; Hennoste 2013) toonud esile mõningaid selle keelevariandi süntaksi erijooni – näiteks mainib ta öeldise ärajätmist ja lausete lühidust – , kuid ühtegi põhjalikumat käsitlust ei ole.

Inglise keele puhul on teemat rohkem uuritud. Reeglipõhiselt ei ole autorile teadaolevalt püütud süntaktilist analüüsi teha, kuid statistikapõhiste meetoditega on katsetanud mitmed, sealhulgas korraldati 2012. aastal *workshop* SANCL 2012 (*Syntactic Analysis of Non-Canonical Language*), kus osalenud rühmade ülesandeks oli luua internetikeele parsimiseks sobiv masinõppel põhinev süsteem (SANCL 2012). Osales 11 rühma, kes kokku esitasid 20 parsimissüsteemi, neist 8 moodustajastruktuuri ja 12 sõltuvusstruktuuri parserit. Kõige edukamad olid kombineeritud lähenemised, mille tulemused jäid vahemikku 80–84%. (Petrov, McDonald 2012). Kokkuvõttes olid

parimad Joseph Le Roux *et al.* esitatud süsteemid DCU-Paris13-1 (moodustajastruktuur, F-skoor 83,52%) ja DCU-Paris13Dep (sõltuvusstruktuur, LAS 83,46%, UAS 87,62%). Mõlema puhul kasutati andmete normaliseerimist, näiteks asendati emotikonid sõltuvalt asukohast lauses punkti või komaga, tavalisemad lühendid ja kirjakeele normidele mittevastavalt kirjutatud sõnad asendati standardsetega (nt *plz* – *please*) jne. Samuti kasutati mõlema puhul parseri täpsuse prognoosi (*parser accuracy prediction*) – selleks parsiti märgendamata andmed Wall Street Journali tekstidel treenitud parseriga ja seejärel, et saada suuremat hulka hästiparsitud internetikeele puid oma treeningandmete hulka, oletati prognoosimistööriistaga, millised laused on hästi parsitud. Prognoosimiseks kasutati erinevaid tunnuseid, näiteks lause pikkust, puu sügavust, sõnade hulka, mida treeningkorpuses ehk Wall Street Journalis ei esinenud jne. Lisaks kasutati klassifitseerijat, mis otsustab teksti põhjal, millisesse etteantud viiest internetikeele valdkonnast tekst kuulub. (Le Roux *et al.* 2012)

Wall Street Journali peal treenitud parseriga on proovinud internetikeelt – täpsemalt Twitteri säutse – parsida Foster *et al.* (2011a). Nad leidsid, et võrreldes kirjakeelega tuleb internetikeele puhul LAS lausa 20 protsendipunkti madalam, kuid sealjuures on suures hulgas parsimisvigades süüdi morfoloogilise analüüsi vead. Parseri kohandamiseks kasutasid nad protseduuri *uptraining.* (Foster *et al.* 2011a) See tähendab, et vaatluse all olevat sõltuvusstruktuuri parserit treenitakse täpsema, kuid aeglasema moodustajastruktuuri parseri väljundiga. *Uptraining* on võimalik tänu sellele, et sõltuvusstruktuuri parserid on valdkonna muutusele tundlikumad kui moodustajastruktuuri omad ja seetõttu langevad võõra valdkonna puhul nende näitajad rohkem. (Petrov *et al.* 2010: 705) Valdkonnapõhise *uptraining'*u abil saavutasid Foster *et al.* LASi tõusu 67,33%-lt 71,94%-le (2011a).

Foster *et al.* (2011b) on kasutanud parseri internetikeelele kohandamiseks ka teist masinõppe meetodit, *self-training'*ut, mis tähendab, et parserit treenitakse selle sama parseri väljundi peal. PCFG-LA *self-training'*uga saavutasid nad foorumitekstide puhul F-skoori kasvu 5,5 protsendipunkti, Twitteri säutsude puhul oli kasv väiksem, ligi 2 protsendipunkti. (Foster *et al*. 2011b)

Nii Foster (2010) kui ka Foster *et al.* (2011a) ja Foster *et al.* (2011b) toovad konkreetse

probleemina välja koordinatsiooni: kõigis internetikeele parsimise katsetustes kipuvad pikemate koordineeritud üksuste puhul tekkima vead, mille vastu ei aita ka parseri kohandamine.

Khan *et al.* (2013) on uurinud, kas paremini toimib sõltuvussüntaksi parser, mis on treenitud suurema andmehulga ehk nii Wall Street Journali kui English Web Treebanki peal, või väiksema, aga rohkem sihttekstiga ühtiva andmehulga ehk ainult English Web Treebanki peal. Nad leidsid, et esimene variant annab paremaid tulemusi, veel paremaid tulemusi annab aga see, kui Wall Street Journalist on võetud sama suur hulk teksti kui sisaldub English Web Treebankis: sel juhul saavutati LAS 83,67% ja UAS 86,41%. (Khan *et al.* 2013)

Slav Petrov ja Ryan McDonald (2012) loetlevad internetikeele süntaktilise analüüsi takistajatena ebajärjekindlat kirjavahemärgistust ja suurtähtede kasutust, slängi, tehnikažargooni, täheortograafiavigu ja ebagrammatilisi lauseid. Samuti mainivad Petrov ja McDonald (2012), et mõned süntaktilised konstruktsioonid on internetikeeles tavalisemad kui näiteks ajakirjanduses – küsimused, käsud, pikad nimede loetelud ja lausefragmendid. Kõik need erinevused on põhjuseks, miks kirjakeele jaoks välja töötatud keele automaattöötluse vahendid internetikeele puhul tunduvalt kehvemaid tulemusi annavad.

Eesti keele puhul on kirjakeele automaatse pindsüntaktilise analüüsi jaoks välja töötatud kitsenduste grammatika reeglistik (Müürisep 2000), millest oli juttu juba käesoleva peatüki eelmistes alaosades. 2013. aastal saadi selle reeglistikuga (mida on vahepeal pidevalt edasi arendatud) pindsüntaktilise analüüsi saagiseks 92,6% ja täpsuseks 72% (Muischnek jt 2013: 339).

Kaili Müürisep ja Heli Uibo (2005) on kohandanud seda reeglistikku transkribeeritud suulise keele pindsüntaktiliseks analüüsiks. Peamiselt oli neil vaja selleks kirjutada uued osalausepiiride määramise reeglid, kuna kirjakeeles põhineb osalausepiiride tuvastamine suuresti kirjavahemärgistusel, ent suulise keele transkribeerimisel ei kasutata kirjavahemärke nii nagu kirjakeeles. Samuti muutsid nad mõningaid pindsüntaktiliste märgendite määramise reegleid, mis kasutasid kirjavahemärke. Lisaks kirjakeeles olemasolevatele märgenditele kasutati partiklite tähistamiseks märgendit @B. Reeglite

kohandamise järel saavutasid nad pindsüntaktilise analüüsi saagiseks 97,3% ning täpsuseks 89,2%, kasutades tekste, mille morfoloogiline analüüs oli käsitsi parandatud. (Müürisep, Uibo 2005)

Transkribeeritud suulisele keelele kohandatud reeglistiku kohandasid Liina Lindström ja Kaili Müürisep (2009) omakorda transkribeeritud murdekeele analüüsimiseks. Kuna suulise keele transkribeerimisel oli kirjavahemärkide abil transkribeeritud intonatsioon, aga murdekeele puhul mitte, siis oli taas vaja muuta peamiselt osalausepiiride määramise reegleid, samuti mängisid rolli erinevused sõnavaras. Reeglite kohandamise tulemusel said nad saagiseks 96–97% ning täpsuseks 87–89%. (Lindström, Müürisep 2009)

Sõltuvussüntaktilise analüüsiga on eesti keele puhul tegeletud vaid kirjakeelega. Muischnek jt (2014a) on kasutanud selleks nii kitsenduste grammatikat kui ka andmepõhist MaltParseri süsteemi. MaltParseriga said nad tulemuseks LA 88,6%, LAS 80,3% ja UAS 83,4%. Paremaid tulemusi saavutasid Muischnek jt neid kaht lähenemist kombineerides: nii saavutati kuni 1,5 protsendipunkti kõrgem LAS. Samas nentisid nad siiski, et keerulisema süntaktilise struktuuriga (osa)lausete analüüsil tekkis probleeme mõlema süsteemi puhul. (Muischnek jt 2014a)

3. Pindsüntaktiline analüüs kohandamata analüsaatoriga

3.1. Kasutatud andmestik ja tarkvara

Andmestikuna on käesolevas töös kasutatud jututubade tekste, mis on pärit Uue Meedia korpusest (Jututubade korpus), sealt on võetud ka kõik töös kasutatud näitelaused. Oma bakalaureusetöös leidis autor, et internetikeele automaatse morfoloogilise ühestamise täpsuseks on 74,7% ning saagiseks 80,6% (Särg 2012: 16). Nagu seletatud eelmises peatükis, kasutatakse süntaktilise analüüsi teostamisel morfoloogilise analüüsi märgendeid, seega otsustati käesolevas töös võtta aluseks jututubade tekstid, mille morfoloogiline analüüs on eelnevalt käsitsi parandatud, kuna niivõrd ebakorrektse morfoloogilise analüüsiga reeglipõhine süntaktiline analüüs tõenäoliselt mingit mõistlikku tulemust ei anna.

Jututubade teksti lausestamisel on ühikuks võetud kasutaja korraga öeldud repliik, mis ei pruugi tegelikult kattuda kirjakeele mõistes lausega. Suulise keele uurimisel on Hennoste (2000: 2226) kasutanud selle tähistamiseks terminit „lausung“, mida Muischnek jt (2011: 113) on kasutanud ka internetikeele puhul. Hennoste ise seejuures kasutab internetikeelest rääkides sõna „saadetis“ (2012: 37).

Jututoa lausung (või saadetis) võib sisaldada kirjakeele seisukohast mitut lauset nagu näites 5, kus kasutaja korraga edastatud lausungi võiks jagada kolmeks lauseks: esimest lauset lõpetaks küsimärk, teine oleks *m22ratud parentiks ja kõik* ning kolmas oleks *ot ma l2hen otsin selle faili ylesse*. Samas võib üks kirjakeele mõistes lause olla jagunenud mitme lausungi vahel: näites 6 võiks kasutaja Markos kaks lausungit moodustada ühe lause *bill varastas idee ja tegi „aknad“*. Kuna aga kirjavahemärkide kasutamine jututubades on juhuslik ja vähene, siis pole võimalik ega otstarbekas kasutajate repliike kirjakeele reeglite järgi lauseteks jagama hakata.

(5)

```
"<speaker=konn>"
"<s>"

"<mis>"
	"mis" L0 P inter rel sg nom @NN>
	"mis" L0 P inter rel sg nom <0> @SUBJ
"<millise>"
	"milline" L0 P rel sg gen @NN> @ADVL @OBJ
"<configa>"
	"conf" Lga S com sg kom @NN> @<NN @ADVL
"<?>"
	"?" Z Int CLB
"<m22ratud>"
	"määra" Ltud V main partic past imps <NGP-P> @ADVL
	"määra" Ltud V main partic past imps <NGP-P> <0> @IMV
"<parentiks>"
	"parent" Lks S com sg tr @ADVL
"<ja>"
	"ja" L0 J crd CLBC @J
"<k6ik>"
	"kõik" L0 P det sg nom @OBJ
"<ot>"
	"oot" L0 B @B
"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ
"<l2hen>"
	"mine" Ln V main indic pres ps1 sg ps af <FinV> @FMV
"<otsin>"
	"otsi" Ln V main indic pres ps1 sg ps af <FinV> <NGP-P> @FMV
"<selle>"
	"see" L0 P dem sg gen @NN>
"<faili>"
	"fail" L0 S com sg gen @ADVL
"<ylesse>"
	"ülesse" L0 D @<DN @ADVL
"</s>"
```

(6)

```
"<speaker=Markos>"
"<s>"

"<bill>"
	"bill" L0 S prop sg nom @SUBJ
"<varastas>"
	"varasta" Ls V main indic impf ps3 sg ps af <FinV> <0> @FMV
"<idee>"
	"idee" L0 S com sg gen @OBJ
"</s>"
```

```
"<speaker=[carramba]>"
"<s>"

"<jep>"
	"jep" L0 B @B
"</s>"

"<speaker=Markos>"
"<s>"

"<ja>"
	"ja" L0 J crd @J
"<tegi>"
	"tege" Li V main indic impf ps3 sg ps af <FinV> <NGP-P> <0> @FMV
"<">"
	"""" Z Quo
"<aknad>"
	"aken" Ld S com pl nom @SUBJ
"<">"
	"""" Z Quo
"</s>"
```

Nagu eelmises peatükis öeldud, kasutati süntaktiliseks analüüsiks VISL-CG3 parserit (VISL-CG3) ning Kaili Müürisepa ja Tiina Puolakaineni eesti kirjakeele jaoks väljatöötatud kitsenduste grammatika reegleid, mida autor töö käigus internetikeele analüüsimiseks sobivamaks kohandas. Kohandatud reeglistik on veebist kättesaadav (vt Lisa), reeglistiku kommentaarides on märgitud, millised reeglid käesoleva töö autor lisas ja milliseid muutis. Süntaktilise analüüsi täpsuse ja saagise hindamiseks kasutati Tiina Puolakaineni koostatud Pythoni skripte.

3.2. Pindsüntaktilise analüüsi esialgsed tulemused

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. Tulemuste hindamiseks kontrollis töö autor ~100 000 sõna suurusest süntaktiliselt analüüsitud korpusest käsitsi

üle alamhulga, mille suuruseks ilma kirjavahemärkideta oli 19 809 sõnet (kuna kirjavahemärkide süntaktiline analüüs on n-ö triviaalne ülesanne, nagu on nentinud ka Muischnek jt (2014a), siis on nii siin kui edaspidi antud sõnede hulk kirjavahemärkideta ja neid pole ka arvestatud täpsuse ja saagise arvutamisel). Vastavasse alamkorpusesse kuulus 10 tekstikatkendit 9 erinevast jututoast, katkendite pikkus oli 1928–2082 sõnet.

Tabelis 2 on esitatud korpuse pindsüntaktilise analüüsi vigade jaotus. Nagu tabelist näha, sai jututubades kohandamata parseriga väära pindsüntaktilise analüüsi 1587 sõnet käsitsi kontrollitud 19 809 sõnest ehk 8,01%. Kõige enam vigu tekkis subjekti (32,07% kõigist vigadest), öeldistäite (21,55%) ning adverbiaali (10,08%) tuvastamisel.

Tabel 2. **Pindsüntaktilise analüüsi vigade jaotus**

Funktsioon	**Esinemisi tekstis**	**Vigu**	**% kõigist vigadest**
@IMV	346	4	0,25
@FMV	2885	22	1,39
@ICV	1	-	-
@FCV	292	10	0,63
@NEG	402	1	0,06
@SUBJ	3401	509	32,07
@OBJ	983	120	7,56
@PRD	666	342	21,55
@ADVL	4650	157	10,08
@NN>	709	98	6,36
@<NN	110	39	2,46
@AN>	364	21	1,32
@<AN	22	14	0,88
@DN>	128	102	6,43
@<DN	82	54	3,40
@INFN>	27	-	-
@<INFN	10	4	0,25

@KN>	1	1	0,06
@<KN	11	1	0,06
@P>	139	15	0,95
@<P	36	9	0,57
@Q>	3	3	0,19
@<Q	110	12	0,76
@Vpart	203	26	1,64
@J	848	7	0,44
@B	2134	10	0,63
@E	920	-	-
@??	326	-	-
KOKKU	**19 809**	**1587**	**100**

Pindsüntaktilise märgenduse täpsuseks oli 83,97% ning saagiseks 91,99%, seega F1-skoor oli algselt 87,80%. Täpsus oli seejuures erinevate jututubade tekstide puhul suhteliselt suuresti varieeruv: kõige madalam 81,47%, kõrgeim 86,30%. Saagis oli erinevates jututubades sarnasem, ulatudes 89,39%-st 93,76%-ni ning F1-skoorid varieerusid 85,67%-st 89,88%-ni.

84,39% jututubade sõnedest said kirjakeele reeglistiku rakendamise järel ühese analüüsi, mis on väiksem osakaal kui Liina Lindströmi ja Kaili Müürisepa (2009) poolt saadud tulemus murdekeele puhul: kirjakeele analüsaatoriga jäi ühese analüüsi saanud sõnede osakaal 89–92% vahele. Ka kirjakeele tekstide puhul saab ühese analüüsi 88–90% sõnedest (Muischnek jt 2014a: 115).

Tuleb mainida, et vaatamata sellele, et jututubade tekstide morfoloogiline analüüs oli eelnevalt käsitsi üle kontrollitud, ei ole ta siiski sajaprotsendiliselt korrektne. Selleks on kaks põhjust. Esiteks, tekste on kontrollinud ja parandanud ainult üks inimene (töö autor) ühe korra, parandamist vajanud tekstides oli aga vigu üle 25% (Särg 2012). Samas tekkisid mõningad ebatäpsused aga ka teisenduste käigus: nimelt oli morfoloogiline analüüs tehtud OÜ Filosoft analüsaatoriga t3mesta, mille väljund ei vasta täpselt kitsenduste grammatika parseri sisendi kujule. Selle illustreerimiseks on

toodud näited 7 (t3mesta väljund) ja 8 (kitsenduste grammatika parseri sisend). Näiteks ei tehta t3mesta analüüsi puhul vahet lihtmineviku ainsuse 2. pöörde ja mitmuse 3. pöörde vahel, kuna tegu on homonüümsete vormidega: näites 7 on sõnal *hakkasid* üks morfoloogiline analüüs *// _V_ sid, //*, mis tähistab mõlemat nimetatud vormi, näites 8 on samal sõnal kaks erinevat morfoloogilist analüüsi *Lsid V main indic impf ps2 sg ps af* ja *Lsid V main indic impf ps3 pl ps af*, millest esimene tähistab ainsuse 2. pööret ning teine mitmuse 3. pööret. Vale pindsüntaktilise analüüsi saanud sõnedest oli väär/mitmene morfoloogiline analüüs ~8%-l, mille parandamise ega ühestamisega käesoleva töö raames ei tegeletud.

(7)

```
<sp><speaker>
    kitu_kokkab
</speaker><p>
krt
   krt+0 //_B_ //
seal
   seal+0 //_D_ //
olid
   ole+id //_V_ sid, //
kohe
   kohe+0 //_D_ //
kaamerad
   kaamera+d //_S_ pl n, //
kohal
   kohal+0 //_D_ //
kui
   kui+0 //_J_ //
veidi
   veidi+0 //_D_ //
vabamalt
   vabamalt+0 //_D_ //
end
   ise+0 //_P_ sg p, //
tundma
   tund+ma //_V_ ma, //
hakkasid
   hakka+sid //_V_ sid, //
ju
   ju+0 //_D_ //
</p></sp>
```

(8)

```
"<speaker=kitu_kokkab>"
"<s>"

"<krt>"
	"krt" L0 B @B
"<seal>"
	"seal" L0 D @ADVL
"<olid>"
	"ole" Lid V main indic impf ps3 pl ps af <FinV> <Intr> <PhVerb> <0> @FMV
"<kohe>"
	"kohe" L0 D @ADVL
"<kaamerad>"
	"kaamera" Ld S com pl nom @SUBJ
"<kohal>"
	"kohal" L0 D @Vpart
"<kui>"
	"kui" L0 J crd CLB @J
	"kui" L0 J sub CLB @J
"<veidi>"
	"veidi" L0 D @ADVL
"<vabamalt>"
	"vabamalt" L0 D @ADVL
"<end>"
	"ise" L0 P pos det refl sg part @OBJ
"<tundma>"
	"tund" Lma V main sup ps ill <Part-P> <0> @IMV
"<hakkasid>"
	"hakka" Lsid V main indic impf ps2 sg ps af <FinV> <Intr> <Ad> <Tr> <mdainf> @FMV
	"hakka" Lsid V main indic impf ps3 pl ps af <FinV> <Intr> <Ad> <Tr> <mdainf> @FMV
"<ju>"
	"ju" L0 D @ADVL #13->11
"</s>"
```

3.2.1. Subjekt

Kõige sagedasemaks veaks oli subjektile vale märgenduse andmine – seda tüüpi vead moodustasid kõigist vigadest 32%. Kõigist vale märgenduse saanud subjektidest omakorda pea poole (47%) moodustasid ütted, mida kirjakeele analüüsireeglite kohaselt subjektiks märgendatakse, kuna kirjakeeles lihtsalt esineb ütet väga harva ja seega pole olnud vajadust eraldi märgendi jaoks. Jututubades on ütte kasutamine aga põhimõtteliselt ainus võimalus näidata, kellele tekst suunatud on, seetõttu kasutatakse seda palju. Samas tekitab ütte kasutamine samasse osalausesse sageli mitu koordineerimata subjekti nagu näites 9: nii sõnad *dodge* kui *taskud* peaksid saama subjekti märgendi, alguses see mõlemale sõnale lisataksegi, kuid siis eemaldatakse

sõnalt *taskud* just seetõttu, et samas osalauses on juba eespool üks subjekti märgendiga sõna (nimelt *dodge*) olemas. Sõna *dodge* olemasolevate märgendite hulgast valitakse hiljem välja aga hoopis eestäiendi tõlgendus. Seetõttu võib öelda, et üte on märgendusreeglite seisukohast pigem erandlik juhtum.

(9)

```
"<s>"

"<dodge>"
	"dodge" L0 S prop sg nom @NN>
"<sul>"
	"sina" Ll P pers ps2 sg ad @ADVL
"<on>"
	"ole" L0 V main indic pres ps3 pl ps af <FinV> <Intr> <0> @FMV
"<suured>"
	"suur" Ld A pos pl nom @AN>
"<taskud>"
	"tasku" Ld S com pl nom @PRD
"<?>"
	"?" Z Int CLB
"</s>"
```

3.2.2. Predikatiiv

Teiseks sagedamini esinevaks veaks on predikatiivile vale märgendi andmine: sellised juhtumid moodustasid 22% kõigist vigadest. Kirjakeeles esineb predikatiiv vaid koos väga piiratud hulga verbidega (*olema, näima, paistma* jne, EKG II: 55) ning kuna märgendusreeglid toetuvad predikatiivi määramisel just vastavate verbide olemasolule lauses, siis tekivadki vead enamasti siis, kui lausungist verb välja jäetud on. Näidetes 10 ja 11 puudub finiitses vormis *olema*-verb ning seetõttu on predikatiivid vale märgendi saanud. Näites 10 anti sõnale *paha* seetõtttu järeltäiendi märgend predikatiivi märgendi asemel, näites 11 sai sõna *õnnetu* adverbiaali märgendi, kuid peaks olema predikatiiv.

(10)

```
"<speaker=anna31>"
"<s>"

"<srry>"
	"sry" L0 B @B
"<dodgekene>"
```

```
        "dodge=kene" L0 S prop sg nom <0> @SUBJ
    "<,>"
        "," Z Com
    "<ma>"
        "mina" L0 P pers ps1 sg nom @SUBJ
    "<täna>"
        "täna" L0 D @ADVL
    "<nii>"
        "nii" L0 D @ADVL
    "<paha>"
        "paha" L0 A pos sg nom @<AN
    "</s>"
```

(11)

```
    "<speaker=jessicaz>"
    "<s>"

    "<kudagi>"
        "kuidagi" L0 D @ADVL
    "<ei>"
        "ei" L0 V aux neg @NEG
    "<lää>"
        "mine" L0 V main indic pres ps neg <FinV> <0> @FMV
    "<ma>"
        "mina" L0 P pers ps1 sg nom @SUBJ
    "<suht>"
        "suhteliselt" L0 D @ADVL
    "<õnnetu>"
        "õnnetu" L0 A pos sg nom @ADVL
    "</s>"
```

3.2.3. Täiend

Kolmanda suurema veagrupina saaks välja tuua täiendid, millele vale märgendi andmine moodustab kokku 21% kõigist vigadest. Nagu näidatud peatükis 2.2.1, esineb märgendussüsteemis aga tegelikult viit liiki täiendeid: nimisõnaline, omadussõnaline, adpositsioonifraas, infinitiivtäiend ning adverbiline täiend, lisaks sisaldab täiendi märgend veel infot selle kohta, kas peasõna asub täiendist lauses vasakul või paremal pool. Seega, kui täiendid vastavalt märgendussüsteemile liikideks jagada, tuleb välja, et kõige sagedamini jäävad ära tundmata adverbilised täiendid, mis moodustab 10% kõigist vigadest. Üheks selliste vigade põhjuseks on ortograafia: kui adverbist ja käändsõnast koosnev liitsõna on lahku kirjutatud nagu näites 12 *kokku tulek,* siis peaks adverb *kokku* jääma käändsõna täiendiks ja saama märgendi @DN>. Kirjakeeles aga

kirjutatakse sellised väljendid liitsõnana ja adverbilisele komponendile seega eraldi süntaktilist märgendit lisada vaja ei ole.

(12)

```
"<speaker=@Crizzu>"
"<s>"

"<kuivikute>"
	"kuivik" Lte S com pl gen @NN>
"<kokku>"
	"kokku" L0 D @ADVL
"<tulek>"
	"tulek" L0 S com sg nom <0> @SUBJ
"<:P>"
	":P" E @E
"</s>"
```

Samas ei puuduta ortograafiavigade tõttu äratundmata jäänud täiendid mitte ainult adverbe – näites 13 peaks sõna *kaubandus* saama täiendi märgendi, aga kuna konstruktsioon kirjakeele reeglitele ei vasta, siis on see loetud hoopis subjektiks.

(13)

```
"<s>"

"<miskis>"
	"mingi" Ls P indef sg in @NN>
"<kaubandus>"
	"kaubandus" L0 S com sg nom <0> @SUBJ
"<keskuses>"
	"keskus" Ls S com sg in @<NN @ADVL
"<?>"
	"?" Z Int CLB
"</s>"
```

3.2.4. Muu

Vale märgendi saanud adverbiaalid moodustasid 10%, objektid 8% vigadest ning ülejäänud vealiigid kokku (adpositsiooni ja kvantori laiendid, ühendverbi koosseisu kuuluvad afiksaaladverbid, predikaadid, millel mitmeid erinevaid märgendeid vastavalt sellele, kas tegu on põhi- või abiverbi ja finiitse või infiniitse vormiga, konjunktsioonid

ja muud tüüpi vead) moodustasid kokku alla 8% kõigist vigadest. Loomulikult ei tähenda väljatoodud protsendid seda, et kvantori laiendite määramine tegelikult oleks näiteks nii palju korrektsem kui objekti määramine, vaid kvantori laiendeid lihtsalt esinebki tekstis väga vähe ja seega tekitavad nad ka vähe vigu. Öeldisi esineb aga loomulikult tekstis palju, nende märgendamine on see-eest suhteliselt lihtne, kuna põhi- vs abiverbi märgendamine ning finiitse ja infiniitse vormi määramine põhineb väga otseselt morfoloogilisel analüüsil.

3.2.5. Mitmesused

1487 sõna ehk 7,5% jututubade korpuse sõnadest said külge nii oma õige märgendi kui mõne(d) vale(d) märgendi(d) ehk nende analüüs jäi mitmeseks. Enamik (58%) sellistest sõnadest on määrused, samuti tuleb seda ette subjektide (17%), objektide (9%) ja nimisõnaliste eestäienditega (7%), ülejäänud sõnaliikide osakaal kokku on alla 10% mitmeseks jäänud analüüsiga sõnadest.

Mitmeseks jäänud sõnadele anti enamasti (1223 juhul ehk 82% juhtudest) kaks erinevat märgendit, 233 juhul anti kolm märgendit ning 29 juhul neli eri märgendit (tegu numbrite ja lühenditega). Tabelis 3 on ära toodud kümme kõige sagedasemat mitmesuse liiki. Sealt on näha, et kõige enam mitmesusi tekib sellest, et ei suudeta valida määruse ja järeltäiendi tõlgenduse vahel. Niisugune mitmesus jääb alles nii määrsõnade kui nimi-, ase- ja arvsõnade puhul. Näites 14 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. Näites 15 on samal põhjusel antud sõnale *tehnikamaailmas* antud @ADVL ja @<NN märgendid.

(14)

```
"<speaker=marqs>"
"<s>"

"<tegelt>"
      "tegelikult" L0 D @ADVL
"<winil>"
      "win" Ll S prop sg ad @ADVL
"<oli>"
```

```
        "ole" Li V main indic impf ps3 sg ps af <FinV> <Intr> <0> @FMV
"<subpixeli>"
        "subpixel" L0 S com sg gen @NN>
"<värk>"
        "värk" L0 S com sg nom @SUBJ
"<isegi>"
        "isegi" L0 D @ADVL @<DN
"<enne>"
        "enne" L0 D @ADVL
"<olemas>"
        "ole" Lmas V main sup ps in <Intr> @ADVL
"<vist>"
        "vist" L0 D @ADVL
"<,>"
        "," Z Com CLBC
"<kui>"
        "kui" L0 J crd CLB @J
        "kui" L0 J sub CLB @J
"<X'il>"
        "X" Ll S prop sg ad cap @ADVL
"</s>"
```

(15)

```
"<speaker=konn>"
"<s>"

"<progemine>"
        "programmeeri=mine" L0 S com sg nom <mine> "programmeeri" @SUBJ
"<on>"
        "ole" L0 V main indic pres ps3 sg ps af <FinV> <Intr> <0> @FMV
"<m6ttetu>"
        "mõttetu" L0 A pos sg nom @PRD
"<,>"
        "," Z Com CLB
"<proged>"
        "programmeeri" Ld V main indic pres ps2 sg ps af <FinV> @FMV
"<paar>"
        "paar" L0 N card sg nom l @ADVL
"<aastat>"
        "aasta" Lt S com sg part @<Q
"<ja>"
        "ja" L0 J crd CLB CLBC @J
"<siis>"
        "siis" L0 D @ADVL
"<oled>"
        "ole" Ld V main indic pres ps2 sg ps af <FinV> <Intr> @FMV
"<yleyldises>"
        "üle_üldine" Ls A pos sg in @AN>
"<m6ttes>"
        "mõte" Ls S com sg in @NN> @ADVL
"<jobu>"
        "jobu" L0 S com sg nom @PRD
```

```
"<tehnikamaailmas>"
	"tehnika_maa_ilm" Ls S com sg in @<NN @ADVL
"</s>"
```

Tabelist 3 võib järeldada, et enamik mitmesustest tekib nimi-, ase- ja arvsõnadest: kui välja arvata esimene mitmesuse liik @ADVL @<DN, mis saab esineda vaid määrsõnade puhul, siis kõik ülejäänud mitmesused on võimalikud vaid nimi-, arv- ja asesõnade puhul.

Tabel 3. **Sagedasemad mitmesuse liigid**

Mitmesus	**Esinemisi**	**% kõigist mitmesustest**
@ADVL @<DN	335	22,53
@ADVL @<NN	243	16,34
@ADVL @NN>	128	8,61
@SUBJ @OBJ	113	7,60
@SUBJ @NN>	112	7,53
@ADVL @SUBJ @OBJ	90	6,05
@ADVL @SUBJ	61	4,10
@NN> @<NN @ADVL	52	3,50
@NN> @OBJ	43	2,89
@OBJ @ADVL	26	1,75

4. Sõltuvussüntaktiline analüüs kohandamata reeglistikuga

Sõltuvussüntaktilise analüüsi algtaseme tulemuste saamiseks kasutati sama andmestikku, tarkvara ja metoodikat, mida pindsüntaktilise analüüsi jaoks ja mida on kirjeldatud peatükis 3.1. UAS ehk sõltuvussüntaktiliste märgendite 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 ehk selliste sõnade saagis, millel on õige nii pindsüntaktiline kui sõltuvusmärgend, oli korpusel 71,86%, seejuures loeti õigete hulka ka need juhud, kui pindsüntaktilisi märgendeid oli antud mitu ja üks nendest õige oli.

4.1. Vigade jaotus

Tabelis 4 on toodud vigade jaotus kogu korpuses. Sõnadest, mis ei saanud õiget ülemuse märgendit ja mis loeti seega vigade hulka, jäid 30% üldse sõltuvusmärgendita ning 70% said vale märgendi. Seejuures olid kirjakeele analüüsireeglitele siiski lisatud kõige põhilisemad reeglid uute pindsüntaktiliste märgenditega (@E ja @B vastavalt emotikoni ja partikli jaoks, vt alapeatükk 3.2) toimetulekuks – nimelt märgendati nende ülemuseks lause kõige kõrgem ülemus (verb või subjekt).

Nagu tabelist 4 näha, moodustavad ligi poole kõigist vigadest partiklid ja adverbiaalid, esimesed 24,69% ning teised 24,63%. Lisaks on muudest lauseliikmetest suurem vigade osakaal veel subjektidel (13,62%) ja finiitsetel öeldistel (8,26%). Kuna tegu on ka korpuses kõige enam esindatud lauseliikmetega, siis on see ootuspärane.

Tabel 4. **Sõltuvussüntaktiliste vigade jaotus**

Funktsioon	Esinemisi korpuses	Vigu	% kõigist vigadest
@IMV	346	69	1,39
@FMV	2885	410	8,26
@ICV	1	-	-
@FCV	292	26	0,52
@NEG	402	3	0,06
@SUBJ	3401	676	13,62
@OBJ	983	218	4,39
@PRD	666	171	3,45
@ADVL	4650	1222	24,63
@NN>	709	148	2,98
@<NN	110	68	1,37
@AN>	364	21	0,42
@<AN	22	8	0,16
@DN>	128	63	1,27
@<DN	82	57	1,15
@INFN>	27	-	-
@<INFN	10	7	0,14
@KN>	1	1	0,02
@<KN	11	9	0,18
@P>	139	11	0,22
@<P	36	9	0,18
@Q>	3	2	0,04
@<Q	110	12	0,24
@Vpart	203	7	0,14
@J	848	159	3,20
@B	2134	1225	24,69
@E	920	360	7,26
@??	326	-	-
KOKKU	**19 809**	**4962**	**100**

4.2. Vigade põhjused

4.2.1. Ülemuse märgendita sõnad

Ülemuse märgendita jäänud sõnadest enamik ehk 64% olid partiklid ning 17% emotikonid. Nimelt puudusid reeglid, mis määraksid ülemusi partiklitele ja emotikonidele, kui nad esinevad kas üksi või kombineeruvad omavahel (vt näited 16 ja 17: kui sõna ülemuseks on märgitud tema enda järjekorranumber, siis tähendab see, et ükski ülemuse määramise reegel pole rakendunud ehk ülemus on jäänud määramata) või olid muudes taolistes konstruktsioonides, mida kirjakeeles iseseisva lausena ei esine. 10% ülemuse märgendita jäänud sõnadest olid adverbiaalid ning 3% objektid, ülejäänute osakaal jäi kõigil 0–2% vahemikku.

(16)

```
"<speaker=Viviann>"
"<s>"

"<heia>"
        "heia" L0 B @B #1->1
"<:))>"
        ":)" E @E #2->2
"</s>"
```

(17)

```
"<speaker=angel>"
"<s>"

"<uh>"
        "uh" L0 B @B #1->1
"<jah>"
        "jah" L0 B @B #2->2
"</s>"
```

4.2.2. Vale ülemuse märgendiga sõnad

Vale ülemuse märgendi saanud sõnadest 33%-l oli ka vale pindsüntaktiline analüüs ning kuna sõltuvussüntaktilise märgendamise reeglid lähtuvadki peamiselt pindsüntaktilisest märgendusest, siis seda tüüpi vigade vastu on raske midagi ette võtta.

Sõnadest, millel oli õige pindsüntaktiline analüüs, kuid määratud vale ülemus, oli kõige enam adverbiaale (34%), suhteliselt palju ka öeldisi (21%), subjekte (15%) ning partikleid (10%). Kuna märgendussüsteemis ei tehta vahet lause- ja fraasiadverbiaalidel ning üldlaienditel, siis võivad adverbiaalid laiendada väga erinevaid lauseliikmeid, seetõttu on nendele õige ülemuse määramine keeruline (vt näide 18 – sõna *peaaegu* ülemuseks on märgitud öeldis *elan,* ent siin võiks olla tegelikult ülemuseks *yxinda*).

(18)

```
"<speaker=armsake>"
"<s>"

"<ma>"
    "mina" L0 P pers ps1 sg nom @SUBJ #1->2
"<elan>"
    "ela" Ln V main indic pres ps1 sg ps af <FinV> <Intr> <In> <Ad> <0> @FMV #2->0
"<ka>"
    "ka" L0 D @ADVL #3->2
"<peaaegu>"
    "pea_aegu" L0 D @ADVL #4->2
"<yxinda>"
    "üksinda" L0 D @ADVL #5->2
"</s>"
```

Öeldiste puhul tekivad probleemid osalausete omavaheliste suhete määramisest, seda eriti näiteks juhtudel, kui lause kõige kõrgemaks ülemuseks ei ole määratud mitte öeldis, vaid subjekt. See on juhtunud näites 19: subjekti märgendiga sõna *vaikija_lambakarjus* on saanud ülemuse märgendiks 0, mis tähendab, et tegu on lause kõige kõrgema ülemusega. Tegelikult on see sõna antud näites aga üte ja selles lauses peaks kõige kõrgemaks ülemuseks olema määratud öeldis *oskad*, millele alluks ka sõna *vaikija_lambakarjus.*

(19)

```
"<speaker=Katrin>"
"<s>"

"<vaikija_lambakarjus>"
	"vaikija_lambakarjus" L0 S <0> prop sg nom @SUBJ #1->0
"<,>"
	"," Z Com CLB #2->2
"<sa>"
	"sina" L0 P pers ps2 sg nom @SUBJ #3->4
"<oskad>"
	"oska" Ld V main indic pres ps2 sg ps af <FinV> <Part> <InfP> <mdainf> <0> @FMV #4->1
"<vaid>"
	"vaid" L0 D @DN> #5->6
"<karvu>"
	"karv" Lu S com pl part @OBJ #6->7
"<ajada>"
	"aja" Lda V main inf <NGP-P> @OBJ #7->4
"</s>"
```

Nii öeldiste kui subjektide ülemuse määramise vigade põhjuseks on sageli ka osalausepiiride puudumine. Näites 20 on nii subjekt *ma* kui ahelverbi *pean elama* koosseisu kuuluv *pean* määratud tegelikult eelmisesse osalausesse kuuluva öeldise *tundub* külge, kuna sõna *et* eest on puudu koma ja seetõttu on jäänud osalausepiir paika panemata.

(20)

```
"<speaker=+tibu>"
"<s>"

"<tundub>"
	"tundu" Lb V main indic pres ps3 sg ps af <FinV> <Intr> <0> @FMV #1->0
"<et>"
	"et" L0 J sub @J #2->9
"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ #3->1
"<pean>"
	"pida" Ln V mod indic pres ps1 sg ps af <FinV> <NGP-P> <mdainf> @FCV #4->1
"<oma>"
	"oma" L0 P pos det refl sg gen @NN> #5->7
"<raske>"
	"raske" L0 A pos sg gen @AN> #6->7
"<kõhuga>"
	"kõht" Lga S com sg kom @ADVL #7->9
```

"<edasi>"
	"edasi" L0 D @ADVL @<DN #8->9
"<elama>"
	"ela" Lma V main sup ps ill <Intr> <In> <Ad> @IMV #9->1
"</s>"

5. Pindsüntaksi reeglistiku kohandamine internetikeelele

Eesti keele pindsüntaktiline analüüs ESTKG grammatikaga toimub järgmiste etappidena:

1. morfoloogiline analüüs
2. osalausepiiride määramine
3. morfoloogiline ühestamine
4. ühendverbide tuvastamine
5. pindsüntaktiline märgendamine

Kuna käesolevas töös kasutatud tekstide morfoloogiline analüüs oli juba eelnevalt käsitsi korrastatud, nagu peatükis 3 mainitud, siis morfoloogilise analüüsi ja ühestamise reeglite kohandamine jäeti vahele.

Kohandamisel kasutati n-ö arenduskorpusena peatükis 3.2 kirjeldatud korpuse 5909 sõnast koosnevat alamkorpust, mis sisaldas tekste kolmest jututoast ning mille esialgsed täpsus ja saagis olid sarnased kõigi 10 jututoa pealt saadutega (nagu mainitud peatükis 3.2, siis jututoati olid nimetatud näitajad suhteliselt varieeruvad). Testkorpusena kasutati 5821 sõne suurust teist 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 loomulikult nad ka omavahel ei kattunud.

Reeglite kohandamise protsess nägi välja selline, et esmalt parsiti arenduskorpus kirjakeele reeglistikuga, kõrvutati tulemust arenduskorpuse versiooniga, mille süntaktiline analüüs käsitsi parandatud oli, ning leitud vigade põhjal tehti reeglistikku muudatusi. Arvutati välja ka täpsus ja saagis. Seejärel parsiti arenduskorpus uuesti kohandatud reeglistikuga ning korrati sama protsessi, kuni täpsuses ja saagises enam paranemist saavutada ei õnnestunud.

5.1. Osalausepiiride reeglite kohandamine

Osalausepiirid on info, mida kasutavad paljud süntaktiliste funktsioonide määramise kitsendused, seetõttu on nende eelnev kindlaksmääramine vajalik. Transkribeeritud suulise keele puhul mainivad Kaili Müürisep ja Heli Uibo (2005) koguni, et kõige rohkem oligi parseri kohandamisel suulise keele jaoks abi osalausepiiride reeglite muutmisest – sel viisil saavutasid nad transkribeeritud suulise keele jaoks isegi parema tulemuse kui oli tookord kirjakeele puhul (täpsus 89,2% ja saagis 97,3%), konstateerides samas, et parem tulemus võis tulla tänu sellele, et suulise keele laused on lühemad ja lihtsamad.

Jututubade laused on samuti pigem lühikesed ja koosnevadki tegelikult enamasti ühestainsast osalausest. Mitmest osalausest koosnevate lausete puhul sageli ei kasutata korrektset kirjavahemärgistust. Kuna kirjakeele osalausestamisreeglid baseeruvad aga suures osas just kirjavahemärgistusel, siis oligi vaja lisada juurde reegleid, mis aitaksid paika panna osalausepiiri juhul, kui kasutaja kirjavahemärke kasutanud ei ole. Selleks lisasin reeglistikku olemasolevale 67 reeglile (paljud neist on väga spetsiifiliste juhtude jaoks ning jututubade arenduskorpusele rakendusid neist tegelikult vaid 28) kümmekond täiendavat osalausepiiride määramise reeglit, lisaks lisasin mõnedele reeglistikus kasutatud sõnaloenditele ja märgiloenditele omalt poolt jututubades kasutatavaid vorme (näiteks sõna *miuke*, kolm punkti (...) osalauseid eraldava kirjavahemärgina jms).

Uute reeglitega lahendatud tüüpjuhud olid näiteks järgmised:

- lisada alistavale sidesõnale osalausepiiri märgend sõltumata sellest, kas talle koma eelneb või mitte (v.a sõna *nagu* puhul) – kirjakeeles kasutatakse vastavas kohas koma, ent internetikeeles on komade ärajätmine vägagi levinud, alistav sidesõna on aga piisavalt kindlaks osalausepiiri indikaatoriks ka ise. Näites 21 on toodud üks niisugune juhtum: sõna *et* ees oleks kirjakeeles koma, mille külge pandaks osalausepiiri märgend (CLB või CLBC). Jututoa lausungis on aga koma panemata jäetud ja seetõttu on vaja osalausepiir märkida osalause

esimese sõna ehk sõna *et* külge, nagu uue reegli rakendumisel tehtigi.

(21)

```
"<speaker=+Adi>"
"<s>"

"<mis>"
    "mis" L0 P inter rel sg nom
"<rahvas>"
    "rahvas" L0 S com sg nom
"<teeb>"
    "tege" Lb V main indic pres ps3 sg ps af <FinV> <NGP-P> <0>
"<et>"
    "et" L0 J sub CLB
"<ei>"
    "ei" L0 V aux neg
"<räägi>"
    "rääki" L0 V main indic pres ps neg <FinV> <Part-P> <El>
"<?>"
    "?" Z Int CLB
"</s>"
```

- lisada osalausepiir muudele sageli uue osalause algust tähistavatele sõnadele (nagu *kes*, *mis*, *seega*, *järelikult* jne), kui ümbritsev kontekst seda kuidagi otseselt ei takista (nt ei eelne ega järgne kirjavahemärk, mille külge osalausepiir juba pandud on). Kirjakeeles eelneb nendelegi sõnadele osalausepiiri tähistav koma, jututubades jällegi sageli mitte.
- lisada osalausepiiri märgend teisele verbile, kui kõrvuti on kaks finiitset verbivormi, mõlemad on põhiverbid ja esimene pole *minema* ega *tulema* (et vältida osalausepiiride lisamist konstruktsioonidele nagu *lähen teen*, mida Mati Erelt (2013: 56) nimetab rindverbideks. Seega näites 22 lisati vastava reegli rakendumisel osalausepiir sõna *ütlege* külge, kuna ka talle eelnev sõna *on* on finiitne põhiverb.

(22)

```
"<speaker=BaltMan>"
"<s>"

"<palju>"
```

```
        "palju" L0 D
"<rahvast>"
        "rahvas" Lt S com sg part
"<siin>"
        "siin" L0 D
"<on>"
        "ole" L0 V main indic pres ps3 sg ps af <FinV> <Intr> <0>
"<ütlege>"
        "ütle" Lge V main imper pres ps2 pl ps af <FinV> <NGP-P> <0> CLBC
"<JAA>"
        "jaa" L0 B <ccap>
"</s>"
```

Kuna internetikeeles on liitlauseid aga võrdlemisi vähe, siis suurt paranemist täpsuses ja saagises osalausepiiride reeglite kohandamine ei andnud. Testkorpuse algseks 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 tuvastamise reeglite kohandamine

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

Ühendverbid internetikeeles on üldjoontes samad kirjakeele omadega, kuid kaks väikest muudatust reeglistikus siiski tehti. Ühendverbide tuvastamine põhineb nii leksikonil kui 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* (*järgi vaatama*, aga *järele proovima*). Internetikeeles (ja üleüldse kõnekeeles) neil aga enamasti vahet ei tehta, kasutatakse näiteks nii vorme *järgi vaatama* kui *järele vaatama*. Seetõttu

otsustasin need kaks loendit ühendada.

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

5.3. Pindsüntaktilise analüüsi reeglite kohandamine

5.3.1. Üte

Üte on nimetavas käändes nimisõnafraasiline kiil, mille eesmärgiks on näidata, kelle poole pöördutakse (EKK: 434). Nagu öeldud peatükis 3.2.1, märgendati kirjakeele reeglistiku järgi ütted subjektideks, kuna kirjakeeles on ütete osakaal tekstis marginaalne. Internetikeeles aga esineb ütteid palju ning seetõttu otsustas töö autor ütte subjektist eristada ning võtta kasutusele eraldi märgendi @VOC, nagu on tehtud näiteks saami keele puhul (Sami Syntactic Tags).

Kirjakeeles eristatakse üte muust tekstist komadega, internetikeeles aga ainult sellest lähtuda ei saa. Kuna enamasti kasutatakse üttena kas pärisnime või teatavaid n-ö mingisse rühma kuulumist märkivaid nimisõnu (*mehed, rahvas, inimesed, sõbrad* jne), siis otsustaski töö autor ütetena kaaluda vaid nimetavas käändes pärisnimesid (sealhulgas jututoa kasutajanimed, mis moodustavadki suurema osa ütetest) ning vastavasse defineeritud loendisse kuuluvaid nimisõnu. Ütte märgendamine käis üldiselt sõnastatuna järgmiselt:

- nimetavas käändes pärisnimedele ja n-ö ütteks sobivatele sõnadele lisati märgend @VOC;
- kui samas osalauses juba üheselt määratud subjekt olemas oli, siis jäeti @VOC märgendi saanud sõnadele alles vaid ütte tõlgendus;
- kui sõnale oli antud nii ütte kui subjekti märgend ja järgmine sõna oli verb (v.a imperatiivne verb), siis eemaldati ütte märgend;

- kuna kirjakeele reeglistikus märgendati üte subjektiks, siis oli mugav teatud juhtudel kasutada asendusreegleid: näiteks asendada subjekti märgend ütte omaga, kui nimetavas käändes pärisnimele eelneb või järgneb partikkel (näide 23), lausung koosnebki ainult nimetavalisest pärisnimest ning emotikoni(de)st (mis on jututubades võrdlemisi tavaline viis tervitamiseks), ütteks sobivat sõna eraldavad muust lausest komad vms;

(23)

```
"<speaker=konn>"
"<s>"

"<markos>"
	"markos" L0 S prop sg nom <0> @VOC
"<kle>"
	"kle" L0 B @B
"</s>"
```

- samuti asendati subjekti märgend ütte omaga väga internetikeelespetsiifilisel juhul: nimelt kui sellele eelnes @-märk (vt näide 24: sõna *Adi* saab @VOC märgendi, kuna eelneb @). Tegu on ainult internetis kasutusel oleva pöördumise vormiga ning loomulikult puudub see variant igasugustes eesti keele süntaksi kirjeldustes, kuid käesoleva töö autori meelest sarnaneb konstruktsioon piisavalt üttega, et selleks liigitatud saada.

(24)

```
"<speaker=+Sigitis>"
"<s>"

"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ
"<tean>"
	"tead" Ln V main indic pres ps1 sg ps af <FinV> <NGP-P> <InfP> <0> @FMV
"<@>"
	"@" Z Gnr #3->3
"<Adi>"
	"Adi" L0 S prop sg nom cap @VOC
"</s>"
```

- kui nimetavas käändes pärisnimi oli lisaks ütte märgendile saanud täiendi, objekti või adverbi märgendi, siis need eemaldati, kuna vastav kasutus on võrdlemisi ebatõenäoline võrreldes üttega.

5.3.2. Predikatiiv

Nagu öeldud alapeatükis 3.2.2, põhineb predikatiivi märgendamine kirjakeeles sobilike verbide olemasolul lauses. Kuna aga *olema*-verbi ärajätmine on internetikeeles võrdlemisi tavapärane, siis vajas predikatiivi tuvastamine täiendamist just (osa)lausetes, kus verb puudub. Selleks lisati reegel, mis lisab nimetavas käändes adjektiivile predikatiivi märgendi, kui samas osalauses verbi pole. Vastava märgendi lisamine ei tähenda muidugi, et kõik nimetavas käändes adjektiivid öeldiseta osalausetes oleksidki lõpuks predikatiivi märgendusega jäänud: näiteks rakendub hiljem reegel, mis kõigi lisatud tõlgenduste hulgast jätab alles omadussõnalise eestäiendi oma, kui vahetult järgneb nimetavas käändes nimisõna.

Predikatiiv saab küll olla ka nimisõnafraas, kuid verbita lauses on nimetavas käändes nimisõnafraas sagedamini siiski subjekt (milleks kirjakeele reeglistiku järgi üksi esinev nimisõnafraas märgendataksegi) või üte ning sellistel juhtudel predikatiivimärgenditega lisasegadust tekitama hakata ei tundunud otstarbekas.

5.3.3. Täiend

Osas 3.2.3 mainiti, et täiendite tuvastamine tekitas palju vigu. Samas, kuna täiendite jaoks on kümme erinevat märgendit, siis kõiki neid eraldi vaadates olid vead pigem suhteliselt juhuslikku laadi. Järgnevalt mõningad kohandused, mis täiendite tuvastamise osas tehti:

- sõnade *täna, eile, homme* puhul valida alati adverbilise eestäiendi märgend, kui järgneb mingit päevaaega tähistav sõna: näites 25 järgneb sõnale *täna* päevaaega tähistav *päeval*, seega valitakse sõnale *täna* antud adverbiaali ja adverbilise täiendi hulgast välja viimane. Kuna jututubades räägitakse sageli oma igapäevaelust, siis esineb seda sorti konstruktsioone suhteliselt palju;

(25)

```
"<speaker=Y2K>"
"<s>"

"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ
"<täna>"
	"täna" L0 D @DN>
"<päeval>"
	"päev" Ll S com sg ad @ADVL
"<peaks>"
	"pida" Lks V mod cond pres ps af <NGP-P> <mdainf> @FCV
"<saama>"
	"saa" Lma V main sup ps ill <NGP-P> <0> @IMV
"</s>"
```

- sõna *kus* lisati sõnade loendile, mis kohamäärusele eelnedes adverbilise täiendi moodustavad, seega näites 26 saab *kus* adverbilise eestäiendi märgendi, enne aga määrati ta määruseks;

(26)

```
"<speaker=armsake>"
"<s>"

"<Renka>"
	"Renka" L0 S <0> prop sg nom cap @VOC @SUBJ
"<,>"
	"," Z Com CLB
"<kus>"
	"kus" L0 D @DN>
"<kandis>"
	"kant" Ls S com sg in @NN> @ADVL
"<sa>"
	"sina" L0 P pers ps2 sg nom @SUBJ
"<elad>"
	"ela" Ld V main indic pres ps2 sg ps af <FinV> <Intr> <In> <Ad> <0> @FMV
"<:P>"
	":P" E @E
"</s>"
```

- sõna *mõni* puhul valida täiendi tõlgendus, kui järgneb samas käändes nimisõna – vastav reegel kehtib küll erinevatesse alaliikidesse kuuluvate pronoomenite kohta (possessiiv-, demonstratiiv-, determinatiivpronoomenid), kuid mitte täpsemalt liigitamata pronoomenite kohta nagu *mõni*.

5.3.4. Muu

Väikseid kohandusi tehti ka muude märgendite osas, näiteks lisati reegel, mis annab prepositsiooni *ilma* järel olevale abessiivis nimi- või asesõnale kaassõnafraasi laiendi märgendi (@<P) – mitte kuidagi internetikeelega seotud reegel, vaid tõenäoliselt on kirjakeele reeglistikus kogemata käsitletud sõna *ilma* postpositsioonina. Samuti täiendati mõningaid juba olemasolevate reeglite hulgast, lisades uute märgendite – ütte ja partkliga – seotud kitsendusi.

5.4. Tulemused kohandatud reeglitega

Reeglite kohandamise järel oli testkorpusel saagiseks 93,35% ja täpsuseks 85,16%, F1-skoor seega 89,07%. Seejuures tuleb arvesse võtta, et algsete näitajate arvutamisel ei eristatud ütet subjektist: kui kohandatud reeglitega pandud märgenduse puhul mitte lugeda vigadeks juhte, kus subjekt on märgendatud ütteks või vastupidi, siis oleks saagis 94,43% ning täpsus 86,54%.

Tabelis 5 on esitatud erinevate lauseliikmete puhul tekkinud vigade arvud ning vastavate märgenditega sõnade saagised testkorpuses enne ja pärast analüüsireeglite kohandamist. Nagu näha, siis ka pärast reeglite kohandamist esineb 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 ütte puhul 91,18% ning subjektide (ütted välja arvatud) puhul 85,04%. Predikatiivide saagis tõusis 49,49%-lt 70,92%-le. Teiste suurema esindatusega märgendite osas vähenes veaprotsent kuni paari protsendipunkti võrra, välja arvatud adverbiaalide puhul, mille veaprotsent kasvas poole protsendipunkti võrra – seda seetõttu, et adverbiaalide klassi käsitletakse n-ö ülejäägina: kui sõna vastab määruse tingimustele ning talle pole muid funktsioone määratud, siis loetakse ta määruseks (Müürisep 2000: 66).

Tabel 5. **Vead testkorpuses enne ja pärast analüüsireeglite kohandamist internetikeelele**

Süntaktiline funktsioon	Esinemisi korpuses	Kohandatud reeglistik			Kohandamata reeglistik		
		Vigu	Saagis (%)	% kõigist vigadest	Vigu	Saagis (%)	% kõigist vigadest
@IMV	117	2	98,29	0,52	3	97,44	0,64
@FMV	810	7	99,14	1,84	7	99,14	1,49
@ICV	-	-	-	-	-	-	-
@FCV	89	1	98,88	0,26	1	98,88	0,21
@NEG	96	1	98,96	0,26	1	98,96	0,21
@SUBJ	695	104	85,04	27,30	181	83,07	38,43
@VOC	374	33	91,18	8,66			
@OBJ	278	33	88,13	8,66	34	87,77	7,22
@PRD	196	57	70,92	14,96	99	49,49	21,02
@ADVL	1434	49	96,58	12,86	42	97,07	8,92
@NN>	208	18	91,35	4,72	18	91,35	3,82
@<NN	26	8	69,23	2,10	6	76,92	1,27
@AN>	102	8	92,16	2,10	8	92,16	1,70
@<AN	4	4	0,00	1,05	4	0,00	0,85
@DN>	47	28	40,43	7,35	36	23,40	7,64
@<DN	24	11	54,17	2,89	11	54,17	2,34
@INFN>	-	-	-	-	-	-	-
@<INFN	1	1	00,00	0,26	1	0,00	0,21
@KN>	-	-	-	-	-	-	-
@<KN	4	-	100,00	-	-	100,00	-
@P>	40	5	87,50	1,31	5	87,50	1,06
@<P	13	2	84,62	0,52	3	76,92	0,64

@Q>	1	1	0,00	0,26	1	0,00	0,21
@<Q	31	2	93,55	0,52	2	93,55	0,42
@Vpart	49	3	93,88	0,79	4	91,84	0,85
@J	204	-	100,00	-	0	100,00	-
@B	624	3	99,52	0,79	3	99,52	0,64
@E	218	-	100,00	-	-	100,00	-
@??	136	-	100,00	-	-	100,00	-
KOKKU	5821	381	93,35	100	471	92,15	100

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 (vt peatükk 3.2.5), kuid nende osakaalud vähenesid pisut. Uusi mitmesusi lisandus ütte märgendi lisamise tõttu süsteemi, enim niisugusi, kus antud märgenditeks olid subjekt ja üte.

Võrreldes teiste eesti keele variantidega, millele reeglistikku kohandatud on – nimelt transkribeeritud suulise keele ja murdekeelega – , on internetikeelel saadud tulemused pisut madalamad. Põhjusi võib otsida sellest, 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.

Vigade jaotust on võimalik võrrelda Tiina Puolakaineni (2013) leitud ajalehetekstide vigade jaotusega. Nagu näha tabelist 4, moodustavad internetikeeles ka kohandatud reeglistikuga subjekti ja ütte märgendamise vead kokku 35,96% kõigist vigadest, järgnevad predikatiivide vead (14,96%) ning adverbiaalid (12,86%). Ka ajalehetekstides on kõige sagedasemaks veaks subjektile (mille alla on arvatud ka ütted, kui neid esineb) vale märgendi andmine, kuid see moodustab vaid 16,3% vigadest. Pea sama palju vigu tekib ajalehetekstides adverbiaalide (16,03%) ning nimisõnaliste eestäiendite tõttu

(14,55%). (Puolakainen 2013) Internetikeeles seevastu tekib nimisõnalistest eestäienditest vaid 4,72% kõigist vigadest. Seega võib öelda, et internetikeeles ja kirjakeeles on vigade jaotus võrdlemisi erinev.

6. Sõltuvussüntaktilise analüüsi reeglite kohandamine

Kuna sõltuvussüntaktiline märgendamine põhineb peamiselt pindsüntaktilise analüüsi märgenditel, siis tehakse see pindsüntaktilise analüüsi järel. Ka käesoleva töö raames järgnes sõltuvussüntaksi reeglite kohandamine pindsüntaksi reeglite kohandamisele. Arendus- ning testkorpusena kasutati samu korpusi, mida peatükis 5, kuid pindsüntaktiline märgendamine oli tehtud juba kohandatud reeglistikuga. Peatükis 4 mainiti, et erinevates jututubades oli esialgne sõltuvussüntaktiliste märgendite saagis võrdlemisi kõikuv, kuid nii arendus- kui testkorpuse saagis oli sellegipoolest väga lähedane kogu korpuse esialgsele saagisele (74,95%): arenduskorpusel 75,32% ning testkorpusel 75,03%.

Järgnevates alapeatükkides kirjeldatakse lauseliikmete kaupa sõltuvvussüntaksi reeglite kohandamist internetikeelele.

6.1. Üte

Kuna pindsüntaktiliste reeglite kohandamise käigus lisati üttele eraldi märgend @VOC, siis oli vaja sellele ka 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 üldjuhul määrata ütte ülemuseks selle (osa)lause kõige kõrgema ülemuse, mille juurde ta kuulub: näites 27 on ütte *rallikas* ülemuseks määratud seega esimese osalause öeldis *tuled*.

(27)

```
"<speaker=+mamamia>"
"<s>"

"<rallikas>"
	"rallikas" L0 S prop sg nom @VOC #1->3
"<äkki>"
	"äkki" L0 D @<DN @ADVL #2->3
"<tuled>"
	"tule" Ld V main indic pres ps2 sg ps af <FinV> <Intr> <0> @FMV #3->0
"<ise>"
	"ise" L0 P pos det refl sg nom @ADVL #4->3
"<siia>"
	"siia" L0 D @ADVL #5->3
"<ja>"
	"ja" L0 J crd CLB CLB CLB CLBC @J #6->7
"<koristad>"
	"korista" Ld V main indic pres ps2 sg ps af <FinV> <NGP-P> @FMV #7->3
"<?>"
	"?" Z Int CLB #8->8
"</s>"
```

Mõnel puhul võib küll tunduda, et üte sarnaneb vägagi järellisandiga ja võiks olla märgendatud pigem subjektile alluvaks. Näites 28 on üks niisugune juhtum toodud: sõnale *Tegija* on siin antud ütte märgend @VOC ja märgitud tema ülemuseks öeldis *oled*. Ütte ja järellisandi vahe tuleb sellest, et järellisand käändub koos oma ülemusega, üte jääb nimetavasse käändesse (EKK: 434). Kuna näites 28 on aga on subjekt *Sina* ka nimetavas käändes, siis pole võimalik öelda, kummaga tegu on: praegu toodud märgenduse asemel võiks sõnal *Tegija* olla ka nimisõnalise järeltäiendi märgend @<NN, sel juhul oleks vaja ülemuseks määrata subjekt *Sina*. Teiste käänete puhul sellist mitmeti tõlgendamise 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.

(28)

```
"<speaker=Miik>"
"<s>"

"<Sina>"
```

```
        "sina" L0 P pers ps2 sg nom cap @SUBJ #1->3
"<Tegija>"
        "Tegija" L0 S prop sg nom cap @VOC #2->3
"<oled>"
        "ole" Ld V main indic pres ps2 sg ps af <FinV> <Intr> <0> @FMV #3->0
"<alati>"
        "alati" L0 D @ADVL #4->3
"<minu>"
        "mina" L0 P pers ps1 sg gen @P> #5->6
"<vastu>"
        "vastu" L0 K post <gen> @ADVL #6->3
```

Kui öeldis puudub, nagu jututubades sageli juhtub, siis märgiti ütte ülemuseks samuti (osa)lause kõige kõrgem ülemus, näite 29 puhul sõna *Inetu-tujutu* ülemuseks subjekt *nimi*.

(29)

```
"<speaker=Mitte_andekas_isane_kass>"
"<s>"

"<Inetu-tujutu>"
        "Inetu-tujutu" L0 S prop sg nom cap @VOC #1->7
"<kallis>"
        "kallis" L0 A pos sg nom @AN> #2->3
"<naiksa>"
        "naiska" L0 S com sg nom @<NN #3->1
"<miks>"
        "miks" L0 D @ADVL @<DN #4->7
"<taas>"
        "taas" L0 D @ADVL #5->7
"<selline>"
        "selline" L0 P dem sg nom @NN> #6->7
"<nimi>"
        "nimi" L0 S com sg nom <0> @SUBJ #7->0
"<?>"
        "?" Z Int CLB #8->8
"</s>"
```

Juhtudel, kus lausung koosnes partiklist ja üttest (näide 30), otsustasin määrata partikli ütte ülemuseks, mitte vastupidi, kuna lausungi sisulist tähendust – nii palju kui seda taolistes lausungites üldse on – annab edasi partikkel. Esialgses käsitsi märgendatud korpuses olid need juhud lahendatud küll vastupidi, kuid põhjus oli selles, et üttele anti subjekti märgend ning öeldise puudumisel on lause kõrgeima ülemuse esimene

kandidaat subjekt. Nüüd, kus üttel on eraldi märgend, ei ole aga enam põhjust teda partikli ülemuseks pidada.

(30)

```
"<speaker=raux>"
"<s>"

"<hendric>"
	"hendric" L0 S prop sg nom <0> @VOC #1->2
"<ommik>"
	"omik" L0 B @B #2->0
"</s>"
```

Seega sai üte ise lause kõrgeimaks ülemuseks vaid siis, kui ta esineb üksi või koos kirjavahemärkide või emotikonidega (vt ka osa 6.3) või teda laiendavad täiendid (nagu näites 29 üttele *Inetu-tujutu* allub järeltäiend *naiksa*, millele omakorda allub eestäiend *kallis*). Samuti kehtivad koordineeritud ütete puhul tavapärased koordinatsioonireeglid, mille kohta on toodud näide 31: 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*.

(31)

```
"<speaker=SunMan>"
"<s>"

"<tehke>"
	"tege" Lke V main imper pres ps2 pl ps af <FinV> <NGP-P> <0> @FMV #1->0
"<siis>"
	"siis" L0 D @ADVL #2->1
"<Caspar>"
	"Caspar" L0 S prop sg nom cap @VOC #3->1
"<ja>"
	"ja" L0 J crd @J #4->5
"<operaator-k6ps>"
	"operaator-k6ps" L0 S prop sg nom @VOC #5->3
"</s>"
```

6.2. Partikkel

Esialgses katses (vt peatükk 4) oli partikli ülemuseks määratud alati lause kõrgeim ülemus – üldiselt predikaat, erandjuhtudel subjekt. Laias laastus on lähenemine õige, kuna sarnaselt üttega on ka partikkel enamasti muu lausega grammatiliselt sidumata ega kuulu ühegi lauseliikme juurde. Sellele leiduvad siiski mõned erandid, nt näites 32 on partikkel *vot* subjekti rõhutaja, mitte ei kuulu üleüldiselt kogu lause juurde, seega lisati reeglistikku vastav reegel.

(32)

```
"<speaker=konn>"
"<s>"

"<vot>"
      "vot" L0 B @B #1->2
"<see>"
      "see" L0 P dem sg nom @SUBJ #2->4
"<mind>"
      "mina" Ld P pers ps1 sg part @OBJ #3->4
"<vihastabki>"
      "vihasta" Lbki V main indic pres ps3 sg ps af <FinV> <0> @FMV #4->0
"</s>"
```

Nagu mainitud eelmises alapeatükis, määrati vaid üttest ja partiklist koosnevas lauses ülemuseks partikkel. Lisaks üttele võib aga mõnedele partiklitele alluda ka kaassõnafraas: näites 33 toodud lausungi puhul on partikkel *ait2h* lausungi keskseks osaks, kaassõnafraas *pikema kommentaari eest* laiendab teda ja oleks põhimõtteliselt võimalik lausungist ka ära jätta. Seega on näites 33 märgitud lause kõrgeimaks ülemuseks partikkel *ait2h* ja kaassõna *eest* allub partiklile.

(33)

```
"<speaker=kiti>"
"<s>"

"<aaa>"
	"aa" L0 B @B #1->3
"<...>"
	"..." Z Ell CLB #2->2
"<ok>"
	"ok" L0 B @B #3->5
"<,>"
	"," Z Com #4->4
"<ait2h>"
	"aitäh" L0 B @B #5->0
"<pikema>"
	"pikem" L0 A comp sg gen @AN> #6->7
"<kommentaari>"
	"kommentaar" L0 S com sg gen @P> #7->8
"<eest>"
	"eest" L0 K post <gen> <0> @ADVL #8->5
"</s>"
```

Vaid partiklist (või partiklist ja kirjavahemärkidest/emotikonidest) koosnevates lausetes määrati partikkel loomulikult samuti lause kõrgeimaks ülemuseks, mitmest partiklist koosnevate puhul viimane partikkel (vt näide 34).

(34)

```
"<speaker=+icebaby>"
"<s>"

"<no>"
	"no" L0 B @B #1->2
"<kle>"
	"kle" L0 B @B #2->0
"</s>"
```

6.3. Emotikon

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

6.4. Predikatiiv

Predikatiivide puhul tuli juurde lisada võimalus, et predikatiiv ise ongi lause kõige kõrgem ülemus – kas siis esinedes täiesti üksinda või pikema lausungi põhilise elemendina nagu näites 35, 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.

(35)

```
"<speaker=SunMan>"
"<s>"

"<üsna>"
	"üsna" L0 D <0> @ADVL #1->2
"<head>"
	"hea" Ld A pos pl nom @PRD #2->0
"<juba>"
	"juba" L0 D @ADVL #3->2
"<tegelt>"
	"tegelikult" L0 D @ADVL #4->2
"</s>"
```

Samuti oli vaja luua reegel juhtudeks, kui predikatiiv esineb öeldiseta lauses ja tuleks märkida subjektile alluvaks (näide 36).

(36)

```
"<speaker=raksu>"
"<s>"

"<ma>"
      "mina" L0 P pers ps1 sg nom <0> @SUBJ #1->0
"<täna>"
      "täna" L0 D @ADVL #2->1
"<varane>"
      "varane" L0 A pos sg nom @PRD #3->1
"</s>"
```

6.5. Adverbiaal

Määrused, millele konkreetset ülemust lausest ei leita, pannakse osalause kõrgeima ülemuse ehk üldjuhul verbi alluvateks. Püüdsin sellest tekkivate vigade hulka vähendada suhteliselt konkreetsete reeglitega, näiteks sõna *kui* määrsõnana ei kuulu tegelikult kunagi verbi juurde, seega lisasin reegli, mis määraks tema ülemuseks adverbiaali, omadussõnalise täiendi või öeldistäite, kui mõni neist järgneb.

Vaja oli luua ka reegel, mis määraks adverbiaalid predikatiivile alluma, juhul kui viimane on lause kõige kõrgemaks ülemuseks määratud (vt näide 35 *juba* ja *tegelt*).

6.6. Muu

Vähesel määral lisati sõltuvusreegleid ka teiste lauseliikmete sõltuvuste määramise parandamiseks. Mitmel puhul (nt objekti, adverbiaali, alistava sidendi ülemuse määramine, vt näites 37 sõna *mida*, näites 38 sõna *kas*) oli vaja täiendada reegleid juhtumite osas, mil 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 ei jää aga nüüd muud üle, kui märkida ta sõna enda külge.

(37)

```
"<speaker=tibike>"
"<s>"

"<mhmh>"
	"mhmh" L0 B @B #1->5
"<,>"
	"," Z Com #2->2
"<laps>"
	"laps" L0 S com sg nom @SUBJ #3->5
"<peab>"
	"pida" Lb V mod indic pres ps3 sg ps af <FinV> <NGP-P> <mdainf> @FCV #4->5
"<saama>"
	"saa" Lma V main sup ps ill <NGP-P> <0> @IMV #5->0
"<mida>"
	"mis" Lda P inter rel sg part CLB @OBJ #6->8
"<laps>"
	"laps" L0 S com sg nom @SUBJ #7->8
"<tahab>"
	"taht" Lb V main indic pres ps3 sg ps af <FinV> <Part-P> <InfP> @FMV #8->5
"<:)>"
	":)" E @E #9->9
"</s>"
```

(38)

```
"<speaker=karvakera>"
"<s>"

"<tegelt>"
	"tegelikult" L0 D @ADVL #1->3
"<ei>"
	"ei" L0 V aux neg @NEG #2->3
"<tegi>"
	"tead" Lgi V main indic pres ps neg <FinV> <NGP-P> <InfP> <0> @FMV #3->0
"<kas>"
	"kas" L0 D CLB @ADVL #4->9
"<selline>"
	"selline" L0 P dem sg nom @NN> #5->6
"<asi>"
	"asi" L0 S com sg nom @SUBJ #6->9
"<ültse>"
	"üldse" L0 D @<DN @ADVL #7->9
"<olemas>"
	"ole" Lmas V main sup ps in <Intr> @ADVL #8->9
"<ongi>"
	"ole" Lgi V main indic pres ps3 sg ps af <FinV> <Intr> @FMV #9->3
"</s>"
```

6.7. Tulemused kohandatud sõltuvussüntaksi reeglistikuga

Sõltuvussüntaktiliste reeglite kohandamise tulemusel tõusis UAS ehk sõltuvusmärgendite saagis testkorpusel peatüki alguses mainitud 75,03%-lt 84,60%-ni ning LAS 72,21%-lt 82,19%-le. Seega on mõlemad näitajad pisut kõrgemad Muischneki jt (2014a) leitud eesti kirjakeele vastavatest näitajatest UAS 83,4% ning LAS 80,3%.

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

Sõltuvusmärgendite vigade arvud ja saagised enne ja pärast reeglistiku kohandamist on esitatud tabelis 6. Nagu näha, on enamiku süntaktiliste funktsioonide puhul saagised tõusnud – langus on toimunud vaid omadus- ja määrsõnaliste täiendite ning kaassõna laiendite puhul, mida aga esineb korpuses kokku suhteliselt vähe.

Kõige enam aitas saagise tõusule kaasa partiklitele põhjalikumate ülemuse määramise reeglite kirjutamine: partiklite sõltuvusmärgendite saagis kasvas testkorpuses 44,71%-lt 88,15%-ni. Samuti tõusis keskmisest rohkem – nimelt 10,71 protsendipunkti võrra ehk 72,96%-lt 83,67%-le – predikatiivide sõltuvusmärgendite saagis. Objektide, adverbiaalide ning öeldiste sõltuvusmärgendite saagised tõusid kõik 4–5 protsendipunkti võrra, samuti subjektide ja ütete oma, kui neid algtulemusega võrdlemise huvides koos vaadelda – nii oleks nende lõppsaagiseks 84,85%. Sidendite sõltuvusmärgendite saagis kasvas 2,95 protsendipunkti ning täienditel 1,44 protsendipunkti võrra.

Tabel 6. **Sõltuvusmärgendite vigade arvud ja saagised testkorpuses**

Süntaktiline funktsioon	**Esinemisi korpuses**	**Kohandatud reeglistik**			**Kohandamata reeglistik**		
		Vigu	**Saagis (%)**	**% kõigist vigadest**	**Vigu**	**Saagis (%)**	**% kõigist vigadest**
@IMV	117	12	89,74	1,40	17	85,47	1,22
@FMV	810	88	88,27	11,07	133	83,58	9,57
@ICV	-	-	-	-	-	-	-
@FCV	89	7	92,13	0,82	7	92,13	0,50
@NEG	96	1	98,96	0,12	1	98,96	0,07
@SUBJ	695	90	87,05	10,49	208	80,54	14,96
@VOC	374	72	80,75	8,39			
@OBJ	278	47	83,09	5,48	58	79,14	4,17
@PRD	196	32	83,67	3,73	53	72,96	3,81
@ADVL	1434	284	80,20	33,10	347	75,80	24,96
@NN>	208	27	87,02	3,15	36	82,69	2,59
@<NN	26	17	34,62	1,98	19	26,92	1,37
@AN>	102	8	92,16	0,93	7	93,14	0,50
@<AN	4	4	0,00	0,47	2	50,00	0,14
@DN>	47	25	46,81	2,91	24	48,94	1,73
@<DN	24	17	29,17	1,98	16	33,33	1,15
@INFN>	-	-	-	-	-	-	-
@<INFN	1	1	0,00	0,12	1	0,00	0,07
@KN>	-	-	-	-	-	-	-
@<KN	4	2	50,00	0,23	2	50,00	0,14
@P>	40	5	87,50	0,58	4	90,00	0,29
@<P	13	2	84,62	0,23	-	100,00	-
@Q>	1	1	0,00	0,12	1	0,00	0,07

@<Q	31	3	90,32	0,35	4	87,10	0,29
@Vpart	49	1	97,96	0,12	1	97,96	0,07
@J	204	38	81,37	4,43	44	78,43	3,17
@B	624	74	88,14	8,62	345	44,71	24,82
@E	218	-	-	-	60	72,48	4,32
KOKKU	**5685**	**858**	**84,60**	**100**	**1390**	**75,03**	**100**

Ka sõltuvusmärgendite vigade osakaalusid on võimalik võrrelda Tiina Puolakaineni (2013) ajalehetekstidel leitud andmetega. Kõige enam vigu (33,10% kõigist vigadest) tekitab jututubades ka pärast reeglite kohandamist adverbiaalile vale ülemuse määramine, sama kehtib kirjakeeles, kus adverbiaali ülemuse määramise vead moodustavad 30,7% kõigist vigadest. Jututubades on järgmisteks suurematekseks veakohtadeks finiitse öeldise (11,07% kõigist vigadest) ja subjekti (10,49% kõigist vigadest) ülemuse määramine. Finiitse öeldise ja subjekti ülemustest tekib ka kirjakeeles suhteliselt palju vigu (vastavalt 12,2% ja 8,5% kõigist vigadest), ent pisut rohkem tekib nimisõnaliste eestäiendite ülemuste määramisel (12,4% kõigist vigadest). (Puolakainen 2013) Internetikeeles seevastu tuleneb nimisõnalistest eestäienditest vaid 3,15% ülemuse määramise vigadest. Seega, kui välja arvata nimisõnalised eestäiendid, on sagedasemate sõltuvusvigade jaotus internetikeeles ja kirjakeeles suhteliselt sarnased.

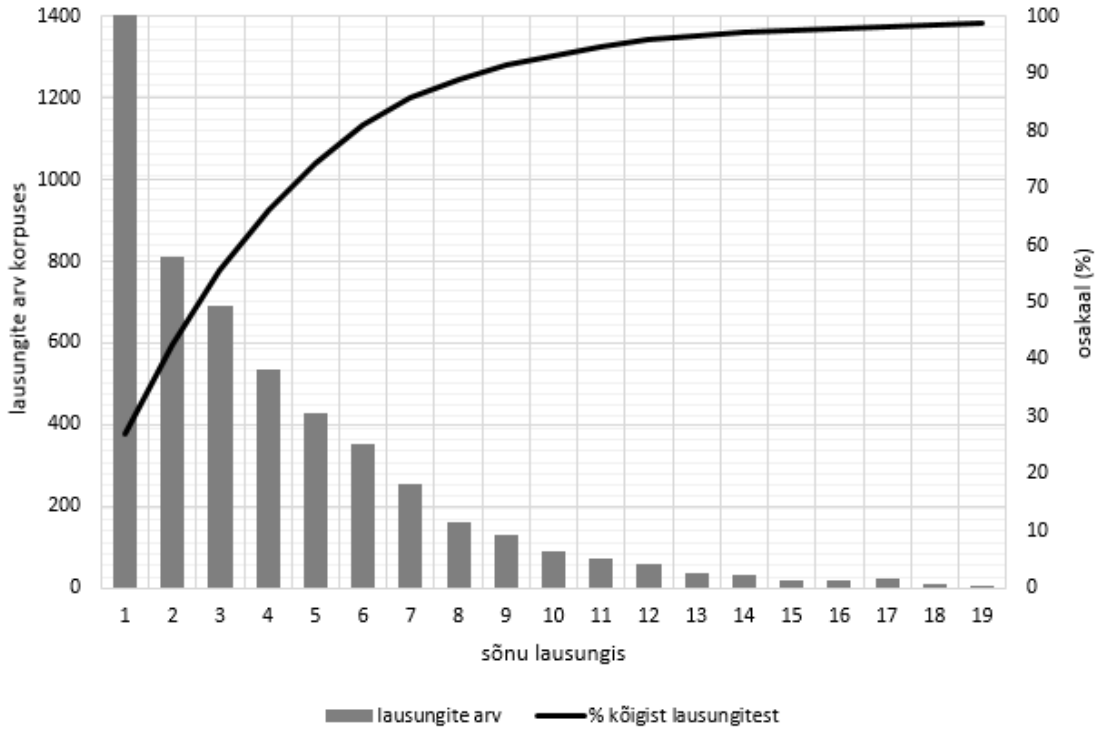
7. Internetikeele süntaksi erijooned

Nagu eelnevates peatükkides kirjeldatu põhjal võib järeldada, ei ole internetikeel süntaktiliselt päris samasugune nagu kirjakeel. Seega annangi käesolevas peatükis kokkuvõtliku ülevaate, mille poolest internetikeele süntaks erineb.

7.1. Lausungi pikkus

Internetikeele lausungid on lühikesed: keskmiseks lausungi pikkuseks (lausungi pikkuse arvutamisel on hulka arvatud ka kirjavahemärgid ja emotikonid) jututubade korpuses on 4 sõna. Eesti keele sõltuvuspuude pangas seevastu, mis sisaldab ajakirjandus-, ilukirjandus- ja teadustekste kokku 430 350 sõna mahus, on kokku 30 299 lauset ehk lause keskmiseks pikkuseks 14 sõna (EKT).

Joonisel 2 on toodud lausungite pikkuse jaotus 5205 lausungi suuruses käsitsi kontrollitud süntaktilise analüüsiga jututubade korpuses. Jooniselt on näha, et 66% lausungeid on 1–4 sõna pikkused ning 90% kuni 9 sõna pikkused. Siiski on korpuses ka 1% lauseid, mille pikkus on 20 või enam sõna. Võrdluseks vaatasin paari ilu- ja ajakirjandusteksti lausete pikkust: vaid 4–6% lauseid olid kuni nelja sõna pikkused ning 25–40% lauseid kuni üheksa sõna pikkused.



Joonis 2. **Lausungite pikkuste jaotus jututubades.** Joonega on tähistatud kumulatiivne protsent kõigist lausungitest korpuses.

7.2. Kirjavahemärgistus

Kirjavahemärgistust kasutatakse vähe ja juhuslikult: jututubade korpuses leidub üks kirjavahemärk iga 10 sõna kohta, samas kui sõltuvuspuude pangas olevates ilu-, aja- ja teaduskirjandustekstides on iga 5 sõna kohta üks kirjavahemärk (EKT). 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 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 (vt näide 39: pandud on kaks koma, ent reeglite järgi oleks vaja komasid ka 1. ja 2. ning 19. ja 20.

sõna vahele).

(39)

```
"<speaker=+tibu>"
"<s>"

"<tundub>"
	"tundu" Lb V main indic pres ps3 sg ps af <FinV> <Intr> <0> @FMV #1->0
"<et>"
	"et" L0 J sub @J #2->9
"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ #3->9
"<pean>"
	"pida" Ln V mod indic pres ps1 sg ps af <FinV> <NGP-P> <mdainf> @FCV #4->9
"<oma>"
	"oma" L0 P pos det refl sg gen @NN> #5->7
"<raske>"
	"raske" L0 A pos sg gen @AN> #6->7
"<kõhuga>"
	"kõht" Lga S com sg kom @ADVL #7->9
"<edasi>"
	"edasi" L0 D @ADVL #8->9
"<elama>"
	"ela" Lma V main sup ps ill <Intr> <In> <Ad> @IMV #9->1
"<,>"
	"," Z Com CLB #10->10
"<keegi>"
	"keegi" L0 P indef sg nom @SUBJ #11->13
"<ei>"
	"ei" L0 V aux neg @NEG #12->13
"<taha>"
	"taht" L0 V main indic pres ps neg <FinV> <Part-P> <InfP> <mdainf> @FMV #13->9
"<seda>"
	"see" Lda P dem sg part @OBJ #14->17
"<vist>"
	"vist" L0 D @ADVL #15->17
"<endale>"
	"ise" Lle P pos det refl sg all @ADVL #16->17
"<osta>"
	"ost" La V main inf <NGP-P> <El> <Abl> @OBJ #17->13
"<,>"
	"," Z Com CLB #18->18
"<kurb>"
	"kurb" L0 A pos sg nom @PRD #19->1
"<a>"
	"aga" L0 J crd @J #20->21
"<lootsin>"
	"loot" Lsin V main indic impf ps1 sg ps af <FinV> <NGP-P> <InfP> <All> @FMV #21->19
"<teie>"
	"sina" L0 P pers ps2 pl gen @P> #22->23
"<peale.>"
	"peale" L0 K post <gen> @ADVL #23->21
"</s>"
```

Erinevusena kirjakeelest võib välja tuua ka selle, et näitamaks, kellele lausung suunatud on, kasutatakse koolonit ning @-märki. Koolonit on kasutatud näites 40, näitamaks, et lausung on suunatud jututoa kasutajale *konn*, @-märk näites 41 enne sõna *Adi* annab teada, et tegu on sellenimelisele kasutajale suunatud tekstiga. Näidetes on toodud märkide tüüpiline kasutus: koolonit kasutatakse reeglina siis, kui adressaat on pandud lausungi algusesse, @-märki siis, kui adressaat on lausungi lõpus.

(40)

```
"<speaker=Markos>"
"<s>"

"<konn>"
	"konn" L0 S <0> prop sg nom @VOC #1->9
"<:>"
	":" Z Col CLBC CLB #2->2
"<kui>"
	"kui" L0 J sub @J #3->6
"<seda>"
	"see" Lda P dem sg part @OBJ #4->6
"<ei>"
	"ei" L0 V aux neg @NEG #5->6
"<kasuta>"
	"kasuta" L0 V main indic pres ps neg <FinV> <Part-P> @FMV #6->9
"<,>"
	"," Z Com CLB #7->7
"<siis>"
	"siis" L0 D @ADVL #8->9
"<pole>"
	"ole" L0 V main indic pres ps neg <FinV> <Intr> <mdainf> <0> @FMV #9->0
"<mõtet>"
	"mõte" Lt S com sg part @PRD #10->9
"<uluda>"
	"ulu" Lda V main inf <Intr> @SUBJ #11->9
"<ka>"
	"ka" L0 D @ADVL #12->11
"<,>"
	"," Z Com CLB #13->13
"<et>"
	"et" L0 J sub @J #14->16
"<aeglane>"
	"aeglane" L0 A pos sg nom @PRD #15->16
"<on>"
	"ole" L0 V main indic pres ps3 sg ps af <FinV> <Intr> @FMV #16->9
"</s>"
```

(41)

```
"<speaker=+Sigitis>"
"<s>"

"<ma>"
      "mina" L0 P pers ps1 sg nom @SUBJ #1->2
"<tean>"
      "tead" Ln V main indic pres ps1 sg ps af <FinV> <NGP-P> <InfP> <0> @FMV #2->0
"<@>"
      "@" Z Gnr #3->3
"<Adi>"
      "Adi" L0 S prop sg nom cap @VOC #4->2
"</s>"
```

Lisaks tavapärastele kirjavahemärkidele kasutatakse internetikeeles emotikone. Nendel on küll vestluses enam funktsioone ja tähendust kui kirjavahemärkidel, kuid süntaksi seisukohast sarnanevad nad siiski enam kirjavahemärkide kui mõne muu lauseliikmega. Enamasti paikneb emotikon kas lausungi lõpus (70% emotikonidest käsitsi kontrollitud süntaktilise analüüsiga korpuses) või moodustab üksinda lausungi (24%).

7.3. Mittetäielik lausung

Suur osa internetikeele lausungitest ei vasta kirjakeele mõistes tüüpilisele lausele, mis peaks sisaldama finiitset verbivormi ja vähemalt ühte käänd- või määrsõna (EKK: 345). Peatükis 3.1 on toodud näited 5 ja 6 selle kohta, kuidas üks jututoa lausung võib sisaldada mitut kirjakeele lauset ning üks kirjakeele mõistes lause võib olla jagunenud mitme lausungi vahel.

Nagu näidatud osas 7.1, on internetikeele lausungid kirjakeele lausetega võrreldes lühemad ning esineb palju väga lühikesi lausungeid. 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 just tervitamisel, kuna jututubadesse lisandub pidevalt uusi vestlejaid, keda tervitatakse ja kes kohalolijaid vastu tervitama hakkavad.

Teiseks väljajätteliste ja väga lühikeste lausungite põhjustajaks on see, et sageli

sarnaneb suhtlus suulise dialoogiga. Hennoste (2012: 37) on MSN-i vestluste kohta märkinud, et nendes vahetatakse sarnaselt suulise kõnega spontaanselt lühikesi kõnevoore. Vaatamata sellele, et jututoas viibib enamasti rohkem kui kaks kasutajat korraga, on sealgi näha lühikestest voorudest koosnevaid dialooge, milles ühest lausungist puuduv info selgub teda ümbritsevast kontekstist. Selle kohta on toodud näide 42: näiteks lausungist *kah hästi säilinud siis* ei selgu, kes või mis on hästi säilinud, vaid seda saab järeldada dialoogi eelmise lausungi põhjal.

(42)

```
"<speaker=raksu>"
"<s>"

"<minu>"
	"mina" L0 P pers ps1 sg gen @P> #1->2
"<käest>"
	"käest" L0 K post <gen> @ADVL #2->3
"<kysitaxe>"
	"küsi" Ltakse V main indic pres imps af <FinV> <NGP-P> <Abl> <0> @FMV #3->0
"<ka>"
	"ka" L0 D @ADVL #4->5
"<veel>"
	"veel" L0 D @ADVL #5->3
"<vahest>"
	"vahest" L0 D @ADVL #6->3
"<papreid>"
	"paber" Lid S com pl part @OBJ #7->3
"</s>"

"<speaker=Y2K>"
"<s>"

"<kah>"
	"kah" L0 D @ADVL #1->2
"<hästi>"
	"hästi" L0 D @ADVL #2->3
"<säilinud>"
	"säili" Lnud V main partic past ps <0> <Intr> @IMV #3->0
"<siis>"
	"siis" L0 D CLB @ADVL #4->3
"</s>"

"<speaker=raksu>"
"<s>"

"<yeap>"
	"yeap" L0 B <0> @B #1->0
"</s>"
```

```
"<speaker=Y2K>"
"<s>"

"<norm>"
	"norm" L0 A pos sg nom <0> @PRD #1->0
"</s>"

"<speaker=raksu>"
"<s>"

"<hea>"
	"hea" L0 A pos sg nom @PRD #1->4
"<nooremaid>"
	"noorem" Lid A comp pl part @AN> #2->3
"<naisi>"
	"naine" L0 S com pl part @OBJ #3->4
"<sebida>"
	"sebi" Lda V main inf <0> @SUBJ #4->0
"</s>"

"<speaker=raksu>"
"<s>"

"<:P>"
	":P" E @E #1->1
"</s>"
```

Internetikeele lausungist võib olla välja jäetud lisaks muudele võimalikele lauseliikmetele ka öeldis – seda on Hennoste (2013) nimetanud „eesti netikeele kõige radikaalsemaks erijooneks“. Niisugune väljajätt toimub enamasti olema-verbiga, kuid võib esineda ka liikumisverbide puhul nagu näites 43.

(43)

```
"<speaker=dodge>"
"<s>"

"<neitsik>"
	"neitsik" L0 S <0> prop sg nom @VOC #1->3
"<...>"
	"..." Z Ell CLB #2->2
"<ma>"
	"mina" L0 P pers ps1 sg nom <0> @SUBJ #3->0
"<homme>"
	"homme" L0 D @DN> #4->5
"<õhtul>"
	"õhtu" Ll S com sg ad @ADVL #5->3
"<vist>"
```

```
	"vist" L0 D @ADVL #6->3
"<tallinna>"
	"Tallinn" L0 S prop sg adit @ADVL #7->3
"<...>"
	"..." Z Ell CLB #8->8
"<muga>"
	"mina" Lga P pers ps1 sg kom @ADVL #9->10
"<kohtama>"
	"kohta" Lma V main sup ps ill <Part> @ADVL #10->11
"<tuled>"
	"tule" Ld V main indic pres ps2 sg ps af <FinV> <Intr> <mdainf> @FMV #11->3
"<?>"
	"?" Z Int CLB #12->12
"</s>"
```

7.4. Internetikeelele iseloomulikud elemendid lausungis

Internetikeele lausungile iseloomulike elementidena võib välja tuua emotikonid, partiklid ning ütte. Kõik need käituvad üldlaienditena, olemata lausungiga grammatiliselt seotud.

Emotikone nimetab Hennoste (2013) „kõige tuntumaks netikeelsuseks“, kirjakeeles neid reeglina ei esine ning suulises kõnes muidugi ka mitte. 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 44 ei lõppeks emotikoniga *:)*, siis ei pruugiks lausungi lugeja aru saada, et ütleja rõõmus või rahul on.

(44)

```
"<speaker=kitu_kokkab>"
"<s>"

"<Bloodhound>"
	"Bloodhound" L0 S prop sg nom cap @VOC #1->4
"<ma>"
	"mina" L0 P pers ps1 sg nom @SUBJ #2->4
"<alles>"
	"alles" L0 D @ADVL #3->4
"<jõudsin>"
	"jõud" Lsin V main indic impf ps1 sg ps af <FinV> <InfP> <All> <Ill> <Ter> @FMV #4->0
```

```
"<koju>"
	"kodu" L0 S com sg adit @ADVL #5->4
"<ja>"
	"ja" L0 J crd CLB CLBC @J #6->7
"<söön>"
	"söö" Ln V main indic pres ps1 sg ps af <FinV> <NGP-P> @FMV #7->4
"<ja>"
	"ja" L0 J crd CLBC @J #8->10
"<siis>"
	"siis" L0 D CLB @ADVL #9->10
"<magama>"
	"maga" Lma V main sup ps ill @ADVL #10->7
"<:)>"
	":)" E @E #11->11
"</s>"
```

Partiklid on internetikeele üheks ühisjooneks suulise keelega. Osas 7.1 mainiti, et väga lühikesed lausungid on sageli moodustunud partikli(te)st, kuid see ei tähenda, et partiklid ainult lühikeste lausungite juurde kuuluksid. Lausungi sisu seisukohast mittevajalike partiklite olemasolu (vt näide 45: *noh* lausungi sisu kuidagi ei muuda) kinnitab, et lausungeid ei genereerita siiski niivõrd ratsionaalselt ja kokkuhoidlikult (aja, tähemärkide osas), nagu võiks arvata paljude väljajättude ja lauselõpumärkide mittekasutamise põhjal.

(45)

```
"<speaker=KIISU17>"
"<s>"

"<noh>"
	"noh" L0 B @B #1->5
"<nunnu>"
	"nunnu" L0 A pos sg nom @AN> #2->3
"<sibelgas>"
	"sipelgas" L0 S prop sg nom <0> @VOC #3->5
"<kuidas>"
	"kuidas" L0 D CLB @ADVL #4->5
"<jääb>"
	"jää" Lb V main indic pres ps3 sg ps af <FinV> <Intr> <0> @FMV #5->0
"</s>"
```

Ütte kasutamine on samuti peamiselt internetikeele ja suulise keele ühine element, 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 kuidagi teisiti märku anda, missugusele kasutajale lausung suunatud on.

7.5. Süntaktiliste funktsioonide osakaalude võrdlus kirjakeelega

Tabelis 7 on toodud süntaktiliste funktsioonide osakaalud internetikeeles ning võrdluseks sõltuvuspangas asuvates ilu-, aja- ja teaduskirjanduse tekstides. Kuna korpuste suurused on väga erinevad – internetikeel 19 809 sõna, ilukirjandus 54 817 sõna, ajakirjandus 225 517 sõna ning teaduskirjandus 80 891 sõna – , siis on andmed esitatud vaid protsendina vastava korpuse kogumahust. Partikli märgendit kirjakeeles ei kasutata, selle asemel on kasutusel interjektsiooni märgend @I.

Nagu arvata võibki, on internetikeeles süntaktiliste funktsioonide jaotus toodud kirjakeele variantidest sarnaseim ilukirjandusega. Suurim erinevus tabelis toodud keelevariantide vahel on täiendite osakaaludes. Jututubades moodustavad kõik täiendiliigid kokku 7,39% korpuse mahust, ilukirjanduses 13,52%, ajakirjanduses 25,70% ning teaduskirjanduses lausa 34,05%. Samuti on jututubade korpuses kaassõna ja kvantori laiendite osakaalud pisut väiksemad kui teistes korpustes.

Nagu juba eelmises alapeatükis välja toodud, esineb internetikeeles lauseliikmeid, mida kirjakeeles on väga vähe või üldse mitte. Jututubade korpusest moodustavad partiklid ja emotikonid vastavalt 10,77% ja 4,64%. Ilukirjanduse korpusest moodustavad küll 0,05% interjektsioonid, kuid võrreldes jututubade partiklite osakaaluga on see üliväike isa. Subjektide ja ütete osakaal kokku (17,17%) jututubades on tunduvalt suurem kui teistes korpustes, kuna kirjakeeles esineb ütteid väga vähe. Lisaks on jututubades pisut suurem osakaal kui teistes korpustes ka öeldistäidetel ning verbi eitustel.

Tabel 7. **Süntaktiliste funktsioonide osakaalud (%) internetikeeles ning ilu-, aja- ja teaduskirjanduse tekstides**

Funktsioon	Jututoad	Ilukirjandus	Ajakirjandus	Teaduskirjandus
@IMV	1,75	2,83	2,68	2,68
@FMV	14,56	15,73	11,26	7,73
@ICV	-	0,03	0,04	0,03
@FCV	1,47	2,14	2,02	2,18
@NEG	2,03	1,63	1,07	0,63
@SUBJ	12,27	13,87	12,04	9,87
@VOC	4,91			
@OBJ	4,96	7,22	7,04	6,00
@PRD	3,36	2,52	2,13	1,79
@ADVL	23,47	26,42	23,66	20,91
@NN>	3,58	6,20	13,82	16,30
@<NN	0,56	0,98	2,13	4,31
@AN>	1,84	5,15	7,78	11,30
@<AN	0,11	0,11	0,07	0,04
@DN>	0,65	0,74	1,38	1,40
@<DN	0,41	0,05	0,06	0,04
@INFN>	-	-	-	-
@<INFN	0,05	0,12	0,21	0,15
@KN>	0,01	0,01	0,04	0,03
@<KN	0,06	0,16	0,21	0,48
@P>	0,70	1,99	2,04	2,29
@<P	0,18	0,53	0,50	0,61
@Q>	0,02	0,02	0,01	-
@<Q	0,56	0,79	1,82	1,42

@Vpart	1,02	2,01	1,18	0,48
@J	4,28	8,49	6,75	6,26
@B/@I	10,77	0,05	-	-
@E	4,64	-	-	-
@??	1,65	0,21	0,03	3,06

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

Kokkuvõte

Magistritöös käsitleti eesti internetikeele automaatset süntaktilist analüüsi. Töö esimesed kaks peatükki olid teoreetilised ning nendes keskenduti tausta kirjeldamisele ja ülevaate andmisele teistest sarnasel teemal ilmunud töödest: esimene peatükk käsitles internetikeelt, selle tuntumaid erinevusi kirjakeelest ja erinemise põhjusi; teine peatükk käsitles süntaktilist analüüsi, töös kasutatud kitsenduste grammatika formalismi ning süntaktilise analüüsi hindamise mõõdikuid.

Töö kolmas ja neljas peatükk andsid ülevaate kirjakeele kitsenduste grammatika reeglistiku rakendamisest käsitsi kontrollitud morfoloogilise analüüsiga jututubade tekstidele, kolmandas peatükis oli vaatluse all pindsüntaktiline ja neljandas peatükis sõltuvussüntaktiline analüüs. Viiendas ja kuuendas peatükis kirjeldati vastavalt pindsüntaktilise ja sõltuvussüntaktilise analüüsi reeglistiku kohandamist internetikeelele ning esitati kohandatud reeglitega saavutatud tulemused.

Viimases, seitsmendas peatükis anti kokkuvõtlik ülevaade internetikeele süntaksi eripäradest, mis eelneva praktilise töö käigus silma olid jäänud, samuti tehti mõningaid kvantitatiivseid võrdlusi kirjakeele süntaksiga.

Kirjakeele reeglistiku rakendamisel internetikeelele tekkis enim pindsüntaktilise analüüsi vigu subjektide (sealhulgas ka ütete), öeldistäidete ning adverbiaalide tuvastamisel, kõige sagedamini jäi mitmeseks tõlgendus adverbiaali ja järeltäiendi vahel. Sõltuvussüntaktilisel analüüsil oli kõige rohkem vigu adverbiaalide, partiklite ja subjektide ülemuste määramisel.

Kitsenduste grammatika reeglite kohandamise tulemusel paranesid nii pind- kui 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.

Internetikeele süntaksi suurimate erinevustena kirjakeelest tulid välja sage partiklite ja ütete kasutamine, lausungite lühidus ja täiendite väike osakaal lauseliikmetest. Samuti on internetikeelele iseloomulikud kirjakeele mõttes mittetäielikud laused, kust lisaks muudele lauseliikmetele võib välja jääda ka verb.

Kuna süntaktiline analüüs on üks automaatse keeletöötluse põhietappidest ning selle korrektsest väljundist on kasu paljude edasiste ülesannete lahendamisel, siis oleks autori hinnangul 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.

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. Vaadatud 19.05.2015.

Buchholz, Sabine, Erwin Marsi 2006. CoNLL-X shared task on Multilingual Dependency Parsing. – Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X). New York City, 149–164. http://ilk.uvt.nl/~emarsi/download/pubs/14964.pdf. Vaadatud 19.05.2015.

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

Dresner, Eli, Susan C. Herring 2010. Functions of the Non-Verbal in CMC: Emoticons and Illocutory Force. – Communication Theory 20, 249–268. http://ella.slis.indiana.edu/~herring/emoticons.pdf. Vaadatud 19.05.2015.

Drouin, Michelle, Claire Davies 2009. R u txting? Is the Use of Text Speak Hurting Your Literacy? – Journal of Literacy Research 41, 46–67. http://jlr.sagepub.com/content/41/1/46.full.pdf+html. Vaadatud 19.05.2015.

Eisenstein, Jacob 2013. What to do about bad language on the internet. – Proceedings of NAACL-HLT 2013. Atlanta, Georgia, 359–369. http://www.aclweb.org/anthology/N/N13/N13-1037.pdf. Vaadatud 19.05.2015.

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

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

Foster, Jennifer 2010. "cba to check the spelling" Investigating Parser Performance on Discussion Forum Posts. – Proceedings of Human Language Technologies: 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, CA, 381–384. http://www.computing.dcu.ie/~jfoster/publications/Foster-NAACL10.pdf. Vaadatud 19.05.2015.

Foster et al. 2011a = Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, Josef van Genabith 2011. #hardtoparse: POS Tagging and Parsing the Twitterverse. – Proceedings of AAAI-11 Workshop on Analysing Microtext, San Francisco, CA. http://www.computing.dcu.ie/~jfoster/publications/Foster-AAAI11.pdf. Vaadatud 19.05.2015.

Foster et al. 2011b = Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, Joseph van Genabith 2011. From News to Comment: Resources and Benchmarks for Parsing the Language of Web 2.0. http://www.computing.dcu.ie/~jfoster/publications/Foster-IJCNLP11.pdf. Vaadatud 19.05.2015.

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

Hennoste, Tiit 2011. Special features of syntax of Estonian used in spontaneous MSN dialogues. – Discourse 2.0 Language & New Media, GURT 2011: Ed. Deborah Tannen, Anna Marie Trester, 73.

Hennoste, Tiit 2012. Enda algatatud eneseparandus eestikeelsetes MSN-i dialoogides. Eesti Rakenduslingvistika Ühingu aastaraamat 8. Tallinn: Eesti Rakenduslingvistika Ühing, 37–54.

Hennoste, Tiit 2013. kuule ma eemale nüüd. Sirp nr 46. http://www.sirp.ee/s1-artiklid/c9-sotsiaalia/2013-12-05-17-05-50. Vaadatud 6.05.2015.

Kaalep, Heiki-Jaan, Kadri Muischnek 2011. Morphological analysis of a non-standard language variety. – NODALIDA 2011 Conference Proceedings. Eds. Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa, 130–137.

http://dspace.utlib.ee/dspace/bitstream/handle/10062/17299/0Kaalep_Muischneek_36.pdf?sequence=1. Vaadatud 19.05.2015

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. Vaadatud 19.05.2015.

Karlsson et al. = Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, Arto Anttila. 1995. Constraint Grammar: a Language Independent System for Parsing Unrestricted Text. Berlin and New York: Mouton de Gruyter.

Khan et al. = Khan, Mohammad, Markus Dickinson, Sandra Kübler 2013. Does Size Matter? Text and Grammar Revision for Parsing Social Media Data. – Proceedings of the Workshop on Language Analysis in Social Media. Atlanta, GA. http://cl.indiana.edu/~md7/papers/khan-et-al13.pdf. Vaadatud 19.05.2015.

Le Roux et al. 2012 = Le Roux, Joseph, Jennifer Foster, Joachim Wagner, Rasul Samad Sadeh Kaljahi, Anton Bryl 2012. DCU-Paris13 Systems for the SANCL 2012 Shared Task. http://doras.dcu.ie/17051/1/LeRoux_dcu12.pdf. Vaadatud 19.05.2015.

Lindström, Liina, Kaili Müürisep 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. Vaadatud 19.05.2015

Muischnek jt 2011 = Muischnek, Kadri, Heiki-Jaan Kaalep, Raul Sirel 2011. Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. – Eesti Rakenduslingvistika Ühingu aastaraamat 7, 111–127.

Muischnek jt 2013 = Muischnek, Kadri, Kaili Müürisep, Tiina Puolakainen 2013. Estonian particle verbs and their syntactic analysis. – Human Language Technologies as a Challenge for Computer Science and Linguistics: 6Th Language & Technology Conference Proceedings. December 7–9, Poznan, Poland. Eds. Zygmunt Vetulani and Hans Uszkoreit. Poznan: Adam Mickiewicz University, 338–342. http://kodu.ut.ee/~kaili/papers/ltc-085-muischnek.pdf.

Vaadatud 19.05.2015.

Muischnek jt 2014a = Muischnek, Kadri, Kaili Müürisep, Tiina Puolakainen 2014. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. – Human Language Technologies – The Baltic Perspective. Eds. A. Utka et al., 111–118. http://ebooks.iospress.nl/volumearticle/38013. Vaadatud 19.05.2015.

Muischnek jt 2014b = Muischnek, Kadri, Kaili Müürisep, Tiina Puolakainen 2014. Ühendverbid eesti keele pindsüntaktilises analüüsis. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 227–240.

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu: Tartu Ülikooli kirjastus.

Müürisep, Kaili, Heli Uibo 2005. Shallow Parsing of Spoken Estonian Using Constraint Grammar. Treebanking for Discourse and Speech. – Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse (eds. Peter Juel Henrichsen and Peter Rossen Skadhauge); Copenhagen Studies in Language 32. Samfundslitteratur, 105–118. http://link.springer.com/chapter/10.1007%2F978-3-642-04235-5_15. Vaadatud 19.05.2015.

Müürisep, Kaili, Helen Nigol 2007. Disfluency Detection and Parsing of Transcribed Speech of Estonian. – Proceedings of 3rd Language & Technology Conference Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznan, Poland: Wydawnictwo Poznanskie, 483–487. http://kodu.ut.ee/~kaili/papers/ltc-114-muurisep.pdf. Vaadatud 19.05.2015.

Oja, Anni 2006. Eesti keel internetis. – Keel ja arvuti. Tartu: Tartu Ülikooli kirjastus, 259–267.

Ooi, Vincent B Y 2002. Aspects of computer-mediated communication for research in corpus linguistics. – New Frontiers of Corpus Research. Eds. P. Peters, P. Collins, A. Smith. Amsterdam-New York, 91–104.

Ooi, Vincent B Y 2009. Computer-mediated Language and Corpus Linguistics. – Corpus Analysis and Variation in Linguistics. Eds. Yuji Kawaguchi, Makoto Minegishi, Jacques Durand. Amsterdam/Philadelphia: John Benjamins

Publishing Company, 103–120.

Petrov, Slav, Ryan McDonald 2012. Overview of the 2012 Shared Rask on Parsing the Web. http://static.googleusercontent.com/external_content/untrusted_dlcp /research.google.com/et//pubs/archive/38278.pdf. Vaadatud 19.05.2015.

Petrov et al. 2010 = Petrov, Slav, Pi-Chuan Chang, Michael Ringgaard, Hiyan Alshawi 2010. Uptraining for Accurate Deterministic Question Parsing. – Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT, Massachusetts, USA, 705–713. http://aclweb.org/anthology//D/D10/D10-1069.pdf. Vaadatud 19.05.2015.

Puolakainen, Tiina 2013. 2013. korpuste automaatse märgendamise väljundi analüüs. Käsikiri.

Roosmaa jt 2001 = Roosmaa, Tiit, Mare Koit, Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Heli Uibo 2001. Eesti keele formaalne grammatika. Tartu: Tartu Ülikooli kirjastus.

Salla, Sigrid 2002. Jututuba kui võrgusuhtlusvorm. Tekstid ja taustad. Tartu: Tartu Ülikooli kirjastus, 128–156.

Seddah et al. 2012 = Seddah, Djamé, Benoit Sagot, Marie Candito, Virginie Mouillero, Vanessa Combet 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. – Proceedings of COLING 2012: Technical Papers, 2441–2458. https://www.aclweb.org/anthology/C/C12/C12-1149.pdf. Vaadatud 19.05.2015

Särg, Dage 2012. Internetikeele automaatse morfoloogilise ühestamise kvaliteedi uuring. Bakalaureusetöö.

Internetimaterjalid:

CG Manual = http://beta.visl.sdu.dk/cg3/single/. Vaadatud 24.04.2015.

EKT = https://www.keeletehnoloogia.ee/et/ekt-projektid/vahendid-teksti-

mitmekihiliseks-margendamiseks-rakendatuna-koondkorpusele/soltuvussuntaktiliselt-analuusitud-korpus. Vaadatud 10.05.2015.

Jututubade korpus = http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/jututoad. Vaadatud 24.04.2015.

Morfosüntaktilised kategooriad = http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=et. Vaadatud 18.05.2015.

Märgendusjuhed = http://kodu.ut.ee/~kaili/papers/syntax.html. Vaadatud 17.05.2015.

Sami Syntactic Tags = http://giellatekno.uit.no/doc/lang/common/docu-sme-syntaxtags.html. Vaadatud 24.04.2015.

SANCL 2012 = https://sites.google.com/site/sancl2012/home. Vaadatud 24.04.2015.

VISL-CG3 = http://beta.visl.sdu.dk/cg3.html. Vaadatud 24.04.2015.

Lisa. Internetikeelele kohandatud kitsenduste grammatika reeglistik

Töö käigus kohandati internetikeele parsimiseks Kaili Müürisepa ja Tiina Puolakaineni väljatöötatud kitsenduste grammatika reeglistik. Internetikeelele kohandatud reeglite failid on kättesaadavad aadressilt http://kodu.ut.ee/~dage/reeglid/.

Kaustas reeglid.rar on kokkupakituna neli reeglitefaili internetikeele süntaktiliseks analüüsiks:

- clo_net.rle: osalausepiirid

- PhVerbs_net.rle: ühendverbid

- pindsyn_net.rle: pindsüntaktiline märgendus

- strukt_net.rle: sõltuvussüntaks

Käesoleva töö autori tehtud muudatused on muust reeglistikust eristatavad selle põhjal, et reegli järel kommentaaris (#-märgi järel) on kirjas "Dage". Kui kirjakeele olemasolevat reeglit internetikeele jaoks vaid pisut muudeti, siis on ka tehtud muudatus reegli järel kommentaaris välja toodud.

Reeglid on VISL-CG3-formaadis, nende rakendamiseks vajalik programm on saadaval http://beta.visl.sdu.dk/download/vislcg3/, juhendid http://beta.visl.sdu.dk/cg3/chunked/.

Syntactic Analysis of Estonian Web Language Using Constraint Grammar

Summary

This master's thesis provides an overview of an attempt to adapt the Estonian constraint grammar rule set for web language. The rule set has been developed by Kaili Müürisep and Tiina Puolakainen for Estonian literary language, and it has previously been adapted for spoken Estonian by Kaili Müürisep and Heli Uibo.

The thesis consists of seven chapters. The first chapter describes web language and its differences from literary language, the second one gives an overview of syntactic analysis and constraint grammar formalism. Chapters three to six present the author's practical work on adapting the rule set. In chapters three and four, the errors that emerge using unadapted rule set for shallow and dependency parsing of web language are discussed. Chapters five and six describe the author's adaptations to the rule set. In the last, seventh chapter, the characteristic traits of the syntax of web language are presented, and some comparisons are made against literary language in terms of sentence length and percentages of syntactic functions.

As a result of adapting the rule set, the results of both shallow and dependency parsing improved. For shallow parsing, the final recall was 93.35% and precision 85.16%. The most error-prone syntactic functions were subjects, predicatives, and adverbials. The final LAS (labeled attachment score) was 82.19% and UAS (unlabeled attachment score) 84.60%. In dependency parsing, the largest number of errors was made in determining the heads of adverbials.

Extensive use of particles and direct addresses, short average sentence length, and small percentage of attributes among the syntactic functions used in text appeared to be the most distinctive features of web language, as well as the large amount of elliptical sentences from which, in addition to other syntactic functions, a predicate can be left out.

As syntactic parsing is one of the basic levels of natural language processing and its correct output is useful in many further tasks, e.g. machine translation or grammar correction, the author of the thesis considers it necessary to continue work on the subject. Taking into account the diverse nature of web language, it might be useful to try data-driven approaches in addition to rule-based parsing.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Dage Särg

(*autori nimi*)

(sünnikuupäev: 06.02.1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Internetikeele süntaktiline analüüs kitsenduste grammatikaga",

(*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, 20.05.2015 (*kuupäev)*