Eesti keele süntaksianalüsaatori märgenditest

Kaili Müürisep
TTÜ Küberneetika Instituut
kaili@phon.ioc.ee

1. Sissejuhatus

Loomuliku keele süntaksianalüsaator on programm, mis saab sisendiks morfoloogiliselt analüüsitud teksti ning väljastab süntaktiliselt analüüsitud teksti. Enamasti esitatakse süntaktiline kirjeldus märgendite abil, st iga sõnavormi juurde kirjutatakse selle sõnavormi morfoloogilisi ja süntaktilisi omadusi kirjeldav märgend või märgendite kombinatsioon.

Süntaktilise analüüsi ülesandeks on lause struktuuri leidmine, erinevad koolkonnad mõistavad aga lause struktuuri erinevalt ja kasutavad selle automaatseks tuvastamiseks erinevaid meetodeid. Struktuur võib näidata, millistest fraasidest lause koosneb ehk, teisisõnu, leitakse lause fraasistruktuur, või kirjeldada lause sõnade sõltuvust üksteisest.

Eesti keele süntaksianalüsaator (Müürisep 2000) põhineb kitsenduste grammatika formalismil (Karlsson jt 1995) ning annab lause igale sõnale pindmise funktsionaalse kirjelduse: analüüsi käigus ei püüta leida lause puukujulist fraasistruktuuri, vaid eraldi iga üksiku sõna funktsiooni lauses (alus, sihitis, määrus jne). Samas jäetakse esitamata sõnadevahelised täpsed sõltuvusseosed, st milline sõna millise sõna juurde kuulub. Fraasis *kulunud kaabu ja jalutuskepiga mees* märgendatakse *kulunud* kui eestäiend, täpsustamata, kas see laiendab ainult sõnavormi *kaabu* või kogu fraasi *kaabu ja jalutuskepiga* või on ta hoopis sõna *mees* eestäiendiks. Selline lähenemine võimaldab jätta lahtiseks mitmed lahendamatud mitmesused.

Kitsenduste grammatika on loomult reduktsionistlik, s.o analüüsi alguses lisatakse igale sõnavormile kõik võimalikud analüüsivariandid ja seejärel hakatakse konteksti mittesobivaid eemaldama. Eemaldamine toimub vastavalt kitsenduste grammatika reeglitele ehk kitsendustele, mis igaüks esitab mõnda spetsiifilist keelereeglilaadset

fakti. Üldisem grammatikareegel kujuneb alles nende koosmõjust.

Niisiis lisab kitsenduste grammatika süntaksianalüsaator igale sõnavormile algul kõik võimalikud süntaktilised märgendid sõnavormi morfoloogilist kirjeldust arvestades. Seejärel hakatakse konteksti sobimatuid märgendeid ükshaaval eemaldama.

Ideaaljuhul jääb analüüsi lõppedes igale sõnavormile üks süntaktiline märgend. Kui sõnal võib olla lauses mitu funktsiooni, antakse need kõik. Mitme märgendiga jäävad ka sõnad, mida analüsaator pole suutnud lõpuni ühestada. Grammatikareeglid on kirjutatud nii, et pigem jäetakse sõna mitme analüüsiga, kui eemaldatakse korrektne märgend.

Analüsaatori täpsus ja korrektsus sõltuvad väga palju sobivast märgenditesüsteemist. Järgmises peatükis vaadeldakse eesti keele kitsenduste grammatika märgendite-süsteemi ja kirjeldatakse eksperimenti selle modifitseerimiseks.

2. Süntaks eesti keele kitsenduste grammatikas

Eesti keele kitsenduste grammatikas (ESTKG) märgendatavad süntaktilised funktsioonid vastavad enam-vähem standardses eesti keele grammatikas (Erelt jt 1993) eristatavatele süntaktilistele funktsioonidele.

Öeldise märgendid eristavad finiitset ja infiniitset öeldist ning eraldi märgendid on põhiverbile ja abi- ning modaalverbidele (@+FMV, @-FMV, @+FCV, @-FCV). Fraasi põhjadest märgendatakse alust, sihitist, öeldistäidet, määrust (vastavalt @SUBJ, @OBJ, @PRD, @ADVL). Laiendite märgendid näitavad põhja leidumise suunda, kuid ei viidata ühelegi sõnale konkreetselt. See tähendab, et on eraldi märgendid ees- ja järeltäienditele (@NN>, @<NN jt), eessõna ja tagasõna laiendile (@<P, @P>) ning kvantori ees- ja järellaiendile (@Q>, @<Q). Täienditest eristatakse omadus-, määr-, kaas-, nimisõnalisi täiendeid ja partitsiipe ning infinitiivseid verbivorme.

Süntaktilise analüüsi esimesel etapil lisatakse igale sõnavormile kõik võimalikud süntaktilised märgendid arvestades sõnavormi morfoloogilist informatsiooni. Joonisel 1 on toodud näitelause pärast kõigi märgendite lisamist.

```
Pööningult
   pööning+lt //_S_ com sg abl  #cap // @ADVL @<Q @NN> @<NN
kukkunud
   kukku=nud+0 //_A_ pos #nud partic //  @VN> @<VN @PRD @ADVL
korstnajalg
   korstna_jalg+0 //_S_ com sg nom  // @SUBJ @PRD @OBJ @NN> @<NN @ADVL @<Q
lõi
   löö+i //_V_ main indic impf ps3 sg ps af  #FinV #NGP-P //  @+FMV
raekoja
   rae_koda+0 //_S_ com sg gen  //  @OBJ @ADVL @NN> @<NN @<Q
võlvlakke
   võlv_lagi+0 //_S_ com sg adit  //  @ADVL @NN> @<NN @<Q
augu
   auk+0 //_S_ com sg gen  //  @OBJ @ADVL @NN> @<NN @<Q
$.
```

Joonis 1. Näitelause pärast kõigi süntaktiliste märgendite lisamist.

Eendatud real paikneb sõnavorm, sellele järgneval real tema grammatiline kirjeldus: kaldkriipsude vahel on morfoloogilised märgendid, rea lõpus süntaktilised. Näiteks nimisõna nimetavas võib olla alus, öeldistäide, sihitis, eestäiend, järeltäiend, määrus või laiendada kvantorifraasi. Märgendite lisamiseks on grammatikas 180 reeglit.

Järgmisel etapil hakatakse märgendeid ükshaaval eemaldama arvestades konteksti. Grammatikas on 1118 märgendite eemaldamise reeglit ehk süntaktilist kitsendust. Joonisel 2 on toodud näitelause pärast süntaktilist ühestamist.

```
Pööningult
   pööning+lt //_S_ com sg abl  #cap // **CLB  @ADVL
kukkunud
   kukku=nud+0 //_A_ pos #nud partic //  @VN>
korstnajalg
   korstna_jalg+0 //_S_ com sg nom  //  @SUBJ
lõi
   löö+i //_V_ main indic impf ps3 sg ps af  #FinV #NGP-P //  @+FMV
raekoja
   rae_koda+0 //_S_ com sg gen  //  @NN>
võlvlakke
   võlv_lagi+0 //_S_ com sg adit  //  @ADVL
augu
   auk+0 //_S_ com sg gen  //  @OBJ
$.
```

Joonis 2. Näitelause pärast süntaktilist ühestamist

Vaatleme lähemalt, kuidas saavutati sõnavormi *augu* analüüs. Kvantori laiendi märgend (@<Q) eemaldati, sest vasakus kontekstis ei leidunud ühtegi kvantorina esineda võivat sõna. Eestäiendi märgend (@NN>) eemaldati, sest paremas kontekstis pole nimisõnu, mida täiendada. Järeltäiend märgend (@<NN) eemaldati, sest omastavas käändes järeltäiend on enamasti kirjavahemärkidega piiritletud, antud lauses kirjavahemärke pole. Määruse märgend (@ADVL) eemaldati, sest tegemist pole kvantorifraasiga, tüüpilise ajamäärusega, paremas kontekstis pole neljas viimases käändes määruseid, millega võiks antud sõna koordinatsioonis olla, samuti pole tegemist võrdlusmäärusega ning sõna ei kuulu partitsiipide ega infinitiividega ühte tarindisse. Ainsana jäi sõnale alles sihitise märgend.

Antud juhul oli tegemist küllaltki lihtsate reeglitega, mis välistamismeetodil leidsid õige analüüsi. Kahjuks ei ole võimalik kõiki sõnu nii hõlpsasti ühestada, automaatsel analüüsil jääb ligikaudu iga kümnes sõna mitme märgendiga. Ilukirjandusliku teksti analüüsi tulemused on toodud tabelis 1.

	Käsitsi ühestatud	***Automaatselt ühestatud***
Saagis	98,53%	96,41%
Täpsus	87,57%	78,09%
Ühesus	89,54%	82,70%

Tabel 1. Analüüsi tulemused.

Teises veerus on toodud tulemused juhul, kui sisendtekst on morfoloogiliselt ühene ja veatu, s.t seda on eelnevalt käsitsi töödeldud. Kolmandas veerus toodud tulemused on saadud täisautomaatsel analüüsil. Saagis näitab, mitu protsenti sõnadest on õige märgendiga, pööramata tähelepanu sellele, kas sõna on ühene või mitte. Täpsus näitab, mitu protsenti kõigist märgenditest on oma õigel kohal ehk siis leitud korrektsete märgendite arvu suhet kõigisse leitud märgendite arvu. Ühesus näitab, mitu protsenti sõnadest on ühese analüüsiga.

Hetkel jääb kõige sagedamini alles mitmesus määruse ja täiendite vahel. See on seletatav asjaoluga, et enamasti saab neid eristada ainult semantilise informatsiooni põhjal. Nt. *Ta kontrollis piisava täpsusega (@ADVL @NN>) emotsioone.*

Analüsaatorile on samuti raske eristada alust ja sihitist, eestäiendit ja sihitist ning määrust ja sihitist. Sihitise analüüsi keerukusel on mitmeid põhjuseid:

- aluse ja sihitise kääne langevad kokku (*Igal juhul ostavad nad aktsiad ära*);
- pole teada verbi transitiivsus/intransitiivsus konkreetses lauses (nt *puutus midagi*, aga *puutus kokku millegagi*);
- tuleb arvestada elliptiliste ja kiillausetega (*Need (@SUBJ @OBJ) aga, kes kodu valmiskujul kätte said, ei tundnud vajadust (@SUBJ @OBJ) midagi (@SUBJ @OBJ) täiendada või isegi korras hoida*);
- omastavas käändes nimisõnade vahel on raske leida fraasipiiri (*Pärtel ajab ta (@NN> @OBJ) oma vaikimisega hauda*);
- kui lauses on mitu sihilist verbi, siis puudub grammatikamudelis võimalus seostada potentsiaalset sihitist konkreetse verbiga (*Mind ahvatles võimalus (@SUBJ @OBJ) püüda üles kirjutada iseennast (@SUBJ @OBJ)*).

Analüüsi tulemused sagedasemate märgendite kaupa on toodud tabelis 2.

Märgend	Saagis	Täpsus
@ADVL	99.64	91.51
@SUBJ	99.60	86.97
@-FMV	100.00	86.60
@NN>	97.35	81.22
@OBJ	96.89	80.83
@PRD	91.30	61.76
@<NN	100.00	13.33

Tabel 2. Analüüsi tulemused mõnede märgendite kaupa.

Tabeli teine veerg näitab, et kõige sagedamini eksitakse öeldistäite ja sihitise märgendi eemaldamisel. Põhjuseid on mitmeid: kiillaused, raskesti määratavad fraasipiirid, keerulised *da*-infinitiivsed konstruktsioonid, aga ka vead reeglites. Kolmas veerg näitab, milline märgend põhjustab enim mitmesusi. Ainult 13% allesjäänud järeltäiendi märgenditest on korrektsed, ülejäänud tekitavad asjatut mitmesust. Samas on väga raske kirjutada reegleid järeltäiendi märgendite eemaldamiseks, mis oleksid lingvistiliselt põhjendatud ja ei põhjustaks massiliselt vigaseid analüüse. Nagu eespool mainitud, põhineb määruse ja määrusliku täiendi

eristamine enamasti semantikal. Samuti põhjustab liigselt mitmesusi öeldistäite märgend, kuid et öeldistäide esineb tekstis harvemini, ei mõjuta see oluliselt üldist statistikat.

3. Märgendite mõjust analüüsi tulemusele

Analüsaatori täpsus ja korrektsus sõltuvad väga palju sobivast märgenditesüsteemist. Esmapilgul tundub, et mida vähem märgendeid, seda lihtsam on korrektset grammatikat kirjutada ja seda suurema efektiivsusega on analüsaator. Tegelikult kujuneb välja nii, et töö käigus saadav süntaktiline informatsioon on sedavõrd napp, et ka neid väheseid üldiseid märgendeid on väga raske ühestada. Eesti keele kitsenduste grammatika esimeses versioonis (Müürisep 1996) ei eristatud ees- ja järeltäiendeid ning kokkuvõttes põhjustasid täiendid massiliselt mitmesusi. Kokkuvõttes said ühese analüüsi ainult 68% sõnadest. Samuti ebaõnnestus nimisõnafraasituvastusgrammatika kirjutamine, milles oli ainult kolm märgendit: fraasis sees, fraasist väljas, fraasi alustaja.

3.1. Objekti märgendi täpsustamine

Sihitise sagedane mitmesus on sageli põhjustatud sellest, et nii öeldise juurde kuuluv sihitis kui ka partitsiibi sihitis on tähistatud ühe märgendiga. Reeglites on aga raske hallata mitme verbi ja mitme sihitise fenomeni ning seetõttu põhjustab sihitise märgend asjatult mitmesusi. Seda saab lihtsustada, kui võtta kasutusele erinevad sihitise märgendid. Tehtud eksperimendis tähistati öeldise juurde kuuluvat sihitist märgendiga @OBJ ja muu verbi juurde kuuluvat sihitist märgendiga @obj. Uute märgendite tõttu tuli ümber kirjutada kogu sihitisegrammatika. Reeglite arv suurenes 33 reegli võrra, paljud neist on lihtsalt pisut muudetud koopiad. Samas muutusid reeglid tunduvalt lihtsamaks.

Saadud tulemused aga ei paranenud, vaid tekkis juurde @OBJ-@obj mitmesus, üheste analüüside osakaal langes 0,3%. Mitmesus kahe sihitise märgendi vahel esines

enamasti lausetes, milles põhiverb oli mitmese süntaktilise analüüsiga, nt *Imetlus pani iga pisiasja (@OBJ @obj) märkama (@-FMV @ADVL).*

Kui aga pärast analüüsi lõppu teisendada mõlemad sihitise märgendid jälle üheks tagasi, õnnestus suurendada üheste analüüside arvu poole protsendi võrra esialgsete tulemustega võrreldes. Kui arvestada, et mitmeste analüüside osakaal on üldse ligikaudu 10%, siis on ka 0,5% hea saavutus. Saadud tulemused on toodud tabelis 3.

	Enne	***@OBJ-@obj mitmesusega***	***Pärast***
Saagis	98,97%	98,92%	98,92%
Täpsus	87,96%	87,33%	88,34%
Ühesus	89,31%	89,06%	89,88%

Tabel 3. Tulemused eksperimendis sihitise märgenditega.

Näiteks muutus üheseks järgmine lause (sulgudes on toodud esialgsed mitmesused): *Osalemine nii Eesti Telefonis kui ka kaabeltelevisoonifirmas võtab riigilt võimaluse täita Euroopa Liidu nõudmised konkurentsi (@OBJ @NN>) loomisel telekommunikatsiooni (@OBJ @NN>) valdkonnas.*

Hoolimata uuest mitmesusest oleks ilmselt siiski mõttekam jätta lõplikku analüüsi mõlemad märgendid alles, sest hiljem saab neid mõne konkreetse rakenduse vajadusi arvestades kergesti teisendada.

Selles eksperimendis kasutati küllaltki väikest korpust (4000-sõnalist reeglite väljatöötamiseks ja 2000-sõnalist testimiseks), suurema treeningkorpuse korral on võimalik reegleid veel täpsustada ja sel teel tulemust parandada.

3.2. Perspektiivid teiste märgendite täpsustamisel

Sarnane eksperiment eraldamaks lause- ja fraasiadverbiaali ilmselt nii head tulemust ei anna, sest määruse vorm ja tema süntaktiline käitumine sõltub väga palju semantikast.

Veel üks võimalus märgendeid täpsustada on lisada põhja märgenditele põhiverbi asukohta näitav sümbol. Näiteks eespool näitena toodud lause *Pööningult kukkunud*

korstnajalg lõi raekoja võlvlakke augu saaks analüüsi *Pööningult (@ADVL>) kukkunud (@VN>) korstnajalg (@SUBJ>) lõi (@+FMV) raekoja (@NN>) võlvlakke (@<ADVL) augu (@<OBJ).* Sellist lähenemist on kasutatud portugali keele kitsenduste grammatikas (Bick 1997) ning saadud analüüsi täpsuseks 98%.

Samas eestikeelset allesjäänud mitmesustega teksti vaadates oli väga raske leida juhtumeid, kus selline märgendus aitaks lauset ühestada.

Mitmesusi aitaks lahendada: verbirektsioonide teadmine, ühend- ja väljendverbide tuvastamine (sihilisus-sihitus võib muutuda) ning mitmetasandiline analüüs, mis võimaldaks kergemini analüüsida sõnu üle osalausepiiride.

Ainsana tasuks kaaluda lisandi ja järeltäiendi märgendite eristamist. Kasu võiks sest tulla ehk teine pool protsenti?

3.3.Märgendite ühendamine

Analüüsides allesjäänud mitmesusi ilmneb, et küllaltki suure osa neist moodustavad sellised, mida ei saagi ühegi reegliga eemaldada. Ehk oleks mõttekas peita sellised mitmesused üldisemate märgendite taha?

Näiteks on suuri probleeme *ma*-infinitiivi süntaktilise rolli määramisega. Ei ole võimalik kirjutada reeglit, mis eristaks järgmistes lausetes *ma*-infinitiivi funktsioone: *Lapsed kippusid õue mängima* (määrus)*; Hing kippus kinni jääma* (öeldis).

Samuti on raskusi partitsiipidel määruse ja öeldise funktsiooni eristamisel. Kui asendada analüüsitud tekstis kõik määruse ja öeldise vahelised mitmesused ühe uue märgendiga (@ADVLFMV), suureneb üheste analüüside osakaal ühe protsendi võrra (enne 89,73%, nüüd 90,71%).

Samuti sõltub *da*-infinitiivi süntaktiline funktsioon lause tähendusest. *da*-infinitiivile lisatakse seitse märgendit: öeldise osa, alus, sihitis, määrus, öeldistäide, ees- ja järeltäiend. Keerukamate lausete korral õnnestub neist eemaldada ainult paar-kolm. Kui asendada *da*-infinitiivile alles jäänud märgendid üheainsaga (lihtsalt @DAINF), tõuseb üheste analüüside osakaal pool protsenti (enne 90,71%, nüüd 91,28%).

Kolmas suur mitmesuste rühm, mida võiks üheks märgendiks üldistada, on valik

järeltäiendi ja määruse vahel. Üldisema märgendi kasutuselevõtt tõstab üheste analüüside osakaalu veel 2,5% (enne 91,28%, nüüd 93,86%).
Kahjuks ei anna sellised märgendid sisuliselt analüsaatori kvaliteedile midagi juurde ning nende kasutamine sõltub konkreetsest rakendusest.

3.4. Lahendamata mitmesused

Allesjäänud mitmesusi (sihitis või eestäiend, alus või sihitis, alus või öeldistäide ning hulga- ja ajamäärusega ning lisandiga seotud mitmesused) ei saa enam kuidagi üldistada: ühest küljest läheks kaduma väga palju süntaktilist informatsiooni, teisalt annaks iga selline üldistus eraldi väga väikese efekti. Näiteks jääb mitmeseks lause: *Nad olid ju kaks* (öeldistäide, määrus) *juhuslikku inimest teineteise kõrval.* Määruse märgendit ei õnnestu eemaldada, sest fraas *kaks juhuslikku päeva* oleks samas lauses kindlasti määrus. Järgmises lauses jäi alles neli aluse ja öeldistäite vahelist mitmesust: *Ning pahatihti oli just Pärtel* (alus ja öt)*, tema tegemised* (alus ja öt) *ja sõnad* (alus ja öt) *ning isegi ütlemata või mõtlematagi mõtted* (alus ja öt) *olnud naise tähelepanu keskpunktis.* Aluse ja sihitise vahelise mitmesuse näiteks on lause: *Nii sündisid nägelemised ja jäid tegemata tööd* (alus või sihitis).
Ilmselt on olemasoleva formalismi piires võimalik analüsaatori täpsust veel suurendada, kui võtta kasutusele täiendavad leksikonid aja- ja hulgamääruste ning kvantorite haldamiseks ning uurida mitmesuste tüüpe suurematel korpustel. Pikemas perspektiivis tuleks siiski minna üle sügavamale süntaksikirjeldusele. Sõnadevaheliste sõltuvuste ilmutatult esiletoomine aitaks muuta tulemust veelgi täpsemaks ning lause mitmetasandiline analüüs võimaldab vältida vigu kiiluga poolitatud lausetes.

4. Kokkuvõte

Käesolevas artiklis käsitletakse eesti keele süntaksianalüsaatori märgendite probleeme. Hetkel jääb süntaksianalüsaatori väljundis ligikaudu 10% sõnadest mitmese analüüsiga. Artiklis uuritakse, kuidas märgendite muutmise teel mitmeste

analüüside osakaalu vähendada. Kui märgendite abil eristada öeldise juurde kuuluvat sihitist muudest, siis suureneb mitmeste analüüside arv tekkinud uue mitmesuse tõttu, kuid väheneb teiste sihitisega mitmesuste protsent. Kui aga üldistada mitmeid märgendeid, siis on võimalik saavutada üheste analüüside osakaalu kasv 4% võrra, kuigi sisuliselt analüüsi kvaliteet sellest ei muutu.

Kirjandus

Bick, Eckhard 1997. Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk. - Sprog og Multimedier. Toim. T. Brøndsted, I. Lytje. Aalborg: 39-57.

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

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.

Müürisep, Kaili 1996. Eesti keele kitsenduste grammatika süntaksianalüsaator. Magistritöö. Arvutiteaduse Instituut, Tartu Ülikool.

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