# Eesti keele formaalne grammatika: mudelist rakenduseni

Kaili Müürisep ja Tiina Puolakainen
Tartu Ülikool ja Eesti Keele Instituut

## Sissejuhatus

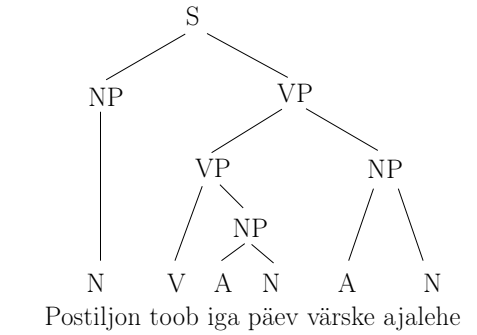
Keele automaattöötluse keskseks osaks on süntaksianalüsaator koos keele formaliseeritud grammatikaga.

Loomuliku keele süntaksianalüsaator on programm, mis saab sisendiks morfoloogiliselt analüüsitud teksti (s.t on leitud iga sõna tüvi, lõpud, sõnaliik, kääne või pööre jms) ning väljastab süntaktiliselt analüüsitud teksti (leitud on igas lauses alus, öeldis, sihitis jt lauseliikmed). Enamasti esitatakse süntaktiline kirjeldus märgendite abil, s.t iga sõnavormi juurde kirjutatakse selle sõnavormi morfoloogilisi ja süntaktilisi omadusi kirjeldav märgend või märgendite kombinatsioon.

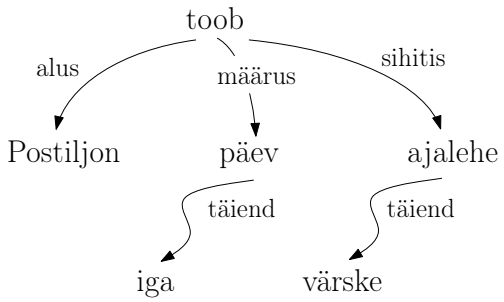
Süntaktilise analüüsi ülesandeks on lause struktuuri leidmine, erinevad koolkonnad mõistavad aga lause struktuuri erinevalt ja kasutavad selle automaatseks tuvastamiseks erinevaid meetodeid. Struktuur võib näidata, millistest fraasidest lause koosneb ehk, teisisõnu, leitakse lause fraasistruktuur, või kirjeldada lause sõnade sõltuvust üksteisest. Joonisel 1 on toodud lause *Postiljon toob iga päev värske ajalehe* fraasistruktuur, joonisel 2 on sama lause sõltuvusstruktuur.

Samuti võib olulisel määral erineda süntaksianalüsaatori tööpõhimõte. Enamasti kasutavad süntaksianalüsaatorid sellist formaliseeritud grammatikat, mis on koostatud inimeste poolt, kuid leidub ka analüsaatoreid, mille n-ö grammatikaks on tõenäosuslik keelemudel, mis on saadud automaatselt suurel tekstihulgal analüsaatorit treenides.

Töö eesti keele süntaksianalüsaatori loomisega algas Tartu Ülikoolis 90ndate keskel kahes magistritöös ning seda on toetanud ka Eesti Teadusfond (projektid 3314 ja 4605). Ühe vaheetapi lõpuks




Joonis 1. Näitelause fraasistruktuur.




Joonis 2. Näitelause sõltuvusstruktuur.

võib pidada valminud doktoritöid [6,7]. Projektis osalesid lisaks käesoleva artikli autoritele ka Mare Koit, Kadri Muischnek, Tiit Roosmaa ja Heli Uibo.

Eesti keele süntaksianalüsaator kasutab lingvistilist keelemudelit ehk formaalset grammatikat, mis põhineb kitsenduste grammatika formalismil [4].

## Kitsenduste grammatikast

Kitsenduste grammatika 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 täp-

sed sõltuvusseosed, s.t 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 mõned mitmeti tõlgendatavused.

Kitsenduste grammatika on loomult reduktsionistlik, s.t 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.

Automaatne süntaktiline analüüs koosneb selle formalismi kohaselt kolmest etapist: morfoloogilisest ühestamisest (kus morfoloogilise analüsaatori poolt leitud mitmest analüüsivariandist valitakse üks, antud konteksti sobiv), osalausepiiride määramisest ja sõnade süntaktiliste funktsioonide määramisest.

Vaatame näitena eestikeelset morfoloogiliselt analüüsitud[1] lauset *Aknas kustus tuli*:

```
Aknas
    aken+s //_S_ com sg in //
    Nimisõna ainsus seesütlev
kustus
    kustu+s //_V_ main indic impf ps3 sg ps af  #Intr //
    Põhiverb kindel kv lihtminevik 3. pööre ainsus
             isikuline tm sihitu
tuli
    tule+i //_V_ main indic impf ps3 sg ps af #Intr //
    Põhiverb kindel kv lihtminevik 3. pööre ainsus
             isikuline tm sihitu
    tuli+0 //_S_ com sg nom //
    Nimisõna ainsus nimetav
$.
    . //_Z_ Fst //
    Punkt
```

Selle lause morfoloogilisel ühestamisel leitakse sõnavormi *tuli*

[1] Eesti keele morfoloogiaanalüsaator on kättesaadav veebiaadressilt: `http://www.filosoft.ee/html_morf_et/`

õige tõlgendus (nimisõna, aga mitte verb) järgmise kitsenduse rakendamisel: eemaldada verbi pöördeline vorm (antud juhul tule+i //_V_ main indic impf ps3 sg ps af #Intr //), kui antud sõnale eelneb vahetult verbi pöördeline vorm, mis on sõnavormi ainus tõlgendus. Morfoloogilise ühestamise tulemusena on igal sõnavormil üks tõlgendus. Kui mingil põhjusel ühese tõlgenduse leidmine ei õnnestu, jäetakse sõnavormile mitu tõlgendust (nt lauses *Keel on võimas realiteet* võib sõnavorm *keel* olla nii nimetavas kui ka alalütlevas käändes).

Osalausepiiride määramine toimub vaheldumisi morfoloogilise ühestamisega ning on oluline eelkõige liitlausete analüüsil.

Süntaktiline märgendamine on jaotatud kaheks: kõigepealt lisatakse sõnavormile kõik võimalikud süntaktilised märgendid, mis sobivad sõnavormi morfoloogiliste tunnustega. Näitelause saab sellel etapil järgmise kuju:

```
Aknas
   aken+s //_S_ com sg in **CLB // @ADVL @<NN @NN>
kustus
   kustu+s //_V_ main indic impf ps3 sg ps af  #Intr// @+FMV
tuli
   tuli+0 //_S_ com sg nom // @SUBJ @OBJ @ADVL @NN> @<NN
$.
    . //_Z_ Fst //
```

Sõnavorm *aknas* võib olla määrus e adverbiaal (märgend @ADVL), järeltäiend (@<NN) või eestäiend (@NN>); sõna *kustus* – finiitne öeldis (@+FMV); sõna *tuli* – alus e subjekt (@SUBJ), sihitis e objekt (@OBJ), määrus, ees- või järeltäiend (@NN> @<NN). Märgend CLB tähistab (osa)lause piiri.

Lõpuks rakendatakse sõnadele süntaktilisi kitsendusi, mis eemaldavad konteksti sobimatud süntaktilised märgendid:

```
Aknas
   aken+s //_S_ com sg in **CLB // @ADVL
kustus
   kustu+s //_V_ main indic impf ps3 sg ps af  #Intr// @+FMV
tuli
   tuli+0 //_S_ com sg nom // @SUBJ
$.
    . //_Z_ Fst //
```

Sõnavorm *aknas* analüüsiti määruseks ja *tuli* aluseks, tegusõnale

*kustus* jäi finiitse öeldise märgend.

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.

## Morfoloogiline ühestamine

Morfoloogiline analüsaator leiab sõnavormile kõik võimalikud morfoloogilised analüüsid, kuid ainult üksikut sõnavormi vaadates ei ole tal võimalust nende hulgast valikut teha. Eesti keele morfoloogiaanalüsaator ESTMORF [2] suudab leida õige morfoloogilise analüüsi enam kui 99,5% sõnavormidest, kuid samas on eesti keele morfoloogilise mitmesuse tase väga kõrge: 45% sõnavormidest saavad mitmese analüüsi (inglise keelel on vastav protsent 40 ja soome keelel 11).

Eesti keele morfoloogilise ühestamise grammatikas on 1240 reeglit, ligikaudu pooled neist käsitlevad konkreetseid sõnavorme.

Morfoloogilise ühestamise reeglite väljatöötamisel ei olnud võimalik kasutada olemasolevaid grammatikakirjeldusi, kuna sellise probleemivaldkonnaga klassikaline lingvistika ei tegele. Reeglite väljatöötamist alustasime sagedasemate mitmesusklasside väljaselgitamisest. Leidsime sagedasemad mitmesed sõnavormid (*oli, ei, on, ta, kui, aga*) ja ka grammatiliste kategooriate mitmesused:

- *nud-* või *tud*-partitsiip omadussõna ainsuse või mitmuse (*loetud raamat, loetud raamatud*), nimisõna (*loetu* mitmuses) ja tegusõna tõlgendustega (*raamatut oli loetud*);

- nimisõna nimetavas, omastavas ja osastavas käändes, nt *maja, loetelu, asjaolu* jt;

- nimisõna omastavas, osastavas ja lühikeses sisseütlevas käändes, nt *jaama, pealkirja, konverentsi*;
- määrsõna ja omadussõna alaltütlevas käändes nt *kindalt, tugevalt* (nt *kindalt pinnalt* või *teadis kindlalt*).

Selline sagedustabel andis vajaliku informatsiooni olulisimate reeglite koostamiseks. Paljud sel teel leitud sagedased mitmesused olid kergesti eemaldatavad (näiteks reeglitega: sõnavorm *ei* ei ole verbi osa, kui järgnev sõna pole verb; eemaldada omadussõna alaltütleva käände tõlgendus, kui paremas kontekstis ei leidu nimisõna alaltütlevas käändes). Saadud reegleid testiti käsitsi märgendatud korpusel e suurel tekstihulgal, kus lingvisti poolt valitud iga õige morfoloogiline tõlgendus oli tähistatud vastava märgendiga. Silumisrežiimis töötav ühestamisprogramm märgib iga analüüsitava sõna juures, milliseid reegleid millises järjekorras kasutati ning lingvisti poolt õigeks peetava tõlgenduse eemaldamise korral kirjutab vigase reegli identifikaatori koos analüüsitava lausega vigade logifaili. Sellise informatsiooni põhjal on võimalik reeglite kontekstitingimusi täiendada ja parandada ning seega vigade hulka vähendada.

Suurim probleem loomuliku keele formaalse grammatika väljatöötamisel on see, et lingvistilised grammatikakirjeldused on (ja peavadki olema) orienteeritud inimesele, mitte arvutile – sageli kasutatakse sõnaliigi, käände jms kindlaksmääramisel semantilist informatsiooni, mis aga on raskesti formaliseeritav.

Eesti keele kitsenduste grammatika koostamisel ja esialgsel testimisel kasutati 20314-sõnalist eelnevalt käsitsi morfoloogiliselt ja süntaktiliselt märgendatud tekstikorpust (edaspidi treeningkorpus).

Analüsaatori töö hindamiseks kasutati käsitsi märgendatud 9663-sõnalist testkorpust, mille abil ei olnud varem grammatikareegleid optimeeritud ega hinnatud.

Nii treening- kui ka testkorpuses vähendas ühestaja rakendamine mitme tõlgendusega sõnade protsenti neli korda. Morfoloogilise ühestamise tulemused treeningkorpuses olid: täpsus[2] 83,39–

[2]Täpsus (ingl k *precision*) – leitud õigete analüüside osakaal kõikide leitud analüüside hulgas.

89,68%, saagis[3] 97,87–99,16%, ühe tõlgendusega sõnade protsent 88,67–91,74.

Morfoloogilise ühestamise tulemused testkorpuses olid: täpsus 85,49-89,16%, saagis 97,95-98,36%, ühe tõlgendusega sõnade protsent 88,67-91,96.

Kõige rohkem vigu tehti nimisõnade (aga ka omadus- ja asesõnade) käände määramisel: raskusi tekitab nimetava, omastava, osastava ja lühikese sisseütleva eristamine. Näiteks fraasis *maailma juhtivad majandusriigid* võib sõnavorm *maailma* olla nii omastavas, osastavas kui sisseütlevas käändes, kõik need võimalused on grammatiliselt korrektsed. Teisel ja kolmandal kohal on partitsiipide määramise probleem omadussõnaks või verbiks ja ka sellega tihedalt seotud olema vormi määramine kas põhi- või abiverbiks (vrdl *Kursor oli liigutatud* ja *Tädi oli liigutatud*).

Olgu siinkohal märgitud, et eesti keele jaoks on praeguseks loodud ka teine, Markovi peitmudelil põhinev morfoloogiline ühestaja [3], mis kasutab oma töös ainult statistilist infot teksti kohta, mitte aga lingvistilisi reegleid. Sellest johtuvalt on ka vigade protsent mõnevõrra suurem kui reeglipõhisel ühestajal (3%).

## Osalausepiiride määramine

Osalausepiirid määratakse sidesõnade, kirjavahemärkide ja verbide põhjal. Osalausepiiride määramise põhireegel on järgmine: kui sõnale eelneb kirjavahemärk ja/või sõna ise on sidesõna ning vasakul ja paremal pool seda sõna leidub verbi pöördeline vorm, siis see sõna on osalause esimene sõna. Nt *Pärtel oleks võinud otsida sõprade seltsi, aga ta oli aja jooksul kogenud, et sõnadest sünnib valesti mõistmine ja arusaamatusest tõuseb tüli.*

See reegel võib mõnede tingimuste osas varieeruda. Nimelt võib koma või rinnastavate sidesõnade *ja*, *ning*, *või*, *ega*, *ehk* abil eraldada mitte ainult osalauseid, vaid ka koondlause korduvaid liikmeid. Seda, millise eraldajaga just konkreetsel juhul on tegu, on ilma sün-

[3]Saagis (ingl k *recall*, eesti keeles on varem kasutatud ka termineid katvus ja korrektsus) – leitud õigete analüüside osakaal kõikide õigete (inimese poolt tehtud) analüüside hulgas.

taktilist informatsiooni teadmata raske otsustada, eriti veel juhul, kui antud sõna lähemas kontekstis ei leidu verbe. Seepärast lisatakse nendele sõnadele üksnes oletatava osalause tunnus. Järgmises lauses on märgendiga CLB näidatud kindlad osalausepiirid ja märgend CLB-C tähistab oletatavat lausepiiri. Nt *"Neist pooled deklareerivad tulusid ametlikult, CLB veerand on allilmategelased ja CLB-C veerand lihtsalt ebaausad ärimehed," CLB väitis ta.*

Grammatikas on 47 osalausepiiride määramise reeglit, paljud neist on väga spetsiifiliste juhtude jaoks.

## Süntaktiline ühestamine

Eesti keele kitsenduste grammatikas märgendatavad süntaktilised funktsioonid vastavad enam-vähem standardses eesti keele grammatikas [1] eristatavatele süntaktilistele funktsioonidele.

Öeldise märgendid eristavad finiitset ja infiniitset öeldist ning eraldi märgendid on põhiverbile ja abi- ning modaalverbidele (@+FMV, @-FMV, @+FCV, @-FCV). Nt *Sellest ei (@NEG) oleks (@+FCV) pidanud(@-FCV) rääkima(@-FMV).* Fraasi põhjadena 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. Järgnevalt on toodud süntaktiliselt analüüsitud näitelause:

```
Eesti                                ;; nimisõnaline eestäiend
   Eesti+0 //_S_ prop sg gen #cap // **CLB  @NN>
vanimad                              ;; omadusssõnaline eestäiend
   vanim+d //_A_ super pl nom  //  @AN>
asukad                               ;; alus
   asukas+d //_S_ com pl nom  //  @SUBJ
saabusid                             ;;öeldis
   saabu+sid //_V_ main indic impf ps3 pl ps af #Intr//@+FMV
siia                                 ;; määrus
   siia+0 //_D_ //  @ADVL
```

```
pärast                              ;; määrus
   pärast+0 //_K_ pre #part //  @ADVL
viimast                             ;; omadussõnaline eestäiend
   viimane+t //_A_ pos sg part  //  @AN>
jääaega                             ;; eessõna laiend
   jää_aeg+0 //_S_ com sg part //  @<P
```

Täiendite *Eesti* ja *vanimad* märgendid näitavad küll põhja *asukad* suunas, kuid otsest viidet nende vahel pole. Ka sõnavormi *jääaega* märgend @<P näitab, et sõna kuulub eessõnafraasi, kuid eessõna ja selle juurde kuuluva nimisõnafraasi põhi pole omavahel ilmutatult seotud.

Süntaktilise analüüsi esimesel etapil lisatakse igale sõnavormile kõik võimalikud süntaktilised märgendid, arvestades sõnavormi morfoloogilist informatsiooni.

Eesti keele kitsenduste grammatikas on 180 süntaktiliste märgendite lisamise reeglit. Nende koostamisel on arvestatud, et võimalikult paljudel juhtudel oleks lisatavate märgendite hulgas ka õige märgend, vältides samal ajal üksikute harva esinevate märgendite lisamist kõigile sõnadele. See on saavutatud kontekstitingimuste lisamisega reeglitele. Nii näiteks lisatakse kaassõna laiendi märgend @P> või @<P ainult sel juhul, kui osalauses leidub vastavat käänet nõudev kaassõna (nt eelmises näites sõnavorm *jääaega*). Pärast märgendite lisamise etappi on süntaktilise mitmesuse protsent väga suur, keskmiselt on ühel sõnavormil 3,8 märgendit. Enamasti on üheselt analüüsitud ainult verbide pöördelised vormid ja sidesõnad. See-eest on vigade osakaal sellel analüüsi etapil väga väike, alla 0,2%. Vigu põhjustavad enamasti omadussõnad, mida kasutatakse nimisõna rollis, nt *Ikka leidus uusi lihtsameelseid, kes ootasid, et ta neid õnge võtaks.*

Järgmisel etapil hakatakse märgendeid ükshaaval eemaldama, arvestades konteksti. Grammatikas on 1118 märgendite eemaldamise reeglit ehk süntaktilist kitsendust.

Automaatsel analüüsil jääb ligikaudu iga kümnes sõna mitme märgendiga. Ilukirjandusliku teksti analüüsi tulemused on toodud tabelis 1. Teises veerus on toodud tulemused juhul, kui sisendtekst on morfoloogiliselt ühene ja veatu, s.t seda on eelnevalt käsitsi töö-

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

Tabel 1. Analüüsi tulemused

deldud. 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*. Enamasti mõistetakse selle lause korral, et kontroll toimus piisava täpsusega, kuid grammatiliselt on täiesti korrektne ka tõlgendus, et piisava täpsusega emotsioonid said kontrollitud.

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*), sel juhul pole selge, kumb on alus ja kumb sihitis;
- pole teada verbi sihilisus/sihitus 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*);

| **Märgend** | **Saagis** | **Täpsus** |
|---|---|---|
| @ADVL (määrus) | 99,64 | 91,51 |
| @SUBJ (alus) | 99,60 | 86,97 |
| @-FMV (öeldise infiniitne vorm) | 100,00 | 86,60 |
| @NN> (eestäiend) | 97,35 | 81,22 |
| @OBJ (sihitis) | 96,89 | 80,83 |
| @PRD (öeldistäide) | 91,30 | 61,76 |
| @<NN (järeltäiend) | 100,00 | 13,33 |

Tabel 2. Analüüsi tulemused mõnede märgendite kaupa

- omastavas käändes nimisõnade vahel on raske leida fraasipiiri (*Ta asetas mantli (@OBJ @NN>) tooli (@OBJ @NN>) seljatoele*);

- 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. 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.

## Rakendusvaldkonnad

Süntaksianalüsaatorit kasutatakse ühe moodulina väga paljudes loomuliku keele töötlust vajavates programmides.

Dokumenditöötluses on oluline koht info otsingul ja et enamasti otsitakse termineid või fraase, siis vajatakse nimisõnafraaside tuvastamist, mis on üks süntaksianalüsaatori ülesannetest. Samuti on vaja dokumentide automaatset liigitamist ja refereerimist, mis eeldab samuti lause analüüsi.

Süntaksianalüsaator on asendamatu tõlkijate abivahendite loomisel ja masintõlkeprogrammides. Eesti keelest näiteks inglise keelde tõlkimisel on lisaks fraaside tõlkevastete leidmisele vaja muuta ka lauseliikmete järjestust.

Kirjutaja abivahenditest vajavad nii grammatikakontrollija kui ka stiilikorrektor süntaktilist analüüsi ning õigekirjakorrektor morfoloogilist ühestamist. Ainult morfoloogilisel analüüsil põhinev eesti keele õigekirjakorrektor ei suuda (ega peagi suutma) leida üles vigaseid sõnu, kui need langevad kokku mõne teise sõnavormiga, nt fraasis *töö praemaks korraldamiseks* on iga sõnavorm eraldi korrektne, aga terve lause analüüsil ilmselt selguks, et nimisõna nimetavas käändes (s.o *praemaks*) ei sobi sellisesse konteksti.

Perspektiivne on süntaksianalüsaatori kasutuselevõtt keeleõppeprogrammides. Süntaksianalüsaator abistab ka tekst-kõne süntesaatoris lause prosoodia kindlaksmääramisel, võimaldades pidada fraaside vahel pikemaid pause ja määrata lause intonatsiooni.

Eesti keele süntaksianalüsaatorit on seni kasutatud kahe programmi koosseisus. Eksperimentaalne nimisõnafraaside tuvastaja[4] kasutab süntaktiliselt analüüsitud teksti fraasipiiride määramiseks. Automaatsel sisukokkuvõtete tegemisel [5] kasutatakse süntaktilist analüüsi lauses oluliste sõnade äratundmiseks.

Ja loomulikult kasutatakse süntaksianalüsaatorit lingvistilises uurimistöös ning uue keeletehnoloogilise tarkvara väljatöötamisel.

[4] `http://www.eki.ee/keeletehnoloogia/projektid/EstNPTool/`

## Kokkuvõte

Eesti keele kitsenduste grammatika koosneb 2500 reeglist, ligikaudu pooled neist tegelevad morfoloogilise ühestamisega, pooled süntaktiliste funktsioonide määramisega. Sellel grammatikal põhinev süntaktiline analüsaator suudab leida ühese analüüsi ligi 83% sõnadest, tehes vigu alla 4%. See on võrreldav tulemus paljude teiste keelte grammatikatega, kuigi erinevate keelte süntaktilise analüüsi probleemide keerukust nii lihtsalt numbritega mõõta ei saagi.

Eesti keele arvutigrammatika kirjutamisel on jõutud etappi, kus olemasolevad võimalused on ennast ammendanud, kuid saadud resultaat on piisavalt hea, et sellest teada anda. Mida teha selleks, et viia eesti keele arvutigrammatika uuele tasandile?

- Grammatikas tuleb laiendada leksikoni osakaalu. Vajalik on teatud semantilist informatsiooni sisaldava leksikoni olemasolu.
- Märgendatud treeningkorpuse maht peab suurenema. Et grammatika täpsus on jõudnud olemasoleval treeningkorpusel 85%-ni, on mitmese analüüsiga jäänud veel vaid mõni tuhat sõna, aga nii väikesel hulgal ei ole võimalik haruldasematele grammatilistele seaduspärasustele jälile jõuda.
- Suuremal treeningkorpusel on võimalik katsetada erinevaid statistilisi meetodeid grammatika automaatseks genereerimiseks. Süntaksianalüsaatori potentsiaalseid rakendusi silmas pidades tuleb leida meetod täiesti ühese analüüsi saavutamiseks, liites kitsenduste grammatika analüsaatori mõne statistilise analüsaatoriga.
- Pikemas perspektiivis minna üle sügavamale süntaksikirjeldusele.

Seda tööd on juba alustatud, suurendatud on nii tekstikorpuse mahtu (200000 sõna) kui ka tehtud algust eesti keele puudepanga loomisega[5].

[5] `http://corp.hum.sdu.dk/tgrepeye_est.html`

## Kirjandus

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

2. Heiki-Jaan Kaalep. ESTMORF. A Morphological Analyzer for Estonian. In *Estonian in the Changing World*, pages 43–98. Tartu, 1996.

3. Heiki-Jaan Kaalep ja Tarmo Vaino. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. *Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1*, pages 87–101. Tartu, 2000.

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

5. Andres Lippur. Eesti keele automaatne sisukokkuvõtete tegemine. Bakalaureusetöö. Käsikiri. Arvutiteaduse Instituut, Tartu Ülikool, 2000.

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

7. Tiina Puolakainen. *Eesti keele arvutigrammatika: morfoloogiline ühestamine.* Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu Ülikooli kirjastus, Tartu, 2001.