Eesti keele süntaksianalüsaatori märgenditest

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

Loomuliku keele süntaksianalüsaator on programm, mis saab sisendiks morfoloogiliselt analüüsitud teksti ja mis väljastab süntaktiliselt analüüsitud teksti. Enamasti esitatakse süntaktiline kirjeldus märgendite abil. Eesti keele süntaksianalüsaator (Müürisep 2000) põhineb kitsenduste grammatika formalismil (Karlsson jt 1995) ning annab lause igale sõnale madala pindmise funktsionaalse kirjelduse: analüüsi käigus ei püüta leida lause puukujulist fraasistruktuuri, vaid eraldi iga üksiku sõna funktsiooni lauses (alus, sihitis, määrus jne).

Kitsenduste grammatika süntaksianalüsaator lisab algul igale sõnavormile kõik võimalikud süntaktilised märgendid sõnavormi morfoloogilist kirjeldust arvestades. Seejärel hakatakse konteksti sobimatuid märgendeid ükshaaval eemaldama. Märgendite lisamine ja eemaldamine toimub vastavalt süntaksianalüsaatori grammatika reeglitele.

Ideaaljuhul jääb analüüsi lõppedes igale sõnavormile üks süntaktiline märgend. Kui sõnal võib olla lauses mitu funktsiooni, antakse need kõik. Näiteks lauses *Naaseme meie teema juurde* võib sõna *meie* olla omastavas käändes täiendiks ja nimetavas käändes aluseks. 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.

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. Fraasi põhjadest märgendatakse alust, sihitist, öeldistäidet, määrust. Laiendite märgendid näitavad põhja leidumise suunda, kuid ei viidata ühelegi sõnale konkreetselt. See tähendab, et on eraldi märgendid ees- ja järeltäienditele, eessõna ja tagasõna laiendile ning kvantori ees- ja järellaiendile. Täienditest eristatakse omadus-, määr-, kaas-, nimisõnalisi täiendeid ja partitsiipe ning infinitiivseid verbivorme.

Tüüpiline analüüs näeb välja järgmine:

Narva	@NN>
jõel	@ADVL
avastati	@+FMV
pühapäeva	@NN>
hommikul	@ADVL
õlireostus	@OBJ

Märgend @NN> tähistab eestäiendit, kuid ei täpsusta, millise sõna juurde täiend kuulub. Näiteks täiend *pühapäeva* võib laiendada nii määrust *hommikul* kui ka sihitist *õlireostus.* Selline grammatiliste seoste esitusviis võimaldab jätta lahtiseks osa

lahendamatuid grammatilisi mitmeti tõlgendatavusi. Näiteks fraasis *lumised teed ja tänavad* ei täpsustata, kas täiend *lumised* kuulub ainult sõna *teed* juurde või laiendab kogu fraasi.

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.

Grammatika praeguses versioonis jääb kõige sagedamini alles mitmesus määruse ja täiendite vahel. See on seletatav asjaoluga, et enamasti saab neid eristada ainult semantilise informatsiooni põhjal. Analüsaatorile on samuti raske eristada alust ja sihitist, eestäiendit ja sihitist ning määrust ja sihitist. Sihitise sagedane mitmesus on põhjustatud sellest, et nii põhiverbi 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. Ettekandes käsitletakse, milliseid probleeme sihitise märgendi mitmekesistamine lahendab ja milliseid juurde toob.

Samuti vaadeldakse ettekandes teisi märgenduse täpsustamise viise. Näiteks hinnatakse, millist kasu tooks põhja märgenditele suunda näitavate sümbolite lisamine, nagu on seda portugali keele kitsenduste grammatikas (Bick 1997) ja kui kaugele jääb sellise lähenemise korral võimalik üleminek sügavamale süntaksikirjeldusele, mis annaks mitmetasandilise süntaktilise analüüsi ja esitaks sõnadevahelised sõltuvused ilmutatult.

Kirjandus

Bick, Eckhard. 1997. Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk. Brøndsted, Tom &Lytje, Inger (eds), *Sprog og Multimedier*, pp. 39-57. Aalborg.

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.