Emakeele Seltsi aastaraamat 62 (2016), 122–145
http://doi.org/10.3176/esa62.04

EESTI KEELE SÕLTUVUSPUUDE PANK JA SELLE KEELETEOREETILISED LÄHTED[1]

KADRI MUISCHNEK,
KAILI MÜÜRISEP

Annotatsioon. Eesti keele sõltuvuspuude panga (EDT) märgendus põhineb sõltuvussüntaksil. Artikkel käsitleb sõltuvussüntaksi keeleteoreetilisi lähteid ja probleemkohti, nende põhimõtete rakendamist ning probleemsete keelendite esitamist eesti keele analüüsil. Näidatakse sõltuvusesituse eeliseid võrreldes fraasistruktuurianalüüsiga ja arutletakse mitteprojektiivsuse näitel selle üle, millist uut infot saab lause ehituse formaalne esitamine anda keeleteadusele. Vaatluse all on ka EDT võimalikud edasiarendused.

Võtmesõnad: sõltuvussüntaks, puudepank, automaatne süntaksianalüüs, eesti keel

1. Sissejuhatus

Puudepangaks (ingl *treebank*) nimetatakse süntaktilise struktuuri (süntaksi-*puude*) suhtes märgendatud keelekorpust. Selles artiklis tutvustatakse eesti keele sõltuvuspuude panka (EDT[2]) ning esitatakse selle loomise lingvistilised alused.

Eesti keele sõltuvuspuude pank on mahukas (sisaldab ligi 400 000 sõna) käsitsi märgendatud tekstikorpus, milles on nii ajakirjandus-, ilukirjandus- kui ka teadustekste.

EDT on märgendatud sõltuvussüntaksi põhimõtete järgi, teksti kujul on esitatud iga sõna morfoloogiline info, süntaktiline funktsioon ja viide sõnale, millest ta otseselt sõltub (ülemustipule).

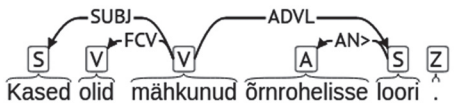
[1] Kirjutise valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus) ning Eesti Haridus- ja Teadusministeerium uurimisprojektiga IUT20-56.

[2] Selle saab alla laadida aadressilt https://github.com/EstSyntax/EDT.

Joonisel 1 on toodud lause tekstikuju korpuses ning joonisel 2 selle üks võimalik visualiseering. Nagu sõltuvussüntaksile omane, on lause süntaktiliseks keskmeks verb ning teised lauseliikmed sõltuvad sellest kas otseselt või kaudselt. Järgmistes osades tutvustatakse järk-järgult, kuidas selline esitusviis on välja kujunenud. Sõltuvussüntaksi põhimõtteid kirjeldatakse lühidalt osas 2. Osas 3 antakse ülevaade varasematest katsetest märgendada eestikeelseid lauseid fraasistruktuurisüntaksi järgi ning näidatakse fraasistruktuuriesituse probleeme eestikeelse lause süntaktilise struktuuri esitamisel. Otsus eelistada eesti keele puhul just sõltuvusesitust lähtus seega nii varasemast kogemusest kui ka arvutilingvistika üldisematest arengusuundadest. Osas 4 esitletakse EDT struktuuri ja märgendust ning osas 5 põhilisi keelendeid, mille märgendusviis sõltuvussüntaksis erineb koolkonniti. Osas 6 käsitletakse mitteprojektiivseid lauseid ning kokkuvõttes kirjeldatakse ka puudepanga edasiarendamise võimalusi.

```
"<s>"
"<Kased>"
        "kask" Ld S com pl nom cap @SUBJ #1->3
"<olid>"
        "ole" Lid V aux indic impf ps3 pl ps af @FCV #2->3
"<mähkunud>"
        "mähku" Lnud V main partic past ps @IMV #3->0
"<õrnrohelisse>"
        "õrn_roheline" Lse A pos sg adit @AN> #4->5
"<loori>"
        "loor" L0 S com sg adit @ADVL #5->3
"<.>"
        "." Z Fst CLB #6->6
"</s>"
```

Joonis 1. Eesti keele sõltuvuspuude panga lause tekstikuju



Joonis 2. Joonisel 1 esitatud EDT lause visualiseering

2. Sõltuvussüntaksi põhimõtted

Puudepanga märgendus põhineb sõltuvussüntaktilisel lähenemisel, mille puhul esitatakse kogu lausestruktuur kahe sõnavormi vahelise ebasümmeetrilise suhtena (siinses artiklis kasutatakse traditsioonilise *põhja* ja *laiendi* asemel termineid *ülemus* ja *alluv*, mis autorite arvates väljendavad tabavalt, kuigi metafoorselt sõltuvussuhte olemust) ning sellel suhtel on süntaktilisest funktsioonist tulenev nimetus. Lausestruktuuri esitamisel mitteterminaalseid sümboleid ei kasutata, st sõltuvussuhted on sõnade vahel, vahesõlmi (fraase, moodustajaid) ei moodustata. Ühel sõnal võib olla mitu alluvat, aga ainult üks ülemus.

Sõltuvussüntaksi põhimõisted sõltuvuspuude ehk sõltuvushargmike kirjeldamisel on *tipp*, *kaar* ja *juur*. Sõltuvuspuus on *tipuks* (ingl *node*) lauses kasutatud sõnavorm; puudepangas, st tegeliku teksti märgendamisel ka lühend, sümbol või kirjavahemärk. Kahte tippu ühendab suunatud *kaar* (ingl *arc*), mille nimetus väljendab informatsiooni süntaktilise suhte ehk süntaktilise funktsiooni kohta. *Juur* ehk *juurtipp* (ingl *root*, *top node*) on lause kõige kõrgem ülemus, st selline tipp, millel endal ülemus puudub. Juurtipp võib olla tegelik, st üks lauset moodustavatest sõnavormidest, või virtuaalne, millele allub lause tegelik juurtipp.

Sõltuvushargmik ei tohi sisaldada tsükleid, st teekonnal juurtipust mistahes terminaalse tipuni (lauses osaleva sõnavormini) peavad kõik vahepealsed tipud esinema täpselt ühe korra. Tsüklite esinemisel (siis ei saa rääkida enam matemaatilisest terminist *puu*, vaid pigem graafist) ei ole võimalik üheselt määrata, milline tipp on millise ülemuse alluv.

Tänapäevase sõltuvussüntaksiteooria rajajaks peetakse Lucien Tesnière'i, kellelt 1959. aastal ilmus postuumselt teos „Éléments de syntaxe structurale", millele selles artiklis viidatakse aastal 2015 ilmunud ingliskeelse tõlke kaudu (Tesnière 2015). Tesnière ise küll mõistet *sõltuvus* ei kasutanud ning väidetavalt on Tesnière'i teooria pigem sõltuvus- ja fraasistruktuuriesituse hübriid, sest selles kasutatakse eksotsentriliste konstruktsioonide, nt verbiahelate ja kaassõnaühendite analüüsil moodustajasarnast struktuuri, mida nimetatakse *jagatud üksuseks* (pr *nucleus dissocié)* (Tesnière 2015: 40; Osborne 2013: 269).

Vahemärkusena olgu öeldud, et Tesnière'i kui sõltuvusteooria alusepanija staatusega ei nõustu siiski mitte kõik. Näiteks väidab Igor Melčuk (1988: 3), et sõltuvussuhted kui süntaktilise struktuuri formaalse esitamise

viis olid Euroopa keeleteaduses, eriti klassikalises ja slaavi keeleteaduses tuntud varemgi ja Tesnière'i 1959. aastal ilmunud teos pole selle algus, vaid kulminatsioon.

Eesti keeles kirjutas sõltuvussüntaksist esimesena Huno Rätsep (1978: 9 jj), kes lähtus lihtlausete mallide koostamisel lause verbikesksuse põhimõttest.

Kui vaadata arvutilingvistika valdkonda, siis selles on viimasel ajal sõltuvussüntaksi populaarsus tunduvalt kasvanud (nt Nivre 2005; Osborne 2015: 1041), põhjuseks vajadus analüüsida keeruka morfoloogia ja vaba sõnajärjega keeli. Arvutilingvistikas loodavad loomuliku keele automaatsed sõltuvussüntaktilised analüsaatorid (parserid) ning ka arvutilingvistilistel eesmärkidel märgendatud puudepangad esitavad enamasti lause sõltuvusesituse, aga ei rakenda ühtegi konkreetset sõltuvussüntaktilist keeleteooriat, kuigi keeleteaduses on sõltuvussüntaksil põhinevaid süntaksiteooriaid mitmeid: *meaning-text theory* (Melčuk 1988), *word grammar* (Hudson 1984), *functional generative description* (Sgall jt 1986) jt. Erandina võiks siinkohal mainida Praha sõltuvuspuude panka (Prague Dependency Treebank), millest tuleb lähemalt juttu osas 3.

Samas on eri sõltuvussüntaktiliste parserite või sõltuvuspuude pankade kasutatav süntaktiliste märgendite repertuaar ja nende rakendamise eeskiri muidugi erinevad.

3. Puudepangad

Selleks, et arendada loomulikku keelt töötlevaid programme, on vaja keelelisi andmeid; süntaksianalüsaatori arendamiseks on vaja suurel hulgal süntaktiliselt märgendatud tekste ehk korpuseid, mida nimetatakse ka puudepangaks.

Puudepankade märgendus jaguneb laias laastus kaheks: moodustajaid esitav fraasistruktuuri märgendus ja funktsionaalne ehk sõltuvusstruktuuri märgendus. Kuid loomulikult leidub ka nende hübriidvariante, lisaks semantilise märgendusega puudepanku.

Puudepankade loomisega hakati intensiivsemalt tegelema 1990ndatel. Kõige kuulsam ja üks vanemaid puudepanku on Penn Treebank[3] (Marcus jt 1993), mis oli algul klassikaline fraasistruktuuripuude pank.

[3] Kättesaadavad on selle hilisemad versioonid Treebank-2 (https://catalog.ldc.upenn.edu/LDC95T7) ja Treebank-3 (https://catalog.ldc.upenn.edu/LDC99T42).

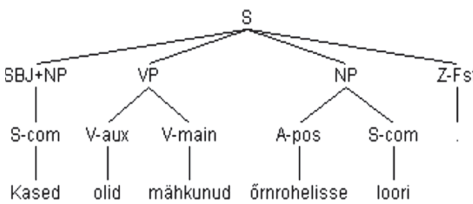
Tuntuim sõltuvuspuude pank on tšehhi keele Prague Dependency Treebank (Hajič 1998). See koosneb kolmest kihist: morfoloogiline kirjeldus, sõltuvussüntaktiline kirjeldus ja tekstogrammatiline ehk semantiline kirjeldus. Selle puudepanga märgendussüsteemi aluseks on Praha koolkonna arendatav keeleteooria *functional generative description* (Sgall jt 1986).

Tuntuim hübriidpank on saksa keele TIGER Treebank (Brants jt 2002), milles fraasistruktuuri tipud on süntaktilise funktsiooni suhtes märgendatud.

Paljud nüüdisaegsed puudepangad on osaliselt automaatselt märgendatud. Käsitsi märgendatud (või vähemalt täielikult lingvisti poolt kontrollitud) sõltuvuspuude pankadest on üks suuremaid Hamburgi sõltuvuspuude pank (Foth jt 2014), milles on ligi neli miljonit märgendatud ühikut (ingl *tokens*) u 262 000 lauses.

Eesti keele puudepanga loomiseks on tehtud mitu katset. Esimeseks võib lugeda 1994–1998 kitsenduste grammatika süntaksianalüsaatori arendamiseks loodud pindsüntaktiliselt märgendatud eesti keele korpust, mis sisaldas 20 000 sõna ja koosnes eri autorite ilukirjanduslikest tekstidest. Seda korpust suurendati hiljem 200 000 sõnani (Uibo 2004). Samas ei saa seda korpust pidada puudepangaks, sest igal sõnal oli küll süntaktilise funktsiooni märgend, kuid sõnadevahelisi alluvussuhteid ei esitatud.

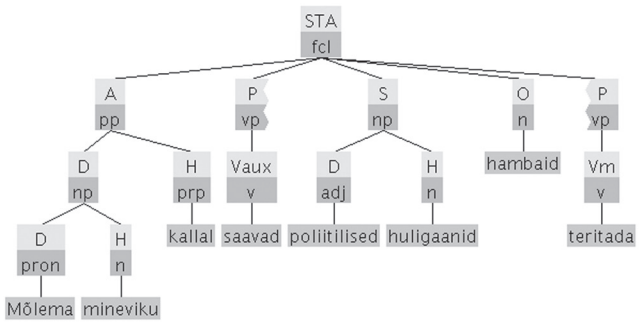
Eestikeelne osa Sofie paralleelsest puudepangast[4] loodi Põhjamaade puudepanga võrgustiku pilootprojekti käigus. See sisaldas esimest peatükki Jostein Gaarderi novellist „Sofie maailm“ (50–100 lauset). Eestikeelne tekst oli märgendatud Penni puudepanga stiilis fraasistruktuuripuudena (Uibo 2004).



Joonis 3. Eestikeelne lause Sofie paralleelses puudepangas. Ekraanitõmmis on pärit Joakim Nivre jt artiklist (2004)

[4] Sofie puudepank asub CLARINi Bergeni keskuse serveris INESS aadressil http://clarino.uib.no/iness/.

Järgmise katsena loodi Eesti keele pindsüntaktilise korpuse põhjal väike käsitsi märgendatud hübriidne puudepank Arborest[5] (150 lauset) (Bick jt 2004), mis sisaldab nii fraasistruktuuri kui ka süntaktiliste funktsioonide märgendeid. Seda puudepanka üritati genereerida ka automaatselt fraasistruktuurigrammatika abil, kuid tulemus ei olnud piisavalt korrektne. Sama meetodit kasutati liikumisverbidega lihtlausete korpuse (370 lauset) märgendamiseks: laused analüüsiti automaatselt ja parandati käsitsi. Saadud korpus teisendati TIGER-formaati, mis võimaldas seda edasi semantiliselt märgendada (Müürisep jt 2008). Arboresti loomisel kasutati Lõuna-Taani ülikoolis väljatöötatud vahendeid (süntaksianalüsaatorite mootoreid ja visualiseerimisserverit), mis olid osutunud sobivaks juba taani, portugali, prantsuse ja esperanto keele puudepankadele[6]. Kuid keeled on erinevad ning eesti keele automaatne fraasistruktuurianalüüs jäi liiga vigaseks, mistõttu oleks tulnud märgendus lisada käsitsi.



Joonis 4. Eestikeelse lause fraasistruktuuripuu puudepangas Arborest

Eesti keele lausete fraasistruktuuriesituse põhiline probleem on see, et kuna põhiverbi laiendite asukoht lauses on vaba, ei teki klassikalist verbifraasi. Eestikeelses tekstis on verbi liitvormi või perifrastilise verbi komponendid sageli üksteisest eraldatud muude moodustajatega, V2-reegli tõttu võib liitöeldise komponentide vahel olla ka subjekt. Probleemi üks lahendus oleks ristuvate kaarte lubamine (vt osa 6). Nende vältimiseks tuleks verbifraasi struktuuri esitada oluliselt lihtsustatuna, mille tulemusel on eestikeelse lause fraasistruktuuripuu tüüpiliselt väga madal. Joonisel

[5] Korpusepäringut saab teha aadressil http://corp.hum.sdu.dk/arborest.html.

[6] Teiste keelte korpused asuvad aadressil http://corp.hum.sdu.dk/.

4 esitatud, Arboresti korpusest pärineva lause *Mõlema mineviku kallal saavad poliitilised huligaanid hambaid teritada* analüüsil on ristuvaid kaari välditud ja hargmik on selle tulemusena madal ja väheinformatiivne: subjektifraas *poliitilised huligaanid* on perifrastilise öeldisverbi finiitsele komponendile *saavad* sama taseme laiend kui perifrastilise öeldise infiniitne osa *teritada.* Katkendlikku verbifraasi on joonisel kujutatud sakiliste kaartega.

Automaatsel analüüsil osutus väga keeruliseks fraasipiiride määramine. Kui pindsüntaktilise analüüsi korral oli kõige raskem eristada adverbiaali ja adverbiaalset atribuuti ning genitiivset objekti ja genitiivatribuuti, siis fraasistruktuurigrammatikas sõltub nendest otsustest hargmiku ülesehitus.

Seetõttu otsustati pärast mitut fraasistruktuuriesituse katset valida eesti keele puudepanga loomisel aluseks sõltuvussüntaktiline lähenemine.

4. Eesti keele sõltuvuspuude pank ja tema märgendus

Eesti keele sõltuvuspuude panga suurus on 400 000 sõna ning ta sisaldab ilukirjandus-, ajakirjandus- ja teadustekste, mis pärinevad tasakaalus korpusest[7]. Igale tekstisõnale on lisatud info tema algvormi, muutetunnuste ja -lõppude, sõnaliigikuuluvuse, temas avalduvate morfoloogiliste kategooriate, süntaktilise funktsiooni ning positsiooni kohta sõltuvuspuus. Märgendatud on poolautomaatselt, st esialgne morfoloogiline ja süntaktiline märgendus on tekstile lisatud automaatselt kitsenduste grammatika analüsaatori abil (Müürisep 2000) ning kahe sõltumatu märgendaja poolt käsitsi üle kontrollitud. Süntaktiliste funktsioonide puhul järgib EDT märgendussüsteem valdavalt eesti keele akadeemilist grammatikat (EKG II).

Joonis 5 esitab lause *Öö jooksul olid hundid kolm lammast maha murdnud* sellisena, nagu see on Eesti keele sõltuvuspuude pangas (EDT), ja joonis 6 sama lause visualiseeritud sõltuvuspuu.

Korpuselause algab lausealguse märgendiga <s> ning lõpeb lauselõpu märgendiga </s>. Tekstisõna ja ta analüüs on eraldi ridadel, analüüs algab taandele järgneva jutumärgistatud algvormiga. Selle järel järgnevad lühendile L muutelõpud ja -tunnused. Järgneb sõnaliigi märgend: S tähistab substantiivi, K adpositsiooni, V verbi, N numeraali ning D adverbi. Sõnaliigi märgendile järgneb osa sõnaliikide puhul täpsustav märgend:

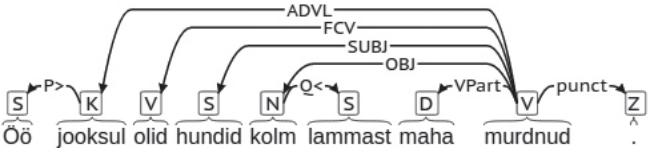
[7] http://www.cl.ut.ee/korpused/grammatikakorpus/.

S com tähistab üldnime, K post postpositsiooni, V aux abiverbi, N card kardinaalarvu ning V main põhiverbi. Kirjavahemärgid saavad sõnaliigi märgendi Z. Nendele järgnevad sõnavormis väljenduvate grammatiliste kategooriate märgendid: arvu tähistavad sg ja pl, käändeid nt gen ja part, verbivormis avalduvatest grammatilistest kategooriatest tähistab näites indic indikatiivi, impf imperfekti, ps3 kolmandat isikut, af afirmatiivi, partic partitsiipi ja past minevikku. Morfoloogilisele infole järgneb sümboliga @ algav süntaktilise funktsiooni tähistaja. @P> märgib paremal pool asuvale kaassõnale alluvat sõna, @ADVL adverbiaali, @FCV öeldise koosseisu kuuluvat finiitset abiverbi, @SUBJ subjekti, @OBJ objekti, @<Q vasakul pool asuvale kvantorile alluvat sõna, @Vpart ühendverbi afiksaaladverbilist komponenti ning @IMV perifrastilise öeldise infiniitset komponenti. Analüüsirida lõpeb sõltuvusstruktuuri esitavate märgenditega. Sümboli # järel on sõna järjekorranumber ning tähise -> järel tema ülemuse number. Lause juurtipu, näitelauses sõnavormi *murdnud* ülemuseks on 0.

Selline tekstikuju saadakse kitsenduste grammatika analüsaatori väljundina (Bick, Didriksen 2015).

```
"<s>"
"<Öö>"
        "öö" L0 S com sg gen @P> #1->2
"<jooksul>"
        "jooksul" L0 K post @ADVL #2->8
"<olid>"
        "ole" Lid V aux indic impf ps3 pl ps af @FCV #3->8
"<hundid>"
        "hunt" Ld S com pl nom @SUBJ #4->8
"<kolm>"
        "kolm" L0 N card sg nom l @OBJ #5->8
"<lammast>"
        "lammas" Lt S com sg part @<Q #6->5
"<maha>"
        "maha" L0 D @Vpart #7->8
"<murdnud>"
        "murd" Lnud V main partic past ps @IMV #8->0
"<.>"
        "." L0 Z Fst CLB #9->9
"</s>"
```

Joonis 5. EDT tekstiline kuju



Joonis 6. Joonisel 5 esitatud EDT lause visualiseering

5. Sõltuvussüntaktilise analüüsi probleemkohti

Sõltuvusstruktuur sobib väga hästi väljendama süntaktilist suhet verbi ja tema laiendi või täiendi ja tema põhisõna vahel, aga on ka keelelisi nähtusi, mis sõltuvusesituse seisukohalt on problemaatilisemad (vt nt Zeman jt 2012). Näiteks on raske koordinatsiooniseoses olevate sõnavormide, verbiahela liikmete või mitmesõnalise pärisnime osade vahel näha sõltuvussuhet. Järgnevalt vaadeldakse selliseid problemaatilisi keelendeid koos eesti keele puudepangas kasutatava lahendusega.

5.1. Kaassõna- ja kvantoriühendid

Nagu sedastab eesti keele akadeemiline grammatika EKG II (: 9, 137–148), on kaassõna- ja kvantoriühendite puhul tegemist eksotsentriliste struktuuridega, mis ei ole sellistena süntaktiliselt ekvivalentsed ei põhja ega laiendiga ning süntaktilist funktsiooni kannab selline ühend tervikuna.

Eksotsentriliste konstruktsioonide analüüsil saab ilmseks Tesnière'i süntaksikäsitluse hübriidne iseloom. Nimelt on tema teoorias kasutusel jagatud üksuse mõiste. Jagatud üksus koosneb vähemalt kahest sõnast, millest üks täidab strukturaalset funktsiooni, teine semantilist. (Tesnière 2015: 40). Kaassõnafraasi käsitleski Tesnière kui jagatud üksust, mille sees adpositsioon allub substantiivile ja on grammatiliseks vahendiks, mis teisendab nimisõna teise sõnaliiki kuuluvaks (Tesnière 2015: 368; Osborne 2013: 265).

Puhta sõltuvussüntaksi põhimõtteid järgivas EDT-s peab kaassõnafraasis emb-kumb sõna olema märgendatud teise sõna ülemusena. Ühelt poolt on kaassõnakonstruktsiooni kuuluval nimisõnal õigus olla ülemus, kuna tema annab konstruktsioonile põhilise tähendussisu ja kaassõna toimib käändelõpu moodi, kandes grammatilist tähendust. Teiselt poolt sõltub nimisõna kääne kaassõnast rektsiooniliselt, mis on jällegi ülemuslik

omadus. EKG II järgi on kaassõnafraasi põhjaks kaassõna, kuna käändsõna vorm on tingitud kaassõna rektsioonist (EKG II: 137). Ka eesti keele sõltuvuspuude pangas on kaassõnaühendid märgendatud nii, et kaassõna on ülemus, millele substantiiv allub.

Sarnane probleem kerkib kvantorifraaside puhul: ka siin tuleb põhiline tähendussisu substantiivilt, mille kääne ja arv aga sõltuvad kvantorist: *viis poissi*, *palju poisse*, *klaas piima.* Analoogselt kaassõnaühendiga on EKG II-s (: 140) ning sellele toetuvas EDT märgendussüsteemis kvantoreid (*viis*, *palju*, *klaas*) käsitletud nendest vormiliselt sõltuvate substantiivide (*poissi*, *poisse*, *piima*) ülemustena.

5.2. Verbiahel

Mitmesõnalise öeldise (verbi liitvorm, perifrastiline verb, ahelverb) analüüsil kerkib küsimus, kas pidada ühendi peasõnaks ehk kõige kõrgemaks ülemuseks a) finiitset verbivormi (***olen*** *teinud*, ***pean*** *tegema*, ***hakkan*** *tegema*), mis ühildub subjektiga ja mille vorm (nt personaal *vs*. impersonaal või imperatiiv) määrab subjekti võimalikkuse lauses, või b) infiniitvormi, mis tegelikult on lause struktuuriliseks keskmeks ja määrab lause argumentstruktuuri (*olen* ***teinud***, *pean* ***tegema***, *hakkan* ***tegema***). Tekstides esineb muidugi verbi liitvormide, ahelverbide ja perifrastiliste verbide kõikvõimalikke kombinatsioone, nt *selle oleks pidanud välja mõtlema.*

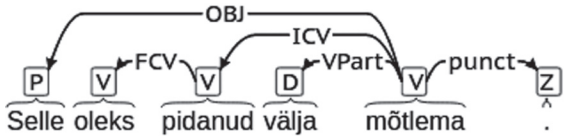
Tesnière nimetas ka abiverbist ja põhiverbist koosnevat struktuuri jagatud üksuseks (Tesnière 2015: 40), kuid enamik nüüdisaegseid sõltuvusteooriaid pooldab seisukohta, mille järgi finiitne (abi)verbi vorm on infiniidi ülemus (Osborne 2013: 264).

Võimalik on ka selline lahendus, et subjekt allub finiitsele verbivormile, millele allub ka infiniitne verbivorm, kuid verbi ülejäänud laiendid alluvad infiniitsele verbivormile (nt Järvinen, Tapanainen 1997).

EDT märgendusskeemi järgi on (osa)lause juurtipuks verbiahela põhitähendust kandev infiniitne verbivorm, millele allub ka subjekt. Sellise märgendamisotsuse taga on tõdemus, et lause argumentstruktuuri, lause võimalikud osalised ja nende keelelise vormistuse määrab ikkagi leksikaalne infiniitne verbivorm.

Joonisel 7 on näha, et lause *Selle oleks pidanud välja mõtlema* juurtipuks on mitmesõnalise öeldise infiniitne komponent *mõtlema*, millele

allub esmalt afiksaaladverb *välja* ning seejärel ahelana modaalverbi vorm *pidanud* ja selle kaudu liitaega väljendav *oleks*.



Joonis 7. Verbiahel EDT-s

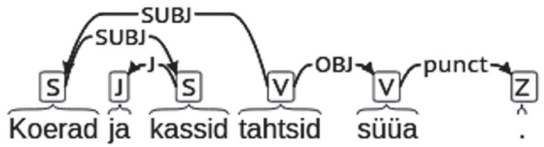
5.3. Koordinatsioon

Koordinatsioon kui samaväärsete elementide ühendus on nähtus, mille puhul kõik sõltuvussüntaktilised esitused tunduvad mõnevõrra kunstlikud. Levinumad lahendusviisid on järgmised.

Esiteks, kuulutada ülemuseks sidend või kirjavahemärk, millele koordineeritud elemendid alluvad. Teiseks, kuulutada ülemuseks üks koordineeritud element (esimene või viimane), millele alluvad ülejäänud koordineeritud elemendid; sidendid riputatakse selle lahenduse puhul tüüpiliselt järgneva koordineeritud elemendi külge. Kui koordineeritud elemente on rohkem kui kaks, on eelmise lahenduse puhul võimalik allutada kõik elemendid ühele või riputada nad riburada pidi üksteise külge.

Tesnière kasutas koordineeritud elementide vahelise suhte puhul terminit *junction*, st ta ei kasutanud koordinatsiooni analüüsiks sõltuvussuhet; ka on tema lähenemises koordinatsiooni puhul lubatud ühele sõnavormile määrata kaks ülemust, mis on omavahel koordinatsiooniseoses (Tesnière 2015: 342–344).

Eesti keele sõltuvuspuude pangas on koordinatsiooniseoses olevate sõnavormide puhul ülemuseks esimene koordineeritud element ning kui koordineeritud elemente on rohkem, märgendatakse nad ahelana, st kolmanda koordineeritud elemendi ülemus on teine element, mitte esimene.



Joonis 8. Koordinatsiooniseoses olevate subjektidega lause EDT-s

Joonis 9. Koordineeritud öeldisverbidega lause EDT-s

Joonis 10. Koordineeritud osalausetega lause EDT-s

Koordinatsiooniseoses olevate subjektidega lause on esitatud joonisel 8. Siit nähtub, et koordineeritud subjektide ülemus, öeldisverb *tahtsid*, on esitatud esimese koordineeritud üksuse ülemusena ning teise subjekti *kassid* ülemuseks on esimene subjekt *koerad*. Joonistelt 9 ja 10 on aga näha, et probleemne on koordineeritud öeldiste ja koordineeritud osalausete eristamine, sest mõlemal juhul märgendatakse esimene öeldisverb teise ülemuseks. Samuti näeme jooniselt 9, et ühe ülemuse põhimõtte tõttu pole võimalik esitada laiendi jagamist, st ei ole võimalik näidata, et adverbiaal *köögis* laiendab mõlemaid koordineeritud öeldisi *keetis* ja *küpsetas*.

5.4. Ellips

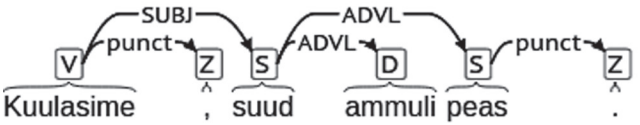
Sõltuvussüntaktilise analüüsi jaoks tekitab raskusi kontekstiellips, mille puhul on lünk ehk ellips tarindis, mis on oma vormilt, positsioonilt ja sisult sarnane lähikontekstis paikneva lüngata tarindiga ning lüngaga konstruktsioon täidetakse analoogia põhjal. Markeerimata rinnastusele iseloomulik rindliikmete süntaktilis-semantiline sarnasus ehk parallelism soodustab sellise ellipsi esinemist ning tavalisim on öeldise või aluse nullväljendus. (EKG II: 223–224)

Kuna sõltuvussüntaks on oma olemuselt verbikeskne, on raskeim probleem öeldise väljajätt. See on küllaltki tavaline koordineeritud osalausete puhul, millest teises on öeldis samasuse alusel kustutatud, nt *Tüdrukud mängisid nukkudega ja poisid autodega.* Sellises struktuuris puudub teises osalauses süntaktiline kese, finiitne verbivorm, millele subjekt *poisid* ja adverbiaal *autodega* peaksid alluma. Võimalikud lahendused on järgmised.

Esimene lahendus on analüüsida sõnavormi *poisid* sõnavormiga *tüdrukud* koordinatsiooniseoses olevana ning sõnavormi *autodega* sõnavormi *nukkudega* koordineeritud elemendina. See lahendus on praegu kasutusel ka eesti keele sõltuvuspuude pangas, vt joonis 11.

Teine võimalus on seada üks alluv elliptilise ülemuse asendajaks; verbi väljajätu puhul seega määrata osalause juurtipuks selle osalause subjekt. Mõnevõrra ebajärjekindlalt on EDT-s selliselt märgendatud ellipsiga sarnanev nähtus, absoluuttarind (nt *suu ammuli peas)*, mille puhul on ülemuseks see tarindi osaline, mille kohta EKG II (: 271) kasutab väljendit *subjektisarnane element* (näites *suu*), vt joonis 12.

Joonis 11. Elliptilise öeldisega osalause märgendamine EDT-s

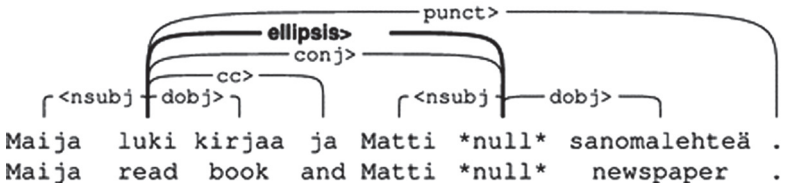


Joonis 12. Absoluuttarindi märgendamine EDT-s

Veel üks võimalik lahendus on nn nulltipu (ingl *null node*) kasutamine, mille abil taastatakse väljajäetud sõna. Sellise lahenduse puudus on see, et rikutakse või muudetakse lause algset süntaktilist struktuuri. Nulltippe kasutavates puudepankades on sageli mitu märgenduskihti ja elliptiliste konstruktsioonide märgendamiseks kasutatakse nulltippe teises kihis. Kõrvalmärkusena olgu öeldud, et puudepankade märgendamisel kasutataksegi üha enam mitut analüüsikihti: esimeses kihis esitatakse puhtalt süntaktiline informatsioon, nulltippe ei kasutata, mitteprojektiivsust välditakse niipalju kui võimalik. Teises kihis esitatakse semantilisem info, kasutatakse nulltippe, lubatakse ühele alluvale mitut ülemust, lõdvendatakse projektiivsuse nõuet jm.

Kahe märgenduskihi ideed näitlikustab Katri Haverineni jt artiklist (2014) laenatud joonis 13. Seal on näha, et rinnastusseoses olevatest osalausetest teises on korduv öeldisverb elliptiline ning lause

sõltuvusstruktuuri märgendamisel on selle asemel lausesse sisestatud nulltipp *null*, mille ülemuseks on esimese osalause verbivorm *luki* kahe süntaktilise sõltuvussuhte – ellipsi (*ellipsis*) ja konjunktsiooni (*conj*) kaudu. Nulltipp ise on kahe sõnavormi – subjekti (*nsubj*) *Matti* ja objekti (*dobj*) *sanomalehteä* ülemuseks.



Joonis 13. Taastatud elliptilise öeldisverbiga lause koos kahe märgenduskihiga korpuses Turku Dependency Treebank. Joonis pärineb Katri Haverineni jt artiklist (2014)

Elliptiliste konstruktsioonide analüüs ja märgendamine ongi siinses osas loetletud probleemidest kõige keerukam ja vajab kriitilist ülevaatamist ka eesti keele sõltuvuspuude pangas. Hea ülevaate elliptiliste konstruktsioonide esitamisest eri puudepankades annab Jan Hajič jt (2015).

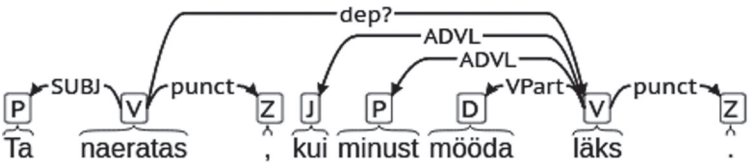
5.5. Osalausetevahelised suhted

Kirjaliku keele, eriti ajakirjandus- ja teaduskeele laused kujutavad endast tüüpiliselt omavahel mitmesuguste süntaktiliste suhetega seotud osalausete võrgustikku. Siiamaani on juttu olnud ühe osalause sees olevate süntaktiliste suhete sõltuvussüntaktilisest analüüsist, kuid lause täieliku süntaktilise struktuuri esitamisel on vaja omavahel siduda ka osalaused ning nendele siduvatele sõltuvussuhetele anda nimed.

Eesti keele sõltuvuspuude pangas on see süntaktilise struktuuri esitamise osa lahendatud vähimate vahenditega, st osalaused on omavahel seotud nende keskmeks olevate finiitverbide kaudu, mis tähendab, et pealause öeldisverb on kõrvallause öeldisverbi ülemuseks. Koordineeritud osalaused on esitatud samal viisil, st esimese koordineeritud osalause finiitne öeldisverb on järgneva koordineeritud osalause finiitse öeldisverbi ülemuseks. Ainsaks erandiks siin on relatiivkõrvallaused, mille (st mille öeldisverbi) ülemuseks on käändsõna, mida kõrvallause laiendab.

Osalausetevahelistel suhetel ei ole EDT-s nimesid (süntaktilise suhte märgendeid), mis on selle süsteemi üks põhipuudusi. See ei kehti sisestatud struktuuride ehk lauselühendite kohta, mille keskmeks olev infiniitne verbivorm on süntaktilise funktsiooni suhtes märgendatud.

Joonistel 14 ja 15 on esitatud kõrvallausega (*Ta naeratas, kui minust mööda läks*.) ja lauselühendiga (*Ta naeratas minust mööda minnes*.) struktuurid. Adverbiaallause *kui minust mööda läks* süntaktiline funktsioon pole EDT analüüsis väljendatud; sisestatud tarindi *minust mööda minnes* adverbiaalne funktsioon on aga esitatud tarindi peasõna *minnes* süntaktilise märgendina.



Joonis 14. Kõrvallause esitamine EDT-s

Joonis 15. Lauselühendi esitamine EDT-s

5.6. Nimed ja muud mitmesõnalised sisemise struktuurita üksused

Loomulikus keeles, eriti kirjalikus tekstis esineb palju sõnaühendeid, millel puudub sisemine süntaktiline struktuur, nii et tegu on pigem mitmesõnaliste leksikaalsete kui süntaktiliste üksustega. Ent puudepanga märgendamisel või automaatsel sõltuvussüntaktilisel analüüsil tuleb mingi analüüs anda igale tekstisõnale ja paratamatult kuulutada üks ühendi osaline teiste ülemuseks.

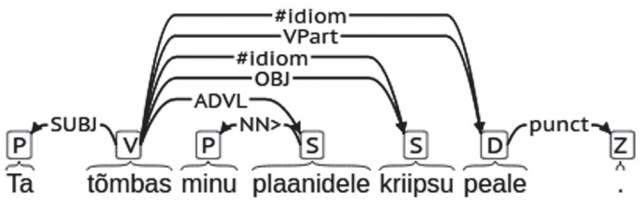
Sellised sõnaühendid on näiteks mitmest sõnast koosnevad pärisnimed ja tekstis esinevad võõrkeelsed fraasid, mis sageli on nimeüksused või pealkirjad. Kui pärisnime *Friedrich Reinhold Kreutzwald* puhul on perekonnanime *Kreutzwald* kuulutamine selle sõnaühendi ülemustipuks põhjendatav sellega, et pärisnime käänamisel liituvad käändelõpud just

tema külge, siis näiteks ansambli nime *Elephants from Neptune* puhul on puhtalt kokkuleppe küsimus, milline nendest võõrkeelsetest sõnavormidest märgendada ülejäänute ülemusena. Eesti keele sõltuvuspuude pangas on selliste sõnaühendite puhul ülemusena märgendatud ühendi viimane sõna.

Mõnes süntaksianalüüsi teoreetilises ja praktilises käsitluses on toodud ka idiomaatiliste mitmesõnaliste ühendite probleem, st on avaldatud arvamust, et otsese ja idiomaatilise tähendusega konstruktsioonid, millel on sama morfosüntaktiline struktuur, peaksid saama erineva süntaktilise analüüsi (nt Urešová jt 2013). Näiteks eestikeelsete väljendite *lõi palli väravasse* ja *lõi end lille* analüüsid peaks selle seisukoha järgi olema erinevad: esimeses sõnaühendis on *palli* objektina ja *väravasse* adverbiaalina verbivormi *lõi* alluvad, teisel juhul peaks idioomi esitama ilma sisestruktuurita üksusena, sõnavormid *end* ja *lille* alluvad samuti verbivormile *lõi*, kuid süntaktilise sõltuvuse nimetus peaks olema „idioom" või „mitmesõnaline leksikaalne üksus".

Alternatiivselt on võimalik väita, et süntaks tegeleb puhtalt lause vormiga, et ka väljendis *lõi end lille* on *end* süntaktiline objekt ja *lille* adverbiaal ning see väljend tuleks märgendada idioomina alles järgmisel, semantilisel tasandil.

EDT märgendamisel on asutudki seisukohale, et süntaktiline märgendus lähtugu eelkõige vormistruktuurist ning idioomide ja muude mitmesõnaliste püsiühendite märgendamine ei peaks toimuma mitte märgenduse süntaktilises põhikihis, vaid selle peale pandavas lisakihis. Joonisel 16 ongi esitatud idiomaatilist väljendit *tõmbas kriipsu peale* sisaldava lause sõltuvussüntaktiline esitus EDT praegusel kujul ning hüpoteetilise teise märgenduskihiga (# sümboliga tähistatud kaared).



Joonis 16. Idiomaatiline väljend EDT-s koos tulevikus lisatava teise märgenduskihiga

6. Projektiivsus

Projektiivsuse all mõeldakse sõltuvushargmiku sellist ülesehitust, mille puhul 1) lauset moodustavate sõnavormide vahelisi sõltuvusi esitavad kaared ei ristu; ja 2) kui ei kasutata virtuaalset juurtippu, siis ükski kaar ei kata lause juurtippu, milleks tavaliselt on öeldisverb. Keeleteaduslikes käsitlustes mööndakse, et mitteprojektiivseid struktuure esineb kõigis keeltes, kuigi nad on marginaalsed ja harvad (nt Melčuk 1988: 35 jj). Automaatsel süntaksianalüüsil on püütud vältida mitteprojektiivseid struktuure, sest pikka aega puudusid efektiivsed algoritmid nende analüüsimiseks (Nivre 2009). Samas, paljude keelte puudepankades on mitteprojektiivsete puude osakaal veerandi ja kolmandiku vahel, mitteprojektiivsete struktuuride rohkuse poolest paistavad silma nt hollandi, tšehhi ja saksa keel (Havelka 2007).

Mitteprojektiivseid lauseid või struktuure ühendab keeleteaduse seisukohast see, et nad kõik on millegi poolest markeeritud – emfaatiliselt, stilistiliselt, kommunikatiivselt (Melčuk 1988: 36).

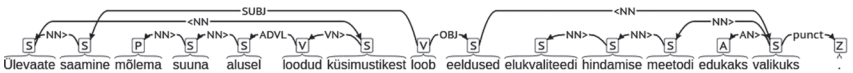
Seega on mitteprojektiivsust ehk ristuvaid kaari põhjustavate struktuuride uurimine oluline nii automaatse süntaksianalüüsi kui ka keeleteaduse seisukohast. Arvutilingvistika jaoks on oluline välja töötada selline sõltuvuste esitamise viis, mis minimeeriks mitteprojektiivsust, keeleteadust võiksid huvitada mitteprojektiivsed struktuurid kui neutraalsest või sagedasimast süntaktilisest konfiguratsioonist hälbivad nähtused.

Eesti keele sõltuvuspuude panga mitteprojektiivsetest lausetest on kirjutatud üliõpilastöö (Torga 2016), millest järgnevad näited pärinevad. Sellest tööst on näha, et Igor Melčuki väidetu – mitteprojektiivsed lausedon markeeritud – kehtib ka eesti keele kohta, kuid EDT-s on mitteprojektiivseid struktuure sisaldavate lausete osakaal märksa väiksem – vaid 3,1%. Oma osa nii suurde erinevusse eesti keele ja germaani keelte vahel võib anda ka see, et EDT puhul pole mitteprojektiivsete struktuuride hulka arvestatud kirjavahemärkidest ning lausega tegelikult mitteseotud elementidest (nt viited teadustekstis) põhjustatud kaarte ristumisi. Koos kirjavahemärkidest põhjustatud ristumistega kasvab mitteprojektiivseid struktuure sisaldavate lausete osakaal EDT-s 7,2%-ni. Ilmselt põhjustab erinevusi ka teistsugune märgendusskeem. Ka saksa keele puudepanga analüüsil on eri märgendusskeemide korral mitteprojektiivsete puude osakaalu erinevus üle kahe korra (12%–27%) (Foth jt 2014).

Teatud osa ristuvaid kaari saaks vältida teistsuguse süntaktilise analüüsiga. Näiteks põhjustab alati ristumise osalause piiril asuv postpositsioonifraas, kui järgnev kõrvallause allub relatiivlausena postpositsiooni juurde kuuluvale sõnale (joonis 17). Nagu juba öeldud, allub EDT märgendussüsteemis substantiiv adpositsioonile, kuid vastupidise analüüsi korral sellist ristumist ei tekiks.

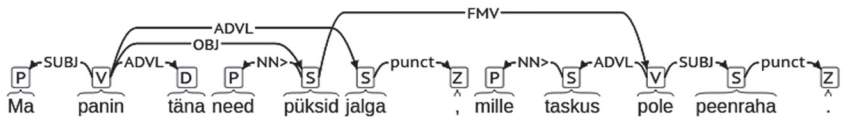
Joonis 17. Relatiivlause laiendab kaassõna juurde kuuluvat sõnavormi, mis ei asu osalausepiiril; tulemuseks mitteprojektiivne struktuur

Hea näide selle kohta, et ristuvad kaared võivad esile tuua lause halva struktuuri, on keerulised ja kohmakad nimisõnafraasid, mis on tihti saadud nominalisatsiooni tulemusel, nt *ülevaate saamine mõlema suuna alusel loodud küsimustikest* (joonis 18).



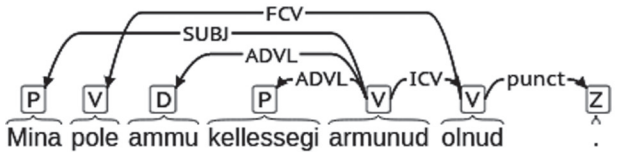
Joonis 18. Ristuvad kaared keerulise struktuuriga nominalisatsioonis

Nagu eespool mainitud, on EDT-s relatiivlause ülemuseks märgendatud see substantiiv, mida relatiivlause tegelikult laiendab. Juhul, kui selline substantiiv ei asu osalause piiril, on kaarte ristumine vältimatu, v.a. muidugi juhul, kui selle substantiivi ja osalausepiiri vahel on ainult selle substantiivi alluvad. Sellisest struktuurist põhjustatud kaarte ristumist esitab joonis 19.



Joonis 19. Relatiivlause laiendab substantiivi, mis ei asu osalause piiril; tulemuseks ristuvad kaared

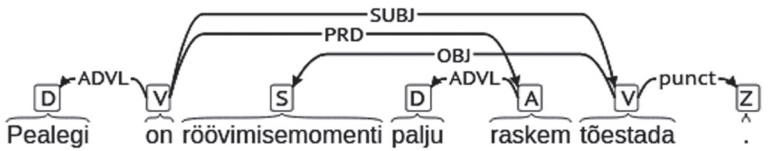
Suure osa kaarte ristumisest põhjustab verbiahela sõnajärg. Kuna mitmesõnaline öeldis märgendatakse ahelana ning öeldisverbi laiendid (sh subjekt) alluvad infiniitsele komponendile, põhjustab verbiahela komponentide sõnajärje muutus kohe ka kaarte ristumise. Joonisel 20 esitatud lauses *Mina pole ammu kellessegi armunud olnud* on rõhutamise tõttu muudetud verbifraasi neutraalset sõnajärge *pole olnud armunud*, tagajärjeks mitteprojektiivne struktuur.



Joonis 20. Verbiahela sõnajärje põhjustatud mitteprojektiivne lause EDT-s

Eriti aldid ristuvaid kaari põhjustama on *da*-infinitiiviga tarindid, mille analüüsil EDT-s on järgitud EKG II (: 232–274) käsitlust.

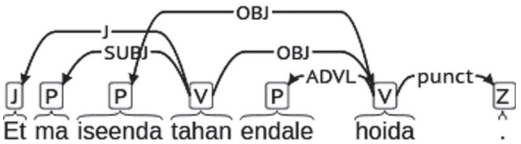
da-infinitiivne subjekt on mitteprojektiivsete lausete hulgas eriti sagedane: Liisi Torga analüüsitud 756 mitteprojektiivsest lausest oli 153-s ristuvate kaarte põhjustajaks *da*-infinitiivne subjekt (Torga 2016: 9). Tüüpiline on lause *Pealegi on röövimisemomenti palju raskem tõestada* joonisel 21: subjektiks analüüsitud *da*-infinitiivsele verbivormile *tõestada* allub objekt *röövimisemomenti*, mis asub aga teisel pool öeldisverbile *on* alluvat predikatiivi *raskem*.



Joonis 21. *da*-infinitiivse subjektiga mitteprojektiivne lause EDT-s

da-infiniitne objekt on samuti sage ristuvate kaarte põhjustaja, kuid mitte nii suurel määral kui *da*-infinitiivne subjekt. Tüüpiliselt asuvad sellises lauses *da*-infinitiiv (joonise 22 näites *hoida* ja temale alluv laiend (näites objekt *iseenda*) teine teisel pool öeldisverbi (näites *tahan*) ning ristuvad

da-infinitiivselt objektilt tema laiendile minev kaar ning öeldisverbi tema alluvaga ühendav kaar.



Joonis 22. *da*-infinitiivse objektiga mitteprojektiivne lause EDT-s

7. Kokkuvõte

Eesti keele lausete sõnajärje määrab suuresti infostruktuur (Lindström 2002), süntaktiliseks analüüsiks sobib sõltuvussüntaktiline lähenemine paremini kui fraasistruktuuriesitus. Samas pole ka sõltuvusanalüüsi rakendamine probleemivaba, eksotsentriliste konstruktsioonide, koordinatsiooni ning muudegi keelendite esitamisel pole ideaalseid lahendusi, valida on rohkem või vähem kunstlike vahel. Siiski võib loota, et eesti keele sõltuvuspuude pangal on pakkuda väärtuslikku keeleteaduslikku uurimismaterjali.

Eesti keele sõltuvuspuude panga loomise eesmärke oli luua keeletehnoloogiline ressurss, mis aitaks kaasa uute eesti keele automaatse analüüsi vahendite väljatöötamisele. Niivõrd suure hulga tekstide ühtlaseks märgendamiseks peab olema heas mõttes formalist ning mitmete keeleliselt põnevate nähtuste märgendamisel minema lihtsustamise teed. Märgendamise käigus tekkis palju küsimusi, millele ammendavat vastus pole veel leitud: miks on eesti keele puudepangas nii vähe mitteprojektiivseid puid võrreldes teiste vaba sõnajärjega keeltega; kuidas määrata kvantorifraasi; mis on ikkagi eksotsentriliste konstruktsioonide põhi; kas lauselühendite käsitlemine on olnud ühtlane; kuidas märgendada osalauseid – need on ainult osa lähemat uurimist vajavatest küsimustest. Õnneks pole tehtud valikud kivisse raiutud ning loodud puudepanga märgendust saab hiljem vajaduse korral täpsustada või muuta. Päris valmis eesti keele sõltuvuspuude pank veel ei ole – kriitilist ülevaatamist ootab ellipsite märgendamine ja on teisigi ühtlustamist vajavaid kohti.

Kui rääkida tulevikust, siis üheks edasiarenduseks on EDT teisendamine Universal Dependenciesi (UD)[8] kujule. UD projekti (McDonald

[8] http://universaldependencies.org/.

jt 2013; Nivre 2015) eesmärk on luua üldine ja keeletüpoloogiliselt põhjendatud märgendusskeem eri keelte morfoloogiliseks ja sõltuvussüntaktiliseks märgendamiseks ning selle baasil arendada keelest sõltumatuid sõltuvussüntaktilisi parsereid. Selle kohta vt täpsemalt Muischnek jt 2016.

EDT loogiliseks edasiarenduseks oleks liikumine puhtalt süntaksilt semantilisema märgenduskihi lisamise poole. Esimeseks sellealaseks katseks on anafooride, st samaviiteliste sõnade märgendamine, täpsemalt asesõnade sidumine nende viitealustega.

Kirjandus

Bick jt 2004 = Eckhard Bick, Heli Uibo, Kaili Müürisep. Arborest – a VISL-style treebank derived from an Estonian Constraint Grammar Corpus. – Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004), Tübingen, December 10–11, 2004. Eds. Sandra Kübler, Joakim Nivre, Erhard Hinrichs, Holger Wunsch. 1–14.

Bick, Eckhard, Tino Didriksen 2015. CG-3 – Beyond Classical Constraint Grammar. – Proceedings of NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. Linköping: LiU Electronic Press, 31–39.

Brants jt 2002 = Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, George Smith. The TIGER treebank. – Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol.

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

Foth jt 2014 = Kilian Foth, Arne Köhn, Niels Beuck, Wolfgang Menzel. Because size does matter: the Hamburg Dependency Treebank. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik. 2326–2333.

Hajič, Jan 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. – Issues of valency and meaning, 106–132.

Hajič jt 2015 = Jan Hajič, Eva Hajičova, Marie Mikulov, Jiří Mírovský, Jarmila Panevová, Daniel Zeman. Deletions and Node Reconstructions in a Dependency-Based Multilevel Annotation Scheme. – Computational Linguistics and Intelligent Text Processing. 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I. Ed. Alexander Gelbukh. LNCS 9041. Springer, 17–31. http://dx.doi.org/10.1007/978-3-319-18111-0_2.

Havelka, Jirí 2007. Beyond projectivity: multilingual evaluation of constraints and measures on non-projective structures. – Proceedings of ACL.

Conference: ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. Eds. John A. Carroll, Antal van den Bosch, Annie Zaenen. The Association for Computational Linguistics, 608–615.

Haverinen jt 2014 = Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, Filip Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. – Language Resources and Evaluation 48 (3), 493–531.

Hudson, Richard A. 1984. Word Grammar. Blackwell.

Järvinen, Timo, Pasi Tapanainen 1997. A dependency parser for English. Technical Report TR-1, Department of General Linguistics, University of Helsinki.

Lindström, Liina 2002. Veel kord subjekti ja predikaadi vastastikusest asendist laiendi järel. – Emakeele Seltsi aastaraamat 47 (2001). Tartu: Emakeele Selts, 87−106.

Marcus jt 1993 = Mitchell P. Marcus, Mary Ann Marcinkiewicz, Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. – Computational Linguistics 19 (2), 313–330.

McDonald jt 2013 = Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, Jungmee Lee. Universal Dependency Annotation for Multilingual Parsing. – 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), August 4–9 August, 2013, Sofia, Bulgaria. Proceedings of the conference, Volume 2: Short papers, 92–97.

Mel'čuk, Igor 1988. Dependency Syntax: Theory and Practice. State University of New York Press.

Muischnek jt 2016 = Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen. Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23–28, Portorož, Slovenia, 1558−1565.

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. (= Dissertationes mathematicae Universitatis Tartuensis 22.) Tartu Ülikooli matemaatikateaduskond. Tartu: Tartu Ülikooli Kirjastus.

Müürisep jt 2008 = Kaili Müürisep, Heili Orav, Haldur Õim, Kadri Vider, Neeme Kahusk, Piia Taremaa. From Syntax Trees in Estonian to Frame Semantics. – The Third Baltic Conference on Human Language Technologies, October 4-5, Kaunas, 211–218.

Nivre, Joakim 2005. Dependency grammar and dependency parsing. – Technical report. Växsjö University, School of Mathematics and Systems Engineering.

Nivre, Joakim 2009. Non-projective dependency parsing in expected linear time. – Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol. 1. Association for Computational Linguistics, 2009, 351–359.

Nivre, Joakim 2015. Towards a Universal Grammar for Natural Language Processing. – Computational Linguistics and Intelligent Text Processing. 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I. Ed. Alexander Gelbukh. LNCS 9041. Springer, 3–16.

Nivre jt 2004 = Joakim Nivre, Koenraad de Smedt, Martin Volk. Treebanking in Northern Europe: a white paper. – Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Ed. Henrik Holmboe. 97–113.

Osborne, Timothy 2013. A Look at Tesnière's Éléments through the Lens of Modern Syntactic Theory. – Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), 262–271.

Osborne, Timothy 2015. Dependency Grammar. Syntax – Theory and Analysis. An International Handbook. Vol 2, 1027–1045.

Rätsep, Huno 1978. Eesti keele lihtlause tüübid. Tallinn: Valgus.

Sgall jt 1986 = Petr Sgall, Eva Hajičová, Jarmila Panevová. The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Reidel Publishing Company, Dordrecht, Netherlands.

Zeman jt 2012 = Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Stepánek, Zdeněk Žabokrtský, Jan Hajič. HamleDT: To parse or not to parse? – LREC, 2735–2741.

Tesnière, Lucien 1959. Éléments de syntaxe structurale. Paris: Klincksieck.

Tesnière, Lucien 2015. Elements of Structural Syntax. Transl. Tymothy Osborne, Sylvain Kahane. John Benjamins. http://dx.doi.org/10.1075/z.185.

Torga, Liisi 2016. Mitte-projektiivsed laused eesti keele sõltuvuspuude pangas. Bakalaureusetöö. Käsikiri Tartu Ülikooli eesti ja üldkeeleteaduse instituudis.

Uibo, Heli 2004. Syntactically annotated corpora of Estonian. – The First Baltic Conference Human Language Technology – the Baltic Perspective, 21–22.

Urešová jt 2013 = Zdeňka Urešová, Jana Šindlerová, Eva Fučíková, Jan Hajič. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. – NAACL HLT 2013, 58–63.

The Estonian Dependency Treebank and its theoretical basis

KADRI MUISCHNEK, KAILI MÜÜRISEP

This article presents the Estonian Dependency Treebank (EDT) and discusses its language-theoretical basis. EDT contains ca 400,000 tokens of fiction, newspaper and science texts. Its syntactic annotation is based on principles of dependency syntax. Previous experiments with annotating Estonian sentences according to the principles of phrase structure syntax have shown that the resulting trees tend to be too shallow and thus do not encode the linguistic information in the best possible way. Therefore dependency-syntactic representation was chosen instead.

Dependency relations are efficient for encoding typical head-dependent relations like verb-argument or head-modifier but are not so suitable for analysing adpositional phrases, verbal chains, multi-word expressions or other constructs without clear internal syntactic relations. In such cases, there are arguments both for and against all possible solutions.

Keywords: dependency syntax, treebank, automatic syntax analysis, Estonian language

Kadri Muischnek
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
kadri.muischnek@ut.ee

Kaili Müürisep
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
kaili.muurisep@ut.ee