TARTU ÜLIKOOL

HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND

EESTI JA ÜLDKEELETEADUSE INSTITUUT

Liisi Torga

ELLIPTILISTE LAUSETE TUVASTAMINE JA MÄRGENDAMINE EESTI KEELE UD PUUDEPANGAS

Magistritöö

Juhendaja dotsent Kadri Muischnek

Tartu 2018

Sisukord

Sissejuhatus

Magistritöö eesmärgiks on anda ülevaade elliptiliste lausete esinemisest eesti keele *Universal Dependencies'* (UD) kujul puudepangas ning esitada nende tuvastamise ja märgendamise meetodid, vahendid ja tulemused. Sealjuures tutvustatakse ellipsi teoreetilist käsitlust, puudepanku, UD märgendusreegleid ning seda, kuidas elliptilisi lauseid nende märgendusreeglite järgi märgendatakse.

Elliptiliste lausete märgendamine puudepangas on motiveeritud peamiselt sellest, et kui lauses on predikaadi väljajätt, mõjutab see oluliselt lause sõltuvussüntaktilist struktuuri, mille tõttu on raske lauseid automaatselt töödelda. See tähendab, et sõltuvussüntaktilise lähenemise järgi peetakse verbi lause keskmeks ja seega muudab öeldisverbi väljajätt (osa)lause raskesti analüüsitavaks: lause osad on olemas, aga puudub neid ühendav ülemus. Kui aga kasutada märgendusviisi, mis teeb elliptilised konstruktsioonid nähtavaks, on neid lauseid oluliselt kergem analüüsida.

Sellega seoses püstitatakse kolm praktilist eesmärki:

1. Tuvastada elliptilised laused eesti keele UD sõltuvuspuude pangas.
2. Märgendada nendes lausetes orvuks jäänud lauseliikmed eesmärgiga luua põhi tulevasele lausekujule, kus kunstlikult tuuakse sisse puuduv osalause pea ning orvuks jäänud lauseliikmed riputatakse selle külge.
3. Luua eesti keele UD puudepangast selline versioon, kus elliptilised konstruktsioonid on UD reeglite järgi esile tõstetud.

Selle tarbeks loodi elliptiliste lausete tuvastamisprogramm[1], mis lisaks lausete leidmisele märgendab need ümber nii, et nad vastavad UD elliptiliste konstruktsioonide märgendusreeglitele. Tööd tehes selgus, et UD kujul eesti keele sõltuvuspuude pank sisaldab vigu, mida raske automaatselt lahendada. Seega otsustati osa vigu parandada käsitsi, eesmärgiga saada võimalikult ühtlane ja korrektne korpus.

Esimeses peatükis antakse ülevaade ellipsi teooriast, teises tutvustatakse puudepanku ning projekti *Universal Dependencies*. Kolmas peatükk keskendub ellipsite märgendamisele eri puudepankade versioonides ning neljas tutvustab varasemaid tehtud töid sellel teemal. Viies peatükk tutvustab korpust, mida töös kasutatakse. Kuues peatükk annab ülevaate

[1] https://github.com/liisitor/UD-EDT_with_orphans/

elliptiliste lausete tuvastamisprotsessist ning seitsmes peatükk esitab saadud tulemused, kirjeldades põhjalikult, millised elliptilised lausetüübid eesti keele sõltuvuspuude pangas leiti. Kaheksandas peatükis antakse ülevaade orbude märgendamisest ning tulemustest. Üheksas peatükk on teoreetiline ning annab ülevaate, millised on võimalused ellipsi täiendava versiooni tegemiseks. Kümnes peatükk võtab kokku kõik olulise magistritöös.

1. Ellips

Ellips on mingi lauseelemendi (nt korduva predikaadi või subjekti) väljajätt, kusjuures väljajäetud elemendi lähtevormi on võimalik leida tekstilise või situatsioonilise konteksti põhjal. Väljajättelise elemendi saab lünka kirjutada nii, et lause tähendus ega süntaktiline struktuur ei muutu. Need kaks omadust on elliptilise konstruktsiooni olulisemateks tingimusteks. (Biber jt 1999: 156) Ellipsi põhifunktsiooniks on väljendusökonoomia – ta on süntaktiline vahend korduste vältimiseks (Hakulinen jt (2004): §1177).

Eristatakse kolme tüüpi ellipseid: grammatiline ellips, milles välja jäetud liiane element tuvastatakse lause grammatilisest infost, kontekstiellips, mille lähtevormi leiab samas lausest või lause ümbrusest ning situatsiooniellips, kus puuduva info leiab situatsioonist. (Erelt 2017*a*: 591)

Lause süntaktilise sõltuvusstruktuuri analüüsi seisukohalt on eriti problemaatilised need kontekstiellipsit sisaldavad laused, kus osalauses puudub predikaat. Eesti keele akadeemilise kirjelduse (Erelt 2017*a*: 592) järgi on eesti keeles võimalik kahte tüüpi predikaadi väljajätt: predikaatverbi ellips (ingl k *gapping*) ning predikaatverbi ja tema laiendite ellips ühendav-kaasavas ja vastandavas rindlauses. Predikaatverbi ellipsi puhul võib puuduv verb olla lähtevormist erinevas isiku- ja arvuvormis (1) ning tihti võib lisaks finiitverbile välja jätta ka verbi liitvormi ja ahelverbi infiniitse osise, samuti ühend- ja väljendverbide mitteverbaalse osise (2). Kui tegemist on adverbiaaliga algava lausega, võib välja jätta ka aluse (3). (Erelt 2017*a*: 598–599)

(1) *Vanemad sõid putru, laps ∅[=sõi] võileiba.*

(2) *Ma ei huvitunud temast ja tema ∅[=ei huvitunud] minust.*

(3) *Koolis ta sööb puuvilju, kodus ∅[=ta sööb] sooja toitu.*

Rindlauses on anafoorne finiitverbi ellips normiks, st rindlause järelosalauses tavaliselt ei korrata eelneva osalause finiitset öeldisverbi või öeldisverbi finiitset osa. Käesoleva töö seisukohalt on oluline tõdemus, et kustutada saab tavaliselt vaid järellaiendiga öeldisverbi, st lause lõpus paiknevat verbi ei saa välja jätta muudes lausetüüpides peale võrdsustava ja vastandava rindlause (Erelt 2017*a*: 598–599). See tähendab, et öeldisverbi ellips toob tavaliselt kaasa vähemalt kaks nn orbu – subjekti ja järellaiendi (sihitise, määruse, öeldistäite).

Öeldisverbi ja tema laiendite, sageli kogu öeldisrühma ellips esineb ühendav-kaasavas ning vastandavas rindlauses. Sellele lausetüübile on iseloomulik fokusseerivate partiklite kasutamine. Ühendavas konstruktsioonis on nendeks partikliteks nt *samuti*, *ka* ning eituse puhul lisandub ka *mitte*. Vastanduse korral on üks lause pool jaatav, teine eitav, milles eituse väljendamiseks kasutatakse partiklit *mitte* (4), jaatuse puhul partiklit *küll* (5). (Erelt 2017*a*: 599–601)

(4) *Ma ei taha joosta, Kati samuti mitte ø[=ei taha joosta].*

(5) *Ma ei tahtnud magama minna, nüüd aga küll ø[=tahan magama minna].*

Piir ühelt poolt ühendav-kaasava ning vastandava rindlause vahel ning teiselt poolt lausemoodustajate vastandava rinnastuse vahel on hajus. Näiteks lauset

(6) *Ta ei ela mitte Tartus, vaid Tallinnas.*

käsitletakse akadeemilises eesti keele kirjelduses (Erelt 2017*c*: 611) kui kahest osalausest koosnevat rindlauset, kusjuures teine osalause on elliptiline. Käesoleva töö seisukohalt vaadeldakse seda aga kui lihtlauset, milles on kaks vastandatud rinnastusseoses olevat kohamäärust.

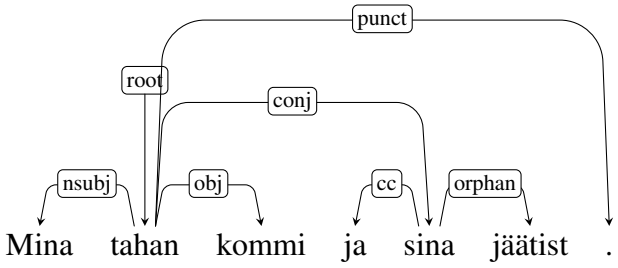
2. Puudepank

Puudepank on süntaktiliselt märgendatud korpus, mis tavaliselt on märgendatud kas moodustajaid esitava fraasistruktuurina või lauset moodustavate sõnavormide süntaktilisi suhteid esitava sõltuvusstuktuurina. Lisaks sellele on osa puudepanku märgendatud ka kaht viisi kombineerides. (Muischnek ja Müürisep 2017*a*) Kuna magistritöös kasutatakse sõltuvussüntaktiliselt märgendatud puudepanka, siis kirjeldatakse edasi lähemalt sõltuvussüntaktilist märgendust.

Sõltuvussüntaktilise märgenduse kirjeldamiseks kasutatakse mõisteid tipp, kaar ja juur. Tipuks on sõna lauses. Kahte sõna ühendab kaar, mis näitab ülemus-alluvus suhet, kus üks sõna on ülemus ja teine alluv. Lause kõige kõrgemaks ülemuseks on tavaliselt öeldis, mida nimetatakse ka juureks või juurtipuks (inglise k *root*). Kuna juurtipp on kõige kõrgem, siis puudub tal ülemus. (Melćuk, 1988, : 23) Niimoodi moodustub puustruktuur, kus juurele ehk öeldisele alluvad subjekt, objekt ja muud öeldise juurde kuuluvad laiendid.

Sõltuvussüntaktiliselt märgendatud puudepank on heaks keeleandmete allikaks erinevate keeletöötlust vajavate ülesannete lahendamiseks. Kui algseks puudepankade loomise motivaatoriks oli keeleteadlaste soov empiiriliselt uurida keele süntaktilisi struktuure, siis hilisemalt on motivatsiooniks saanud puudepankade kasutamine arvutilingvistilises uurimis- ning arendustöös ja arvutilingvistika vajadused on muutunud uute puudepankade loomise peamiseks motivaatoriks. Puudepanku kasutatakse süntaktiliste parserite testmiseks ja masinõppel põhinevate parserite treenimiseks. (Nivre, 2008)

Joonisel 2.1 on kujutatud sõltuvussüntaktiliselt märgendatud lause, kus juureks on predikaat *tahan*, talle allub nii objekt *kommi* kui subjekt *mina*, samuti osalause ülemus *sina* ning punktuatsioonimärk. Teine osalause *sina jäätist* on elliptiline, st korduv predikaat *tahan* on välja jäetud, mille tõttu määratakse osalause kõrgeimaks ülemuseks subjekt *sina*. Selleks, et näidata mittestandardse suhte olemasolu, riputatakse objekt *jäätist* subjekti külge kasutades spetsiaalsuhet *orphan*. Lisaks allub subjektile ka sidend *ja*.



Joonis 2.1: Sõltuvussüntaktiliselt märgendatud lause

2.1. Projekt *Universal Dependencies*

Universal Dependencies (UD) on projekt, mille eesmärgiks on luua ühtselt märgendatud puudepangad paljude keelte jaoks. See tähendab, et samad süntaktilised struktuurid märgendatakse erinevates keeltes ühtemoodi ja erinevad struktuurid erinevalt. (Nivre jt 2016) Selliselt märgendatud puudepangad võimaldavad arendada keelest sõltumatuid või paljude keelte jaoks sobivaid parsereid jm süntaksipõhiseid keeletöötlusvahendeid ja aidata kaasa ka keeletüpoloogilisele uurimistööle.

UD morfoloogilised ja süntaktilised reeglid peavad täitma järgmisi tingimusi[1]:

1. UD märgendussüsteem peab olema piisav iga keele lingvistilise analüüsi jaoks.
2. UD märgendussüsteem peab sobima lingvistilise tüpoloogia kirjeldamiseks, st peab võimaldama leida sarnasusi (ja erinevusi) keelte ja keelegruppide vahel.
3. UD märgendussüsteem peab olema sobiv inimmärgendaja jaoks.
4. UD märgendussüsteem peab olema sobiv täpseks parsimiseks.
5. UD märgendussüsteem peab olema arusaadav mittelingvistile, nt keeleõppurile või inimesele, kellel on vaja teha keeletöötlust (eeldab, et on arusaadav disain, arusaadav terminoloogia).
6. UD märgendussüsteem peab toetama erinevaid keeletöötluse ülesandeid (nt masintõlge).

Projekti *Universal Dependencies* kõige uuemaks versiooniks on 2.2, mis avalikustati 1. juulil 2018. Selles versioonis on loodud 122 puudepanka 71 keeles.[2] Paljudel keeltel on

[1] http://universaldependencies.org/introduction.html
[2] http://universaldependencies.org/

mitu puudepanka. Lisaks tavalisele kirjakeelele on loodud ka näiteks kõnekeele ja viipekeele sõltuvuspuude panku. Samuti on mõnel keelel mitu töörühma, kes on teinud sama keele jaoks erinevaid puudepanku.

Universal Dependencies on käigusolev projekt, see tähendab, et UD puudepankadest tulevad uued versioonid iga 6 kuu järel ja pidevalt käib ka märgenduspõhimõtete ja märgendusjuhendi muutmine, parandamine, täpsustamine ja täiendamine. Seetõttu ei saa UD puudepanku käsitleda kui lõplikke või täielikke.

Projekti *Universal Dependencies* märgendusreeglid põhinevad mitmetel varasematel projektidel: Stanfordi (universaalsetel) sõltuvusreeglitel, Google'i universaalsetel sõnaliigi märgenditel ja Interseti interlingua morfosüntaktilistel märgenditel. UD märgendusreeglite ideeks on sarnaste keelekonstruktsioonide märgendamine eri keeltes samasuguselt. Sellegipoolest lubavad need reeglid ka keelespetsiifilisi erandeid, kui need on vajalikud. (Nivre jt 2016)

UD versioonis 2 kasutatakse 37 süntaktilist suhet sõnadevahelise sõltuvuse näitamiseks.[3] Lisaks neile on olemas veel alamsuhted, mille abil näidatakse keelespetsiifilisi erisusi.[4] Keelespetsiifilisust võib edasi anda ka põhimärgenditega, nt märgendiga *clf* ehk *classifier*[5]. Kodulehel on võimalik näha ka erinevate konstruktsioonide märgendamisjuhendeid.[6]

UD kujul puudepangad on talletatud CoNLL-U formaadis tekstifailides, kus iga lause kohta on detailne info järgmisel kujul:

```
# sent_id = ilu_kivirahk_93
# text =  Esiteks oli siin väga soe.
1 Esiteks  esiteks ADV D _ 5 advmod _ _
2 oli olema AUX V Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
    |Voice=Act 5 cop _ _
3 siin siin ADV D _ 5 advmod _ _
4 väga väga ADV D _ 5 advmod _ _
5 soe soe ADJ A Case=Nom|Degree=Pos|Number=Sing 0 root _ SpaceAfter=No
6 . . PUNCT Z _ 5 punct _ _
```

Rida *# sent_id* tähistab lause failinime ning lause numbrit, rida *# text* aga lauset ennast. Kui mingisugune info puudub, siis tähistatakse vastav lünk alakriipsuga (_). (Nivre jt 2016) Tabel 2.1 kirjeldab lause ühe sõna informatsiooni:

[3] http://universaldependencies.org/u/dep/all.html
[4] http://universaldependencies.org/ext-dep-index.html
[5] http://universaldependencies.org/u/dep/clf.html
[6] http://universaldependencies.org/guidelines.html

Inglise keeles	Eesti keeles	Ühe sõna näide
ID	Järjekorranumber	2
FORM	Sõnavorm	oli
LEMMA	Lemma	olema
UPOS	Sõnaliik	AUX
XPOS	Keelespetsiifiline sõnaliik	V
FEATS	Morfoloogiline info	Mood=Ind|Number=Sing| Person=3|Tense=Past| VerbForm=Fin|Voice=Act
HEAD	Konkreetse sõna ülemus	5
DEPREL	Sõltuvussüntaktiline märgend	cop
DEPS	Konkreetse sõna alluvad	_
MISC	Muud kommentaarid	_

Tabel 2.1: Ühe sõna märgendusinfo

2.2. Projekti *Universal Dependencies* kujul eesti keele sõltuvuspuude pank

Projekti *Universal Dependencies* kujul eesti keele sõltuvuspuude pank saadi UD märgendusreegleid kasutades poolautomaatselt kitsenduste grammatikal (Karlsson jt 1995) põhinevast eesti keele sõltuvuspuude pangast (*Estonian Dependency Treebank* ehk EDT) (Müürisep, 2000). Eesti keele sõltuvuspuude pank koosneb umbes 400 000 sõnast (umbes 30 000 lausest), mis on saadud ilukirjandus-, ajakirjandus- ja teadustekstidest (Muischnek jt 2014).

Kuigi nii eesti keele sõltuvuspuude pank kui UD kujul sõltuvuspuude pank järgivad sõltuvusgrammatikale omaseid reegleid, siis teatud erinevused kahe puudepanga märgenduse vahel siiski on. Märgenduse erinevused tulenevad enamasti kolmest asjaolust:

1. UD-s on põhimõte, et suhted on eelkõige sisusõnade vahel (*primacy of content words*). EDT-s lähtuti rohkem formaalsetest või puhtsüntaktilistest, mitte semantilistest kriteeriumitest. Nt kaassõnaühendis on UD-s ülemus nimisõna, sest tema kannab ühendi põhilist tähendussisu, EDT-s aga kaassõna, sest tema määrab nimisõna käände.

2. EDT-s on osalausete omavaheline sidumine lahendatud askeetlikult: kõrvallause öeldis allub pealause öeldisele, aga süntaktiline märgend on ikka öeldise oma. UD-s allub kõrvallause juurtipp pealause sellele elemendile, mida ta tegelikult laiendab, ja märgend näitab kõrvallause kui terviku süntaktilist rolli pealause või mõne tema elemendi suhtes.

3. Märgendite repertuaaril on erinev keeleteoreetiline ja lingvistiline tagapõhi. UD märgendisüsteem tuleb Stanfordi märgendisüsteemist, mis jällegi on suurel määral mõjutatud LFG-st (*Lexical Functional Grammar*).

Näiteks erinevad sõnaliigi- ja süntaksimärgendid, verbiahelate ja koordinatsiooni märgendus jmt. Kui morfoloogiliste märgendite üleviimine UD kujule on üsnagi lihtne, siis UD süntaktilised reeglid nõuavad tihti lisaks märgendi muutmisele ka puustruktuuri muutmist. Ka kasutab UD kuju oluliselt rohkem süntaktilise suhte märgendeid kui EDT, näiteks võivad EDT subjektimärgendile *@SUBJ* vastata olenevalt lausest UD märgendussüsteemis järgmised märgendid: *nsubj, nsubj:cop, csubj, nummod, csubj:cop, advmod:quant, nmod, compound, advcl, case, cop, acl:relcl, ccomp*. (Muischnek jt 2016)

Sellest tulenevalt erinevad kaks puudepanka ka mahult – hetkel pole ametlikku UD versiooni veel kõiki EDT tekste viidud. Seda seetõttu, et kuigi suures osas saab failid automaatselt ühest versioonist teise viia, siis teatud konstruktsioonides tuleb märgendid käsitsi üle kontrollida ning vajadusel parandada, nii on ka ellipsi märgendusega.

3. Ellipsi märgendus puudepangas

Elliptilise konstruktsiooni esiletoomine lausetes on motiveeritud sellest, et informatsioon oleks kergesti kättesaadav ja üheselt mõistetav. Kuigi sellised konstruktsioonid ei ole väga sagedased, kasutatakse neid tihti faktide edastamiseks. Info ekstraheerimine ja semantiline parsimine aga sageli kasutavad just sellist faktilist infot, vajades elliptiliste konstruktsioonide esiletõstmist. (Schuster jt 2018)

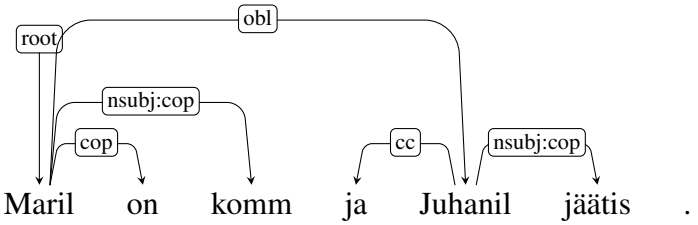
Seetõttu on otsustatud projektis *Universal Dependencies* märgendada ellipsid lähtudes järgmistest tingimustest[1]:

1. Kui väljajättelisel elemendil puuduvad alluvad, siis ei tehta midagi.
2. Kui väljajättelisel elemendil on alluvad, siis üks neist saab kõrgeimaks ülemuseks, millele alluvad teised konstruktsiooni liikmed.
3. Kui välja on jäetud predikaat ning kõrgeim ülemus on tuumargument (subjekt, objekt, seotud infiniittarindiline või osalauseline laiend vmt), siis kasutatakse märgendit *orphan*, et näidata mittestandardset sõltuvussuhet.

Puudepankade märgenduse seisukohalt on oluline just selline ellips, mille tulemusena tekib orb (ingl k *orphan*), st sõna, mille ülemus on kustutatud. Eesti keele puudepanga seisukohalt on oluline eelkõige öeldise ellips, sest sõltuvussüntaks on verbikeskne ja öeldisverbi puudumine on märgenduse seisukohalt suurim probleem: kogu osalause ülemus puudub. Kui puudub alluv, pole märgenduse seisukohalt probleeme.

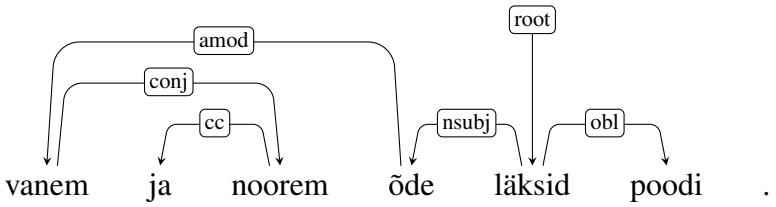
Oluline on märgata, et UD versioonis 2 märgendatakse eesti keele kõige sagedasem verb *olema* koopulana. See tähendab, et osalause kõrgeimaks ülemuseks on mingi muu lause mitteverbaalne osis, millele *olema*-verb allub. (Muischnek ja Müürisep 2017*b*) Seega ei tekita verbi *olema* väljajätt UD 2 märgendusreeglite järgi orbe. Joonisel 3.1 on tegemist omaja-lausega, kus kõrgeimaks ülemuseks on *Maril*. *Olema*-verb märgendatakse tema külge suhte *cop* abil, samuti subjekt *komm*, mille märgendiks saab *nsubj:cop* eristamaks, et tegu on koopulalause subjektiga.

[1] http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis



Joonis 3.1: Koopulalause märgendamine UD versiooni 2 järgi

Sage on ka nimi- või omadussõnafraasis korduva põhja väljajätt, nt *vanem ∅[=õde] ja noorem õde* (vt joonis 3.2). Selline konstruktsioon lahendatakse UD märgendusreeglite järgi nii, et adjektiivid koordineeritakse ehk rinnastatakse ja riputatakse põhja külge, mille tõttu orbe ei teki. (Droganova jt 2017)



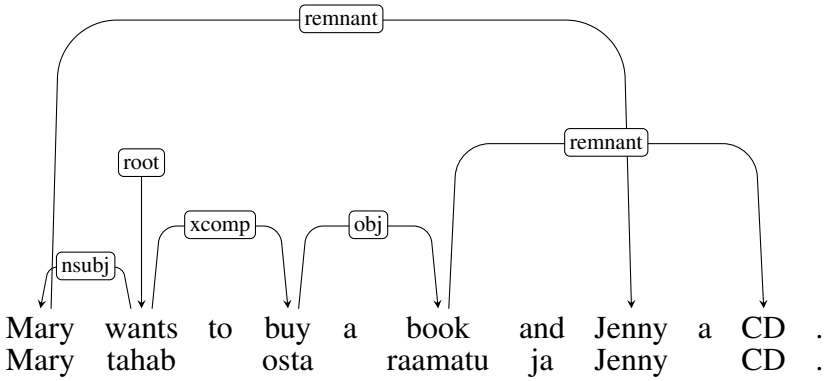
Joonis 3.2: Lauses jäetakse välja korduv põhi *õde*

3.1. Ellipsi märgendamine erinevates UD versioonides

Projekti *Universal Dependencies* märgendusreeglitest on mitu versiooni – UD versioon 1 ja UD versioon 2. Ellipsi ehk lause mingi elemendi väljajätu märgendus kahes versioonis on üsnagi erinev. UD versioonis 1 kasutati elliptiliste konstruktsioonide puhul süntaksimärgendit *remnant*.[2] Sellise märgendusviisiga üritati elliptiline konstruktsioon teha nähtavaks, et verbi tuvastamine oleks kergem, kuid tihti tekitas selline märgendusviis mitteprojektiivsust ehk kaarte ristumisi. (Hajič jt 2015)

Nagu on näha joonisel 3.3, siis süntaksimärgendi *remnant* saavad need lauseliikmed, mis vastavad eelneva osalause samaväärsetele liikmetele, st lauseliikmed elliptilises konstruktsioonis riputatakse eelneva osalause samaväärsete lauseliikmete külge ahelana.

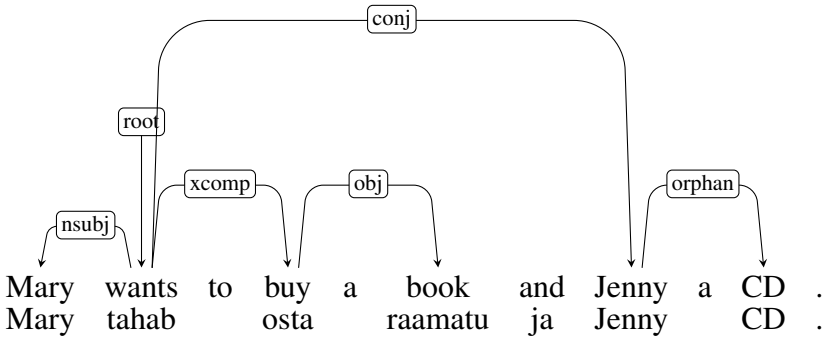
[2]http://universaldependencies.org/docs/u/dep/remnant.html



Joonis 3.3: UD versiooni 1 elliptilise lause märgendus

UD versioonis 2 on *remnant* ära kaotatud ning kasutatakse hoopiski märgendit *orphan* [3], mis võimaldab analüüsida elliptilisi lauseid nii, et lausete tähendus ei muutuks ning mitteprojektiivsust ehk kaarte ristumist esineks vähem. See tähendab, et üks elliptilise konstruktsiooni element määratakse kõrgeimaks ülemuseks ning teised olulised argumendid alluvad süntaksimärgendi *orphan* abil sellele.

Selline spetsiaalsuhe võimaldab kergemini lausest tuvastada lünga. Kui see puuduks, siis võidaks mõne keele puhul lauset valesti tõlgendada, nt keelte puhul, kus lausetes koopulaverbi ei kasutata. (Schuster jt 2017) Ülemus valitakse vastavalt järgmisele hierarhiale: *nsubj* > *obj* > *iobj* > *obl* > *advmod* > *csubj* > *xcomp* > *ccomp* > *advcl*.[3] Nagu on näha joonisel 3.4, siis teise osalause subjekt *Jenny* allub eelneva osalause verbile (selle asemel, et alluda eelneva osalause subjektile *Mary*) ning süntaksimärgendiga *orphan* on märgendatud objekt *CD*, mis on riputatud subjekti külge.



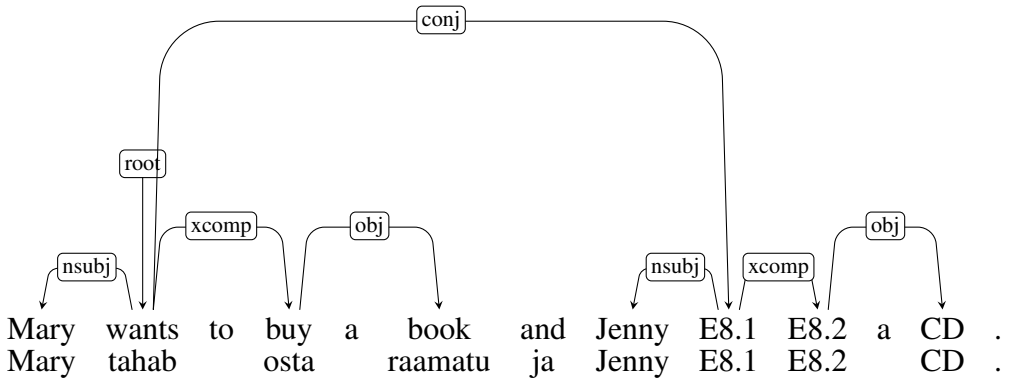
Joonis 3.4: UD versiooni 2 elliptilise lause märgendus

Uueks eesmärgiks on luua ellipsi täiendav sõltuvusmärgendus (*enhanced dependencies*), mille kujul planeeritakse märgendada UD puudepankade eraldi versioonid. See tähendab,

[3] http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis

et lausesse tuuakse sisse lisatipp ehk nulltipp, millega väljendatakse puuduvat infot.[4] Selle eesmärgiks on säilitada võimalikult palju süntaktilist informatsiooni, sest kasutades spetsiaalsuhet *orphan* kaob sõnade algne sõltuvussüntaktiline märgendus (*obj –> orphan* jne). Selle kujuga püütakse jälgida stiili, et kasutatakse nii palju tippe, et kaoksid ära orvud ja mitte rohkem (Schuster jt 2017). Nulltippu tähistatakse spetsiaalse indeksiga kujul *Ea.b*, kus *a* tähistab puuduvale sõnale eelnevat sõna ning *b* käitub loendurina. [5]

Joonisel 3.5 on näha, et puul on kaks lisatippu: *E8.1* tähistab esimese osalause verbi *wants* ja *E8.2 to buy*, kusjuures mõlema puhul arv *8* tähistab neile eelneva sõna *Jenny* indeksit ja vastavalt *1* ning *2* käituvad n-ö järjekorranumbrina.



Joonis 3.5: UD versiooni 2 elliptilise lause laiendatud märgendus

3.2. Ellipsi märgendamise statistika

Elliptiliste konstruktsioonide märgendamine erinevates UD puudepankades on seoses puudepankade arvu suurenemisega samuti suurenenud, kuid pigem rõhutakse kvaliteedile, mitte kvantiteedile (vt ka tabel 3.1). See tähendab, kui aja jooksul on vähemalt ühe *orphan*-märgendi arv protsentuaalselt kasvanud, siis orbude arv 100 lauses aga protsentuaalselt kahanenud (korpuste arv on peaaegu samaks jäänud). Pigem on püütud parandada olemasolevate elliptiliste konstruktsioonide märgendust, kui üritatud uusi konstruktsioone suures mahus märgendada.

[4] http://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis
[5] http://universaldependencies.org/v2/ellipsis.html#predicate-ellipsis-in-enhanced-ud-v2

	UDv1.4		**UDv2.0**		**UDv2.1**		**UDv2.2**	
	arv	%	arv	%	arv	%	arv	%
Puudepankade arv kokku	64	100%	70	100%	102	100%	122	100%
0 märgendit	41	64,06%	29	41,43%	33	32,35%	44	36,06%
Vähemalt 1 märgend	23	35,94%	41	58,57%	69	67,65%	78	63,94%
Vähemalt 100 lauses *orphan*-märgend	12	18,75%	12	17,14%	15	14,71%	13	10,7%

Tabel 3.1: Tabel, mis näitab, kui mitu puudepanka kasutavad vahendeid elliptiliste konstruktsioonide esiletõstmiseks (UDv1.4 vaadati märgendi *remnant* esinemist, UDv2.0, UDv2.1 ja UDv2.2 märgendi *orphan* esinemist)

Droganova ja Zeman (2017) andsid statistilise ülevaate elliptiliste lausete märgendamisest UD versioonis 2.0[6]. Nad tõid välja, et võrreldes versiooniga 1.4[7] on elliptiliste konstruktsioonide märgendamine kasvanud kahekordselt. UD versioonis 2.0 oli 41 puudepangal (58,57%) vähemalt 1 *orphan* suhe kasutusel, ainult 12 puudepangal (17,14%) oli vähemalt 100 lauses *orphan*-märgend. 29 puudepangal (41,43%) puudusid *orphan*-märgendid täielikult. (Droganova ja Zeman 2017)

Magistritöös vaadati ka uuemate versioonide orbude märgendust. UD versioonis 2.1[8] on kokku 102 puudepanka, millest 69 puudepangal (67,65%) on vähemalt üks *orphan* suhe märgitud. Küll aga on vähenema hakanud nende puudepankade protsent (14,71%), millel on vähemalt 100 lauses märgend ellipsi esiletoomiseks. UD versioonis 2.2[9] on kokku 122 puudepanka, millest 78 puudepangal (63,94%) on vähemalt üks *orphan* suhe märgitud. Veelgi vähenenud on nende puudepankade protsent (10,7%), kus on vähemalt 100 lauses *orphan*-märgend.

Tabel 3.2 illustreerib kuuekümne puudepanga *orphan*-märgendi sageduse suhet kõikidesse märgenditesse protsentuaalselt. Nagu näha, siis eesti keele UD puudepank on 57. kohal, mis kõikide puudepankade seisukohalt näitab seda, et eesti keele UD puudepank asub orbude sageduselt napilt esimeses pooles. Küll aga ei saa orbude sagedust päris otsesese kvaliteedinäiduna võtta, sest mõnes keeles võidaksegi kasutada ellipsit rohkem kui teistes. Samuti kasutatakse puudepankades eri tekstižanreid.

[6]UDv2.0 http://hdl.handle.net/11234/1-1983
[7]UDv1.4 http://hdl.handle.net/11234/1-1827
[8]UDv2.1 http://hdl.handle.net/11234/1-2515
[9]UDv2.2 http://hdl.handle.net/11234/1-2837

Puudepank	**arv**	**%**	**Puudepank**	**arv**	**%**
Gothic-PROIEL	222	0,401	Norwegian-Nynorsk	222	0,074
Old_Church_Slavonic	230	0,4	Komi_Zyrian-Lattice	1	0,072
Russian-Taiga	81	0,39	Turkish-PUD	12	0,071
Sanskrit-UFAL	7	0,38	Buryat-BDT	7	0,069
Czech-CAC	1830	0,37	Norwegian-Bokmaal	215	0,069
Latin-PROIEL	712	0,356	Czech-CLTT	22	0,062
Ancient_GreekL	753	0,352	Swedish_Sign_Language	1	0,062
Armenian-ArmTDP	38	0,313	Lithuanian-HSE	3	0,056
Latin-ITTB	861	0,296	French-Sequoia	39	0,055
Czech-PDT	4132	0,274	Italian-PoSTWITA	61	0,049
Czech-FicTree	443	0,265	Dutch-Alpino	100	0,048
Naija-NSC	31	0,241	German-PUD	10	0,047
Hungarian-Szeged	83	0,197	Swedish-PUD	7	0,037
Belarusian-HSE	14	0,173	Swedish-Talbanken	35	0,036
Finnish-TDT	319	0,158	English-PUD	7	0,033
Slovenian-SST	42	0,142	Romanian-RRT	71	0,032
Russian-GSD	137	0,138	Marathi-UFAL	1	0,026
Russian-PUD	26	0,134	Finnish-PUD	4	0,025
Kazakh-KTB	13	0,123	Korean-PUD	4	0,024
Slovak-SNK	128	0,121	Arabic-PUD	5	0,024
Upper_Sorbian-UFAL	12	0,107	French-Spoken	8	0,023
Russian-SynTagRus	1179	0,107	Coptic-Scriptorium	3	0,023
Dutch-LassySmall	101	0,103	English-ParTUT	11	0,022
Telugu-MTG	6	0,093	Italian-PUD	5	0,021
Serbian-SET	79	0,091	Spanish-PUD	5	0,021
Czech-PUD	15	0,081	Uyghur-UDT	8	0,02
Romanian-Nonstan	109	0,079	Estonian-EDT	75	0,02
Arabic-PADT	223	0,079	Breton-KEB	2	0,02
Croatian-SET	152	0,077	Faroese-OFT	2	0,02
Greek-GDT	47	0,074	Kurmanji-MG	2	0,019

Tabel 3.2: *Orphan*-märgendite sagedus UD v2.2 60 puudepangas % järgi järjestatult: arv ehk orbude arv puudepangas, % ehk orvud/kõik märgendid*100

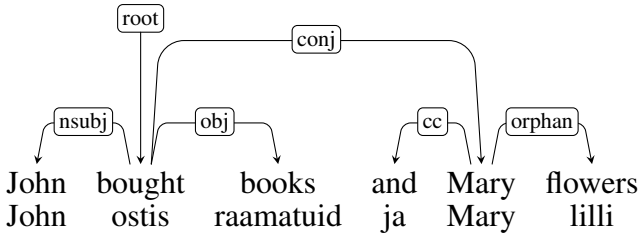
4. Varasemaid teemakohaseid uurimusi

Ellipsite tuvastamist ja nende märgendamist UD kujul on uuritud ka teiste keelte puudepankades. Muhonen ja Purtonen (2012) kirjeldavad oma artiklis reeglipõhist ellipsi tuvastamist koordineeritud osalausetes. Nad toovad välja, et üks tüüpilisemaid ellipsi konstruktsioone esineb osalausete koordinatsioonil, kus teises osalauses on predikaadi väljajätt ning olemasolevad lauseliikmed vastavad eelneva osalause lauseliikmetele.

Selle teadmise põhjal kirjutasid nad elliptilisi konstruktsioone tuvastavad reeglid, mis otsivad selliseid järjestikuseid osalauseid, kus teises osalauses on puudu predikaat ning olemas on nominatiivis subjekt, siis objekt, adverbiaal või predikatiiv samas käändes nagu eelmises osalauses. Reegleid katsetati soomekeelse vikipeedia peal ning tulemuseks saadi 89,9% tuvastamistäpsus.

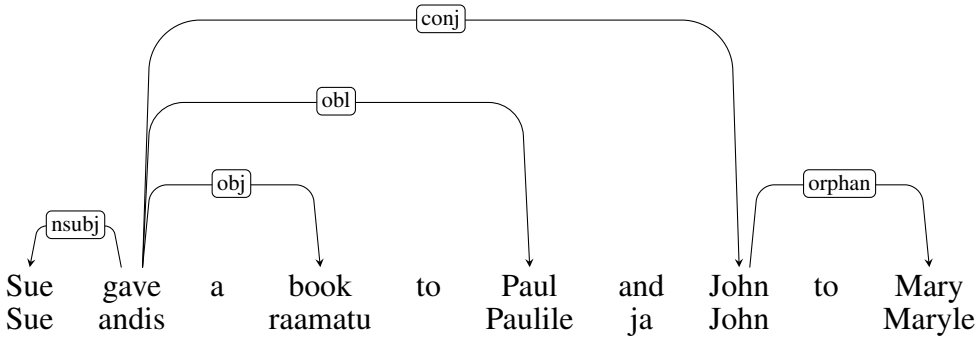
Sarnaselt Muhonenile ja Purtonenile (2012) kirjeldavad ka Schuster jt (2017; 2018) koordineeritud osalausete elliptilisi konstruktsioone. Nad eristavad nelja tüüpi predikaatverbi ellipseid inglise keeles:

1. Üksikverbi puudumine – Kõige tavalisem ellipsi konstruktsioon, kus lünk asub üldjuhul teises osalauses ja lüngale vastav predikaat on esimeses osalauses. Joonisel 4.1 on lüngaks predikaat *bought* (eesti k *ostis*), osalause ülemuseks määratakse subjekt *Mary* ning talle allub objekt *flowers* (eesti k *lilli*) spetsiaalsuhtega *orphan*.



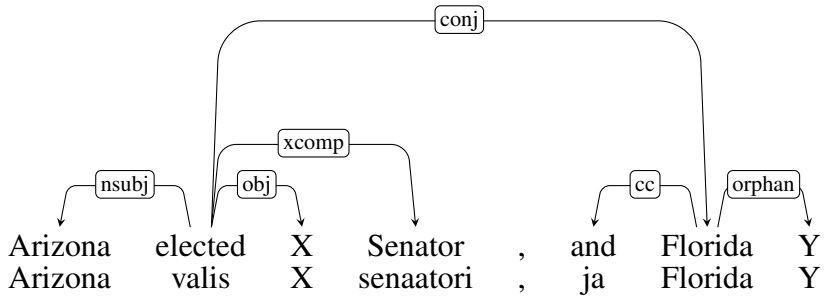
Joonis 4.1: Üksikverbi puudumine inglise keele näitel

2. Verb ja tema argumentide või laiendite ellips – Elliptilises lauses kaotatakse lisaks predikaadile ära ka tema argumendid või laiendid. Joonisel 4.2 on lüngaks *gave a book* (eesti k *andis raamatu*), kus verb on *gave* (eesti k *andis*) ja argumendiks on objekt *book* (eesti k *raamatu*). Teises osalauses määratakse ülemuseks subjekt *John* ja spetsiaalsuhte *orphan* abil riputatakse *Mary Johni* külge.



Joonis 4.2: Verbi puudumine koos argumendiga inglise keele näitel

3. Mittepidevad lüngad – Mõnikord võib lauses olla mitu lünka ning lause tõlgendamine on seetõttu eriti raskendatud. Joonisel 4.3 on lünkadeks *elected* (eesti k *valis*) ja *Senator* (eesti k *senaatori*), mis lahendatakse nii, et üks orbudest (*Florida*) määratakse ülemuseks ja teine (*Y*) pannakse tema külge *orphan* suhtega.



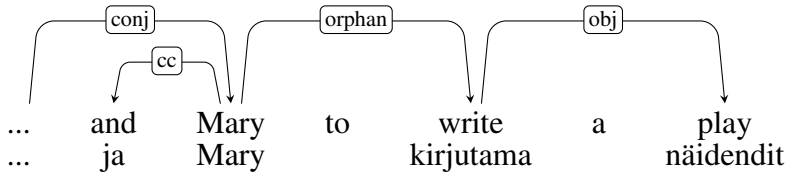
Joonis 4.3: Mitme lünga näide inglise keeles

4. Verb ja osalause või infiniittarind kohustusliku laiendina – Lünga moodustavad verb ning sisestatud infiniittarindid või osalaused. Järgmine näide (7, 8) illustreerib, et võimalik on kustutada infiniittarindeid nii, et lause ei muutu ebaloogiliseks (Lakoff jt 1970).

(7) *I want to try to begin to write a novel and*
... Mary a ___ play.
... Mary ___ to write a play.
... Mary ___ to begin to write a play.
... Mary ___ to try to begin to write a play.

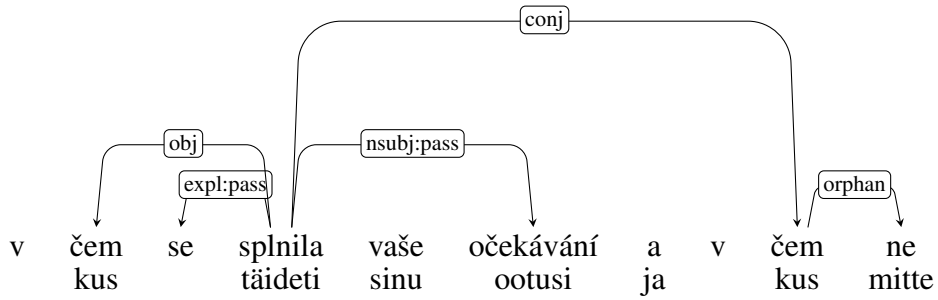
(8) eesti k *Ma tahan hakata kirjutama novelli ja*
... Mary ___ näidendit.
... Mary ___ kirjutama näidendit.
... Mary ___ hakata kirjutama näidendit.
... Mary ___ proovida hakata kirjutama näidendit.

Joonisel 4.4 on näha, et kustutades eelnevad infiniittarindid, määratakse osalause ülemuseks subjekt *Mary* ning sellele allub allesjäänud infiniittarind *to write* (eesti k *kirjutada*) spetsiaalsuhte *orphan* abil.



Joonis 4.4: Verbi ja infiniittarindite kustutamine inglise keele näitel

Droganova ja Zeman (2017) kirjeldavad asendamist (inglise k *stripping*) kui üht elliptilist konstruktsiooni. See tähendab, et tegemist on väljajätuga, mille puhul jääb alles vaid üks argument, mis vastab eelmisele osalausele. Sageli on sellisele orvuks jäänud argumendile lisaks mingisugune adverb (nt *mitte*), mille tõttu pole päris kindel, kas märgendada see orvuna või allesjäänud argumendi juurde kuuluva adverbiaalse laiendina. Joonisel 4.5 on tšehhikeelne näide osalause asendamisest, kus negatiivne partikkel *ne* (eesti k *mitte*) tähistab eelmise osalause verbi *splnila* (eesti k *täideti*) eitust ning on märgendatud orvuna.



Joonis 4.5: Asendamise näide tšehhi keeles

Schuster jt (2018) tõid välja, et kuigi predikaatverbi ellipsit on rohkelt uuritud ja mõnes puudepangas üritatud ka esile tuua (nt Penni puudepank (*Penn treebank*)), siis suurem osa automaatseid parsereid elliptiliste konstruktsioonidega ei tegele. Seetõttu otsustasid nad katsetada ellipsite automaatset parsimist kahel viisil. Üheks oli *orphan*-märgendi sissetoomine, teiseks oli tippude taastamine (vt ka 3.1. pt).

Automaatse parsimise tulemustest selgus, et mõlemad meetodid said elliptiliste lausete märgendamisega samaväärselt hakkama, küll aga oli *orphan*-protseduuri täpsus natuke parem võrreldes tippude taasamise protseduuriga, samas aga sai viimane parema saagise

tulemuse. *Orphan*-protseduuri raskeimaks probleemiks osutus nende osalausete märgendamine, kus sellele vastavas osalauses puudus samaväärne lauseliige. Nulltippude taastamise protseduuri veaks oli lisaks vale struktuuri määramisele ka vale suhte määramine nulltippudel. (Schuster jt 2018)

Näide 9 illustreerib *orphan*-protseduuril probleemseks osutunud lauset, kus teises osalauses on välja jäetud *had left* (eesti k *olid lahkunud*), subjekt *many* (eesti k *paljud*) määratakse osalause ülemuseks, aga kuna fragmendile *for good* (eesti k *lõplikult*) eelmises osalauses samaväärset vastet ei ole, siis parser seda orvuks märgendada ei oska. (Schuster jt 2018)

(9) *They had left the company, many ϕ[= had left] for good.*
(eesti k *Nad olid lahkunud ettevõttest, paljud lõplikult.*)

Droganova ja Zeman (2017) uurisid olemasolevate puudepankade kvaliteeti orbude märgendamisel. Selleks vaatasid nad väikest hulka UD v2.0 puudepanku ning neis olevaid orbusid ning nende märgenduse korrektsust. Nad leidsid viis põhilist veatüüpi:

1. *Conj*-suhte asemel kasutatakse *orphan*-suhet. Sageli tuleb see sellest, et eelmine märgend *remnant* on viidud kujule *orphan*, ilma et oleks muudetud sõnadevahelist struktuuri.
2. Sõltuvussüntaktiline suhe on õige, aga struktuur on vale. Selle puhul riputatakse osalause pea mitte eelmise osalause predikaadi, vaid samaväärse lauseliikme külge (samamoodi nagu märgendati versioonis 1 (vt ka 3.1. pt).
3. Elliptilise osalause pea määramise hierarhias eksitakse.
4. Ühe orvu asemel on kaks või rohkem ning mõlemad riputatakse eelneva osalause pea külge. See tähendab, et selle asemel, et valida osalauses üks uus osalause pea ning riputada teine selle külge, riputatakse mõlemad eelmise osalause pea külge.
5. Struktuur on õige, aga sõltuvussüntaktilised suhted valed (tihti jäetakse *orphan*-suhe üldse kasutamata).

Nad kontrollisid käsitsi *orphan*-märgendiga lauseid ning leidsid, et inglise keele puudepangas on 4,55% *orphan*-suhet sisaldavates lausetes viga, portugali keele puhul oli see protsent aga 33,3%, vene keele puhul 72,33%. (Droganova ja Zeman 2017) Nende leitud veatüübid ning vigade ulatus näitavad seda, et ühelt poolt on ellipsite automaatne märgendamine keeruline protsess ja teisalt seda, et varasem korrektne baaskorpus on väga oluline – kui lause süntaktiline struktuur on vale, on ka automaatne märgendamine oluliselt raskendatud.

5. Materjal ja meetod

Käesolevas töös kasutatakse elliptiliste lausete leidmiseks UD versiooni 2.2 kujul eesti keele sõltuvuspuude panka[1], mille ametlik versioon avalikustati 1. juulil 2018. Puudepangas on kokku 26 197 lauset (umbes 366 000 sõna) ilukirjandus-, ajalehe- ning teadustekstidest. Puudepank on märgendatud UD kujul ning eesti keele sõltuvuspuude pangas kasutatakse 39 erinevat sõltuvussuhet. Nende suhete sagedust illustreerib tabel 5.1. Nagu on näha, siis märgend *orphan* on juba puudepangas kasutusel, kuid kasutatud vaid teatud konstruktsioonides.

Märgend	**Sagedus**	**Märgend**	**Sagedus**	**Märgend**	**Sagedus**
punct	59925	case	7869	csubj	884
obl	32605	nsubj:cop	7831	csubj:cop	746
nmod	32461	acl	6440	compound	479
advmod	27957	advcl	6373	cc:preconj	365
root	26197	det	5758	discourse	292
nsubj	22113	nummod	5233	flat:foreign	154
amod	19583	xcomp	4625	vocative	126
conj	18557	parataxis	4231	fixed	117
obj	17982	compound:prt	3888	orphan	75
cc	13477	flat	3769	goeswith	72
aux	9930	acl:relcl	3214	advmod:quant	63
cop	8836	appos	3122	list	8
mark	8711	ccomp	2280		

Tabel 5.1: Sõltuvussüntaktiliste märgendite sagedus UDv2.2 kujul eesti keele sõltuvuspuude pangas

Elliptiliste lausete tuvastamiseks ja märgendamiseks kirjutati programm *elliptical_ sentences.py* [2] programmeerimiskeeles Python. Programmi töö koosneb kolmest sammust:

[1] http://universaldependencies.org/treebanks/et_edt/index.html
[2] https://github.com/liisitor/UD-EDT_with_orphans/

1. Tuvastab *olema*-verbiga elliptilised laused ning kirjutab need eraldi faili *elliptical_sentences_olema.conllu*.

2. Tuvastab muu verbiga elliptilised laused, märgendab need UD kujul elliptiliseks ning kirjutab need eraldi faili *elliptical_sentences_verbs.conllu*.

3. Kirjutab kogu korpuse uude faili (uutesse failidesse) laiendiga *_with_orphans.conllu*, kus tuvastatud elliptilised laused on muudetud UD kujul nähtavaks.

Programm kasutab ka kahte käsitsi tehtud lisafaili:

1. Tekstifail *exclude_sentences.txt*, kus on käsitsi leitud laused, mida programm UD kujul elliptiliseks ei tohiks märgendada (vt ka 7.3. pt).

2. Tekstifail *exclude_endings.txt*, kus on mõned sagedasemad lausete lõpud, milliseid tuleks ignoreerida.

Programmi jooksutamiseks on vaja Pythoni versiooni 3 [3] ja CoNLL-U parserit [4]. Tegemist on käsureaprogrammiga, mille jooksutamiseks tuleb kirjutada terminaliaknasse järgmine rida:

```
python3 elliptical_sentences.py --files korpus1.conllu
korpus2.conllu --exclusions-file exclude_sentences.txt
--endings-file exclude_endings.txt
```

[3] https://www.python.org/downloads/
[4] https://github.com/EmilStenstrom/conllu

6. Elliptiliste lausete tuvastamine

Elliptiliste lausete tuvastamisel eristatakse kaht tüüpi elliptilisi lauseid. Üheks on laused, milles on välja jäetud korduv o*lema*-verb ning teises välja jäetud korduv verb, mis ei ole *olema*. Taoline eristamine võimaldab paremini jätkata praktilist ellipsite märgendamist UD reeglite järgi, kuna *olema*-verbiga lausete märgendus erineb oluliselt muu verbiellipsiga lausete märgendusest, st *olema*-verb pole lause ülemus ning seega pole selle puudumine UD märgendusreeglite järgi problemaatiline.

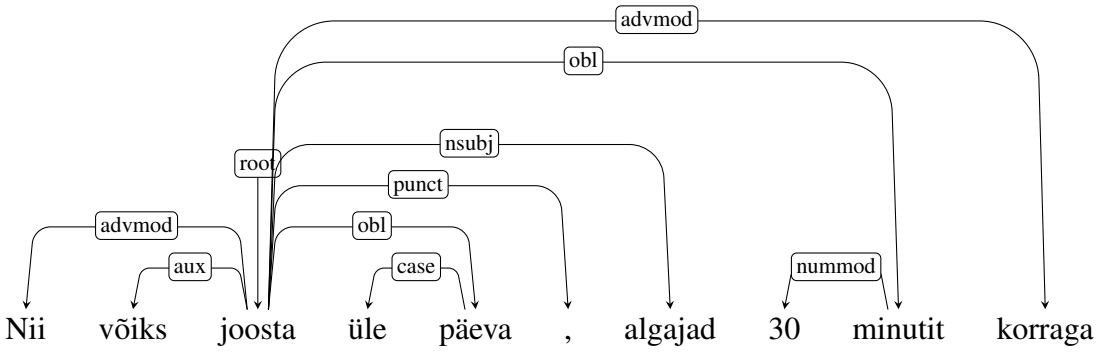
Võttes eelduseks, et osalaused ühendatakse verbiti, saab formuleerida ellipsite tuvastamise põhireegli, mis eeldab, et lauses on verb, mille alluvaks on mitteverb ja tema sõltuvussüntaktiline märgend on *conj*. Selleks, et tegemist poleks koopulaga, tuleb juurde panna tingimus, et selle mitteverbi alluvaks ei oleks sõna, mille sõltuvussüntaktiline suhe on *cop*. Et vältida koordinatsiooni leidmist, tuleks juurde panna ka tingimus, et sellel mitteverbil on alluvaks midagi lisaks punktuatsioonimärgile, sidesõnale või täiendile.

Olema-verbi ellipsite tuvastamise põhireegel käib vastupidiselt, tuleb leida sõna, mis ei ole verb ja mille alluvaks on *olema*-verb sõltuvussuhtega *cop*, selle mitteverbi küljes peab olema mitteverb, mille sõltuvussüntaktiliseks märgendiks on *conj* ning mille alluvaks ei ole *olema*-verb sõltuvussuhtega *cop*.

Tuvastamisprogrammi headuse hindamiseks kasutatakse saagist ja täpsust. Saagis näitab kui mitu asjakohast lauset tuvastatakse kõikidest asjakohastest lausetest. Täpsus näitab seda, kui mitu tuvastatud lausetest on päriselt elliptilise konstruktsooniga. Nende protsentide leidmiseks kasutati testkorpust, mis koosneb 1000 lausest. Tuvastamisprogramm leidis 19 elliptilist lauset 26 lausest. Tuvastamisprogrammi saagiseks saadi 73,1%, täpsuseks 95%. Madalam saagise protsent tuleb sellest, et lausete märgendus on ebaühtlane. Leidub lauseid, kus kaks osalauset ei ole omavahel märgendatud kasutades *conj* suhet, aga ka lauseid, kus osalause pea on riputatud eelmise osalause samaväärse lauseliikme, mitte predikaadi külge (vt ka 4. pt). Küll aga näitab kõrge täpsuse protsent seda, et leitud laused on üldiselt elliptilised.

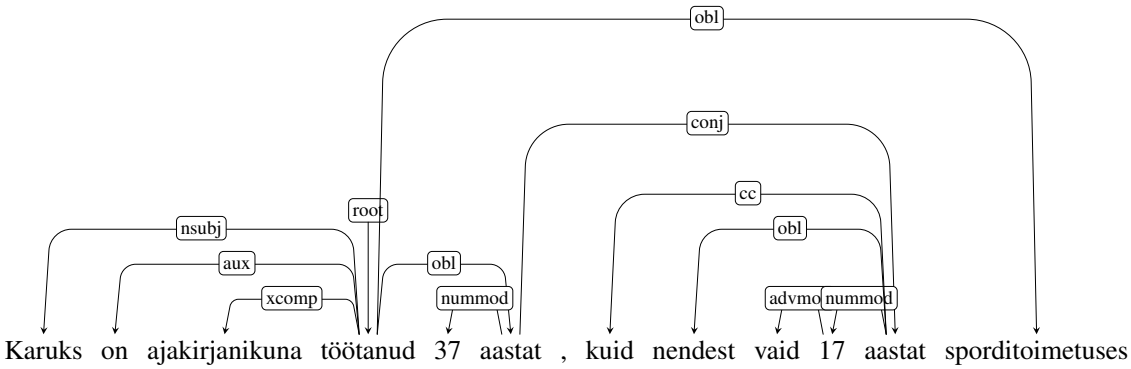
Joonisel 6.1 on tegemist lausega, kus kahte osalauset ühendades ei ole kasutatud *conj* suhet. Sealjuures on teise osalause lauseliikmed kõik riputatud esimese osalause predikaadi

külge, selle asemel, et valida osalause pea (subjekt *algajad*), märgendada see esimese osalause predikaadi külge kasutades sõltuvussüntaktilist märgendit *conj* ning teised elliptilise osalause lauseliikmed riputada selle külge.



Joonis 6.1: Vigase märgendusega lause eesti keele UD puudepangas

Joonisel 6.2 on teise osalause peaks määratud *aastat*, küll aga pole see riputatud mitte esimese osalause predikaadi, vaid samaväärse lauseliikme külge. Lisaks sellele on valesti riputatud ka teise osalause obliikva *sporditoimetuses*.



Joonis 6.2: Vigase märgendusega lause eesti keele UD puudepangas

7. Elliptilised laused eesti keele UD puudepangas

Ellipsite tuvastamise programmi kasutades leiti 359 elliptilist lauset, nendest 29 on ebasobivad. Võttes arvesse teistes keeltes leitud grupid (vt 4. pt), saab tuvastatud konstruktsioonid jagada kaheksasse gruppi: lihtpredikaadi ellips, liitpredikaadi ellips, predikaadi ja tema laiendite ellips, mittepidevad lüngad, *olema*-verbi ellips, eitus/jaatus asendus, *seda* asendus, ebasobilikud laused. Tabel 7.1 illustreerib korpuses esinevaid elliptilisi konstruktsioone:

Grupp	**Lausete arv**	**%**
Lihtpredikaadi ellips	129	36
Liitpredikaadi ellips	22	6,1
Predikaadi ja tema laiendite ellips	23	6,4
Mittepidevad lüngad	19	5,3
Olema-verbi ellips	100	28
Eitus/jaatuse asendus	23	6,4
Seda asendus	14	3,8
Ebasobilikud laused	29	8
Kokku	359	100

Tabel 7.1: Tuvastatud elliptilised laused grupeeritult

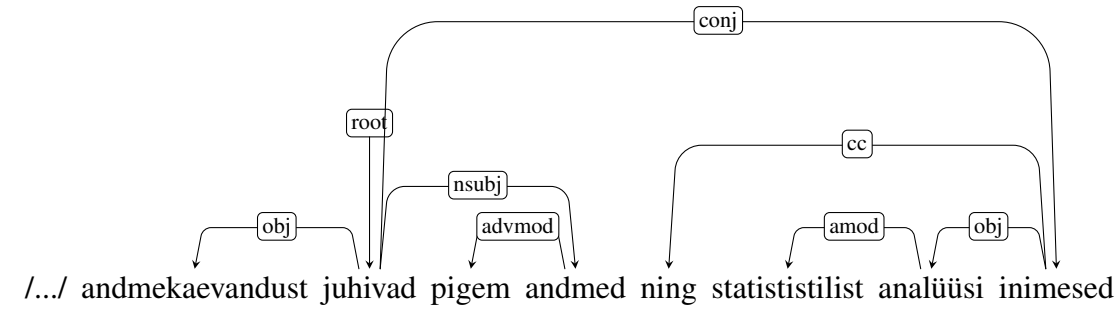
Valdkondade kaupa leiti 193 elliptilist lauset ajakirjandustekstidest, 60 elliptilist lauset ilukirjandustekstidest ning 106 elliptilist lauset teadustekstidest. Need numbrid kinnitavad ka Schusteri jt (2018) tähelepanekut, et tihti kasutatakse elliptilist konstruktsiooni faktide edasiandmiseks, mida suuresti ajakirjandus- ja teadustekstid sisaldavad.

7.1. Predikaatverbi ellips

7.1.1. Lihtpredikaadi ellips

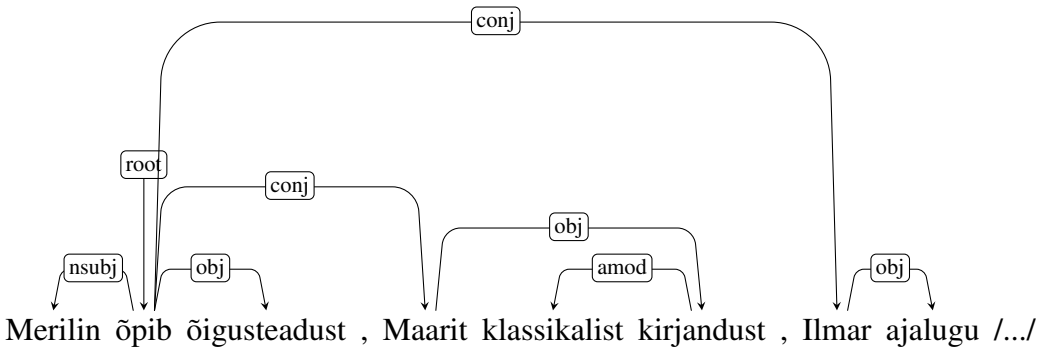
Lihtpredikaadi all peetakse silmas verbi liht- ja liitvorme, mis võivad olla ühesõnalised (nt *ehitas*), aga ka mitmesõnalised (nt *on ehitanud*, *ei ehita*), kus lisaks põhiverbile antakse grammatilist tähendust edasi abiverbiga *olema* või eituspartikliga *ei* (või eitusverbiga *ära, ärgu* jne) (Erelt 2017*b*: 93).

Tulemustest selgus, et eesti keele sõltuvuspuude pangas on kõige levinum lihtpredikaadi väljajätt, nende lausete arv moodustab 36% kõikidest elliptilistest lausetest. Kuna eesti keel on SVO (ehk subjekt, verb, objekt) sõnajärjega, asub olemasolev lihtpredikaat osalauses, millele järgnevad lüngaga osalaused. Joonisel 7.1 on tüüpiline lihtpredikaadi ellipsiga lause, kus kaks osalauset on koordineeritud ning teises osalauses on korduv verbi lihtvorm *juhivad* välja jäetud, osalause ülemuseks on määratud subjekt *inimesed* ning sellele allub objekt *statistilist analüüsi*.



Joonis 7.1: /.../ andmekaevandust juhivad pigem andmed ning statistilist analüüsi ø[= juhivad] inimesed

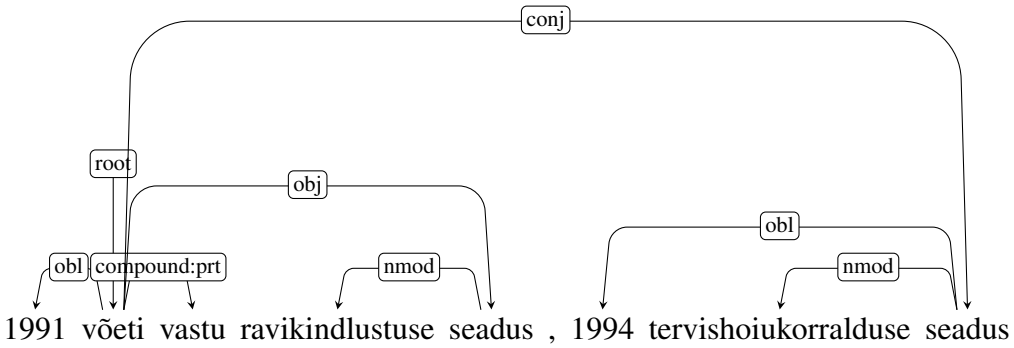
Joonisel 7.2 on tegemist samuti lihtpredikaadi ellipsiga, kuid ühes lauses on mitu elliptilist konstruktsiooni. Nii teine kui kolmas osalause kordavad esimese osalause verbi *õpib*, määrates osalausete ülemuseks subjektid *Maarit*, *Ilmar* ning riputades nende külge objektid *klassikalist kirjandust* ja *ajalugu*.



Joonis 7.2: Merlin õpib õigusteadust, Maarit [ø= õpib] klassikalist kirjandust, Ilmar [ø= õpib] ajalugu /.../

7.1.2. Liitpredikaadi ellips

Liitpredikaadi all peetakse silmas neid verbivormi ühendeid koos kas 1) käändelise verbivormiga, 2) teonime või teise pöördelise verbivormiga, 3) käänd- või määrsõnaga (Erelt 2017*b*: 94). Liitpredikaadi väljajätuga laused moodustavad 6,1% kõikidest elliptilistest lausetest. Joonisel 7.3 on ühendverbi (verbivorm + määrsõna) *võeti vastu* väljajätt, osalause ülemuseks on määratud objekt *seadus* ja sellele allub obliikva *1994*.

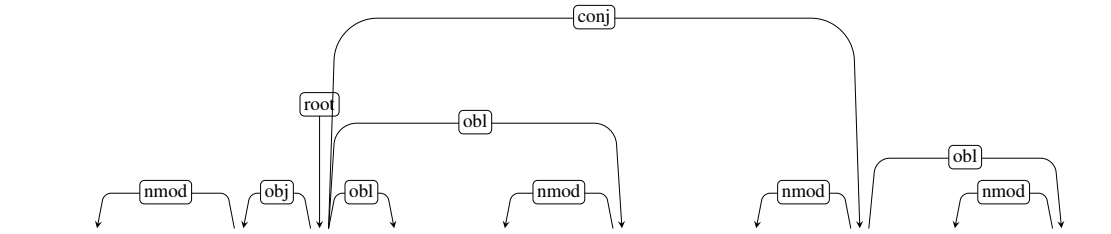


Joonis 7.3: 1991 võeti vastu ravikindlustuse seadus, 1994 ø[= võeti vastu] tervishoiukorralduse seadus

7.1.3. Predikaadi ja tema laiendite ellips

Predikaadi ja tema laiendite ellipsitega laused moodustavad 6,4% kõikidest elliptilistest lausetest. Lisaks predikaadile (mis võib olla nii verbi liht- kui liitvorm) jääb kordamata ka selle laiend, mis võib olla nii subjekt, objekt kui obliikva.

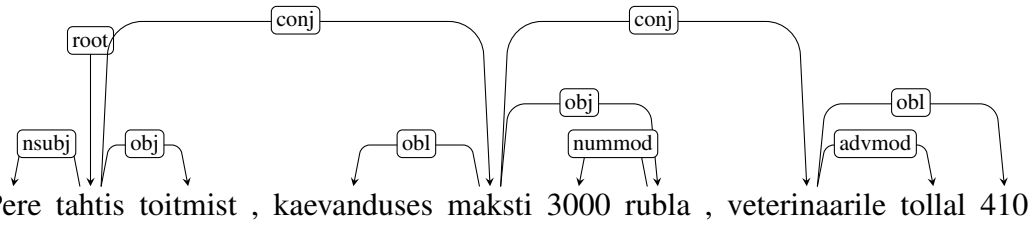
Kõige tüüpilisem näide on joonisel 7.4, kus välja on jäetud predikaat *kata* koos tema obliikvalise laiendiga *plekiga*. Osalause ülemuseks on määratud objekt *hobuveoki rattaid* ning selle külge riputatakse verbi teine laiend *esimeses kvartalis*.



Joonis 7.4: Talisõiduvahendi jalaseid kata plekiga kolmandas kvartalis, hobuveoki rattaid ø[= kata plekiga] esimeses kvartalis

7.1.4. Mittepidevad lüngad

Mittepidevate lünkadega elliptilised laused moodustastavad 5,3% kõikidest elliptilistest lausetest. Tavaliselt jäävad kordamata verb (mis võib olla nii verbi liht- kui liitvorm) ja arvufraasi, omadussõnafraasi vmt põhi. Joonisel 7.5 on välja jäetud korduv predikaat *maksti* ning korduv arvufraasi põhi *rubla*. Osalause ülemuseks määratakse *veterinaarile*, millele allub adverbiaalne modifikaator *tollal* kui obliikva *410*.



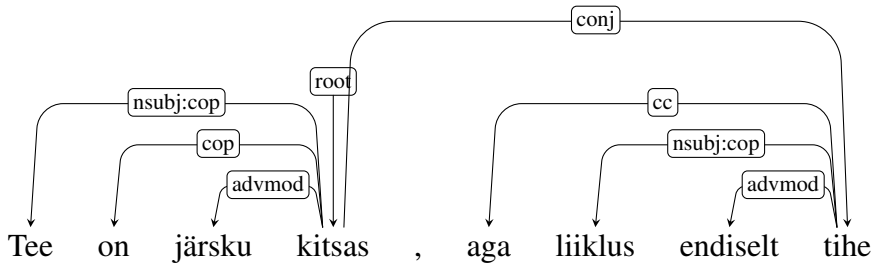
Joonis 7.5: Pere tahtis toitmist, kaevanduses maksti 3000 rubla, veterinaarile ø[= maksti] tollal 410 ø[= rubla]

7.1.5. *Olema*-verbi ellips

Olema-verbi ellipsitega laused moodustavad 28% kõikidest elliptilistest lausetest. UD süsteemis pole *olema*-verbi väljajätt oluline, sest lauses pole mitte *olema*-verb kõrgeimaks ülemuseks, vaid selleks on muu lauseliige[1]. Koopulalausete (eesti keele puhul *olema*-verbiga lausete) erineva märgendamise põhjuseks UD-süsteemis on asjaolu, et ees-

[1] http://universaldependencies.org/v2/copula.html

märgiks on luua ühtne märgendussüsteem paljude keelte jaoks ning osades keeltes võib sellises lauses öeldisverb puududa ja seda ei käsitleta väljajättelise, vaid normaalse lausena. Küll aga pole olulisemates eesti keele süntaksikirjeldustes eristatud koopulalauseid teistest lausetüüpidest (EKG 1993, Erelt jt 2017). Joonisel 7.6 on välja jäetud korduv *olema*-verb. Osalause ülemuseks on adjektiiv *tihe*, millele alluvad nii sidesõna *aga*, subjekt *liiklus* märgendiga *nsubj:cop* kui adverbiaalne modifikaator *endiselt*.

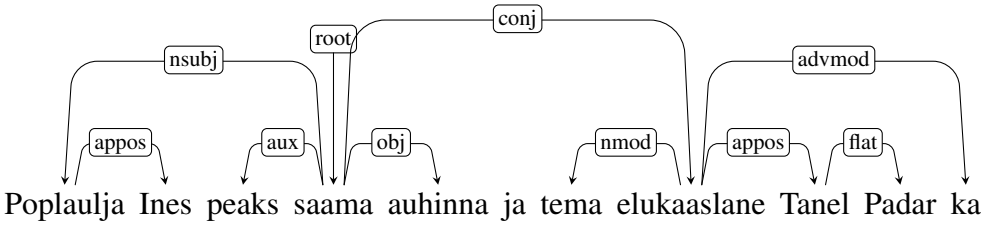


Joonis 7.6: Tee on järsku kitsas, aga liiklus ø[= on] endiselt tihe

7.2. Asendamine

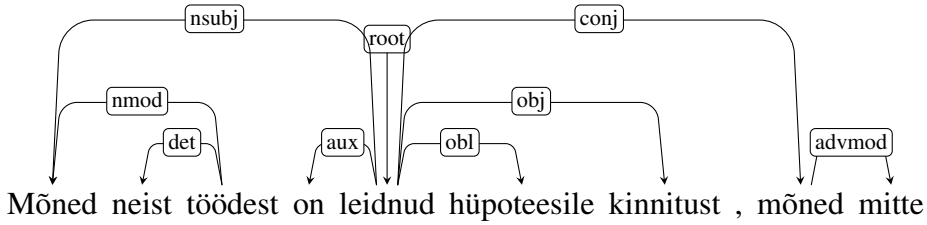
7.2.1. Jaatus/eitus asendus

Jaatuse või eituse partikliga laused moodustavad 6,4% elliptilistest lausetest. Täpsemalt asendatakse partikliga verb (mis võib olla nii verbi liht- kui liitvorm) kui ka muud korduvad lauseliikmed. Jaatuse väljendamiseks kasutatakse partikleid *küll, ka, samuti* ja eituse väljendamiseks partiklit *mitte*. Tavaliselt jääb alles subjekt + *samuti, ka, küll, mitte* või siis impersonaalilauses objekt + *samuti, ka, küll, mitte*. Joonisel 7.7 on näide osalause asendusest kasutades partiklit *ka*, mis sisuliselt väljendab korduvat infot *peaks saama auhinna*. Kuna verb ja tema laiend jääb kordamata, siis määratakse osalause ülemuseks subjekt *elukaaslane* ning partikkel *ka* ripub selle küljes.



Joonis 7.7: Poplaulja Ines peaks saama auhinna ja tema elukaaslane Tanel Padar ka ø[= peaks saama auhinna]

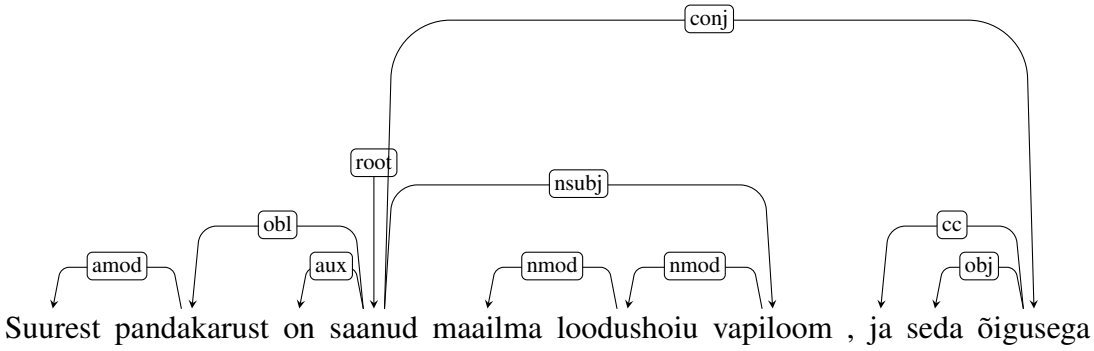
Joonisel 7.8 on näide osalause asendusest kasutades partiklit *mitte*, mis sisuliselt väljendab korduva info eitust *ei ole leidnud hüpoteesile kinnitust*. Kuna verb ja tema laiend jääb kordamata, siis määratakse osalause ülemuseks subjekti osa *mõned* ning partikkel *mitte* ripub selle küljes.



Joonis 7.8: Mõned neist töödest on leidnud hüpoteesile kinnitust, mõned mitte ø[= ei ole leidnud hüpoteesile kinnitust]

7.2.2. *Seda* asendus

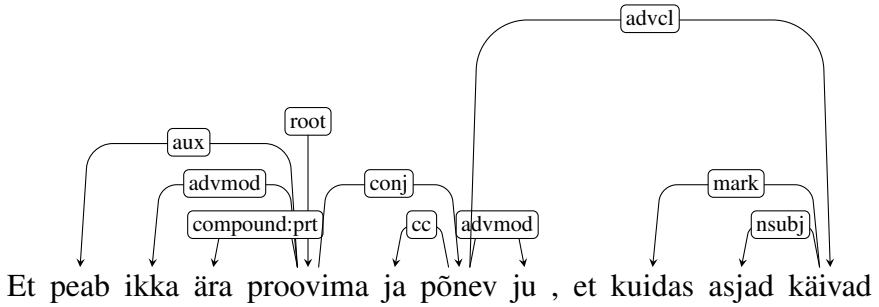
Seda asendusega laused moodustavad 3,8% elliptilistest lausetest. *Seda* asendust võib vaadata kui asendust, kus algne verbifraas on asendatud pronoomeniga *seda* ja üldise tähendusega verbi *tegema/saama* vmt vormiga ning seejärel on see verbivorm välja jäetud. Näiteks joonisel 7.9 asendab *seda* eelmise osalause tervikut *on saanud suurest pandakarust maailma loodushoiu vapiloom*. Osalause ülemuseks on *õigusega*, mille külge riputatakse *seda*.



Joonis 7.9: Suurest pandakarust on saanud maailma loodushoiu vapiloom, ja seda (∼ on saanud suurest pandakarust maailma loodushoiu vapiloom) õigusega

7.3. Ebasobilikud laused

Tuvastamisprogramm tagastas ka praktilise märgendamise jaoks (vt 8. pt) ebasobilikke elliptilisi lauseid, mis moodustavad 8% kõikidest elliptilistest lausetest. Leitud laused osutusid ebasobilikuks, kui lauses on märgendusviga, kui tegemist on lauselühendiga, samuti kui lünka pole võimalik taastada (lünga moodustab verb, mida pole võimalik kontekstist taastada). Joonisel 7.10 on tegemist ilmselt *olema*-verbi puudumisega, aga kuna seda kontekstist pole võimalik leida, osutub leitud lause ebasobilikuks.



Joonis 7.10: Et peab ikka ära proovima ja ø[= on] põnev ju, et kuidas asjad käivad.

8. Orbude automaatne märgendamine eesti keele UD puudepangas

8.1. Orbude märgendamise põhimõte

Nagu kirjeldati 3.1. peatükis, määratakse elliptilises osalauses kõrgeimaks ülemuseks üks lauseliige ning teised osalauses olevad lauseliikmed, mis muidu alluksid predikaadile, alluvad uuele ülemusele, tekitades mõningatel juhtudel ebaloogilise süntaktilise sõltuvuse (nt objekt allub subjektile). Seetõttu kasutatakse UD kujul puudepangas märgendit *orphan* selliste suhete eristamiseks.

Selline märgendusviis on osutunud problemaatiliseks ellipsi täiendava versiooni (*enhanced dependencies*) tegemisel (vt ka 3.1. pt, 9. pt), st tihti on probleemiks see, et raske on taastada orbude algset süntaktilist märgendit (Schuster jt 2018). Kuna ka UD kujul eesti keele sõltuvuspuude panga eesmärgiks on välja jõuda ellipsi täiendava versioonini, siis otsustati jätta alles orvuks jääva sõna algne märgend ning kõik orvuks jäävad lauseliikmed said märgendi kujul *orphan:algne_märgend*.

8.2. Programmi kirjeldus

Programmi tegemisel otsustati, et olemasolevate *orphan*-märgenditega programm ei tegele, sest need on kohati vigaselt märgendatud ning raske taastada. Sellist otsustust motiveeris ka fakt, et tuleviku sisendis enam märgendit kujul *orphan* ei kasutata. Seega otsustati need märgendid parandada käsitsi kujule *orphan:algne_märgend*.

Programm tuvastab need lauseliikmed, mis elliptilises konstruktsioonis on omavahel ühendatud ebaloogiliselt. Sealjuures on oluline vaadata, kas tegemist on võimaliku verbi alluvaga või mitte, seega on programmile ette antud list sõltuvussüntaktilistest suhetest, mida orvuks ei märgendata: *cc, punct, amod, case, appos, flat, acl:relcl, compound, mark, ccomp, det, amod, advcl, conj, parataxis, orphan*. Nagu näha, siis lisaks nimisõna juurde kuuluvatele sõnadele ei märgendata orvuna ka punktuatsiooni, sidesõnu ega muud taolist.

Mõned nimisõna juurde kuuluvad märgendid jäid siiski lubatuks, kuna algne märgendus on teisenduste käigus muutunud vigaseks. Näiteks on mitmetes lausetes obliikva saanud märgendi *nummod* või *nmod*, mis mõlemad peaksid tähistama nimisõna laiendina toimivat arv- või nimisõna. Seetõttu otsustati, et võimalikud predikaadi arv- või nimisõnalised laiendid märgendatakse kujule *orphan:obl*.

Märgendamisprogrammi tegemisel ja hindamisel otsustati ignoreerida korpuses juba olemasolevaid *orphan*-märgendeid. Hindamisele läksid kõik automaatselt tuvastatud (märgendamisprogrammi tulemusel sai uueks sõltuvussüntaktiliseks suhteks näiteks *obl -> orphan:obl*) ja mitte tuvastastud orvud (pidi olema *orphan:obl*, aga jäi kujule *obl*). Nendest tingimustest lähtudes saadi testkorpust kasutades märgendamisprogrammi saagiseks 81,8%, täpsuseks 94,4%. Madalam saagise protsent tuleb nendest samadest vigastest lausetest, mida kirjeldati ka 6. peatükis. See-eest on jällegi märgendamisprogrammi täpsus hea, st need orvud, mis programm leiab, märgendatakse üldjuhul õigesti.

8.3. Orbude märgendamise tulemused

Automaatselt märgendades leiti 278 orbu (lisaks 75 algset *orphan*-märgendit). Tabel 8.1 iseloomustab automaatselt leitud orbude sagedust UD kujul eesti keele sõltuvuspuude pangas.

Märgend	**Sagedus**	**Märgend**	**Sagedus**
orphan:obl	128	orphan:acl	11
orphan:advmod	110	orphan:xcomp	5
orphan	75	orphan:nsubj	4
orphan:obj	19	orphan:compound:prt	1

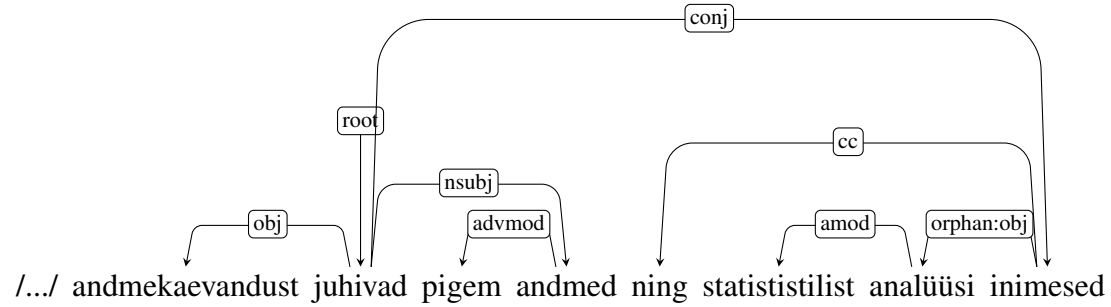
Tabel 8.1: Orbude sõltuvussüntaktilised märgendid ja nende sagedused programmi kasutades

Programmi väljundi käsitsi parandamisel leiti kokku 344 orbu. Tabel 8.2 iseloomustab automaatselt leitud ja käsitsi parandatud orbude sagedust UD kujul eesti keele sõltuvuspuude pangas.

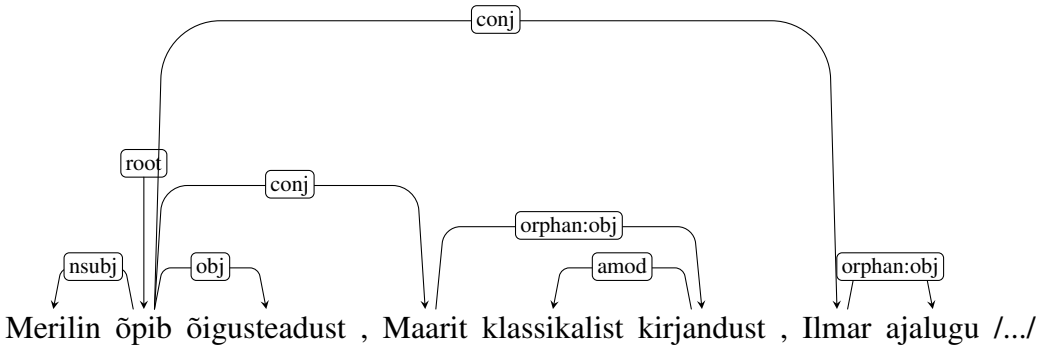
Märgend	Sagedus	Märgend	Sagedus
orphan:obl	151	orphan:nsubj	5
orphan:advmod	121	orphan:advcl	3
orphan:obj	38	orphan:parataxis	2
orphan:xcomp	13	orphan:compound:prt	1
orphan:acl	10		

Tabel 8.2: Orbude sõltuvussüntaktilised märgendid ja nende sagedused programmi väljundi käsitsi parandamisel

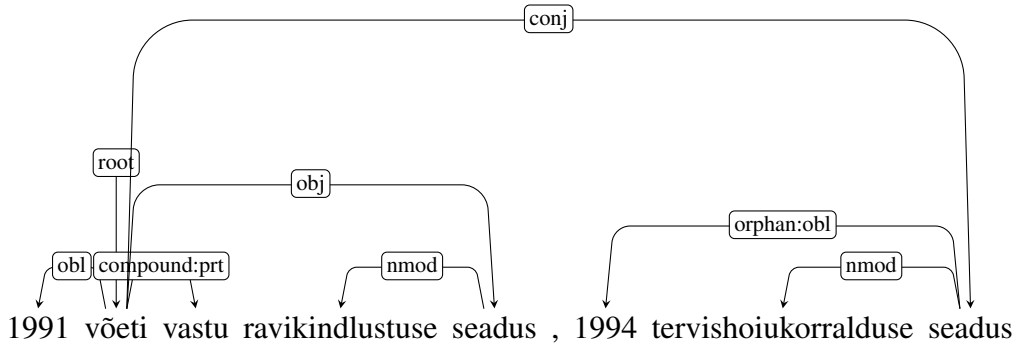
Järgmistel joonistel (8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8) on juba 7. peatükis kujutatud laused, kus elliptilises osalauses orvuks jäänud lauseliikmed on märgendatud kujul *orphan:algne_märgend*.



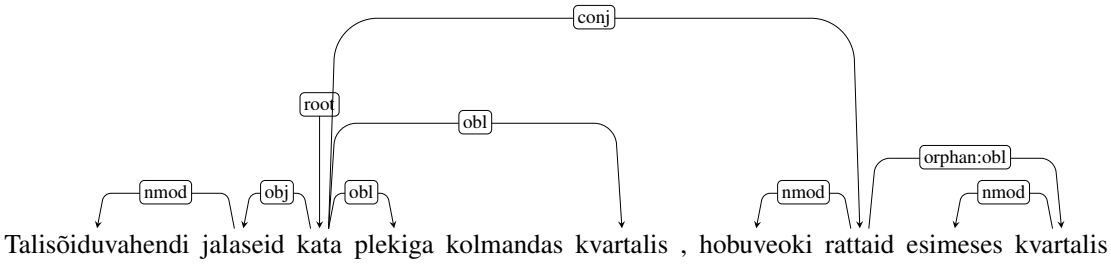
Joonis 8.1: /.../ andmekaevandust juhivad pigem andmed ning statistilist analüüsi ø[= juhivad] inimesed



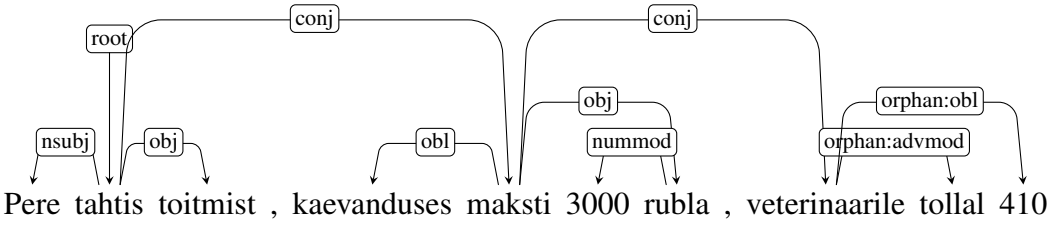
Joonis 8.2: Merlin õpib õigusteadust, Maarit [ø= õpib] klassikalist kirjandust, Ilmar [ø= õpib] ajalugu /.../



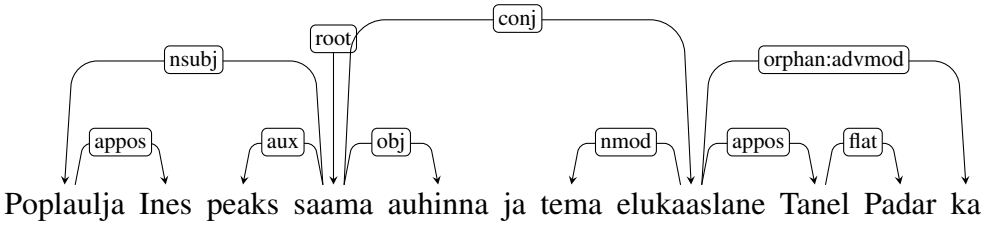
Joonis 8.3: 1991 võeti vastu ravikindlustuse seadus, 1994 ø[= võeti vastu] tervishoiukorralduse seadus



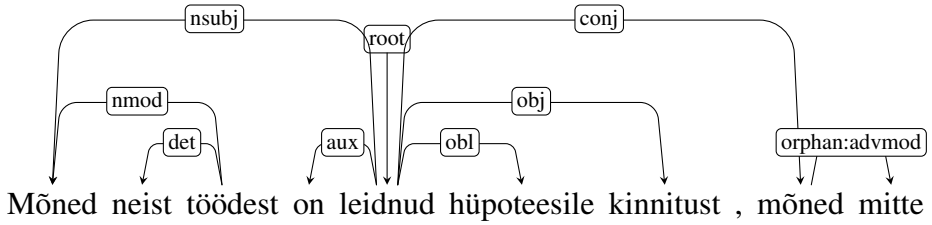
Joonis 8.4: Talisõiduvahendi jalaseid kata plekiga kolmandas kvartalis, hobuveoki rattaid ø[= kata plekiga] esimeses kvartalis



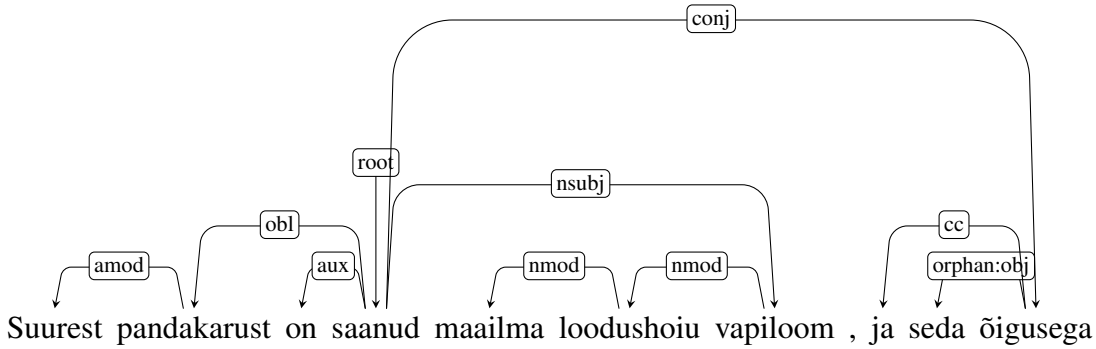
Joonis 8.5: Pere tahtis toitmist, kaevanduses maksti 3000 rubla, veterinaarile ø[= maksti] tollal 410 ø[= rubla]



Joonis 8.6: Poplaulja Ines peaks saama auhinna ja tema elukaaslane Tanel Padar ka ø[= peaks saama auhinna]

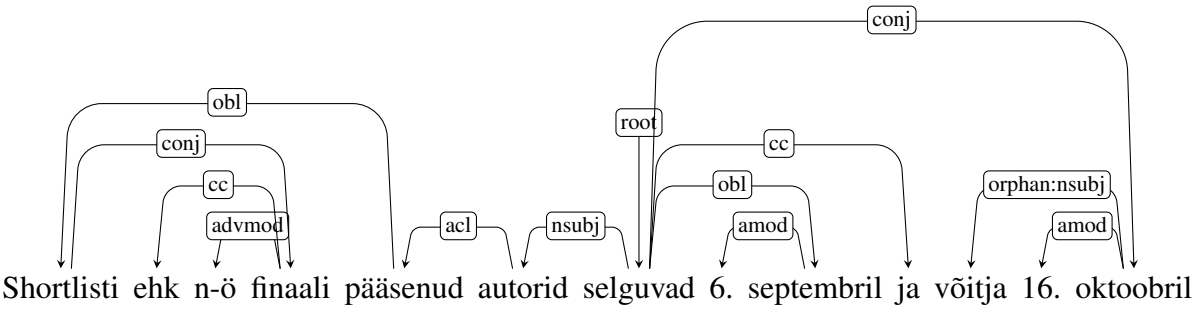


Joonis 8.7: Mõned neist töödest on leidnud hüpoteesile kinnitust, mõned mitte ø[= ei ole leidnud hüpoteesile kinnitust]



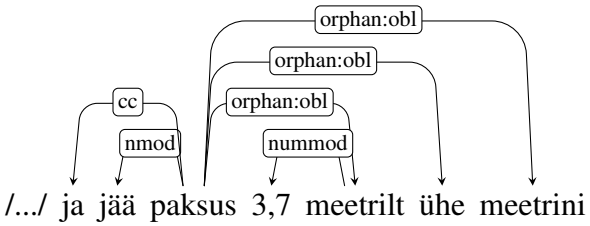
Joonis 8.8: Suurest pandakarust on saanud maailma loodushoiu vapiloom, ja seda (∼ on saanud maailma loodushoiu vapiloomaks) õigusega

Kuigi märgendamistäpsus testkorpusel osutus heaks (94,4%), siis tervet korpust hinnates mõned probleemid siiski tekkisid. Näiteks ei peaks olema orbu *orphan:nsubj*, sest osalause pead määrates peaks hierarhia järgi selleks olema subjekt ise. Küll aga on varasemates teisendustes määratud selleks mingi muu lauseliige. Joonisel 8.9 on teise osalause peaks määratud ajamäärus *oktoobril*, kuigi tegelikult peaks olema selleks subjekt *võitja* ning ajamäärus *oktoobril* peaks sellele alluma suhtega *orphan:obl*.



Joonis 8.9: Elliptiline lause, mille osalause pea on valesti määratud

Tulenevalt varasematest teisendustest tulenevatest vigadest said orvu märgendi ka mõned ebasobilikud lauseliikmed. Joonisel 8.10 sai o*rphan:obl* märgendi arvuline täiend *ühe*, mis peaks kuuluma nimisõna, mitte verbi juurde.

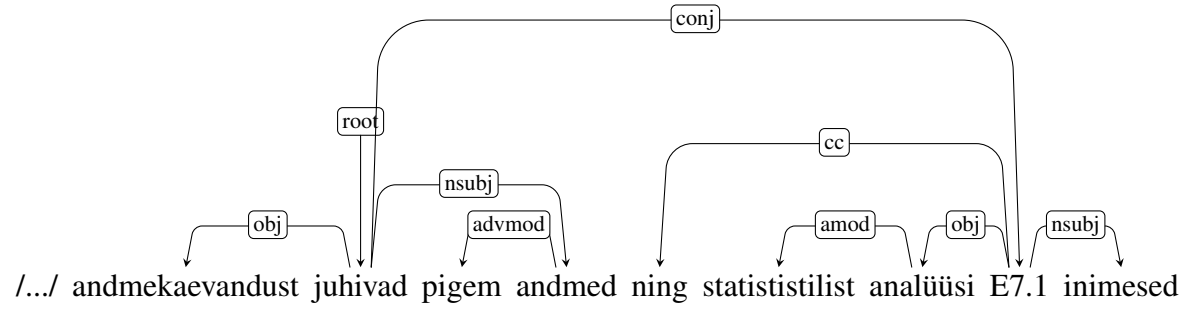


Joonis 8.10: *Koostatud arvutimudeli kohaselt võib jää hulk septembrikuus kahaneda kümne aastaga kuuelt miljonilt ruutkilomeetrilt kahe miljoni ruutkilomeetrini ja jää paksus 3,7 meetrilt ühe meetrini.* – Elliptiline lause, milles märgendati liigne orb

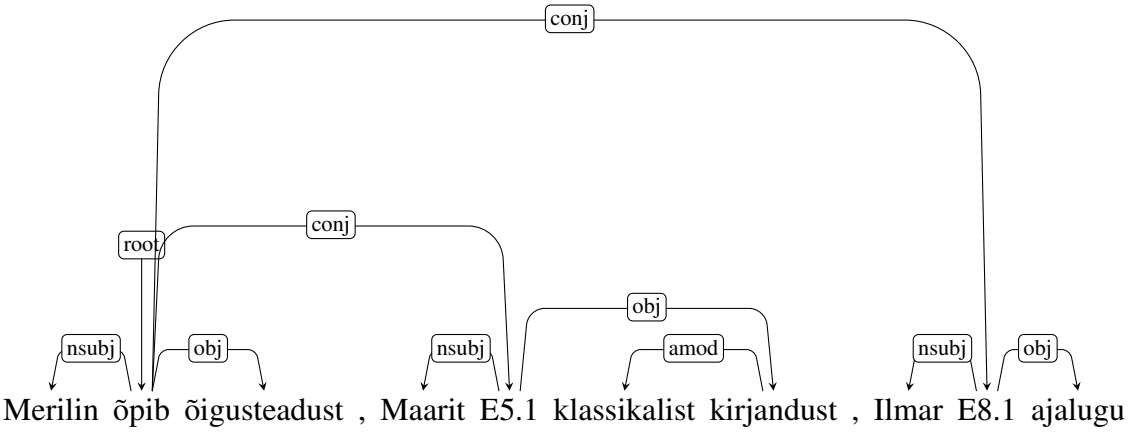
9. Ellipsi täiendav sõltuvusmärgendus

Selle peatüki eesmärgiks on anda informatiivne ülevaade ellipsi täiendavast sõltuvusmärgendusest, selle võimalikest raskustest ning näidata, kuidas peaksid laused selles versioonis olema märgendatud. Kuigi magistritöös märgendatud elliptilised laused annavad hea põhja uue versiooni loomiseks, pole nulltippude sissetoomine lihtne ülesanne. Tuleb arvestada, et taastatakse vaid olulisemad tipud (vt 3.1. pt). Raskendatud on ka nulltipu asukoht ning tippude lisamine CoNLL-U formaati. Lisaks sellele tuleb taastada varasema *conj* märgendiga sõna algne sõltuvussüntaktiline märgend. Järgnevad joonised kujutavad neid samu lauseid, mida kasutati ka peatükkides 7. peatükis ja 8.3. peatükis.

Joonisel 9.1 tähistab lisatipp *E7.1* eelmise osalause lihtpredikaati *juhivad*. Joonisel 9.2 tähistavad mõlemad lisatipud 5.1 ja 8.1 lihtpredikaati *õpib*.

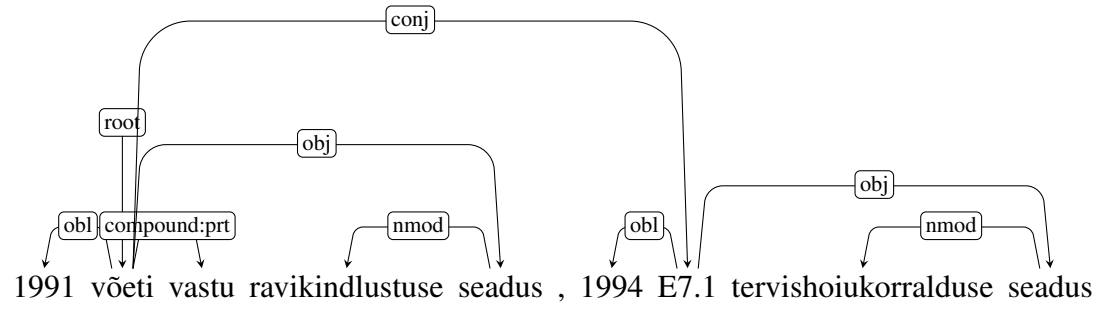


Joonis 9.1: /.../ andmekaevandust juhivad pigem andmed ning statistilist analüüsi E7.1 inimesed



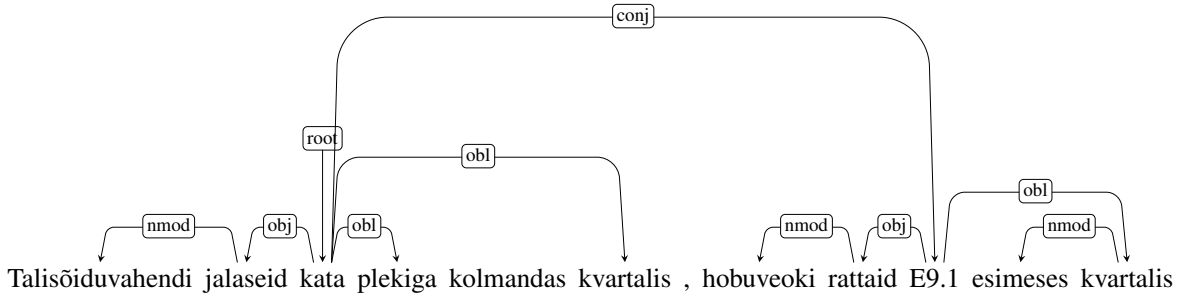
Joonis 9.2: Merilin õpib õigusteadust, Maarit E5.1 klassikalist kirjandust, Ilmar E8.1 ajalugu

Joonisel 9.3 tähistab nulltipp *E7.1* ühendverbi verbi osa *võeti*, määrsõnaline osa *vastu* eraldi tippu ei saa, sest see ei mõjuta teiste osalause liikmete märgendust.



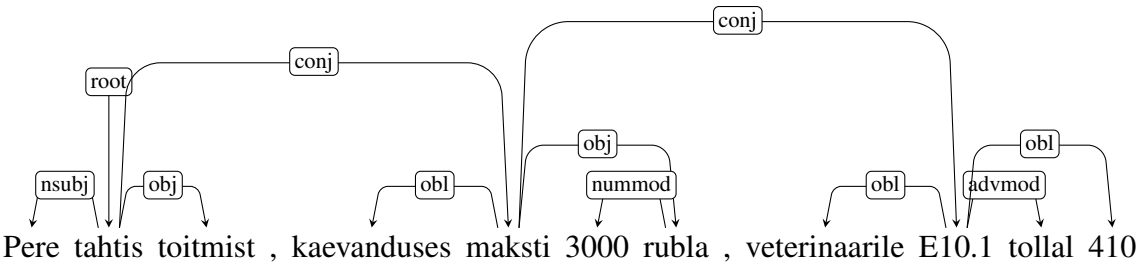
Joonis 9.3: 1991 võeti vastu ravikindlustuse seadus, 1994 E7.1 tervishoiukorralduse seadus

Joonisel 9.4 tähistab nulltipp *E9.1* verbi *kata*, selle juurde kuuluv laiend *plekiga* eraldi tippu ei saa, sest see ei mõjuta teiste osalause liikmete märgendust.



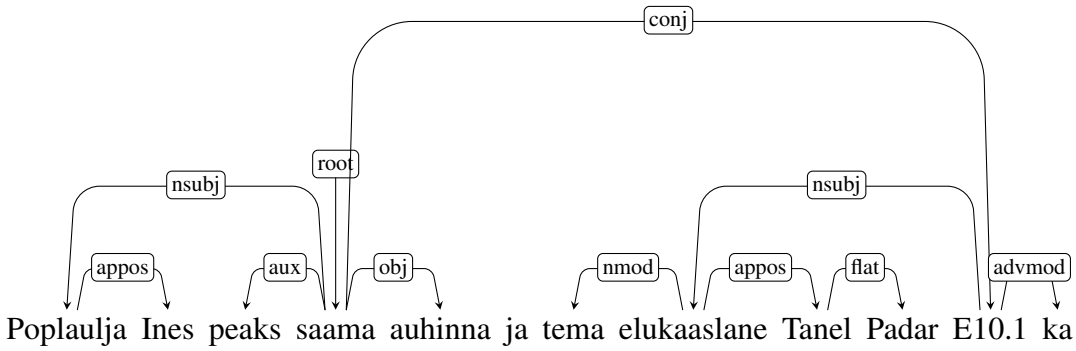
Joonis 9.4: Talisõiduvahendi jalaseid kata plekiga kolmandas kvartalis, hobuveoki rattaid E9.1 esimeses kvartalis

Joonisel 9.5 tähistab nulltipp *E10.1* verbi *maksti*, korduva arvufraasi põhja puudumine eraldi tippu ei saa, sest see ei mõjuta teiste osalause liikmete märgendust.



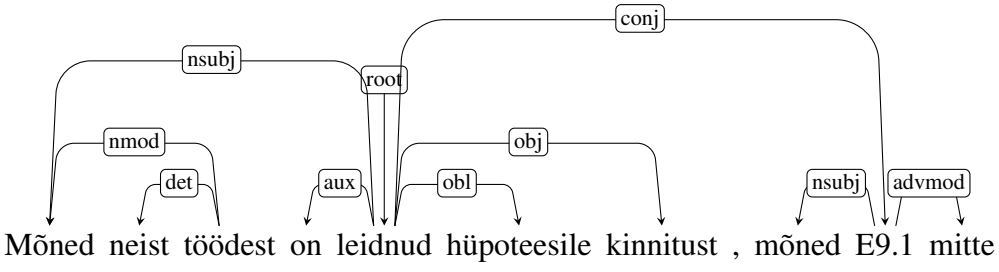
Joonis 9.5: Pere tahtis toitmist, kaevanduses maksti 3000 rubla, veterinaarile E10.1 tollal 410 ø[= rubla]

Joonisel 9.6 tähistab nulltipp *E10.1* perifrastilise verbivormi põhja *saama*, selle küljes olev *peaks* ja objekt *auhinna* eraldi nulltippe ei saa, sest nende olemasolu ei mõjuta teiste osalause liikmete märgendust.



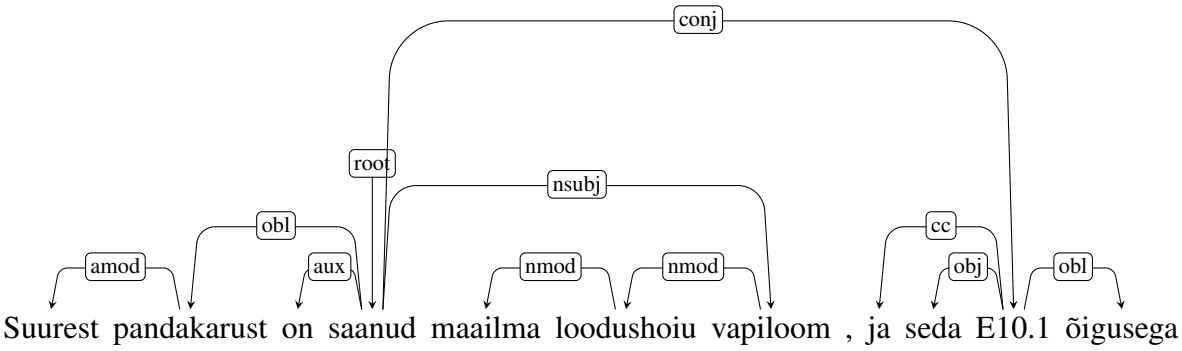
Joonis 9.6: Poplaulja Ines peaks saama auhinna ja tema elukaaslane Tanel Padar E10.1 ka

Joonisel 9.7 tähistab nulltipp *E9.1* lihtpredikaadi põhja *leidnud*, selle küljes olev *ole* ja *ei*, samuti obliikva *hüpoteesile* ja objekt *kinnitust* eraldi nulltippe ei saa, sest nende olemasolu ei mõjuta teiste osalause liikmete märgendust.



Joonis 9.7: Mõned neist töödest on leidnud hüpoteesile kinnitust, mõned E9.1 mitte

Joonisel 9.8 tähistab nulltipp *E10.1* lihtpredikaadi põhja *saanud*, selle küljes olev *on* eraldi nulltippe ei saa, sest selle olemasolu ei mõjuta teiste osalause liikmete märgendust.



Joonis 9.8: Suurest pandakarust on saanud maailma loodushoiu vapiloom, ja seda E10.1 õigusega

10. Kokkuvõte

Magistritöö eesmärgiks oli anda ülevaade elliptiliste lausete esinemisest eesti keele UD kujul puudepangas ning esitada nende tuvastamise ja märgendamise meetodid, vahendid ja tulemused. Sealjuures püstitati kolm praktilist eesmärki:

1. Tuvastada elliptilised laused eesti keele UD sõltuvuspuude pangas.
2. Märgendada nendes lausetes orvuks jäänud lauseliikmed eesmärgiga luua põhi tulevasele lausekujule, kus kunstlikult tuuakse sisse puuduv osalause pea ning orvuks jäänud lauseliikmed riputatakse selle külge.
3. Luua eesti keele UD puudepangast selline versioon, kus elliptilised konstruktsioonid on UD reeglite järgi esile tõstetud.

Magistritöös püsitatud eesmärgid täideti. Tuvastati elliptilised laused puudepangas, sealjuures jagati need gruppidesse ning kirjeldati neid süvitsi. Samuti kirjeldati programmi, millega need laused leiti. Saadud laused märgendati UD kujul elliptilisteks. Kirjeldati orbude märgendamise meetodit, anti ülevaade lausete uuest kujust, kus orvud on sisse toodud ning esitati orbude statistika.

Kokku leiti 359 elliptilist lauset, mis moodustavad kogu korpusest 1,37%. Uues korpuses on kokku 344 orbu (st märgendit *orphan:algne_märgend*), mis moodustab kõikidest märgenditest 0,094%. Kuigi see arv ei ole väga suur, siis vaadates tabelit 3.2 asuks eesti keele puudepank enam mitte tabeli lõpus, vaid esi kolmekümne seas. Samuti tabeli 3.1 järgi oleks versioonis 2.2 eesti keele puudepank nende $\sim$11% seas, kellel on vähemalt 100 lauses *orphan*-märgend. Siinkohal ei tasu neid numbreid vaadata kui 'mida rohkem, seda parem', sest erinevad keeled kasutavad ellipsit erinevalt. Küll aga on magistritöös loodud korpuseversioonis $\sim$5 korda rohkem orbusid kui ametlikus versioonis, mis näitab, et vähemalt eesti keele UD kujul puudepangas oli seni suur hulk elliptilisi konstruktsioone esile tõstmata.

Tulemuste analüüsist selgus ka see, et hulk elliptilisi lauseid jäi tuvastamata, kuna algne märgendus on teisenduste käigus vigaseks muutunud. Ellipsite tuvastusprogrammi saagiseks saadi 73,1%, täpsuseks 95%. Orbude märgendamise saagiseks 81,8%, täpsuseks 94,4%. Et saagist paremaks saada, tuleb kõigepealt parandada korpuses olevad laused,

kas automaatselt või käsitsi. Kaks põhilist puudepangas leiduvat viga on järgmised: 1) osalauseid ei ühendata *conj*-suhtega 2) elliptilise osalause pea on riputatud eelmise osalause samaväärse liikme, mitte predikaadi külge, mille tõttu on raske eristada, kas tegemist on koordinatsiooni või elliptilise osalausega.

Töös anti ka ülevaade võimalikku edasiarendusse, milleks on ellipsi täiendav sõltuvusmärgendus (*enhanced dependencies*). Kuna puudepangas olevad orvud märgendati kujul *orphan:algne_märgend* mitte kujul *orphan*, siis on ellipsi täiendava sõltuvusmägenduse loomine sammu võrra kergem – enam ei pea hakkama taastama orbude algset märgendit, sest see info on korpuses olemas. Küll aga on see siiski keeruline ülesanne, sest tuleb arvestada, et taastatakse vaid olulisemad tipud, samuti tuleb taastada varasema *conj* märgendiga sõna algne sõltuvussüntaktiline märgend.

Magistritöös anti ülevaade elliptilistest lausetest eesti keele UD puudepangas, nende tuvastamismeetoditest ja märgendamisest. Praktilise tulemusena saadi puudepank, kus esile on toodud elliptilised konstruktsioonid, mis on heaks aluseks edasises ellipsite märgendamises, aga võimaldab ka juba praegusel kujul paremat praktilist kasutamist erinevates keeletehnoloogilistes ülesannetes.

Kirjandus

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. et al. (1999), 'Longman grammar of spoken and written english'.

Droganova, K. and Zeman, D. (2017), Elliptic constructions: Spotting patterns in ud treebanks, *in* 'Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)', pp. 48–57.

EKG, I., Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K. and Vare, S. (1993), 'Eesti keele grammatika ii', *Süntaks.[Estonian Grammar II. Syntax.] Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut* .

Erelt, M. (2017*a*), 'Ellips', *Eesti keele süntaks. Tartu: Tartu Ülikool* pp. 591–601.

Erelt, M. (2017*b*), 'Öeldis. liht- ja liitöeldis', *Eesti keele süntaks. Tartu: Tartu Ülikool* pp. 93–107.

Erelt, M. (2017*c*), 'Rinnastus', *Eesti keele süntaks. Tartu: Tartu Ülikool* p. 611.

Erelt, M., Metslang, H. et al. (2017), 'Eesti keele süntaks', *Tartu: Tartu Ülikool* .

Hajič, J., Hajičová, E., Mikulová, M., Mírovskỳ, J., Panevová, J. and Zeman, D. (2015), Deletions and node reconstructions in a dependency-based multilevel annotation scheme, *in* 'International Conference on Intelligent Text Processing and Computational Linguistics', Springer, pp. 17–31.

Hakulinen, A., Korhonen, R., Vilkuna, M. and Koivisto, V. (2004), *Iso suomen kielioppi*, Suomalaisen kirjallisuuden seura.

Karlsson, F., Voutilainen, A., Heikkila, J. and Anttila, A. (1995), *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*, Mouton de Gruyter.

Lakoff, G. and Ross, J. R. (1970), 'Gapping and the order of constituents', *Progress in linguistics: A collection of papers* **43**, 249.

Melćuk, I. A. (1988), *Dependency syntax: theory and practice*, SUNY press.

Muhonen, K. and Purtonen, T. (2012), Rule-based detection of clausal coordinate ellipsis., *in* 'LREC', pp. 1955–1959.

Muischnek, K. and Müürisep, K. (2017*a*), 'Eesti keele sõltuvuspuude pank ja selle keeleteoreetilised lähted', pp. 122–145.

Muischnek, K. and Müürisep, K. (2017*b*), Estonian copular and existential constructions as an ud annotation problem, *in* 'Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)', pp. 79–85.

Muischnek, K., Müürisep, K. and Puolakainen, T. (2016), Estonian dependency treebank: from constraint grammar tagset to universal dependencies., *in* 'LREC'.

Muischnek, K., Müürisep, K., Puolakainen, T., Aedmaa, E., Kirt, R. and Särg, D. (2014), Estonian dependency treebank and its annotation scheme, *in* 'Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)', pp. 285–291.

Müürisep, K. (2000), *Eesti keele arvutigrammatika: süntaks*, Tartu Ülikooli Kirjastus.

Nivre, J. (2008), *Treebanks*, Mouton de Gruyter, 2008, pp. 225–241.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S. and Silveira, N. (2016), Universal dependencies v1: A multilingual treebank collection., *in* 'LREC'.

Schuster, S., Lamm, M. and Manning, C. D. (2017), Gapping constructions in universal dependencies v2, *in* 'Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)', pp. 123–132.

Schuster, S., Nivre, J. and Manning, C. D. (2018), 'Sentences with gapping: Parsing and reconstructing elided predicates', *arXiv preprint arXiv:1804.06922* .

Lisad

Programm ja selleks vajalikud failid, samuti orbudega märgendatud korpus asuvad siin: https://github.com/liisitor/UD-EDT_with_orphans

Summary

IDENTIFICATION AND ANNOTATION OF ELLIPTICAL SENTENCES IN ESTONIAN UD TREEBANK

The goal of this master thesis was to give an overview of the occurrence of elliptical sentences in the Estonian UD treebank and provide methods, means and results of their identification and annotation. Three main practical goals were set:

1. Identify elliptical sentences in the Estonian UD treebank.
2. Annotate orphaned sentence nodes with the aim of creating a basis for the new sentence format which artificially includes the missing head of the clause and orphaned nodes are attached to it.
3. Create a version of the Estonian UD dependency treebank in which the elliptical constructions are highlighted based on UD rules.

The goals set for the master thesis were achieved. The elliptical sentences were identified, grouped and more thoroughly analyzed. In addition, the program that was created for identifying the sentences was described and included along with this thesis. The identified sentences were annotated as elliptical sentences in UD format. As part of this thesis, a method for annotating orphans was described, an overview of a new representation of sentences was given along with statistics for identified orphans.

359 elliptical sentences were found in total which made up 1.37% of the corpus. The new corpus contains a total of 344 orphans (i.e *orphan:original_dependency_relation*) which make up 0.094% of all of the orphans. Although this is not a very great number, if we were to look at table 3.2, the new version of Estonian dependency treebank would now rank in the top thirty instead of at the bottom of the table. Moreover, according to table 3.1 the Estonian treebank version 2.2 would place among the top $\sim$11% percent that have orphan annotation in at least 100 sentences. In this case, these figures may be misleading and the notion of 'the more, the better' is not appropriate as different languages make use of ellipsis in different ways. Still, there are about 5 times as many orphans in the corpus that was created as part of this thesis than in the official version, which indicates that at

least in the Estonian UD treebank there was a great amount of previously not highlighted elliptical constructions.

During the analysis of the results, it was found that some of the elliptical sentences still remained unidentified, because the original annotations had become corrupted. The recall for the elliptical sentence identification program was 73.1%, accuracy 95%, for the annotation of orphans these figures were 81.8% and 94.4% respectively. In order to further improve the recall, the sentences must first be fixed either by hand or automatically. The two main mistakes were found to be the following: 1) clauses are not joined with the *conj*-relation 2) elliptical sentence clause head is connected to the equivalent node in the previous clause, rather than the predicate within the same clause.

The thesis also outlined possible future work by means of enhanced dependencies of ellipsis. As the identified orphans in the treebanks were annotated in the form *orphan:original_dependency_relation* (instead of *orphan*), enhanced annotation should be simplified because the step of having to restore the original annotation of orphans can now be skipped, as the corpus already contains this information. It must be noted that this is still a difficult task, because the fact that only the most important nodes are restored must be taken into account. Additionally, the node with the previously annotated *conj* dependency relation must be restored to its original dependency relation.

This thesis gave an overview of elliptical sentences in the Estonian UD treebank, the methods for identification and annotation. As a practical result, a new treebank was created which highlights elliptical constructions. This is a good basis for further annotation of elliptical sentences, but which also already enables improved practical application for natural language processing tasks in its current form.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Liisi Torga

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   'Elliptiliste lausete tuvastamine ja märgendamine eesti keele UD puudepangas"

   mille juhendaja on Kadri Muischnek

   (a) reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile;

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **27.08.2018**