Eesti *Universal Dependencies'* sõltuvuspuude panga versioon 1.4

detsember 2016

NB! Kuna Eesti UD puudepank on osa parsimisvõistluse (*shared task*) „Multilingual Parsing from Raw Text to Universal Dependencies" http://universaldependencies.org/conll17/ andmestikust, siis on Eesti UD versioonides 1.3 ja 1.4 kuni 15. maini 2017 eemaldatud testandmestikku kuuluvad laused. Vajadusel küsige tervet puudepanka kaili.muurisep@ut.ee või kadri.muischnek@ut.ee.

Sissejuhatus

Universal Dependencies' projekti eesmärgiks on luua üldine ja keeletüpoloogiliselt põhjendatud märgendusskeem erinevate keelte morfoloogiliseks ja sõltuvussüntaktiliseks märgendamiseks. Täpsemalt vt http://universaldependencies.org/

Tartu Ülikoolis loodud Eesti *Universal Dependencies'* sõltuvuspuude pank on saadud eesti keele sõltuvuspuude panga (EDT) https://github.com/EstSyntax/EDT teisendamisel *Universal Dependencies'* formaati.

Eesti UD puudepanga loomiseks on kõigepealt EDT puudepanga morfoloogiline ja süntaktiline märgendus automaatselt teisendatud UD formaati ning tulemus käsitsi ühestatud ja parandatud.

See eesti keele UD sõltuvuspuude pank vastab UD märgendusskeemi esimesele versioonile, UD v1. Detsembris 2016 anti välja uus UD märgendusskeem, UD v2. Sellele vastav eesti keele UD sõltuvuspuude pank valmib 2017. aasta lõpuks.

Sõnestamine

Sõnana käsitletakse tühikute ja/või kirjavahemärkidega eraldatud stringe.

CoNLL-U-formaat

Eesti UD puudepank on CoNLL-U-formaadis; selle kohta vt täpsemalt siit: http://universaldependencies.org/format.html

Kasutatud sõnaliigi märgendid

ADJ adjektiiv
ADP adpositsioon
ADV adverb
AUX abiverb
CONJ koordineeriv sidend
INTJ interjektsioon
NOUN substantiiv
NUM numeraal
PRON pronoomen
PROPN pärisnimi
PUNCT punktuatsioon
SCONJ alistav sidend
SYM sümbol (nt *C5* ühendis *Citroen C5*)

VERB verb
X muu

Morfoloogilised kategooriad
Abbr=Yes lühend
AdpType=Post adpositsiooni liik: postpositsioon
AdpType=Prep adpositsiooni liik: prepositsioon
Case=Abe kääne: abessiiv
Case=Abl kääne: ablatiiv
Case=Add kääne: aditiiv (illatiivi lühike vorm)
Case=Ade kääne: adessiiv
Case=All kääne: allatiiv
Case=Com kääne: komitatiivCase=Ela kääne: elatiiv
Case=Ess kääne: essiiv
Case=Gen kääne: genitiiv
Case=Ill kääne: illatiiv
Case=Ine kääne: inessiiv
Case=Nom kääne: nominatiiv
Case=Par kääne: partitiiv
Case=Ter kääne: terminatiiv
Case=Tra kääne: translatiiv
Connegative=Yes eitavas vormis verb; selle märgendi saab nt sõnavorm *tee* ühendis *ei tee*; *ei* saab märgendi
Degree=Cmp võrdlusaste: komparatiiv
Degree=Pos võrdlusaste: positiiv
Degree=Sup võrdlusaste: superlatiiv
Foreign=Yes võõrkeelne sõna
Hyph=Yes 1)sidekriipsuga lõppev sõna, nt politsei- ja piirivalveamet; 2) sidekriips kirjavahemärgi alaliigina
Mood=Cnd kõneviis: konditsionaal
Mood=Imp kõneviis: imperatiiv
Mood=Ind kõneviis: indikatiiv
Mood=Qot kõneviis: kvotatiiv
Negative=Neg kõneliik: eitav
Number=Plur arv: mitmus
Number=Sing arv: ainsus
NumForm=Digit arvsõna: numbritena
NumForm=Letter arvsõna: sõnana
NumForm=Roman arvsõna: Rooma numbritena
NumType=Card arvsõna: põhiarv
NumType=Ord arvsõna: järgarv
Person=1 isik: 1
Person=2 isik: 2
Person=3 isik: 3

Poss=Yes possessiivne (ainult asesõna kohta)
PronType=Dem asesõna: demonstratiiv
PronType=Ind asesõna: indefiniitne
PronType=Int,Rel asesõna: interrogatiiv-relatiivne
PronType=Prs asesõna: personaalne
PronType=Rcp asesõna: retsiprookne
PronType=Rel asesõna: relatiivne
PronType=Tot asesõna: totaalne e kollektiivne
Reflex=Yes refleksiivne (ainult asesõna kohta)
Tense=Past aeg: minevik
Tense=Pres aeg: olevik
VerbForm=Fin verbi vorm: finiitne
VerbForm=Ger
VerbForm=Inf verbi vorm: infiniitne
VerbForm=Part verbi vorm: partitsiip
VerbForm=Sup verbi vorm: supiin
Voice=Act tegumood: aktiiv
Voice=Pass tegumood: impersonaal ja passiiv

Sõltuvussüntaktiline märgendus
Sõltuvussüntaktilise analüüsi puhul esitatakse kogu lausestruktuur kahe sõnavormi vaheliste ebasümmeetrilise suhtena (põhi e ülemus - laiend e alluv) ja sellel suhtel on nimi (süntaktiline funktsioon). 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.

UD üldpõhimõtted, lühidalt. Pikemalt vt viidatud UD lehekülge.
Universal Dependencies' süntaktiline märgendus esitab sõnadevahelised sõltuvussuhted koos nende süntaktiliste funktsioonide nimetustega. Sõltuvuste nimetuste (süntaktiliste funktsioonide) taksonoomia aluseks on eristus tuumargumentide (subjektid, objektid, seotud infiniittarindilised või osalauselised laiendid (*clausal complements*)) ja ülejäänud argumentide e seotud laiendite vahel. Samas ei üritata eristada seotud obliikva laiendeid vabadest laienditest. Obliikvalised argumendid ja vabad laiendid märgendatakse vastavalt nende sõnaliigilisele kuuluvusele. Nii näiteks saavad nimisõnaline täiend (*vurrudega kass*) ja nimisõnaline määrus (*läksin poodi*) mõlemad märgendi `nmod`; ka kaassõna juurde kuuluv nimisõna saab sama märgendi ning kaassõna riputatakse nimisõna külge märgendi `case` abil, sest UD süsteemi järgi on semantiline põhi ka süntaktiline põhi.
On keelelisi konstruktsioone, mille jaoks sõltuvusesitus sobib väga hästi ja ka neid, mille puhul ühte konstruktsioonis osalevat sõnavormi teise alluvaks või ülemuseks kuulutada on mõnevõrra kunstlik. Sellisteks konstruktsioonideks on näiteks kaassõna- või kvantoriühendid, verbiahelad, koordinatsioon. Nende keelendite analüüsil lähtub UD süsteem rohkem semantikast kui näiteks eesti keele sõltuvuspuude panga märgendamisel kasutatud kitsenduste grammatika (CG) märgendussüsteem. Nimelt:
- kaassõna ülemuseks on käändsõna (*laua all*);

- kvantori ülemuseks on käändsõna (*kolm meest, pudel piima*);
- verbiahela ülemuseks on leksikaalne verb, mitte finiitne abiverbi, modaalverbi jms vorm (*pean tegema*), kolmest ja enamast komponendist koosnevat verbiahelat (*oleks pidanud ette nägema*) ei märgendata „ahela" vaid „põõsana";
- koordineeritud üksused (*Luik, haug ja vähk*) on CG süsteemis samuti esitatud „ahelana" ning UD süsteemis „põõsana". Koordineeritud sõnavormide ülemuseks on esimene koordineeritud element.

Lisaks on väga oluline erinevus öeldistäitelausete e predikatiivlausete märgendamisel: koopulaga predikatiivlauses (J*üri on õpilane, Jüri on pikk*) on CG süsteemis *olema*-verb ülemus ja (osa)lause juurtipp, UD süsteemis on selleks predikatiiv (*õpilane, pikk*) ning koopulana toimiv *olema*-verbi vorm allub predikatiivile ja saab abiverbi märgendi `cop`, ka subjekt märgendiga `nsubj:cop` allub predikatiivile.
UD süsteem ei erista osalauseid ja infiniittarindeid (lauselühendeid), näiteks saavad sama märgendi täiendkõrvallause ja täiendina kasutatav infiniitne verbivorm.
Ka võrdsustab see süsteem verbi infiniitsed laiendid ja EKG II mõistes ahelverbi infiniitsed osad, st verbiahelaid (v.a. verbi liitvormid ja modaalkonstruktsioonid) ei üritatagi jagada ahelverbideks ja verb + laiend konstruktsioonideks. Kasutusel on küll abiverbi märgend `aux`, mille saavad verbi *olema* vormid liitaegades ning verbid *saama, võima* ning *pidama* modaalkonstruktsioonides.
Ülejäänud finiitverbi ühendites infiniitsete verbivormidega saavad infiniidid märgendi `xcomp` või `ccomp`.

Eesti keele UD sõltuvuspuude pangas on kasutatud järgmisi **süntaktiliste funktsioonide märgendeid**:

`root` – lause juurtipp, öeldisverb: *Kass <u>nägi</u> koera*, mitmesõnalise öeldise puhul infiniitne komponent: *Kass oli koera juba <u>näinud</u>, Kass võis koera <u>näha</u>.* Predikatiiv- e öeldistäitelause juurtipuks on öeldistäide: J*üri on <u>õpilane</u>, Jüri on <u>pikk</u>.*

Tuumargumendid

`nsubj` – käändsõnaline subjekt, nt *<u>Kass</u> nägi koera.*

`nsubj:cop` – predikatiivlause käändsõnaline subjekt, nt *<u>Kass</u> on triibuline.*

`csubj` – infiniitne või osalauseline subjekt, nt *Tüdrukule meeldib <u>tantsida</u>. Tundus, et oleme asjast aru <u>saanud</u>.*

`csubj:cop` – predikatiivlause infiniitne või osalauseline subjekt, nt *Laenu on kerge <u>võtta</u>. Tema sõnul on väheusutav, et vaatajate arv edaspidi <u>tõuseks</u>.*

`dobj` – käändsõnaline objekt, nt *Kass nägi <u>koera</u>*. da-infinitiivne objekt saab märgendi `xcomp`.

`xcomp` – Eesti UD selles versioonis sisuliselt kõik verbi seotud infiniitsed laiendid, v.a. da-infinitiivne öeldistäide, mis saab märgendi `ccomp`. Märgendi `xcomp`

saavad mh:
ahelverbi infiniitsed osad, välja arvatud modaalverbide *saama, võima* ja *pidama* laiendid, nt *hakkan tegema, jäi magama, ajab nutma,*
da-infiniitsed objektid, nt *tahan teha,*
da-infiniitsed verbid otstarbelause öeldisena, nt *tahtis proovida oma tiivakesi, et teada saada.*
Lisaks saavad märgendi `xcomp` ka translatiivsed predikatiivadverbiaalid, nt *President nimetas Juhani ministriks; Ta tahtis saada rikkaks; Need majad on luksusliku eluviisi võrdpildiks* ning essiivsed predikatiivadverbiaalid, nt *See tundus meile olulisena.*

`ccomp` – Märgendi `ccomp` saavad eelkõige komplementlausete tipud: sihitiskõrvallause öeldis ja sihitisena toimiv infiniittarind, öeldistäitelise kõrvallause öeldis ja öeldistäitena toimiv infiniittarind + muude rektsiooniliselt seotud kõrvallausete öeldis ja vastavad infiniittarindid, v.a. aluslaused ja -tarindid, nt *Ma arvan, et Santiagos sajab vihma. Ma tahan, et Santiagos sajaks vihma. Ta küsis, kas ma varsti tulen. Muusikat võib proovida sõnadesse panna.*

Enam-vähem samatähendusliku lause *Santiagos sajab vihma, arvas ta.* puhul on tegemist süntaktilise suhtega nimega `parataxis` ja lause juurtipuks on *sajab.*

Otsekõne puhul on ülemuseks otsekõnelise lauseosa kõrgeim ülemus *sajab* ja süntaktiline suhe on samuti `parataxis.` *„Santiagos sajab vihma", ütles ta.*

Märgendi `ccomp` saab ka da-infinitiivne öeldistäide: *Mõlema hobi on kassipilte netti riputada.*

Muud laiendid
`acl` täiendkõrvallause või täiendina toimiv infiniittarind: *See oli rohkem kui 10 protsenti Hansapanka paigutatud rahast. Feeri otsus suusakeskus üle võtta oli emotsionaalne.*

`acl:relcl` relatiivkõrvallause *See on liigutav lugu kanakarjast, kes otsustab farmist jalga lasta;* ka komplement-relatiivlaused: *Ta välistas täielikult võimaluse, et pangast oleks raha saanud kaduda.*
Ka pealauses korrelaadiga kõrvallaused loetakse relatiivlauseteks: *Meie turustame seda, mida enamus nõuab.*

`advcl` määruskõrvallause või määrusena toimiv infiniittarind: *Politsei ei tee midagi, kuna talle pole teatatud. Ta surus oma tahtmise läbi kellegagi arvestamata.*

`advmod` määrsõnaline laiend (määrus); ka kas kas-küsimuste algul: *Olen seda korduvalt rõhutanud. Kas sa tuled juba?*

`advmod:quant` endine CG süsteemi kvantorfraasi põhi, nt *palju õpilasi, pudel piima*. NB! arvsõnaline kvantor saab märgendi `nummod`, nt viis õpilast.

amod adjektiivne täiend, nt *Triibuline kass lõi nurru.*

appos lisand: *Siseminister Jüri Mõis; firma Sarved ja Sõrad.* Selle märgendi saavad ka pealkirjad ja muud sellesarnased struktuurid: *Kasutusel on termin „laisk raha”. Peeter Sauteri novell „Kõhuvalu”... ;* ka järellisand: *Päikest, ühte sagedamini esinevat kujundit, pole kunagi analüüsitud.*

case kaassõna, nt *Kass ronis diivani alla. Kass hüppas üle diivani.*

det selle märgendiga on puudepanga käesolevas versioonis nimisõna laiendina järgmised sõnad: s*ee, too, ise, oma, kõik, esimene* (asesõnana) , *teine* (asesõnana), *kolmas* (mitmuses, nt *kolmandad riigid*), *miski, nihuke, sihuke, siuke, teistsugune, minusugune, meiesugune, temasugune.*

nmod nimisõnaline määrus, nt *Kass põõnas diivanil*; ka koos kaassõnaga, nt *Kass põõnas palmi all*; nimisõnaline täiend nt *Kassi toidukauss on tühi*; ka koos kaassõnaga, nt *Maja mere ääres on müüa*.

nmod:poss possessiivtäiend; puudepanga selles versioonis omastavas käändes isikulised asesõnad täiendina, nt *minu raamat*

nummod arvsõnaline (sh ka numbritega kirjutatud) laiend või kvantor, nt *aastal 2016. Paadis istus kolm meest. Orkaan tappis sadu inimesi. Selles asulas on 15 800 elanikku.* Viimases näites saab 15 märgendi compound.

Verbiahela osad

aux abiverb: *olema* verbi liitaegades; modaalverbid *saama, pidama, võima* modaalkonstruktsioonides. Ülemuseks on infiniitne leksikaalne verb, nt *olin teinud; saan teha, võin teha, pean tegema.*

cop koopula, verb *olema* öeldistäitelauses, kus öeldistäide (v.a infinitiivne või osalauseline) saab märgendi root ja verbi *olema* vorm allub sellele, nt *Kass on triibuline. See raamat on minu oma.*
Kui koopula on verbi *olema* liitvorm (*Maja oli kunagi olnud punane*), siis ei ripu verbivormid üksteise küljes vaid kumbki eraldi *punase* küljes.

neg *ei* verbi eitava vormi osana; *ära, ärge, ärgu* ja *ärgem* verbi käskiva kõneviisi eitava vormi osana.

Koordinatsioon

cc koordineeriv sidend, ülemuseks on esimene koordineeritud element nt *Luik, haug ja vähk vedasid vankrit.*

cc:preconj lahksidendi esikomponent. Praeguse seisuga saavad selle märgendi: *nii* | *niihästi* |

niivõrd (järelkomponent: *kui*); *kas* (*või*); *küll* (*küll*); *nii | sellepärast* (*et*); *selle asemel | vaatamata | hoolimata* (*et*); *siis | samal ajal* (*kui*); *nii* (*nagu*)

`conj` koordineeritud elemendid. Nende puhul märgendatakse esimene element oma süntaktilise funktsiooni märgendiga ning ülejäänud koordineeritud elemendid alluvad sellele märgendiga `conj`, nt *Luik, haug ja vähk vedasid vankrit*.

Mitmesõnalised sisemise sõltuvusstruktuurita keelendid

`compound` mitmesõnalised arvud, nt *kolm tuhat seitsesada kaheksakümmend viis* märgendatakse nii, et ühendi viimane osis saab ühendi kui terviku süntaktilise funktsiooni märgendi ja ülejäänud osised on selle otsesed alluvad märgendiga `compound.` Nii on märgendatud ka (osaliselt) numbritega kirjutatud arvud, nt *28 miljonit* või *50 000.*

`compound:prt` ühendverbi afiksaaladverbiline osis, nt *leidis üles.*

`name` pärisnime osad. Pärisnime viimane osis märgendatakse pärisnime kui terviku süntaktilise funktsiooniga ja nime ülejäänud osad märgendatakse selle otseste alluvatena ning nad saavad märgendi `name:` *New York, Carl Robert Jakobson, Cantrade Private Bank.*

Muu

`foreign` võõrkeelsed sõnad, nt *Transgeensete ja knock out hiireliinide loomine ...*

`discourse` hüüundid ja diskursuspartiklid nagu *tere, ahah, noh, nojah, appi, aitäh, mhmh* jms.

`list` loendis elementide järjekorranumbrid vms tähised (nt *a,b,c*)

`mark` alistavad sidendid osalause algul; küsisõnad küsilause algul (v.a. *kas,* mis saab märgendi `advmod`), *kui, otsekui, justkui* võrdlustarindites, nt *Supp on kuumem kui päike.*
Selle märgendi saavad järgmised sõnad: *ehkki , et , justkui , kuhu , kui , kuidas , kuigi , kuivõrd , kuna , kuni , kus , kusjuures, kust , kustkohast , miks, mil , millal , milleks , nagu , otsekui , selmet , sest* ja kõrvallause alustajana lühend *st* .

`parataxis` – nõrgalt seotud lauseosa, kõrvuasend. Kasutatud otse- ja kaudkõne saatelause märgendamiseks, sellise lause juurtipp on otse- või kaudkõnes lauseosa kõrgeim ülemus. *Santiagos sajab vihma, arvas ta. „Santiagos sajab vihma”, ütles ta.*
Enam-vähem samatähenduslike komplementkõrvallausetega struktuuride puhul on juurtipuks saatelause põhiverb (*arvan, tahan*) ja komplementlause põhiverbi (*sajab, sajaks*) süntaktilise funktsiooni märgend on `ccomp`: *Ma arvan, et Santiagos sajab vihma. Ma tahan, et Santiagos sajaks vihma.*

`punct` punktuatsioon

`vocative` üte: *Head uut aastat, kallid tartlased!*