***Universal Dependencies*' eesti keele märgendusskeem**

Universal Dependencies on projekt, mille eesmärgiks on luua ühtne, tüpoloogiliselt relevantne morfoloogiline ja sõltuvussüntaktiline märgendussüsteem võimalikult paljude keelte jaoks.

Eesti keele *Universal Dependencies* sõltuvuspuude pank on loodud Eesti keele sõltuvuspuude panga teisendamise teel *Universal Dependencies*' kujule.

UD morfoloogiline märgendus

Sõnaliigid

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
VERB verb
X muu

Morfoloogilised kategooriad

AdpType=Post adpositsiooni liik: postpositsioon

AdpType=Prep adpositsiooni liik: prepositsioon

Abbr=Yes lühend

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: komitatiiv

Case=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

Degree=Cmp võrdlusaste: komparatiiv

Degree=Pos võrdlusaste: positiiv

Degree=Sup võrdlusaste: superlatiiv

InfForm=Inf infiniitne verbivorm: da-infinitiiv

InfForm=SupAbe infiniitne verbivorm: supiini abessiiv

InfForm=SupEla infiniitne verbivorm: supiini elatiiv

InfForm=SupIll infiniitne verbivorm: supiini illatiiv

InfForm=SupIne infiniitne verbivorm: supiini inessiiv

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 asesõna: interrogatiivne

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=Inf verbi vorm: infiniitne

VerbForm=Part verbi vorm: partitsiip

VerbForm=Sup verbi vorm: supiin

VerbForm=Trans verbi vorm: transgressiiv

VerbType=Intr intransitiivne verb

VerbType=NGP aspektiverb

VerbType=Part partitiivne verb

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), sellel suhtel võib olla 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 (*Jüri 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.

UD märgendusskeemis on rikkalik märgendite repertuaar mitmesõnaliste leksikaalsete üksuste jaoks (mwe, compound, name); selle poolest erinevad UD märgendid positiivselt eesti keele kitsenduste grammatika märgenditest.

Eesti keele UD märgendid

Tuumargumendid

nsubj – käändsõnaline subjekt, nt *Kass nägi koera*.

nsubj:cop – predikatiivlause käändsõnaline subjekt, nt *Kass on triibuline.*

csubj – infiniitne või osalauseline subjekt. Eesti UD käesolevas versioonis saavad selle märgendi da-infiniitsed subjektid, nt *Tüdrukule meeldib tantsida.*

csubj:cop – predikatiivlause infiniitne või osalauseline subjekt, nt *Noor olla on kevadet rinna sees kanda.* Sellises lauses on praegu, ebajärjekindlalt, aga tulenevalt UD juhendist koopula st *olema*-verbi vorm (osa)lause juurtipp.

dobj – käändsõnaline objekt, nt *Kass nägi koera.* 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* jne,

da-infiniitsed objektid *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 tegi selle raskeks.* ning essiivsed predikatiivadverbiaalid verbide *näima, paistma, tunduma, näikse, püsima, säilima, seisma, toimima, funktsioneerima, esinema, käituma, avalduma, teenima, töötama, käibima, kehtima, nägema, teadma, tundma* laiendina.

ccomp – eesti UD selles versioonis saab selle märgendi ainult da-infinitiivne öeldistäide, nt *Tema eesmärk on ellu jääda. Olema*-verb sel juhul on osalause juurtipp märgendiga root.

Muud laiendid

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.*

appos – lisand. Lisandina on selles versioonis märgendatud nimisõnaline nominatiivne lisand, nt *professor Tammele* ning nimisõnaline ühilduv lisand, nt *ülikoolilinnas Tartus*.

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.

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

advcl – infiniitne määruslik laiend, nt *Koer jooksis saba liputades mööda tänavat*. *Pikalt mõtlemata asus ta asja kallale*; ka võrdlustarind, nt *ta on tuntud kui läänemeelne poliitik*

advmod – määrsõnaline laiend (määrus); ka *kas* kas-küsimuste algul

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

neg – *ei* verbi eitava vormi osana (*ära* ja *ärge* saavad märgendi aux)

acl – nimisõna infiniitne täiend, sh ka partitsiiptäiendid, nt *Õpetaja andis talle loa koju minna. Ema küpsetatud kook maitseb hea. Haukuv koer ei hammusta.*

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

Special clause dependents

vocative – üte, nt *Mari, tule palun siia!.* Eesti keele sõltuvuspuude panga kitsenduste grammatika märgendussüsteemiga (CG) märgendatud versioonis olid ütted märgendatud subjektideks, nüüd on nad eraldi märgendiga.

aux – abiverb: *olema* verbi liitaegades; modaalverbid *saama, pidama, võima* modaalkonstruktsioonides; *ära* ja *ärge* verbi käskiva kõneviisi eitavates vormides. Ülemuseks on infiniitne leksikaalne verb, nt *olin teinud*; *saan teha, võin teha, pean tegema; ära tee, ärge tehke.*

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.

mark – alistavad sidendid osalause algul; küsisõnad küsilause algul, k*ui, otsekui, justkui* võrdlustarindites, nt *Supp on kuumem kui päike.*

Hetkel saavad selle märgendi sõnad *nagu, kui, siis, miks, kuidas, millal, mil, kus*.

discourse – hüüundid nagu *tere, ahah, noh, nojah, appi, aitäh* jms.

Koordinatsioon

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.*

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* | *enam* (*et*); *siis* | *samal ajal* (*kui*); *nii* (*nagu*)

punct – punktuatsioon

Muu

root – lause juurtipp, pealause öeldisverb, verbi liitvormi või ahelverbi puhul põhitähendust kandev komponent, nt *Sa oled palju ära teinud. Võid nüüd sööma hakata.* Öeldistäitelauses on juurtipuks öeldistäide, nt *Kass on triibuline.*

UD süsteemis ei ole da-infiniitset öeldistäidet, da-infiniitne verbivorm lausetes nagu *Minu eesmärk on hea välja näha.* või *Noor olla on kevadet rinna sees kanda.* saab märgendi ccomp ning nende lausete juurtipp on *olema*-verbi vorm.

dep – spetsifitseerimata sõltuvus; praeguses eesti UD versioonis on sellega ühendatud kõik osalaused.

Mitmesõnalised üksused (sisemise struktuurita sõnaühendid)

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 muud numbrijadad, nt lauses *Kohtumine lõppes seisuga 1:2* on '2' '1' alluv märgendiga compound.

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, nt *New York, Carl Robert Jakobson*. Praeguses versiooni märgendataksegi suhtega name ainult isikunimed ja väike hulk kohanimesid.

Nõrgalt seotud suhete märgendid **(loose joining relations)**

parataxis – kasutatakse otsekõne saatelause põhiverbi märgendamiseks, nt „*Kuidas elad?” küsis Mari.*

Väljajättelised struktuurid

Täislause juurtipuks on öeldisverb, predikatiivlausetes predikatiiv. Kuid loomulikus tekstis esineb palju väljajättelisi struktuure või mittetäielikke lauseid, milles öeldisverb puudub. Nendes on juurtipuks „tähtsaim” moodustaja – kui subjekt on olemas, siis see, kui “lauseks” on ainult nimisõnafraas, siis

märgendatakse selle põhi kui juurtipp.

Koordineeritud osalauseid, milles teise osalause verb on välja jäetud (nt *Kassid sõid puru ja koerad konservi*) märgendatakse nii, et teise osalause „tähtsaim” moodustaja, näitelauses subjekt *koerad* „ülendatakse” selle osalause tipuks, mis allub eelmise osalause öeldisverbile *sõid*.

Verbita lauselühendis (*kepp käes, kott üle õla*) on see, mille kohta EKK ütleb „subjektisarnane element” (st *kepp, kott*), subjekt ja selle osalause tipp ning teine osaline selles konstruktsioonis (*kott, õla*) on subjekti laiendav nmod.