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.
Eesti keele Universal Dependencies veebitekstide sõltuvuspuude pank EWTB on loodud tekste käsitsi märgendades.

UD sõltuvuspuude pankades on märgendatud:
1. lemma e algvorm
2. sõnaliik
3. morfoloogilised kategooriad (*features*)
4. ülemus sõltuvuspuus
5. sõltuvussuhe (*relations*)
6. Osades failides osades versioonides on märgendatud ka nn täiustatud sõltuvused (*enhanced dependencies)*
7. sõnaortograafia vead

CONLLU formaat

Kooditabel on UTF-8. Ridu on kolme tüüpi:

1. Sõnad, kirjavahemärgid jm tekstiüksused koos nende märgendusega: reas on 10 tabelimärgiga eraldatud välja, täpsemalt allpool.
2. Tühjad read lausepiiride märkimiseks.
3. # märgiga algavad kommentaariread.

Laused koosnevad ühest või rohkematest ridadest. Lauset moodustavate üksuste read koosnevad järgmistest väljadest:

- ID: sõna indeks e järjekorranumber lauses. Iga lause esimene sõna on numbriga 1
  - Kui identifikaator on naturaalarvude vahemik, siis on tegemist mitmesõnalise üksusega ning järgmised read kirjeldavad seda üksust sõne kaupa. Kui identifikaator on kujul naturaalarv . naturaalarv, siis on tegemist elliptilise lausega, milles on puuduv sõna või puuduvad sõnad märgendusse lisatud. Erandkorras võib lisatav sõna paikneda ka lause algul, siis algab indeks arvuga 0.
- FORM: Sõnavorm st tekstisõna, punktuatsioonimärk või muu sümbol
- LEMMA: algvorm
- UPOS: Sõnaliik, vt altpoolt
- XPOS: Sõnaliik nagu need on määratletud siin https://filosoft.ee/html_morf_et/morfoutinfo.html#2
- FEATS: Morfoloogilised kategooriad (*features*), vt altpoolt. Kui sellel üksusel neid pole, siis alakriips _
- HEAD: Ülemus sõltuvuspuus. Lause juurtipu ülemus on 0
- DEPREL:sõltuvussuhte nimi (*relation*), vt altpoolt
- DEPS: Täiustatud sõltuvused (*enhanced dependency graph*) ülemus-alluv paaride loendina. Kui pole märgendatud, siis alakriips _
- MISC: Muu märgendus. Kui pole, siis alakriips _

Väljad peavad vastama järgmistele tingimustele:

1. Väljad ei tohi olla tühjad. Kui vastavat infot pole, siis on väljal alakriips _
2. Muud väljad peale FORM, LEMMA ja MISC ei tohi sisaldada tühikuid.
3. UPOS, HEAD, ja DEPREL väljad ei tohi olla täitmata, välja arvatud juhud, juhul kui need paiknevad mitmesõnaliste üksuste vahemikku kirjeldaval real, siis on nad vaikeväärtusega “_”.

Sõnestamine (*tokenization*)

Üldiselt sõnestatakse nn tavalisel viisil: sõnapiiriks on tühik või reavahetus, kirjavahemärgid tõstetakse sõnadest lahku.
Tühikuteta kuupäev stiilis 07.06.03 on üks sõne.
Valesti kokku kirjutatud sõnad on märgendamise käigus lahutatud ning esimene sõna märgendatud MISC-väljal märgendite kombinatsiooniga SpaceAfter=No|CorrectSpaceAfter=Yes

Kui sõna sisaldab sidekriipsu ja see on korrektne, siis märgendatakse sõna vastavalt tema morfoloogilise informatsioonile, nt

18 must-valge must-valge NOUN S Case=Nom|Number=Sing 15 conj _ _

Kui sõna on koordinatsioonis olev liitsõna esimene pool (nt riist- ja tarkvara), siis märgendatakse sõna selle morfoloogilise märgendusega, mis on selle sõnapoole kohta teada

14 riist- riist NOUN S Hyph=Yes 17 nmod _ _

UD morfoloogiline märgendus

Lemma

- Lemma määramise erijuhud:
- trüki- või kirjaveaga sõna lemma on õige sõna lemma.
- „tsenseeritud" sõnavormi lemma on tsenseerimata sõna, nt *****iidi* lemma on *perseiid*. Kui õige lemma kontekstist üheselt ei selgu, siis jääb muidugi „tsenseeritud" variant.

Sõnaliigid

ADJ adjektiiv. Täiendi positsioonis olevad mineviku partitsiibid on sõnaliigilt omadussõnad, aga omavad ka verbi morfoloogilisi kategooriaid, nt
tehtud tehtud ADJ A Degree=Pos|Tense=Past|VerbForm=Part|Voice=Pass
ADP adpositsioon, ees- ja tagasõna eristatakse tunnuse AdpType=Pre ja AdpType=Post abil.
ADV adverb
AUX abiverb

olema liitaegades, modaalid: *saama*, *võima*, *pidama*
olema koopulalausetes
ei, ära vormid verbi eitava liitvormi koosseisus

CCONJ koordineeriv sidend, nendena on märgendatud *aga, ega, ehk, elik, ent, ja, kui, kuid, kuni (kolm kuni neli kuud), nagu, nii* (liitsidendis *nii ... kui), ning, vaid, või*
DET määratleja (*determiner*). Määratleja kohta eesti keeles vt Erelt ja Metslang „Eesti keele süntaks" lk 382 jj.
Määratlejatena on eesti keele UD puudepankades märgendatud: *see, too, seesama, toosama, sama, esimene, teine, mis* (tähenduses *milline,* nt kasutustes *mis vahe on ...*, *mis asi on ...*, *mis tähtsust sellel on* jne), *iga, kõik, kogu, keegi, miski, üks, ükski, mingi, terve* (tähenduses 'kogu', nt *keetsin terve potitäie suppi*), *muu, mõni, paljud, igasugu, igasugune, mitmesugune, niisugune, niisamasugune, samasugune, selline, seesamune, seesugune, nihuke, sihuke, siuke, säherdune, säärane, selletaoline, taoline*

Määratleja on alati täiendi positsioonis:
See maja on suur – see on DET ; *See on suur maja – see* on PRON
INTJ interjektsioon. Ka üneemid (*ah, mh, no* jm) on märgendatud interjektsioonidena.
NOUN substantiiv
NUM numeraal
PRON pronoomen
PROPN pärisnimi
PUNCT punktuatsioon
SCONJ alistav sidend, sellena on märgendatud *ehkki, et, justkui, kui* (v.a. liitsidendis *nii ... kui*), *kuigi, kuna, kuni, nagu, otsekui, selmet, sest.*

SYM sümbol, nt 50 %, Saab 340B, Ansip & Co. Märgendi SYM saavad ka emotikonid veebitekstides, nt :) ;D ja adressaati tähistav @ EWTB mõnedes foorumitekstides.
VERB verb
X muu, selle märgendi saavad mh muukeelsed sõnad. Kuid kui sõna on süntaktilises mõttes osa eestikeelsest lausest, saab ta oma süntaktilisest funktsioonist tuleneva sõnaliigi.
Nt lauses *mis eelmise kahe basho grotesksete edutamiste kõrval eriliselt silma torkab* on *'basho'* nimisõna ainsuse omastavas, kuid lauses *Mis teksti sisu on (do not retell)* on sulgudes fraasi liikmed sõnaliigi märgendiga X.

Morfoloogilised kategooriad (*features*)

AdpType=Post adpositsiooni liik: postpositsioon

AdpType=Prep adpositsiooni liik: prepositsioon

Abbr=Yes lühend (sõnaliik vastab tähendusele/kasutusele, nt *jne* on ADV)

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

Connegative=Yes: verbi eitava liitvormi osa

Degree=Cmp võrdlusaste: komparatiiv

Degree=Pos võrdlusaste: positiiv

Degree=Sup võrdlusaste: superlatiiv

Mood=Cnd kõneviis: konditsionaal

Mood=Imp kõneviis: imperatiiv

Mood=Ind kõneviis: indikatiiv

Mood=Qot kõneviis: kvotatiiv

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

Polarity=Neg kõneliik: eitav (jaatavat pole märgendatud)

Poss=Yes possessiivne. Praegu on nii märgendatud ainult asesõna *oma*

Märgend PronType võib olla ka muul sõnal kui asesõnal.

PronType=Dem: demonstratiiv: *see, too, seesama, toosama, sama, esimene, teine*

PronType=Ind: indefiniitne: *keegi, miski, üks, igaüks, muu, mõned, paljud*

PronType=Int , Rel: interrogatiiv-relatiivne: *kes, mis, kumb, missugune, milline, mitu, mitmes*

PronType=Prs: personaalne: *mina, sina, tema, meie, teie, nemad, oma, ise, iseenda, omaenese*

PronType=Rcp: retsiprookne: *üksteise, teineteise*

PronType=Tot: totaalne e kollektiivne: *kõik, iga, kogu, terve*

Reflex=Yes refleksiivne: *ise*, *enda, iseenese, iseenda,*

Tense=Past aeg: minevik

Tense=Pres aeg: olevik

Typo=Yes sõnaortograafia viga. Praegu kasutusel ainult osades EWTB failides.

VerbForm=Fin verbi vorm: finiitne

VerbForm=Inf verbi vorm: infiniitne (da-infinitiiv)

VerbForm=Part verbi vorm: partitsiip

VerbForm=Sup verbi vorm: supiin (ma-infinitiiv)

VerbForm=Conv verbi vorm: konverb (des-vorm)

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 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 https://universaldependencies.org/u/overview/syntax.html.

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.

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)*
- 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.

UD süsteem ei erista osalauseid ja infiniittarindeid (lauselühendeid), näiteks saavad sama märgendi täiendkõrvallause (relatiivlause) 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 ja koopulalausetes 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 (fixed, flat, compound); selle poolest erinevad UD märgendid positiivselt eesti keele kitsenduste grammatika märgenditest.

Koopulalaused

Kui lause põhiverbiks on verb *olema*, loetakse lause koopulalauseks ja juurtipuks ei ole mitte *olema* vorm, vaid mingi teine element lauses ning koopulana toimiv *olema*-verbi vorm allub sellele ja saab abiverbi sõnaliigi märgendi AUX ning koopula süntaktilise märgendi cop. cop-l ei tohi olla alluvaid, st kõik, mis muidu on öeldise alluvad, on nüüd selle juurtipuks määratud sõna alluvad.

Koopulalauseteks EI OLE järgmised *olema*-verbi sisaldavad laused:

1. Need, kus on ainult *olema*-verb ja selle subjekt (pluss viimase täiendid); võib olla ka veel nn modaaladverb (*Oli tore õhtu*, *Raha ei ole ju, Raha küll ei ole*).
2. Need, kus *olema*-verb on ühendverbi osa (talle allub sõna märgendiga compound:prt)

   olema-ga ühendverbid: t*arvis, vaja, alles, üle, läbi, kohal, juures olema* Seda hulka võib vajadusel täiendada

3. mas-vormis alluvaga verbi *olema* vorm, kus mõlemad osalised on sõnaliigiga VERB ja

olema vorm on osalause juurtipp. (mitte *ta on söömas tüüp,* aga *ta oli vette hüppamas tüüp*). Sagedasim selline konstruktsioon on *olemas olema*. mas-vorm selles konstruktsioonis on xcomp.

4. kui öeldistäide on da-infinitiiv (mis siis on ccomp)

Öeldistäitena esinevad infinitiivid ka lauseis, nagu *Olla tõeline sõber on abistada kaaslast hädas. Taganemine on jätta võiduvõimalus vaenlasele.*

5. Lause koosneb küsisõnast süntaktilise märgendiga mark, *olema*-verbist ja alusest: *Kus on kirves?*

Koopulalause juurtipp määratakse vastavalt järgmisele hierarhiale:

1. öeldistäide. *Kadri on inimene. Maja on suur.*

2. öeldistäitemäärus (t*a oli Valgas õpetajaks, Hiinlased on teistsuguse psühholoogiaga*)

3. öeldistäitesarnane määrsõna (*Kõik on halvasti, Nad olid kahekesi; Tal on klapid peas*). Viimane, välise omajaga lause on praegu nii märgendatud, aga võib-olla mõtleme ümber.

ka: *tulemus oli 5 %. Tulemus oli üle viie protsendi. Aeg esimesel ringil oli 3:20.*

4. omaja ja kogeja (*Tal oli kodus kass; Tal oli kodus külm*)

5. Koht (*Ta oli õhtul kodus; ka Ta oli õhtul õnnetuna kodus*)

6. Aeg (*See oli möödunud aastal*)

7. Viis (*See on nii, et ...*, ka *Sellega on nii, et...*)

Väljajättelised struktuurid ehk ellipsid

Üldpõhimõtted.

Kui väljajäetud elemendil ei ole alluvaid, ei tehta midagi.

Kui väljajäetud elemendil on alluvad, üks neist „ülendatakse" väljajäetu asemele (=saab tema süntaktilise funktsiooni) ja teised, kui neid on, alluvad talle. Nt lauses *Ostsin ühe kollase pliiatsi ja kaks roosat.* on *roosat* conj pliiatsi küljes ja *kaks* allub *roosale*.

Kui väljajäetud element on osalause tipp (öeldis), aga mõni abiverb on alles, siis abiverb ülendatakse põhiverbi kohale.

Muul juhul, kui väljajäetud element on osalause tipp (öeldis) siis ülendatakse üks tema alluvatest osalause tipuks vastavalt hierarhiale nsubj > obj > iobj > obl > advmod > csubj > xcomp > ccomp > advcl > dislocated > vocative. Kui lauses on mitu sama funktsiooniga „orvukest", ülendatakse

osalause tipuks see, mis on lausealgulisem.

Need sõnad, mille ülemus peaks olema see väljajätteline öeldis, on „ülendatud" moodustaja alluvad märgendiga orphan, millel on alaliigid orphan:obj, orphan:obl jne vastavalt sellele, mis oleks nende funktsioon väljajätteta lauses. Erandiks funktsioonisõnad, nt sidendid, nemad ei saa märgendit orphan, vaid oma tavalise märgendi.

Verbita lauselühendid (*kepp käes, kott üle õla*) on analüüsitud kui koopulalaused: see, mille kohta EKK ütleb „subjektisarnane element" (st *kepp, kott*), on subjekt ja käändeline vorm on tarindi juur, tüüpiliselt funktsiooniga advcl.

Muud sõltuvusstruktuuri küsimused

Finiitse ja infiniitse verbi ühendites tekib sageli küsimus, kummale verbile ülejäänud lauseliikmed peaksid alluma. Mõnes lauses on see täiesti selge, mõnes mitte. Subjekt allub igal juhul finiitsele verbile. Lausetüübis *keelama/käskima/laskma/paluma* + kellelgi + da-infinitiiv, nt *keelasin koeral maad kraapida on a*lalütlevas moodustaja *koeral* sisuliselt mõlema verbi alluv, aga paneme ta finiitverbi (*keelan, käsen*, jne) alluvaks.

Mõned näited.

Ometi suutis ta ennast ka nõnda maailma maleeliiti suruda. - *ometi* ja *ta* alluvad *suutis*-le, *ka, nõnda, maleeliiti* alluvad *suruda*-le.

Ta hoiab selle enese teada. - *ta* ja *selle* alluvad *hoiab*-le, *enese* allub *teada*-le.

Eesti keele UD süntaktilised märgendid (*relations*)

Tuumargumendid

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

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

csubj – infiniitne või osalauseline subjekt. Infiniitidest saab subjektiks olla ainult da-infinitiiv: *Tüdrukule meeldib tantsida.*

Osalauselise subjekti näide: *Aga mulle tundub , et kogu maailm ootab muusikamaailmalt midagi erutavalt uut minimalismi kõrvale.*

csubj:cop – koopulalause infiniitne või osalauseline subjekt, nt *Imelik, et ma seda veel näidata*

julgen.

obj – käändsõnaline objekt, nt *Kass nägi koera.* da-infinitiivne objekt saab märgendi xcomp.
xcomp – Märgendi xcomp saavad:

1. ahelverbi infiniitsed osad, välja arvatud modaalverbide *saama, võima ja pidama* laiendid, nt *hakkan tegema*, *jäi magama, ajab nutma* jne,
2. da-infiniitsed objektid *tahan teha,*
3. da-infiniitsed verbid otstarbelause öeldisena, nt *tahtis proovida oma tiivakesi, et teada saada.*
4. 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.

Kuid verbi vaba laiend translatiivis või essiivis saab märgendi obl, nt *Juttu jätkus kauemaks.*

Verbi laiendite jaotus xcomp ja obl vahel vt sellest tabelist:
https://docs.google.com/spreadsheets/d/1ADNYGgymecXIiSBHdB5eizoiq4WXXVHmD5sPSjgUE0M/edit#gid=0

ccomp

1. verbi laiendav komplementlause, nt *Ta ütles, et tuleb homme.* NB! *Tulen homme, ütles ta* – parataxis *Ta ütles: „Tulen homme.”* - parataxis
2. da-infinitiivne öeldistäide, nt *Tema eesmärk on ellu jääda. Olema*-verb sel juhul on osalause juurtipp.

Muud laiendid

obl – nimisõnaline (sh asesõnaline) määrus, nt *Kass põõnas diivanil*; ka koos kaassõnaga, nt *Kass põõnas palmi all*.

nmod - nimisõnaline (sh asesõnaline täiend)

appos – lisand. Lisand saab praegu UD-s oma ülemusele ainult järgneda, mitte eelneda.

appos märgendi on saanud:

1. nimed, pealkirjad, jm, kui on olemas eelnev liigisõna: *arhitekt Boulle on üks minu kangelasi*

   K*ust tuli mõte kirjutada ooper " Writing to Vermeer "? Jällegi täissaalile lugesid oma luulet*

ja tõlkeid marilane Vladimir Kozlov , komilane Niina Obrezkova , liivlane Valt Ernstreit , soomlane Kari Sallamaa jt .

2. „meie" mõistes järellisand: *Keegi küsis , kuidas võidi Sallamaa kutsuda Ižkari detsembris , kõige tröötitumal aastaajal .*

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

NB! nummod on ainult täiendi märgend, muus funktsioonis arvsõnad märgendatakse vastavalt oma funktsioonile: nt *jagas kolmeks* on xcomp

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

advcl

1. määruskõrvallause, nt *Kui sa tuled, too mul lilli.*
2. infiniitne määruslik laiend, nt *Koer jooksis saba liputades mööda tänavat*. *Pikalt mõtlemata asus ta asja kallale*
3. 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*. NB! arvsõnaline kvantor saab märgendi nummod, nt *viis õpilast.*

acl – nimisõna infiniitne täiend, sh ka partitsiiptäiendid, nt *Õpetaja andis talle loa koju minna. Ema küpsetatud kook maitseb hea. Haukuv koer ei hammusta. See, et päike tõuseb iga päev, teda ei lohuta.*

acl:relcl Täiendkõrvallaused: *Mees, kes seal seisab, on minu isa.*

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

Special clause dependents

vocative – üte, nt *Mari, tule palun siia!*

Hooligan88 , kas sul on sidemeid-tutvusi mille kaudu see Viimsi muuseumi külastus kokku leppida?

Käskivas kõneviisis verbi subjekt ei ole üte, vaid subjekt, nt *Sina ära mind käsuta!*

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

cop – koopula, verb *olema* koopulalausetes, kus öeldistäide (v.a infinitiivne või osalauseline) vm moodustaja saab märgendi root ja verbi *olema* vorm allub sellele, nt *Kass on triibuline*. *See raamat on minu oma. Mari on kodus.*

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, *nagu*, k*ui, otsekui, justkui* võrdlustarindites, nt *Supp on kuumem kui päike.* Sõnaliigiliselt on need adverbid ADV või alistavad sidendid SCONJ.

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

Samuti saavad selle märgendi nn partiklid (*Tõesti või icicic!*) ja emotikonid. Samuti adressaati märkiv sümbol @ teatud foorumitekstides EWTB-s.

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 järgnev koordineeritud element nt *Luik, haug ja vähk vedasid vankrit.*

Ka lause alguses olev *aga* on cc. *Aga ilm on täna ilus.*

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. Punktuatsioon ei ole muidugi tegelikult lause süntaktilise struktuuri osa, nende ülemuste määramine käib järgmiselt. Lauselõpumärk allub juurtipule, välja arvatud juhul, kui sellest tekkiks ristuv kaar. Sulud, jutumärgid jm paariskirjavahemärgid alluvad nende vahel oleva konstruktsiooni kõrgeimale ülemusele, v.a. juhul, kui sellest tekiks ristuv kaar.

Sidendite ja punktuatsioonimärkide ülemuseks on vahetult järgnev konjunkt.

Muu

root – lause juurtipp, pealause öeldisverb, verbi liitvormi või ahelverbi puhul põhitähendust kandev komponent, nt *Sa oled palju ära teinud. Võid nüüd sööma hakata.* Koopulalause juurtipu määramise kohta vt Koopulalaused.

dep – spetsifitseerimata sõltuvus. St alluvussuhe on selgelt olemas, aga funktsiooni pole võimalik

määrata.

Nt EWTB-s konstruktsioonid ‘kes te’ lausetes nagu *Lumehelbekesed, kes te tahate täielikku karantiini, kolige …*

on ‘te’ ‘tahate’ subjekt ning ‘kes’ allub samuti ‘tahate’-le, aga suhtega dep Pole siiski kindel, kas nii on õige.

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.

Aga tühikuga numbritega kirjutatud arvud (100 000) tuleb märgendada märgendiga goeswith

compound:prt ühendverbi afiksaaladverbiline osis, nt *leidis <u>üles</u>*. Ühendverbid on avatud hulk.

flat Sellega märgendatakse eksotsentrilised, st selge ülemuseta sõnaühendid. Märgendi flat puhul on ülemuseks alati mitmesõnalise üksuse esimene komponent ja teised alluvad talle. Märgendi flat saavad mh

- pärisnime osad. Pärisnime esimene osis märgendatakse pärisnime kui terviku süntaktilise funktsiooniga ja nime ülejäänud osad märgendatakse selle otseste alluvatena märgendiga flat, nt *New York, Carl Robert Jakobson*. Praeguses versiooni märgendataksegi suhtega name ainult isikunimed ja väike hulk kohanimesid. Kuid kui nimel on süntaktiline struktuur, märgendatakse teda selle struktuuri järgi, nt *Tartu Ülikool*.

- Võõrkeelsed fraasid, nt *siis mõeldi just entry <u>level</u> kaamerate hindu*; nn *süsinikneutraalsele (carbon <u>neutral</u>) tasemele*

See on ka see märgend, millega märgendada mitmesõnalisi üksusi, mis compound ja fixed alla ei mahu. St kui on kahtlusi, on märgend flat.

fixed – sellega märgendatakse grammatiseerunud sõnaühendeid, mis süntakstiliselt „töötavad” funktsioonisõnade või määrustena.

Sellena on märgendatud (alati tähendab "alati siis, kui nende vahel pole koma")

SIDENDID

ainult et - alati

enam kui - osad

enne kui - alati

eriti kui - alati

ilma et - alati

isegi kui - alati

just kui - alati

just nagu - alati

nii et - alati

nii kui - alati

nii nagu - alati

niipalju kui - alati

nõnda et - alati

nõnda nagu - alati

peaaegu et - alati

rohkem kui - osad

samas kui - alati

samuti kui - osad

samuti nagu - osad

seeasemel et - alati

seni kuni - alati

ükskõik kuhu alati

ükskõik kui alati

ükskõik kuidas - alati

ükskõik kus - alati

ükspuha kus - alati

vaat et - alati

vaata et - alati

vähem kui - osad

MUU, st mitte-sidendid

- pseudoühilduvad väljendid sõnadega 'pool', 'poole', 'poolt', mh

igale poole, igal pool, igalt poolt - NB! 'iga' on siin PRON 'iga', mitte DET

- ühendid sõnadega 'iganes' ja 'tahes', mh *mis iganes, mis tahes*, *kus iganes, kus tahes*, jne

tahes tahtmata

- ühendid sõnaga 'teab', mh *teab mis, teab kus, teab milline, teab mitmes*
- *seda enam* konstruktsioonides *seda enam, et ..* aga muidugi mitte konstruktsioonis *seda enam ei tehta*
- *eks ju* (EWTB-s)
- *mis siis* lausetes nagu *mis siis sellest*, *mis siis ikka*
- *kas või* lausetes nagu *Meenutagem kas või noort Peeter Volkonskit.*

Nõrgalt seotud suhete märgendid (*loose joining relations*)

parataxis

1. Kasutatakse otsekõne saatelause põhiverbi märgendamiseks, nt „*Kuidas elad?” küsis Mari.*

   Ka selline kaudkõne nagu *Tulen homme, ütles ta.*

2. Kõrvuasendiga osalaused, mille vahel on mingi muu kirjavahemärk kui koma, nt *Eesti Pank tunnistab: meie käed on liiga lühikesed! Narvas tegutsevad fašistid -- nii väidab 14. mai ajaleht Molodjž Estoni. Kuid tegelik probleem on üldtuntud: 1996.-1997. aasta tarbimispidu hakkab kätte maksma.*
3. Sulgudes osalaused ja lauseosad, nt *Samuti on ajaleht avaldanud vastulause juures oma kommentaari (08.01.99).Antud küsimuses laiema kultuurilise hoiaku (mis asi see on?) saavutamiseks* ... Samas: kui sulgudes on tõesti sama asi teise sõnaga öeldud, nt TÜ (Tartu Ülikool) või Tartu Ülikool (TÜ), siis on suhe appos.
4. EWTB foorumites kõneleja/kirjutaja riputatakse lause juurtipu külge märgendi parataxis abil: *anarchy: Aga ongi ju.*

list UD juhendi järgi peaks sellega märgendatama loendeid, aadresse jm loendisarnaseid asju. Praegu on EDT-s märgendi list saanud aadressid, nt *Tartu, Kreutzwaldi 1* või struktuurid nagu *Teenuse kasutamiseks tuleb saata SMS sõnum kujul "PEATUS peatuse nimi liini number" lühinumbrile 1311.*

Üldiselt on siiani märgendit ’list’ kasutatud vähe ja ebajärjekindlalt.

orphan sellega märgendatakse elliptilise öeldisega lausete elemente, mis muidu peaksid alluma öeldisele. Täpsemalt vt väljajättelised struktuurid.

Märgendil orphan on alaliigid, mis näitavad ’orvu’ süntaktilist funktsiooni juhul, kui öeldis oleks alles: orphan:obj, orphan:obl, orphan:advmod jne.

goeswith sellega märgendatakse sõnad, mis ortograafiareeglite järgi peaksid olema kokku kirjutatud. Lisaks märgendatakse sellega ka numbritega kirjutatud arvud, mille keskel on tühik, nt 100 000. Ülemuseks on esimene komponent. Selle alla käivad ka tühikutega telefoninumbrid.

Täiustatud sõltuvused (*Enhanced Dependencies*)

Enhanced märgendus on järgmiste nähtuste lisamärgendus:

1. nn nulltipud (*null nodes*) elliptiliste predikaatide jaoks
2. konjunktsiooniseoses olevate elementidele jagatud laiendite lisamine (*propagation of conjuncts*)
3. nn kontrolli ja tõste konstruktsioonides samasubjektilisuse kaarte lisamine (*additional subject relations for control and raising constructions*)
4. samaviitelisuse märgendamine relatiivlausetes (*coreference in relative clause constructions*): relatiivkõrvallauset alustavale asesõnale, lisatakse kaar, mis näitab temaga samaviitelist sõna pealauses.