UD2 käsitsi kontrollimine

Taust. On olemas sõltuvussüntaktiliselt märgendatud korpus: Eesti keele sõltuvuspuude pank (https://github.com/EstSyntax/EDT)
Ja on olemas Universal Dependencies - http://universaldependencies.org/ - aktiivselt arendatav keelest sõltumatu sõltuvussüntaktiline märgendusskeem ja selle järgi märgendatud sõltuvuspuude pangad rohkem kui 40 keeles. Aktiivselt arendatav tähendab mh ka seda, et märgenduse põhimõtted ja märgendite repertuaar on pidevas muutumises.
Eesti keele puudepanka on teisendatud UD 1.0 kujule ja UD 1.3 kujule ja ... ja UD 2.0 kujule ja õpuks on ta jõudnud kujule UD 2.1. Kuna suur osa teisendusi on tehtud automaatselt, siis on muidugi sisse tulnud vigu, st teatud keelendid tuleb käsitsi üle kontrollida. Lisaks tuleb UD 2.1. järgi märgendada ja eristada osa keelendeid nii, et seda automaatselt teha ei saagi ning need eristused tuleb teha käsitsi.

Sellel juhendil on kaks osa: tehniline ja sisuline pool.
Tehniline.
Kasutame tarkvara Annotatrix https://github.com/jonorthwash/ud-annotatrix
Pluss: visualiseerib ja esitab lause puu ning tekstilise kuju paralleelselt. Saab parandada tekstilist kuju ning tulemus on puus kohe näha. Tarkvarana on ta „toores", tarkvara loojad tahavad vigade ja puuduste kohta tagasisidet saada.
Selle juhendi kirjutamise ajal olid ka „sissetulevad" ja „väljaminevad" kaared halvasti eristatud, aga selle lubasid tarkvara autorid lähiajal ära parandada.
Miinus on ka see, et ei saa failis otsida. Nii et ilmselt tuleb otsida mingis tekstiredaktoris, vaadata sealt lause nr ja siis parandada Annotatrixis. Kogemuse tekkimisel ilmselt õnnestub väikese ulatusega nähtusi parandada ka ainult tekstiredaktoriga, saab rutem. Kindlasti on Annotatrix vajalik siis, kui tuleb muuta lausepuud, nt kui koopulalause märgendamisel määratakse lausele uus juurtipp.
Ristuvaid kaari (jah, need said EDT-s parandatud, aga teisendamised tekitasid uued) tuvastab programm ~/svnkrpsoft/trunk/kasitoo/syntax/kaared_risti.py
Ristumised võivad olla paratamatud, aga võivad aidata ka tuvastada märgenduse vigu. Sageli läheb risti mõne kirjavahemärgi kaar, see tuleb alati ümber sättida nii, et ristumist ei tuleks.
Programmi kaared_risti.py tulemuse saab lasta tekstifaili, parandatud ristumised tekstifailist kustutada. Kui nüüd allesjäänud, paratamatuid ristumisi sisaldav fail anda programmile ette lipuga -e, siis ta jätab need read väljundisse kirjutamata. Juhul, kui ristuvate kaarte info suunata faili, siis ekraanile väljastab programm ikkagi muud veateated (puuduv tabulaator reas, puuduv reanumber jne).

Sisuline.
UD märgendisüsteemi kirjeldus on siin: http://universaldependencies.org/guidelines.html
Kontrollida tuleb:
ManualCheck=Yes
nmod nime laiendina – kas on õige nmod (täiend) või peaks olema märgend appos (lisand)
Põhierinevus eesti traditsioonilisest lisandikäsitlusest on see, et fraasis 'direktor Karl Kask' on eesti tradistioonis direktor lisand, UD-s aga vastupidi: direktor on põhi, Karl lisand (märgendiga appos) ja Kask nime osa (märgendiga flat, ülemus Karl).
NB: appos saab oma peasõnale ainult järgneda, mitte eelneda. appos kasutusnäiteid vt ka UD dokumentatsioonist.
Eelnevaga seostuv **appos** kontroll: kõik appos märgendid tuleb üle kontrollida,sellele viitab ManualCheck=Appos
Sõnaliiki AUX kuuluv sõna, millel on alluvad. Neid ei tohi olla.

AUX on nii modaalverb, *olema* liitaegades kui *ei* verbi eitavas vormis. Kui verbile allub mitu AUXi, peavad nad rippuma verbi küljes põõsana, mitte ahelana: *ei(*AUX; ülemus: teinud*) ole*(AUX: ülemus:teinud) *teinud*
Koopulalause puhul peavad kõik olema-verbi tegelikud laiendid olema selle osalause mitte-verbilise juurtipu küljes: *Möödunud aastal ei olevat suve üldse olnud.* - lause juurtipp on *aastal* ning talle alluvad *ei, olevat, üldse, olnud; möödunud* muidugi ka
ManualCheck=Appos

ToDo
nmod – kontrollida, kas on nmod või peaks olema obl. nmod on nimisõna laiend, st täiend. obl on verbi, adverbi või adjektiivi laiend, st eesti tradistiooni määrus. UD V1-s seda polnud. Lähemalt vt http://universaldependencies.org/u/dep/obl.html
Kui tegelik verbi laiend koopulastruktuuris käib nimisõna külge (sest koopulal alluvaid olla ei tohi), siis on ta ikkagi obl.
nt *Ta on sel aastal maal külavanemaks* on *aastal* ja *maal* obl, sest nende kuigi nende formaalne ülemus on *külavanemaks*, on nad sisuliselt ikkagi olema-verbi laiendid. *sel* on aastal täiend ja tema märgend on ikka nmod.

On veel mõned erijuhud. *teisel pool kappi*: on praeguse seisuga märgendatud nii, et nii *teisel* kui *pool* mõlemad alluvad *kapile*; *pool* kui case ja *teisel* kui nmod (kuigi *teisel* „päris" ülemus peaks olema *poo*l)

punct-in-coord ja **cc-in-coord** – kas punktuatsiooni ülemus on määratud nagu siin http://universaldependencies.org/u/dep/punct.html nõutud ja kas koordineeriva konjunktsiooni (cc) ülemus on määratud nagu siin http://universaldependencies.org/u/dep/cc.html nõutud.
NB! cc ja punct ülemuse määramisel jälgida seda, et kaared ei ristuks.
NB! Praegu on Annotatrixis bug, mille tõttu ta näitab punasena kaari, mis viivad punktuatsioonimärgi juurde, mis asub oma ülemusest eespoole. Programmi autorid lubasid selle ära parandada.
cc-without-conj
See veateade näitab, et koordineeriv sidend on, aga konjunkti ei järgne või cc (coordinating conjunct) ei allu sellele konjunktile. Tavalisim viga on see, et sidendi ülemus on vale. Sel juhul tasub vaadata ka selle sidendi ees olevat koma, et kas tema ülemus on õige.

Mitmeosalised pärisnimed
veateate märgendit ei ole ja kuna teisendusskriptiga ei õnnestunud pärisnimede kogu variatiivsust hallata, siis tuleb kõik mitmesõnalised pärisnimed üle kontrollida. Sisuliselt üle vaadata kõik PROPN sõnaliigi märgendiga read.
Mitmeosalise pärisnime puhul on kaks süntaktilise struktuuri väljendamise võimalust:
1. isikunimi ja võõrkeelsed või muud (eesti keele seisukohalt) sisemise süntaktilise struktuurita nimed: ülemus on esimene komponent, ülejäänud on selle esimese komponendi küljes suhtega flat
Siin elas Ferdinand (nsubj->elas) *Johann* (flat->Ferdinand) *Wiedemann* (flat->Ferdinand)
Sama: *International Jugglers Association*, *Baker Street* jm
2. Nimeüksus on eestikeelne fraas: märgendatakse nagu tavalist fraasi, säilib tema süntaktiline struktuur, nt *Tartu Ülikool*, *Suur Siiditee, Makedoonia Aleksander, Hansa Liit, Hiina Rahvavabariik* - esimene komponent on nmod (sest sõnaliik on PROPN).
Ka Vargamäe Andres – Vargamäe on nmod
NB! Kui muudad nime peasõna, siis tuleb ka täiendid jm ümber tõsta, flatil alluvaid ei tohi olla.

SEDA PRAEGU EI TEE
Asesõnad, mille UD sõnaliik on milles on ADJ\s*P, ehk siis puudepangas asesõna, meil ADJ. Valdavalt tuleb nad ümber liigitada sõnaliigiliselt PRON-ks ja muuta ka süntaktilist funktsiooni nmod-ks või (enamasti) det-ks.
Täiendina on alati det järgmised sõnad:
asesõna terve – kas ta on kasutusel asesõna (ja süntaktiliselt määratlejana) või omadussõnana.
terve ja kogu: kui nad on det funktsioonis, siis muuta sõnaliik PRON, kustutada morfoloogiline märgend Degree ja lisada PronType=Tot
sama – kas ta on asesõna (ja siis süntaktiliselt det) või määrsõna:
asesõna: *Mul oli alguses sama küsimus*
Määrsõna: *See on sama hea*

Üksikjuhte.
keegi teine - *keegi* on det
kõik see – kõik on det
see kõik – see on det
*kõik need asja*d – nii *kõik* kui *need* on det ja mõlema ülemus on *asjad*
missugune neist – missugune on põhi ja *neist* on *nmod*
Meis kõigis on nii – praegu on *meis* ruut ja *kõigis* det

Mõttega tuleb vaadata see ja nad omastavas, sest võib juhtuda, et see on ekslikult nmodi asemel detiks analüüsitud.

Siin lõpeb osa ’SEDA PRAEGU EI TEE’

Kaassõnad, mille süntaktiline märgend on muu kui case . Nad kas pole kaassõnad või on nende süntaktiline märgend vale.

Koopulad – kõige aeganõudvam probleem
UD versioonis 2.1 tuleb kõik *olema*-laused märgendada koopulalausetena, v.a.
1) need, kus on ainult *olema*-verb ja selle subjekt (pluss selle täiendid); võib olla ka veel nn modaaladverb (*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 paistavad olema: *tarvis, vaja, alles, üle olema*
3) kui öeldistäide on da-infinitiiv (mis siis on ccomp)
4) mas-vormis alluvaga *olema* , kus mõlemad osalised on sõnaliigiga VERB ja *olema* 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.
5) Lause koosneb küsisõnast süntaktilise märgendiga mark, *olema*-verbist ja alusest:
Kus on kirves?

Ümbermärgendamist vajavad laused leiab otsides neid *olema*-verbe, mille sõnaliik on VERB.
Kui muudate osalause ülemust, tuleb üle kontrollida kõik, mis enne sellele *olema*-verbile allus, eriti sidesõnad ja kirjavahemärgid, et need saaks tõstetud uue ülemuse külge. Selle tegemisel on sellest Annotatrixi graafilisest esitusest tõesti kasu. Tuleb ära muuta ka *olema* sõnaliik VERB -> AUX
kui olema-verbile allus nimisõnaline või asesõnaline laiend märgendiga obl ja uus ülemus on nimisõna või nimisõnaline asesõna, siis selle peab muutma märgendiks nmod

Milline moodustaja saab koopulalauses ruuduks? Hierarhia on selline
1. öeldistäitemäärus (*ta oli Valgas **õpetajaks**, Hiinlased on teistsuguse **psühholoogiaga***)
2. öeldistäitesarnane määrsõna (*Kõik on **halvasti**, Nad olid **kahekesi**; Tal on klapid **peas***)
ka: *tulemus oli **5%**. Tulemus oli üle viie **protsendi**. Aeg esimesel ringil oli **3:20**.*
3. omaja ja kogeja (***Tal** oli kodus kass; **Tal** oli kodus külm*)
4. Koht (*Ta oli õhtul **kodus**;* ka *Ta oli õhtul õnnetuna* kodus)
5. Aeg (*See oli möödunud **aastal***)
6. Viis (*See on **nii**, et ..., ka Sellega on **nii**, et...*)

Andrielale: Kõige halvemad on need nn välise omajaga laused (*Tal on klapid peas*). Ma pole kindel, kas otsus, et peas on ülemus, on õige.

Ja muidugi tekib siin palju kahtlasi kohti. Olen siia kokku kogunud mõned märgendamisotsused:
nagu oleks tal põrgukoerad kannul – *tal* on juurtipp (root)
mis neil viga on – root on *neil*
olin magamata – panin *magamata* ruuduks
Kus on Madis – jätsin *on* ruuduks
Kas võib olla veel kaunimat vaatepilti? - siin jätsin *olema* ruuduks.
... millel on sügavad juured – panin *milllel* ruuduks. Sama: mille ümber kammimisel oli olnud suurem elektriväli
Ainult siin on neid tunduvalt rohkem – panin *rohkem* ruuduks
ega siis kannatusi palju pole – panin *palju* ruuduks
ärkvel olles – *ärkvel* on ülemus
Mida seal varastada oli – jätsin *oli* ruuduks
Kindlasti poleks olnud Euroopa ja Hiina vastandamist. - jätsin *olnud* ruuduks.
No on ju sama trepp – jätsin *on* ruuduks
nii ei saa lihtsalt olla – panin *nii* ruuduks
miks see nii on – panin *nii* ruuduks
Miks oli nii kindel tunne – *oli* jäi ruuduks
või oli ta hoopis – jätsin *oli* ruuduks
kas seal pole midagi sees – panin *sees* ruuduks
Nagu polegi midagi teha – *polegi* jäi ruuduks
Mina olin Hiinas teist korda – panin *korda* ruuduks
eks need eesti rahad ole kõik nagu nad on - *ole* on ruut ja mõlemad *olema*d on verbid
teiselt poolt ega mingeid poliittunde ka polnud – jätsin *polnud* ruuduks
Meil oli siin Eestis palju aastaid nii – ruut on *nii*