Eesti keele puudepanga ud-teisenduse kirjeldus ja kvaliteedihinnang

Süntaktilised teisendused tehakse automaatselt spetsiaalsete kitsenduste grammatika reeglite abil (320 reeglit).

Teisenduste kvaliteeti hinnati 3000-sõnelise ilu_indrikson.tasak korpusefailil. Tulemused on toodud tabelis 1.

accuracy / Metric:LA	accuracy / Metric:LAS	accuracy / Metric:UAS	Token
0.991	0.985	0.992	Row mean
3032	3032	3032	Row count

Tabel 1. Teisenduse kvaliteedi hinnang. LA (label accuracy) näitab süntaktiliste märgendite teisendamise korrektsust, UAS (unlabeled accuracy score) näitab seoste teisendamise korrektsust ning LAS (labeled accuracy score) näitab, mitu protsenti tulemusest on nii õige märgendi kui seosega.

Arvestada tuleb:

- LA sisaldab ka punktuatsiooni teisendusi, mis on triviaalsed. Samas punktuatsiooni seoste teisendamine sõltub juurtipu määramisest ning see enam triviaalne ei ole.
- osalausete vahelised seosed on märgendatud märgendiga dep, mis ei täpsusta osalause funktsiooni. Märgend dep esines korpuses 51 korral. Selle täpsustamine toimub uue teisendusreeglite versiooni väljatöötamisel.
- UD märgenduse rahvusvahelises versioonis on veel palju ebaselgust, mistõttu ei ole alati üheselt arusaadav, kuidas peaks konkreetset märgendit (ehk mõistet) eesti keele jaoks määratlema, seetõttu ei ole UD märgendus alati piisavalt järjepidev. Näiteks UD ei püüa kohati eristada nähtusi, kus on nn hallid piirid, nt verb+verb ühendeid ei jagata öeldiseks ja verb+laiendiks, see ühelt poolt kaotab infot aga teiselt poolt teeb tulemuse ühtlasemaks. Funktsioonide ccomp, xcomp ja advcl täpsed piirid on hetkel veel ebaselged
  Täpsustumine ja defineerimine jätkub järgmiste versioonidega.

Kindlasti on antud versioon korrektsem UD puudepangas hetkel väljas olevast eestikeelsest versioonist nii korrektsuse kui ka mahu poolest.