Eesti uue meedia keele universaalsõltuvuste puudepank

Eesti uue meedia keele universaalsõltuvuste (Universal Dependencies, UD) puudepank (EWT) arenes välja uue meedia keele kitsenduste grammatika puudepangast, mis koosnes üksikutest failidest etTenTeni korpusest. Korpus märgendati käsitsi kitsenduste grammatika märgendusskeemiga ja seejärel konverteeriti poolautomaatselt universaalsete sõltuvuste formaati.
Korpuse esialgne maht oli u 27000 sõnet ning see koosnes veebist korjatud tekstidest, mis sisaldasid nii toimetatud uudiseid, toimetamata blogitekste kui ka foorumitekstide katkeid. Korpuse arendamise järgmises etapis lisati foorumitekste, mis käsitlevad nii hariduselu, söögi-joogi-, ulmekirjanduse- kui ka tehnika-alaseid arutelusid. See korpuse osa koosneb u 32000 sõnest.
2020. aastal lisati korpusesse Covid-19 alaseid kommentaariumi tekste kokku u 12000 sõnet. Neid tekste ka anonümiseeriti, kuna kohati sisaldasid tekstid reljeefseid väljendeid. Seejärel lisati korpusele u 8500 sõnet Redditi foorumi ja militaarfoorumi tekste ning lõpuks 9000-sõnelise osa foorumitekstide korpusest, mis on alamosa Tartu Ülikooli eesti ja üldkeeleteaduse instituudis arendatavast pragmaatikakorpusest. See korpuseosa sisaldab aianduse ja autode temaatikat käsitlevaid foorumitekste.
Kogu korpuse maht on 90600 sõnet.

kg-teisendus	27274
foorumid	32779
pandeemia kommentaarium	12725
reddit ja militaar	8492
pragmaatika	9331
Kokku	90601

Korpus on jagatud treening-, testimis- ja arendusosadeks, suhtega vastavalt 74,5% 14,5% ja 10%.

Korpus on märgendatud vastavalt universaalsõltuvuste standardile, lisaks on märgendatud ellipsid ja asesõnade viitesuhted laiendatud sõltuvuste tasandil ning täiendavalt nimeüksused. Märgenduse kohta saab täpsemalt lugeda EDT dokumentatsioonist.

Hetkel on puudepanga aktiivse versiooni number 2.11, see on kättesaadav aadressidelt https://github.com/EstSyntax/EstUD/tree/master/vers2_11 ja pärast versiooniuuendust ka https://universaldependencies.org/treebanks/et_ewt/index.html