

## Anafooride suhtes märgendatud korpus

30. mai 2019

Anafooride suhtes märgendatud korpus on praegu ca 200 000 sõna mahus tekste, milles on u 7800 märgendatud asesõna, millest u 6000 on ühendatud oma viitealusega, ülejäänud u 1800-l asesõnal viitealus tekstis puudub.

Tekstideks on ajalehe- ning ajakirjatekstitid ning üks teadustekst (ajakirja Eesti Arst 2004. aasta aastakäik).

Märgendatud on järgmised asesõnad kõigis käändevormides ja nende viitealused:

- isikulised asesõnad (*mina/ma, sina/sa, tema/ta, meie/me, teie/te, nemad/nad*). Kokku on korpus 3180 isikulist asesõna, neist 2576 on ühendatud viitealustega.
- näitav asesõna *see* esineb korpus 2844 korral, neist 2132 korral on tal tekstis olemas viitealus.
- siduvad asesõnad *kes* ja *mis* esinevad tekstis kokku 1981 korda, neist 1611 juhul on neil olemas viitealus tekstis.

Viitelauseks võib olla tekstis olev üksiksõna, nt järgmises lauses on asesõna *talle* viitelauseks pärisnimi Mari: *Mari tahtis, et vanaema talle külla tuleks.*

Viitelauseks võib olla ka terve osalause, sellisel juhul on viitealusena korpus 3180 märgendatud selle osalause öeldisverb. Järgneva näite teises lauses on asesõna *seda* viitab esimese lause teisele osalausele *et laseb koera kohe lahti* ja korpus 3180 on viitelauseks märgitud selle osalause öeldisverb *laseb*: *Koerajuht karjus kõrkjatesse, et laseb koera kohe lahti. Seda aga polnudki vaja, sest peagi oli kahtlusalune pilliroo vahelt väljas.*

Kui eelpoolloetletud asesõnale ei ole viitealust märgitud, tähendab see, et seda tekstis ei ole või pole see piisavalt selgelt väljendatud. Näiteks ei ole tekstis selget viitealust asesõnal *meie* järgmises lauses: „*See asi on väga oluline meie kõigi jaoks*”, rääkis Kuressaare linnapea.

Korpuse formaat on Eesti keele sõltuvuspuude panga (EDT) oma, kuhu on lisatud asesõnade ja nende viitealuste märgendid. EDT märgendite kohta vt

<https://github.com/EstSyntax/EDT/blob/master/syntmargendus.pdf>

Järgnevas näitelause on viitealustena märgitud sõnavormid *McDonald's* ja *klientide* ning asesõnadena *nad*, mis viitab lause esimesele sõnale *McDonald's*, veel kord *nad*, mis viitab jälle lause esimesele sõnale *McDonald's* ning asesõna *neile*, mis viitab lause kaheksandale sõnavormile *klientide*. Asesõnavormil *mida* ei ole tekstis viitealust.

"<s id="32">"

"<McDonald's>"

"McDonald's" Ls S com sg in @SUBJ #1->4 {Viitealus}

"<peab>"

"pida" Lb V mod indic pres ps3 sg ps af @FCV #2->4

"<hoolikamalt>"

"hoolikamalt" L0 D @ADVL #3->4

"<kontrollima>"

"kontrolli" Lma V main sup ps ill @IMV #4->18

"<,>"

"," Z Com #5->5

"<kuidas>"

"kuidas" L0 D @ADVL #6->10

"<nad>"

"tema" Ld P pers ps3 pl nom @SUBJ #7->10 {Pronoomen} {Coref:32.1}

"<klientide>"

"klient" Lde S com pl gen @NN> #8->9 {Viitealus}  
"<soove>"  
"soov" Le S com pl part @OBJ #9->10  
"<täidavad>"  
"täit" Lvad V main indic pres ps3 pl ps af @FMV #10->4  
"<ning>"  
"ning" L0 J crd @J #11->15  
"<mida>"  
"mis" Lda P inter rel sg part @OBJ #12->15 {Pronoomen}  
"<nad>"  
"tema" Ld P pers ps3 pl nom @SUBJ #13->15 {Pronoomen} {Coref:32.1}  
"<neile>"  
"tema" Lle P pers ps3 pl all @ADVL #14->15 {Pronoomen} {Coref:32.8}  
"<pakuvad>"  
"pakku" Lvad V main indic pres ps3 pl ps af @FMV #15->4  
"<, >"  
", " Z Com #16->16  
"<">"  
"" Z Quo #17->17  
"<ütles>"  
"ütle" Ls V main indic impf ps3 sg ps af @FMV #18->0  
"<Jacksoni>"  
"Jackson" L0 S prop sg gen @NN> #19->20  
"<advokaat>"  
"advokaat" L0 S com sg nom @NN> #20->22  
"<Timothy>"  
"Timothy" L0 S prop sg nom @NN> #21->22  
"<Houston>"  
"Houston" L0 S prop sg nom @SUBJ #22->18  
"<. >"  
"." Z Fst #23->23  
"</s>"