Corpus of Estonian texts, annotated for anaphora
Jan 9, 2018

This corpus containing ca 10 000 words of running text, is annotated for pronouns and their antecendents. There are 4200 annotated pronouns, among them 3200 are linked with their antecendents. The remaining 1000 pronouns have no clearly identifiable antecendent in text. Majority of the texts come from Estonian newspapers plus one scientific (medical) text, namely an issue of journal „Eesti Arst" (Estonian Doctor).

All case forms of the following pronouns are annotated:
- personal pronouns (*mina/ma* 'I', *sina/sa* 'sg you', *tema/ta* 's/he', *meie/me* 'we', *teie/te* 'pl you', *nemad/nad* 'they'). The corpus contains 1734 personal pronouns, 1320 of them are connected with their antecendents.
- demonstrative pronoun *see* 'it, that' occurs 1489 times in the corpus, 1084 occurences have an identifiable antecendent in text.
- relative pronouns *kes* 'who' and *mis* 'what' occur 1053 times in text, 851 occurences have an antecendent in text.

An antecendent is most often a noun, but it can also be a whole clause. In the latter case, the main verb of this clause is annotated as an antecendent.

The corpus format is that of the Estonian Dependency Treebank (EDT), with additional annotation of pronouns and their antecendents. Tags and labels used in EDT annotation are described here: https://github.com/EstSyntax/EDT/blob/master/syntmargendus.pdf

In the following example sentence, the word forms *McDonald's* and *klientide* are tagged as antecendents {Viitealus}. Pronoun {Pronoomen} *nad* is coreferential {Coref} with proper noun *McDonald's*, and another pronoun *nad* is coreferential with the same proper noun. Pronoun *neile* is coreferential with the 8. word in the sentence, *klientide.* Pronoun *mida* has no antecendent in text.

```
"<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>"