Anafooride suhtes märgendatud eesti keele sõltuvuspuude pank

English version below

Anafooride suhtes märgendatud korpusest on kaks versiooni: kitsenduste grammatika (Constraint Grammar, CG) märgendusega ja *Universal Dependencies*' (UD) märgendusega (UD versiooni 2.4 baasil). Mõlema korpuseversiooni tekstid ja märgendatud viitesuhted on samad.

Mõlema korpuseversiooni süntaktiline märgendus põhineb sõltuvussüntaksil, mille järgi süntaktilised seosed on üksiksõnade, mitte fraaside vahel. Nii on ka viitesuhe kahe tekstisõna vaheline suhe, ka siis, kui viitealuseks on terve osalause: näiteks lauses *Kõik see, mis koerast peegeldub, on koerajuhi töö tulemus.* viitab näitav asesõna *see* kogu täiendkõrvallausele *mis koerast peegeldub*, aga korpuses on see märgendatud nii, et *see* viitealuseks on osalause juurtipp *peegeldub*.

Anafooride suhtes märgendatud korpuses on u 253 000 sõna u 17 500 lauses. Märgendatud on u 8350 asesõna, millest u 7100 on ühendatud oma viitealusega, ülejäänud asesõnadel viitealus tekstis puudub. 482-l asesõnal on mitu viitealust. Tekstideks on ajalehetekstid ning üks teadustekst (ajakirja Eesti Arst 2004. aasta aastakäik). Märgendatud on järgmised asesõnad kõigis käändevormides ja, kui on tekstis olemas, nende viitealused:

- isikulised asesõnad (*mina/ma, sina/sa, tema/ta, meie/me, teie/te, nemad/nad*),
- näitav asesõna *see*,
- siduvad asesõnad *kes* ja *mis*.

Programmid, mis teisendavad puudepanga formaadis faili brati (https://brat.nlplab.org/index.html) märgendustööriistale sobivaks ja tagasi (pronoomentykeldaja.pl ja brat2inforem) on kataloogis tools. Programmide autorid on Kaili Müürisep ja Katrin Tsepelina.

Korpuse näide: UD-kuju, anafooride märgendus on kollasel taustal.

```
# sent_id = aja_ee199920_11
# text = "Tegutsesime nende võimaluste piirides, mis meile on antud," ütleb ta.
1	"	"	PUNCT	Z	_	2	punct	_	SpaceAfter=No
2	Tegutsesime	tegutsema	VERB	V	Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	_
3	nende	see	DET	P	Case=Gen|Number=Plur|PronType=Dem	4	det	_	Coref=11.10|CorefType=Cl
4	võimaluste	võimalus	NOUN	S	Case=Gen|Number=Plur	5	nmod	_	Antecedent=11.4
5	piirides	piir	NOUN	S	Case=Ine|Number=Plur	2	obl	_	SpaceAfter=No
6	,	,	PUNCT	Z	_	10	punct	_	_
7	mis	mis	PRON	P	Case=Nom|Number=Plur|PronType=Int,Rel	10	obj	_	Coref=11.4|CorefType=N
8	meile	mina	PRON	P	Case=All|Number=Plur|Person=1|PronType=Prs	10	obl	_	Coref=8.2|CorefType=N
```

```
9	on	olema	AUX	V
Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	10	aux	_
	_
10	antud	andma	VERB	V	Tense=Past|VerbForm=Part|Voice=Pass	5
acl:relcl	_	Antecedent=11.10|SpaceAfter=No
11	,	,	PUNCT	Z	_	2	punct	_	SpaceAfter=No
12	"	"	PUNCT	Z	_	2	punct	_	_
13	ütleb	ütlema	VERB	V
Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	2	parataxis
	_	_
14	ta	tema	PRON	P	Case=Nom|Number=Sing|Person=3|PronType=Prs	13
	nsubj	_	Coref=9.6|CorefType=N|SpaceAfter=No
15	.	.	PUNCT	Z	_	2	punct	_	_
```

Korpuse näide: CG-kuju

```
"<s id="11">"
"<">"
	""" Z Quo #1->1
"<Tegutsesime>"
	"tegutse" Lsime V main indic impf ps1 pl ps af @FMV #2->13
"<nende>"
	"see" Lde P dem pl gen @NN> #3->4 {Pronoomen} {Coref:11.10}
"<võimaluste>"
	"võimalus" Lte S com pl gen @NN> #4->5 {Viitealus}
"<piirides>"
	"piir" Ldes S com pl in @ADVL #5->2
"<,>"
	"," Z Com #6->6
"<mis>"
	"mis" L0 P inter rel pl nom @OBJ #7->10 {Pronoomen} {Coref:11.4}
"<meile>"
	"mina" Lle P pers ps1 pl all @ADVL #8->10 {Pronoomen} {Coref:8.2}
"<on>"
	"ole" L0 V aux indic pres ps3 pl ps af @FCV #9->10
"<antud>"
	"and" Ltud V main partic past imps @IMV #10->5 {Viitealus}
"<,>"
	"," Z Com #11->11
"<">"
	""" Z Quo #12->12
"<ütleb>"
	"ütle" Lb V main indic pres ps3 sg ps af @FMV #13->0
"<ta>"
	"tema" L0 P pers ps3 sg nom @SUBJ #14->13 {Pronoomen} {Coref:9.7}
"<.>"
	"." Z Fst #15->15
"</s>"
```

Estonian Treebank annotated for pronominal anafora

This corpus comes in two versions: with Constraint Grammar and Universal Dependencies morphological and syntactic annotations. Both versions contain the same texts and coreference annotations.

Corpus contains ca 253 000 tokens in 17 500 sentences. Ca 8350 pronouns have been annotated, among them ca 7100 are linked with their antecedents, the remaining 1250 pronouns have no clearly identifiable antecedent in text. 482 pronouns have more than one antecedent.

Majority of the texts come from Estonian newspapers, plus one scientific (medical) text, namely an issue of the journal “Eesti Arst” (*Estonian Doctor*).

All case forms of the following pronouns have been annotated:

- ☐ personal pronouns *mina/ma* (I, 1. person singular), *sina/sa* (you, 2. person singular), *tema/ta* (he/she, 3. person singular), *meie/me* (we, 1. person plural), *teie/te* (you, 2. person plural), *nemad/nad* (they, 3. person plural)
- ☐ demonstrative pronoun *see* ‘it’
- ☐ relative pronouns *kes* ‘who’ ja *mis* ‘what’.

Programs to convert Estonian dependency trees (VISLCG format) to brat (https://brat.nlplab.org/index.html) annotations and back (pronoomentykeldaja.pl and brat2inforem) are in the tools folder, authors Kaili Müürisep and Katrin Tsepelina.

Example: UD Format
Annotation is in the misc-field. Antecedents have been annotated with keyword Antecedent=Sent_No.Word_No, and words referring to them have the keyword Coref and the address of the antecedent. CorefType indicates whether the antecedent is a noun or a clause.

```
# sent_id = aja_ee199920_11
# text = "Tegutsesime nende võimaluste piirides, mis meile on antud," ütleb ta.
1	"	"	PUNCT	Z	_	2	punct	_	SpaceAfter=No
2	Tegutsesime	tegutsema	VERB	V	Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	_
3	nende	see	DET	P	Case=Gen|Number=Plur|PronType=Dem	4	det	_	Coref=11.10|CorefType=Cl
4	võimaluste	võimalus	NOUN	S	Case=Gen|Number=Plur	5	nmod	_	Antecedent=11.4
5	piirides	piir	NOUN	S	Case=Ine|Number=Plur	2	obl	_	SpaceAfter=No
6	,	,	PUNCT	Z	_	10	punct	_	_
7	mis	mis	PRON	P	Case=Nom|Number=Plur|PronType=Int,Rel	10	obj	_	Coref=11.4|CorefType=N
8	meile	mina	PRON	P	Case=All|Number=Plur|Person=1|PronType=Prs	10	obl	_	Coref=8.2|CorefType=N
```

```
9	on	olema	AUX	V
Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	10	aux	_
	_
10	antud	andma	VERB	V	Tense=Past|VerbForm=Part|Voice=Pass	5
acl:relcl	_	Antecedent=11.10|SpaceAfter=No
11	,	,	PUNCT	Z	_	2	punct	_	SpaceAfter=No
12	"	"	PUNCT	Z	_	2	punct	_	_
13	ütleb	ütlema	VERB	V
Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	2	parataxis
	_	_
14	ta	tema	PRON	P	Case=Nom|Number=Sing|Person=3|PronType=Prs	13
	nsubj	_	Coref=9.6|CorefType=N|SpaceAfter=No
15	.	.	PUNCT	Z	_	2	punct	_	_
```

Example: CG Format

```
"<s id="11">"
"<">"
	""" Z Quo #1->1
"<Tegutsesime>"
	"tegutse" Lsime V main indic impf ps1 pl ps af @FMV #2->13
"<nende>"
	"see" Lde P dem pl gen @NN> #3->4 {Pronoomen} {Coref:11.10}
"<võimaluste>"
	"võimalus" Lte S com pl gen @NN> #4->5 {Viitealus}
"<piirides>"
	"piir" Ldes S com pl in @ADVL #5->2
"<,>"
	"," Z Com #6->6
"<mis>"
	"mis" L0 P inter rel pl nom @OBJ #7->10 {Pronoomen} {Coref:11.4}
"<meile>"
	"mina" Lle P pers ps1 pl all @ADVL #8->10 {Pronoomen} {Coref:8.2}
"<on>"
	"ole" L0 V aux indic pres ps3 pl ps af @FCV #9->10
"<antud>"
	"and" Ltud V main partic past imps @IMV #10->5 {Viitealus}
"<,>"
	"," Z Com #11->11
"<">"
	""" Z Quo #12->12
"<ütleb>"
	"ütle" Lb V main indic pres ps3 sg ps af @FMV #13->0
"<ta>"
	"tema" L0 P pers ps3 sg nom @SUBJ #14->13 {Pronoomen} {Coref:9.7}
"<.>"
	"." Z Fst #15->15
"</s>"
```