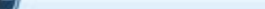


Eesti keele puudepanga semantilisest märgendusest

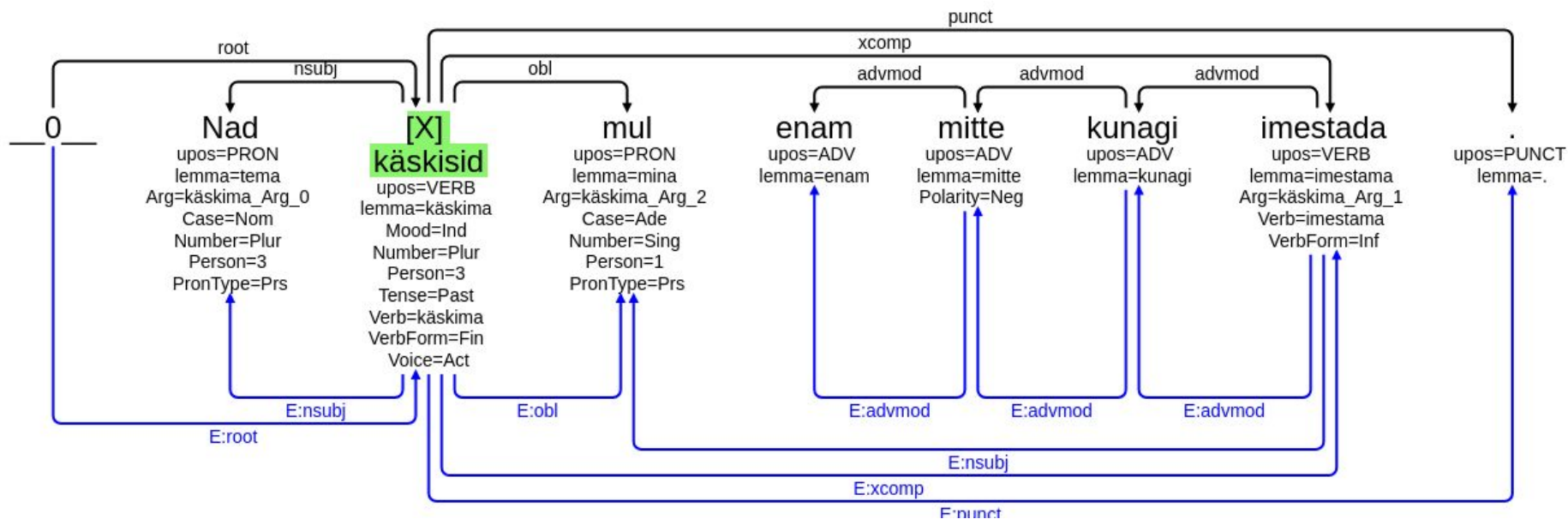
Kadri Muischnek, Kaili Müürisep, Andriela Rääbis
Tartu Ülikool

Taustast

- Meil on küllaltki suur süntaktiliselt märgendatud universaalsete sõltuvuste puudepank nii standardse kirjaliku kui uue meedia keele jaoks.
- Selles on märgendatud nii baassüntaks kui ka laiendatud sõltuvused kui ka lausesisesed anafoorid.
- Samas puudus süstemaatiliselt semantiliselt märgendatud korpus
- <https://universaldependencies.org/> (de Marneffe jt 2021)

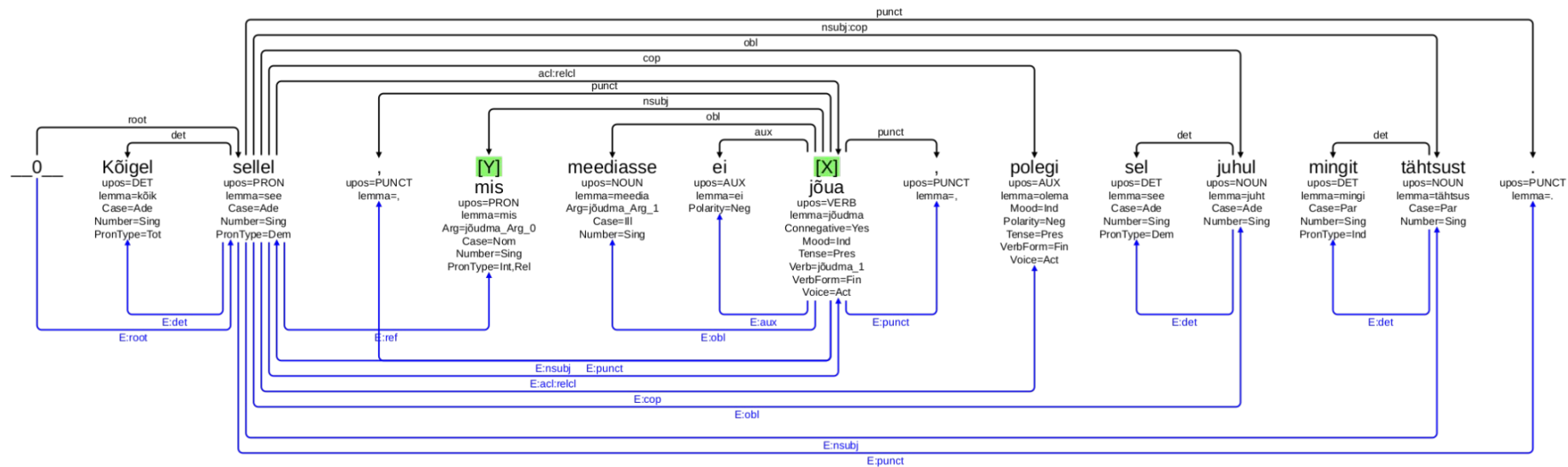
▶		English	11	760K		IE, Germanic
▶		Erzya	1	20K		Uralic, Mordvin
▶		Estonian	2	529K		Uralic, Finnic
▶		Faroese	2	50K		IE, Germanic
▶		Finnish	4	397K		Uralic, Finnic
▶		French	7	635K		IE, Romance

UD puudepanga näide



Nad käskisid mul enam mitte kunagi imestada.

vt https://universal.grew.fr/?corpus=UD_Estonian-EDT@2.15



Kõigel sellel, mis meediasse ei jõua, polegi sel juhul mingit tähtsust.

Eesmärgid

EKTB75 "Semantilise analüüsi baasvahendid" (01.01.2023–31.12.2023)

- Leida sobiv viis semantiliste rollide lisamiseks süntaktilisele märgendusele
- Luua märgenduskeem
- ja märgendada standardkirjakeele puudepank
- Kaardistada probleemid

Semantilised rollid

Semantilised rollid kirjeldavad predikaadiga väljendatud sündmuse osalisi ja nende osalemise viisi.

Predikaat määrab, millised osalised sündmusel saavad olla.

Semantiliste rollide täpset nimekirja pole, kuid enamkasutatavamaid rolle (agent, patsient, saaja jne) on võimalik loetleda, vt nt Lindström ja Taremaa (2013:40-42).

Semantiliste rollide hulk sõltub (ka) sellest, kui detailselt soovitakse rolle eristada.

Semantilised rollid: proto-rollid

Lahendamaks küsimust, kui detailselt tuleks semantilisi rolle eristada, on välja pakutud mitmetasandilisi hierarhilisi süsteeme.

nt Dowty (1991) tõi sisse proto-rollid: proto-agendi ja proto-patsiendi

lauses *Laps sööb saia* on laps agent, st teadlik toimija

lauses *Lapsele meeldivad koerad* on laps kogeja

Samas on *laps* mõlemas lauses aktiivseim osaline, proto-agent, hoolimata sellest, et esimeses lauses on ta süntaktiliselt subjektiks ja teises määruseks.

Täpsemalt semantilistest rollidest (keeleeuurimisel) vt nt Lindström ja Taremaa (2013)

Semantilised rollid arvutilingvistikas/keeletehnoloogias

Semantilistel rollidel on oluline koht üleminekul lause süntaktiliselt esituselt semantilisele esitusele.

Semantilised rollid aitavad vastata küsimusele "Kes tegi kellele mida; milleks, kus, millal ja kuidas?"

st nad aitavad lahendada infoeraldamise ülesannet ja sündmuse tuvastamise ning formaalse esitamise ülesannet.

Millises vormis semantilisi rolle esitada?

Automaatanalüüs vajab kindlat formaati ehk reeglistikku - märgendite loetelu ja nende rakendamise eeskirja.

Rahvusvaheliselt on olemas kolm põhilist semantiliste rollide/argumentstruktuuri esitamise standardit: **Framenet** (Fillmore jt 2003), **Verbnet** (Kipper-Schuler 2006) ja **PropBank** (Palmer jt 2005)

Valisime neist **PropBanki**, sest see tundus olevat kõige lihtsamini süntaksiga ühendatav.

Lisamotivatsiooni pakkus asjaolu, et on olemas rahvusvaheline initsiatiiv *Universal Proposition Bank* (Jindal jt 2022).

PropBank ehk Proposition Bank

Semantilistel rollidel ei ole nimesid, vaid on numbrid.

Seega pole vaja rollide nimekirja ja välditakse rollide kuhjumise probleemi.

Igal verbil või polüseemse verbi igal tähendusel on oma rollide nummerdus.

Siiski märgendatakse proto-agent e tegevussubjekt kui Arg0 ja proto-patsient e tegevusobjekt, samuti automatiivverbi teema kui Arg1.

Teistele Arg antakse numbrid verbipõhiselt, püüeldes verbiülestest üldistuste poole: saaja on Arg2; eesti keele liikumis- ja liigutamisverbide korral on lähtekoht Arg2, sihtkoht Arg3 jne. Eesti keele PropBankis kasutame 6 argumendimärgendit..

Soome PropBank

sentence-text: Olin kyllä nähnyt muualla, mitä metsälle saattoi tapahtua yhdessä päivässä.

1	Olin	olla	AUX	V	/.../	3	aux	3:PBArgM_mod	_	
2	kyllä	kyllä	ADV	Adv	_	3	advmod	3:PBArgM_dis	_	
3	nähnyt	nähdä	VERB	V	/.../	0	root	_	PBSENSE=nähdä.1	
4	muualla	muualla	ADV	Adv	_	3	advmod	3:PBArgM_loc	SpaceAfter=No	
5	,	,	PUNCT	Punct_		9	punct_	_		
6	mitä	mikä	PRON	Pron	Case=Par Number=Sing PronType=Int	9	nsubj	9:PBArg_1	_	
7	metsälle	metsä	NOUN	N	Case=All Number=Sing	9	nmod	9:PBArg_2	_	
8	saattoi	saattaa	AUX	V	/.../	9	aux	9:PBArgM_mod	_	
9	tapahtua	tapahtua	VERB	V	/.../	3	ccomp	3:PBArg_1	PBSENSE=tapahtua.1	
10	yhdessä	yksi	NUM	Num	Case=Ine Number=Sing NumType=Card	11	nummod	_	_	
11	päivässä	päivä	NOUN	N	Case=Ine Number=Sing	9	nmod	9:PBArgM_tmp	_	
12	.	.	PUNCT	Punct_		3	punct_	_		

Soome PropBank

vt https://turkunlp.org/Finnish_PropBank/ (Haverinen jt 2015)

tapahtua.1

Happen, occur

(tags: model:happen.01)

Tapahtua, sattua

A1	Thing happening		Asia, joka tapahtuu
A2	Beneficiary	(tags: rare)	Taho, jolle arg1 tapahtuu

Argumendistruktuuri märgendus eesti keele puudepangas

- Milline moodustaja saab semantilise rolli märgendi ehk kuidas me eristame seotud ja vabu laiendeid?
 - Lähtume seisukohast, et seotud ja vabad laiendid moodustavad pigem skaala (nt Croft 2001: 273-274), mitte kaks selgepiirilist hulka.
 - Kindla verbi laiendite jagamisel seotud (=saavad Arg märgendi) ja vabadeks (=ei saa Arg märgendit) lähtusime laiendi sagedusest märgendatavas korpuses.
 - Näiteks on verbil *silmitsema* märgendatavas korpuses sage viisi väljendav laiend ja kuigi viisimäärus on “klassikaline” vaba laiend, on ta märgendatud kui Arg2:

Kääbus (Arg0) silmitses meid (Arg1) pikalt (Arg2).

- Eristasime vaid väga ilmseid ning argumendistruktuuri puudutavaid verbitähendusi.
- Pole seotud inglise keele PropBanki kirjetega.

Näide

sent_id = ilu_kivirahk_593

text = Ott ajas end jalule ja läks.

1	Ott	Ott	PROPN	/.../	2	nsubj	2:nsubj 6:nsubj	Arg=ajama_Arg_0
2	ajas	ajama	VERB	/.../	0	root	0:root	Verb=ajama_2
3	end	ise	PRON	/.../	2	obj	2:obj	Arg=ajama_Arg_1
4	jalulejal		NOUN	/.../	2	obl	2:obl	Arg=ajama_Arg_3
5	ja	ja	CCONJ	_	6	cc	6:cc	
6	läks	minema	VERB	/.../	2	conj	0:root 2:conj	Verb=minema_1
7	.	.	PUNCT	_	2	punct	2:punct	

246	ajama	ajama_2	Nominal#nsubj#	Nominal#obj#Pa -	Nominal#obl#All -	-	!	-	ajas	-	ilu_ki	Ott ajas end jalule ja läks.	
247	ajama	ajama_2	Nominal#nsubj#	Nominal#obj#Pa -	Nominal#obl#All -	-	!	-	ajas	-	ilu_ki	Siis ajas ta end jalule ning sõnas:	
248	ajama	ajama_2	Nominal#nsubj#	Nominal#obj#Pa -	Nominal#obl#Ill -	-	!	-	ajab	-	aja_e	Flegmaatiline inimene ajab vardasse	
249	ajama	ajama_3	-	Nominal#obj#Ge	Nominal#obl#Ad -	-	-	!	-	ajada	-	aja_p	Siseminister Tarmo Loodus on selles
250	ajama	ajama_3	-	Nominal#obj#Nc -	-	-	-	!	-	aetud	-	ilu_in	Kella kaheks peab asi aetud olema!

Töö käik

1. Päringute põhjal esialgse kava koostamine, mitu tähendust, millised on argumendid.
 - a. Ainult sagedasemad lihtverbid, ühend- ja väljendverbid tulevikus.
 - b. Sagedane ehk esines EDT puudepangas rohkem kui 14 korda ja siis veel natuke, kokku 570 erinevat verbi.
 - c. Nende 848 erinevat tähendust.
2. Korpusest vastava verbiga lausete allalaadimine, märgendamine, üleslaadimine (u 37000 verbi ja nende 58000 argumenti)
3. Vigade kontroll, dokumentatsioon, valideerimine
4. Järeldused

Erinevatel etappidel 4 struktuurikirjelduse loojat, 2 märgendajat, 1 skriptikirjutaja.

Näide

```
# sent_id = ilu_kanep_901
```

```
# text = Nad käskisid mul enam mitte kunagi imestada.
```

1	Nad	tema	PRON	/.../	2	nsubj	2:nsubj	Arg=käskima_Arg_0
2	käskisid	käskima	VERB	/.../	0	root	0:root	Verb=käskima
3	mul	mina	PRON	/.../	2	obl	2:obl 7:nsubj	Arg=käskima_Arg_2
4	enam	enam	ADV	_	5	advmod	5:advmod	_
5	mitte	mitte	ADV	/.../	6	advmod	6:advmod	_
6	kunagi	kunagi	ADV	_	7	advmod	7:advmod	_
7	imestada	imestama	VERB	/.../	2	xcomp	2:xcomp	Arg=käskima_Arg_1
Verb=imestama								

Dokumentatsioon

<https://github.com/EstSyntax/EstSemRol>

ajama

ajama_1

kedagi/midagi seisundisse ajama

Arg0: nsubj#SPN#Nom (agent, põhjus)

Arg1: obj#SPN#Nom | obj#SPN#Gen | obj#SPN#Par (objekt)

Arg2: xcomp#Verb#Sup#Ill | xcomp#SPN#Tra | obl#SPN#Add | advmod#Adv#(püsti | nurja | segamini | pärani) (seisund)

Arg3:

Arg4:

Arg5:

Näited

- Punavõim aeti tigidaks kui herilane ja ühele korraldajale Karuksile lõppes asi eetrikeeluga.
- Aga ma ajasin nende ülla kava nurja -- laulsin ja tegin sihukest möllu, et ...
- Siristavad ja kudrutavad preilid, kelle rinnad ja istmikud korsetiliste abinõude varal kummi olid aetud /.../
- Aga soe suitsuräim ei aja praegu kellelgi suud vett jooksmas.

Mis edasi?

- Mudel
- Järjekindlus
- Tähtsaim küsimus - kas selle järgi on tänapäeval vajadust?
- Kui palju LLM suudaks aidata?

Croft, W. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective* (1st ed.). Oxford:Oxford University Press.

Dowty, D. 1991. Thematic proto-roles and argument selection. - *Language* 67, 547–619.

Fillmore, C. J., C. R. Johnson, ja M. R. L. Petruck. 2003. Background to FrameNet. *International journal of lexicography*, 16(3):235–250.

Haverinen, K.; Kanerva, J.; Kohonen, S.; Missilä, A.; Ojala, S.; Viljanen, T.; Laippala, V. & Ginter, F. The Finnish Proposition Bank. 2015. Language Resources and Evaluation, Volume 49, Issue 4, pp 907-926

Jindal, I., A. Rademaker,, M. Ulewicz, H. Linh, H. Nguyen, K. Tran, H. Zhu ja Y. Li. 2022. Universal Proposition Bank 2.0. - *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1700–1711.

Kipper-Schuler, K. 2006. *Verbnet: a broad-coverage, comprehensive verb lexicon*. PhD Thesis, University of Pennsylvania.

Lindström, L., P. Taremaa. 2013. Semantilised rollid keeleuurimise vahendina. - *ESUKA – JEFUL* 2013, 4 – 1: 35 – 52

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman (2021). Universal Dependencies. *Computational Linguistics* 47(2): 255–308.

Palmer, M., D. Gildea, P. Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. - *Computational Linguistics Journal*, 31:1.