Eesti keele universaalsõltuvuste (Universal Dependencies, UD) puudepankade EDT ja EWT nimeüksuste (*named entities, NE*) suhtes märgendatud versioonid

Formaat

Märgendus on UD CONLLU formaadis, mis näeb välja selline:

```
# sent_id = aja_ml200247_1651/et
# text = Festivali põhiprogramm avatakse reedel, 29. novembril kl 19 Sakala Keskuses Tom Tykweri filmiga "Taevas" (Saksa-USA).
1	Festivali	festival	NOUN	S	Case=Gen|Number=Sing	2	nmod	2:nmod:gen	_
2	põhiprogramm	põhi_programm	NOUN	S	Case=Nom|Number=Sing	3	obj	3:obj	_
3	avatakse	avama	VERB	V	Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Pass	0	root	0:root	_
4	reedel	reede	NOUN	S	Case=Ade|Number=Sing	3	obl	3:obl:ade	SpaceAfter=No
5	,	,	PUNCT	Z	_	7	punct	7:punct	_
6	29.	29.	ADJ	N	Case=Ade|Number=Sing|NumForm=Digit|NumType=Ord	7	amod	7:amod	_
7	novembril	november	NOUN	S	Case=Ade|Number=Sing	3	obl	3:obl:ade	_
8	kl	kl	NOUN	Y	Abbr=Yes	3	obl	3:obl	_
9	19	19	NUM	N	Case=Gen|Number=Sing|NumForm=Digit|NumType=Card	8	nummod	8:nummod	ToDo=nummod_err
10	Sakala	Sakala	PROPN	S	Case=Gen|Number=Sing	11	nmod	11:nmod:gen	NE=B-Loc
11	Keskuses	Keskus	PROPN	S	Case=Ine|Number=Sing	3	obl	3:obl:ine	NE=I-Loc
12	Tom	Tom	PROPN	S	Case=Nom|Number=Sing	14	nmod	14:nmod:nom	NE=B-Per
13	Tykweri	Tykwer	PROPN	S	Case=Gen|Number=Sing	12	flat	12:flat	NE=I-Per
14	filmiga	film	NOUN	S	Case=Com|Number=Sing	3	obl	3:obl:com	_
15	"	"	PUNCT	Z	_	16	punct	16:punct	SpaceAfter=No
16	Taevas	taevas	NOUN	S	Case=Nom|Number=Sing	14	appos	14:appos	NE=B-Prod|SpaceAfter=No
17	"	"	PUNCT	Z	_	16	punct	16:punct	_
18	(	(	PUNCT	Z	_	19	punct	19:punct	SpaceAfter=No
19	Saksa-USA	Saksa-USA	PROPN	Y	Abbr=Yes	14	parataxis	14:parataxis	NE=B-Gep|SpaceAfter=No
20	)	)	PUNCT	Z	_	19	punct	19:punct	SpaceAfter=No
21	.	.	PUNCT	Z	_	3	punct	3:punct	_
```

Nimeüksused (NE) on näites kollase taustaga esile tõstetud.
NE=B tähistab nimeüksuse algussõna, NE=I algusmärgendile B järgnevaid samasse nimeüksusesse kuuluvaid sõnu.
Loc, Per jne märgendis tähendavad nimeüksuse liike.

Eristatavad nimeüksuste liigid:

Per isik, elusolendite nimed. Inimeste, aga ka nt kasside jm nimed.
Loc koht, nt *Emajõgi, planeet Maa, (keegi viidi) Rakvere haiglasse, (avarii toimus) Võsu - Vergi teel*
vt ka Loc vs Gep näiteid dokumendi lõpuosast

Gep geopoliitiline üksus, koht, mis käitub organisatsioonina, õigemini küll mida esitatakse keeleliselt toimijana, elusana, personifitseerituna: *Moskva saatis kirja, Prantsusmaa otsustas nii, Ja üha irratsionaalsemaks muutuv Venemaa. Iirimaa võitis Eesti 2:0*

Org organisatsioon. *BBC andmetel ..., Riigikogu võttis vastu otsuse…*

PRod artefakt st “tehtud asi”, toode, ka teos, nt *“Püha õhtusöömaaeg”*, *“Tõde ja Õigus”*, *ajakiri Horisont, Berliini müür*

Event sündmus, nt *teatrifestival Kuldne Mask* , *Külm sõda, Rakenduslingvistika Ühingu aastakonverents, Paide arvamusfestival, näitus "100 maailma kirjeldavat objekti"*

Muu nimeüksus ei kuulu ühtegi eelnevasse kategooriasse

Unk ei saa liigitada ebapiisava info tõttu

Nime ulatus

Liigisõna (*Tartu **linn,** Ülikooli **tänav,***) on nime osa. Mõned näited:
ajakiri Horisont - *ajakiri* on liigisõna ja on nime osa
Tangoko looduskaitseala - *looduskaitseala* on liigisõna ja nime osa, liik on Loc
Sulawesi tuttmakaak - *tuttmakaak* pole nime osa, *Sulawesi* on Loc
Tartu bussiliinid - *bussiliinid* pole nime osa, *Tartu* on Gep
Minolta kaamera - kaamera on nime osa

Tiitel ei ole nime osa: *härra Greenaway*: *härra* pole nime osa

Pärisnimega algavad **liitsõnad** *Pariisi-reis, Nokia-vaimustus* ei ole NE.

Üksteise sees olevaid nimesid ei erista, st *Tartu ülikooli jalgpallimeeskond FC Fauna* on üks NE, mille liik on Org

Kirjavahemärgid
Sellistel juhtudel nagu näiteks *"Inimesed, ma kardan teid" - niisugune veidi ehmatava pealkirjaga raamat,...* on nimeüksuse osaks ka jutumärgid ja koma.

EDT-s ja EWT-s kasutati erinevat lähenemist otsustamaks, kas sõna või sõnaühend on nimeüksus või mitte. EDT-s lähtuti ortograafiast, st oletati, et nimeüksuse koosseisus on pärisnimi, mida kirjutatakse suure algustähega. Veebitekstides me sellest oletusest ei lähtu, st nimeüksus EWT-s ei pruugi sisaldada ühtegi suure algustähega sõna.
nt *kui koidulas koolis käisin 10 aastat tagasi,...*
… aga dodo pizzas on ananassi, pohlade ja kondenspiimaga pitsa.

Võrreldes nn Kairit Sirtsu NE-märgendatud korpusega (https://github.com/TartuNLP/EstNER_new) ei ole meie korpuses märgendatud kuupäevi, ajaväljendeid, tiitleid, rahaühikuid ja protsente, sest need ei ole rangelt võttes nimeüksused. Nimetatud korpuses on kasutatud hierarhilist märgendamist, meie hetkeversioonis mitte.