Creation of a balanced multilayer text corpus: trees, frames, entities

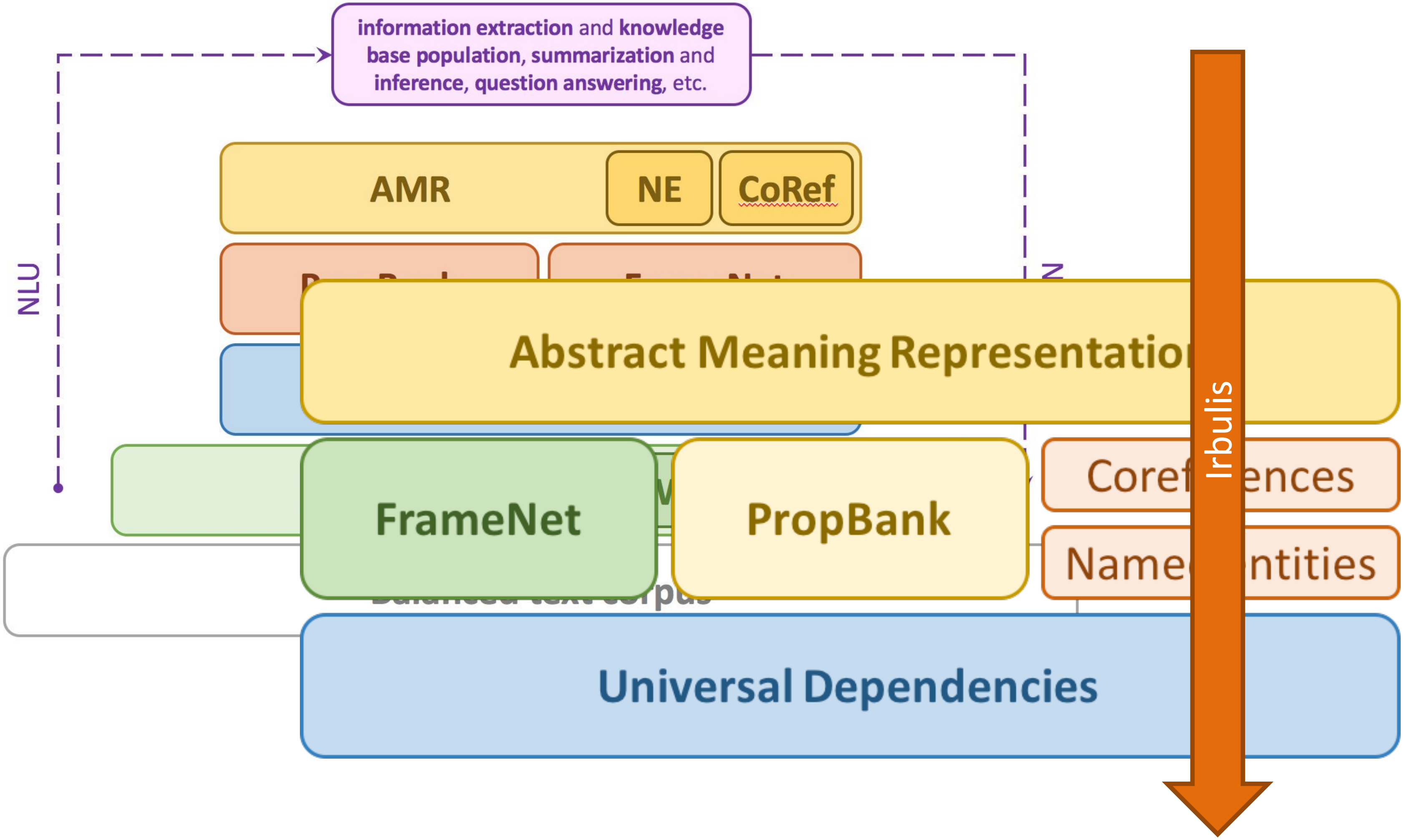
Normunds Grūzītis, Pēteris Paikens, Lauma Pretkalniņa
Ilze Auziņa, Guntis Bārzdiņš, Roberts Darģis, Mikus Grasmanis, Kristīne Levāne-Petrova,
Gunta Nešpore-Bērzkalne, Laura Rituma, Inguna Skadiņa, Baiba Valkovska, Artūrs Znotiņš

University of Latvia
Institute of Mathematics and Computer Science
Artificial Intelligence Laboratory

LU 76. zinātniskās konferences Datorlingvistikas sekcija
2018. gada 1. martā

NATIONAL DEVELOPMENT PLAN 2020
EUROPEAN UNION
European Regional Development Fund
INVESTING IN YOUR FUTURE
Grant agreement No. 1.1.1.1/16/A/219

Full Stack of Latvian Language Resources
for Natural Language Understanding and Generation


information extraction and knowledge base population, summarization and inference, question answering, etc.
NLU
N
AMR
NE
CoRef
Abstract Meaning Representation
FrameNet
PropBank
Coref ences
Name ntities
Balanced text corpus
Universal Dependencies
Irbulis

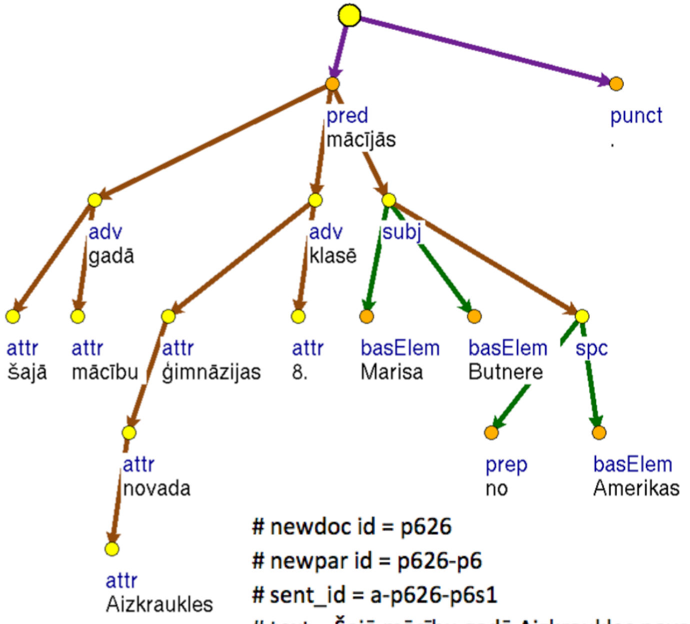
Approach

- **UD** is derived from **hybrid** dependency-constituency trees
  - Paragraphs, sentences and tokens are uniformly split across all layers (CoNLL-U)
  - Merging is based on the document, paragraph, sentence and token identifiers
- **FrameNet** annotations are added on top the underlying **UD**
- The **PropBank** layer is derived from the **FrameNet** and **UD** layers
- Semi-automatic annotation of **named entities** (NE), as well as NE linking
  - **Coreference** annotations are added afterwards, on top of **NE**, consulting **UD** if necessary
- Draft **AMR** graphs are *to be* derived from the **UD**, **PropBank**, **NE** and **coreference** layers, with the potential to integrate FrameNet into AMR
  - The semantically richer **FrameNet** helps to acquire more accurate AMR graphs

Balanced Data Set

- Aiming at a medium-sized corpus – around **10,000** sentences
  - **Balanced** in terms of **genres** and writing styles, and **lexical units** (LU)
- Fundamental design decision: a text unit is an isolated **paragraph**
- Representative paragraphs are **manually selected** in different proportions from a balanced 10-million-word text corpus
  - 60% news sources, 20% fiction, 10% legal texts, 5% spoken language, 5% misc.
- Our goal is to cover **1,000** most frequently occurring verbs
  - The number of LUs (verb senses w.r.t. FrameNet and PropBank frames) will be larger
  - Around 10 example sentences per LU (on average)
  - Paragraphs are selected based on verbs they contain, not randomly
  - Curators are constantly updated on the current balance / imbalance of the corpus

UD – Universal Dependencies

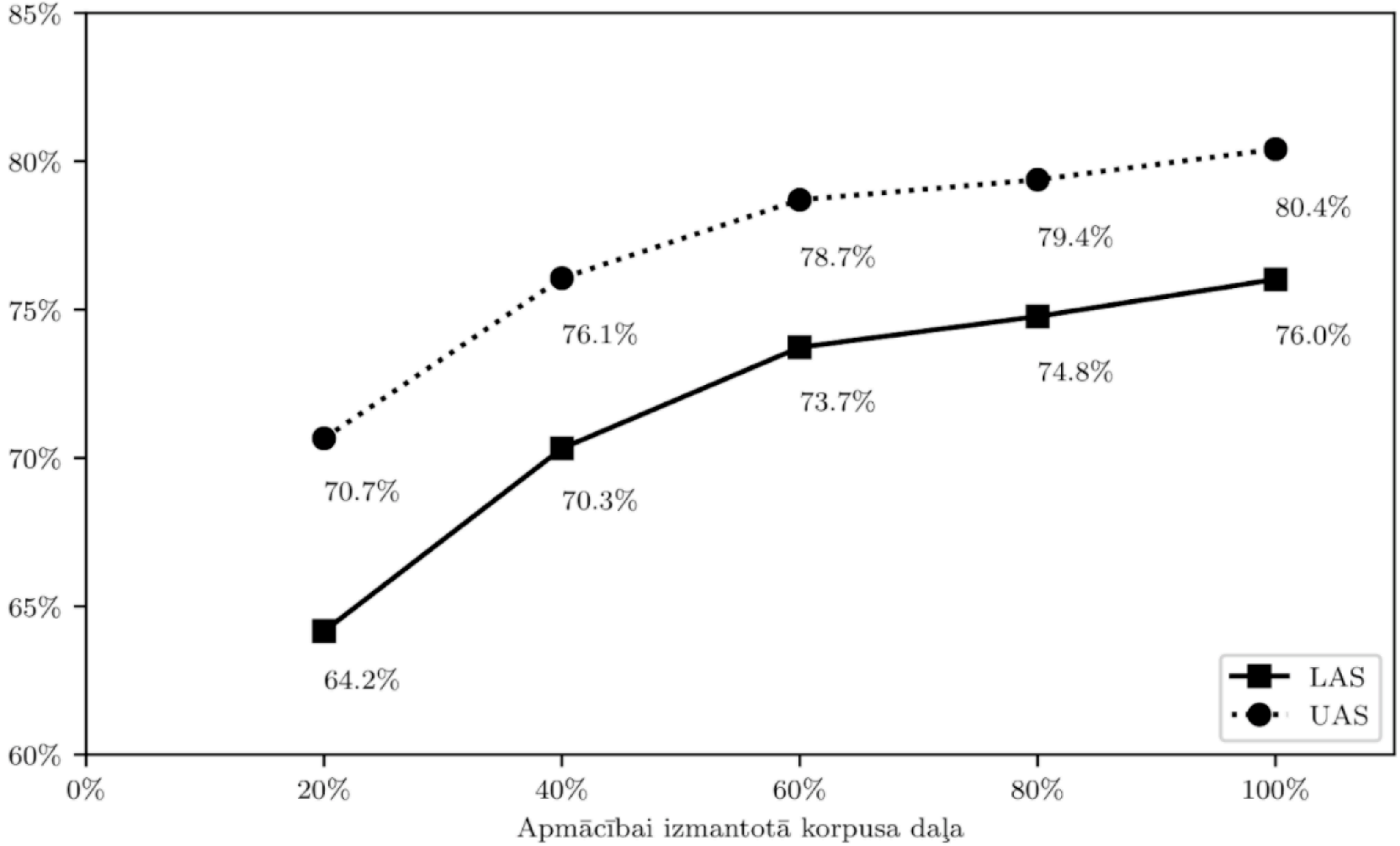


Genre	Trees	Percent	Aim	ToDo	Total
p	2774	49.6%	60%	4027	6801
d	2267	40.5%	20%	0	2267
n	27	0.5%	10%	1106	1133
s	57	1.0%	5%	509	566
c	280	5.0%	3%	60	340
z	188	3.4%	2%	38	226
Total	5593			5740	11333

newdoc id = p626
newpar id = p626-p6
sent_id = a-p626-p6s1
text = Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.

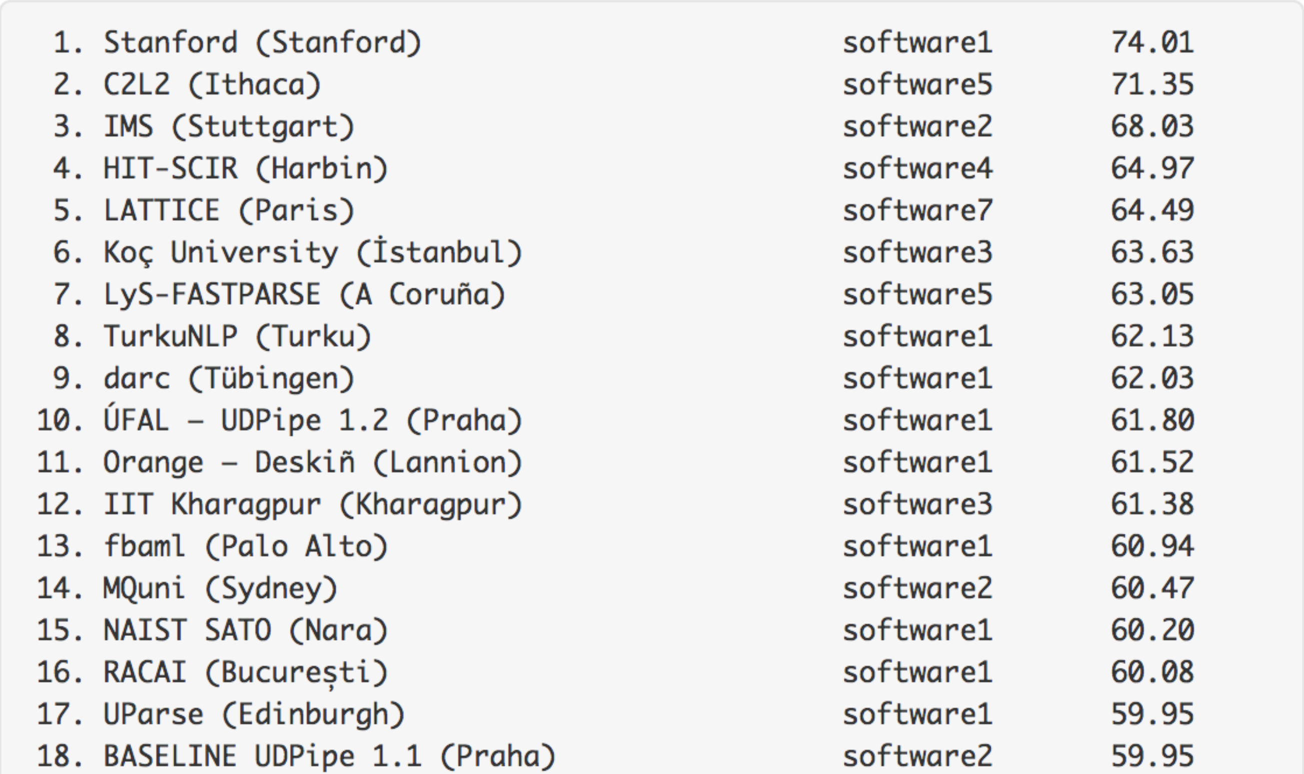
ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS
1	Šajā	šis	DET	pd0fsln	Case=Loc|Gender=Fem|Number=Sing|PronType=Dem	3	det	3:det
2	mācību	mācība	NOUN	ncfpg4	Case=Gen|Gender=Fem|Number=Plur	3	nmod	3:nmod:gen
3	gadā	gads	NOUN	ncmsl1	Case=Loc|Gender=Masc|Number=Sing	9	obl	9:obl:loc
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	Case=Gen|Gender=Fem|Number=Sing	5	nmod	5:nmod:gen
5	novada	novads	NOUN	ncmsg1	Case=Gen|Gender=Masc|Number=Sing	6	nmod	6:nmod:gen
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	Case=Gen|Gender=Fem|Number=Sing	8	nmod	8:nmod:gen
7	8.	8.	ADJ	xo	NumType=Ord	8	amod	8:amod
8	klasē	klase	NOUN	ncfsl5	Case=Loc|Gender=Fem|Number=Sing	9	obl	9:obl:loc
9	mācījās	mācīties	VERB	vmyisi330an	Evident=Fh|Mood=Ind|Person=3|Polarity=Pos|Reflex=Yes	0	root	0:root
10	Marisa	Marisa	PROPN	npfsn4	Case=Nom|Gender=Fem|Number=Sing	9	nsubj	9:nsubj
11	Butnere	Butnere	PROPN	npfsn5	Case=Nom|Gender=Fem|Number=Sing	10	flat:name	10:flat:name
12	no	no	ADP	spsg	_	13	case	13:case
13	Amerikas	Amerika	PROPN	npfsg4	Case=Gen|Gender=Fem|Number=Sing	10	nmod	10:nmod:no
14	.	.	PUNCT	zs	_	9	punct	9:punct

UD – Universal Dependencies



CoNLL 2017 Shared Task

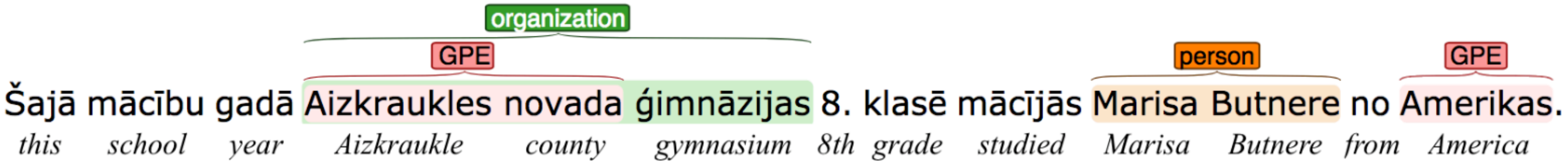
- Latvian among the **big** treebanks
  - ar, bg, ca, cs, cs_cac, cs_cltt, cu, da, de, el, en, en_lines, en_partut, es, es_ancora, et, eu, fa, fi, fi_ftb, fr, fr_sequoia, gl, got, grc, grc_proiel, he, hi, hr, hu, id, it, ja, ko, la_ittb, la_proiel, **lv**, nl, nl_lassysmall, no_bokmaal, no_nynorsk, pl, pt, pt_br, ro, ru, ru_syntagrus, sk, sl, sv, sv_lines, tr, ur, vi, zh
- Labeled Attachment Score (LAS)

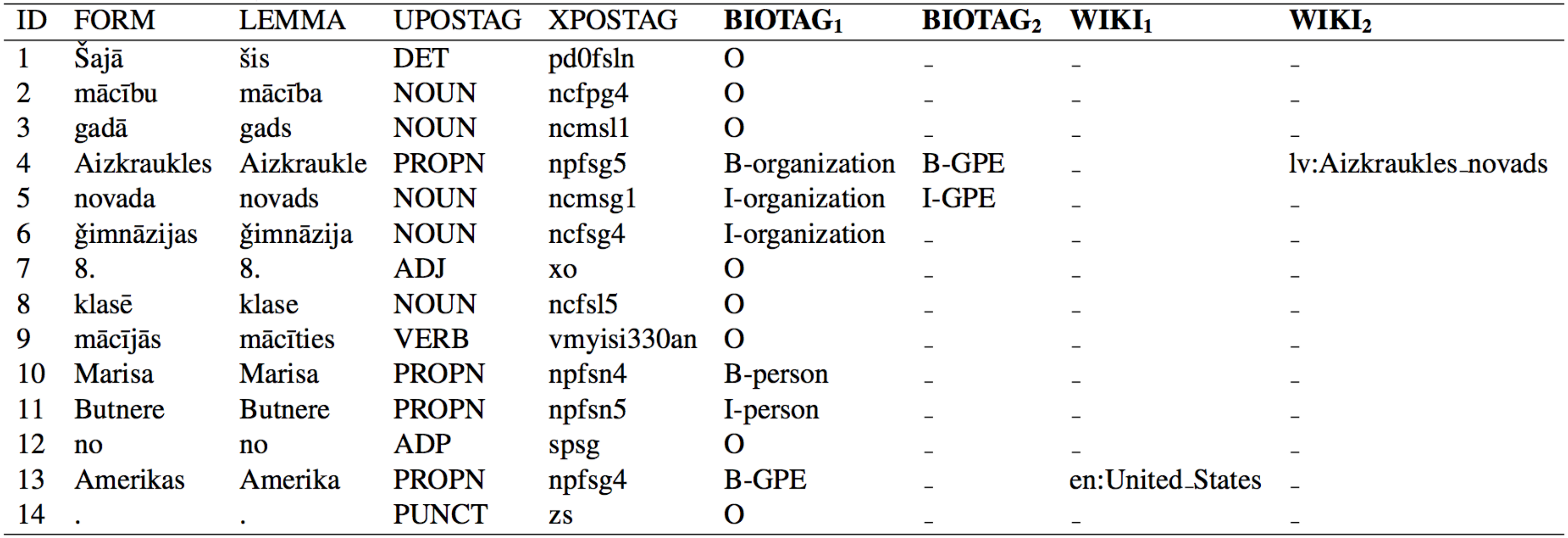


1.	Stanford (Stanford)	software1	74.01
2.	C2L2 (Ithaca)	software5	71.35
3.	IMS (Stuttgart)	software2	68.03
4.	HIT-SCIR (Harbin)	software4	64.97
5.	LATTICE (Paris)	software7	64.49
6.	Koç University (İstanbul)	software3	63.63
7.	LyS-FASTPARSE (A Coruña)	software5	63.05
8.	TurkuNLP (Turku)	software1	62.13
9.	darc (Tübingen)	software1	62.03
10.	ÚFAL – UDPipe 1.2 (Praha)	software1	61.80
11.	Orange – Deskiñ (Lannion)	software1	61.52
12.	IIT Kharagpur (Kharagpur)	software3	61.38
13.	fbaml (Palo Alto)	software1	60.94
14.	MQuni (Sydney)	software2	60.47
15.	NAIST SATO (Nara)	software1	60.20
16.	RACAI (București)	software1	60.08
17.	UParse (Edinburgh)	software1	59.95
18.	BASELINE UDPipe 1.1 (Praha)	software2	59.95

Named Entities

- Categories (MUC + AMR): person, organization, GPE, location, product, time, event, and entity

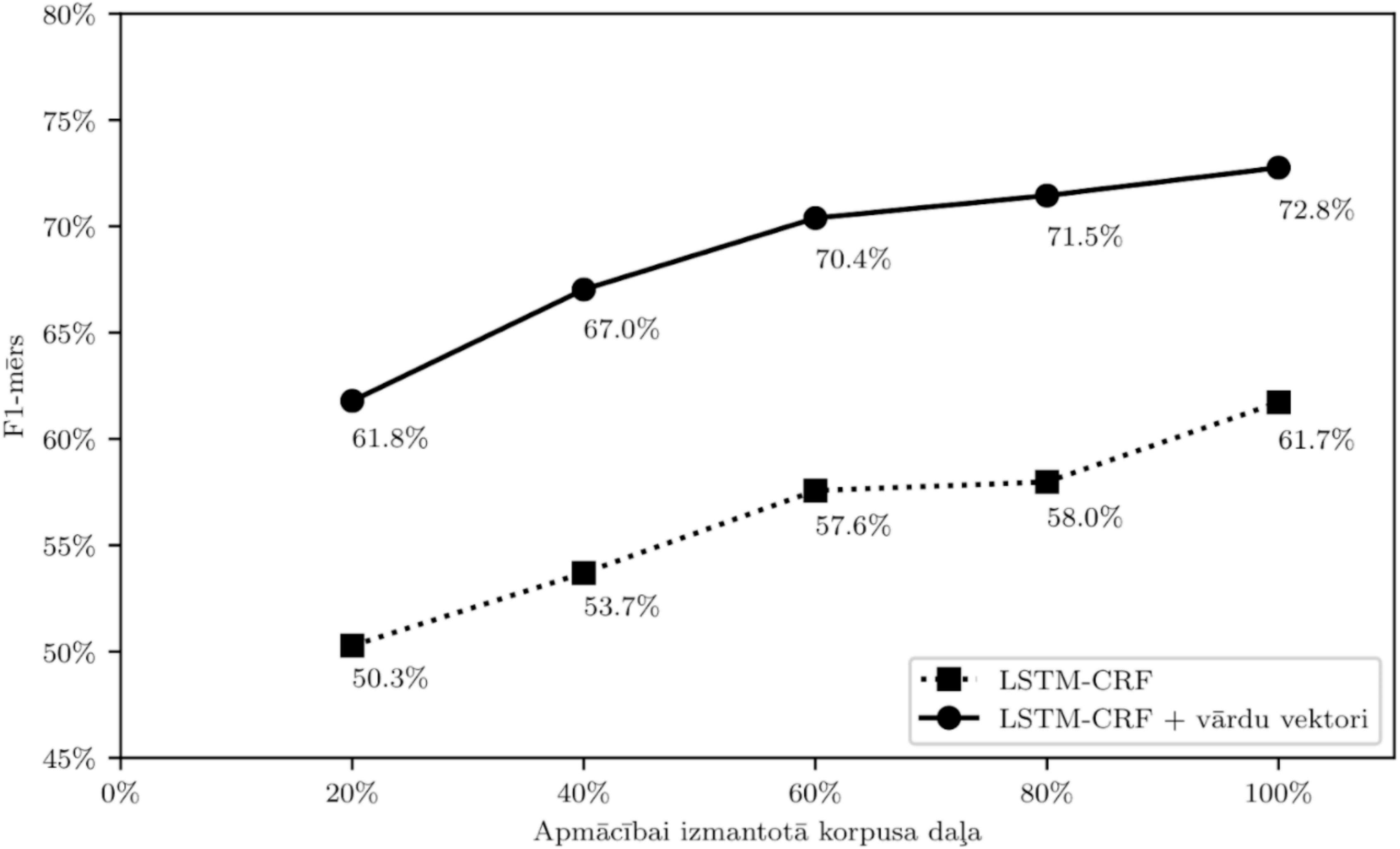

organization
GPE
person
GPE
Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.
this school year Aizkraukle county gymnasium 8th grade studied Marisa Butnere from America



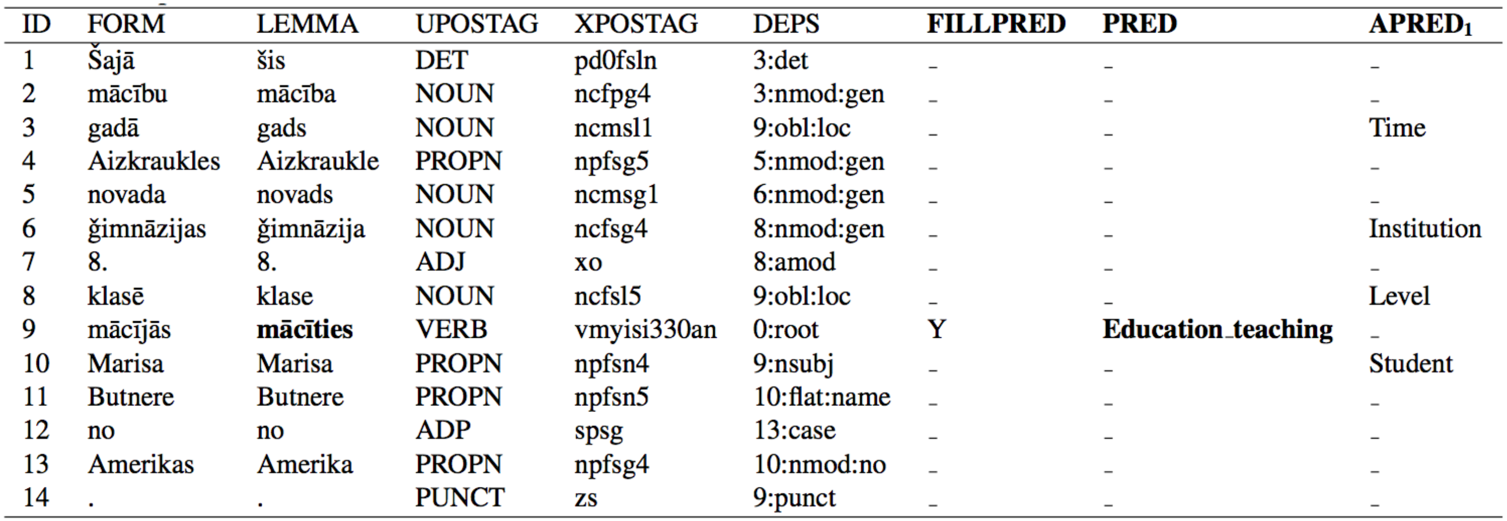
ID	FORM	LEMMA	UPOSTAG	XPOSTAG	**BIOTAG$_1$**	**BIOTAG$_2$**	**WIKI$_1$**	**WIKI$_2$**
1	Šajā	šis	DET	pd0fsln	O	_	_	_
2	mācību	mācība	NOUN	ncfpg4	O	_	_	_
3	gadā	gads	NOUN	ncmsl1	O	_	_	_
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	B-organization	B-GPE	_	lv:Aizkraukles_novads
5	novada	novads	NOUN	ncmsg1	I-organization	I-GPE	_	_
6	ğimnāzijas	ğimnāzija	NOUN	ncfsg4	I-organization	_	_	_
7	8.	8.	ADJ	xo	O	_	_	_
8	klasē	klase	NOUN	ncfsl5	O	_	_	_
9	mācījās	mācīties	VERB	vmyisi330an	O	_	_	_
10	Marisa	Marisa	PROPN	npfsn4	B-person	_	_	_
11	Butnere	Butnere	PROPN	npfsn5	I-person	_	_	_
12	no	no	ADP	spsg	O	_	_	_
13	Amerikas	Amerika	PROPN	npfsg4	B-GPE	_	en:United_States	_
14	.	.	PUNCT	zs	O	_	_	_

- Statistics (outer/inner): 1005/16 *person*, 808/57 *organization*, 593/200 GPE, 329/5 *time*, 193/24 *location*, 69/1 *event*, 64/1 *product*, 68/0 *entity* (**3129**/304)

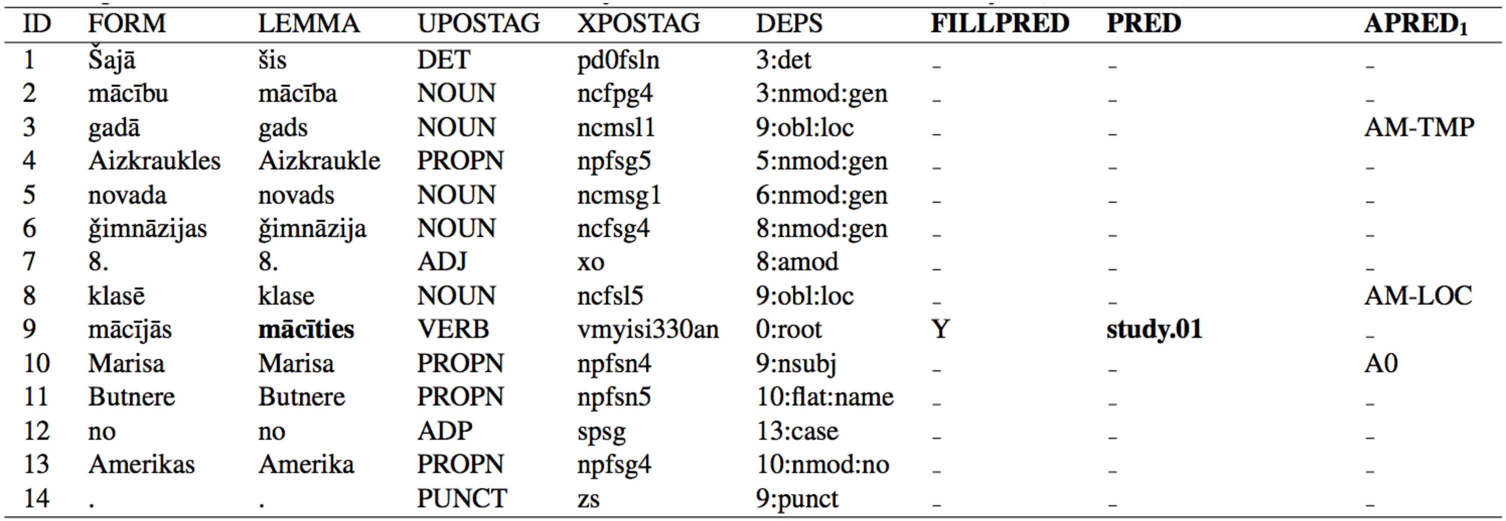
Named Entities



Semantic Frames

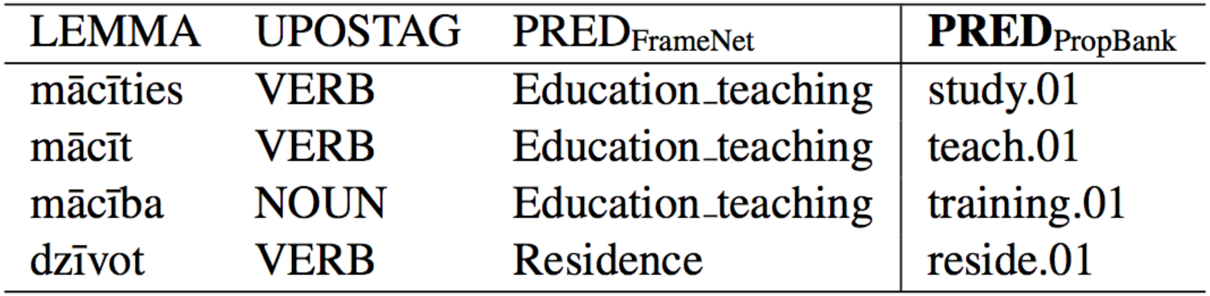


ID	FORM	LEMMA	UPOSTAG	XPOSTAG	DEPS	**FILLPRED**	**PRED**	**APRED$_1$**
1	Šajā	šis	DET	pd0fsln	3:det	_	_	_
2	mācību	mācība	NOUN	ncfpg4	3:nmod:gen	_	_	_
3	gadā	gads	NOUN	ncmsl1	9:obl:loc	_	_	Time
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	5:nmod:gen	_	_	_
5	novada	novads	NOUN	ncmsg1	6:nmod:gen	_	_	_
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	8:nmod:gen	_	_	Institution
7	8.	8.	ADJ	xo	8:amod	_	_	_
8	klasē	klase	NOUN	ncfsl5	9:obl:loc	_	_	Level
9	mācījās	**mācīties**	VERB	vmyisi330an	0:root	Y	**Education_teaching**	_
10	Marisa	Marisa	PROPN	npfsn4	9:nsubj	_	_	Student
11	Butnere	Butnere	PROPN	npfsn5	10:flat:name	_	_	_
12	no	no	ADP	spsg	13:case	_	_	_
13	Amerikas	Amerika	PROPN	npfsg4	10:nmod:no	_	_	_
14	.	.	PUNCT	zs	9:punct	_	_	_

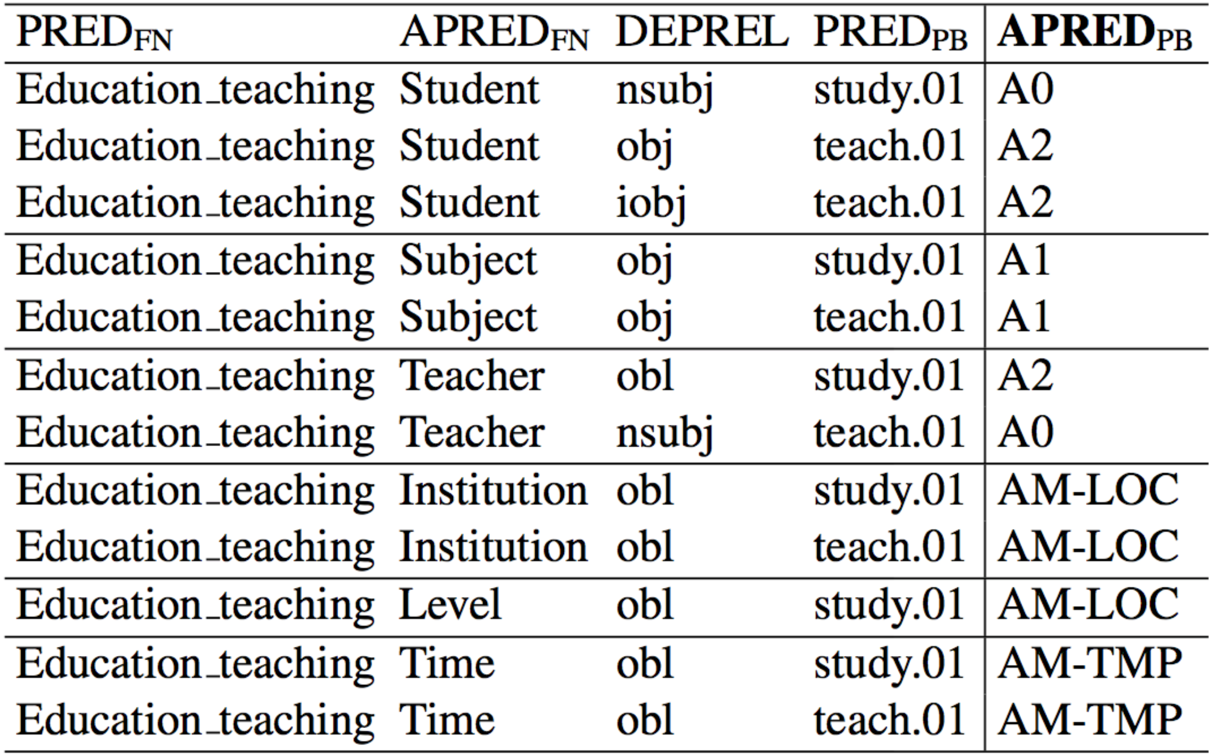


ID	FORM	LEMMA	UPOSTAG	XPOSTAG	DEPS	**FILLPRED**	**PRED**	**APRED$_1$**
1	Šajā	šis	DET	pd0fsln	3:det	_	_	_
2	mācību	mācība	NOUN	ncfpg4	3:nmod:gen	_	_	_
3	gadā	gads	NOUN	ncmsl1	9:obl:loc	_	_	AM-TMP
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	5:nmod:gen	_	_	_
5	novada	novads	NOUN	ncmsg1	6:nmod:gen	_	_	_
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	8:nmod:gen	_	_	_
7	8.	8.	ADJ	xo	8:amod	_	_	_
8	klasē	klase	NOUN	ncfsl5	9:obl:loc	_	_	AM-LOC
9	mācījās	**mācīties**	VERB	vmyisi330an	0:root	Y	**study.01**	_
10	Marisa	Marisa	PROPN	npfsn4	9:nsubj	_	_	A0
11	Butnere	Butnere	PROPN	npfsn5	10:flat:name	_	_	_
12	no	no	ADP	spsg	13:case	_	_	_
13	Amerikas	Amerika	PROPN	npfsg4	10:nmod:no	_	_	_
14	.	.	PUNCT	zs	9:punct	_	_	_

From FrameNet to PropBank



LEMMA	UPOSTAG	PRED$_{\text{FrameNet}}$	**PRED**$_{\text{PropBank}}$
mācīties	VERB	Education_teaching	study.01
mācīt	VERB	Education_teaching	teach.01
mācība	NOUN	Education_teaching	training.01
dzīvot	VERB	Residence	reside.01



PRED$_{\text{FN}}$	APRED$_{\text{FN}}$	DEPREL	PRED$_{\text{PB}}$	**APRED**$_{\text{PB}}$
Education_teaching	Student	nsubj	study.01	A0
Education_teaching	Student	obj	teach.01	A2
Education_teaching	Student	iobj	teach.01	A2
Education_teaching	Subject	obj	study.01	A1
Education_teaching	Subject	obj	teach.01	A1
Education_teaching	Teacher	obl	study.01	A2
Education_teaching	Teacher	nsubj	teach.01	A0
Education_teaching	Institution	obl	study.01	AM-LOC
Education_teaching	Institution	obl	teach.01	AM-LOC
Education_teaching	Level	obl	study.01	AM-LOC
Education_teaching	Time	obl	study.01	AM-TMP
Education_teaching	Time	obl	teach.01	AM-TMP

Current statistics (FN):

- 5093 annotation sets
- 319 frames
- 871 lexical units

ToDo: evaluation (SRL)

Availability

- https://github.com/LUMII-AILab/FullStack (to appear)
- Under a CC BY-NC-SA 4.0 license

- More details:
  - Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, Peteris Paikens. ***Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU***. Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), May 2018
  - Normunds Gruzitis, Gunta Nespore-Berzkalne, Baiba Saulite. ***Creation of Latvian FrameNet based on Universal Dependencies***. Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons, May 2018