



Multilayer Corpus and Toolchain for Full-Stack NLU in Latvian

Normunds Grūzītis normunds.gruzitis@lumii.lv Artūrs Znotiņš arturs.znotins@lumii.lv

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



EUROPEAN UNION European Regional Development Fund



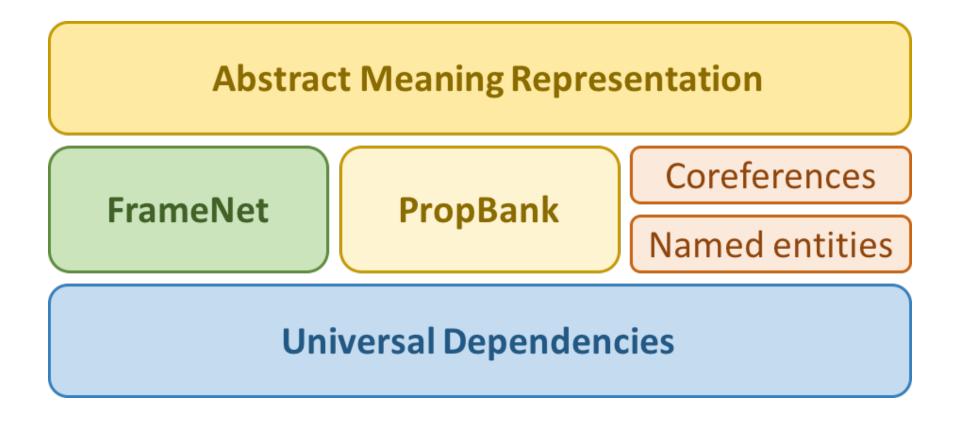
CLARIN Annual Conference 10 October 2018

INVESTING IN YOUR FUTURE

Grant agreement No. 1.1.1.1/16/A/219

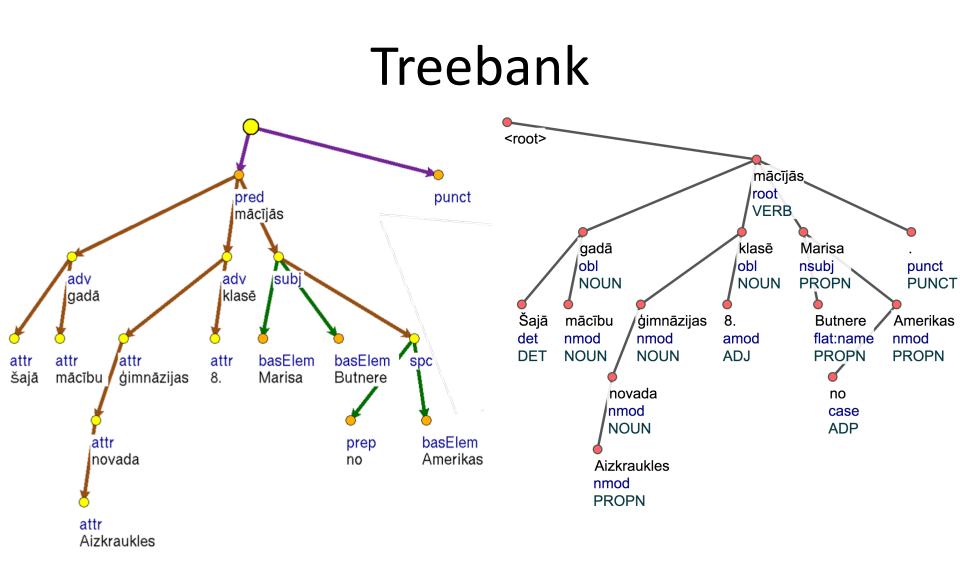
Grant agreement No. 1.1.1.5/18/I/016 Sub-activity of CLARIN Latvia

Multilayer Corpus



Balanced Data Set

- Goal of around 10-15K sentences
- Texts from a balanced text corpus of Latvian
- Isolated paragraphs as the main text unit
- Selected to cover 1000 most common verbs
- Balanced in terms of:
 - Genres
 - Lexical units
 - Writing styles

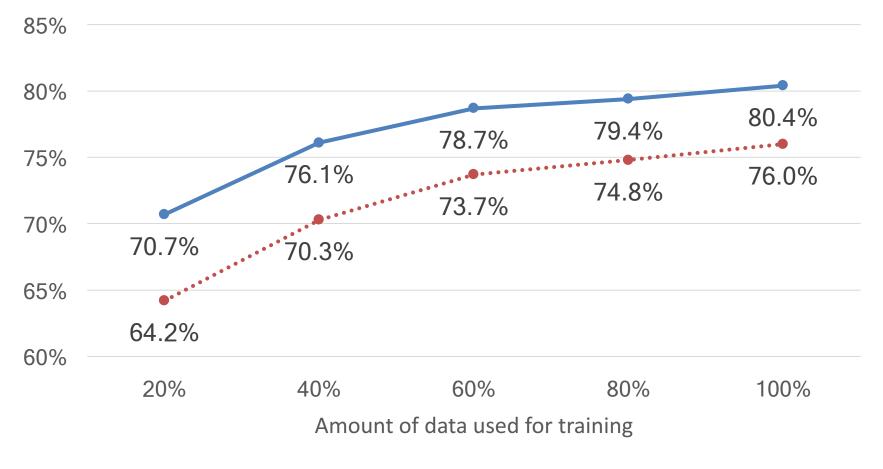


Hybrid dependency-constituency grammar model



Universal dependencies Enhanced universal dependencies

Treebank: Draft Tree Generation



-UAS ··•··LAS

Treebank: Statistics

Genre	Trees	Percent	Aim	ToDo	Total
news	3,992	45.5%	60%	3,931	7,923
fiction	2,641	30.1%	20%	0	2,641
academic	786	9.0%	7%	138	924
legal	297	3.4%	6%	495	792
spoken	648	7.4%	5%	12	660
other	405	4.6%	2%	-140	265
Total	8,769			4,436	13,205

Treebank: CONLL-U

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	хо	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

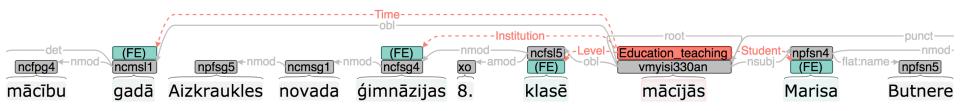
Named Entities and Coreferences

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

				anization				_
			GPE				pers	Son GPE
Šaj	jā mācību	gadā <mark>Aizk</mark>	raukles n	ovada ģim	nāzijas 8. kl	asē mācījā	ās <mark>Marisa E</mark>	Butnere no Amerikas
thi	s school	year Aiz	kraukle	county gyn	nnasium 8th gr	ade studie	d Marisa	Butnere from America
ID	FORM	LEMMA	UPOSTAG	XPOSTAG	BIOTAG ₁	BIOTAG₂	WIKI ₁	WIKI ₂
1	Šajā	šis	DET	pd0fsln	0	_	_	-
2	mācību	mācība	NOUN	ncfpg4	0	_	_	_
3	gadā	gads	NOUN	ncmsl1	0	_	_	_
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	хо	0	_	_	_
8	klasē	klase	NOUN	ncfsl5	0	-	_	_
9	mācījās	mācīties	VERB	vmyisi330an	0	_	_	_
10	Marisa	Marisa	PROPN	npfsn4	B -person	_	_	_
11	Butnere	Butnere	PROPN	npfsn5	I-person	_	_	_
12	no	no	ADP	spsg	0	_	_	-
13	Amerikas	Amerika	PROPN	npfsg4	B-GPE	_	en:United_Sta	ites _
14	•	•	PUNCT	ZS	0	-	-	-

Statistics (outer/inner): 1563/19 *person*, 980/90 *organization*, 752/222 **GPE**, 484/10 *time*, 282/38 *location*, 81/3 *event*, 104/1 *product*, 97/8 *entity* (total: 4343/391)

Semantic Frames



Current statistics	
Target words (verb lexemes)	781 (cf. 1,000)
Annotation sets (frame instances)	7,143
Berkeley FrameNet frames covered	435
Lexical units (lexeme-frame pairs)	1,440 (cf. 2,000)
Annotation sets per lexeme (avg.)	9.15
Annotation sets per lexical unit (avg.)	4.96

Semantic Frames: Concordance View



FrameNet and PropBank

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	ğimnāzijas	ğimnāzija	NOUN	ncfsg4	8:nmod:gen	-	-	Institution
7	8.	8.	ADJ	хо	8:amod	-	-	_
8	klasē	klase	NOUN	ncfs15	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		šis	DET	pd0fsln	3:det	_	-	_
	Šajā mācību			pd0fs1n ncfpg4	3:det 3:nmod:gen	-	-	-
1	Šajā	šis	DET	•			-	- - AM-TMP
1 2	Šajā mācību	šis mācība	DET NOUN	ncfpg4	3:nmod:gen			-
1 2 3	Šajā mācību gadā	šis mācība gads	DET NOUN NOUN	ncfpg4 ncmsl1	3:nmod:gen 9:obl:loc			-
1 2 3 4 5	Šajā mācību gadā Aizkraukles	šis mācība gads Aizkraukle	DET NOUN NOUN PROPN	ncfpg4 ncmsl1 npfsg5 ncmsg1	3:nmod:gen 9:obl:loc 5:nmod:gen			-
1 2 3 4 5	Šajā mācību gadā Aizkraukles novada	šis mācība gads Aizkraukle novads	DET NOUN NOUN PROPN NOUN	ncfpg4 ncmsl1 npfsg5	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen		- - - -	-
1 2 3 4 5 6	Šajā mācību gadā Aizkraukles novada ğimnāzijas	šis mācība gads Aizkraukle novads ğimnāzija	DET NOUN NOUN PROPN NOUN NOUN	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen		- - - - -	-
1 2 3 4 5 6 7	Šajā mācību gadā Aizkraukles novada ğimnāzijas 8.	šis mācība gads Aizkraukle novads ğimnāzija 8.	DET NOUN NOUN PROPN NOUN NOUN ADJ	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4 xo	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen 8:amod	- - - - - Y	- - - - - - study.01	- AM-TMP - - -
1 2 3 4 5 6 7 8	Šajā mācību gadā Aizkraukles novada ğimnāzijas 8. klasē	šis mācība gads Aizkraukle novads ğimnāzija 8. klase	DET NOUN NOUN PROPN NOUN NOUN ADJ NOUN	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4 x0 ncfsl5	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen 8:amod 9:obl:loc	- - - - -	- - - - -	- AM-TMP - - -
1 2 3 4 5 6 7 8 9	Šajā mācību gadā Aizkraukles novada ğimnāzijas 8. klasē mācījās	šis mācība gads Aizkraukle novads ģimnāzija 8. klase mācīties	DET NOUN NOUN PROPN NOUN NOUN ADJ NOUN VERB	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4 xo ncfsl5 vmyisi330an	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen 8:amod 9:obl:loc 0:root	- - - - -	- - - - -	- AM-TMP - - - - AM-LOC
1 2 3 4 5 6 7 8 9 10	Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa	šis mācība gads Aizkraukle novads ģimnāzija 8. klase mācīties Marisa	DET NOUN NOUN PROPN NOUN NOUN ADJ NOUN VERB PROPN	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4 x0 ncfsl5 vmyisi330an npfsn4	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen 8:amod 9:obl:loc 0:root 9:nsubj	- - - - -	- - - - -	- AM-TMP - - - - AM-LOC
1 2 3 4 5 6 7 8 9 10 11	Šajā mācību gadā Aizkraukles novada ğimnāzijas 8. klasē mācījās Marisa Butnere	šis mācība gads Aizkraukle novads ģimnāzija 8. klase mācīties Marisa Butnere	DET NOUN NOUN PROPN NOUN NOUN ADJ NOUN VERB PROPN PROPN	ncfpg4 ncmsl1 npfsg5 ncmsg1 ncfsg4 xo ncfsl5 vmyisi330an npfsn4 npfsn5	3:nmod:gen 9:obl:loc 5:nmod:gen 6:nmod:gen 8:nmod:gen 8:amod 9:obl:loc 0:root 9:nsubj 10:flat:name	- - - - -	- - - - -	- AM-TMP - - - - AM-LOC

From FrameNet to PropBank

LEMMA	UPOSTAG	PRED _{FrameNet}	PRED _{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 _{FN}	APRED _{FN}	DEPREL	PRED _{PB}	APRED _{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

Toolchain

- Core processing tools:
 - Tokenization
 - Morphological tagging
 - Dependency parsing
 - Named Entity recognition
 - Coreference resolution
 - FrameNet parsing
- NLP-PIPE library

Toolchain Demo

\leftrightarrow \rightarrow C i nlp.ailab.lv	☆ = :
Arī otra figūra "Daimler" lietā ir Bojāra ārštata padomnieks, un sens eksmēra drau laikiem – Armands Zeihmanis.	ugs no armijas
Go tokenizer × morpho × parser × ner × Image: NER CONLL JSON	
Arī otra figūra <mark>" Daimler " organization</mark> lietā ir <mark>Bojāra person</mark> ārštata pado un sens eksmēra draugs no armijas laikiem – <mark>Armands Zeihmanis persor</mark>	

Availability

- Final release in the end of 2019
- Corpus
 - <u>https://github.com/LUMII-AILab/FullStack</u>
 - Dual license (CC BY-NC-SA 4.0 and commercial)
 - CLARIN Latvia
- Toolchain
 - <u>http://nlp.ailab.lv</u>
 - <u>http://hub.docker.com</u> (to appear)
 - Dual license (GPL v3 and commercial)
 - CLARIN Latvia