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

NATIONAL DEVELOPMENT PLAN 2020

EUROPEAN UNION
European Regional Development Fund

INVESTING IN YOUR FUTURE

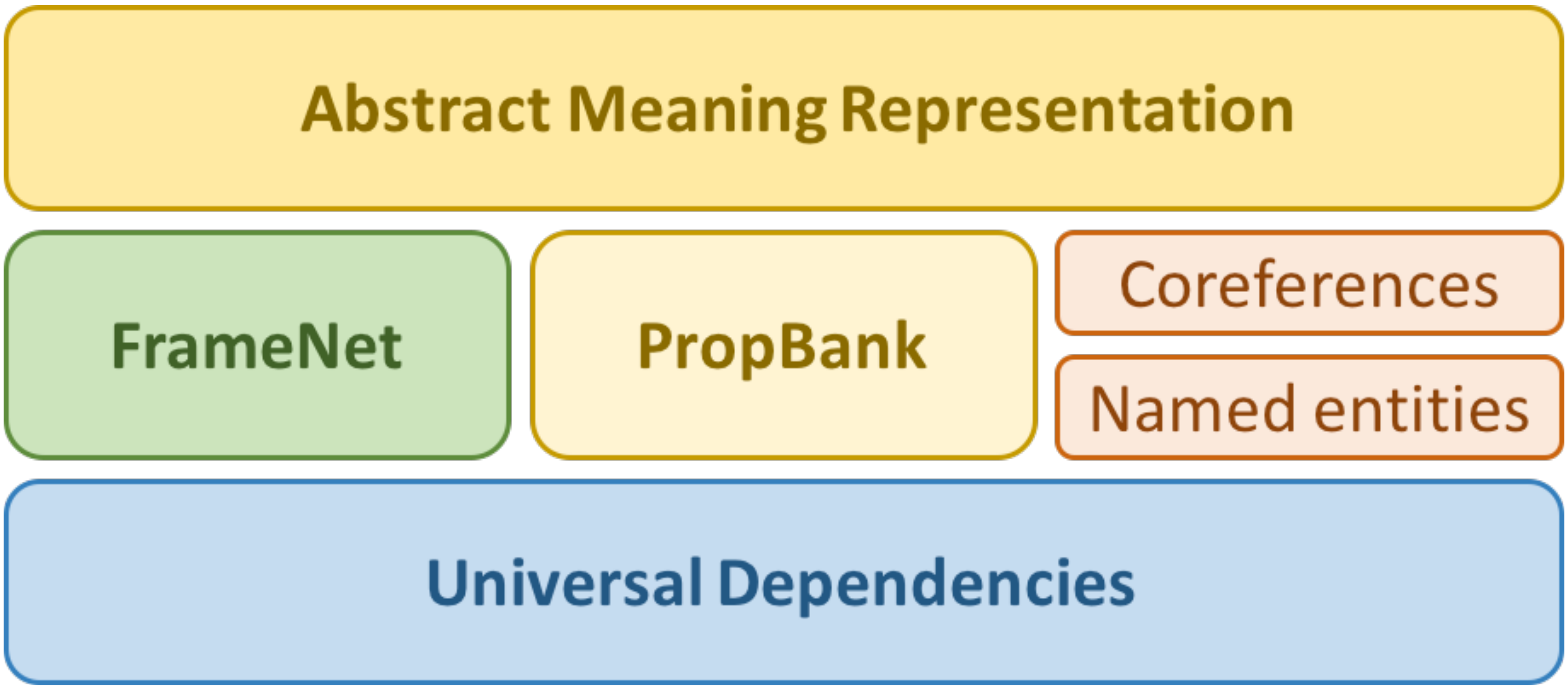
Grant agreement No. 1.1.1.1/16/A/219

CLARIN

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

CLARIN Annual Conference
10 October 2018

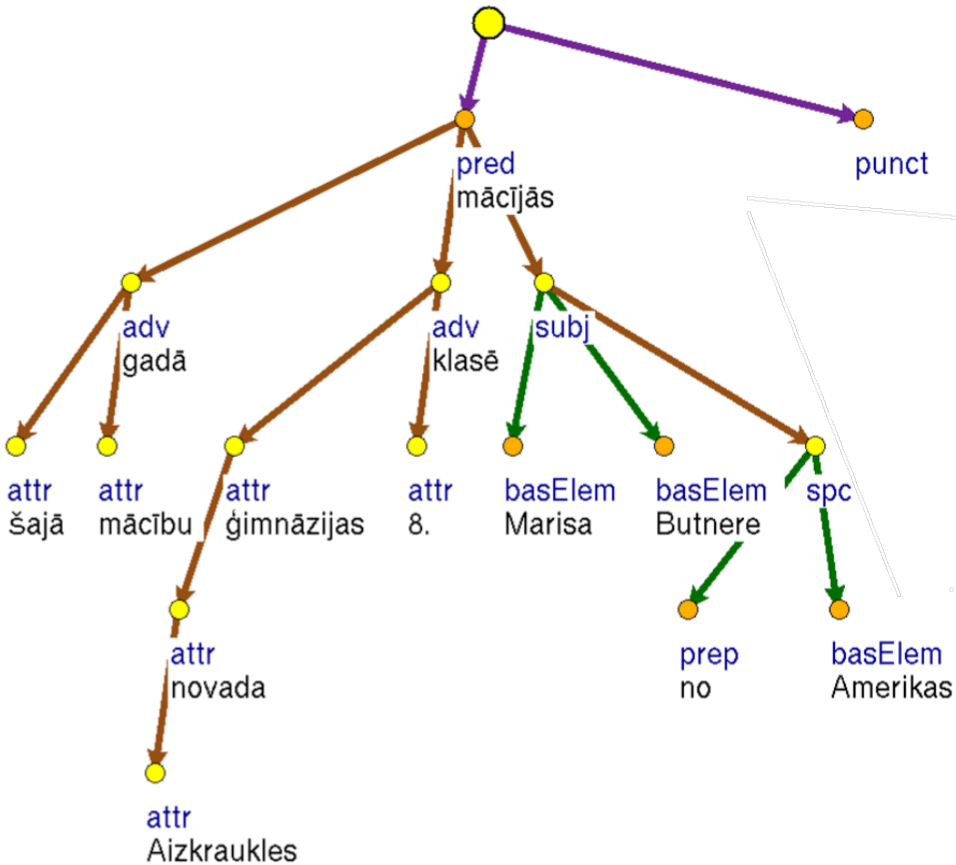
Multilayer Corpus

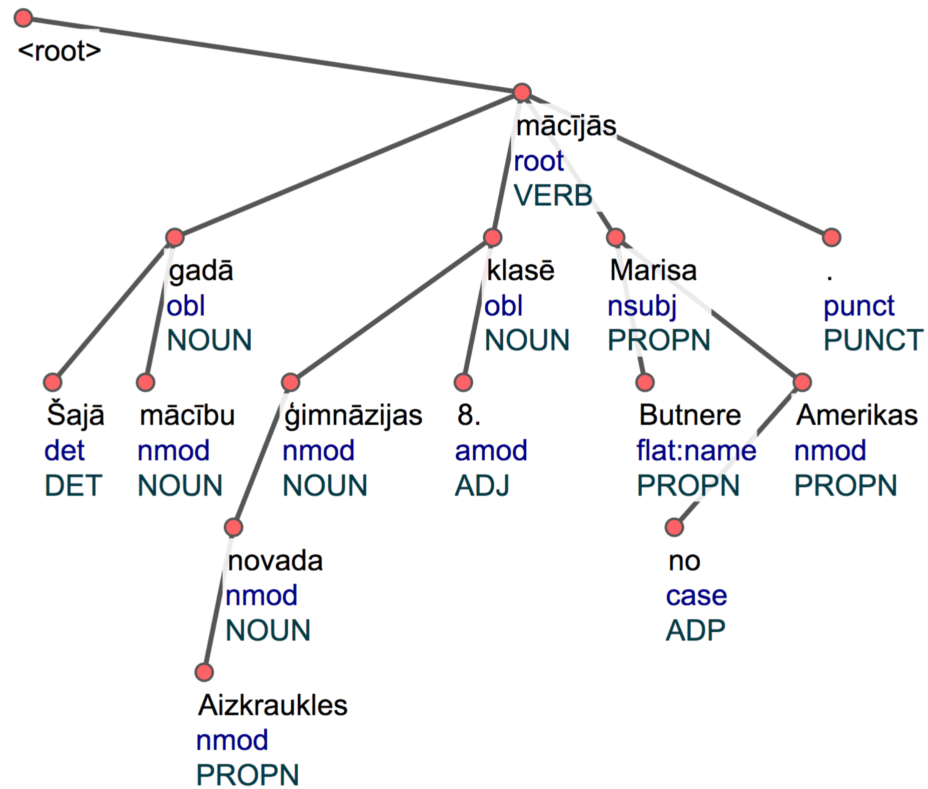


Balanced Data Set

- 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

Treebank



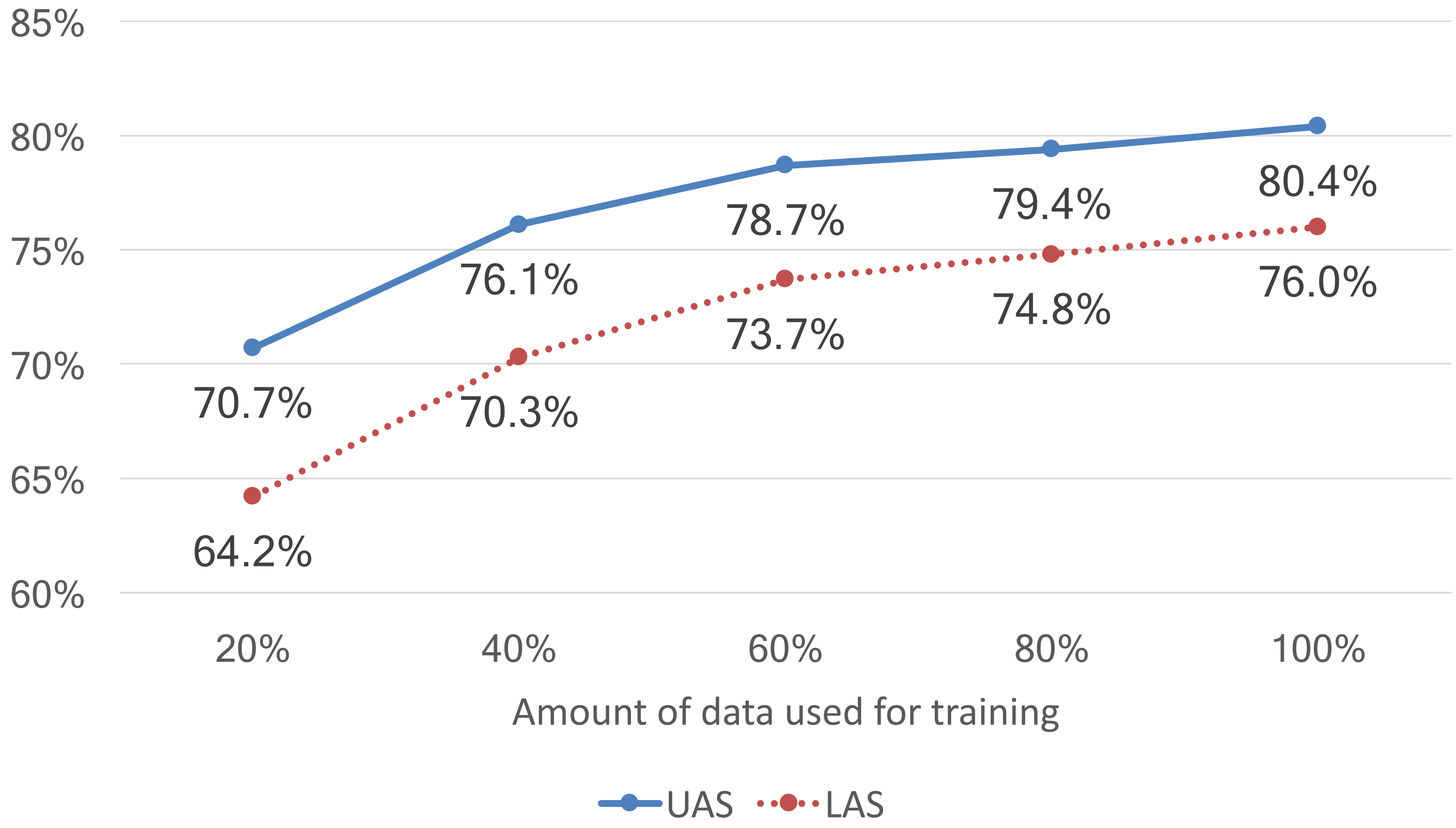


Hybrid dependency-constituency grammar model

Universal dependencies

Enhanced universal dependencies

Treebank: Draft Tree Generation



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	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

Named Entities and Coreferences

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

organization; GPE; person; GPE

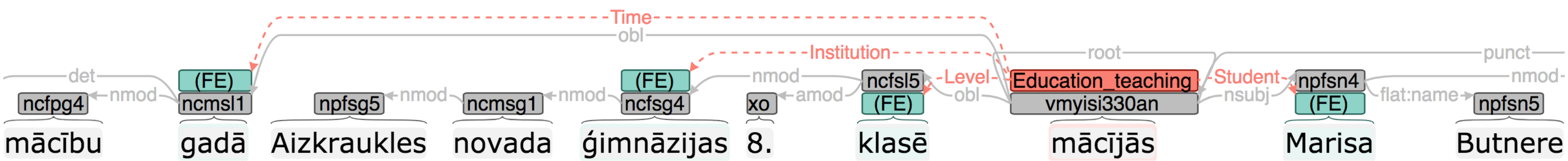
Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.

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	ģimnāzijas	ģimnāzija	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): 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	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

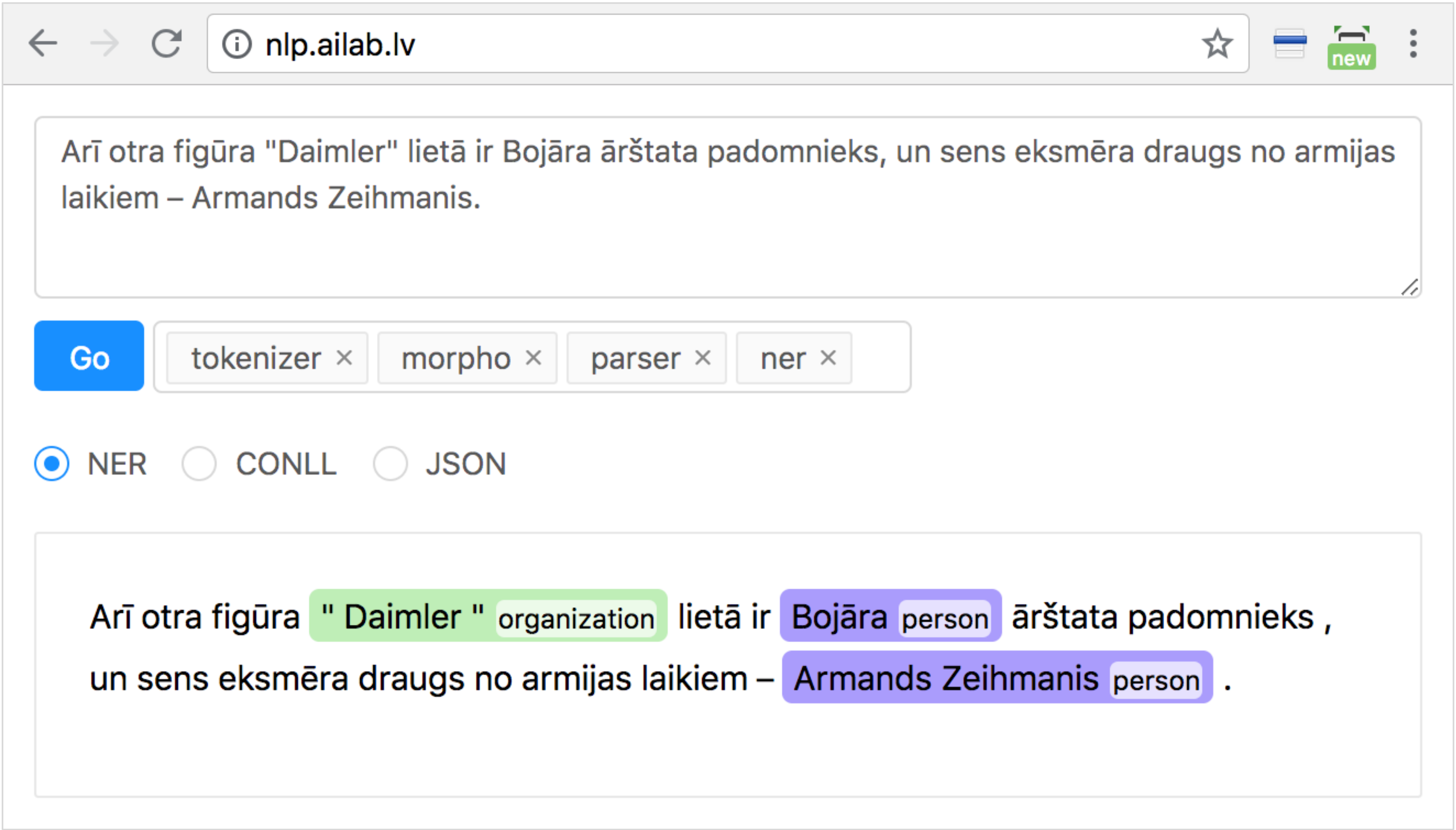
LEMMA	UPOSTAG	$\mathrm{PRED}_{\mathrm{FrameNet}}$	$\mathbf{PRED}_{\mathrm{PropBank}}$
mācīties	VERB	Education_teaching	study.01
mācīt	VERB	Education_teaching	teach.01
mācība	NOUN	Education_teaching	training.01
dzīvot	VERB	Residence	reside.01

$\mathrm{PRED}_{\mathrm{FN}}$	$\mathrm{APRED}_{\mathrm{FN}}$	DEPREL	$\mathrm{PRED}_{\mathrm{PB}}$	$\mathbf{APRED}_{\mathrm{PB}}$
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



Availability

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