

Latvian National Corpora Collection – Korpuss.lv

Baiba Saulīte, Roberts Dargis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa,
Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns,
Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, Jānis Ziediņš

Institute of Mathematics and Computer Science, University of Latvia

National Library of Latvia

Culture Information Systems Centre

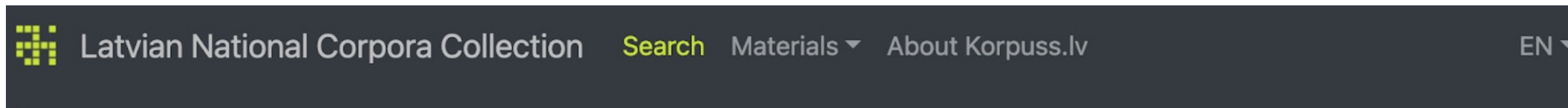
roberts.dargis@lumii.lv, anda.baklane@lnb.lv, janis.ziedins@kis.gov.lv

Abstract. Latvian National Corpora Collection (LNCC) is a diverse collection of corpora representing both written and spoken language. LNCC covers various use cases and all the important text types and genres. It is a continuous multi-institutional and multi-project effort, supported by the Digital Humanities and Language Technology communities in Latvia. All corpora of LNCC are annotated with a uniform morpho-syntactic annotation scheme which enables federated search and consistent linguistics analysis in all the LNCC corpora and allows to select and mix various corpora for pre-training large Latvian language models like BERT and GPT.

(1) The current compilation of LNCC

Code name	Full name	Size	Type	Release
<i>Written language text corpora</i>				
LVK2018	Balanced Corpus of Modern Latvian (Levāne-Petrova and Dargis, 2018)	12M tokens	text, general, representative	2016–2018
UDLV-LVTB	Latvian UD Treebank, subset of LVK2018, part of UD v2.9 (Zeman et al., 2021)	266k tokens (16k sent.)	text, general, representative, manually annotated	2015–2021
Hugo.lv	Hugo.lv Parallel Corpora	10.5M tokens	text, general, culture	2018
Timeklis2007	Latvian Web Corpus (Džeriņš and Džonsons, 2007)	123.5M tokens	text, web	2006–2007
Timeklis2020	Latvian Web Corpus	492.6M tokens	text, web	2020–2022
Vikipēdija	Latvian Wikipedia	27.7M tokens	text, specialised	2022
Emuāri	Latvian Blog Corpus	8M tokens	text, specialised	2014–2015
Barometrs	Corpus of News Portal Comments	447.3M tokens	text, specialised	2011–2021
Saeima	Corpus of Latvian Parliament Debates (Auziņa et al., 2018)	24M tokens	text, specialised:parliamentary	2013–2018
Likumi	Corpus of Legal Acts of the Republic of Latvia	116.2M tokens	text, specialised	2022
LiLa	Lithuanian-Latvian-Lithuanian Parallel Text Corpus (Utka et al., 2013)	5.7M tokens	text, parallel, representative	2011–2013
MuLa	Corpus of Contemporary Latgalian Texts (Sperga et al., 2013)	1.3M tokens	text, specialised:dialect, representative	2011–2013
LaVA	Latvian Language Learner Corpus (Auziņa et al., 2021)	241k tokens	text, specialised:learner, manually annotated, error annotation	2018–2021
Pārspriedumi	Corpus of Students Essays (Levāne-Petrova et al., 2021)	226k tokens	text, specialised	2018–2021
Disertācijas	Corpus of Latvian PhD Theses	23.4M tokens	text, specialised	2022
LatSenRom	Corpus of Latvian Early Novels	3.3M tokens	text, specialised:literary	2019–2021
Rainis	Corpus of Texts Written by Rainis (Spektors et al., 2018)	2.3M tokens	text, specialised:literary	2018
Senie	Corpus of Early Written Latvian Texts (Andronova et al., 2002)	2.7M tokens	text, specialised:diachronic	2002–2021
<i>Spoken language text corpora</i>				
LRK2013	Latvian Speech Recognition Corpus	975k tokens (100 hours)	spoken, general, representative	2013
Subtitri	Latvian Subtitles of Public Broadcasting	10.8M tokens (1200 hours)	spoken, specialised	2020–2022
LVMED	Latvian Medical Speech Corpus	157k tokens (35 hours)	spoken, specialised	2022
LNCC:	21 corpora	1.3B tokens		Jun 2022

(2) The Korpuss.lv website



general (6) web (2) learner (3) literary (2) parallel (1) parliamentary (1) diachronic (1)
specialised (6) representative (7) text (18) speech (4) morphology (17) syntax (3) semantics (1)
error annotation (2) manually annotated (4)

Lvk2018

The Balanced Corpus of Modern Latvian
2016–2018, 10M words (12M tokens)

Developers: IMCS UL

f More

🔍 Search

LaVa

Latvian Language Learner Corpus
2018–2021, 192k words (241k tokens)

Developers: IMCS UL

f More

🔍 Search

LaVa

Latvian Language Learner Corpus

The corpus includes more than 1000 texts created by foreign Latvian language learners studying at Latvian higher education institutions for the first or second semester. The morphologically annotated texts have been checked manually; the language learners' errors have been manually annotated.

Publication to be cited:
R. Dargis and I. Auzina & K. Levene-Petrova and I. Kaļa
Quality Focused Approach to a Learner Corpus Development
Proceedings of The 12th Language Resources and Evaluation Conference (LREC), 382–398, 2020
[PDF](#)

test (20)
learner (13)
morphology (6)
error annotation (2)
manually annotated (6)

Corpus size	192k words (241k tokens)
Development period	2018–2021
Developers	Institute of Mathematics and Computer Science UL
Funding	The project "Development of learner corpus of Latvian: methods, tools and applications" (project No. lzp-2018/1-0527) is financed by Latvian Council of Science.
Homepage	http://lava.korpusi.lv/
CLARIN	http://hdl.handle.net/20.500.12574/449
Other publications	I. Kaļa and I. Auzina Data collection for learner corpus of Latvian; copyright and personal data protection Selected papers from the CLARIN Annual Conference 2018, 41–47, 2020 PDF DOI

(3) Federated search result for the query *sirds** ('heart*')

sirds* ? Search

2530 occurrences (1541 per million)	LatSenRom Corpus of Latvian Early Novels	565 occurrences (426 per million)	MuLa Corpus of Contemporary Latgalian Texts
1734 occurrences (305 per million)	LiLa Lithuanian-Latvian-Lithuanian Parallel Text Corpus	9 occurrences (37 per million)	LaVA Latvian Language Learner Corpus
2128 occurrences (256 per million)	Emuāri Latvian Blog Corpus 2015	157 occurrences (137 per million)	LRK2013 Latvian Speech Recognition Corpus
3321 occurrence (270 per million)	LVK2018 The Balanced Corpus of Modern Latvian	510 occurrences (2257 per million)	Pārspriedumi Corpus of Students' Essays
13 occurrences (5 per million)	Senie Corpus of Early Written Latvian Texts	2752 occurrences (1196 per million)	Rainis Corpus of Texts Written by Rainis
2922 occurrences (121 per million)	Saeima Corpus of the Saeima (the Parliament of Latvia)	1230 occurrences (114 per million)	Subtitri Latvian Subtitles of Public Broadcasting
19 970 occurrences (162 per million)	Timeklis2007 Latvian Web Corpus 2007	164 504 occurrences (334 per million)	Timeklis2020 CommonCrawl of Latvian 2020

(4.1) Common tagset: morphology

pateicas	Dievām	grozās	sirdās	par	zeņķiem	kas	spļauj
0an/pateikties	ncmsd1/dievs	zcl/, vmyip_330an/grozīties	zcl/, vmyip_330an/sirdities	spdp/par	ncmpd2/zeņķis	zcl/, pr000n/kas	vmini130an/spļaut
uzkozdams	kas	gadījās	sirdīgi	pārmeta	:	Tu	Mairita
isn0000n/uzkost	zcl/, pr000n/kas	vmyis_330an/gadīties	zcl/, rp_1sirdīgi	vmini130an/pārmest	zcl/: zcl/,	pu20snn/tu	zcl/, nplsn4/Mairita
šc	notiek	sirdsdarbības	apstāšanās	Sirds	darbojas	pateicoties	tam
šc	vmini130an/notikt	ncfsg4/sirdsdarbība	ncfsg4/apstāšanās	ncfns6/sirds	vmyip_230an/darboties	zcl/, vmymp000000n/pateikties	pdm3s
šam	ilgi	bija	apslēpts	Sirdsbals	gan	sacīja	priekšā
ī/pats	rp_/ilgi	vminis_330an/būt	vmpndmsnpsnpr/apslēpt	zcl/?	cc/gan	vminis330an/sacīt	r0_/priekšā
ēnsībās	kurās	ir	aktiera	sirdslieta	Patriks	piekritis	sievas
šacensība	zcl/, pr0īpn/kura	vncipi30an/būt	ncmsg2/aktieris	ncfns4/sirdslieta	zsl/, nprnsn1/Patriks	vmpndmsnpsnpr/piekrīst	ncfsg4/sieva
weat-heart	(	Ar	labu	nakti	)	—	vai šoreiz mājās
weet-heart	zb/(	zq/"	spsa/ar	afssanp/labs	ncfsa6/nakts	zcl/,	cc/vai
acis	—	Visuspēcīgais	Dievs	sirdsžēlīgais	Tēvs	Nu es	zinu
1c1cpa6/acis	zcl/, zcl-	afmsynp/visuspēcīgs	ncmsn1/dievs	zcl/, afmsynp/sirdsžēlīgs	ncmsn1/tēvs	zcl/, q/nru	pp10snns/evs

(4.2) Common tagset: Universal Dependencies

patēlās	Dievam	,	grozās	,	sirdās	par	zenķiem	,	kas	splauj	sa
VERB/root	NOUN/obj	PUNCT/punct	VERB/root	PUNCT/punct	VERB/root	ADP/case	NOUN/obj	PUNCT/punct	PRON/nsbj	VERB/act	NOUN/obj
ozdams			kas	gadījās	sirdīgi	pārmēta	:	–	Tu		
ERB/act	PUNCT/punct	PRON/nsbj	VERB/act	PUNCT/punct	ADV/advmood	VERB/root	PUNCT/punct	PUNCT/punct	PRON/vocative	PUNCT/punct	
.	Kāpēc	notiek	sirdsdarbības	apstāšanās	Sirds	darbojas	,	pateicoties	tam	,	
conj	ADV/advmood	VERB/root	NOUN/nmod	NOUN/nsbj	NOUN/nsbj	VERB/root	PUNCT/punct	VERB/advcl	PRON/obj	PUNCT/punct	
pašam	ilgi	bija	apslēpts	?	Sirdsbalss	gan	sacīja	priekšā	,	bet	
PRON/obj	ADV/advmood	VERB/aux	VERB/ccomp	PUNCT/punct	NOUN/nsbj	CCONJ/cc	VERB/root	ADV/advmood	PUNCT/punct	CCONJ/cc	
icēnsībās			kuras	ir	aktiera	sirdslīdētā					
NOUN/obl	PUNCT/punct	PRON/nsbj	AUX/act	NOUN/nmod	NOUN/nmod	NOUN/nsbj					
"	Ar	labu	nakti		sirdsmīļā	"	)		–	vai	
CT/punct	ADP/case	ADJ/amod	NOUN/discourse	PUNCT/punct	ADJ/conj	PUNCT/punct	PUNCT/punct	PUNCT/punct	PUNCT/punct	CCONJ/cc	
net	–	Visuspēctīgais	Dievs		sirdsžēlīgais	Tēvs		Nu	es	zinu	
conj	PUNCT/punct	ADJ/amod	NOUN/nsbj	PUNCT/punct	ADJ/amod	NOUN/act	PUNCT/punct	PRON/nsbj	PRON/nsbj	VERB/root	

Benefits:

- An open-ended national reference corpus
- Easily accessible otherwise separate and diverse corpora
- Quick quantitative comparison across the whole corpora collection

Future tasks:

- Full-fledged Latvian speech corpora in LNCC
- A new, extended (100M tokens) version of LVK
- Other endpoints and selected corpora
- 10B token corpora collection