Latvian National Corpora Collection – Korpuss.lv

Baiba Saulīte, Roberts Darģis, 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

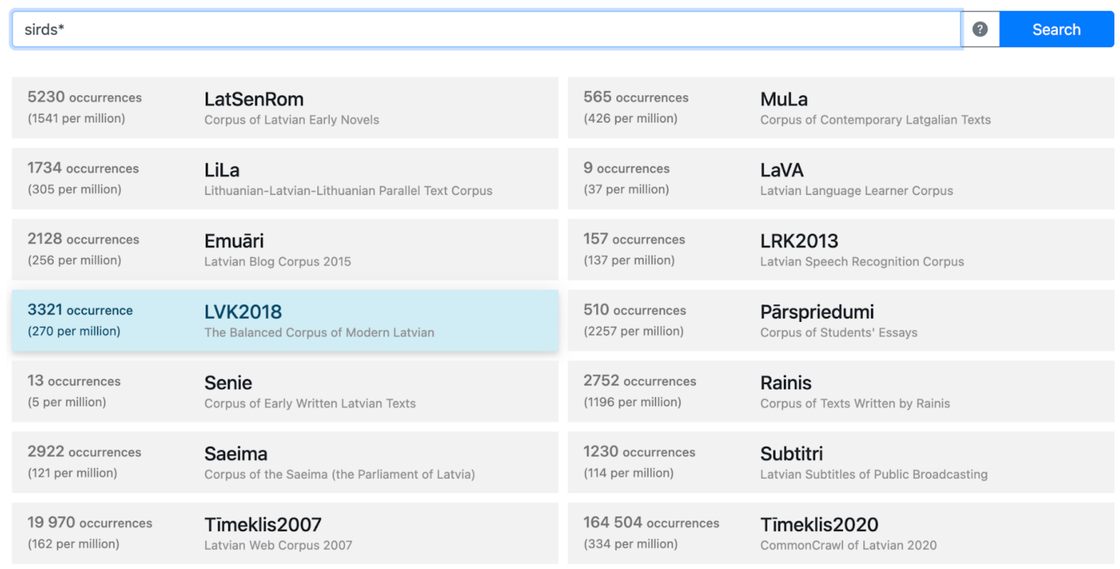
Abstract. Latvian National Corpora Collection (LNCC) is a diverse collection of corpora representing both written and spoken language. LNCC covers various use cases and all the important text types and genres. It is a continuous multi-institutional and multi-project effort, supported by the Digital Humanities 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
Written language text corpora				
LVK2018	Balanced Corpus of Modern Latvian (Levāne-Petrova and Darģis, 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
Tīmeklis2007	Latvian Web Corpus (Džeriņš and Džonsons, 2007)	123.5M tokens	text, web	2006–2007
Tīmeklis2020	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
Spoken language text corpora				
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**

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



(4.1) **Common tagset: morphology**

pateicas	Dievam	,	grozās	,	**sirdās**	par	zeņķiem	,	kas	spļauj
0an/pateikties	ncmsd1/dievs	zc/,	vmyip_330an/grozīties	zc/,	vmyip_330an/sirdīties	sppd/par	ncmpd2/zeņķis	zc/,	pr000nn/kas	vmnipi130an/spļaut
uzkozdams	,	kas	gadījās	,	**sirdīgi**	pārmeta	:	–	Tu	,	Mairita
ısn0000n/uzkost	zc/,	pr000nn/kas	vmyis_330an/gadīties	zc/,	rp_/sirdīgi	vmnist130an/pārmest	zo/:	zd/-	pp20snn/tu	zc/,	npfsn4/Mairita
ēc	notiek	sirdsdarbības	apstāšanās		**Sirds**	darbojas	,	pateicoties	tam
ēc	vmnipi130an/notikt	ncfsg4/sirdsdarbība	ncfsgr/apstāšanās		ncfsn6/sirds	vmyip_230an/darboties	zc/,	vmypu0000000n/pateikties	pd3msd
šam	ilgi	bija	apslēpts	?	**Sirdsbalss**	gan	sacīja	priekšā	,	bet	Dāvis
ı/pats	rp_/ilgi	vmnis_i30an/būt	vmnpdmsnpsnpn/apslēpt	zs/?	ncfsn6/sirdsbalss	cc/gan	vmnist330an/sacīt	r0_/priekšā	zc/,	cc/bet	npmsn2/Dāv
ensībās	,	kuras	ir	aktiera	**sirdslieta**	.	Patriks	piekritis	sievas	not
‘sacensība	zc/,	pr0fpnn/kura	vcnipii30an/būt	ncmsg2/aktieris	ncfsn4/sirdslieta	zs/.	npmsn1/Patriks	vmnpdmsnasnpn/piekrist	ncfsg4/sieva	ncn
eet-heart	(	„	Ar	labu	nakti	,	**sirdsmīļā**	”	)	,	–	vai	šoreiz	mājās	vajadzēs
‘weet-heart	zb/(	zq/"	spsa/ar	affsanp/labs	ncfsa6/nakts	zc/,	affsnyp/sirdsmīļš	zq/"	zb/)	zc/,	zd/-	cc/vai	r0_/šoreiz	ncfpl4/māja	vmnift330an/
acis	.	—	Visuspēcīgais	Dievs	,	**sirdszēlīgais**	Tēvs	,	Nu	es	zinu	,	ka
ıcfpa6/acs	zs/.	zd/-	afmsnyp/visuspēcīgs	ncmsn1/dievs	zc/,	afmsny_/sirdszēlīgs	ncmsn1/tēvs	zc/,	q/nu	pp10snn/es	vmnipt31san/zināt	zc/,	cs/

(4.2) **Common tagset: Universal Dependencies**

pateicas	Dievam	,	grozās	,	**sirdās**	par	zeņķiem	,	kas	spļauj	sau
VERB/root	NOUN/iobj	PUNCT/punct	VERB/root	PUNCT/punct	VERB/root	ADP/case	NOUN/iobj	PUNCT/punct	PRON/nsubj	VERB/acl	NOU
ozdams	,	kas	gadījās	,	**sirdīgi**	pārmeta	:	–	Tu	,
ERB/acl	PUNCT/punct	PRON/nsubj	VERB/acl	PUNCT/punct	ADV/advmod	VERB/root	PUNCT/punct	PUNCT/punct	PRON/vocative	PUNCT/
.	Kāpēc	notiek	sirdsdarbības	apstāšanās	**Sirds**	darbojas	,	pateicoties	tam	,
unct	ADV/advmod	VERB/root	NOUN/nmod	NOUN/nsubj	NOUN/nsubj	VERB/root	PUNCT/punct	VERB/advcl	PRON/iobj	PUNCT/punc
pašam	ilgi	bija	apslēpts	?	**Sirdsbalss**	gan	sacīja	priekšā	,	bet	D
RON/iobj	ADV/advmod	VERB/aux	VERB/ccomp	PUNCT/punct	NOUN/nsubj	CCONJ/cc	VERB/root	ADV/advmod	PUNCT/punct	CCONJ/cc	P
ıcensībās	,	kuras	ir	aktiera	**sirdslieta**	.	Patriks	piekritis	sievas	note
NOUN/obl	PUNCT/punct	PRON/nsubj	AUX/acl	NOUN/nmod	NOUN/nsubj	PUNCT/punct	PROPN/nsubj:pass	VERB/root	NOUN/nmod	NC
„	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	CCON
.	—	Visuspēcīgais	Dievs	,	**sirdszēlīgais**	Tēvs	,	Nu	es	zinu
nct	PUNCT/root	ADJ/amod	NOUN/nsubj	PUNCT/punct	ADJ/amod	NOUN/acl	PUNCT/punct	PART/discourse	PRON/nsubj	VERB/roc

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

http://www.korpuss.lv

This work has received financial support from the State Research Programmes under grant agreements No. VPP-IZM-DH-2020/1-0001 and No. VPP-LETONIKA-2021/1-0006, from the Latvian Language Agency under grant agreement No. 4.6/2019-029, and from ERDF under the grant agreement No. 1.1.1.1/18/A/153.