CLARIN-LV: Many Steps till Operation

Inguna Skadiņa, Ilze Auziņa, Roberts Darģis,
Eduards Lasmanis and Arnis Voitkāns

CLARIN

CLARIN-LV

NATIONAL DEVELOPMENT PLAN 2020

EUROPEAN UNION
European Regional Development Fund

INVESTING IN YOUR FUTURE

CLARIN ERIC

- Preparatory phase (2008-2010): puting everything in place

- Construction phase: (2011-2015): building and populating with tools and resources

- Exploitation phase (2016 -):

CLARIN in full service

CLARIN-LV

2008-2011: Preparatory Phase. MoU signed.

2016: Latvia joins CLARIN ERIC

2018-2022: Construction phase

Outline: Many Steps till Operation

- Centre and Repository
  - building
  - filling
  - certification
- Consortium
  - building
  - collaboration
  - strengthening collaboration with digital humanities community
- Teaching and training

CLARIN-LV

CLARIN-LV | About | Participants | Repository | News | Publications | Contacts

About

CLARIN stands for "Common Language Resources and Technology Infrastructure". It is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences.

CLARIN K CENTRE

In 2012 CLARIN ERIC was established and took up the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. Currently CLARIN provides easy and sustainable access to digital language data (in written, spoken, or multimodal form) for scholars in the social sciences and humanities, and beyond. CLARIN also offers advanced tools to discover, explore, exploit, annotate, analyse or combine such data sets, wherever they are located. This is enabled through a networked federation of centres: language data repositories, service centres and knowledge centres, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work.

The CLARIN infrastructure is fully operational in many countries, and a large number of participating centres are offering access services to data, tools and expertise. At the same time, CLARIN continues to be constructed in some countries that joined more recently, and CLARIN's datasets and services are constantly updated and improved. On the services page we show the services accessible at this moment and we explain how and by whom the various services can be accessed.

CLARIN-LV organizes the 10th Conference Human Language Technologies - the Baltic Perspective

The Tenth International Conference "Human Language Technologies - the Baltic Perspective" will take place on October 6-7 at the University of Latvia. The aim of the conference is to provide a forum for the sharing of new ideas and recent advances in human language processing and to promote cooperation between the research communities of computer science, linguistics and digital humanities from the Baltic states and the rest of the world.

Conference program.

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020
EIROPAS SAVIENĪBA
Eiropas Reģionālās attīstības fonds
IEGULDĪJUMS TAVĀ NĀKOTNĒ

Activities of CLARIN Latvia are supported by project of the European Regional Development Fund Nr. 1.1.1.5/18/I/016 University of Latvia and institutes in the European Research Area - Excellency, activity, mobility, capacity

www.clarin.lv
info@clarin.lv

© 2022 CLARIN Latvia

Repository

CLARIN DSpace

- After analysis of different options used by other CLARIN centres, CLARIN Dspace (formerly known as LINDAT Dspace) repository system was chosen.
- The main reasons for this choice were:
  - availability of installation resources and documentation
  - rather easy installation and documented
  - approval by many CLARIN centers
  - reliability
  - technical support

www.clarin.lv
info@clarin.lv

Core Trust Seal and B-centre certification

[Source: CoreTrustSeal Requirements 2020-22]

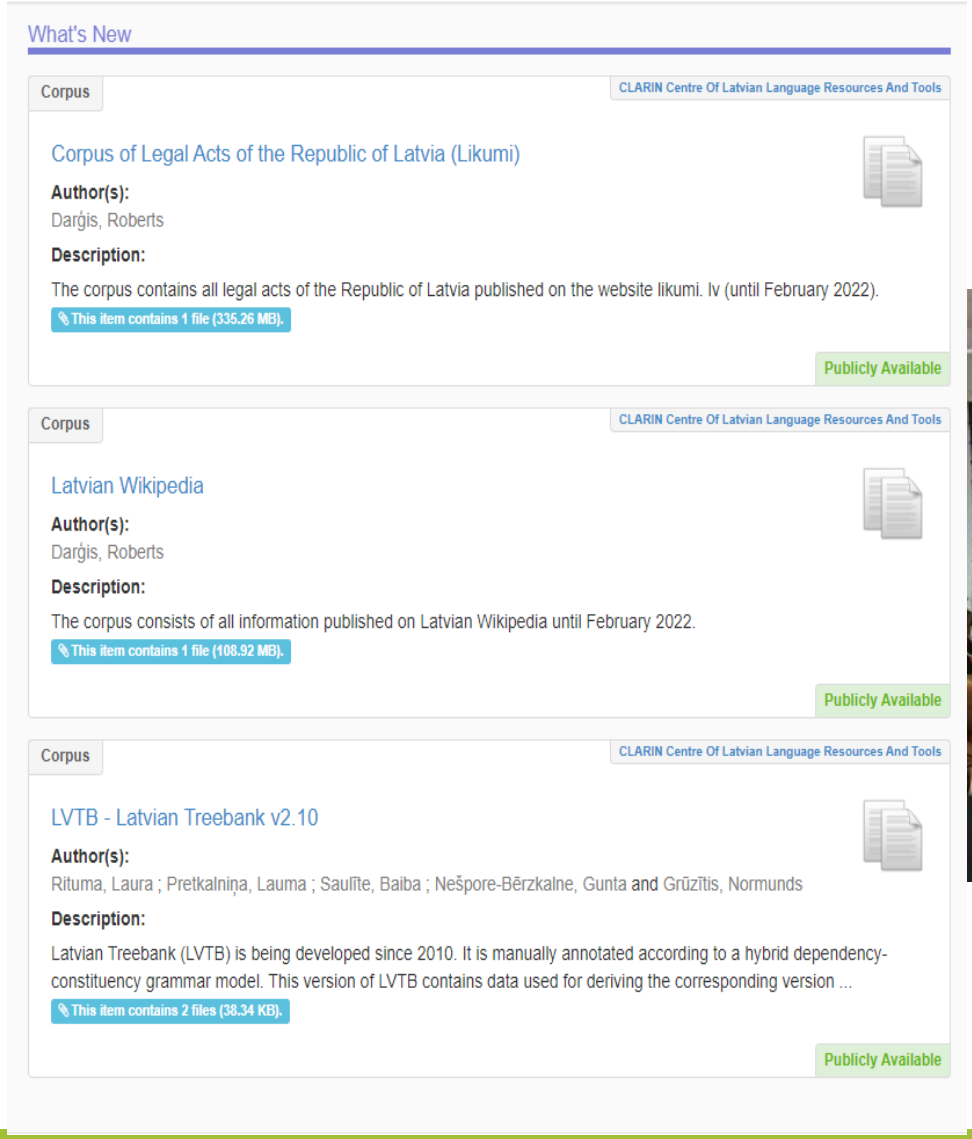
CoreTrustSeal self-assessment template

The CoreTrustSeal self-assessment includes 16 requirements (R1–R16) plus R0 providing contextual information on the repository. This template contains all requirements along with guidance for formulating self-assessment statements and gathering related evidence. Along with the self-assessment statements, the applicant must indicate a compliance level for R1–R16:

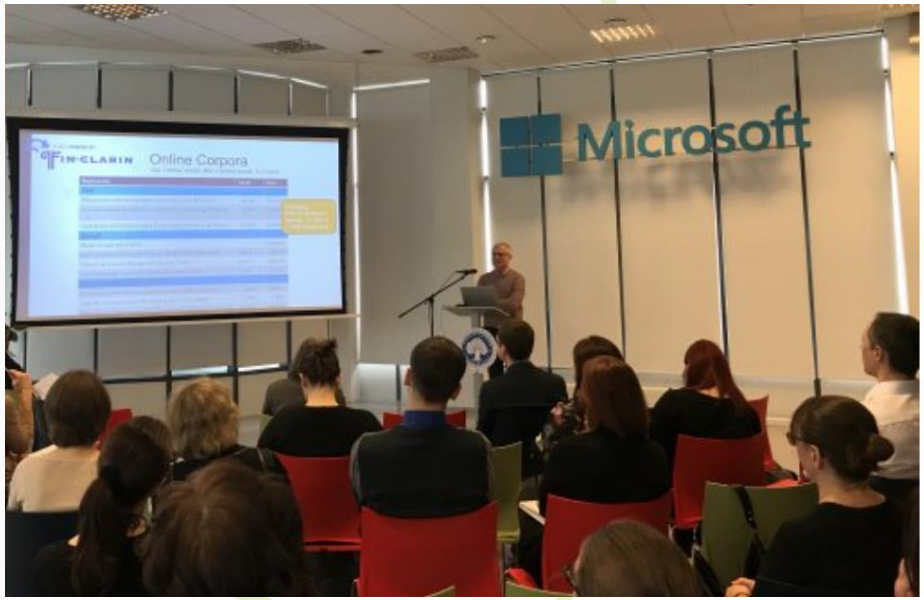
0. Not applicable
1. The repository has not considered this yet
2. The repository has a theoretical concept
3. The repository is in the implementation phase
4. The guideline has been fully implemented in the repository

The CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022 document contains more information on e.g. missing evidence, understandability, non-English language documentation, sensitive/internal documentation, structure and length of responses, and further guidance for certain requirements. Please read it before starting!

CLARIN-LV repository

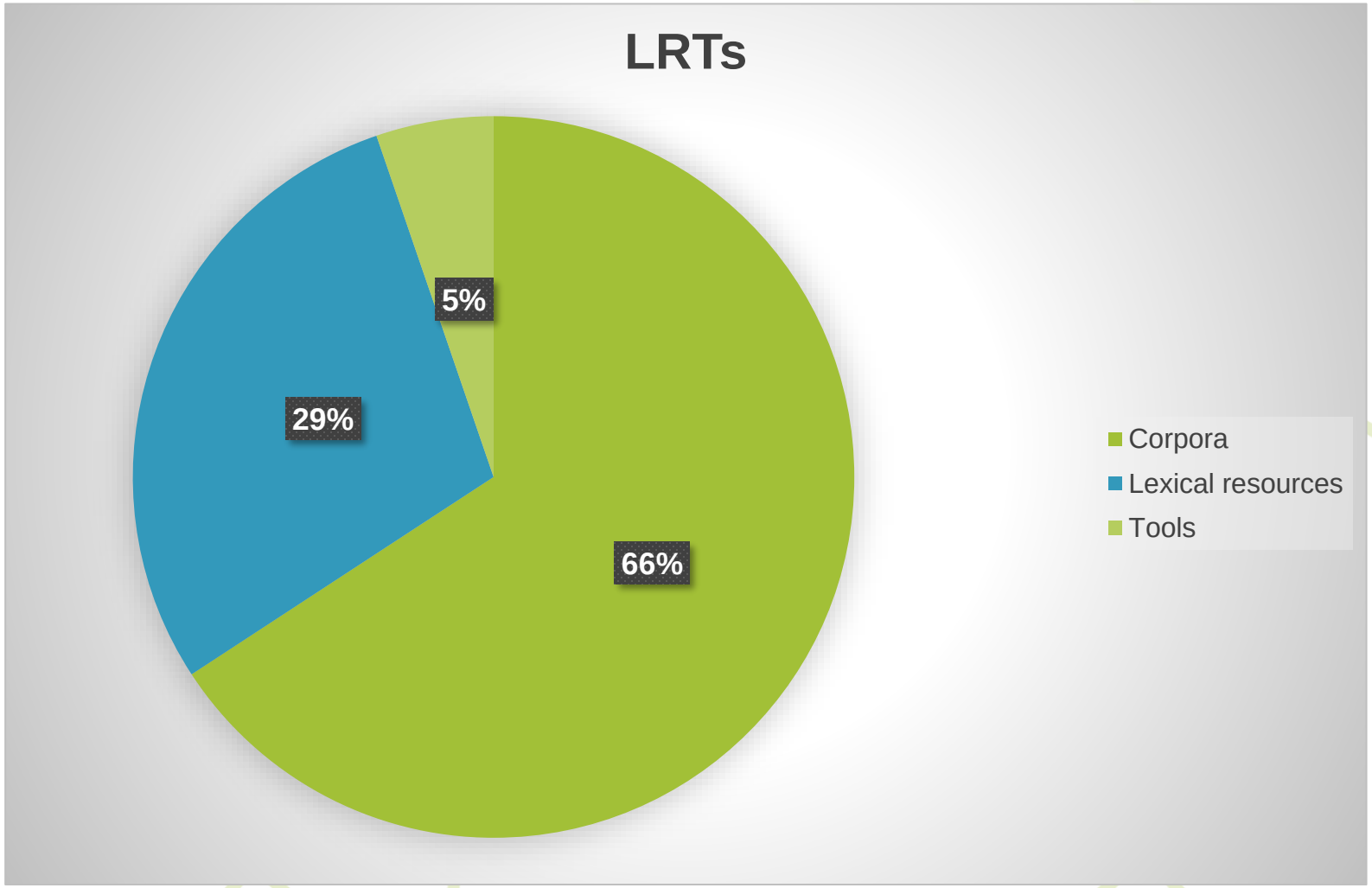


The repository was set up in March, 2020

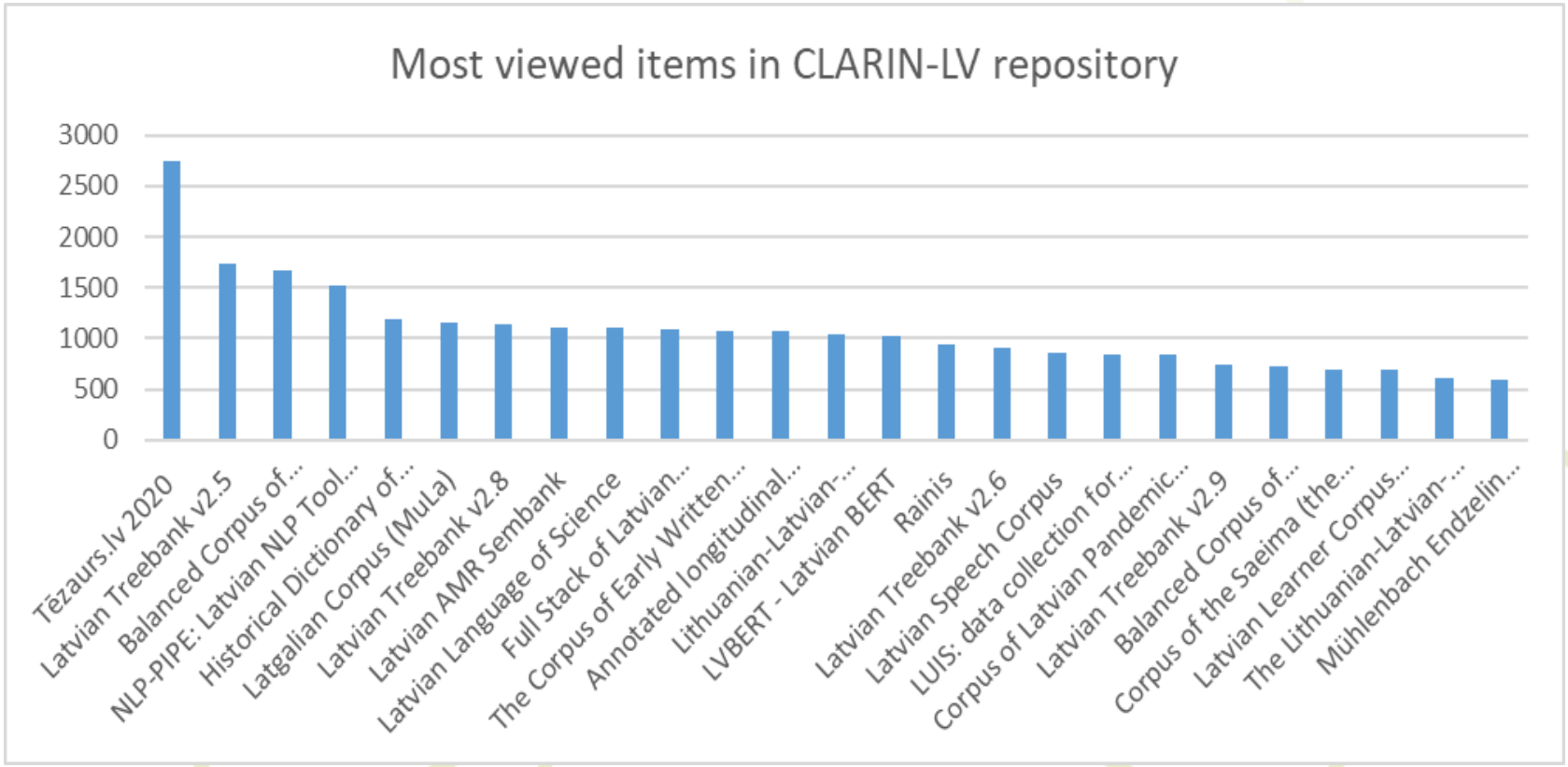


repository.clarin.lv

Language Resources and Tools in CLARIN-LV



Top 25 most viewed language resources and tools in CLARIN-LV repository



Corpora

general (6) learner (2) literary (4) parliamentary (1) diahronic (1) specialised (12) representative (7) parallel (1) text (24)
speech (5) morphology (22) syntax (3) semantics (1) error annotation (2) manually annotated (4) web (2)

LVK2018
The Balanced Corpus of Modern Latvian
2016–2018, 10M words (12M tokens)
Developers: IMCS UL
More Search

MuLa2022
Corpus of Contemporary Latgalian Texts 2022
2020–2022, 2M words (2.8M tokens)
Developers: RAT, IMCS UL
More Search

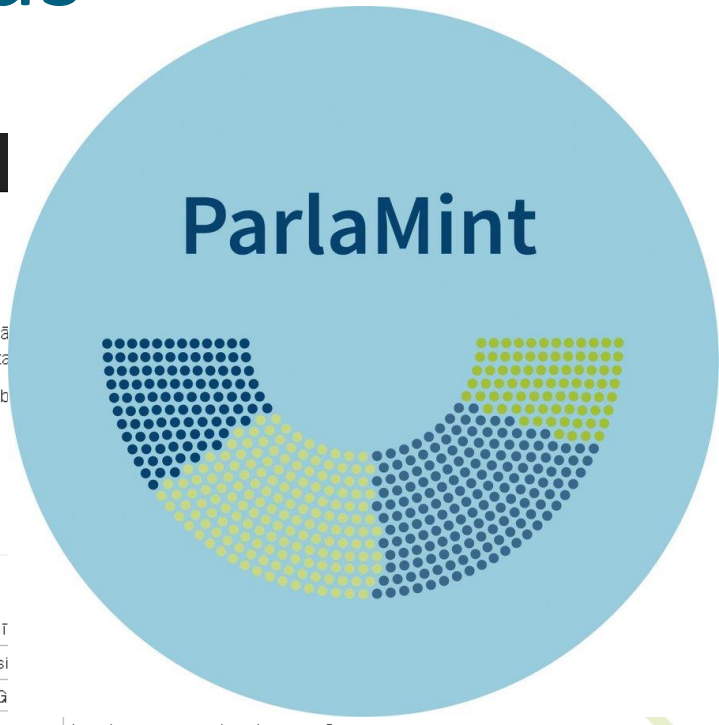
LaVA
Latvian Language Learner Corpus
2018–2021, 192k words (241k tokens)
Developers: IMCS UL
More Search

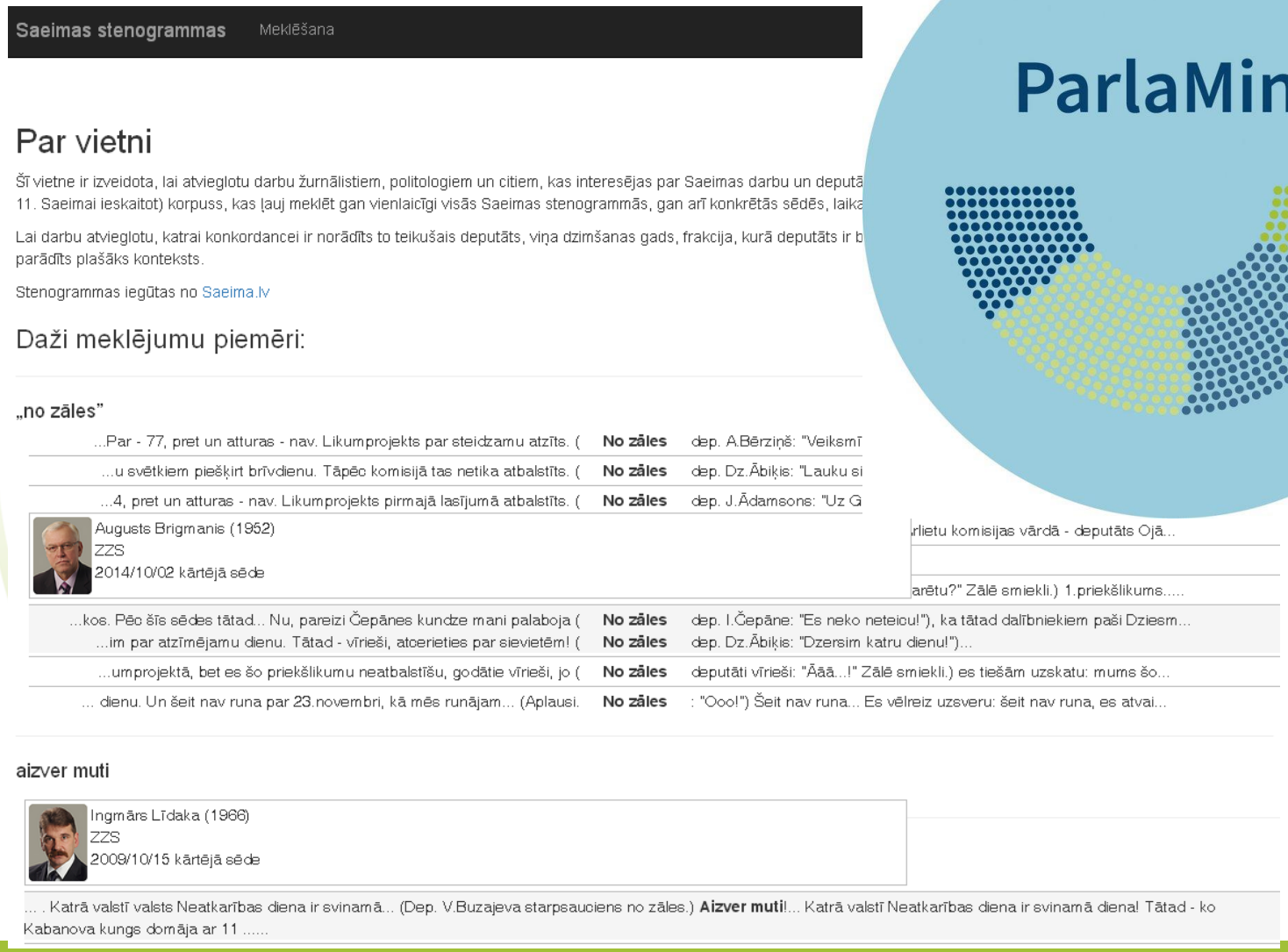
LVTB
Latvian Treebank
2010-2022, 16803 sentences (282 167 tokens)
Developers: IMCS UL
More Search

Barometrs
Corpus of News Portal Comments
2011–2022, 26M comments (642M tokens)
Developers: RSU, IMCS UL
More Search

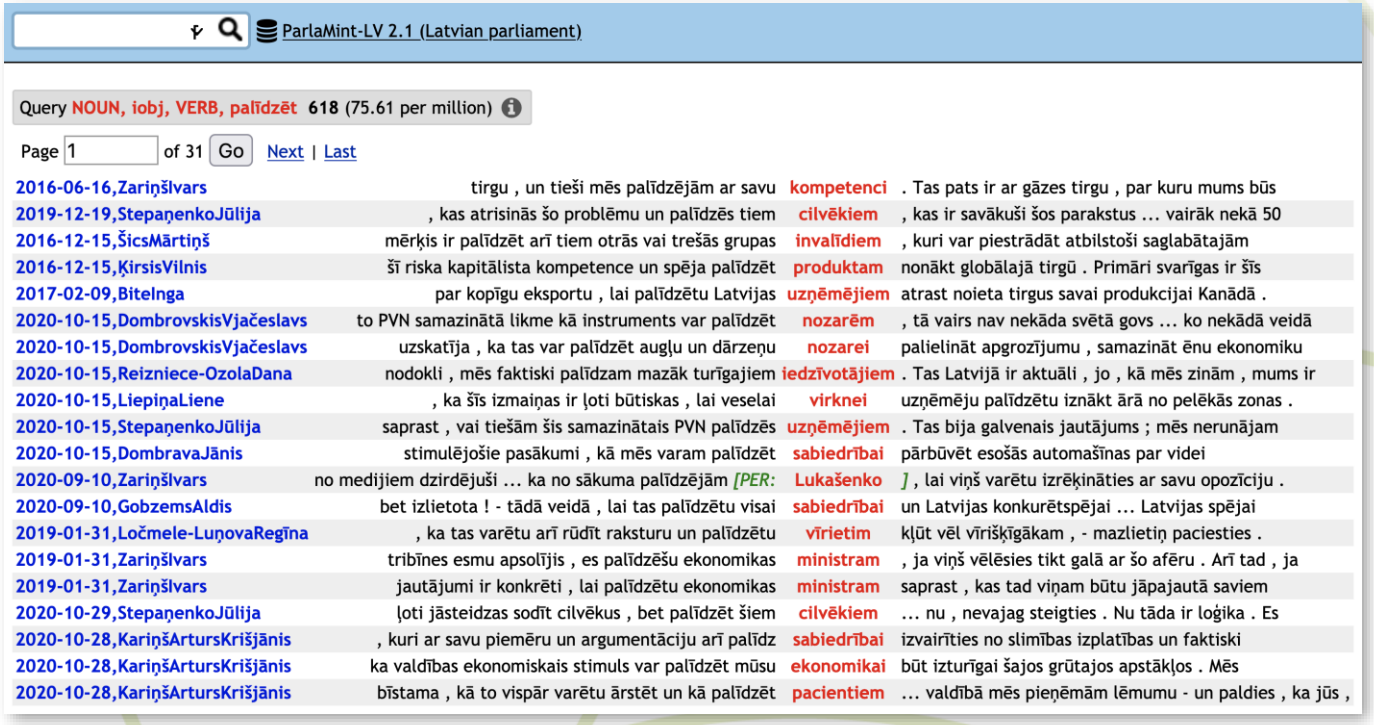
Disertācijas
Corpus of Latvian PhD Theses
2022, 16.7M words (23.4M tokens)
Developers: IMCS UL
More Search

Parliament Data Corpus





ParlaMint: UD-annotated comparable corpora



CQL query:

```
[pos="NOUN"
 & dep="iobj"
 &
dep_head_pos="VERB"
 & dep_head_lemma=

"palīdzēt"]
```


Erjavec, T., Ogrodniczuk, M., ..., Darģis, R., ..., Krilavičius, T., et al. ParlaMint: Comparable Corpora of European Parliamentary Data. Proceedings of the CLARIN Annual Conference, 2021

https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

Lexical Resources

LexicalConceptualResource — CLARIN Centre Of Latvian Language Resources And Tools

Historical Dictionary of Latvian

(Latvian Language Institute of the University of Latvia / 2016-12-05)

Author(s):

Andronova, Everita ; Frīdenberga, Anna ; Siliņa-Piņķe, Renāte ; Trumpa, Anta and Vanags, Pēteris

This item contains no files.

LexicalConceptualResource — CLARIN Centre Of Latvian Language Resources And Tools

Dictionary of Contemporary Latvian Language (MLVV)

(Latvian Language Institute of the University of Latvia; AiLab IMCS UL / 2019)

Author(s):

Jērāne, Santa ; et al.

▶ show everyone

This item contains 1 file (2.68 MB).

Publicly Available

LexicalConceptualResource — CLARIN Centre Of Latvian Language Resources And Tools

Language consultations: electronic database

(Latvian Language Agency / 2017)

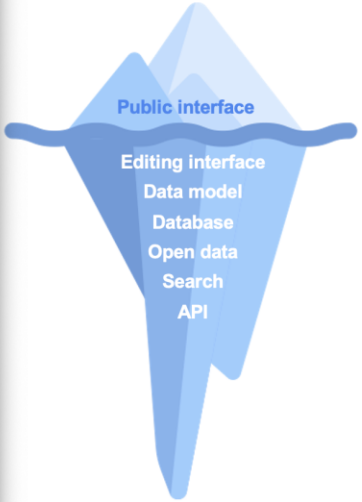
Author(s):

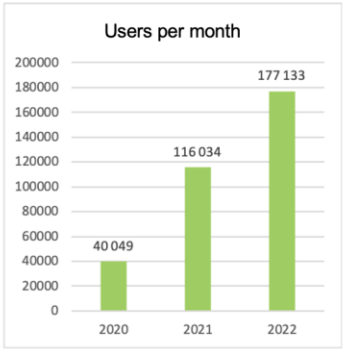
---, ---

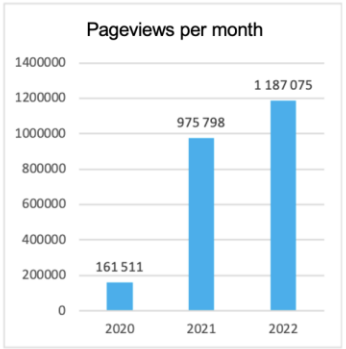
This item contains no files.

LexicalConceptualResource — CLARIN Centre Of Latvian Language Resources And Tools

The Tēzaurs.lv platform

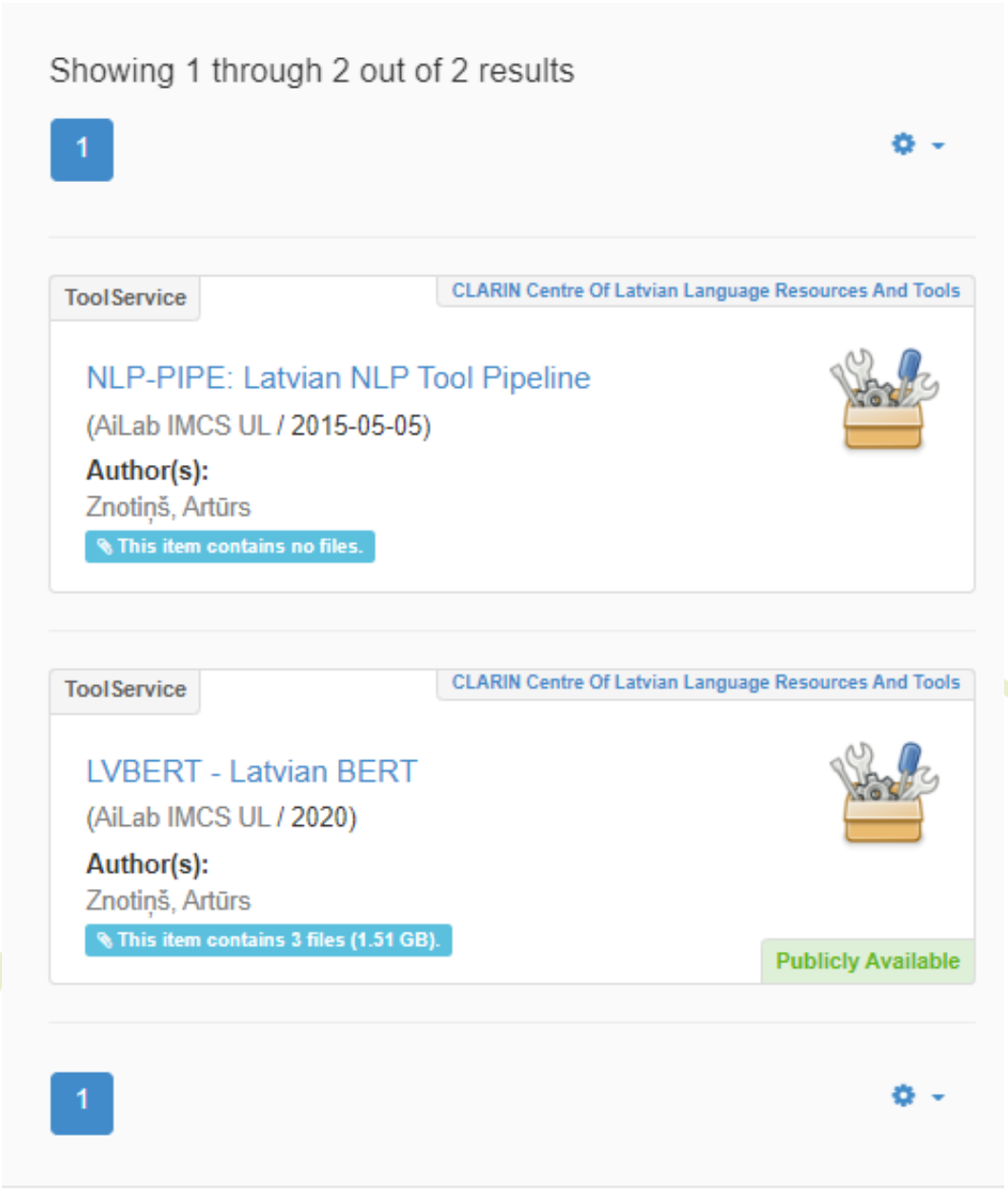




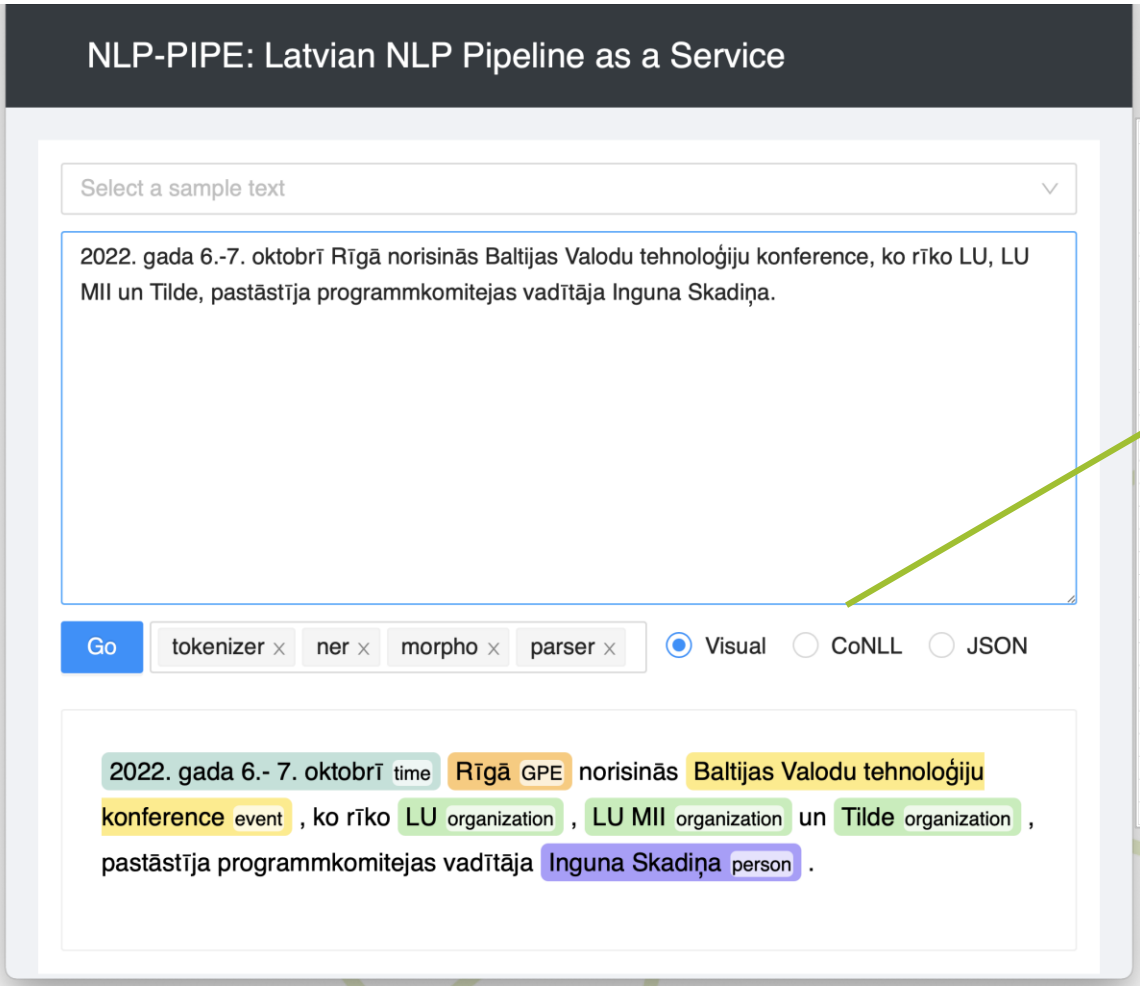


https://tezaurs.lv

Tools


Showing 1 through 2 out of 2 results
1
Tool Service
CLARIN Centre Of Latvian Language Resources And Tools
NLP-PIPE: Latvian NLP Tool Pipeline
(AiLab IMCS UL / 2015-05-05)
Author(s):
Znotiņš, Artūrs
This item contains no files.
Tool Service
CLARIN Centre Of Latvian Language Resources And Tools
LVBERT - Latvian BERT
(AiLab IMCS UL / 2020)
Author(s):
Znotiņš, Artūrs
This item contains 3 files (1.51 GB).
Publicly Available
1

Toolchain for Latvian Language Processing: NLP-PIPE



Used by: Korpuss.lv, National Library of Latvia, News agency LETA, IT students, etc.

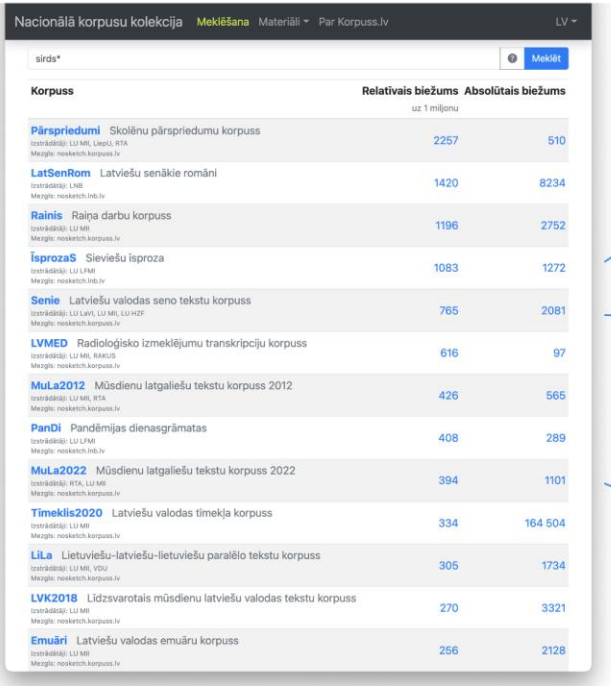
INDEX	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
#text=2022. gada 6.- 7. oktobrī Rīgā norisinās Baltijas Valodu tehnoloģiju konference , ko rīko LU , LU MII un Tilde , pastāstīja programmkomitejas vadītāja Inguna Skadiņa .							
1	2022.	2022.	ADJ	xo	NumType=Ord	2	amod
2	gada	gads	NOUN	ncmsg1	Case=GenlGe	3	nmod
3	6.-	6.-	NUM	xn	NumType=Car	5	amod
4	7.	7.	ADJ	xo	NumType=Ord	5	amod
5	oktobrī	oktobris	NOUN	ncmsl2	Case=LoclGen	7	obl
6	Rīgā	Rīga	PROPN	npfsl4	Case=LoclGen	7	obl
7	norisinās	norisināties	VERB	vmyip_330an	Evident=FhlMc	22	ccomp
8	Baltijas	Baltija	PROPN	npfsg4	Case=GenlGe	9	nmod
9	Valodu	valoda	NOUN	ncfpg4	Case=GenlGe	10	nmod
10	tehnoloģiju	tehnoloģija	NOUN	ncfpg4	Case=GenlGe	11	nmod
11	konference	konference	NOUN	ncfsn5	Case=NomlGe	7	nsubj
12	,	,	PUNCT	zc	_	14	punct
13	ko	kas	PRON	pr000an	Case=AcclPro	14	obj
14	rīko	rīkot	VERB	vmnipt230an	Evident=FhlMc	11	acl
15	LU	LU	PROPN	yp	Abbr=Yes	14	nsubj
16	,	,	PUNCT	zc	_	17	punct
17	LU	LU	PROPN	yp	Abbr=Yes	15	conj
18	MII	mii	X	y	Abbr=Yes	17	iobj
19	un	un	CCONJ	cc	_	20	cc
20	Tilde	tilde	NOUN	ncfsn5	Case=NomlGe	18	conj
21	,	,	PUNCT	zc	_	7	punct
22	pastāstīja	pastāstīt	VERB	vmnist330an	Evident=FhlMc	0	root
23	programmkomitejas	programmkomiteja	NOUN	ncfsg4	Case=GenlGe	24	nmod
24	vadītāja	vadītāja	NOUN	ncfsn4	Case=NomlGe	26	nmod
25	Inguna	Inguns	PROPN	npmsg1	Case=GenlGe	26	nmod
26	Skadiņa	Skadiņa	NOUN	ncfsn4	Case=NomlGe	22	nsubj
27	.	.	PUNCT	zs	_	22	punct

http://nlp.ailab.lv

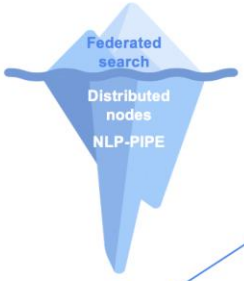
Consortium and Collaboration with Digital Humanities

Collaboration with digital humanities

The Korpuss.lv platform



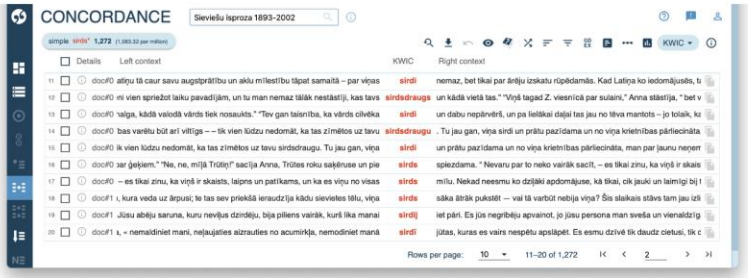
20+ corpora, 10+ institutions, 3+ nodes

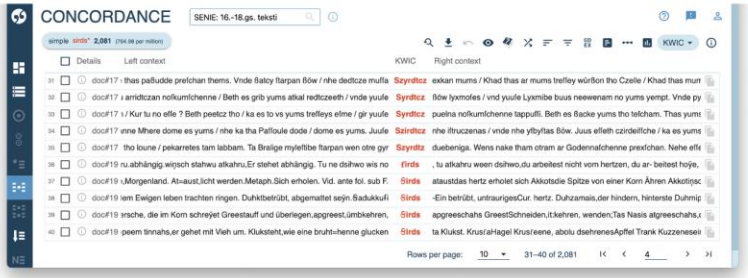


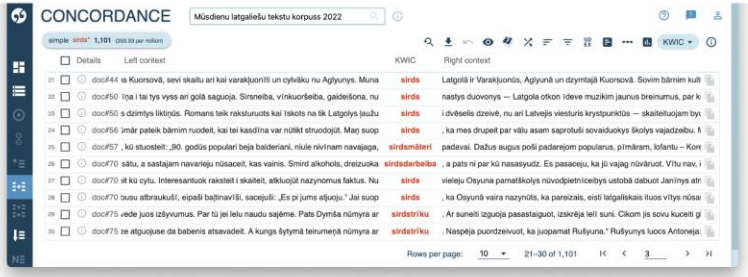
Developer: LU LFMI
Node: nosketch.lnb.lv

Developer: LU LaVI, LU MII, LU HZF
Node: nosketch.korpuss.lv

Developer: RTA, LU MII
Node: nosketch.korpuss.lv







The project “Digital Resources for Humanities: Integration and Development” (VPP-IZM-DH-2020/1-0001)

Consortium (January, 2022)

Institute of Mathematics and Computer Science
University of Latvia

User involvement

- The conference "Digital Latvian Language Resources and Tools in Common Research Infrastructure"
- The seminar “Tools and Resources for Digital Humanities Research”
- The practical seminar “Balanced Corpus of Modern Latvian “
- The practical seminar “Balanced Corpus of Modern Latvian “
- The practical seminar “ How to use Latvian treebank in linguistic studies“
- The conference «Latvian language resources and tools for education»
- The conference «Human Language Technologies - the Baltic Perspective»

Workshop “How to Register Language Resources in the CLARIN Repository” (January, 2021)

Item submission

The Lithuanian-Latvian-Latgalian Dictionary

Project URL	http://hipilatlit.ru.lv/eng/products/dictionary_info.html
Demo URL	http://hipilatlit.ru.lv/dictionary/eng/dictionary.html
Date issued	2013
Type	lexicalConceptualResource
Publisher	RTA

Show full item record

Resume Cancel Share submission

SAFMORIL Knowledge Centre

- SAFMORIL - CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages
- Collaboration between researchers and developers from the University of Helsinki, University of Tromsø, Vytautas Magnus University and the Institute of Mathematics and Computer Science (University of Latvia) in the area of computational morphology and its application to language processing.

Teaching activities

Teaching

Today most of teaching activities concentrate in **the University of Latvia** (Faculty of Humanities, Faculty of Education, Psychology and Art, and Faculty of Social Science)

- In 2017 new *Computational Linguistics* course for Master Students in English philology was started
- In Autumn, 2020 *Computational Linguistics* course started for Master Students in Baltic Philology
- In Autumn, 2020 *Digital Humanities and Social Sciences* for Master students in several study programmes
- In Autumn, 2022 *Language Corpora in the Educational Process* for the professional program for teachers
- Several corpus linguistics courses for Master and Doctoral students in English philology

CLARIN in the Classroom

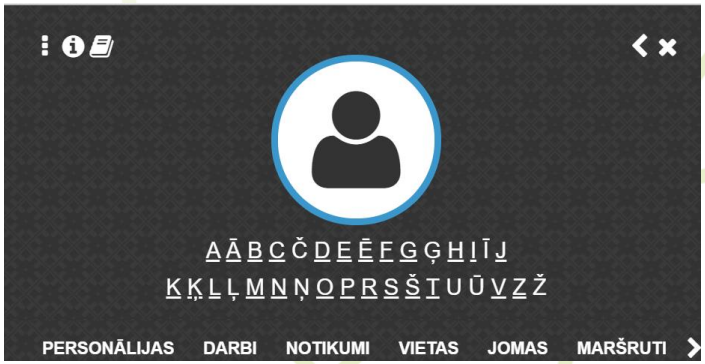
- In lecture we introduce students with CLARIN research infrastructure, Virtual Language Observatory and resource families.
- We also highlight some LRTs that are more relevant to the research interests of students for particular year of studies.
- In the seminar students present their findings – tool or resource they found in CLARIN VLO and that seemed interesting for her/him.
- We ask students explore other resources than simple corpus

CLARIN-LV

Student’s findings (Autumn, 2022)

CLARIN-D

ASV Online Toolbox



ELAN - Linguistic Annotator

CLARIN-LV

Conclussion and next steps

Being among the smallest countries of CLARIN infrastructure and joining CLARIN only in 2016, has some positive and some negative consequences:

- On one hand, many administrative and technical issues were solved when we joined, thus we can use reliable and stable solutions developed by the consortium.
- On the other hand, being among the newest members of CLARIN, clearly revealed a gap in needs, understanding and readiness to serve the SSH community.

Conclussion and next steps

While the targeted project for setting up CLARIN-LV node will end in December, 2022, CLARIN-LV plans to continue activities through several language technology related projects:

- Recently started project “Research on Modern Latvian Language and Development of Language Technology”' of the State Research Programme “Letonika – Fostering a Latvian and European Society'' will continue supporting cooperation with digital humanities researchers, the development of new resources and their subsequent inclusion in the CLARIN-LV repository.
- Additional funding is also expected from the Latvia's Recovery and sustainability plan, which envisions activities for the development and implementation of high-level skills in language technology.
- New project proposal submitted for the the National Research Programme "Digital Resources of the Humanities"

Thank you!

CLARIN Activities of CLARIN Latvia are supported by project of European Structural funds Nr. 1.1.1.5/18/I/016.