

10 Years of Support for Lithuanian Language Research:

An Overview of the CLARIN-LT Repository

7th International Conference of Applied Linguistics
Workshop III

Andrius Utkas

Institute of Digital Resources and
Interdisciplinary Research (SITTI)

Vytautas Magnus University

CLARIN





Plan

- ✓ What is CLARIN ERIC and CLARIN-LT
- ✓ Timeline: 10 Years of Support for Lithuanian Language Research
- ✓ Overview of repository and repository contents
- ✓ How to deposit a resource





<https://clarin.eu>

Research
Infrastructure

Research

Europe

Resources

Language

CLARIN

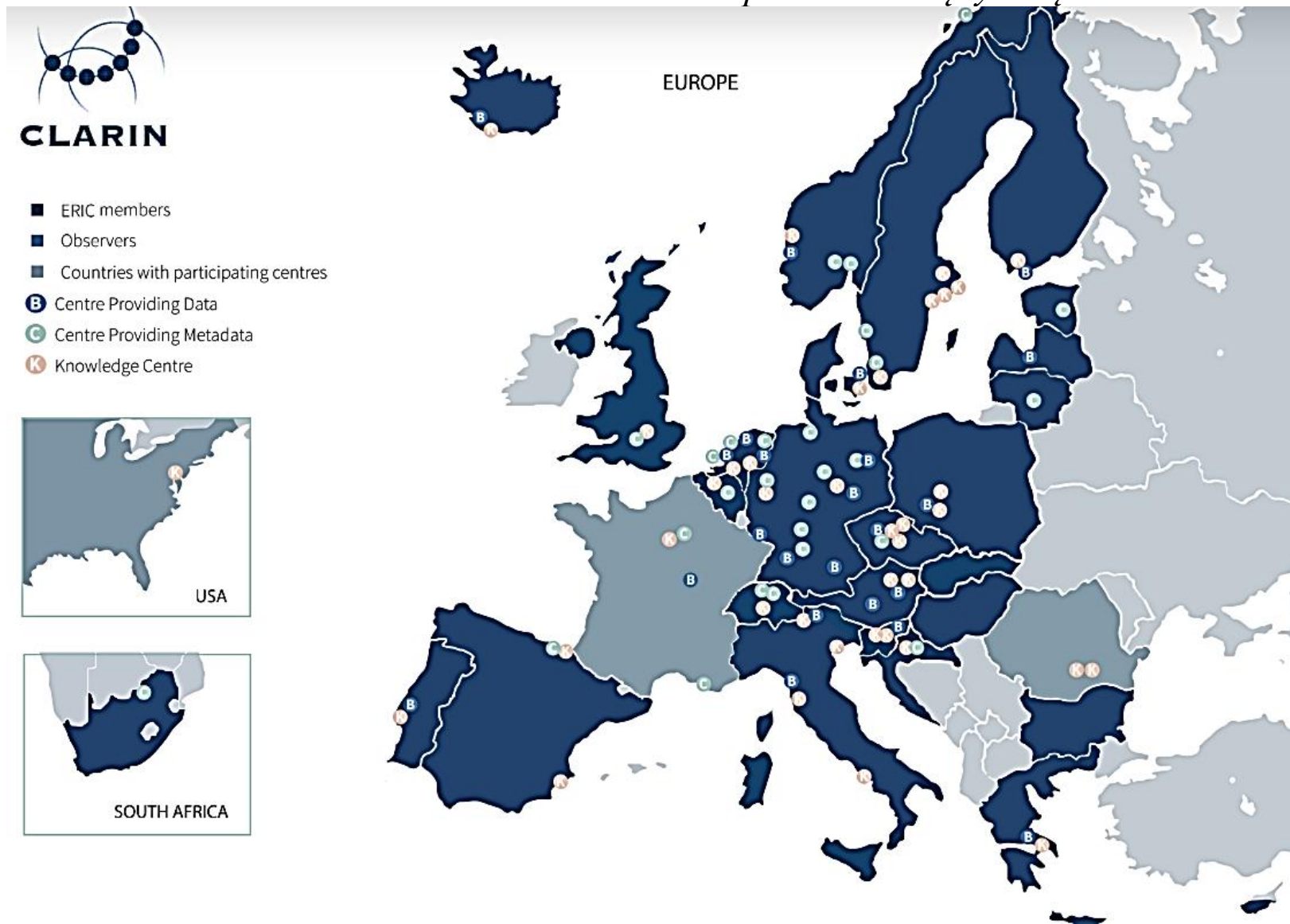


Common Language Resources and Technology Infrastructure

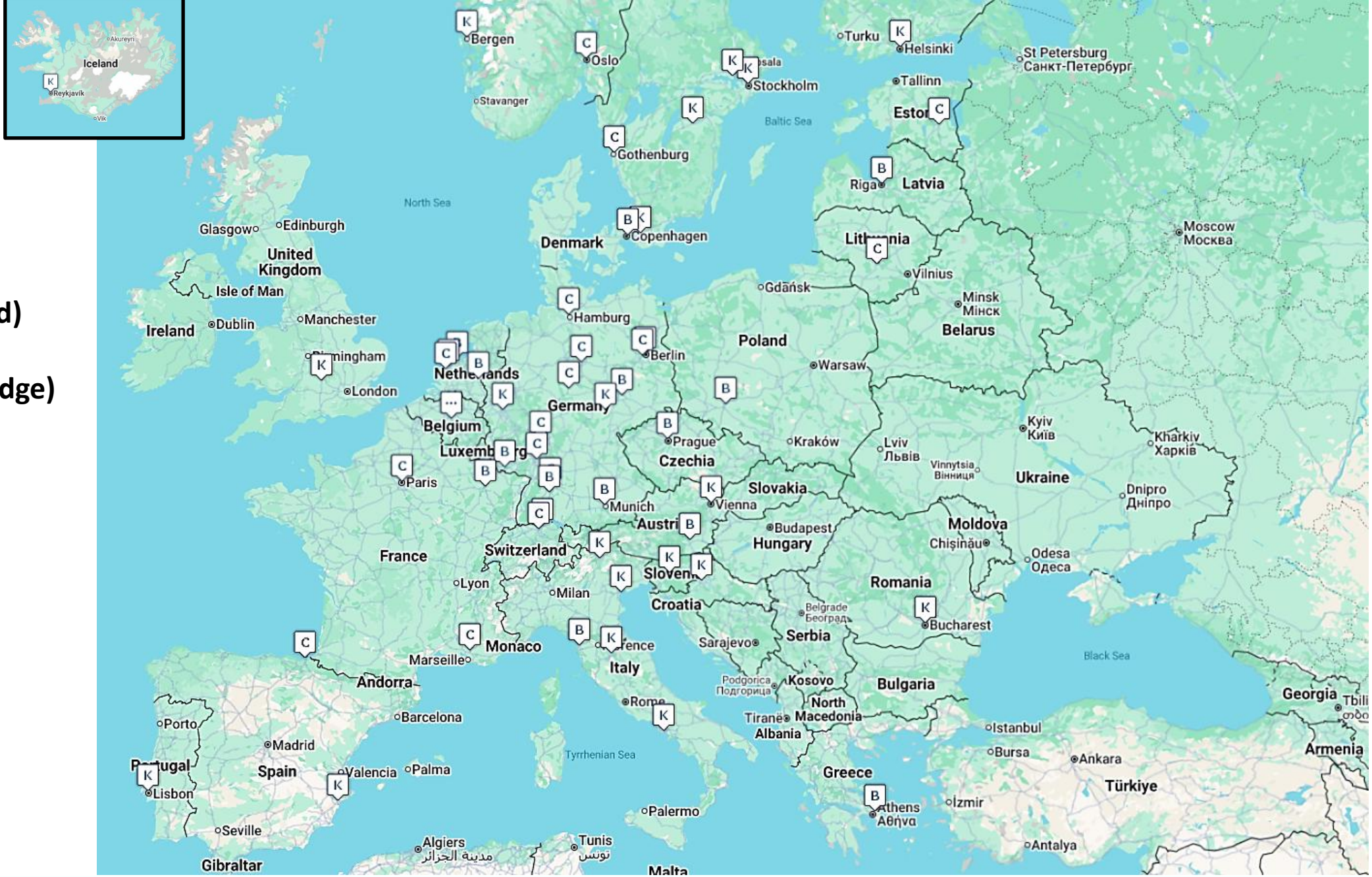
Bendroji kalbinių išteklių ir technologijų infrastruktūra

European Research Infrastructure Consortium

Europos mokslinių tyrimų konsorciumas



- 26 – consortia
- 75 – centres
- B – 23 (certified)
- C – 31
- K – 39 (knowledge)





Search through 928,299 records



Showing all records (528,185 results)

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Lithuanian

Lithuanian (312)

Lithuanian Sign Language (1)

more...

Collection

CLARIN-LT

CLARIN-LT (60)

more...

<< < 1 2 3 4 5 6 > >>

Lithuanian 2-gram dataset

(Part of CLARIN-LT)

Dataset of 2-grams with frequencies extracted from Delfi.lt corpus (~ 70 million words, period: March 2014 - November 2016). Firstly corpus was split into sentences, then symbol analysis as well as analysis of intended structures made of symbols were performed. Also, dictionary of abbreviations was used in order to pre...

Lithuanian

Landing page for this record

1 1

VCR

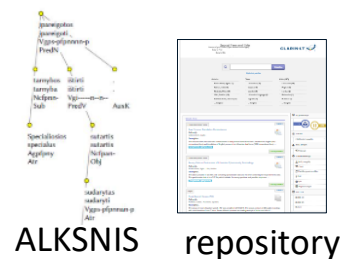
English-Lithuanian Comparable Vaccination Corpus

(Part of CLARIN-LT)

Two news portals were selected for comparable corpora building: the Lithuanian portal DELFI and the English portal The Guardian. The compiled corpora comprise 135 Lithuanian articles from DELFI portal and 135 English articles from the Guardian portal. The main criterion for article extraction from the portals was the p...

English Lithuanian

1 1



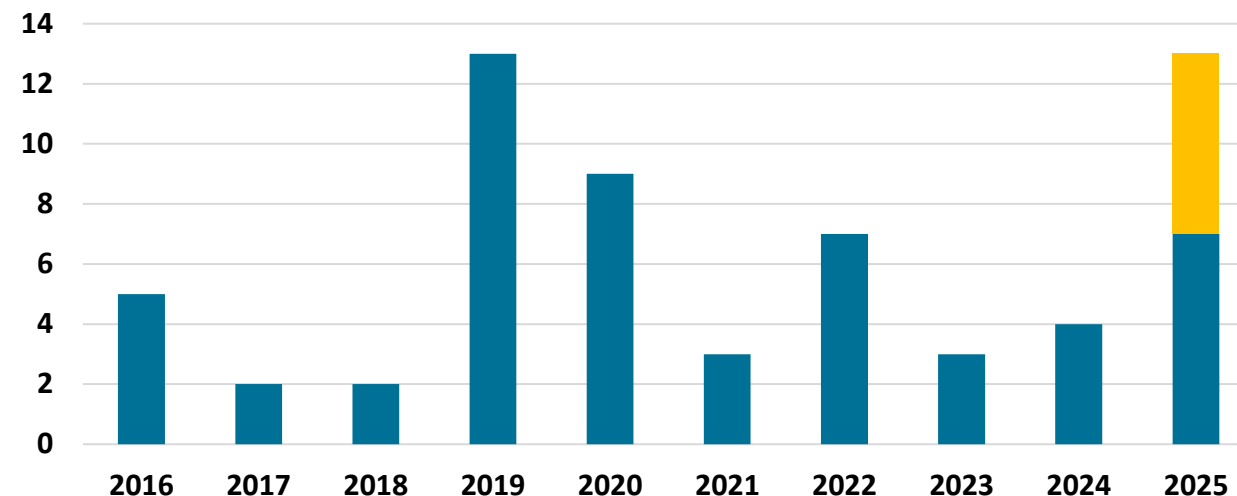
R. Petrauskaitė



WWW



J. Vaičenonienė



2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

2014

October 28
Lithuania has
joined
CLARIN ERIC

CLARIN-LT
project

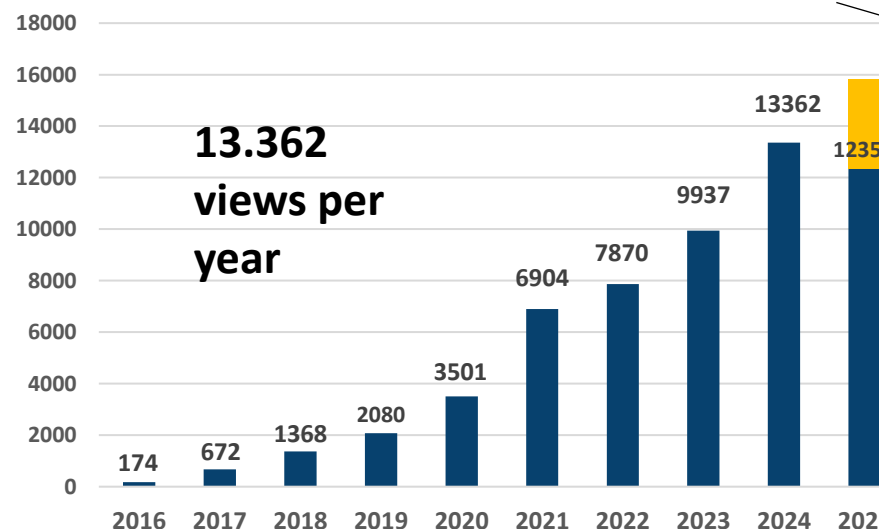
VMU
KTU
VU

PASTOVU

Semantika-2

ARKA // DVITAS

Digires



13.362
views per
year

MRU
BPTI
KU BRIAI

2025



8 thousand
downloads

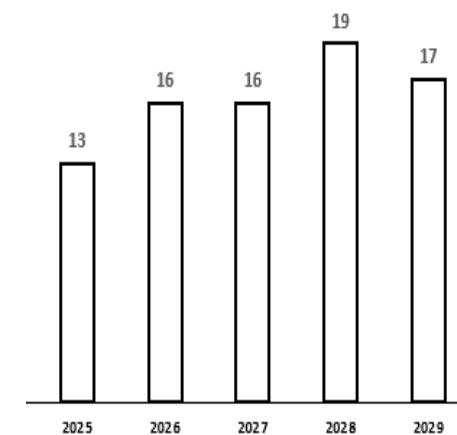
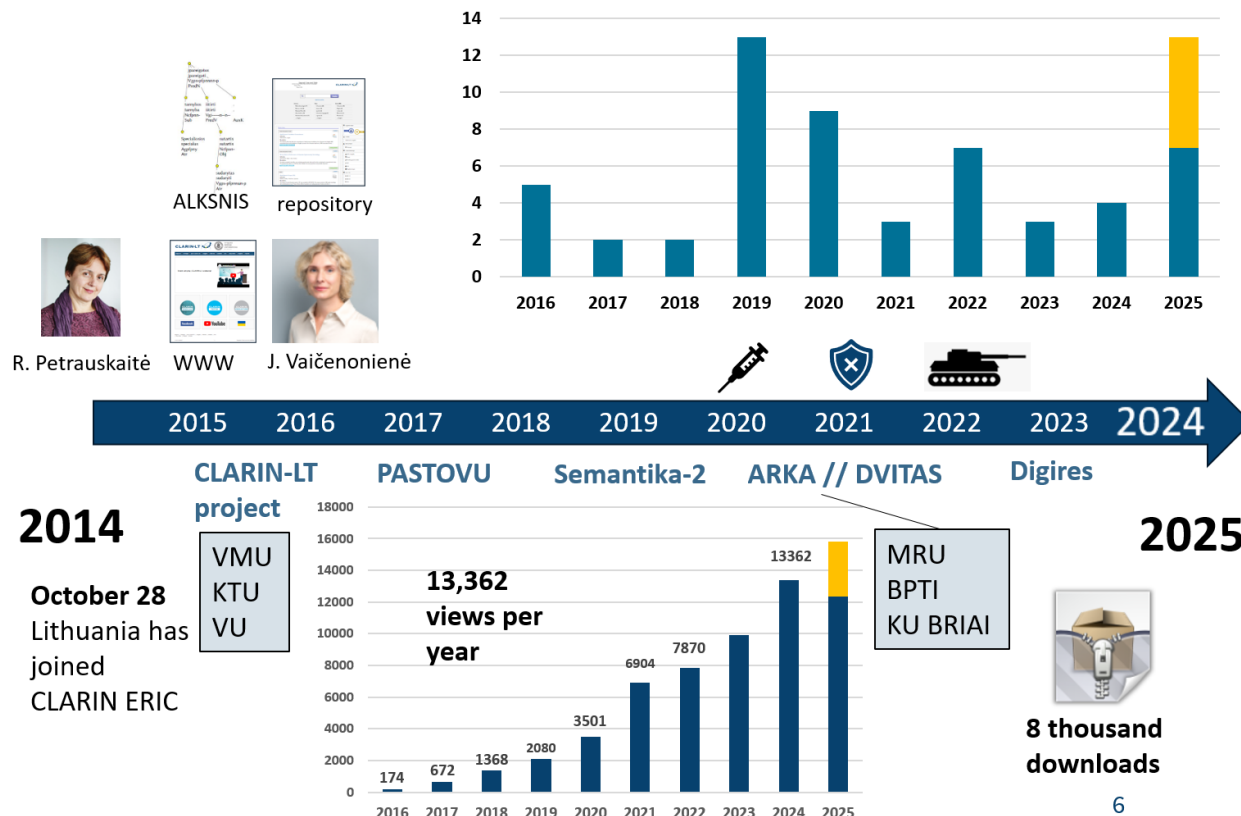
CLARIN





CLARIN-LT plan 2024-2029

- Organising trainings and seminars / involvement of new users
- Certified B centre / CoreTrustSeal
- New resources



RRF project
resources

2029





S. Daukanto 27, Kaunas,
Lithuania

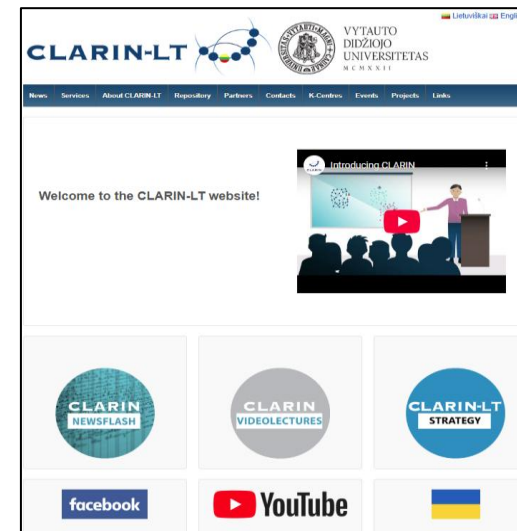


<https://www.clarin.eu/>

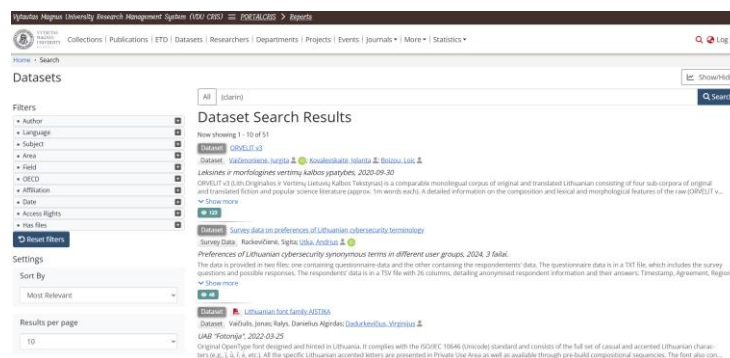


Release
metadata

Where does
CLARIN-LT live?



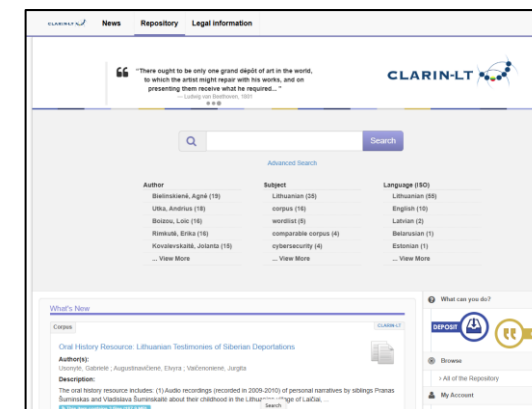
<http://clarin-lt.lt/>



<https://www.vdu.lt/cris/explore/dataset>



<https://sitti.vdu.lt/>



<https://clarin.vdu.lt/xmlui/>

CLARIN



<https://clarin.vdu.lt/xmlui/>

Deposit Free and Safe

License of your Choice (Open licenses encouraged)

Easy to Find

Easy to Cite



Search

[Advanced Search](#)

Author

Bielinskienė, Agnė (19)

Utkā, Andrius (18)

Boizou, Loïc (16)

Rimkutė, Erika (16)

Kovalevskaitė, Jolanta (15)

... View More

Subject

Lithuanian (35)

corpus (16)

wordlist (5)

comparable corpus (4)

cybersecurity (4)

... View More

Language (ISO)

Lithuanian (55)

English (10)

Latvian (2)

Belarusian (1)

Estonian (1)

... View More

What's New

Corpus

CLARIN-LT

Oral History Resource: Lithuanian Testimonies of Siberian Deportations

Author(s):

Usonytė, Gabrielė ; Aujustinavičienė, Elvyra ; Vaičenonienė, Jurgita



? What can you do?

DEPOSIT



CITE



© Browse





Lithuanian Hate Speech Corpus v.1



Please use the following text to cite this item or export to a predefined format:

BIBTEX

CMDI

Butkienė, Rita; Edgaras, Dambrasukas; Algirdas, Šukys and Voldemaras, Žitkus, 2025, *Lithuanian Hate Speech Corpus v.1*, CLARIN-LT digital library in the Republic of Lithuania, <http://hdl.handle.net/20.500.11821/69>.



CLARIN-LT

✎ Authors	Butkienė, Rita ; Edgaras, Dambrasukas ; Algirdas, Šukys ; Voldemaras, Žitkus
📅 Date issued	2025-05-05
📁 Type	corpus
📏 Size	3514 items, 2 files
🚩 Language(s)	Lithuanian
📄 Description	This corpus consists of (1) examples of hate speech based on ethnicity, nationality, or race, and (2) a collection of neutral comments, including both general comments and comments mentioning nationality in a positive or neutral context. All comments are written in Lithuanian and collected from news portals and social networks. The corpus is intended for the development of hate speech detection solutions and research.
🏢 Publisher	Kaunas University of Technology
📄 Acknowledgement	Funded under The Recovery and Resilience Facility (RRF) and the state budget of the Republic of Lithuania Project code: No. 02-002-P-0001



What can you do?

DEPOSIT



CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Statistics

Piwik Statistics

BETA

General Information

Deposit

Cite

Submission Lifecycle



Subject(s)

[hate speech](#)

[offensive language](#)

[comments](#)

[data annotation](#)

Collection(s)

CLARIN-LT

[Show full item record](#)

Files in this item

[Download all files in item \(803.65 KB\)](#)

This item is **Publicly Available** and licensed under:
PUB_CLARIN-LT_End-User-Licence-Agreement_EN-LT

Name EN_About.pdf
Size 196.59 KB
Format PDF
Description Data description



[Download file](#)

Name LT_Apie.pdf
Size 273.22 KB
Format PDF
Description Duomenų aprašas



[Download file](#)

Name DATASET_No_1_Ethnicity_nationality_race_HS.v1.zip
Size 165.27 KB
Format application/zip
Description Dataset 1



[Download file](#)

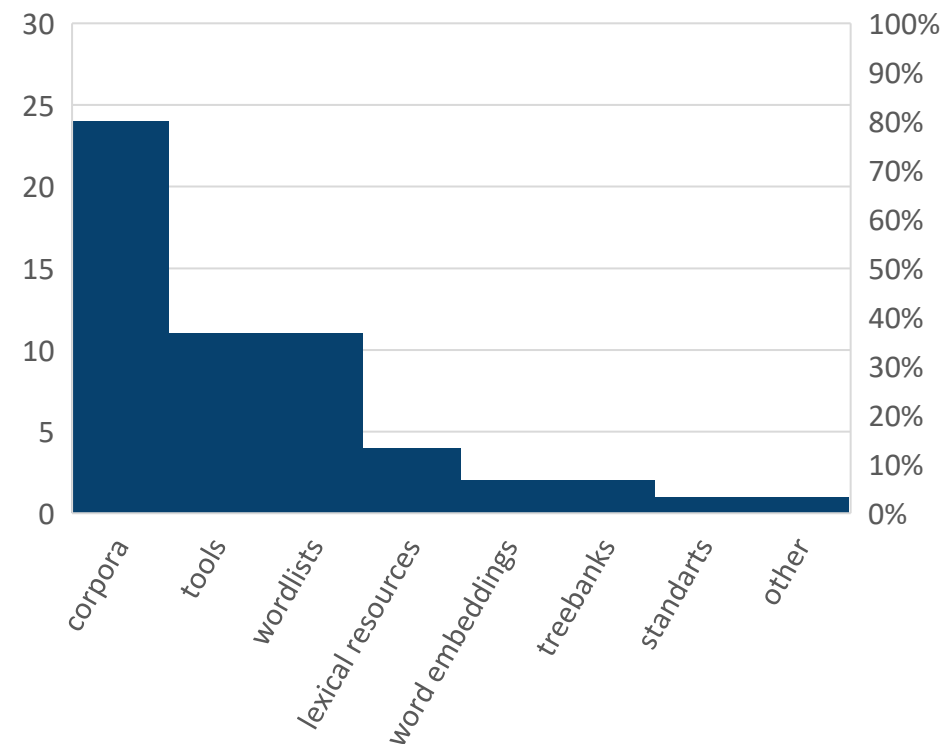
Name DATASET_No_2_Ethnicity_nationality_race_N.v1.zip
Size 168.57 KB
Format application/zip
Description Dataset 2



[Download file](#)

Contents of the Repository

RESOURCE	NUMBER	PUB <i>lic</i>	ACA <i>demic</i>	links
corpora	24	15	7	2
tools	11	7	0	4
wordlists	11	10	1	0
lexical resources	4	4	0	0
word embeddings	2	2	0	0
treebanks	2	2	0	0
standarts	1	1	0	0
other	1	1	0	0
TOTAL	57	43	8	6



Most Popular Resources

No.	Resource	Type	Licence	Views	Dwnlds
1	Lithuanian Spelling Checker V.1.0.42 for macOS	tool	PUB	103.01	34.01
2	Lithuanian speech-to-text Transcriber	tool	link	98.52	0
3	Lithuanian Hunspell dictionary	tool	PUB	89.13	8.67
4	Oral History Resource: Lithuanian Testimonies of Siberian Deportations	corpus	PUB	70.35	10.46
5	Lithuanian keyboard for macOS users	tool	PUB	69.66	13.26
6	Transliteration list of foreign person names into Lithuanian v.1	lexical resource	PUB	33.00	2.78
7	Lithuanian morphologically annotated corpus – MATAS	corpus	ACA	32.55	0.28
8	Lithuanian morphologically annotated corpus - MATAS v1.0	corpus	PUB	32.37	4.08
9	Lithuanian Spelling Checker V.1.0.42 for LibreOffice and OpenOffice	tool	PUB	31.64	6,74
10	Lithuanian font family AISTIKA	tool	PUB	29.56	6.88

Most Downloaded Resources

No.	Resource	Type	Licence	Views	Dwnlds
1	Lithuanian Spelling Checker V.1.0.42 for macOS	tool	PUB	103,01	34,01
2	Lithuanian keyboard for macOS users	tool	PUB	69,66	13,26
3	Oral History Resource: Lithuanian Testimonies of Siberian Deportations	corpus	PUB	70,35	10,46
4	Lithuanian Hunspell dictionary	tool	PUB	89,13	8,67
5	Lithuanian font family AISTIKA	tool	PUB	29,56	6,88
6	Lithuanian Spelling Checker V.1.0.42 for LibreOffice and OpenOffice	tool	PUB	31,64	6,74
7	JABLONSKIS tagset v2	standart	PUB	18,86	6,50
8	Language Technology Research Bibliography for Lithuanian 2016-2020	other	PUB	6,45	4,54
9	Wordlist of Lemmas from the Joint Corpus of Lithuanian	wordlist	PUB	19,30	4,29
10	Dual Pronoun Translation Concordances	lexical resource	PUB	17,52	4,24



We invite you to be a part of CLARIN!

- If you have written an article or a thesis and have a useful resource or tool to share with language community ...
- ... you are welcome to do it via CLARIN-LT repository or any other CLARIN repository
- For CLARIN-LT, contact Andrius or Jurgita, and they will help you!

Andrius.Utka@vdu.lt

Jurgita.Vaicenoniene@vdu.lt





How to Deposit a Resource (basic principles)

- Open standards are preferred over proprietary standards.
- Formats and protocols should be:
 - Well-documented
 - Verifiable
 - Proven (being used in practice).
- Text-based formats are (where possible) preferred over binary formats.
- In the case of digitisation of an analogue signal, using no or lossless compression is recommended.

<https://www.clarin.eu/content/standards-and-formats>





Other things to consider

- Do you have the right to upload the resource? If yes:
- Choose appropriate **license**:
 - **PUB**, public, if the data do not contain restricted material;
 - **ACA**, academic, if data can be used only for academic purposes;
 - **RES**, restricted, if access to data can only be given with special conditions.
- Write a clear **description** of the resource, contents, how it is useful, and how to use it;
- Who is the **author/authors**?





How to Deposit a Resource (help)

What's New

Corpus

CLARIN-LT

Oral History Resource: Lithuanian Testimonies of Siberian Deportations

Author(s):
Usonytė, Gabrielė ; Augustinavičienė, Elvyra ; Vaičenonienė, Jurgita

Description:
The oral history resource includes: (1) Audio recordings (recorded in 2009-2010) of personal narratives by siblings Pranas Šuminskas and Vladislava Šuminskaitė about their childhood in the Lithuanian village of Laičiai, ...
[This item contains 3 files \(187.9 MB\).](#)

Publicly Available

Corpus

CLARIN-LT

LITUND corpus v1

Author(s):
Dambrauskas, Edgaras ; Utka, Andrius

Description:
LITUND contains two comparable corpora: 1. Unreliable news texts. 147 full-text articles (100,678 words) identified as misleading by professional fact-checkers. The corpus includes metadata file with the following ...
[This item contains 3 files \(3.34 MB\).](#)

Academic Use

Corpus

CLARIN-LT

Lithuanian Hate Speech Corpus v.1

Author(s):
Butkienė, Rita ; Edgaras, Dambrauskas ; Algirdas, Šukys ; Voldemaras, Žitkus

Description:
This corpus consists of (1) examples of hate speech based on ethnicity, nationality, or race, and (2) a collection of neutral comments, including both general comments and comments mentioning nationality in a positive or ...
[This item contains 4 files \(803.65 KB\).](#)

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

RSS Feed

RSS 1.0

RSS 2.0

Atom





How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at clarin.eu and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.

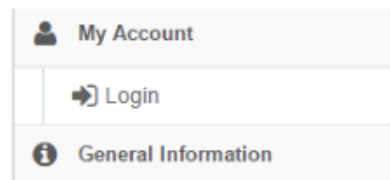


Fig1. Menu Login

Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.

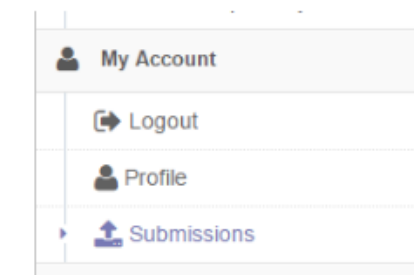


Fig2. Menu Submissions

Now you should be on the main Submission and Workflow tasks page where you can view your incomplete/archive submissions. Click on the 'Start another submission' link to start a new Submission.

Submissions & workflow tasks

Unfinished submissions

These are incomplete item submissions. You may also [start another submission](#).



What can you do?



Browse

> All of the Repository

My Account

Login

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk





Login



CLARIN-LT

News

Repository

Legal information

CLARIN-LT Repository Home / News

CLARIN-LT Repository News

hosted at CLARIN-LT, Kaunas, the Republic of Lithuania

22.08.2025 A new resource of [Oral History Resource: Lithuanian Testimonies of Siberian Deportations](#) was deposited.

25.06.2025 A new resource of [LITUND corpus v1](#) was deposited.

04.06.2025 A new resource of [Lithuanian Hate Speech Corpus v.1](#) was deposited.

24.04.2025 A new resource of [Transliteration list of foreign person names into Lithuanian v.1](#) was deposited.

06.03.2025 A new resource of [Corpus of Transcriptions - part 1](#) was deposited.

28.02.2025 A deeply annotated corpus of Lithuanian poet Kristijonas Donelaitis (1714-1780) has been deposited to the repo. It contains the full Toolbox project, ELAN files, and TEI-XML files [CorDon](#).

26.02.2025 A new resource of [Lithuanian Hunspell Dictionary](#) was deposited.

31.12.2024 A new resource of [English-Lithuanian Parallel Cybersecurity Corpus - DVITAS v2.0](#) was deposited.

31.12.2024 A new resource of [EMVAKA](#) was deposited.

20.12.2024 A new resource of [Lithuanian morphologically annotated corpus - MATAS v3.0](#) was deposited.

04.11.2024 A new resource of [Dual Pronoun Translation Concordances](#) was deposited.

04.10.2024 A new resource of [Survey Data on Preferences of Lithuanian Cybersecurity Terminology](#) was deposited.

13.03.2024 A new resource of [Wordlist of the Contemporary Corpus of Lithuanian Language in the Face of War in Ukraine](#) was deposited.

25.01.2024 A new resource of [JABLONSKIS tagset v2](#) was deposited.

13.04.2023 A new resource of [Lithuanian-English Cybersecurity Termbase v.0.1](#) was deposited.

Login

LT | EN

Sign in to CLARIN-LT Repository

Login via Your home institution (e.g. university)

Vilnius University

Lithuania

University of the Highlands and Islands

Constructor University Bremen

Klaipeda State University of Applied Sciences

Lithuania

Vilnius Gediminas Technical University

Lithuania

ArtEZ University of the Arts

ISM University of Management and Economics

Lithuania

BPP University

Vytautas Magnus University

Lithuania

University of Liverpool

Carnegie Mellon University

University

Please help, I cannot find my provider

Locate me and show nearby providers

Show providers in Lithuania show all countries

CLARIN-LT

News

Repository

Legal information

CLARIN-LT Repository Home / Submissions

Search

Q

Submissions & workflow tasks

Submissions

You may [start a new submission](#).

The submission process includes describing the item and uploading the file(s) comprising it. Each community or collection may set its own submission policy.

Archived Submissions

These are your completed submissions which have been accepted into DSpace.

	Date accepted	Title	Collection
☐	2022-06-21	The Database of Lithuanian multiword expressions	CLARIN-LT

Add new version

Submissions being reviewed

These are your completed submissions which are currently being reviewed by collection curators.

Title	Collection	Status
Test tekstynas	CLARIN-LT	Awaiting final editor's attention

CLARIN-LT

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk



CLARIN-LT

NewsRepositoryLegal information

CLARIN-LT Repository Home / Item submission

Search

Item submission

1. Basic Info

2. Who's involved

3. Describe

4. Upload

5. License

6. Note

7. Review

8. Complete

Submission Info



Corpus



Lexical conceptual



Language description



Technology / Tool / Service



Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Lexical Conceptual Resource" includes lexica, ontologies, dictionaries, word lists etc. "Language Description" covers language models and grammars. "Technology / Tool / Service" is used for tools, systems, system components etc.

Title

 Enter the main title of the item in English.

Project URL

 Enter the project url.

Demo URL

 A url with samples of the resource or, in the case of tools, of samples of the output.

CLARIN-LT

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

<https://clarin.vdu.lt/xmlui/page/deposit>





Good luck depositing!

