

CLARIN

Vienota valodas resursu un tehnoloģiju infrastruktūra

Inguna Skadiņa
CLARIN-LV nacionālā koordinatore

 Latvijas Universitātes
Matemātikas un informātikas institūts

Zinātnieku brokastis

RSU

2023. gada 6. decembrī

CLARIN



Vienota valodas resursu un tehnoloģiju infrastruktūra

CLARIN (Common Language Resources and Technology Infrastructure) ir digitāla **Eiropas pētniecības infrastruktūra** (ERIC)

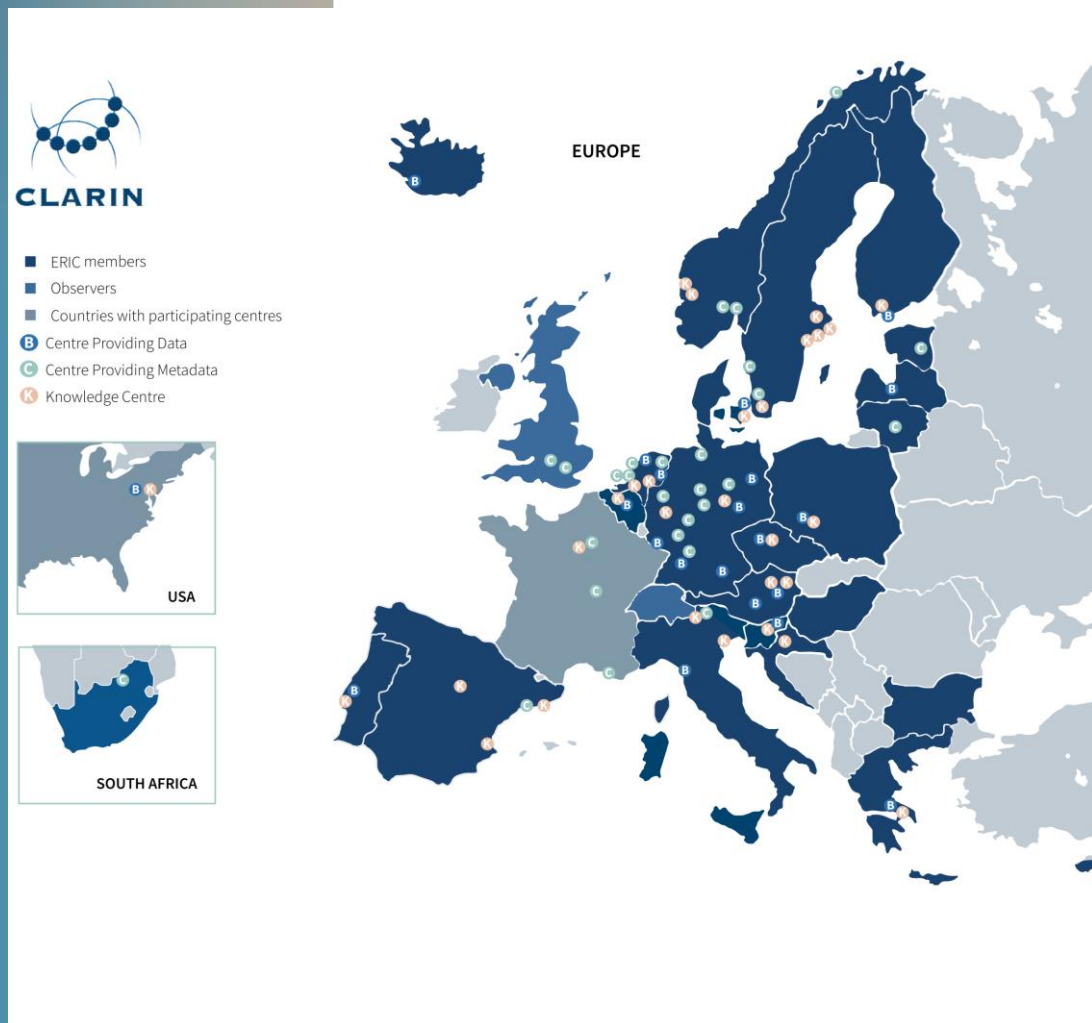
Tās pamatā ir ideja, ka

visi **digitālie valodas resursi un rīki** Eiropā un citur pasaulē ir pieejami

visu jomu, it īpaši **humanitāro un sociālo zinātņu** zinātniekiem un studentiem, izmantojot **vienotu pierakstīšanos**.



CLARIN – daļīta Eiropas pētniecības infrastruktūra

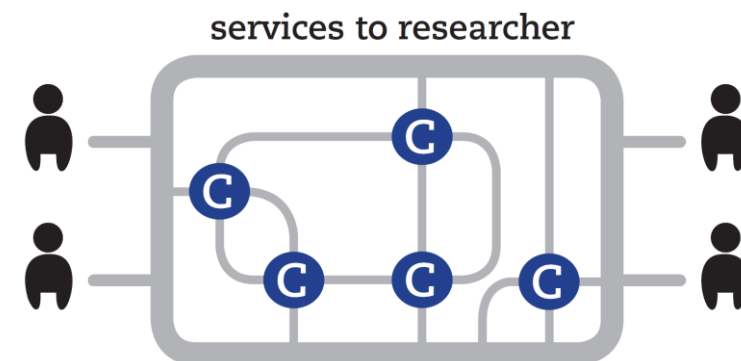


- 23 + 3 valstis
- > 60 reģistrētu centru
- Zināšanu centru tīkls (K-centri)
- Vienkārša un **ilgtspējīga** piekļuve **digitālajiem valodas datiem**
- Mūsdienīgi **valodas apstrādes rīki**, valodas datu izpētei un analīzei neatkarīgi no to atrašanās vietas
- Cieša sasaiste ar Atvērtās zinātnes programmu, fokuss uz **sadarbspēju**



FAIR principles for open language resources

- **F** – *findability*:
existence at a data **provider**, discovery,
persistence, long-term **preservation**
- **A** – *accessibility*:
open data / access
- **I** – *interoperability*:
common **terminology**, data **models** and **formats** (for
both humans & machines)
- **R** – *reusability*:
content, **documentation**, provenance, **versioning**,
licencing



Par CLARIN

CLARIN (Common Language Resources and Technology Infrastructure –Vienota valodas resursu un tehnoloģiju infrastruktūra) ir dalīta Eiropas pētniecības infrastruktūra (ERIC), kurā integrētas zināšanas, valodas resursi un rīki, repozitoriji un servisi, ko piedāvā daudzie CLARIN centri un nacionālie konsorcijs.

CLARIN mērķis ir novērst sadrumstalotību valodas resursu jomā un padarīt valodas resursus pieejamus visu jomu, it īpaši humanitāro un sociālo zinātņu zinātniekiem un studentiem. Lai to īstenotu, izveidota kopīga izkliedēta infrastruktūra, kurā integrēti valodas resursi un rīki no visas Eiropas. CLARIN piedāvā ilgtermiņa risinājumus un servisi digitālo valodas datu un rīku izvietojšanai, savienojšanai un analīzei.

Latvija CLARIN iniciatīvā piedalās kopš tās pirmsākumiem. 2008. gadā LU MII kļūst par CLARIN sagatavošanas posma projekta konsorcijs dalībnieku, būtisku finansiālu atbalstu dalībai projektā sniedz Izglītības un zinātnes ministrija. 2016. gada 1. jūnijā Latvija pievienojas CLARIN ERIC. Izglītības un zinātnes ministrija uzticējusi LU Matemātikas un informātikas institūtam pārstāvēt Latviju CLARIN ERIC un vadīt Latvijas CLARIN konsorcijs. 2019. gadā CLARIN-LV pievienojas CLARIN zināšanu centram morfoloģiski bagātām valodām SAFMORIL. 2020. gada martā tiek izveidots CLARIN-LV repozitorijs, kurā tiek reģistrēti latviešu valodas resursi un rīki. Esot CLARIN ERIC konsorcijs, CLARIN-LV ir apņēmusies ilgstoši rūpēties un saglabāt repozitorijs deponētos digitālos valodas resursus un rīkus. CLARIN-LV īsteno CLARIN ERIC misiju Latvijā, vācot, dokumentējot, kurējot un nodrošinot ilgtermiņa piekļuvi digitālajiem latviešu valodas resursiem un rīkiem. 2022. gada janvārī, konsorcijs apvienojoties septiņiem partneriem, tiek noslēgts CLARIN nacionālā konsorcijs līgums, kurā CLARIN-LV konsorcijs vienojas par CLARIN-LV darbību un turpmāku attīstību. 2023.gada janvārī CLARIN-LV repozitorijs iegūst B-centra statusu un saņem CoreTrustSeal sertifikātu kā uzticams datu repozitorijs.

CLARIN-LV regulāri rīko konferences un seminārus, kuros iepazīstina ar dažādiem valodas resursiem un rīkiem un to lietojumu pētniecībā un valodu tehnoloģiju izveidē. Latvijā paveiktais apkopots vairākās publikācijās.

Kopš 2018. gada CLARIN Latvija darbību atbalsta ERAF projekts "Latvijas Universitāte un institūti Eiropas pētniecības telpā - ekselence, aktivitāte, mobilitāte, kapacitāte", kā arī projekti "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (5.10.2020-5.10.2022) , "Atvērtas un FAIR principi atbilstīgas digitālo humanitāro zinātņu ekosistēmas attīstība Latvijā" (kopš 2022. gada decembra), "Mūsdienu latviešu valodas izpēte un valodas tehnoloģiju attīstība" (kopš 2022. gada) un "Valodu tehnoloģiju iniciatīva" (kopš 2023. gada).



CLARIN-LV piedalās ikgadējā CLARIN konferencē

No 16. oktobra līdz 19. oktobrim Lēvenē notika gadskārtējā Eiropas pētniecības infrastruktūras CLARIN konference.



CLARIN
B-centre 

CLARIN
K CENTRE 

www.clarin.lv
info@clarin.lv

CLARIN-LV valodas resursu repozitorijs

Repozitorijs

Meklēšana korpusā

Par vietni



Atrast

Lingvistiskie dati un valodas apstrādes rīki
Citēšanas atbalsts (ar pastāvīgajiem ID)



Meklēt

Detalizētā meklēšana

Autors

Grūzītis, Normunds (21)

Darģis, Roberts (20)

Saulīte, Baiba (15)

Pretkalniņa, Lauma (13)

Rituma, Laura (12)

... Skatīt vairāk

Temats

text (17)

morphology (14)

dictionary (12)

manual annotation (8)

constituency (7)

... Skatīt vairāk

Valoda (ISO)

Latvian (58)

English (5)

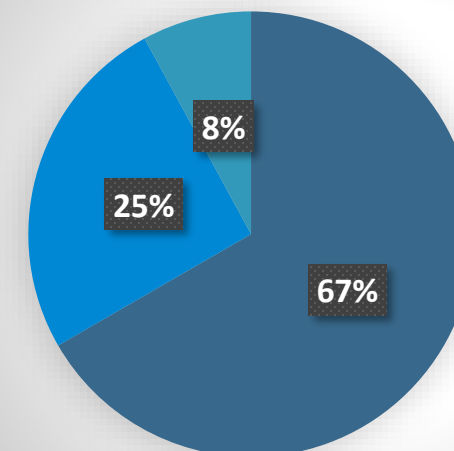
German (4)

Lithuanian (4)

Latgalian (3)

... Skatīt vairāk

Valodas resursi un rīki



- Korpusi
- Leksikoni
- Rīki

CLARIN-LV valodas resursu repozitorijs

Kas jauns

LexicalConceptualResource

CLARIN Centre Of Latvian Language Resources And Tools

Dictionary of Contemporary Latvian Language (MLVV) (2023-09-21)



Autors (-i):

Jērāne, Santa ; et al.

► show everyone

Apraksts:

“Contemporary dictionary of Latvian language” (MLVV), which is developed by the UL Latvian Language institute, is a new explanatory dictionary based on Latvian language materials obtained during the last decade. The analysis ...

Šajā vienumā ir 1 faili (7.05 MB).

Publicly Available

Corpus

CLARIN Centre Of Latvian Language Resources And Tools

Corpus of Latvian PhD Theses (Disertācijas)



Autors (-i):

Darģis, Roberts

Apraksts:

The corpus consists of PhD theses and abstracts published in the University of Latvia, Riga Technical University, Riga Stradins University and Liepaja University until 2020.

Šajā vienumā ir 3 faili (127.73 MB).

Publicly Available

CLARIN-LV valodas resursu repozitorijs

Corpus

CLARIN Centre Of Latvian Language Resources And Tools

LVMED: Latvian Speech Transcripts of the Medical Domain

(AiLab IMCS UL / 2021-09-30)

Autors (-i):

Auziņa, Ilze ; et al.

► show everyone

Šajā vienumā ir 1 fails (267.18 KB).

Publicly Available



SINGLE-WORDS ✓

MULTI-WORDS ✓



reference corpus: Saeima

Word	Word	Word	Word	Word
1 plauša ...	11 CT ...	21 patoloģisks ...	31 pleira ...	41 perēklis ...
2 limfmezgls ...	12 urīnpūslis ...	22 žultspūslis ...	32 dobums ...	42 iekava ...
3 akna ...	13 virsniere ...	23 daiva ...	33 mililitrs ...	43 sienīga ...
4 centimetrs ...	14 liesa ...	24 gluds ...	34 kontrastēt ...	44 aksiāls ...
5 homogēns ...	15 dziedzeris ...	25 izmeklējums ...	35 ieslēgums ...	45 žultsvads ...

CLARIN-LV valodas resursu repozitorijs

Corpus

CLARIN Centre Of Latvian Language Resources And Tools

LVMED: Latvian Speech Transcripts of the Medical Domain

(AiLab IMCS UL / 2021-09-30)

Autors (-i):

Auziņa, Ilze ; et al.

► show everyone

Šajā vienumā ir 1 fails (267.18 KB).

Publicly Available



vecuma normas robežās. </s><s> labajā pusē vecs l
iāta pa kreisi, truncus pulmonalis trīs komats trīs cent

23	<input type="checkbox"/>		doc#89	svaigu infiltrāciju neredz . </s><s> pneimofibroze. </s><s>	sirds	palielināta pa kreisi, aorta vecuma robežā . </s><s> trahejā	
24	<input type="checkbox"/>		doc#91	sklerozētu sienu. </s><s> un tad secinājumā lūdzu rakstām	sirds	neizvērtējama punkts </s><s> un tad apr* secinājumā lūdzu	
25	<input type="checkbox"/>		doc#94	s> pastiprināts plaušu zīmējums piesakņu rajonā . </s><s>	sirds	, aorta vecuma normā . </s><s> sinusi brīvi. </s><s> vēder:	
26	<input type="checkbox"/>		doc#103	><s> nākamā rindā lūdzu rakstām nākamajā rindā </s><s>	sirds	paplašināta šķērsizmērā , vairāk pa kreisi. </s><s> vēl atzīn	
27	<input type="checkbox"/>		doc#118	<s> aorta nav paplašināta. </s><s> nākamā rindā </s><s>	sirds	nav paplašināta šķērsizmērā . </s><s> un tas arī viss paldie	
28	<input type="checkbox"/>		doc#134	><s> saknes sabiezētas , salīdzinoši kreisā sakne. </s><s>	sirds	palielināta pa kreisi, aorta vecuma robežā . </s><s> kontrol	
29	<input type="checkbox"/>		doc#138	hiperplastisks etmoidīts un punkts </s><s> plaušu lauki un	sirds	bez redzamas novirzes no normas. </s><s> paraksts </s><	
30	<input type="checkbox"/>		doc#141	rencējas punkts šķidrumu pleiras telpās nekonstatē punkts	sirds	nav paplašināta šķērsizmērā . </s><s> aorta nav paplašināt	
31	<input type="checkbox"/>		doc#145	s> plašu emfizēma. </s><s> difūza pneimofibroze. </s><s>	sirds	nav palielināta. </s><s> aorta plata, sklerozēta. </s><s> vir:	
32	<input type="checkbox"/>		doc#154	:glu kalcinācija kreisā pusē. </s><s> tālāk rakstām </s><s>	sirds	šķērsizmērā nav paplašināta. </s><s> aorta sklerotizēta. </:	

CLARIN-LV valodas resursu repozitorijs

Tool Service

CLARIN Centre Of Latvian Language Resources And Tools

RUTA:MED – Dual Workflow Medical Speech Transcription Pipeline and Editor

(AiLab IMCS UL / 2022)

Author(s):

Znotiņš, Artūrs ; Dargis, Roberts ; Grūzītis, Normunds ; Bārzdiņš, Guntis and Goško, Didzis

This item contains no files.



✓ Saglabāt ○ 5s **B** *I* U x^2 x_2 ☰

MR prostatai ar k/v.

Prostata 3,9 x 3,3 x 3,8 cm, tilpums 26,15 cm³.

PSA līmenis nav zināms.

Dziedzera pārejas daļā redzami labdabīgas hiperplāzijas mezgli.

Urīnpūslis gludām sieniņām, bez intralūmenāliem ieslēgumiem.

Patoloģiska lieluma l/m nesaskatu.

Slēdziens

Labdabīga prostatas hiperplāzija.

Pašreiz nav datu par augstas malignitātes tumora audiem prostatā.



00:00:06 / 00:01:32

CLARIN-LV valodas resursu repozitorijs

Kas jauns

LexicalConceptualResource

CLARIN Centre Of Latvian Language Resources And Tools

Dictionary of Contemporary Latvian Language (MLVV) (2023-09-21)



Autors (-i):

Jērāne, Santa ; et al.

► show everyone

Apraksts:

"Contemporary dictionary of Latvian language" (MLVV), which is developed by the UL Latvian Language institute, is a new explanatory dictionary based on Latvian language materials obtained during the last decade. The analysis ...

📎 Šajā vienumā ir 1 fails (7.05 MB).

Publicly Available

Corpus

CLARIN Centre Of Latvian Language Resources And Tools

Corpus of Latvian PhD Theses (Disertācijas)



Autors (-i):

Darģis, Roberts

Apraksts:

The corpus consists of PhD theses and abstracts published in the University of Latvia, Riga Technical University, Riga Stradins University and Liepaja University until 2020.

📎 Šajā vienumā ir 3 faili (127.73 MB).

Publicly Available

Dictionary of Contemporary Latvian Language (MLVV) (2023-09-21)

Author(s):

Jērāne, Santa ; et al.

► show everyone

Description:

"Contemporary dictionary of Latvian language" (MLVV), which is developed by the UL Latvian Language institute, is a new explanatory dictionary based on Latvian language materials obtained during the last decade. The analysis ...



This item contains 1 file (7.05 MB).

Publicly Available

This item is **Publicly Available** and licensed under:
Creative Commons - Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Name	mlvv_2023_4_tei.zip
Size	7.05 MB
Format	application/zip
Description	MLVV open data in the TEI/XML format (https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html)
MD5	cef0a6b5399382c98b294b4406c2d77e



Download file

Preview

Licences un to pārvaldīšana

ToolService

CLARIN Centre Of Latvian Language Resources And Tools

[Ilvars - Latvian Male VITS Text-to-Speech Model \(vers. 2023\)](#)




Autors (-i):

Darģis, Roberts and Auziņa, Ilze

Apraksts:

A neural model for text-to-speech (TTS) synthesis in Latvian. Trained using VITS on a 25-hour speech corpus of audiobooks read in a male voice. Available for academic and non-commercial purposes via an API. To get access ...

Šajā vienumā ir 1 fails (376.17 MB).

Restricted Use  



Licences un to pārvaldīšana

ToolService

ci Licenču pārvaldīšana

Ilvars - Latvian Male VITS Text-to-Speech Model (vers. 2023)

Autors (-i):

Darģis, Roberts and Auziņa, Ilze

Apraksts:

A neural model for text-to-speech (TTS) synthesis in Latvian. Trained using audiobooks read in a male voice. Available for academic and non-commerc

Šajā vienumā ir 1 fails (376.17 MB).

All Licenses

Define License

Define License Label

License Name:

CLARIN RESTRICTED

Definition (URL):

https://www.kielipankki.fi/lic/example-license-clarin-res-id-by-ŗ

Confirmation:

Ask always

Label:

RES

Extended Label(s):

BY

SA

NC

ReD

Lietotājs saņems e-pasta ziņojumu ar lejupielādes instrukcijām.

Lietotāja vārds

Dzimšanas datums

Adrese

Valsts

Papildu nepieciešamā lietotāja informācija:

Vaicāt lietotājam citu e-pasta adresi

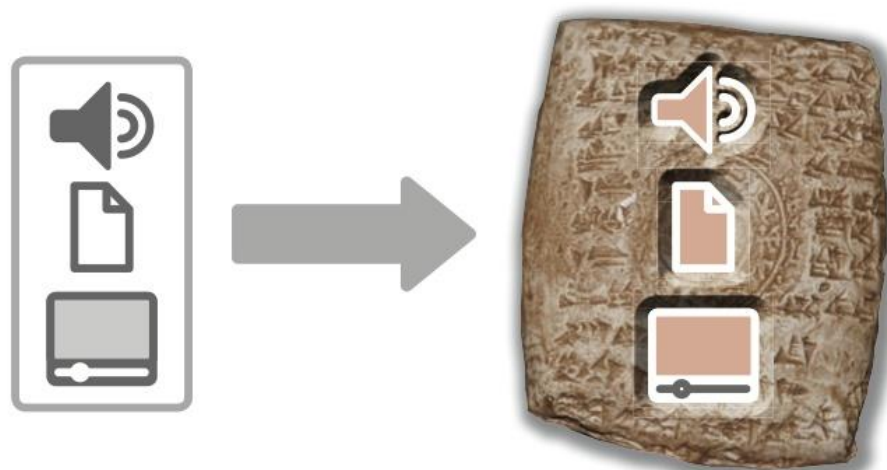
Jautāt lietotājam par organizāciju (nav obligāti).

Jautāt lietotājam par organizāciju (obligāti).

Pajautāties lietotājam par viņa nodomiem ar vienumu.

Datu uzglabāšanas servisi: deponēšana

- Viens no CLARIN pamatuzdevumiem ir nodrošināt, ka valodas resursi tiek arhivēti, tie ir **pieejami ilgtermiņā** un tie ir **uzticami**
- Tāpēc CLARIN centri piedāvā valodas resursu uzglabāšanas (deponēšanas) pakalpojumu



https://www.clarin.lv/attachments/CLARIN%202020_resursu_iesniegsana.pdf

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!



Citēšana un unikālie identifikatori

Universal Dependencies 2.0



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


BIBTEX

CMDI

Nivre, Joakim; et al., 2017, *Universal Dependencies 2.0*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-1983>.



 This resource is also integrated in following services:

Share:  

PML-TQ



LINDAT / CLARIAH-CZ



Citēšana

Universal Dependencies 2.0

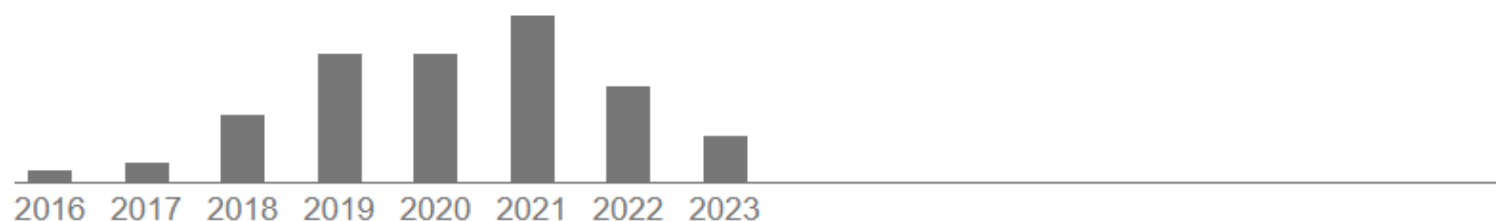
Autori Joakim Nivre, et al.

Publicēšanas
datums 2017/3

Avots <http://hdl.handle.net/11234/1-1983>

Apraksts UNIV-ST-ETIENNE| ENS-LYON| UNIV-LYON3| UNIV-PARIS7| ENS-PARIS| UNIV-TLN| PRES_CLERMONT| CNRS| INRIA| UNIV-AMU| UNIV-LYON2| UNIV-PARIS3| LATTICE| MODYCO| LLF| LPP| CERHAC| UNIV-LORRAINE| INRIA2| LORIA| LORIA-NLPKD| PSL| USPC| LIS-LAB| IHRIM| UNIV-PARIS-LUMIERES| UDL| UNIV-PARIS| UP-SOCIETES-HUMANITES| UNIV-PARIS-NANTERRE| IFRA-NIGERIA

Atsauču
kopskaits **Minēts 312**

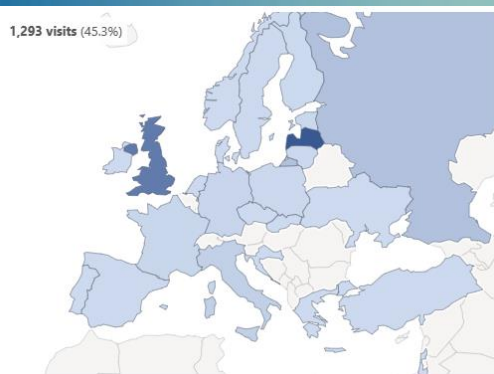


Populārākie valodas resursi

Visitor Map



1,293 visits (45.3%)



CLARIN-LV

Item/Handle	Number of views
Tēzaurš.lv 2020 (Spektors, Andrejs et al) (20.500.12574/9)	5352 [+bots: 6275]
Latvian Treebank v2.5 (Rituma, Laura et al) (20.500.12574/10)	3447 [+bots: 4109]
Tēzaurš.lv 2022 (Spektors, Andrejs et al) (20.500.12574/66)	3296 [+bots: 3494]
Balanced Corpus of Modern Latvian (LVK2018) (Levāne-Petrova, Kristīne et al) (20.500.12574/11)	3093 [+bots: 3206]
LVBERT - Latvian BERT (Znotiņš, Artūrs) (20.500.12574/43)	2967 [+bots: 3574]
NLP-PIPE: Latvian NLP Tool Pipeline (Znotiņš, Artūrs et al) (20.500.12574/4)	2685 [+bots: 2857]
The Corpus of Early Written Latvian (Andronova, Everita et al) (20.500.12574/12)	2514 [+bots: 2882]
Rainis (Spektors, Andrejs et al) (20.500.12574/41)	2413 [+bots: 2820]
Corpus of Latvian Pandemic Diaries 2020–2021 (Reinsone, Sanita et al) (20.500.12574/48)	2403 [+bots: 2800]
Latvian AMR Sembank (Znotiņš, Artūrs et al) (20.500.12574/40)	2360 [+bots: 2705]

Latviešu valodas resursi citu valstu CLARIN repozitorijos

Latvian Treebank v2.5

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

Rituma, Laura; Pretkalniņa, Lauma; Saulīte, Baiba; Nešpore-Bērzkalne, Gunta and G
Treebank v2.5, CLARIN-LV digital library at IMCS, University of Latvia, <http://hdl.han>

This resource is also integrated in following services:

PML-TQ

The screenshot shows the LINDAT TreeQuery interface. The browser address bar indicates the URL: `lindat.mff.cuni.cz/services/pmltq/#!/treebank/lvtb25/query/IYWgdg9gJgpgBAKANpwLYHoAuWdmcB+cA5AHphEA0icNtoks1tcKG2ehpADkXALr8A3AiA/result/sv`. The interface includes a navigation menu with options like Repository, Corpus Search, TreeQuery, Treex, More Apps, About, and CLARIN. The current page is titled "LVTB - Latvian dependency-constituency treebank v 2.5". A search query is entered: `a-node [m/tag ~ '^n', a-node [m/tag ~ '^p']];`. The results show 22 items, with 1 selected (a-node) and 2 others (a-node). The selected item is a dependency-constituency tree for the sentence: "Kādā atpūtas brīdī, kad imperators sēdējis zem tējas koka un dzēris vārītu ūdeni, sacēlies vējš un imperatora krūzē iepūtis pāris tējas koka lapas." The tree diagram shows a root node "sent" (ID: a-c2-p4s2) branching into "pred" and "punct". The "pred" node further branches into "adv", "subj", "conj", and "crdPart". The "adv" node branches into "attr", "attrC1", and "attrC2". The "subj" node branches into "crdPart" and "auxVerb". The "conj" node branches into "adv" and "basElem". The "crdPart" node branches into "adv", "basElem", and "auxVerb". The "attr" node branches into "attr" and "attrC1". The "attrC1" node branches into "attrC1" and "attrC2". The "attrC2" node branches into "attrC2" and "attrC3". The "attrC3" node branches into "attrC3" and "attrC4". The "attrC4" node branches into "attrC4" and "attrC5". The "attrC5" node branches into "attrC5" and "attrC6". The "attrC6" node branches into "attrC6" and "attrC7". The "attrC7" node branches into "attrC7" and "attrC8". The "attrC8" node branches into "attrC8" and "attrC9". The "attrC9" node branches into "attrC9" and "attrC10". The "attrC10" node branches into "attrC10" and "attrC11". The "attrC11" node branches into "attrC11" and "attrC12". The "attrC12" node branches into "attrC12" and "attrC13". The "attrC13" node branches into "attrC13" and "attrC14". The "attrC14" node branches into "attrC14" and "attrC15". The "attrC15" node branches into "attrC15" and "attrC16". The "attrC16" node branches into "attrC16" and "attrC17". The "attrC17" node branches into "attrC17" and "attrC18". The "attrC18" node branches into "attrC18" and "attrC19". The "attrC19" node branches into "attrC19" and "attrC20". The "attrC20" node branches into "attrC20" and "attrC21". The "attrC21" node branches into "attrC21" and "attrC22". The "attrC22" node branches into "attrC22" and "attrC23". The "attrC23" node branches into "attrC23" and "attrC24". The "attrC24" node branches into "attrC24" and "attrC25". The "attrC25" node branches into "attrC25" and "attrC26". The "attrC26" node branches into "attrC26" and "attrC27". The "attrC27" node branches into "attrC27" and "attrC28". The "attrC28" node branches into "attrC28" and "attrC29". The "attrC29" node branches into "attrC29" and "attrC30". The "attrC30" node branches into "attrC30" and "attrC31". The "attrC31" node branches into "attrC31" and "attrC32". The "attrC32" node branches into "attrC32" and "attrC33". The "attrC33" node branches into "attrC33" and "attrC34". The "attrC34" node branches into "attrC34" and "attrC35". The "attrC35" node branches into "attrC35" and "attrC36". The "attrC36" node branches into "attrC36" and "attrC37". The "attrC37" node branches into "attrC37" and "attrC38". The "attrC38" node branches into "attrC38" and "attrC39". The "attrC39" node branches into "attrC39" and "attrC40". The "attrC40" node branches into "attrC40" and "attrC41". The "attrC41" node branches into "attrC41" and "attrC42". The "attrC42" node branches into "attrC42" and "attrC43". The "attrC43" node branches into "attrC43" and "attrC44". The "attrC44" node branches into "attrC44" and "attrC45". The "attrC45" node branches into "attrC45" and "attrC46". The "attrC46" node branches into "attrC46" and "attrC47". The "attrC47" node branches into "attrC47" and "attrC48". The "attrC48" node branches into "attrC48" and "attrC49". The "attrC49" node branches into "attrC49" and "attrC50". The "attrC50" node branches into "attrC50" and "attrC51". The "attrC51" node branches into "attrC51" and "attrC52". The "attrC52" node branches into "attrC52" and "attrC53". The "attrC53" node branches into "attrC53" and "attrC54". The "attrC54" node branches into "attrC54" and "attrC55". The "attrC55" node branches into "attrC55" and "attrC56". The "attrC56" node branches into "attrC56" and "attrC57". The "attrC57" node branches into "attrC57" and "attrC58". The "attrC58" node branches into "attrC58" and "attrC59". The "attrC59" node branches into "attrC59" and "attrC60". The "attrC60" node branches into "attrC60" and "attrC61". The "attrC61" node branches into "attrC61" and "attrC62". The "attrC62" node branches into "attrC62" and "attrC63". The "attrC63" node branches into "attrC63" and "attrC64". The "attrC64" node branches into "attrC64" and "attrC65". The "attrC65" node branches into "attrC65" and "attrC66". The "attrC66" node branches into "attrC66" and "attrC67". The "attrC67" node branches into "attrC67" and "attrC68". The "attrC68" node branches into "attrC68" and "attrC69". The "attrC69" node branches into "attrC69" and "attrC70". The "attrC70" node branches into "attrC70" and "attrC71". The "attrC71" node branches into "attrC71" and "attrC72". The "attrC72" node branches into "attrC72" and "attrC73". The "attrC73" node branches into "attrC73" and "attrC74". The "attrC74" node branches into "attrC74" and "attrC75". The "attrC75" node branches into "attrC75" and "attrC76". The "attrC76" node branches into "attrC76" and "attrC77". The "attrC77" node branches into "attrC77" and "attrC78". The "attrC78" node branches into "attrC78" and "attrC79". The "attrC79" node branches into "attrC79" and "attrC80". The "attrC80" node branches into "attrC80" and "attrC81". The "attrC81" node branches into "attrC81" and "attrC82". The "attrC82" node branches into "attrC82" and "attrC83". The "attrC83" node branches into "attrC83" and "attrC84". The "attrC84" node branches into "attrC84" and "attrC85". The "attrC85" node branches into "attrC85" and "attrC86". The "attrC86" node branches into "attrC86" and "attrC87". The "attrC87" node branches into "attrC87" and "attrC88". The "attrC88" node branches into "attrC88" and "attrC89". The "attrC89" node branches into "attrC89" and "attrC90". The "attrC90" node branches into "attrC90" and "attrC91". The "attrC91" node branches into "attrC91" and "attrC92". The "attrC92" node branches into "attrC92" and "attrC93". The "attrC93" node branches into "attrC93" and "attrC94". The "attrC94" node branches into "attrC94" and "attrC95". The "attrC95" node branches into "attrC95" and "attrC96". The "attrC96" node branches into "attrC96" and "attrC97". The "attrC97" node branches into "attrC97" and "attrC98". The "attrC98" node branches into "attrC98" and "attrC99". The "attrC99" node branches into "attrC99" and "attrC100".

CLARIN



CLARIN virtuālā valodas resursu krātuve (VLO)

latvian

Showing 1 to 10 of 2,733 results for latvian Results per page: 10

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

Language

Collection

Type to filter or search for more

- Europeana - National Library of Latvia: Collection 748774 (1332)
- Europeana newspapers full-text (574)
- Europeana - National Library of Latvia: Collection 569516 (396)
- CLARIN:EL Catalogue (145)
- Language resources and tools of AiLab IMCS UL (51)
- LRT + Open Submissions Data & Tools (42)
- LINDAT / CLARIAH-CZ Data & Tools (41)
- Europeana - National Library of Latvia: Collection 57760 (31)
- CLARIN.SI data & tools (18)
- Rigascher Almanach für das Jahr ..., das 700ste seit Gründung der Stadt - 1858-1944 Rigascher Almanach für das Jahr ..., das 700ste seit Gründung der Stadt - 1858-1944 (16)

more...

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Latvian Wikipedia
(Part of Language resources and tools of AiLab IMCS UL)

The corpus consists of all information published on **Latvian** Wikipedia until February 2022.

Latvian

Landing page for this record

TITUS Latvian
(Part of LRT + Open Submissions Data & Tools)

ca. 10.000 tokens; linked with relational database; XML-encoding in progress

Latvian

Landing page for this record

Bulgarian, Latvian
(Part of Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC))

Newton's list comments: Balkanton (Bulgarian), **Latvian**, fables in Lettish Archive tape note: 00.00 AIATSIS announcement. Field Tape LT10, side 1. Balkanton, **Latvian** and Lettish languages. 25.35 End of Field Tape LT10, side 1. Tape LT10, side 2. 63.00 End of Field Tape LT10, side 2. AIATSIS Identifier: A16994. Language as given: Balkanton, Lettish

Bulgarian Latvian

<https://vlo.clarin.eu>

CLARIN valodas resursu kopas

Corpora

- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources

- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

[Resource Families | CLARIN ERIC](#)

CLARIN RESOURCE FAMILIES

Parliamentary corpora are a very important multidisciplinary language resource that can be approached from many research perspectives, including not only political science, but also sociology, history, psychology, and applicative approaches to linguistics, for instance, critical discourse analysis. The good availability of parliamentary proceedings in digitized form and granted access rights to public information in the EU countries have motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora.

PARLIAMENTARY CORPORA

18 parliamentary corpora

1 MULTILINGUAL: *Europarl* (21 languages)

17 MONOLINGUAL IN 14 LANGUAGES:

1 Czech	1 Finnish
1 Danish	3 German
2 English	1 Greek
1 French	1 Lithuanian
1 Estonian	2 Norwegian
	1 Polish
	1 Portuguese
	1 Slovenian
	1 Swedish

AVAILABILITY

3 through a concordancer

9 for download

4 both

SIZE

7 small (<10 million tokens)

7 medium (10–100 million tokens)

3 large (>100 million tokens)

ANNOTATION

8 PoS-tagged

7 lemmatised

www.clarin.eu/resource-families

CLARIN valodas resursu kopas

Corpora

- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources

- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries

Tools

- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

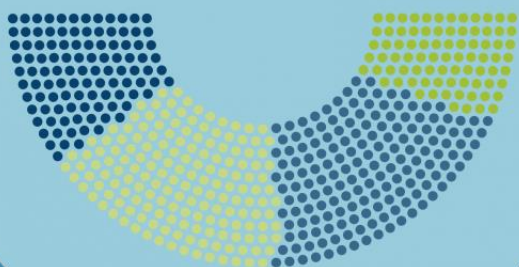
Corpora of Academic Texts

Corpora of academic texts contain scholarly writing, such as research papers, essays and abstracts published in academic journals, conference proceedings, and edited volumes, theses written by students at undergraduate and graduate levels, and scientific monographs.

The CLARIN ERIC infrastructure gives access to 24 corpora of academic texts, 2 of which are multilingual and 22 monolingual. The available corpora contain scholarly texts in the following 11 languages: Czech, English, Estonian, Finnish, French, German, Greek, Russian, Slovenian, Spanish, and Swedish. More than 15 different scholarly disciplines are represented, with the most prominent being linguistics, computer science, economics, and **medicine**. The majority of the corpora are richly tagged and are available under public licences.

Sadarbība, veidojot kopīgus valodas resursus

ParlaMint



Corpus

CLARIN.SI Data & Tools

Multilingual comparable corpora of parliamentary debates ParlaMint 2.0



(CLARIN ERIC / 2021-05-10)

Author(s):

Erjavec, Tomaž ; Ogrodniczuk, Maciej ; Osenova, Petya ; Ljubešić, Nikola ; Simov, Kiril ; Grigorova, Vladislava ; Rudolf, Michał ; Pančur, Andrej ; Kopp, Matyáš ; Barkarson, Starkaður ; Steingrímsson, Steinþór ; van der Pol, Henk ; Depoorter, Griet ; de Does, Jesse ; Jongejan, Bart ; Haltrup Hansen, Dorte ; Navarretta, Costanza ; Calzada Pérez, María ; de Macedo, Luciana D. ; van Heusden, Ruben ; Marx, Maarten ; Çöltekin, Çağrı ; Coole, Matthew ; Agnoloni, Tommaso ; Frontini, Francesca ; Montemagni, Simonetta ; Quochi, Valeria ; Venturi, Giulia ; Ruisi, Manuela ; Marchetti, Carlo ; Battistoni, Roberto ; Sebők, Miklós ; Ring, Orsolya ; Dargis, Roberts ; Utkā, Andrius ; Petkevičius, Mindaugas ; Briedienė, Monika ; Krilavičius, Tomas ; Morkevičius, Vaidas

This item contains 17 files (2.03 GB).

Publicly Available

Showcase 2: 'COVID' keywords

Italy	Poland	Slovenia	UK
pandemic	pandemic	epidemic	covid
covid	<i>bell</i>	ventilator	coronavirus
covid-19	coronavirus	coronavirus	furlough
coronavirus	covid-19	covid-19	lockdown
lockdown	mask	virus	pandemic
mask	epidemic	quarantine	distancing
recovery	quarantine	corona	ppe
European Stability Mechanism (ESM)	Kukiz15	mask	<i>inaudible</i>
distancing	the Left	pandemic	
fund	shield	anti-corona (adj)	
serologic	anti-crisis	/paramilitary group/	
virus	covid	/name/	

[Parliamentary Debates in the COVID Times](#)

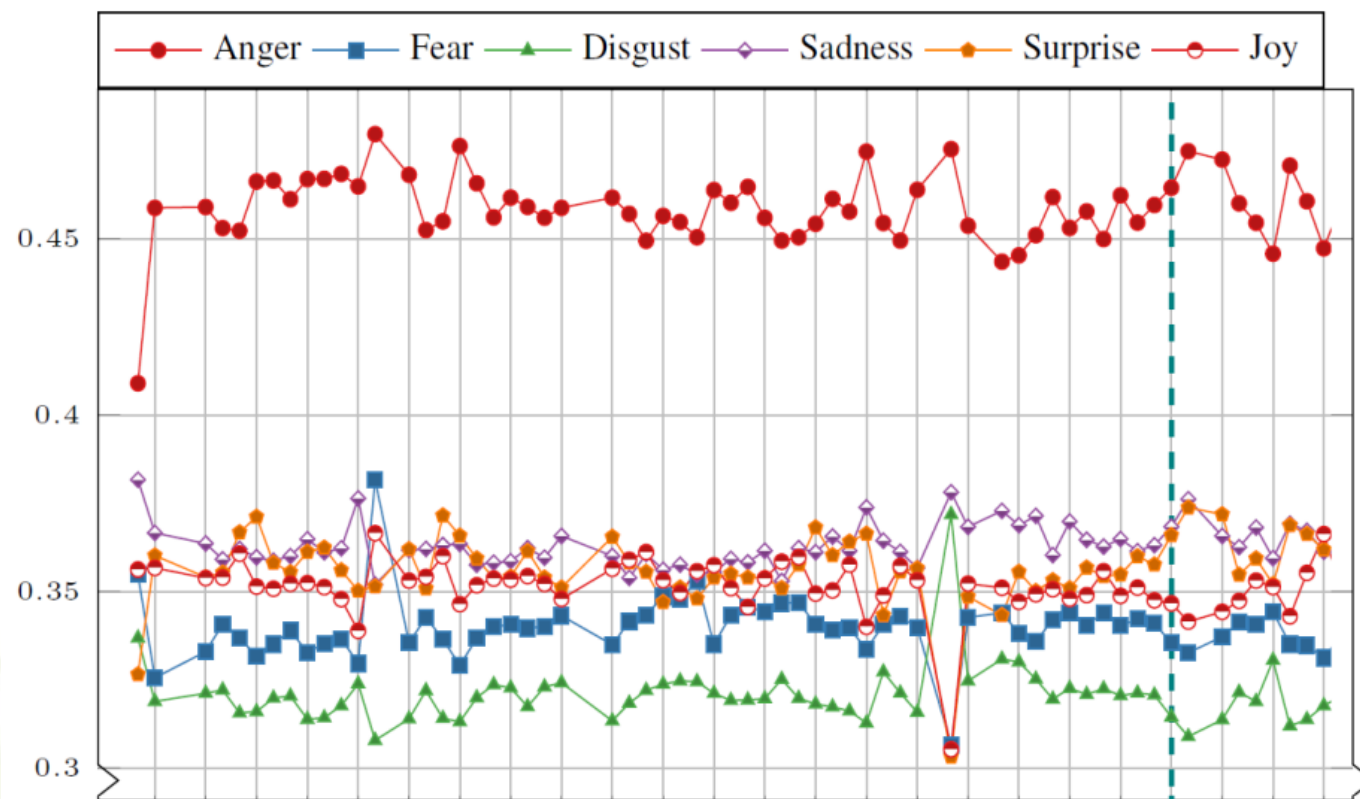
Helsinki DH Hackathon 21



Showcase 6: Emotions Running High

Kurtoğlu and Çöltekin (2022).
[Emotions Running High?](#)

- Investigating polarization of politics by assigning emotion scores to speeches
- anger being the dominant emotion
- the ruling party showing more stable emotions compared to the opposition



CLARIN zināšanu centrs morfolóģiski bagātām valodām SAFMORIL

- SAFMORIL ir CLARIN zināšanu centrs (knowledge centre) morfolóģiski bagātām valodām
- SAFMORIL konsultē digitālo humanitāro zinātņu pētniekus un valodu tehnolóģiju izstrādātāju par valodas resursiem un rīkiem morfolóģiski bagātām valodām
- SAFMORIL apvienojušās Helsinku Universitāte, Trumses Universitāte, Latvijas Universitātes Matemātikas un Informātikas institūts un Vītauta Dižā Universitāte

Konference "Jaunākie valodas resursi un rīki digitālajām humanitārajām zinātnēm"

2023. gada 12. septembrī Latvijas Universitātes Matemātikas un informātikas institūts sadarbībā ar digitalhumanities.lv un DHELI projektu rīkoja ikgadējo CLARIN-LV konferenci "Jaunākie valodas resursi un rīki digitālajām humanitārajām zinātnēm". Konferences materiāli:

- *Kaspars Bērziņš (VPC)*. Augstākās izglītības un zinātnes IT koplietošanas pakalpojumu centrs (VPC) un datu kuratoru projekts
- *Maciej Ogrodniczuk (CLARIN ERIC)*. ParlaMint: Multilingual Comparable Corpora of Parliamentary Debates for Digital Humanities
- *Edward Gray (DARIAH EU)*. The Digital Research Infrastructure for the Arts and Humanities: How DARIAH Helps Enable SSH Researchers
- *Inguna Skadiņa (LU MII)*. CLARIN-LV - solis pretī valodu digitālai vienlīdzībai
- *Anta Trumpa un Sanda Rapa (LU LaVI)*. Valodas digitalizācija: no pirmajiem rakstu avotiem līdz mūsdienām
- *Anda Baklāne (LNB)*. Digitālo pakalpojumu aktualitātes Latvijas Nacionālajā bibliotēkā
- *Antra Kļavinska (RTA)*. Latgaliešu valodas digitālie resursi: paveiktais, kļūšanas akmeņi un iespējas
- *Sanita Reinsone un Sandis Laime (LU LFMI)*. Humma.lv – ceļā uz digitālo platformu humanitārajām un mākslas zinātnēm
- *Valts Ernštreits (LU Lībiešu institūts)*. Lībiešu valodas datu ieguve un tās nākotnes perspektīvas
- *Baiba Saulīte un Ilze Auziņa (LU MII)*. Tēzauris.lv un Korpusis.lv jaunumi

Seminārs "Runas korpusu izveide un to izmantojums"

2023. gada 18. maijā notika digitalhumanities.lv seminārs sadarbībā ar Clarin-LV un DARIAH-EU "Runas korpusu izveide un to izmantojums". Semināra materiāli:

- Ilze Auziņa. Runas korpusu nozīme valodu tehnoloģiju izstrādē.
- Antra Kļavinska. Mūsdienai latgaliešu runas korpusa izveide mazāk lietoto valodu dokumentēšanas kontekstā.
- Ilze Auziņa un Guna Rābante-Buša. Runas datu apstrāde programmā ELAN.



Citi būtiski CLARIN piedāvājumi

Learning Hub

- Learning and Training Resources
- Training and Workshops
- Digital Humanities Course Registry
- Notebooks in Education

[Learning Hub | CLARIN ERIC](#)

Funding Hub

- Workshop Funding
- User Involvement Funding
- Teaching with CLARIN
- Trainer Network Programme
- Mobility Grants
- Resource Families Project Funding

[CLARIN Funding Hub | CLARIN ERIC](#)

CLARIN konference

- iespēja piedalīties ikvienam konsorcijs partnerim (atbalsts referātiem)
- Īpašs atbalsts doktorantiem



Konsorcijs



Paldies par uzmanību!



VPP-LETONIKA-2021/1-0006



Finansē
Eiropas Savienība
NextGenerationEU



2.3.1.1.i.0/1/22/I/CFLA/002