Disseminating and Teaching CLARIN in the time of Corona. Challenges and opportunities

Francesca Frontini

CLARIN ERIC & ILC-CNR

Latvian Language Digital Resources And Tools For Education

June 2, 2021



June 2nd 1946

Outline

- What is CLARIN
- What is the CLARIN Knowledge infrastructure
- What it offers to researchers …
- … and teachers, lecturers and educators

CLARIN in eight bullets

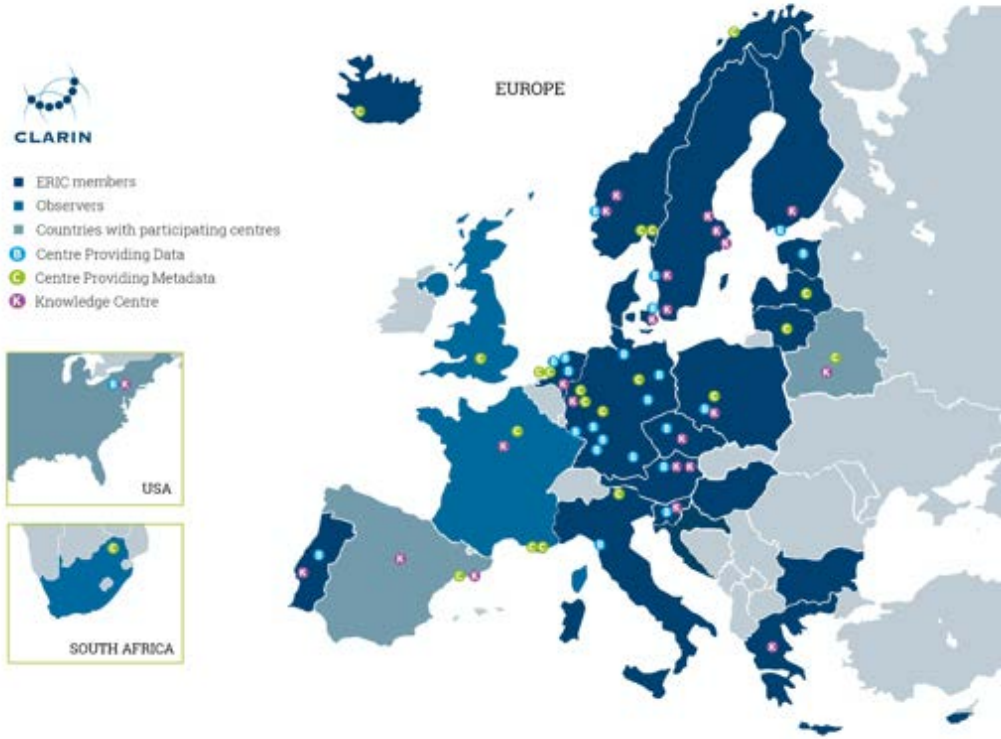
- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** environment
- that serves as an ecosystem for **knowledge exchange**
- and: ready for **integration in EOSC** (European Open Science Cloud; link)

CLARIN Value Proposition: https://www.clarin.eu/content/value-proposition (link to pdf)

CLARIN ERIC in members and centres

A consortium of:

- 21 members: AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, **LV**, NL, NO, PL, PT, SE, SI
- 3 observers: FR, UK, ZA
- 1 linked party: CMU
- >60 centres (incl. 24 CTS certified data centres)
- Support for linguistic diversity
  - Data covering more than 1500 languages
  - Tools for many languages
  - Language resources in all modalities

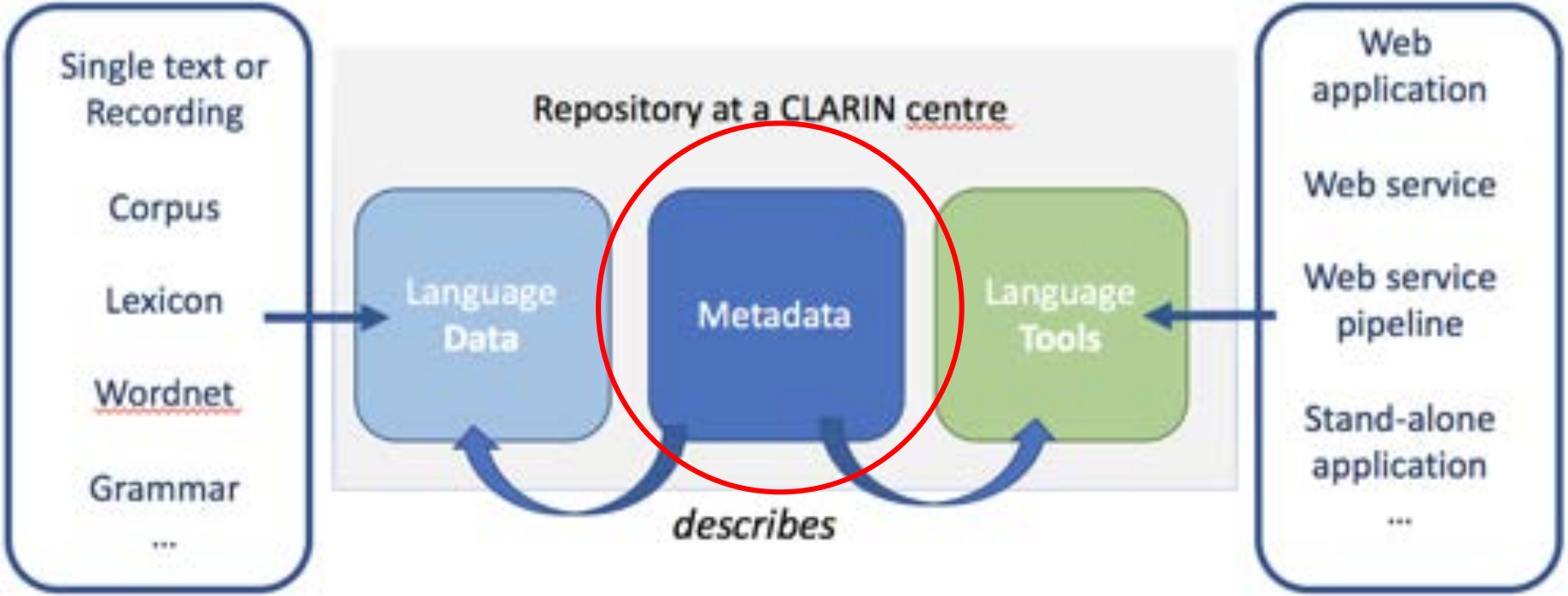


CLARIN data types and communities of use in SSH and beyond

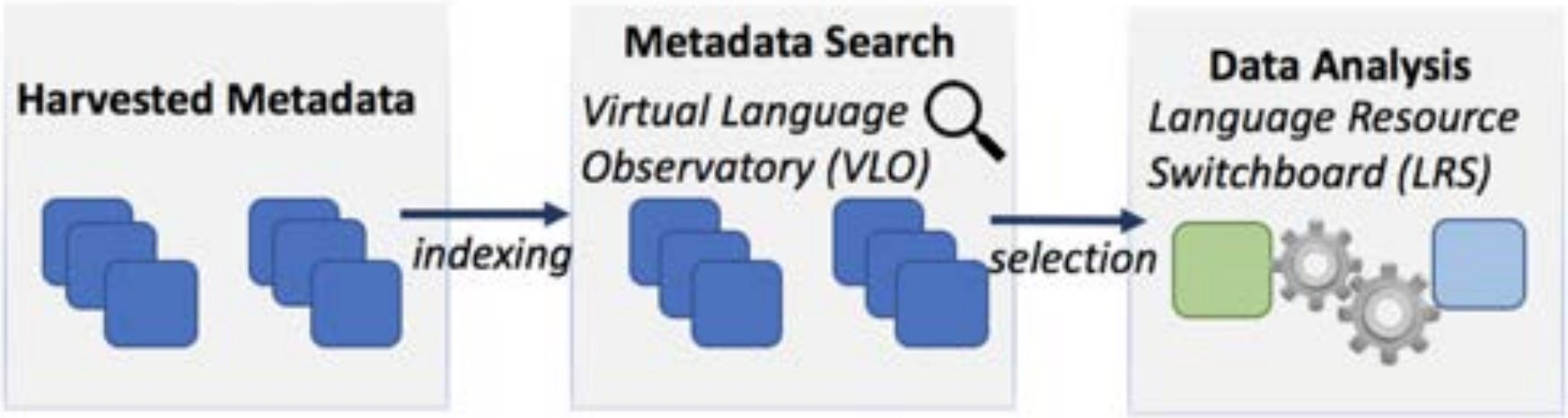
- Newspaper archives
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- Patient recordings
- Excavation reports

- Digital humanities
- Linguistics and Philology
- Data Science /AI
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public

The CLARIN data architecture:
a network of distributed centre repositories



The CLARIN data architecture:
central processing of metadata

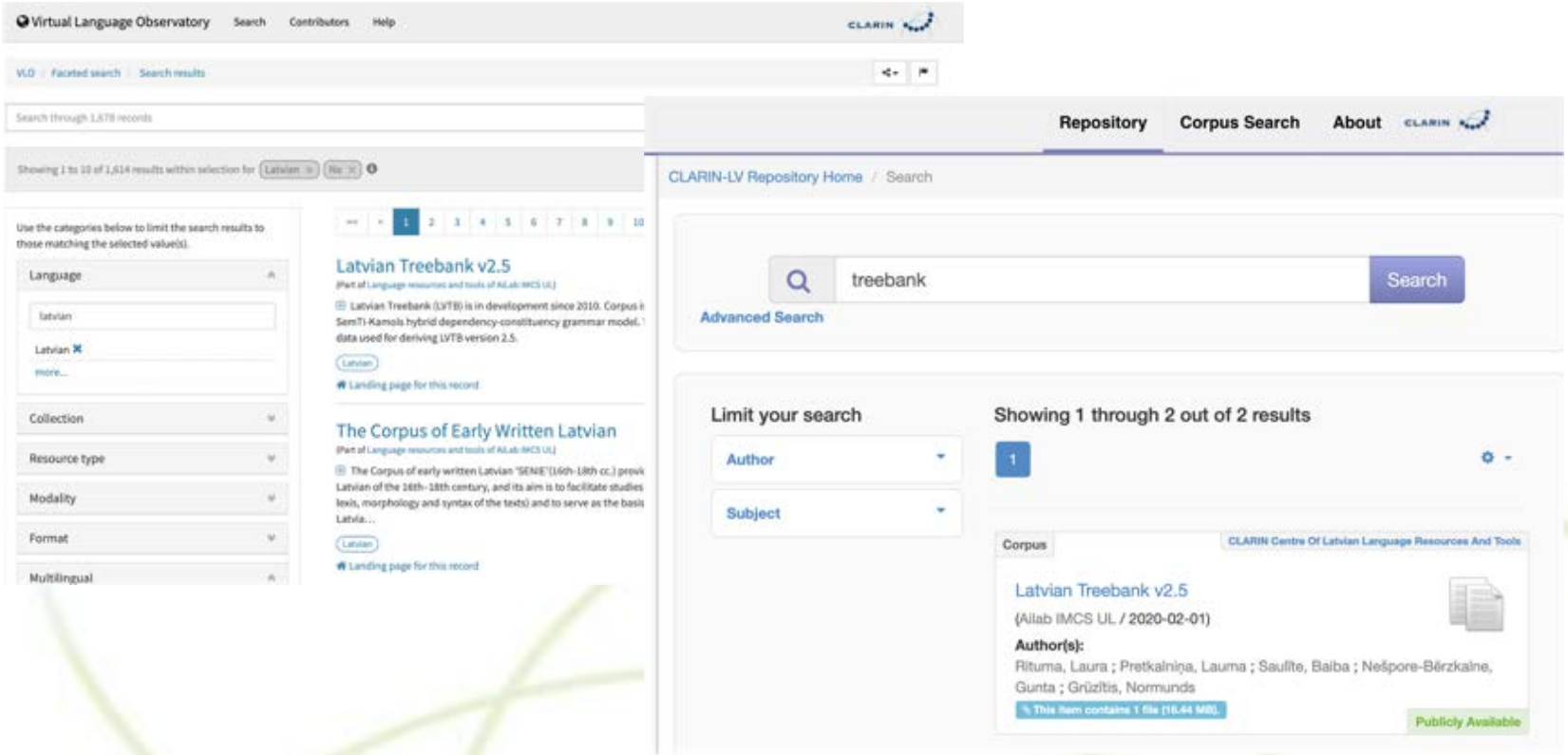


Findable Accessible Interoperable Reusable

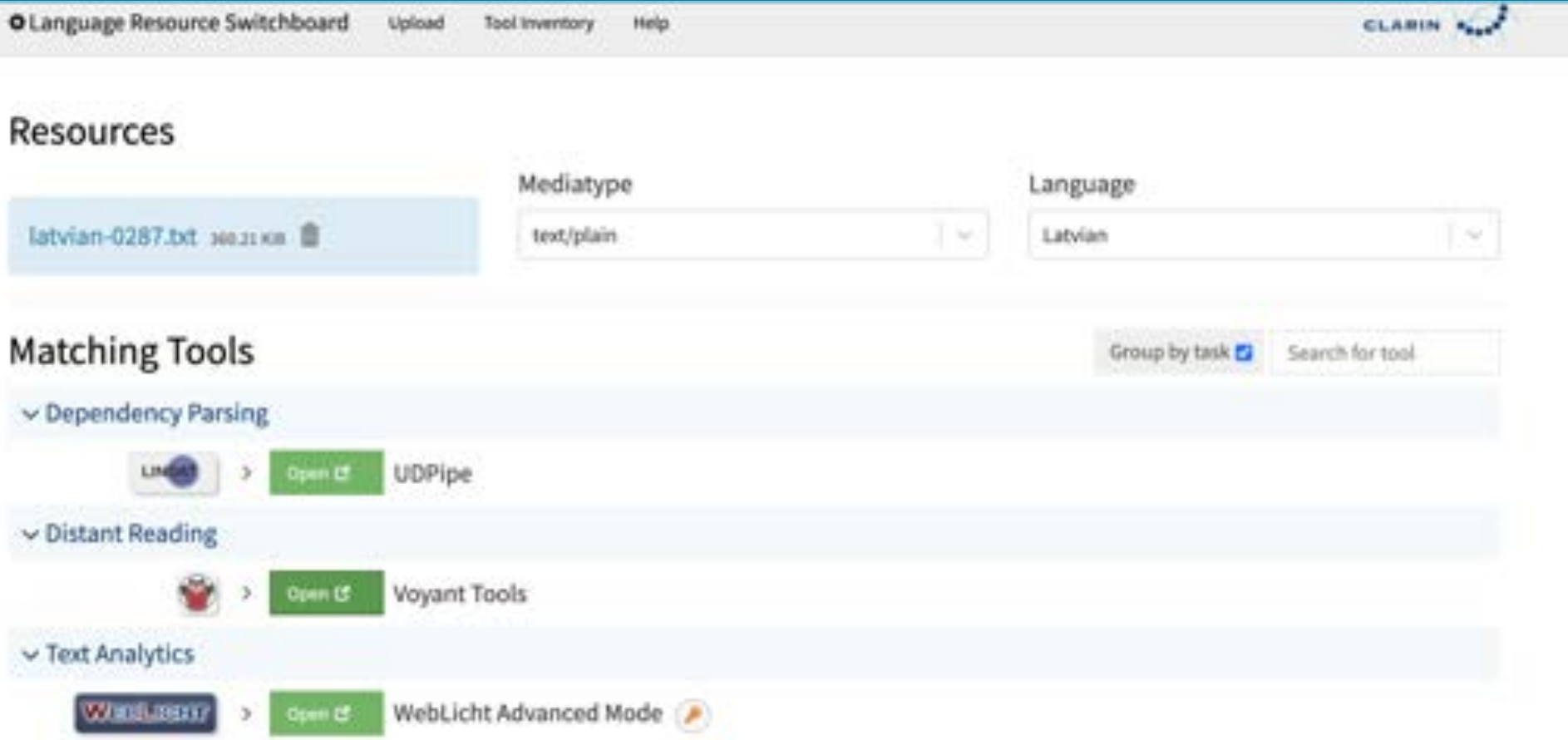
clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu



vlo.clarin.eu



switchboard.clarin.eu

CLARIN as ecosystem for knowledge exchange



CLARIN Knowledge Centres (abbreviated K-centres) are a cornerstone of the CLARIN knowledge infrastructure. K-centres are institutions that have agreed to share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure.

CLARIN as ecosystem for knowledge exchange



- Individual languages (e.g. Danish, Czech, Portuguese), language families (e.g. South Slavic) or groups of languages (e.g. morphologically rich languages, the languages of Sweden)
- Written text and modalities other than written text (e.g. spoken language, sign language)
- Linguistic topics (e.g. language diversity, language learning, diachronic studies)
- Language processing topics (e.g. speech analysis, building treebanks, machine translation)
- Data types other than corpora (e.g. lexical data, word nets, terminology banks)
- Using or processing families of language data that will exist for most languages (e.g. newspapers, parliamentary records, oral history)
- Generic methods and issues (e.g. data management, ethics, IPR, OCR)

CLARIN as ecosystem for knowledge exchange



In collaboration with DARIAH ERIC.
The DH Course Registry contains a selection of DH courses offered by European academic organizations.

DIGITAL HUMANITIES
COURSE REGISTRY

STORIES INFO LOGIN

FILTER SORTING

NAME	UNIVERSITY	PLACE	DATE	TYPE
Digital Humanities	Universität Basel	Basel, Switzerland	20 Sept 2 years	Master Programme
Data Science with Minor in Digital Humanities	Friedrich-Alexander-Universität Erlangen-Nürnberg	Erlangen, Germany	18 Oct 6 semester	Bachelor Programme
Master Mittelalter und Frühe Neuzeit, Vertiefung Digital Hu ...	**Friedrich-Alexander-Universität Erlangen-Nürnberg**	**Erlangen, Germany**	**18 Oct 4 semester**	**Master Programme**
Digital Humanities	Friedrich-Alexander-Universität Erlangen-Nürnberg	Erlangen, Germany	15 Oct 4 semester	Master Programme
Digitale Geistes- und Sozialwissenschaften	Friedrich-Alexander-Universität Erlangen-Nürnberg	Erlangen, Germany	1 Oct 6 semester	Bachelor Programme
Linguistic and Literary Computing	Technische Universität Darmstadt	Darmstadt, Germany	1 Oct 4 semester	Master Programme
Introduction to Computational Linguistics (DHS)	University of Latvia	Riga, Latvia	5 Sept 2021 1 semester	Course
Cultural Data Studies	Philipps-Universität Marburg	Marburg, Germany	1 Oct 4 semester	Master Programme
Digital Humanities and Social Sciences	University of Latvia	Riga, Latvia	1 Sept 2021 1 semester	Course
Fakultätsweites Basismodul Digitale Geisteswissenschaften ...	Karl-Franzens-Universität Graz	Graz, Austria	3 Oct 1 semester	Course
Doktoratsstudium "Digitale		Graz	1 Oct	



Imprint ©2014-2021 CC-BY 4.0

CLARIN as ecosystem for knowledge exchange





CLARIN as ecosystem for knowledge exchange



Get to know the different types of resources in different languages CLARIN makes available for researchers from digital humanities, social sciences and human language technologies.

https://www.clarin.eu/resource-families

Dictionaries in the CLARIN infrastructure

Size: 83,000 entries **Licence:** CC-BY 4.0		download and for online browsing through the Language Bank of Finland.	Download
Dictionary of Modern Latvian (MLV) **Size:** 45,877 entries **Licence:** browsable	Latvian	The resource is available for online browsing.	Browse
Dictionary of Standard Latvian Language (LLVV) **Size:** 64,200 entries **Licence:** browsable	Latvian	This is a dictionary that contains examples of the standard Latvian language from the 1870s to the 1990s. The resource is available for online browsing.	Browse
Historical Dictionary of Latvian (16th-17th Centuries) (LVVV) **Size:** 1,800 entries **Licence:** Terms of Use	Latvian	This is a dictionary of Latvian from the period between the 16th and 17th centuries with a reference to the "Historical Corpus of the Latvian Language". The resource is available for online browsing. For a related publication, see Andronova et al. (2016).	Browse
Rendering Dictionary of Personal Names **Size:** 4,000 entries **Licence:** Terms of Use	Latvian	This dictionary contains person names in four languages (Czech, French, Italian and German) and their renderings in Latvian in accordance with the rendering rules of proper nouns. The resource is	Browse

Table of Contents

Introduction

Dictionaries in the CLARIN infrastructure

Monolingual resources

Multilingual resources

Publications

CLARIN as ecosystem for knowledge exchange







CLARIN Going Virtual

- 11 CLARIN cafés
- Support for virtual events
- 1 virtual conference (with another coming up)
  - Largest attendance ever!





https://www.clarin.eu/content/clarin-responses-covid-19

https://www.clarin.eu/events



Reaching more communities and researchers

- More virtual events and asynchronous access (recordings)
- More opportunities to engage with CLARIN (no limitations due to travel, lack of funding)
  - Lecturers
  - Early career researchers
  - Students



New website, tomorrow!



CLARIN in the classroom



In order for CLARIN to have a lasting impact, it is necessary for both methodological and infrastructural issues to become integrated in university curricula.



CLARIN in the classroom



Survey 22 member and observer countries (2008-2020)

CLARIN used in 55 individual courses, already taught at 31 different universities in 20 countries



What teachers think of CLARIN - Advantages

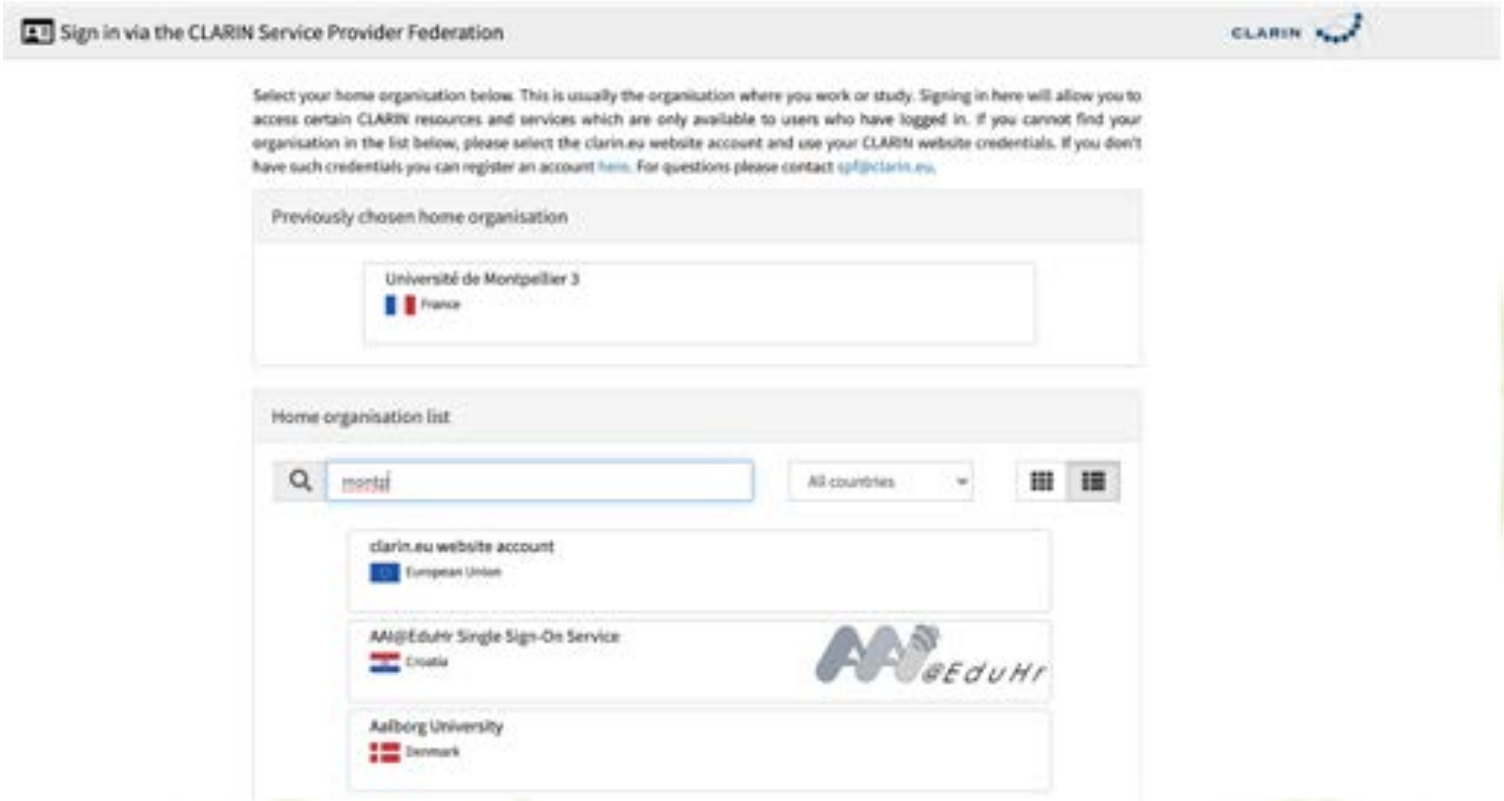
- Accessibility
  - *”The interfaces are usable, reliable, state of the art; students like the idea that they can access services that are hosted in other countries with their university identifier.”*
- Findability of resources
  - *”Search the VLO for concrete corpora and to show the power of well curated metadata.”*
- Reuse of resources
  - *”Students use CLARIN corpora and tools in their theses, projects and internships with companies.”*
- Archiving of teaching materials
  - *”The integration of CLARIN content in teaching motivated some of the lecturers to archive their materials and make them accessible to the research community and the general public.”*
- Support for research activities
  - *”A student who explored Fin-CLARIN said that it is paradise for researchers."*

Reusable, integrated online services from the Switchboard



Weblicht, a lifeline when courses moved online.

Single sign on, no need to install software



Key issues

- Lack of resources and/or tools for a specific language
  - *"Lack of resources for French, in particular of corpora and NLP modules."*
- Lack of standardization
  - *"Lack of standardization that hinders reuse of resources and tools."*
- Lack of teacher expertise
  - *"There is a potential to use lots of other tools that I am insufficiently familiar with. There is insufficient documentation of available resources and no tutorial on how to effectively use them in the classroom."*
- Lack of training materials
  - *"Lack of time and other resources to locate or create materials for all potentially relevant technologies."*
- Too difficult for students
  - *"Many corpora are in TEI, and there are not many tools which allow graphical exploration or analysis of TEI corpora."*

Working towards a CLARIN Training Suite





First workshop – November 2019

Home > Workshop: CLARIN@universities

Workshop: CLARIN@universities

Monday, November 4, 2019 - 12:00 to Tuesday, November 5, 2019 - 13:30

The CLARIN@universities workshop will take place on 4 and 5 November 2019 in Utrecht (lunch-to-lunch).

This workshop on the experience with the integration of CLARIN content into university programmes is organised by the Knowledge Infrastructure Committee.

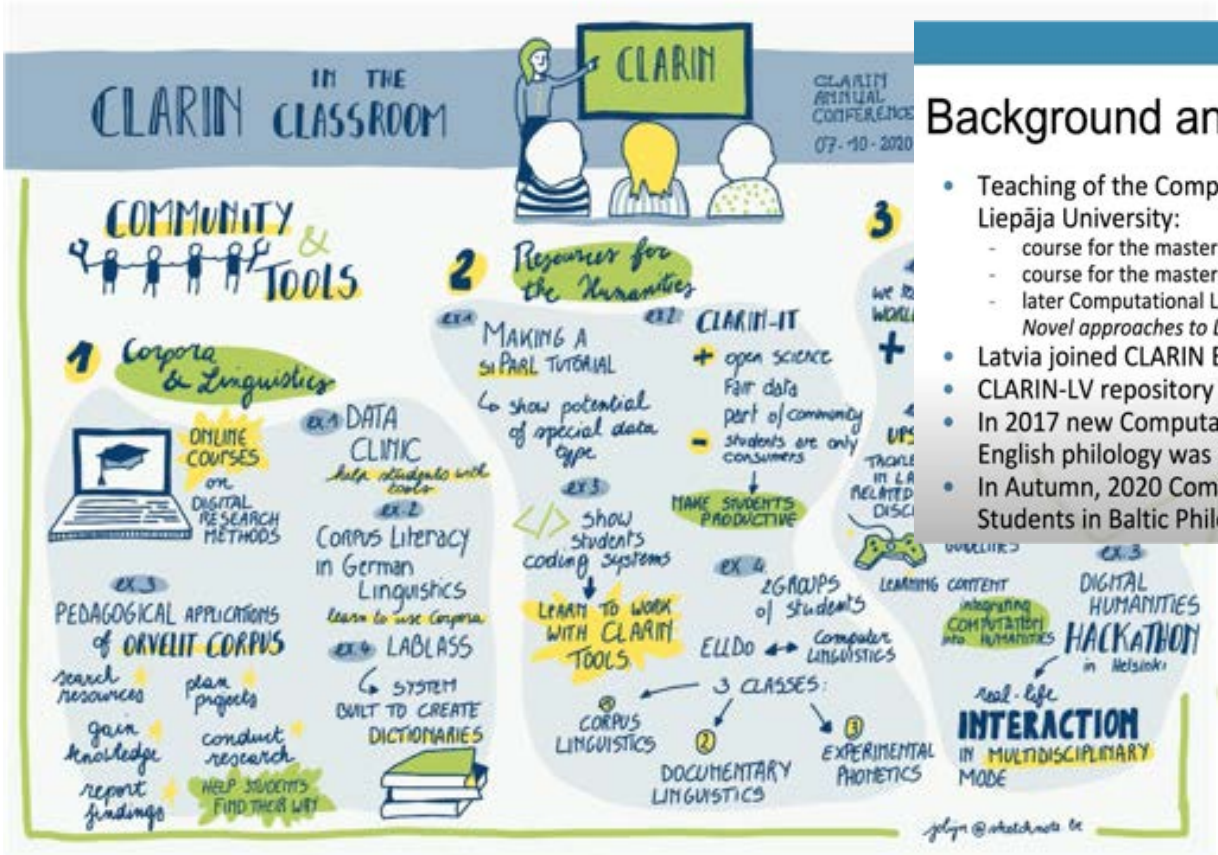
Background | Approach | Workshop participants | Workshop programme | Workshop materials | Getting there

Important note for those travelling from Schiphol to Utrecht between 1 and 3 November: Between Friday 1 November and Sunday 3 November, the railway company (NS) is working on the tracks, and the direct train connection from Schiphol to Utrecht Central Station cannot be taken. Instead you need to take a train from Schiphol to Amsterdam Central Station and from Amsterdam Central Station you can take the intercity train to Utrecht Central Station.

CLARIN café on higher education, June 2020



CLARIN in the Classroom (October 2020)



Background and Current Activities

- Teaching of the Computational Linguistics course started in 2003 at Liepāja University:
  - course for the master students in linguistics
  - course for the master students in computer science
  - later Computational Linguistics was included in a course for doctoral students *Novel approaches to Linguistics* at Liepāja University
- Latvia joined CLARIN ERIC in 2016
- CLARIN-LV repository has been registered in March, 2020
- 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

Training officer

Introducing Iulianna van der Lek: New Training and Education Officer CLARIN ERIC

CLARIN *ERIC* is pleased to welcome Iulianna van der Lek as the new Training and Education Officer.

Iulianna will work closely with the CLARIN community to develop the CLARIN Training Suite and the Trainer Network. She will collaborate with university lecturers from the humanities and social sciences disciplines to accelerate the further integration of CLARIN into university curricula. In addition, she is coordinating CLARIN's activities in the UPSKILLS project, an ERASMUS+ initiative that aims at upgrading the skills of Linguistics and Language Students.

Read more



The Upskills project

https://upskillsproject.eu/



Need analysis seeks to develop a detailed professional profile for graduates of linguistics and language-related courses, with the associated learning objectives (knowledge, skills and competences), typical tasks and responsibilities

More comparable resources



Parliamentary corpora in the CLARIN infrastructure

Corpus	Language	Description	Availability
Multilingual comparable corpora of parliamentary debates ParlaMint 2.0 **Size:** 3.3 million utterances, 462 million words **Annotation:** tokenised, MSD-tagged, syntactically parsed (Universal Dependencies) **Licence:** CC BY 4.0	Bulgarian, Croatian, Czech, Danish, Dutch, English, French, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Polish, Slovenian, Spanish, Turkish	ParlaMint is a multilingual set of comparable corpora containing parliamentary debates mostly starting at the end of 2015 and extending to mid-2020, with each corpus being about 20 million words in size. The sessions in the corpora are marked as belonging to the COVID-19 period (after October 2019), or being "reference" (before that date). The corpora have extensive meta-data about the speakers (name, gender, party affiliation, MP status), are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches also contain marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc. The corpus is available for download from the CLARIN.SI repository and through the concordancer noSketchEngine.	Concordancer Download
Croatian parliamentary corpus ParlaMeter-hr 1.0 **Size:** 14.1 million tokens **Annotation:** tokenised, MSD-tagged, lemmatised, named	Croatian	The corpus contains minutes of the National Assembly of the Republic of Croatia and currently covers its VIth mandate from 15 November 2016 to 21 November 2018. The corpus contains speaker metadata (gender, age, education, party affiliation). The corpus is available for download from the CLARIN.SI repository and through the concordancers KonText and	Concordancer Download

Table of Contents

- Introduction
- Parliamentary corpora in the CLARIN infrastructure
- Other parliamentary corpora
- CLARIN-funded project: ParlaMint
- Additional materials
- Publications on the parliamentary corpora

Parliamentary data are a for comparative study of public debate.

https://www.clarin.eu/resource-families/parliamentary-corpora

More comparable resources

- Fill gaps in Resource Families



Parliamentary corpora in the CLARIN infrastructure

Corpus	Language	Description	Availability
Multilingual comparable corpora of parliamentary debates ParlaMint 2.0 **Size:** 3.3 million utterances, 462 million words **Annotation:** tokenised, MSD-tagged, syntactically parsed (Universal Dependencies) **Licence:** CC BY 4.0	Bulgarian, Croatian, Czech, Danish, Dutch, English, French, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Polish, Slovenian, Spanish, Turkish	ParlaMint is a multilingual set of comparable corpora containing parliamentary debates mostly starting at the end of 2015 and extending to mid-2020, with each corpus being about 20 million words in size. The sessions in the corpora are marked as belonging to the COVID-19 period (after October 2019), or being "reference" (before that date). The corpora have extensive meta-data about the speakers (name, gender, party affiliation, MP status), are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches also contain marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc. The corpus is available for download from the CLARIN.SI repository and through the concordancer noSketchEngine.	Concordancer Download
Croatian parliamentary corpus ParlaMeter-hr 1.0 **Size:** 14.1 million tokens **Annotation:** tokenised, MSD-tagged, lemmatised, named	Croatian	The corpus contains minutes of the National Assembly of the Republic of Croatia and currently covers its VIth mandate from 15 November 2016 to 21 November 2018. The corpus contains speaker metadata (gender, age, education, party affiliation). The corpus is available for download from the CLARIN.SI repository and through the concordancers KonText and	Concordancer Download

Table of Contents

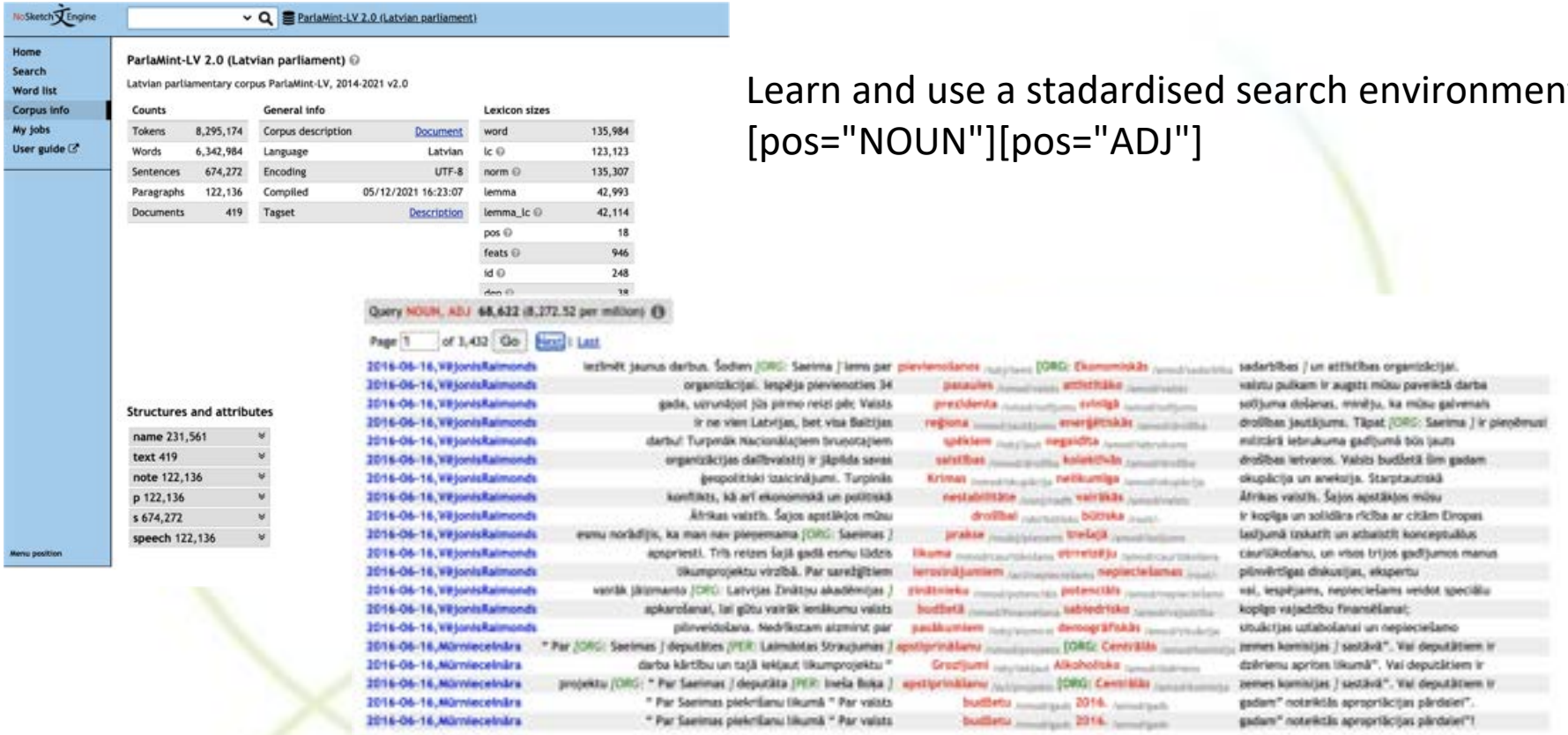
- Introduction
- Parliamentary corpora in the CLARIN infrastructure
- Other parliamentary corpora
- CLARIN-funded project: ParlaMint
- Additional materials
- Publications on the parliamentary corpora





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

ParlaMint data, Reusable by Design



Learn and use a stadardised search environment

[pos="NOUN"][pos="ADJ"]



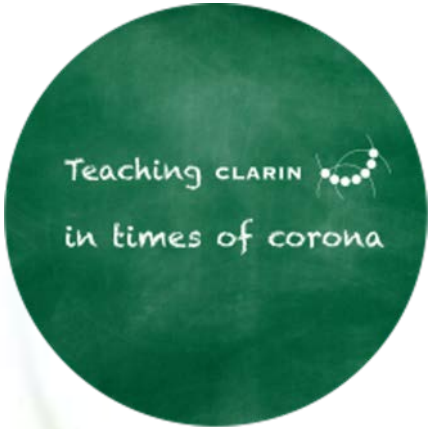
Yankelevich, Tanya, Fiser, Darja, Lenardic, Jakob, Gorgaini, Elisa, & Braukmann, Ricarda. (2020, June). **LIBER 2020 - Workshop: SSHOC Train-the-Trainer Bootcamp for Librarians.** Zenodo. http://doi.org/10.5281/zenodo.3970799

Slovenian,Croatian → Italian, English, Polish

Helsinki DH Hackathon 2021
https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/

Frontini, Francesca, 2021.
Presentation on CLAIRN at **Pisa summer school**.

"Teaching CLARIN in Times of Corona"
Deadline ~~01/06/2021~~ 15/06/2021



All relevant submissions will be showcased on the CLARIN website, for the purpose of sharing best practices and examples suitable for reuse. In addition, three submissions will be selected for the Teaching CLARIN award, and will receive the sum of 350 euros each. The award ceremony will be hosted during CLARIN Annual Conference 2021.

Getting involved in CLARIN

- Join our NewsFlash
  - https://www.clarin.eu/content/newsflash
- Check out our events
  - https://www.clarin.eu/events
- Open calls
  - https://www.clarin.eu/content/funding-opportunities

- Follow us on Twitter @CLARINERIC
- And stay tuned for the next cafés
  - https://www.clarin.eu/content/clarin-cafe
  - **ParlaMint Café - 28th June 14:00-16:00 CEST**

www.clarin.eu
or
francesca.frontini@ilc.cnr.it

