



Co-funded by the  
Erasmus+ Programme  
of the European Union



# CLARIN & UPSKILLS

By Iulianna van der Lek and Darja Fišer,  
CLARIN ERIC



The Second Swiss UPSKILLS Multiplier  
Event, Geneva, 21-22 April 2023

# Outline

- Why UPSKILLS & CLARIN need each other
- UPSKILLS Learning Content: *Introduction to Language Data: Standards and Repositories*

# Why UPSKILLS and CLARIN Need Each Other

- Skills and knowledge about **linguistic research data repositories and related standards** were **not explicitly mentioned** in the learning outcomes of the analysed European language and linguistics degrees
- The data and services offered by **research infrastructures** (e.g. CLARIN) are **seldom** used in teaching of language-related disciplines

*Survey of lecturers in language-related programmes (May-July 2021)*

# Why UPSKILLS and CLARIN Need Each Other

## Data Discovery

- **Little usage of repositories** to find published language resources to use in teaching
  - E.g. institutional, CLARIN national repository, Linguistic Data Consortium, OPPUS, Corpora Mailing List, Language Resource Families, DGT Translation Memories, ELRA, Meta-Share

# Why UPSKILLS and CLARIN Need Each Other

## Data Storage

- Dropbox, Google drive
- Own computer

## Data Archiving

- Institutional repository
- Domain-specific repositories:
  - CLARIN national repository, Meta-share, ELRA Catalogue, The Language Archive, Language Data Consortium
- General repositories: Zenodo, FigShare
- Github
- **Moodle**

# Why UPSKILLS and CLARIN Need Each Other

## Challenges

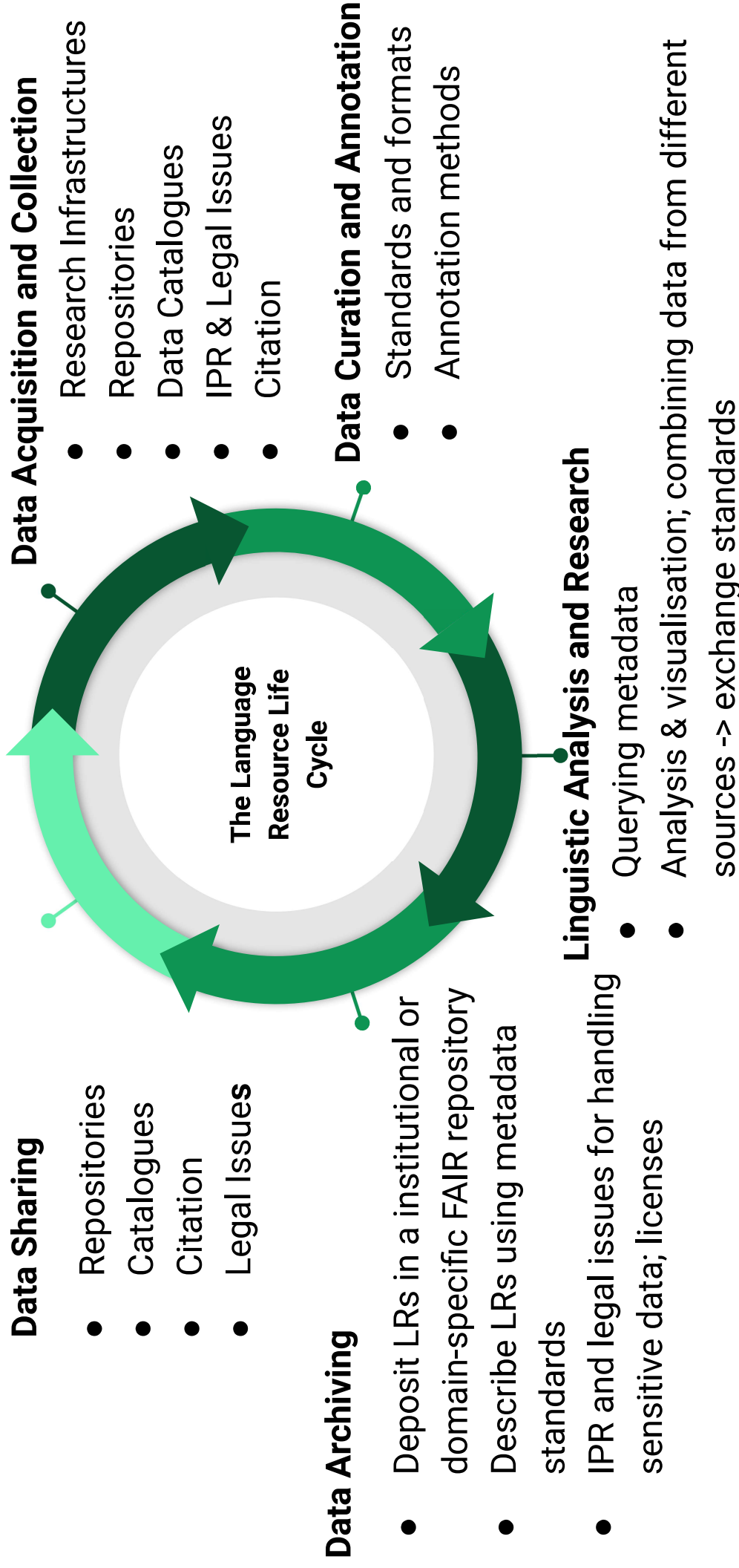
- Technical challenges, limited space, little IT support
  - Administrative load and costs
  - Issues with IPR, students need **help with the interpretations of the legal requirements**
  - Protection of **data privacy** in the case of spoken language and multilingual recordings
  - Students' **low level of digital literacy**
-

# Why Use Research Infrastructures

## Advantages

- By interacting with research infrastructures and repositories (e.g CLARIN, DARIAH, Meta-Share), **both teachers and students** gain awareness of **open science, FAIR data guiding principles, research data management** and may engage in collaborative research, teaching and training
- May open new career paths for students, e.g. **language data manager** (clean, curate and manage data) or **data steward** supporting LR and LT research and development

# Why Use Research Infrastructures





# Example - Research

## Barack Obama's identity-building in the health care debate: A corpus-assisted discourse study

### AUTHOR

[Katherina Riesner](#)

### Summary, in English

In this study, I demonstrate that identity-building is an important discursive strategy for President Barack Obama in the seven-year long debate surrounding the Affordable Care Act (ACA). The data for the study comes from a 6-million word corpus of speeches that were held by Obama between January 2009 and January 2016, all published by the White House. The speeches are classified according to genre, audience, topic and date of delivery. Throughout the paper, I adopt the notion that identity is intentionally constructed by the speaker and strategically exploited for his communicative goals. With the help of two methodological approaches, I investigate what kind of identities Obama builds. The purely qualitative part of the study deals with three central corpus speeches from a discourse-analytic perspective. In the second, more quantitative part, I use a group of seven verbs with epistemic meaning to trace the usage of two predominant discursive identities in the ACA debate. The results suggest that President Obama repeatedly constructs the identities of father and teacher to persuade his audience. I argue that his use of these identities constitutes an attempt to reach the argumentative goals of effectiveness and reasonableness.

### Department/s

Master's Programme: Language and Linguistics

### Publishing year

2016

### Language

English

### Full text

[Available as PDF](#) - 2 MB

[Download statistics](#)

### Document type

Student publication for Master's degree (two years)

### Topic

Languages and Literatures

- IMDI Corpora
- Lund Corpora
- Eline Visser
- ESST
- Eye-Tracked Frog Stories
- LACOLA
- LANG-KEY
- LUNDIC
- REaCHes
- SpaceH
- Strömqvist-Richthoff
- Swedia2000
- Tactile Reading
- Test
- ThaiSweVideo
- The Barack Obama Corpus**
- the\_barack\_obama\_corpus\_information.txt
- 2009
- 2009
- 2010
- 2010
- 2011
- 2011
- 2012
- 2012
- 2013
- 2013
- 2014
- 2014
- 2015
- 2015
- 2016
- 2016
- USE
- VOKART

METADATA SEARCH    CONTENT SEARCH    MANAGE ACCESS

REQUEST ACCESS    CITATION

BOOKMARK

**Corpus**

**Name** The Barack Obama Corpus  
**Title** The Barack Obama Corpus

**Description**  
the\_barack\_obama\_corpus\_information.txt

**Description**  
The Barack Obama Corpus (BOC) consists of 6,215,948 words (tokens), which are sourced from nearly 3,500 different texts, dating from January 2009 to January 2016. The texts, all taken from the White House Archives, comprise all speeches held by Barack Obama in his official capacity as 44th President of the United States of America. The earliest speech in the BOC is President Obama's inauguration speech and the last is his final State of the Union speech (January 2016). In total, the corpus includes 34,967 word types, which leads to a type/token-ratio of 0.56.

The files, which display the original titles given to them by the White House, have been tagged for genre, audience type, date and location of delivery, and principal topics. The genres include remarks, addresses, statements, press conferences, debates and question-

**Description**  
How to cite this resource:  
Riesner, Katherina (2017). The Barack Obama Corpus [Data set]. <http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4#view>

## Appropriate data citation and PID

Riesner, Katherina (2017). The Barack Obama Corpus [Data set].

<http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4#view>



Search bar containing the text 'obama' and a search icon.

Showing 5 results for obama x Results per page: 10

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

- Language >>
- Collection >>
- Modality >>



### The Barack Obama Corpus

(Part of Lund University Humanities Lab)



the\_barack\_obama\_corpus\_information.txt; The Barack Obama Corpus (BOC) consists of 6,215,948 words (tokens), which are sourced from nearly 3,500 different texts, dating from January 2009 to January 2016. The texts, all taken from the White House Archives, comprise all speeches held by Barack Obama in his official capac...

Landing page for this record at [corpora.humlab.lu.se](http://corpora.humlab.lu.se)



metadata

[vlo.clarin.eu](http://vlo.clarin.eu)

# Example - Teaching

Erjavec, Tomaž; et al., 2021, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1431>

- Findable via CLARIN VLO and CLARIN.SI
- Persistent Identifier, handle
- Described with consistent metadata
- Well-documented
- Can be cited
- Reference to a publication
- Available for exploration in integrated infrastructure tools: KonText and NoSketchEngine
- Licensed under CC-BY-4.0
- The files can be downloaded

FAIR Repository  
FAIR Corpora

# Example - Teaching

## What's on the agenda? Topic modelling parallel corpora before and during pandemic

### Goals and Objectives

The main goal of this tutorial is to introduce applying the Latent Dirichlet Allocation

### Learning Outcomes

By following this tutorial, the students will be able to:

- independently perform topic modeling on parallel corpora;
- understand the pitfalls of topic modeling.

### Author(s)

Ajda Pretnar Žagar  
Researcher

### TABLE OF CONTENTS

CLARIN.SI repository / View Item

Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1

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

Erjavec, Tomaž; et al., 2021, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1431>.

**Integrate** This resource is also integrated in following services:

**KonText** **noSketch**

**Authors** Erjavec, Tomaž; et al.  
▶ show everyone

**Item identifier** <http://hdl.handle.net/11356/1431>

**Project URL** <https://www.clarin.eu/content/parlamint/>

**Demo URL** <https://github.com/clarin-eric/ParlaMint/>

**Referenced by** <https://doi.org/10.1007/s10479-021-09574-0>

**Date issued** 2021-06-18

Search

Q



**Browse**

> All of the Repository

**My Account**

Login

Statistics

Plwik Statistics

BETA

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

CLARIN.SI Data & Tools

BIBTEX

CMDI



Share: **f** **t**

## What's on the agenda tutorial

# CLARIN IN UPSKILLS

- Tasks
- *Demo Introduction to Language Data: Standards and Repositories*

---



# Tasks

- Needs Analysis
- Include specific RI-related learning outcomes into the **RBT guide**
- **Quick Guide** to CLARIN: how to use the main services and where to find relevant information (optional)
- Two courses:



# CLARIN in UPSKILLS

- **Guidelines for Students' Projects** (available soon on Zenodo) and showcase projects
  - Collaboration with **UniBo** to integrate the infrastructure into the students' projects
    - a. How to find and query existing corpora
    - b. How to archive corpora in a repository
- > Launched an **internship programme**



# CLARIN Internships – Meet the Interns

Submitted by Julia Misersky on 22 January 2023

CLARIN has launched a remote internship programme for students interested in learning how to use the research infrastructure to access and engage with digital language data with the help of advanced technologies. The internship aims to help the students enhance their language data handling and processing skills, and show them how to manage language resources according to the FAIR Data Principles.

Our first two interns are Elton Pistollia and Lesley Messori, who are second-year Translation Technology MA students at the University of Bologna.



**Elton Pistollia**

Elton has a BA in Foreign Languages for Tourism and International Mediation and is now in his second year of an MA in Translation Technology at the University of Bologna. He has a strong background in linguistics and translation and is keen to develop his programming and machine-learning skills. He speaks Italian, English, Greek, Albanian, German, Russian and Portuguese. Elton was selected as the best participant for the language combination EN-EL in the eMT Challenge 2022. After his MA, Elton plans to do a PhD in machine translation.



**Lesley Messori**

Lesley has a BA in Linguistic and Intercultural Mediation and is now in her second year of the MA in Translation Technology programme at the University of Bologna. During her MA, she acquired knowledge of corpus linguistics, terminology management, translation technology tools and learnt how to adapt machine translation systems. Lesley speaks five languages, Italian (mother tongue), English, Russian, Danish and Slovak and has experience in audiovisual translation, translation and post-editing. In the future, she sees herself working as a freelance translator because she simply loves anything related to languages.

**Project:** *Improving the discoverability of the language resources in the Virtual Language Observatory*

- GitHub Notebook to parse the XML files from the VLO, detect languages and NER with spaCy
- Create a gold standard to evaluate the VLO metadata set
- Annotation of named entities
- Translation
- Sharing, depositing and archiving in CLARIN-IT

**Project:** *CLARIN Vocabulary to tag web content and academic outputs*

- Corpus & extraction of terms candidates using SketchEngine
- Term validation guidelines & 3 external validators
- Selection of final terms, concept mapping, definitions and translations
- SKOS and TBX formats
- Sharing, depositing and archiving in CLARIN-IT

# CLARIN in UPSKILLS

- **Support with dissemination of the UPSKILLS materials**
  - Overview of public and open repositories and registries

The screenshot displays the SSH OpenCluster website interface. At the top, there is a navigation menu with links for Home, Resources, Curricula, Topics, Sources, Course registry, and Documentation. Below the menu, a search bar is visible with the text "Search for sources". A prominent blue button labeled "Explore" with a right-pointing arrow is positioned below the search bar. The main content area features a large banner with the text "Explore. Create. Collaborate." and a sub-header "SSH Training Discovery Toolkit". The banner also includes the text "OER Commons is a public digital library of open educational resources. Explore, create, and collaborate with educational world to improve curriculum." and a search bar with filters for Subject, Education Level, and Standard. The banner is overlaid with a semi-transparent dark box containing the text "Look DARIAH-CAMP".

The screenshot displays the zenodo website interface. At the top, there is a navigation menu with links for Sign in/Register, Upload, and Communities. Below the menu, a search bar is visible with the text "Search". A prominent blue button labeled "zenodo" is positioned below the search bar. The main content area features a large banner with the text "Featured communities" and a sub-header "Transform to Open Science". The banner also includes the text "Transform to Open Science (TOPS) is a \$40 million, 5-Source Science initiative. Within the TOPS mission, N Initiative to spark change and inspire open science..." and "curated by: nasatransformtoopen". The banner is overlaid with a semi-transparent dark box containing the text "Look DARIAH-CAMP".

# Guide

## CLARIN in the CLASSROOM

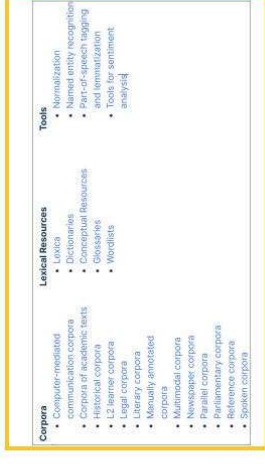
### A Guide for Teachers and Students



1 Find and(re)use published language resources



2 Collect, cite and share collections of virtual resources



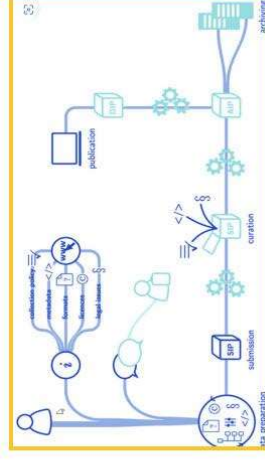
3 Find and query high-quality corpora



4 Search for specific patterns collections of resources



5 Find a matching tool to process your text file (s)



6 Archive and share your language resources

1. About the Authors
  - 1.1. Context and Motivation
  - 1.2 Aims of this Guide
2. What are European Research Infrastructures?

### 3. What is CLARIN?

### 4. Accessing CLARIN

### 5. How to Use CLARIN for Language and Linguistic Research

- 5.1. Searching and Finding Published Language Resources
- 5.2. Searching across Text Collections
- 5.3. Collecting and Citing Language Resources
- 5.4. Finding and Querying Large Collections of Corpora
- 5.5. Finding a Language Processing Tool or Service
- 5.6. Using Published Language Resources and Datasets
- 5.7. Archiving and Sharing Language Resources
- 5.8. Guidance on the Use of Standards and File Formats
- 5.9. Guidance on Licenses and Legal issues in Data Reuse

### 6. How to Use the Knowledge Infrastructure

### 7. Lesson Plans

### 8. Conclusions

### Contribution and Maintenance

### Bibliography



# Tutorials & Learning Activities

In this unit block, you will learn how to locate and use language resources and technology applications through the CLARIN research infrastructure and available national research data repositories.

**Workload:** 4-5 ECTS (estimation, needs to be piloted)

**Designers:** Iulianna van der Lek and Darja Fišer with contributions from Francesca Frontini, Alexander König and Willem Elbers (CLARIN ERIC)

**Proofreader:** Karina Berger (CLARIN ERIC)

## Learning outcomes


By the end of this unit block, learners will be able to:

- Explain the main concepts related to research infrastructures for language resources and technology and the role they play in the research data lifecycle in the context of Open Science and FAIR
- Find and use certified research data repositories to search, find and access, archive and share language resources and datasets
- Find and use online applications to process, annotate, and analyse different types of corpora according to standards and formats used by the community



Introduction to Language  
Data: Standards and  
Repositories

# Tutorials & Learning Activities

 Course Materials

 Unit Block Overview and How to Reuse It



Introduction to Language  
Data: Standards and  
Repositories



 1. Introduction to the Language Resource Management Lifecycle


 2. How Research Data Repositories Make Language Data FAIR

 3. Finding and (Re)using Language Resources in CLARIN Repositories

 4. Citing Language and Linguistic Data

 5. (draft) Legal and Ethical Issues in Language Research

 6. (draft) Sharing and Archiving Language Resources

 7. Student Project

 8. Unit Glossary

# Approach

- Learn by doing
- Interactive content slides in H5P and learning activities
- Take-home assignments and learning resources for self-study
- Modular: lessons can be picked and combined

# Piloting

MA Computational Corpus Analysis  
(5ECTS)

University of Leiden  
The Netherlands

---

# Piloting - Where?

## Computational Corpus Analysis

Vak 2022-2023

### Admission requirements

None

### Description

Due to a relatively recent surge of large-scale text digitization, corpus linguists, digital humanities scholars, but also communication scientists (rather than experimentally elicited) linguists, are increasingly ready to be queried and analyzed. Yet, a large-scale corpus is not only growing in number but also in size, it is no longer feasible to process (manual) of data retrieval, annotation and analysis. In this course, students will be familiarized with a range of computational corpus research and the questions they can address. The course will focus on practical aspects of working with large corpora, that represent computational corpus analysis in fields such as sociolinguistics, and historical linguistics. Finally, ethical issues related to large-scale corpus analysis will also be discussed.

Periode:

Docenten:

C. Tiberius

Dr. J. Fonteyn

### Course objectives

By the end of this course, students will have gained:

- In-depth knowledge of recent developments in computational corpus analysis and digital/computational humanities;
- Insight into the importance of 'found' (i.e corpus) data in linguistic theory;
- Insight into the relevance of using computational methods in linguistic analysis;
- Insight into the various aspects of working with large corpora;
- Refined analytic skills;
- Practical skills for collecting, processing and analyzing corpus data;
- The ability to structure and write a detailed corpus research report.

brigittepace

10

<https://studiegids.universiteitleiden.nl/courses/15783/computational-corpus-analysis>



# Piloting - What?

## Lecture 1 – Collecting and/or finding corpus data for your research

- 3.1. How to use CLARIN
- 3.4. How to create a Virtual Collection with Datasets from Repositories

Self-study material on Brightspace

- 2.1 Introduction to Research Data Repositories
- 2.2 Metadata
- 2.3 FAIR principles

## Lecture 2: Corpus processing and annotation

*I introduce them to the CLARIN Switchboard and let them explore various tools in the switchboard, as one possible resource that they can use for annotation. I referred them to the material from 3.3. How to find tools in CLARIN to process language resources.*

# Piloting - What?

In the next weeks, the teacher will also pilot some activities from

5. *Legal and Ethical Issues in Language Research*
6. *Sharing and Archiving Language Resources*

# Additional Resources and Links

For more info, questions, suggestions for collaborative projects, new ideas, workshops, training, email

**Lulianna van der Lek**

[training@clarin.eu](mailto:training@clarin.eu)

[LinkedIn](#)



## Training Materials

Browse CLARIN's wide range of open-access teaching and learning resources designed by teachers, lecturers and trainers in our network.

[Explore →](#)



## Training Events

Join our free workshops and training events to learn how to use the CLARIN services and natural language processing tools for language and linguistic research.

[Explore →](#)



## Teaching Award

The Teaching with CLARIN Award is for teachers and lecturers who successfully integrate CLARIN resources into the classroom and university curricula.

[Explore →](#)

## CLARIN Impact Stories

In this series we showcase high-quality and innovative research that uses CLARIN tools and resources. These impact stories illustrate the huge variety of disciplines that use the CLARIN infrastructure, highlight the excellent research linked to it, and demonstrate the wider impact that CLARIN and the social sciences and humanities have on broader societal issues.



### Ukrainian History Course in Response to the War in Ukraine

Read more about the distant learning course 'Ukrainian History', which was developed in response to the invasion of Ukraine and was supported by CLARIN.

[Read more →](#)



### CLASP Re-evaluating Child Language Assessment Measures

Dr Nan Bernstein Ratner presents her latest project and discusses how spoken language data is increasingly being used by the ASR industry.

[Read more →](#)



### Discovering Slovenian Language Structure Using Corpora

Jakob Lendič's PhD project combined theoretical and corpus linguistics to explore the subtle characteristics of Slovenian language structure.

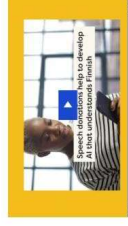
[Read more →](#)



### Open Language Resources for Smarter Artificial Intelligence



### Navigating the GDPR with Innovative Educational Materials



### Donate Speech Database to Boost Development of AI Applications

## Teaching with CLARIN Call

- Continuous call for training material that use CLARIN services, tools and resources in teaching and training
- Presentations at CLARIN Annual Conference
- Showcasing the materials in the Learning Hub
- Inclusion of the materials in the SSH Open Marketplace



### **Teaching Award**

The Teaching with CLARIN Award is for teachers and lecturers who successfully integrate CLARIN resources into the classroom and university curricula.

[Explore →](#)



## Training Materials

Browse CLARIN's wide range of open-access teaching and learning resources designed by teachers, lecturers and trainers in our network.

[Explore](#) →

Applied Language Technology

**Author:** Tuomo Hiippala

Faculty of Arts, University of Helsinki, Finland

**Keywords:** *language technology, digital humanities, tutorial, beginner, spaCy, Stanza, Universal Dependencies, introduction*

Archilochus of Paros: Elegiac Fragments - XML Archive

**Author:** Anika Nicolosi and Beatrice Nava

University of Parma, Italy

**Keywords:** *Ancient Greek, Fragmentary poetry, Textual criticism, Text annotation, Data science*

Computational Morphology with HFST

**Author:** Erik Axelsson

Faculty of Arts, University of Helsinki, Finland

**Keywords:** *morphology, weighted finite-state networks, two-level rules, xfst, lexc, twolc*

GATE, an Open-Source Toolkit for Natural Language Processing

**Author:** Diana Maynard

Faculty of Engineering, University of Sheffield

**Keywords:** *Natural Language Processing; Machine Learning; GATE; social media analysis; disinformation; online abuse detection; Python; Deep Learning; information extraction; digital humanities; corpus linguistics; annotation*

# Training Materials





## Training Events

Join our free workshops and training events to learn how to use the CLARIN services and natural language processing tools for language and linguistic research.

Explore →

# Training Events

## Previous Training Events

Event	Date	Location
<p><b>CLARIN Café: Exploring the Potential of Digital Tools for Online Learning</b></p> <p>This crash course is for academics, researchers and teaching assistants who are interested in learning how to conceptualise and design their own online training materials on a topic of their choice.</p>	3 March 2023	Online
<p><b>Full-text Resource Processing Training Workshop</b></p> <p>This training workshop demonstrated the use of Jupyter notebooks to education professionals working in an academic context and provided them with hands-on experience adapting and extending pipelines for NLP processing of text resources. Participants are guided through Jupyter notebooks that select and pre-process resources making use of metadata, run an NLP task on the selected resources, and further process and present the results.</p>	15 June 2022	Online
<p><b>Teaching with CLARIN Workshop at TALC 2022</b></p> <p>This workshop introduced a number of ways in which CLARIN can support teachers by helping with the discovery, use and sustainability of resources, and by providing materials for teaching.</p>	13 July	University of Limerick, Ireland
<p><b>UPSKILLS ME Utrecht, 4 November</b></p> <p>This one-day event shared the guidelines and best practices for incorporating research and research infrastructures into teaching developed in the UPSKILLS project. The event included a demonstration of how the learning content produced in the UPSKILLS project can be reused and integrated into the university curricula.</p>	4 November 2022	Utrecht and Online

# CLARIN Cafés

A blue poster for a CLARIN Café event. The text is white and yellow. At the top, it says 'CLARIN Café'. Below that, the main title is 'Do Chatbots Dream of Copyright?' and the subtitle is 'Copyright in AI-generated Language Data'. The date '11 April 2023' is in yellow. At the bottom right is the CLARIN logo, which consists of a globe with dots and the word 'CLARIN' below it.

**CLARIN Café**

**Do Chatbots Dream of Copyright?**

Copyright in AI-generated Language Data

11 April 2023

**CLARIN**

# CLARIN Annual Conference 2023

## **Calls**

### **Call for Abstracts**

- Read the call for extended abstracts at [this link](#)
- **Extended deadline: 28 April 2023**

### **Call for Nominations Steven Krauwer Awards 2023**

- Read the call for nomination at [this link](#)
- Submission deadline: 28 April 2023 at 5:00 PM CET

### **PhD Student Session**

- Read more about the PhD Student session at [this link](#)
- Deadlines:
  - Decision on the candidates within the national consortium: Mid July 2023
  - Submission deadline: 8 September 2023

### **Teaching with CLARIN Call**

- Read the call [here](#)
- Submission deadline 15 July 2023



# Open calls for funding and support

