CLARIN Café: Towards Guidelines for Integrating CLARIN into teaching Lessons Learnt from UPSKILLS

CLARIN ERIC
01 December 2021
Organisers

This edition of the CLARIN Café is organized by CLARIN-UPSKILLS team

CLARIN host is Iulianna van der Lek, Training and Education Officer

Technical support by David

The event is recorded for further dissemination purposes. Questions and comments? Put them in the chat box.
Schedule

14:00-14:05 Opening and CLARIN 101, Francesca Frontini (ILC-CNR and CLARIN)

14:05-14:10 UPSKILLS in a Nutshell, Stavros Assimakopoulos (Project Leader UPSKILLS)

14:10-14:20 CLARIN in UPSKILLS, Darja Fišer (CLARIN ERIC)

14:20-14:30 A new Profile for Language Data and Project Specialists, Maja Miličević Petrović (UNIBO)

14:30-14:45 Best-Practice Guidelines for Integrating CLARIN in Teaching and Training, Iulianna van der Lek (CLARIN ERIC)

Break (5 min)

14:50-15:15 Group Discussions

15:15-15:30 Plenary

- Short Presentations on the Outcomes of the Group Discussions
- Wrap-up and Next Steps.
CLARIN 101
https://www.clarin.eu/content/clarin-in-a-nutshell
CLARIN ...

- is the Common Language Resources and Technology Infrastructure
- has the ESFRI ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the humanities and social sciences and beyond
  - to digital language data (in written, spoken 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
- serves as an ecosystem for knowledge sharing
- is an integral part of the European Open Science Cloud
  - See clarin.eu/eosc
CLARIN today

- **70 centres**
- **22 members**: (AT, BE, 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
The Technical Infrastructure

Harvested Metadata → Metadata Search (Virtual Language Observatory (VLO)) → Data Analysis (Language Resource Switchboard (LRS))

clarin.eu/fair   vlo.clarin.eu   switchboard.clarin.eu
The Knowledge Infrastructure

Knowledge centres

Digital Humanities Course Registry

Tour de CLARIN

Impact Stories

Teaching with CLARIN

VideoLectures

Funding Opportunities

Support for EU-funded projects

https://www.clarin.eu/content/clarin-for-researchers

https://www.clarin.eu/content/knowledge-sharing
Teaching with CLARIN

Introduction to Speech Analysis
Author: Mietta Lennes
Faculty of Humanities, University of Helsinki, Finland

Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research
Authors: Darja Fiser and Kristina Pahor de Maiti
Faculty of Arts, University of Ljubljana, Slovenia

GATE, an Open-Source Toolkit for Natural Language Processing
Author: Diana Maynard
Faculty of Engineering, University of Sheffield

Introduction to Digital Humanities
Author: Zuzana Neverilova
Faculty of Arts, Masaryk University, Czech Republic

Oral Archives for Sociolinguistic Research
Author: Silvia Calamai and Rosalba Nodari
Faculty of Languages for Intercultural and Business Communication, University of Siena, Italy

Privacy by Design in Research: How You Can Do a Data Protection Impact Assessment for an Innovative Research Scenario Involving Speech Data
Authors: Esther Hoorn and Henk van den Heuvel
University of Groningen, the Netherlands

Applied Language Technology
Author: Tuomo Hiippala
Faculty of Arts, University of Helsinki, Finland

Archilochus of Paros: Elegiac Fragments - XML Archive
Author: Anika Nicolosi and Beatrice Nava
University of Parma, Italy

Computational Morphology with HFST
Author: Erik Axelson
Faculty of Arts, University of Helsinki, Finland

Winner of the Award
Teaching with CLARIN

Teaching with CLARIN | CLARIN ERIC
CLARIN Resource Families

• What is CRF?
  – user-friendly overviews per data type of the available language resources in CLARIN
  – 12 corpora families, 5 families of lexical resources, and 4 tool families
  – the most important metadata and brief descriptions as well as links to download pages and concordancers

• Who is CRF for?
  – aimed at the needs of researchers from digital humanities, social sciences and human language technologies
  – facilitates comparative research
  – also a useful teaching material
The Café
UPSKILLS:
What is it all about?

Stavros Assimakopoulos
Project Leader - UPSKILLS

info@upskillsproject.eu
What is UPSKILLS?

Funded as an Erasmus+ Strategic partnership

Consortium of 8 partners:

- University of Malta (coordinator)
- University of Belgrade
- University of Bologna
- CLARIN ERIC
- University of Graz
- University of Rijeka

and with funding from Movetia:

- University of Geneva
- University of Zurich
Main aim

→ Tackle skills gaps and mismatches in students of language-related disciplines to create a better workforce.

Rationale:

Graduates of linguistics and language-related degrees are needed in research and industry jobs. Yet, they sometimes lack:

- critical thinking and problem solving skills
- knowledge of research design and data analysis
- project management skills
- digital skills
Main aim

→ Tackle skills gaps and mismatches in students of language-related disciplines to create a better workforce.

How:

- Modular and blended learning
- Innovative pedagogies such as online educational games
- Real-world applications (task-based learning)
- Integrating existing research and research infrastructures into teaching
Outputs and activities

Intellectual outputs:

- Needs analysis ✓
- Guidelines on research-based teaching
- Consolidation and creation of learning content
- Educational games for active learning

Multiplier events planned for each intellectual output:

Bologna, Utrecht, Graz, Malta

→ Summer School in Serbia - July 2023
Main expected results

- Prepare students for the reality of the job market.
- Sensitise academics with respect to what skills employers are looking for.
- Raise awareness among employers about the skills and aptitude of graduates of linguistics and language-related degrees.
- Promote active - task- and research-based - learning.
- Create engaging modular learning content that will be freely accessible to everyone.
Learning content output

- Introduction to scientific research
- Analytical thinking and problem solving
- Project management
- Theorising about language
- Language data science
- Text and speech processing
- Collecting data from human subjects
- Introduction to programming
- Introduction to machine learning
- Data standards and repositories
CLARIN in UPSKILLS

Darja Fišer
CLARIN ERIC
Motivation

Empower the user communities with knowledge, skills and competences in using the CLARIN services, tools and data throughout the research lifecycle.

CLARIN has an established record of supporting training and education for linguists in Europe as a long time supporter for major summer school in Digital Humanities.

UPSKILLS

- A preliminary survey -> there is a need for further integration of CLARIN into the university curricula to support the training of future language data professionals.
Tasks

Best-Practice Guidelines for Integrating Research Infrastructures into Teaching

**Deadline:** September 2022

**Target audience:** lecturer/teacher/trainer/curriculum designer/assistant...

**How:**

- Analyse the insights from related SSH projects & previous CLARIN and UPSKILLS surveys.
- Align the UPSKILLS competence profile with the CLARIN main knowledge areas, services, tools and existing training materials (developed in CLARIN/SSH context).
- Provide guidelines for data collection, annotation and encoding, licensing and depositing the developed data sets to a trusted data repository.

**Workshop for lecturers planned for November 2022**
Tasks

Develop Learning Content for 5-6 ECTS:

• **Text and Speech Processing**
  – Forced Alignment and Automatic Speech Recognition (Henk van den Heuvel, Louis ten Bosch)

• **Language Data Standards and Repositories**
  – Using existing language resources and tools, developing new ones with FAIR principles and GDPR, data management standards, licencing and depositing new resources and tools (Iulianna van der Lek)

• **Deadline**: September 2022

How:

• Identify existing training content and define needed adaptation
• Identify content that could be suitable for gamification
• Define learning outcomes & assessment strategies (5-6 ECTS)
• Define assessment strategies & rubrics
• Select materials, content format & activities, create content
• Define evaluation strategies, pilot and adapt
Approach

1. Draft outlines for both guidelines and learning content
2. Collect feedback and input from the community
3. Write the guidelines and develop the learning content
4. Pilot & adapt
5. Publish
6. Track use and benefits
7. Continuous improvement and encourage contribution
A new Profile for Language Data and Project Specialists

Maja Miličević Petrović
University of Bologna
Language data and project specialist
- knowledge, skills and competences -

DISCIPLINARY
- Knowledge of specific languages
- Ability to conduct linguistic analysis
- Translation, interpreting, post-editing, localisation

(INTER)CULTURAL
- Awareness of specific cultural contexts and cultural differences
- Cultural agility

TRANSVERSAL
- Creative and innovative thinking
- Problem-solving skills
- Presentation skills
- Writing for different audiences

TECHNICAL
- Understanding and use of language technologies and resources
- Understanding of computational linguistics / NLP
- Knowledge of a programming language

DATA-ORIENTED
- Ability to collect, manage, curate, analyse language data
- Knowledge of statistics
- Familiarity with data standards

RESEARCH-ORIENTED
- Knowledge of research design
- Analytical, logical and hypothetical thinking
- Accessing and processing information critically

ORGANISATIONAL
- Entrepreneurship
- Project management
- Quality control
- Planning
- Teamwork

Illustration by Freepik Storyset, https://storyset.com/data
Language data and project specialist
- typical tasks and responsibilities -

**Linguistic data collection**
- Transcription of audio files
- Linguistic annotation
- Language data exploration and analysis

**Data-oriented**

**Research-oriented**
- Language data research
- Research on business processes and market needs

**Technical**
- Translation, interpreting, localisation, post-editing
- Development, analysis and testing of software and technological tools
- Development, testing and improvement of machine learning models

**Organisational**
- Communication with teams, clients and/or vendors
- Project management
- Process evaluation
Language data and project specialist
- four sub-profiles -

Language data analyst
language data collection, annotation, analysis

Language data scientist
language data and processes research

Language data manager
language data cleaning, curation, management

Language project manager
language project and workflow coordination
Towards Guidelines for Integrating CLARIN into Teaching

Presentation of Work in Progress

Iulianna van der Lek

CLARIN ERIC
A little bit more about the context…

Sources for the identification of current practices:

1. CLARIN@University workshop (2019)
2. CLARIN in the Classroom @CLARIN2020
3. UPSKILLS Survey of researchers who teach in language-related disciplines (2021)
4. Teaching with CLARIN (2021)
General outline of the guidelines (Group 1)

1. Introduction
   a. Definitions
   b. A brief introduction to CLARIN
   c. Motivation
   d. Aims of the guidelines
   e. How to use the guidelines

2. The current context, insights, problems and opportunities

3. CLARIN knowledge areas, skills and competences vs. UPSKILLS profile for Language Data and Project Specialist

4. Examples of CLARIN Reusable Learning Objects and how to integrate them in one lesson/course/assignments

5. How to track students research projects and experiments

6. How to FAIRify training materials in the context of open science

7. Other useful resources and tips

8. Conclusions and call for action
CLARIN Data-Oriented Competences (Group 2)
What learning outcomes would be appropriate to define at the BA level for the identified data-oriented competences?

<table>
<thead>
<tr>
<th>General Data-Oriented Topics</th>
<th>Mapping to CLARIN services, resources, tools &amp; existing training material</th>
<th>Targeted Learning Outcomes for BA level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Science, Open Access and Open Data</td>
<td>● From language data to insight: The CLARIN use case (CLARIN-EOSC integration)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● Open Access Resources and Tools for Language (video presentation)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● PARTHENOS Tutorial</td>
<td></td>
</tr>
<tr>
<td>Intro to Research Infrastructures</td>
<td>● PARTHENOS Tutorial</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● CLARIN for Linguists (2014 course; 2015 course)</td>
<td></td>
</tr>
<tr>
<td>FAIR Principles in Data Management</td>
<td>● CLARIN and FAIR (general info; training video)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● The Core TrustSeal (tutorial; CLARIN perspective - slides)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● FAIR Principles and Trusted Repositories (tutorial)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● CLARIN Certified Centres (info)</td>
<td></td>
</tr>
</tbody>
</table>
One easy approach to support the integration of CLARIN into teaching could be the creation of small **Reusable Learning Objects (RLO) on core CLARIN services** that teachers can easily integrate with their lessons/assignments or compile them to form a whole course.

The main characteristics of RLOs suggested by various authors (Wiley, 2002; Barritt and Alderman, 2004; McGreal, 2004; and Koohang and Harman, 2007, Redmond et al., 2018; Konstantinidis et al. 2019, Lazarinis et al. 2019; Costea et al. 2018):

1. **Digital / web-based – 24/7 accessible**
2. **Reusable –** RLOs can be used in multiple contexts; for multiple purposes; at multiple times (e.g. RLOs can be used to teach an undergraduate or graduate course, short course, certificate course, or extension/outreach)
3. **Self-contained –** each RLO focuses on a specific topic/learning objective
4. **Small in size –** to focus learner's attention (2-15 minutes)
5. **Standardized –** RLOs follow the same organizational structure; free of look- and feel of formatting to be reused in multiple delivery media
6. **Searchable –** RLOs are tagged with metadata (information that describes the RLO)
7. **Flexible –** RLOs are easy to update; provide access to quality teaching and learning resources for a wide range of learners
## Example of a full tutorial

### Training and Education Impact stories

<table>
<thead>
<tr>
<th>Voices of the Parliament A Corpus Approach to Parliamentary Discourse Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The tutorial</strong></td>
</tr>
<tr>
<td>The aim of this tutorial is to show the potential of richly annotated language corpora for research of the socio-cultural context and changes over time that are reflected through language use. We adopt a step-by-step approach to guide the reader from formulating queries and analytical procedures to the interpretation of the results. In addition, we supply screencasts for each task that demonstrate the use of a concordancer which helps the user in carrying out the showcased procedures independently. The tutorial is intended for self-study purposes and it takes about 5 hours to complete.</td>
</tr>
<tr>
<td>While the tutorial uses Slovene corpus data, the analyses demonstrated in the tutorial can also be performed on similar parliamentary corpora in other languages as well as generalized to investigate other types of linguistic corpora. On the one hand, this encourages international comparison of parliamentary culture and discourse, and on the other hand, promotes cross-disciplinary exchange of methodological approaches. In order to reach the international audience, there is a Slovene and English version of the tutorial available.</td>
</tr>
<tr>
<td><strong>Target audience</strong></td>
</tr>
<tr>
<td>Students and scholars of modern languages, as well as users from other fields of digital humanities and social sciences who are interested in the study of socio-cultural phenomena through language, to engage with user-friendly digital tools for the analysis of large text collections. No prior experience in using language corpora and corpus querying tools is required in order to follow this tutorial.</td>
</tr>
<tr>
<td><strong>CLARIN Resources and Tools</strong></td>
</tr>
<tr>
<td>The siParl 2.0 corpus can be queried online through the NoSketchEngine or KonText concordancers at CLARIN.SI, the Slovenian node of CLARIN, the European research infrastructure for language resources and technology. The siParl 2.0 corpus can also be downloaded from the CLARIN.SI repository and then further analysed with other corpus or text mining tools. Related materials are available on the PARTHENOS training portal.</td>
</tr>
<tr>
<td><strong>What students say</strong></td>
</tr>
<tr>
<td>This multimedia tutorial enables the users to acquire fundamental methodological and terminological knowledge needed for conducting high quality corpus analysis. It also introduces the users to discourse analysis which can be at the core of the study or used only as a complementary analytical technique. The strong points of this tutorial are its systematic outline and accessible, yet highly informative descriptions that provide enough details also for the advanced users. The tutorial enables students and researchers to go from learning to doing basic corpus analysis in no time. (Mladen Zobec, doctoral student at the University of Graz)</td>
</tr>
</tbody>
</table>

### Analysis of large text collections with digital tools

- Modern languages
  - Digital Humanities
  - Social Sciences
  - Corpus query
- English
- Slovenian

**Beginner/Intermediate**

**Darja Fišer**
**Associate Professor**
Department of Translation, Faculty of Arts, University of Ljubljana

**License:**
CC BY-NC-ND 4.0

**Last modified:** June 2021

**URL:**
https://sidih.github.io/voices/index.html
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
See you at the next café

The Research Data Infrastructure Text+: A new research data initiative in Germany
by Andreas Witt, Erhard Hinrichs, Thorsten Trippel

... to be announced, end of January 2022.

Stay tuned: https://www.clarin.eu/content/

Share your #clarincafe impressions with @CLARINERIC