



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



Integrating research infrastructures into teaching: Recommendations and best practices

The case of CLARIN

(UPSKILLS Intellectual output 2.2)

Compiled by:

Iulianna van der Lek^{*}, Darja Fišer^{*}

Tanja Samardzic[†], Marko Simonovic[‡], Stavros Assimakopoulos[¶], Silvia
Bernardini[§], Maja Milicevic Petrovic[§], Genoveva Puskas^{||},

^{*} CLARIN ERIC

[†] University of Zurich

[‡] University of Graz

[¶] University of Malta

[§] University of Bologna

^{||} University of Geneva

UPSKILLS: UPgrading the SKILLS of Linguistics and Language Students

Erasmus+ Programme

Key Action 2: Cooperation for Innovation and the Exchange of Good Practices Action

KA203: Strategic Partnerships for Higher Education



Grant Agreement Number: 2020-1-MT01-KA203-074246

UPSKILLS Consortium:



&

with financial support from



Integrating research infrastructures into teaching: Recommendations and best practices © 2023 by the authors (UPSKILLS Consortium) is licensed under Attribution-ShareAlike 4.0 International. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>



Disclaimer:

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Executive Summary	3
Acknowledgements	4
List of Abbreviations	5
List of Figures	5
List of references to Moodle learning resources	6
List of Symbols	6
1. Introduction	7
1.1. Context and motivation	7
1.2. How to use this guide	8
1.2.2 How the guide came about	9
2. What are European Research Infrastructures?	10
2.1. Identifying research data repositories for teaching language research	12
3. What is CLARIN?	18
3.1. Accessing CLARIN	19
3.2. Identifying Knowledge Centres of expertise	20
4. Using CLARIN for teaching linguistic research	25
4.1. General recommendations	26
4.1.1. Explore best practices in data-driven learning & corpus pedagogy	26
4.1.2. Build up your corpus literacy and pedagogical knowledge	28
4.1.3. Know your students	30
4.1.4. Identify and select language resources and tools	31
4.1.5. Curate, adapt or create learning content	31
4.1.6. Teaching in the classroom	31
4.1.7. Tracking research projects	33
4.1.8. Evaluate and share your experience	34
4.2. Overview of services	35
4.3. Searching, selecting and using corpora	37
4.3.1. Browsing the Virtual Language Observatory to find language data	38
4.3.2. Searching for language data across multiple corpora collections	43
4.3.3. Locating and querying corpora in the resource families	44
4.3.3.1. Computer-Mediated Corpora (CMC)	46
4.3.3.2. L2 Learner Corpora	47
4.3.3.3. Comparable Corpora	48
4.3.3.4. Parallel Corpora	51
4.3.3.5. Multilingual Web Corpora	52
4.3.3.6. Spoken Corpora	53
4.4. Compiling, sharing and citing virtual collections	55

4.5. Finding tools for data processing and analysis	56
4.5.1. Using the Language Resource Switchboard for text processing	57
4.5.1.1 Corpus analysis and visualisation	59
4.5.1.2 Automatic annotation of texts	59
4.5.1.3 Manual annotation of texts	61
4.5.1.4 Annotation of speech	63
4.5.2. Other NLP tools and applications	64
5. Teaching Linguistic Research Data Management	67
5.1. Examples of learning outcomes for FAIR data skills	69
5.2. Example of student project	71
6. How to Get Involved in the Community	73
7. Conclusions and how to contribute	75
Bibliography	77
Annex A: Resources for teachers	89
A1. Research Tracker Tool	89
A2. Glossary	89
A3. List of Open Corpora	89
Annex B: Questionnaire of lecturers - Current practices of integrating research into teaching	92
Annex C: Accompanying Moodle learning block	95

Executive Summary

According to the UPSKILLS Needs Analysis (Gledić et al., 2021), linguistics and language-related programs seldom include **language data standards and research data repositories** in their learning outcomes. This gap underscores a significant opportunity to integrate essential knowledge and skills students need to engage in **data-driven research and learning**. In response to this gap, Simonovic, M. et al. (2023) developed [guidelines](#) for integrating ongoing academic research and research infrastructures into the curricula of language-related programmes. Specific learning outcomes related to the use of infrastructures for **obtaining, analysing, managing and sharing data** have been included in the course design template of RBT courses. This document complements the RBT guidelines by providing practical guidance on **selecting infrastructures, language resources and tools in the language classroom**.

The most relevant **research infrastructure** for the UPSKILLS project is [CLARIN](#), a European network of knowledge centres of expertise and language research data repositories. The technical infrastructure provides researchers, scholars, teachers and students with a **single access point** to digital language resources published in the research data repositories of the member countries. This guide is a practical resource for lecturers **new to the CLARIN research infrastructure** who want to use it in their programmes, courses or training workshops. It provides an overview of the core services, language resources and tools, which can be used in students' research projects at different stages like data collection, processing, analysis and depositing data in a reliable repository. The guide also includes **insights and testimonials from the lecturers** who engaged in the UPSKILLS project and events, tips and links to other valuable learning resources on data-driven teaching and learning topics.

Furthermore, the guide accompanies the UPSKILLS learning content block on Moodle entitled [Introduction to Language Data: Standards and Repositories](#). The learning content is general enough to be integrated into any programme that involves language research and further extended with more discipline-specific learning activities. Teachers can download the units or individual learning activities and further adapt them for classroom use.

Teachers and trainers can use the guide in the following ways:

- To identify CLARIN centres of expertise, research data repositories, language resources (mainly corpora), services and natural language processing (NLP) tools that they can use to enhance their research and teaching.
- To identify learning content and activities in the accompanying course on Moodle, which they can use in two ways: (1) further educate themselves on a specific topic and (2) repurpose for classroom use to help students improve their data discovery, handling, sharing and archiving skills.

Teachers and trainers who already use the CLARIN infrastructure for language research or research data management in their courses are encouraged to share and contribute to future versions of this guide with their teaching and learning activities examples.

While aimed primarily at teachers and trainers in linguistics and language-related fields, this guide can also benefit anyone in humanities and social sciences disciplines, including curriculum designers, policymakers, librarians, data stewards, and industry professionals seeking to use the infrastructure for research and training.

This work in UPSKILLS aligns with international initiatives like the European Open Science Cloud ([EOSC](#)) and [FAIRisFAIR](#), which promote the adoption of **open science**¹ and **research data management** based on the **FAIR guiding principles**² (Wilkinson et al., 2016) for scientific data management across all domains, disciplines and levels.

Acknowledgements

This guide and learning content is also a community effort. We would like to thank Francesca Frontini, Pawel Kamocki, Esther Hoorn, Alexander König, Mietta Lennes, Jurgita Vaičenonienė, Tanja Wissik, and Willem Elbers for their valuable contributions. We are also grateful to Carole Tiberius and Satu Saalasti, who piloted parts of the Moodle learning content in their courses and workshops. Special thanks to Marie Berthouzoz from the University of Geneva, who helped us build the research tracker prototype.

¹ Open Science. European Commission research and innovation. Available at: https://ec.europa.eu/info/sites/info/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-open-science_2019.pdf. Accessed 30 October 2020

² [FAIR Principles - GO FAIR \(go-fair.org\)](#)

List of Abbreviations

Abbreviation	Definition
CLARIN	Common Language Resources and Technology Infrastructure
CESSDA	Consortium of European Social Sciences Data Archives
DARIAH-EU	Digital Research Infrastructure for the Arts and Humanities
EOSC	European Open Science Cloud
E-RHIS	European Research Infrastructure for Heritage Science
ESS	European Social Survey
ESFRI	European Strategy Forum on Research Infrastructures
FAIRisFAIR	Fostering FAIR Data Practices in Europe
IO	Intellectual Output
RDM	Research Data Management
SHARE	Survey of Health, Ageing and Retirement in Europe
TaDiRAH	Taxonomy of Digital Research Activities

List of Figures³

<i>Figure 1: Searching for research data repositories in linguistics</i>	14
<i>Figure 2: Overview of CLARIN Core Services</i>	36
<i>Figure 3: Faceted Search in the VLO</i>	38
<i>Figure 4: Searching for Language Data in the Virtual Language Observatory</i>	40
<i>Figure 5: Sending plain text files to Switchboard and Virtual Collection Registry</i>	41
<i>Figure 6: Collections of corpora browsable in Content Search sorted by language</i>	43
<i>Figure 7: Searching corpora collections through Federated Content Search</i>	44
<i>Figure 8: Finding and Querying Corpora</i>	45
<i>Figure 9: OpenSoNaR Search Interface</i>	54
<i>Figure 10: Collecting Data from the VLO and Creating a Virtual Collection</i>	55

³ All images and the quick step-by-step guides have been produced with the [TechSmith tools](#).

<i>Figure 11: NLP tools accessible via the Language Resource Switchboard</i>	57
<i>Figure 12: Text processing and analysis with Switchboard</i>	58
<i>Figure 13: Dependency parsing in WebLicht</i>	61
<i>Figure 14: Workflow of a WebAnno project (Castilho, 2014)</i>	62
<i>Figure 15: Examples of annotations in INCEpTION</i>	62
<i>Figure 16: Finding NLP Tools in Language Resource Switchboard</i>	63

List of references to Moodle learning resources

<i>1: Introduction to Research Infrastructures</i>	12
<i>2: FAIR Repositories & Language Resources</i>	17
<i>3: Introduction to CLARIN</i>	20
<i>4: Use Case - Privacy in Research</i>	24
<i>5: Scientific Research & Text Processing</i>	38
<i>6: Citing Language and Linguistic Data</i>	39
<i>7: Finding Data in the VLO</i>	42
<i>8: Finding a Parallel Corpus</i>	52
<i>9: Introduction to Automatic Speech Recognition</i>	54
<i>10: Creating and Citing Virtual Collections</i>	56
<i>11: Basic Introduction to NLP Annotation</i>	58
<i>12: Annotation with WebLicht</i>	60
<i>13: Introduction to Automatic Speech Recognition</i>	63
<i>14: How to Find and Use NLP Tools</i>	64
<i>15: Linguistic Research Data Management</i>	69
<i>16: Example of Student Project</i>	73

List of Symbols



: This icon points to learning resources on Moodle that teachers can reuse in the classroom. ⁴



: This icon encourages the reader to leave the guide and interact with the infrastructure services through an active learning activity. ⁵



: This icon refers to additional tips, information, and learning resources. ⁶

⁴ Online learning content icons created by photo3idea_studio - [Flaticon](#).

⁵ Study icon created by Freepik - [Flaticon](#).

⁶ Info icons created by Roundicons - [Flaticon](#).

1. Introduction

1.1. Context and motivation

The motivation for writing this guide arises from several factors identified within UPSKILLS's remit. The lecturers who participated in our needs analysis (Gledic et al., 2021) expressed interest in integrating research infrastructure data and services into their teaching. However, they faced several challenges, such as difficulty identifying optimal resources and tools to explore specific linguistic aspects, lack of resources for certain languages, insufficient documentation and tutorials on effectively incorporating resources in the classroom, and the absence of discipline-specific best practices and guidelines. An additional questionnaire (see [Annex B](#)) conducted as part of Intellectual Output 2 revealed some technical, financial, and administrative challenges that both lecturers and students encountered when using repositories for language data discovery, reuse and archiving. Furthermore, students' low level of digital literacy was cited as a barrier to using infrastructure and tools in data collection and archiving.

At the same time, informal talks with the UPSKILLS consortium partners and lecturers outside the consortium revealed a need for more general awareness about the wealth of language resources and knowledge available through the CLARIN research infrastructure and national consortia in their respective countries. This lack of general awareness has also been acknowledged at the European level by initiatives such as EOSC and FAIRsFAIR. There is evidence of slow integration of **FAIR data principles, open science and data-related topics** at the Bachelor and Master levels⁷. Therefore, the FAIR Competence Framework for Higher Education proposes a set of core competences for FAIR data education that universities can use to design and integrate research data management (RDM) and FAIR-data-related skills in their curricula and programmes (Demchenko et al., 2021). Students, scholars, teachers and researchers from all disciplines are encouraged to acquire fundamental skills for **open science**, including the **ability to effectively interact with federated research infrastructures** and open science tools for collaborative research. To further support the integration of these skills into the university curricula, FAIRsFAIR published an adoption handbook "*How to be FAIR with your data - A teaching and training handbook for higher education institutions*" (Engelhardt et al., 2022), which contains ready-made lessons plans on a variety of topics, including the use of repositories, data creation and reuse. We hope this UPSKILLS guide and learning content is a valuable addition to these European initiatives but from a discipline-specific angle.

As Gledic et al. (2021) reveal, employers in the digital business sector increasingly seek to hire graduates from language-related programmes with **data-oriented**⁸ and

⁷ EC report, <https://data.europa.eu/doi/10.2777/59065>

⁸ Ability to collect, manage, curate, clean and analyse different language data types.

research-oriented skills.⁹ Such skills have become even more important as new job profiles continue to open up to language and linguistic graduates in the age of AI and ChatGPT revolution, such as computational linguists, machine translation specialists, data curators, data stewards, data annotators, knowledge engineers and terminologists.

This guide shows how teachers and trainers can leverage the CLARIN infrastructure to help students enhance their **data collection, processing and analysis, and archiving** skills. By integrating research infrastructures into teaching, educators can bridge the gap between theoretical knowledge and **practical aspects of linguistic research data management**, equipping students with the necessary skills and competences to thrive in the evolving landscape of open science and data-driven research.

1.2. How to use this guide

This guide complements the [Research-Based Teaching: Guidelines and Best Practices](#) by providing a practical introduction to the CLARIN research infrastructure for lecturers, teachers, trainers and curriculum designers **interested in integrating learning resources (e.g. corpora), language technology tools and research data repositories in their university curricula, summer schools, and workshops.** A browsable-friendly version of **this guide** is available on the [UPSKILLS project website](#).

Along with this guide, an accompanying learning block titled [Introduction to Language Data: Standards and Repositories](#) is available on Moodle as part of **Task 3.2. UPSKILLS Learning Content**. This learning block consists of a modular structure that provides lecturers with a comprehensive collection of learning resources and activities they can use to educate themselves and as reference materials for the classroom. Each major topic in this guide links to relevant learning resources on Moodle to help the teachers identify what they need.



Note that to access the UPSKILLS learning blocks on Moodle, you first need to [create an account](#). Once on Moodle, teachers can browse through all the learning blocks developed in UPSKILLS and download the contents of each block as a Moodle .mbz file, which they can then import into other Moodle systems (version 3.8+). More instructions on reuse are available on Moodle in the *Like What You See* tile at the end of each block, as well as in the CLARIN.SI repository, where the content has been archived for long-term preservation: <http://hdl.handle.net/11356/1865>.

⁹ Critical processing of information, research design, problem-solving, logical thinking, and evaluation technologies

Although the guide targets instructors from linguistics and language-related programmes, it can be adopted by other stakeholders interested in using the CLARIN central services in training and education. 1.2.1 What this guide is not and other guidelines

This guide does not aim to teach how to design, plan and evaluate a course. To this effect, teachers, instructors and curriculum designers may benefit from other guidelines developed in the project.



- Please consult the *UPSKILLS Learning Content Creation Guidelines* for **general guidance about course design** and writing learning outcomes (Gledić et al. 2021). We would also like to highlight this CLARIN Cafe: [Exploring The Potential of Digital Tools for Learning](#), which provides an introduction to instructional design, a template for course design, and tips regarding the selection and integration of digital tools to make courses more engaging and interactive.
- To learn how to **design, run, and evaluate a research-based course**, see [Research-Based Teaching: Guidelines and Best Practices](#) by Simonović et al. (2023). The guidelines include 16 examples of research-based courses piloted by the UPSKILLS consortium partners at their universities.
- For inspiration on **formulating students' projects**, reporting formats and publishing project outputs, see *Guidelines for the Students' Projects and Research Reporting Formats* (Simonović et al., 2021)

** You will encounter many technical terms. A glossary of terms can be found in our Moodle learning block, and it can also be downloaded from [UPSKILLS Glossary - Introduction to Language Data - Standards and Repositories](#).

1.2.2 How the guide came about

The recommendations throughout this guide are based on the insights we obtained from various sources:

- Brief literature review of the state-of-the-art in data-driven learning and corpus pedagogy
- Questionnaire of Lecturers: *Current Practices of Integrating Research into Teaching* (Annex B)

- The third UPSKILLS Multiplier Event: *Guidelines and Best Practices for Research-Based Teaching* (Utrecht, 2023)¹⁰
- The CLARIN CAFE: *Towards Guidelines for Integrating CLARIN into Teaching-Lessons Learnt from UPSKILLS*¹¹
- Analysis of the previous *CLARIN @Universities* events¹²
- Informal discussions with the UPSKILLS consortium partners and personal teaching experience of the authors

2. What are European Research Infrastructures?

The [European Commission](#) defines **research infrastructures** as:

Facilities, resources and services used by the science community to conduct research and foster innovation. They include major scientific equipment, resources such as collections, archives or scientific data, e-infrastructures such as data and computing systems, and communication networks. They can be used beyond research, e.g. for education or public services and they may be single-sited, distributed¹³, or virtual.¹⁴ (European Commission, 2016)

Research infrastructures (RIs) are based on national consortia of research institutes, universities, libraries, museums and archives and **support researchers** in managing the “data lifecycle” in their research projects by providing guidelines for the creation of **data management plans and formats** to facilitate long-term preservation, access and reuse of research data in the context of **Open Science, Open Access** and **FAIR data principles**. The availability of open research data offers valuable resources (e.g. digital text collections, corpora) for students and educators to explore and analyse real-world datasets and it helps them engage in collaborative research across various disciplines. The [Helsinki Digital Humanities Hackathon](#) represents an excellent example of interdisciplinary collaboration, data-driven research, and teaching. Many educational institutions and organisations across

¹⁰ <https://upskillsproject.eu/events/3rd-multiplier-event-2/>

¹¹ <https://www.clarin.eu/event/2021/clarin-cafe-towards-guidelines-integrating-clarin-teaching-lessons-learnt-upskills>

¹² <https://www.clarin.eu/event/2019/workshop-clarinuniversities>

¹³ A distributed research infrastructure is an organisation that enables the research community to use specific facilities, resources and services that are geographically scattered. Definition source: [RI-VIS Communication Toolkit for European Research Infrastructures](#) (2020).

¹⁴ The service is provided electronically.

Europe have recognised the importance of open science and have incorporated training programs and courses to educate students and researchers about open science principles.



To help students understand the benefits of open science, encourage them to play the [Open Up Your Research Game](#) developed by the University of Zurich. The game scenario centres around a PhD candidate who needs to choose between adopting an open science or a traditional approach to conducting research.

Since 2002, about 50 European research infrastructures (RIs) have been set up under the auspices of the [European Strategy Forum on Research Infrastructures \(ESFRI\)](#) in a wide variety of disciplines: e-Infrastructures, Energy, Environment, Health and Food, Physical Sciences & Engineering, and Social and Cultural Innovation. Examples of research infrastructures in the **Social and Cultural Innovation** sector are [CESSDA](#), [CLARIN ERIC](#), [DARIAH-EU](#), [E-RHIS](#), [ESS](#), and [SHARE](#)¹⁵. These European RIs aim to provide open, fair and transparent access to their facilities and services for researchers, scholars and students across Europe and beyond. Moreover, in 2018, the European Commission launched the [European Open Science Cloud \(EOSC\)](#) initiative, which aims to aggregate the services provided by these research infrastructures into one open virtual environment that shares scientific data across borders and disciplines.



To learn how the CLARIN services are integrated into the EOSC platform, watch this video: [A Study on the Use of Nouns by Female and Male Members of Parliament](#). Teachers can replicate this small study with their MA or PhD students using this [tutorial](#) from the CLARIN website.

The two most relevant research infrastructures in digital humanities are the CLARIN ERIC¹⁶ and DARIAH-EU infrastructures. While DARIAH-EU has a broader focus on the arts and humanities disciplines, facilitating knowledge exchange and collaboration through working groups, CLARIN ERIC focuses on the collection, management and long-term archiving of **language resources** and technologies in social sciences and humanities. These two infrastructures collaborate closely on various topics related to training and education through the [Digital Humanities Course Registry working group](#), and by supporting and co-organising summer schools and workshops in digital humanities. **This guide will mainly focus on CLARIN** and how teachers can use it for language and linguistic research.

¹⁵ For more details, see the [ESFRI Roadmap](#).

¹⁶ CLARIN is an ERIC, that is, a European Research Infrastructure Consortium (a legal entity established by the European Commission in 2009), whose main tasks are to build, operate, coordinate and maintain the CLARIN infrastructure.

Besides the CLARIN infrastructure, there are also other infrastructures used to discuss, host and disseminate language resources and technologies, such as [ELRC-SHARE](#), European Language Resources Association ([ELRA](#)), European Language Grid ([ELG](#)), and [META-SHARE](#). For more details and the state-of-the-art on language resources and technology developments, refer to Agerri et al. (2023).

To sum up, the European landscape of research infrastructures is closely intertwined with the principles of open science. Research infrastructures help promote open access, data sharing, collaboration, and interdisciplinary research, which has significantly influenced European training and education. By embracing open science and engaging with European research infrastructures and the communities built around them, students, educators and researchers can access a wealth of knowledge, collaborate across multiple disciplines, and contribute to advancing scientific research and innovation.



Teaching & Learning Resources on Moodle

To give students a general introduction to research infrastructures for language resources and technologies, you can use the following learning content from Moodle:

[Introduction to Language Data: Standards and Repositories](#) learning block

→ Presentations:

- ◆ *1.4. The Role of Research Infrastructures for Science and Research*
- ◆ *2.1. Introduction to Research Data Repositories for Language Resources and Technologies*
- ◆ *1.5. CLARIN: An Example of a Research Infrastructure*

→ Assignment:

- ◆ *1.6. Impact of Language Resources - Reading and Writing Activity.*

Moodle learning resource 1: Introduction to Research Infrastructures

2.1. Identifying research data repositories for teaching language research

Language data is essential for linguistic research and the advancement of natural language processing tools. However, locating relevant language resources and tools for specific tasks can be challenging due to their dispersion across different types of research data repositories (general, domain-specific, institutional) and researchers' personal computers. Thus, we recommend that educators and learners acquaint themselves with current research facilities and language research data repositories in their community. These repositories

provide beneficial avenues for storing, exchanging, and preserving linguistic resources to enhance their discoverability and reusability for the research community.

To ensure that language resources and datasets can be found, are accessible, interoperable and reusable, researchers are recommended to deposit language resources, tools and associated **metadata** in a **FAIR** and **trustworthy research data repository**. FAIR is a set of guiding principles for scientific data management developed by Wilkinson et al. (2016) to help improve the **findability, accessibility, interoperability and reuse** of digital assets. These principles have started to be adopted for data, software, and even training materials¹⁷. FAIR repositories help researchers make their language resources **findable** and **accessible** for a long time by assigning **persistent identifiers (PID)**, such as DOI or handle, which will help retrieve a resource even when it has been moved to another server or domain. Furthermore, repositories provide a fixed set of **metadata elements** (or schema) to describe the deposited language resources consistently, e.g. Dublin Core, OLAC, or CMDI. A free **metadata search engine** (e.g. [Virtual Language Observatory](#) or [OLAC](#)) may be included in the infrastructure to allow users to search and locate language resources useful for a specific (research project). In addition, repositories use standardised **authentication and authorisation procedures** (e.g. single-sign-on¹⁸) and support **different licence models** to enable controlled **access permissions** for different user groups (i.e. use a resource for academic or commercial purposes). Finally, repositories promote **interoperability** through open and standardised formats and facilitate reusability of the language resources and datasets by providing guidelines and best practices for **research data management**. To ensure that a research data repository is trustworthy, it must undergo a quality assessment procedure and get a seal of quality, e.g. [CoreTrustSeal](#), [CLARIN B-Centre Certificate](#), [ISO Standard 16363:2012](#). Data repositories can also serve as **backups** during rare but devastating events where data is lost to the researcher and must be retrieved.



Go to the [CLARIN Centre Registry](#) and see if you can find a CoreTrustSeal-certified centre in your country.

For teachers and students who have never used research data repositories in research or teaching, the [re3data repository registry](#) is a good platform to start exploring available linguistic data repositories and the type of language data they host. This cross-disciplinary directory contains the metadata of over 2000 repositories in different disciplines. Each

¹⁷ There are several task forces at the EU level developing guidelines for making training materials FAIR. As an example, please take a look at the [FAIR Training Handbook](#), produced by ELIXIR FAIR training focus group. CLARIN contributed to chapter 5. *Get a persistent identifier for your training material.*

¹⁸ An authentication method that allows users to sign in using one set of credentials to multiple independent software applications.

repository entry is described with a consistent set of metadata and it can be cited. See Figure 1 for a quick guide on how to use the registry.

Searching for Research Data Repositories in Linguistics

Re3data.org Registry

- 1 Go to **re3data.org** and search for repositories in the linguistics domain.
- 2 Use the **filters** to narrow the search results or sort them by name or last update.
- 3 Check the repository **metadata**, subjects, keywords, certification type.
- 4 Select a repository and check the **Terms** - database and data access, licenses, restrictions for data upload.
- 5 Check the **Standards** used: PID, versioning, citation, metadata standards, quality management.
- 6 Access the repository to search for data or login to deposit data. **Cite** the repository in your research.

Created by I. van der Lek | 12 July 2023

CLARIN in UPSKILLS
Made with TechSmith

Figure 1: Searching for research data repositories in linguistics

When searching for a repository to use for either linguistic research or for sharing and archiving research outputs, teachers could use the re3data registry to teach students how to evaluate and compare different linguistic research data repositories using the checklist that follows:

- What types of policies, documentation and guidelines does the repository offer?
- Is the repository certified (e.g. CoreTrust Seal¹⁹ or CLARIN certification)?

¹⁹ It stands for Core Trustworthy Data Repository Requirements. It is an international, community-based organisation that assesses the quality of a data repository through a certification process. <https://www.coretrustseal.org/>

- Can users search for and access the research data in open access, or are they required to register?
- Are registration and membership required to upload research data to the repository?
- What type of licences does the repository support? (e.g. Creative Commons Licences)
- Does it support common metadata standards and formats (e.g. Dublin Core²⁰)?
- Does it support citation standards and attribution (e.g. implementation of persistent identifiers, such as handles and/or DOIs)?
- Does the repository offer documentation and guidelines for finding and using research data, or support to help create research data management plans?
- Does the repository support the integration of language resources, tools and services in educational settings?
- Does the repository comply with ethical guidelines for data sharing, especially when dealing with sensitive or personal information?

Through the questionnaire in Annex B and testimonials of the lecturers participating in the UPSKILLS projects and events, we have collected a few examples of repositories used for teaching purposes in different language-related disciplines. **Student teaching assistants who start teaching may find them useful** as these repositories provide valuable resources for both teaching and research purposes.

- ❖ **First and Second Language Acquisition:** The [IRIS](#) database, the [ReLDI](#) repository for data collection instruments, or the SLA Speech Tools²¹ repository are often used to search and download learning activities to help students improve their pronunciation in the classroom. The CMU-TalkBank repositories, which give access to well-documented [CHILDES](#) corpora are used to teach students how to analyse child language corpora. For example, the corpora have been widely used in undergraduate courses in language development to create handouts containing sample transcripts that students need to analyse and answer specific questions about language, or make more thorough analysis at the phonological, morphosyntax and discourse levels. Other teachers use the CHAT transcription program to teach students how to correct transcription errors or collect, record and transcribe child language data themselves. For more examples, please refer to the CHILDES [Teaching Resources](#).
- ❖ **Translation Studies and Translation Technology Research:** For parallel corpora, terminology databases or research tools for the study of translated texts, take a closer

²⁰ Dublin Core is a standard for describing information resources, e.g., documents, digital materials, and linguistic resources.

²¹ [Tools for Second Language Speech Research and Teaching](#)

look at the CLARIN repositories²², EuroParl²³, [ELRA-ELDA catalogues](#), [META-SHARE](#),²⁴ [EuroTermBank](#),²⁵ and [TAPoR](#).²⁶ For example, these repositories can be used to find and download parallel corpora in .tmx²⁷ format to use for a translation assignment in the Computer-Aided Translation classroom, extract domain-specific terminology to create a bilingual glossary or use the corpus to train a machine translation engine in AI/Machine Learning programmes. Furthermore, many corpora available in CLARIN are directly integrated into concordancers, such as NoSketch Engine, KonText, Corpuscle, and Korp, which can be freely used in the classroom for linguistic research. More information will be available in [Section 4.3](#).

- ❖ **Language and Speech Technologies:** To search for speech data and tools, the following repositories may be useful: the Bavarian Archive for Speech Signals ([BAS](#)), [Speech Data & Technology platform](#), [Open SLR](#),²⁸ and the [Linguistic Data Consortium](#). The BAS repository shares speech resources of contemporary German and [detailed information](#) about what standards to use when compiling speech corpora and templates for informed consent forms for speakers participating in a research project. Moreover, the **Speech Data and Tech** platform offers a **Transcription Portal** for automatic transcription in English, German, Dutch and Italian.
- ❖ **Language Documentation:** The [ELAR](#) (Endangered Language Archives) repository provides multimedia [collections](#) of endangered languages from all over the world, which can be browsed free of charge. For more advanced use of the resources, registration is needed. The archive also provides documentation, guidelines and [training](#) on research data management, using ELAN to create, transcribe and translate files and Lameta²⁹ to create consistent quality metadata.

After identifying a suitable repository, teachers are advised to check whether it offers integrated, easy-to-use services and tools that can be used in the classroom and whether the

²² <https://vlo.clarin.eu/>

²³ It includes parallel corpora in 21 EU languages, which can be used to train machine translation systems. The corpora are available at: <https://opus.nlpl.eu/Europarl-v3.php>.

²⁴ An open network of repositories for sharing and exchanging language data, tools, and web services.

²⁵ A federated network of terminology collections in 45 languages. Some collections can be downloaded in various formats in .xls and .tbx formats and imported into Computer-Aided Translation Tools to use for translation.

²⁶ The Text Analysis Portal for Research provides a range of tools and resources that support textual analysis.

²⁷ .tmx stands for “Translation Memory eXchange” file format, and it is used in translation and localisation to store and exchange translation memory data (pairs of source and target sentences).

²⁸ Open Speech and Language Resources repository

²⁹ Metadata creation with Lameta:

https://www.elararchive.org/uncategorized/SO_7557c4dd-1c83-485c-b95c-c0a292cfc42c/

repository has an **active user community** and **provides training and support** to its users, e.g.

- Does the repository provide training materials, workshops and webinars to teachers and trainers on how to use the services in education and training?
- Does the repository guide on creating data management plans (DMPs), which outline strategies for data collection, organisation, sharing and preservation?
- Does the repository have an active user community? Do other colleagues use it in your field for research or teaching?
- Would the repository provide support if you were to develop a language resource with your students as part of a research project?

Once teachers become familiar with research data repositories and their services, they can help students choose suitable repositories for their projects, data type and research goals.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

To introduce students to research data repositories and the FAIR data principles, the following teaching resources can be used.

→ Presentations:

- ◆ 2.3. *How Repositories Help Make Your Language Data FAIR*
- ◆ 2.2. *Metadata Standards for Language Resources* (please note that this presentation does not aim to teach how to create metadata but to make learners aware of different types of metadata used to develop and archive language resources)
- ◆ See 2.6. *Finding Experimental Data in Language Acquisition*, which shows how teachers can use research data repositories (e.g. re3data.org, Virtual Language Observatory, Tromso Repository of Language and Linguistics) to find child language data and reuse it to investigate the acquisition of definiteness in Latvian. (Andreassen 2019, March 04).

→ Self-study materials:

- ◆ 5.2. *Data Protection in Research Practice*

→ Assignments:

- ◆ 2.4. *FAIR Analysis of Language Data Repositories*
- ◆ 2.5. *How Findable are Corpora?*

Moodle learning resource 2: FAIR Repositories & Language Resources

3. What is CLARIN?

After a short overview of the current research infrastructure landscape, this section introduces the [CLARIN](#) research infrastructure, an established European Research Infrastructures for “Common Language Resources and Technology Infrastructure.” CLARIN offers online access to an extensive range of written, spoken, or multimodal language resources, which can be used for research, training and education, and developing language technology applications. For an overview of the state-of-the-art in language technology (LT) development, refer to Agerri et al. (2023).

The term linguistic resource refers to (usually large) sets of language data and descriptions in machine-readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems. Examples of linguistic resources are written and spoken corpora, lexical databases, grammars, and terminologies, although the term may be extended to include basic software tools for the preparation, collection, management, or use of other resources.

(Godfrey & Zampolli, 1997, p. 441)

The infrastructure operates through a distributed network of [centres](#) and [services](#), allowing academic users from various fields, particularly in the humanities and social sciences, to use integrated applications to **discover, explore, exploit, annotate, analyse or combine language datasets** to answer new research questions. Researchers and students alike can deposit, share and archive new language resources they create as part of a (research) project in one of the national CLARIN data repositories, ensuring their long-term sustainability. To promote accessibility and usability, all CLARIN repositories and services adhere to the Open Science and FAIR data principles (Wilkinson et al., 2016), making the deposited language data **findable, accessible, interoperable and reusable**.



If you have never used the CLARIN infrastructure in research and/or teaching, you may want to watch [this short introductory video](#).

Member countries contribute to the ERIC financially and in kind, for instance, by hosting a [CLARIN centre](#). Researchers, students and teachers from the **member countries** have access to several central core services and opportunities, which will be showcased throughout this guide. **Users from non-member countries (e.g. Serbia, Slovakia, Malta,**

USA) can also access and explore all the central services and the metadata of the language resources and tools in the repositories freely, without having to log in.



Check the [list of participating consortia](#) to learn if your country is a member of CLARIN. If your country is not a member of CLARIN, but you are interested in using resources with restricted access or depositing language resources in a CLARIN repository, please contact the central CLARIN office at clarin@clarin.eu.

3.1. Accessing CLARIN

The technical infrastructure ensures that academic users in all [participating countries](#) can discover and use the language resources made available and hosted by the various local data [centres](#) through a single sign-on access³⁰ using a [federated identity](#) (i.e. university credentials or CLARIN account).

1. All users can freely **explore** the [CLARIN core services](#) to search for language resources (data and tools) and expertise on specific language research and documentation topics.
2. Due to license restrictions, some resources and services are only available for academic use. To access these resources, login is required through the [CLARIN Service Provider Federations](#), using your institutional or CLARIN website credentials.
3. If your university or academic institute is not listed in the list of organisations, you can request a CLARIN account [here](#).

If help is required to access specific corpora, please check the articles on the [CLARIN Knowledge Base](#). To introduce students to the CLARIN research infrastructure and the added value of developing and sharing language resources and tools, use the learning content recommended below.

³⁰ Access to multiple systems by logging in once.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

To introduce students to the CLARIN infrastructure, the following learning content from Moodle can be used:

Presentations:

- 1.5. *CLARIN: An Example of Research Infrastructure*
- 1.1. *What are Language Resources?*

Assignment:

- 1.6. *Impact of Language Resources*, which refers the students to the [CLARIN Impact Stories](#). This assignment is a nice and easy way to make BA-level students aware of the impact language resources and technologies can have on society.

Moodle learning resource 3: Introduction to CLARIN



To learn more about CLARIN and its history, we recommend these two articles from the CLARIN anniversary book:

- Krauwer, S. & Maegaard, B. (2022). CLARIN – How It Started. In D. Fišer & A. Witt (Ed.), *CLARIN: The Infrastructure for Language Resources* (pp. 1-30). Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110767377-001>
- Jong, F., Van Uytvanck, D., Frontini, F., Bosch, A., Fišer, D. & Witt, A. (2022). Language Matters. In D. Fišer & A. Witt (Ed.), *CLARIN: The Infrastructure for Language Resources* (pp. 31-58). Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110767377-002>

3.2. Identifying Knowledge Centres of expertise

This section gives an overview of the national [CLARIN Knowledge Centres](#) (K-centres) that teachers and educators may find useful. These knowledge centres provide expertise, guidance and training on linguistic topics, data types and tools available via the infrastructure, language processing and linguistic research data management.



Use the [keyword-based search](#) or the direct links below to locate those K-centres with expertise in language resources and methods that might be relevant for teaching and research area:

- [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, intellectual property rights, digitalisation of texts using optical character recognition technology)

Below, we highlight a few CLARIN K-Centres located in the countries of the UPSKILLS consortium partners to help raise awareness about the expertise of these centres and their added value for research and teaching. A **full overview** of all the CLARIN centres can be found [here](#).



Please remember that even if a country is not a member of CLARIN, teachers, students and researchers can still benefit from all the resources and tools hosted by other national repositories that are publicly available.

Austria

In Austria, there are two knowledge centres:

1. [CLARIN Knowledge Centre for Terminology Resources and Translation Corpora \(TRTC\)](#) (hosted by the Centre for Translation Studies at the University of Vienna) provides material and training about the preparation and documentation of [terminology resources](#) (mono-, bi- and multilingual) and translation corpora, including helpdesk support.

- a. The terminology databases are publicly available and can be queried via a user-friendly terminology management system, [Kaleidoscope](#). Users can send feedback on the terms and request to add new ones and/or translations.
 - b. The corpora are available via the [Austrian Language Resource Portal](#) and can be used for translating Austrian public administration documents from Austrian German into English and for intercultural communication.
2. [Phonogrammarchiv K-centre](#), hosted by the Austrian Academy of Sciences, provides audio and video archives from all research fields, including linguistics, focusing on Austrian scholars and institutions. The recorded materials are annotated to facilitate different modes of analysis and can be accessed via the [Academy-CATalogPlus](#) archive for scientific purposes only.

Germany

In Germany, [SAW Leipzig](#).³¹ a Text + / CLARIN Centre, focuses on preserving language and lexical resources for underrepresented languages and offers various research-based applications to explore them. For example, the Vocabulary Portal gives access to over 30 million sentences of German-language newspaper corpora crawled from the Web. In addition, the centre offers [a corpora portal](#) that offers a search interface to more than 1000 corpus-based monolingual dictionaries in 293 languages. The corpora can be downloaded³². The platform can be used to search for words with similar context, examples of use, and neighbour cooccurrences and visualise the relations in a word graph.

Italy

The [CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media corpora](#) provides researchers and students with knowledge, support and training for developing and managing CMC corpora. More specifically, the centre provides FAIR guidelines on [data management](#), such as using [standards](#) and formats, and advice on legal and ethical issues. Researchers, teachers and students can contact the centre via their [helpdesk](#). More information about CMC corpora and how to work with this type of corpora is provided in 4.1.3.4. *Computer-Mediated Corpora*. [ILC4CLARIN.IT](#) offers [services](#) for browsing and querying corpora, file conversion, term extraction, lexical editing and text annotation, focusing on the Italian language.

Norway

Teachers of **syntax and morphology** can use language resources and training materials from INESS, the Norwegian Infrastructure for the Exploration of Syntax and Semantics. This centre is part of the CLARINO Bergen Centre and the [CLARIN Knowledge Centre for Treebanking](#). It provides access to treebanks, which are databases of syntactically and

³¹ Saxon Academy of Sciences and Humanities in Leipzig, a Text+ / CLARIN Centre Leipzig

³² [Download Corpora English \(uni-leipzig.de\)](#)

semantically annotated sentences. The platform is **language-independent** and can be used for building, accessing, searching and visualising treebanks. Data from INESS has also been used in a Master's course on computational language models to show how empirical corpus data can strengthen or challenge hypotheses about grammar. Students quickly learn to use the system in projects for term papers and master's theses. Victoria Troland's master's thesis, for instance, used INESS to extract syntactic markers from syntactically analysed Norwegian novels and subsequently used these markers as a basis for an author identification model. To teach linguists how to query treebanks, see the [INESS Search Walkthrough](#) and the [Parseme tutorial: Studying MWE annotations](#) in treebanks.



To learn how to query linguistically annotated corpora, see Kuebler, S., & Zinsmeister, H. (2014). *Corpus Linguistics and Linguistically Annotated Corpora*, London: Bloomsbury.

South-Slavic Countries

There are currently no K-centres in Croatia³³ and Serbia³⁴, but the [CLASSLA](#) K-centre provides expertise and training on the development of language resources and technologies for **South Slavic Languages**, including Slovenian, Slovene, Croatian, Bosnian, Serbian, Montenegrin, Macedonian, and Bulgarian. The platform gives access to research and tools for language processing, such as information extraction, language understanding, named entity recognition, processing of morphologically rich languages and speech recognition. The centre is managed by [CLARIN.SI](#), [the Institute of Croatian Language and Linguistics](#), and [CLADA.BG](#). Moreover, it offers documentation on how to use the CLARIN.SI infrastructure in Slovene, Croatian and Serbian. A detailed description of the centre is available in [Tour de CLARIN](#).



In June 2023, the centre launched CLASSLA-web, a new collection of large web corpora for [Slovenian](#), [Croatian](#) and [Serbian](#), which can be queried using the [CLARIN.SI concordancers](#) (noSketch engine). Use [this tutorial](#) to learn how to query these corpora.

³³ Croatia is a member of CLARIN ERIC and has a [national consortium](#). The K-centre for the Croatian language is under preparation.

³⁴ Serbia is not a member of CLARIN ERIC; however, researchers, scholars and trainers can use those services, language resources, tools and training materials that are available under a non-restrictive licence.

The Netherlands

Teachers and researchers of **bilingual language development and sign language** are referred to the [CLARIN K-centre for Atypical Communication Expertise](#) (ACE), hosted by Radboud University in Nijmegen. Internationally, the centre is linked to the [DELAD](#) task force, an initiative that provides guidelines for data acquisition,³⁵ processing and sharing of corpora and datasets that contain sensitive data (e.g. speech, audio and transcripts collected from people with language disorders). Moreover, the centre collaborates closely with both the [Language Archive](#) and the [TalkBank](#) repositories to host and give access to corpora of speech disorders securely and in a GDPR-compliant way. Educators can find some examples of well-documented speech corpora on the K-centre [website](#), which can be used to teach students how to analyse disordered speech. Conversely, researchers can use the available corpora to refine the analysis methods and formulate and test hypotheses. For teaching and learning materials related to the impact of GDPR on language research and how to handle sensitive research data collected from human subjects, see this impact story: [Navigating GDPR with Innovative Educational Materials](#). Students and researchers working with patient data learn how to perform a [Data Protection Impact Assessment \(DPIA\)](#) through role plays and use cases. For examples of information sheets and consent forms for collecting speech data from children, download the templates from the [DELAD website](#).



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

To make students aware of key **ethical issues in data collection and sharing of pathological speech data**, see the guidelines on the DELAD website and use the following tutorial on Moodle:

- 5.9. *Use Case - Privacy in Research - Voice Recognition and Parkinson*. (by Esther Hoorn and Henk van den Heuvel)

Moodle learning resource 4: Use Case - Privacy in Research

Switzerland

There is currently no CLARIN K-centre in Switzerland. Still, several universities that are part of the CLARIN-CH consortium, develop language resources, datasets, and tools and provide expertise and training in language sciences.³⁶ Teachers and students searching for language resources are referred to the collection of [CLARIN-CH Resources](#) and the national [Linguistic Corpus Platform](#) hosted by LiRI. Here, we highlight the computer-mediated communication

³⁵ <https://delad.ruhosting.nl/wordpress/guidelines-annotations/>

³⁶ For an overview of language resources and research projects in Switzerland, see the inventory of CLARIN-CH: <https://www.whatsup-switzerland.ch/resources/start>.

(CMC) corpora collected in the *What's up, Switzerland?* project. It contains 617 WhatsApp chats in all four national languages of Switzerland and their varieties, and they are freely available for linguistic research. To learn how to use and query the corpus, see the project website: <https://whatsup.linguistik.uzh.ch/start>. This project is also a good example for students and researchers who want to learn how to process, handle and annotate CMC corpora.



More information about the CLARIN K-Centres is available in [Tour de CLARIN](#) and [Impact Stories](#), which showcase innovative research and educational projects.

To summarise this section, teachers can use the CLARIN local networks to:

- Contact the [CLARIN National Coordinator](#) and/or [national helpdesks](#) to learn more about what CLARIN has to offer in a specific country;
- Contact a [K-centre](#) to find support for a university course/programme, check for training opportunities and seek guidance in getting access to the resources and tools in their country;
- Apply for a [mobility grant](#) to visit a centre or set up a teacher exchange and training programme;
- Search for other [funding opportunities](#), e.g. to organise events or train-the-trainer workshops.

4. Using CLARIN for teaching linguistic research

“The integration of infrastructures into teaching should be seen more as a journey, rather than a goal in itself.” (Vesna Lušicky, Lecturer in Translation Studies, University of Vienna)

When creating a research-based course in linguistics and language-related disciplines, extensive planning and research are necessary to ensure that the course meets students' needs and achieves desired learning outcomes. It is crucial to access research infrastructures, repositories, and language resources that teachers and students can freely use to collect, process, analyse, and deposit language data in the classroom. Free resources and tools can be used to design an engaging and practical hands-on curriculum.

This part of our guide showcases how CLARIN core services, collections of open corpora and online natural processing tools can be used to enhance the teaching of language data discovery, analysis, and archiving skills. As corpora are one of CLARIN's most valuable

language resources, we begin with general recommendations in [Section 4.1](#) for teachers who consider integrating data-driven learning and corpus-based pedagogy in language-related disciplines. Then, in [Section 4.2](#), we provide a brief overview of the CLARIN central services. Should this overview not be sufficient, we invite you to explore the services in more detail in [Section 4.3](#), which provides quick step-by-step guides and references to additional teaching and learning content on Moodle. Finally, [Section 4.4](#) shows how to find natural language processing tools in the infrastructure, which are often used or may be suitable to explore in educational settings.

4.1. General recommendations

👉 NB: *For teachers and trainers who want to start using corpora and corpus technologies in the linguistic, language learning and translation classroom, reading the general recommendations below may be helpful before delving into CLARIN infrastructure (or any other language resources and technology infrastructure).*

The insights collected from the lecturers participating in the UPSKILLS events revealed that the successful implementation of infrastructures, language resources and tools in linguistics and language-related programmes depends on several factors:

- Teachers' own perception, attitude, and confidence in applying data-driven learning and corpus technology tools in the classroom;
- Students' background and their study load;
- The flexibility of the curriculum.

Hence, this section provides **general recommendations on how teachers can improve their perception and skills in applying corpus-based pedagogy and technologies in the classroom** through the CLARIN knowledge infrastructure and wider linguistic community.

4.1.1. Explore best practices in data-driven learning & corpus pedagogy

One of CLARIN's most essential language resources are corpora of various types, modalities and languages. Corpora are often used to answer research questions in both social sciences and humanities domains³⁷, and in **data-driven learning (DDL)**³⁸ to encourage students to analyse corpus data independently using corpus query platforms, create hypotheses, identify linguistic patterns and formulate rules, and verify the validity of grammatical rules. DDL can benefit, for example, foreign language learners by providing them with access to corpora containing authentic texts and user-friendly tools to explore

³⁷ McCarthy & O'Keefe (2010) gives a good overview of the linguistic areas in which corpus linguistics is used.

³⁸ The concept was proposed by Johns (1991).

linguistic patterns and trends in language use and reach their own conclusions (Bernardini, 2002; Boulton, 2009 & 2017). This approach gives students autonomy over their learning process, helps them develop critical thinking skills and turns them into “researchers”.

According to the literature (Boulton 2009; Gilquin and Granger, 2010), data-driven learning based on corpora has not yet been widely adopted in some language-related programmes because teachers are often unaware of the benefits of using corpora for pedagogical purposes. Second, teachers need a high level of **corpus literacy** (Mukherjee, 2006) to be able to develop **corpus-based pedagogy (CBP)**, “the ability to integrate corpus linguistics technology into classroom language pedagogy to facilitate language teaching” (Ma et al., 2021, p 2). Furthermore, teachers need to learn to switch to a “less central role” in the classroom than in traditional teaching and guide the students through the learning process (Gilquin and Granger, 2010). Finally, DDL may not suit all learner types because it requires certain technical expertise to work with the corpus technologies. It can also be time-consuming, as learners need to analyse concordance results and draw their own conclusions³⁹.

Testimonial:

While students appreciate the use of corpus concordancers in the translation classroom, it may take a long time to teach them how to collect enough evidence to be able to make generalisations. Students may also find it difficult to understand how to transform knowledge from one language to another, connect several language resources (e.g. WordNet vs. Valency Lexicon), apply linguistic tests, and make decisions regarding translation equivalents.

Petya Osenova

(Professor and Researcher in Syntax, Morphology and Corpus Linguistics,
Faculty of Slavonic Languages, St. Kl. Ohridski University, Sofia, Bulgaria)

Teachers who never used corpora and corpus technologies in the classroom are recommended to first delve into the literature to understand whether incorporating **corpus-based pedagogy (CBP)** and **data-driven learning** in their curriculum would benefit their specific linguistic sub-discipline, teaching style and the students’ background, level and learning style.

³⁹ We also experienced this challenge during the small-scale corpus research project that we organised during the UPSKILLS Summer School (Petnica, July 2023). A group of 6 students at the BA/MA level with no or limited experience in corpus linguistics had to find comparable corpora in multiple languages to analyse and interpret the meaning of near-synonyms. In one week, the students were introduced to basic corpus methods analyses, learned how to compile a corpus and worked on a research project. The students needed more time than expected to learn how to work with the concordancers and find and interpret the concordance results.



For a good literature review of how corpora are used as a pedagogical tool in various areas of linguistics, see “Part III. Corpora, language pedagogy and language acquisition” in the *Routledge Handbook of Corpus Linguistics* (O’Keeffe & McCarthy 2022).

Other scenarios in which corpora are used in educational settings are teaching **corpus linguistics** as an academic subject, using corpora to **inform syllabus design** and development of educational resources (e.g. dictionaries and grammars) and involving students in the **development of language resources** (collect, design, and compile corpora) (Cheng & Lam, 2022). Finally, students can also be taught how **to share and archive a corpus** at the end of the project, solving all the issues related to handling personal and sensitive data. See Unit 6 in the Moodle UPSKILLS learning content, *Introduction to Language Data: Standards and Repositories*.

4.1.2. Build up your corpus literacy and pedagogical knowledge

Before integrating corpora or any other language resource or tools in the classroom, teachers should first gain hands-on experience themselves. For example, learning basic methods in corpus linguistics, different types of corpora and corpus technologies, extracting relevant data from corpora, count occurrences of phenomena, and doing statistical analyses (Lin, M. H., 2019). Some of the active learning content developed in UPSKILLS and available on Moodle can be a good starting point:

- [First Steps into Scientific Research](#): Early-career teachers or trainers with limited knowledge of scientific research can use this learning block to learn about the main steps in the research process as applied to language, and how to formulate research questions and hypotheses.
- [A Glimpse into Language Data Science](#): This learning block will help learners understand the core concepts related to language data science, such as data description, visualisation, testing of hypotheses and inference.
- [Processing Texts and Corpora](#): This learning block shows teachers how to teach, design, compile, process and analyse corpora for linguistic research. Teachers will find corpus-based activities they can repurpose for their classroom. A final research-based student project has been included for an additional 1-2 ECTS, which also requires students to share and archive their corpus at the end of the project.

Besides proficient use of corpora and corpus technology, teachers will also need **pedagogical knowledge** to design corpus-based activities and assessments that match their course's overall

goals and students' learning needs (Ma et al., 2021). Below, we recommend⁴⁰ established workshops, summer schools and free online courses, which could help teachers build up both their technical and pedagogical skills:

- Lancaster University organises workshops, summer schools and online courses for continuing professional education, e.g.
 - ◆ The Lancaster [Summer Schools in Corpus Linguistics](#), usually include a summer school for language learning, teaching and testing⁴¹, which is aimed at teachers, researchers and students who are interested in language data analysis and research using corpus methods. The programme combines lectures and practical hands-on sessions in computer labs, demonstrating how to explore the British National Corpus (BNC) with corpus-based methods and develop corpus-based materials for language teaching.
 - ◆ [Corpus Linguistics: Method, Analysis, Interpretation](#) is an eight-week course offered via Future Learn, starting each September. The course demonstrates the use of corpora in discourse analysis, sociolinguistics, and language learning and teaching. It suits teachers, researchers, and students who start working with corpora.
 - ◆ The '[Corpus for Schools project](#)' offers ideas and activities for English Language teachers.
- The [Corpus-Aided Platform for Language Teachers \(CAP\)](#) is offered by the Department of Linguistics and Modern Language Studies at the Education University of Hong Kong. The platform offers a great collection of teaching activities and regularly organises training and workshops for teachers to help them design corpus-based activities for the classroom..
- The free [online Moodle course](#) by Agnieszka Lenko-Szymanska at the Institute of Applied Linguistics, University of Warsaw⁴² is a great example of how to prepare corpus-based teaching materials and class activities for teaching languages, vocabulary, phraseology, grammar, language for special purposes, discourse, and language skills.
- Although not free of charge, it is worth mentioning the Boot Camp provided by SketchEngine, both online and face-to-face: [Boot Camp | Sketch Engine](#). The Boot Camp does not teach any theoretical background of corpus linguistics, text analysis or NLP but focuses on using the interface, understanding the functionalities and the search results.

⁴⁰ We also recommend keeping an eye on the upcoming [CLARIN workshops](#), which aim to enhance, among other things, the inclusion of CLARIN corpora and tools in academic curricula.

⁴¹ [Language learning, teaching & testing | Lancaster Summer Schools in Corpus Linguistics \(lanes.ac.uk\)](#)

⁴² Note that you might get a warning that the link is not secure.



For a gentle introduction to **Corpus Query Language (CQL)**, **Sketch Engine** offers free tutorials and short videos for beginners and advanced users, e.g. [CQL – basics | Sketch Engine](#). The same query techniques can be used in NoSketch Engine, the open-source variant, which has been implemented in the CLARIN.SI infrastructure to allow users to query the available corpora.

4.1.3. Know your students

Sometimes, classes in Applied Linguistics, such as Corpus Linguistics and Computational Linguistics, and Translation Studies consist of students with mixed backgrounds (language and humanities vs. computer scientists) and coming from different countries. Such classes pose additional challenges in teaching and learning because the teacher needs to find ways to motivate language and humanities students to work with computational methods and tools. In contrast, the computer scientists need to learn about linguistics. Furthermore, access to multilingual language data repositories and resources is needed to design learning activities in the students' preferred languages. For example, the Virtual Language Observatory and CLARIN Resource Families may be useful in multilingual language technology classes because they provide access to different types of corpora and datasets in multiple languages. Therefore, it is important to **identify students' backgrounds, languages, technical skills, research interests and their level of interest in technology** to be able to tailor the language resources and research methods chosen for the classroom. If the learning activities cater to the needs of the students' linguistic concerns or research interests, they will feel more motivated to engage with technology. Whenever possible, classes should be tailored to certain student groups instead of attempting to meet the needs of a varied group of students in each course (Baldrige & Erk, 2008).

Testimonial:

Try to split the RBT course into feasible portions with clear learning outcomes and consider students' own interests in making the connections between linguistics and the more technical parts (e.g. programming in Python, data handling). You could start with the linguistic questions, then use technical knowledge to address the questions and translate the answers back into the research domain.

Louis ten Bosch
(Associate Professor of Deep Learning and Automatic Speech Recognition,
Radboud University, the Netherlands)

4.1.4. Identify and select language resources and tools

Teachers are invited to use this guide and the accompanying course on Moodle, [Introduction to Language Data Standards and Repositories](#), to get acquainted with the CLARIN central services and **identify those language resources and tools suitable to include in teaching**. The exploration may be easier if it is based on a clear research question, study or project that teachers intend to formulate in the classroom. Consider the type of data the corpus should have, register, languages, size, availability, the time period that the corpus covers, etc. When testing and selecting corpus technology tools, it is recommended to test them properly to identify those functions that will help achieve a specific goal in the classroom. The final choices and implementation of language resources and tools in the classroom are often influenced by teachers' own perceptions, attitudes, and confidence in applying corpus technology in the classroom (Leńko-Szymańska, 2017; Ma et al., 2022).

4.1.5. Curate, adapt or create learning content

Once appropriate resources and tools have been identified for classroom use, the learning outcomes of the course should be adapted to target practical skills related to the use of infrastructure, creating and curating learning materials and designing corpus-based activities and assessments. To save time and effort, teachers are recommended to pick and choose units of individual learning activities from the UPSKILLS Moodle platform, build on them and share them with their students. Instructions on how to export and reuse learning content are included within each learning block.

When **creating learning content for technical-oriented tasks** and using different tools, it is important to keep in mind that tools change rapidly, so updating the content will require time and effort. Therefore, to save time, consider reusing tutorials provided by the infrastructure and tool providers, and adapting them to match the students' learning objectives and levels. If learning materials need to be created from scratch, the focus should be on making them as modular as possible so that other teachers can easily update and reuse them. Moreover, in multidisciplinary courses, learners come from different fields and backgrounds, and some may not yet have all the required skills. This can place an extra burden on the teacher. In **online teaching**, the issue might be partially solved by creating small, self-contained and well-described learning objects that can be used flexibly in many different courses and replaced or removed when the content becomes obsolete.

4.1.6. Teaching in the classroom

Effective planning and preparation are crucial when introducing language resources and technologies in the classroom. **Before the class**, it is useful to set up a collaborative online space like Google Drive, Colab, or GitHub, containing all the files and instructions for the students, including handouts and tutorials. Such a space will help increase teaching

efficiency and minimise disruptions. Moreover, setting up a forum where students can ask questions to teachers and peers regarding the technical issues they may encounter during homework and class assignments can also ease the teacher's workload. This way, students learn to solve problems independently, with the teacher intervening only when necessary.

During the class, allocate sufficient time for questions and to help students with lower digital literacy.

Testimonial:

Switching to a more student-centred approach, including peer review and project-based evaluations, might also ease the lecturer's workload because students sometimes come up with great solutions on their own. Using scaffolding instruction techniques and creating a safe environment for the students to experiment and discover new knowledge on their own (e.g. use the infrastructure to find metadata and standards for their own language combinations) can make both the students' learning experience and the task of the lecturer more enjoyable.

Vesna Lusicky

(Research Associate and Lecturer at the Centre for Translation Studies,
University of Vienna, Austria)

Below, we have compiled a general framework of helpful teaching strategies from Sripicharn (2010), Tribble (2010), and Lessard-Clouston & Chang (2014) that can be referenced when starting to teach with corpora regardless of the linguistic sub-discipline.

1. First, assess students' prior knowledge⁴³.
2. Introduce corpora and data-driven learning.
3. Identify/create corpus-based tasks/activities for the classroom⁴⁴.
4. Identify and select relevant corpora and tools, providing access to them (preferably accessible online).
5. Show how to use corpus analysis tools, e.g. concordancers, and demonstrate types of different queries. A possible process to teach students how to use and read concordance lines, could involve the following steps:
 - a. *Initiate* a search for patterns in a set of concordance lines.
 - b. *Interpret* the concordance line results.

⁴³ Based on our experience at CLARIN with organising workshops, the assessment of students' prior knowledge should preferably be done two weeks before the class so that the instructor has enough time to prepare the class, choose the suitable corpora and tools, design learning activities based on the students' level, knowledge, skills and domain interest.

⁴⁴ An overview of corpus-based activities per linguistic sub-discipline is available in the *Routledge Handbook of Corpus Linguistics* (O'Keeffe, A., & McCarthy, M.J. (eds.) 2022).

- c. *Consolidate the results* by looking for additional patterns and report them to others by explicitly writing down your observations.
 - d. *Recycle* your results by looking for further information, patterns and other contexts.
 - e. *Check* what other relevant sources (e.g. dictionaries) say about the patterns you identify.
6. Help students interpret search results and repeat the process with more data from the corpus until they can work independently.

After students have increased their corpus literacy skills, they can be introduced to data-driven learning by challenging them to work on a small-scale research project and answer a specific research question. Bennett (2010) proposes a general framework for using corpora in language teaching, similar to the one above but starting from a research question.

1. Have a clear research question.
2. Determine the register on which your students are focused.
3. Select a corpus appropriate for the register (e.g. *spoken or written; general or specialised; contemporary, historical or diachronic; standard or non-standard*).
4. Use a concordance program for quantitative analysis.
5. Engage in qualitative analysis.
6. Create exercises for students.
7. Engage students in a whole-language activity.



See [Section 5.2](#) for an example of a student project designed in collaboration with the UPSKILLS consortium partners from the University of Bologna.

4.1.7. Tracking research projects

Undertaking and monitoring research projects involving language resources can be challenging and require continuous guidance and feedback to help students improve their work. With many changes that may occur during the research process, it is easy to get lost and confused. However, **keeping track of information at different stages** can help clarify ideas, resulting in faster progress. In line with this, our consortium partners from the University of Zurich have developed an interactive **research tracking tool** that enables students to track their progress during projects. Students fill in the template using the student version of the tool. Teachers then use the **teacher's version of the tool** to provide brief feedback in the designated field. At the project's onset, it is up to the teachers and students to agree on the reporting frequency. Each submitted report should receive short feedback from

the teacher. The preferred sharing mode (email, online drive, etc.) will also be decided. The research tracker can be downloaded from the UPSKILLS project website⁴⁵.



To learn more about student project design and reporting formats, please see the *Guidelines for the Students' Projects and Research Reporting Formats* (Simonović et al., 2021).

4.1.8. Evaluate and share your experience

Language resources and technologies can support teaching and research across various disciplines to teach students how to develop a data-driven mindset and draw insights and conclusions from large volumes of text. While testing and using various language resources and tools corpus classroom use, collecting feedback from students on the usability of the resources, tools and infrastructure and informing the resource and tool developers and the infrastructure managers of their usefulness in educational settings is beneficial.

For example, Lusicky and Wissik (2016) evaluated the usability of language resources like corpora and translation memories, disseminated through research data catalogues such as Virtual Language Observatory, Meta-Share, and ELRA, for translation studies scholars and students. According to the authors, repositories could help researchers make their language resources more FAIR and meaningful for translation studies by including specific metadata catering to the field. For instance, in the case of parallel corpora, it is essential to know the original language, the reliability of the source texts, whether the translation was obtained via post-editing machine-translation output and the translators' names and their native languages.

The following criteria could be used to evaluate your experience with the tools in the classroom:

- *Accessibility* (e.g. User interface design, clarity of instructions, ease of navigation)
- *Efficiency* (e.g. Search speed, data retrieval, overall tool performance)
- *Functionality* (e.g. Does the tool include the features you need to help you achieve your goal? e.g. linguistic processing, analysis, data manipulation and visualisation?)
- *Interoperability* (e.g. Can you upload a dataset or a digital text collection you found in a repository into a tool of your choice for further exploration or more sophisticated analysis?)
- *Quality of the documentation and other support, training materials provided with the resource and/or tool* (e.g. Does the tool creator provide clear documentation, tutorials, and use cases on how the tool can be used for research or teaching?)

⁴⁵ See Integrating Research Infrastructures into Teaching, Resources for Teachers: [Research Infrastructures | UpSkills Project](#).

All the feedback, suggestions for improvements or additional features could then be shared with the infrastructure and tools developers to improve the subsequent versions of their tools. See the testimonials below as an example.

Testimonials:

The language resources available via the CLARIN infrastructure are very important for my teaching. It is vital that the same materials are persistently available and that they can be cited consistently. I try to teach first-year students to search for corpora and other resources that can be relevant to them via CLARIN platforms, and the more advanced students can benefit from tools, good practices and guidelines that help them to process and make available the resources they create. Praat and ELAN have been around for a long time, and teachers can be confident they will remain accessible in the coming years. University students need good examples of reproducibility, scientific references and citation practices for tools and data since these will be the building blocks they will work with in the future

Mietta Lennes
(Lecturer of Speech Technologies, University of Helsinki, Finland)

The students perceive corpora as something entirely new; they tend to get scared initially and need some time to familiarise themselves with different types of queries, and they tend to struggle with regex. But these challenges can be overcome.

Anonymous lecturer
(Questionnaire of Lecturers - Annex B)

4.2. Overview of services

In a nutshell, teachers can use the CLARIN core services in the classroom to teach language data discovery, (re)use, sharing, citing and archiving. See the overview of services in [Fig. 2](#), followed by a brief description.

The CLARIN Core Services

An Overview

1 Browsing for LRs in the Virtual Language Observatory

2 Processing LRs with NLP tools from Language Resource Switchboard

3 Creating, sharing and citing LRs with the Virtual Collection Registry

4 Searching for linguistic patterns across multiple corpora collections at the same time

5 Searching & using corpora, lexical resources and NLP tools from the Resource Families

6 Depositing, sharing and archiving LRs in a FAIR research data repository

Created by I. van der Lek | 13 August 2023

CLARIN in UPSKILLS

Figure 2: Overview of CLARIN Core Services

1. Use the [Virtual Language Observatory \(VLO\)](#) to search for full-text language resources of different types, languages, modalities, time periods, formats and licences.
2. Find a matching natural language processing tool via [Language Resource Switchboard \(LRS\)](#) to process language resources or texts and perform more advanced linguistics tasks, such as different types of automatic annotation, morphological analysis, distant reading, terminology and keywords extractions, topic modelling, etc.
3. Collect the resources discovered in the VLO or any other research data repository in a virtual collection in the [Virtual Collection Registry \(VCR\)](#) that can be cited and shared with other teachers or students. This service allows users to save resources for later exploration and processing. In contrast with Zotero or Zenodo repositories, the VCR allows you to add multiple resources to a collection and cite the entire collection using persistent identifiers, such as **handle** or **DOI**.
4. Use the [Federated Content Search](#) to search for specific linguistic patterns across several collections of corpora in several repositories simultaneously. You can use CQL for queries and download the search results in different formats.

5. Search for corpora for specific registers and languages in the [Language Resource Families](#). Most corpora are freely available, can be cited, and downloaded from the repository where it is located. Some corpora are directly available for query in online concordancers, such as KonText, Korp and NoSketchEngine.
6. Language resources created collaboratively as part of a research-based project or in the context of a thesis can be deposited, shared and archived through a suitable CLARIN repository. The repositories adhere to the FAIR guiding principles for research data management and sharing. See the [depositing services](#) for general depositing guidelines and an overview of the centres providing support in this process.



If are new to CLARIN, watch [this video](#) to learn how the **Virtual Language Observatory** and the **Language Resource Switchboard** are integrated to enable language resource discovery and reuse for research purposes.

To help teachers evaluate the suitability of the CLARIN infrastructure (or any other infrastructure listed in section 2) for teaching language and linguistic research, we recommend first **exploring and testing the core services, some tools and language resources** with the help of this guide and accompanying learning content on Moodle, [Introduction to Language Data: Standards and Repositories](#), especially *Unit 3: Finding and (Re)using Language Resources in CLARIN Repositories*.



Teachers already acquainted with the CLARIN infrastructure may also go directly to the learning content on Moodle and explore the learning content.

4.3. Searching, selecting and using corpora

This subsection presents the CLARIN central services, which teachers and students can use in the planning and data collection phases of a research project to search and locate language resources that can help answer specific research questions, replicate a dataset, build a corpus, or train a language model. The services for data discovery are the **Virtual Language Observatory**, **Federated Content Search**, **Resource Families**, and the **Virtual Collection Registry**. While many language resources are accessible through CLARIN, we will mainly demonstrate how to search, locate and use corpora of different types, languages and modalities.

After getting acquainted with the basic functionalities of each service and understanding what they can be used for, try to test them by formulating a research question

for a specific register and language(s). This would make the searches more focused and the identification of appropriate corpora and tools easier.



Teaching and Learning Resources on Moodle

- To teach students the main steps in the research process as applied to language and how to formulate good research questions and hypotheses, see the [First steps into scientific research](#) learning block.
- Teachers who never worked with corpora or cannot find a corpus for a specific register and language can create their own corpus for classroom use by collecting texts from the web with tools such as BootCat or SketchEngine. See the UPSKILLS [Processing Texts and Corpora](#) learning block to learn how to design, compile, process and analyse a corpus in a concordancer.

Moodle learning resource 5: Scientific Research & Text Processing

4.3.1. Browsing the Virtual Language Observatory to find language data

The [Virtual Language Observatory \(VLO\)](#) central catalogue automatically harvests metadata on language resources contributed by researchers in CLARIN member and observer countries. It offers advanced search functionalities that facilitate the easy discovery of language resources, such as corpora, lexica, grammars, multimedia recordings, digitised texts such as books, articles, and transcripts of parliamentary debates, software & web applications, and even training materials.

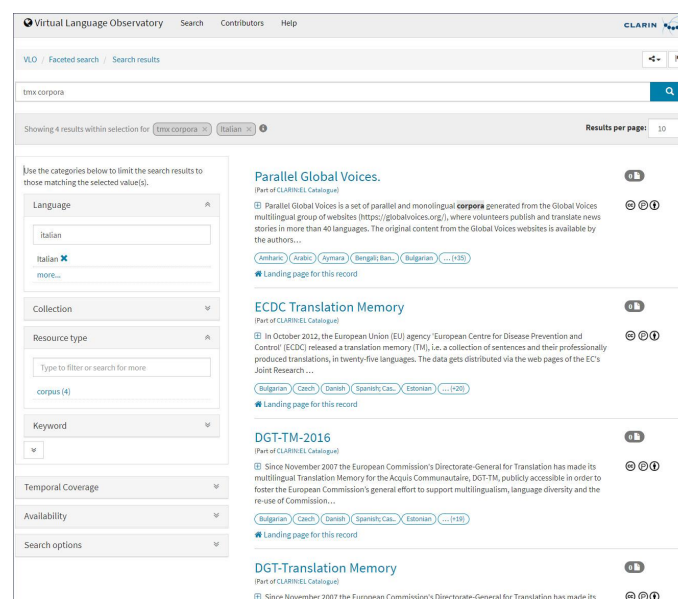


Figure 3: Faceted Search in the VLO

Because of the large amount of data, there are multiple ways of exploring the VLO, e.g., full-text search, facet browsing, or geographic overlay. The advanced filters can help narrow down the search results and find resources or text collections in a specific *language*, *resource type* (text, audio, dataset, corpus, software, video, etc.), *modality* (spoken, writing), *format* (text, audio, image, specific keywords), *temporal coverage* or *availability* (for public, academic, restricted use). See Fig. 3 for an example of a faceted search.

When searching for resources, remember that the search results might contain duplicate entries, and incorrect or incomplete titles/descriptions. The flag icon on the top right corner can be used to report issues via the VLO feedback form. The service is continuously improved to facilitate easy data discovery.

Each resource in the VLO can be accessed directly via the **unique/persistent identifier**, i.e. **handle**, pointing to the landing page of the repository where the resource creator initially deposited the resource (see Fig. 4). The handle can be used to reference the landing page online and in publications. CLARIN endorses the Data Citation Principles.⁴⁶



Teaching and Learning Resources on Moodle

- If you are unfamiliar with the current practices in citing language data, see **Unit 4 on Moodle of our [Introduction to Language Data: Standards and Repositories](#) on Moodle: *Citing Language and Linguistic Data*.**

Moodle learning resource 6: Citing Language and Linguistic Data

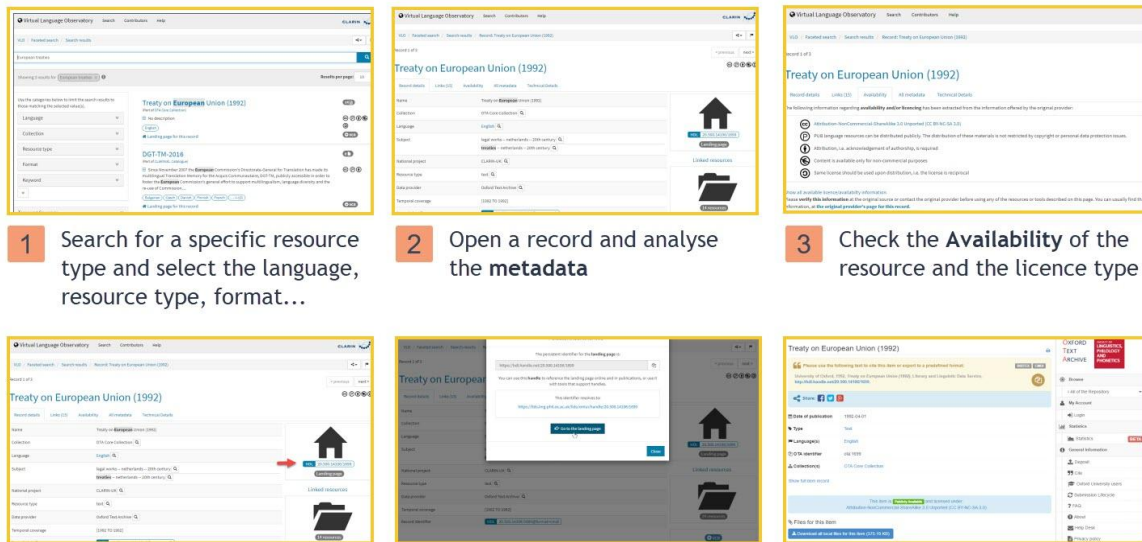


Assuming you are searching for a digitalised text collection of European treaties to build a corpus, go to the VLO and perform the steps described in Fig. 4. Was it easy to use, find and access the Treaty of European Union collection?

⁴⁶ Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 <https://doi.org/10.25490/a97f-egy>

Searching for Language Data

Browsing the Virtual Language Observatory



- 1 Search for a specific resource type and select the language, resource type, format...
- 2 Open a record and analyse the metadata
- 3 Check the **Availability** of the resource and the licence type
- 4 Click on the **Handle** to learn in which repository the resource is hosted
- 5 Go to the **landing page** of the resource
- 6 Preview or download the files in the resource. Copy the **citation format** to refer to the dataset in your research.

Created by I. van der Lek | 06 August 2023

CLARIN in UPSKILLS

Figure 4: Searching for Language Data in the Virtual Language Observatory

When you find a language resource in the VLO that might be interesting to explore further either for research or teaching, **carefully read and interpret the metadata** fields used to describe the resource, especially the **record details, links** and their **availability**. Each resource is described based on **metadata standards** (e.g. CMDI, Dublin Core, OLAC), which provide helpful information about who created the resource and how. It is also advisable to review the metadata of the resource **from the perspective of your specific linguistic sub-discipline**. For instance, the study by Lusicky & Wissik (2016), mentioned earlier, assessed the usability of language resources such as corpora, translation memories, terminology resources, and lexica from the VLO and other research data repositories (META-SHARE, ELRA) for translation studies scholars and students.

Further, the VLO records contain information about the **licence** assigned to the resource and **terms and conditions of use**. Although CLARIN advocates for open science and access, data creators may restrict access to data collections due to sensitive data. It is important to note that resources and tools in CLARIN are assigned either a **public (PUB)**, **academic (ACA)** or **restricted (RES)** licence. Public resources can be reused without copyright restrictions, whereas academic or restricted resources can only be used under specific conditions.

Finally, the digital text collections in the VLO can be processed and analysed with integrated **Natural Processing Tools**. For example, suppose you find a resource in the VLO in plain text format. You can process it directly using one of the **Language Resource Switchboard** tools, e.g. use **WebLicht** to annotate the plain text file automatically, use **UDPipe** to produce syntax trees or use the **LINDAT machine translation** service to translate the file into another language. You will learn more about Switchboard in [Section 4.5. Finding Tools for Data Processing and Analysis](#). If you do not want to process the file immediately, you can choose to **queue it for submission to a Virtual Collection**. The following figure shows how to access the Switchboard and Virtual Collection Registry directly from the **Links** area of the VLO record.

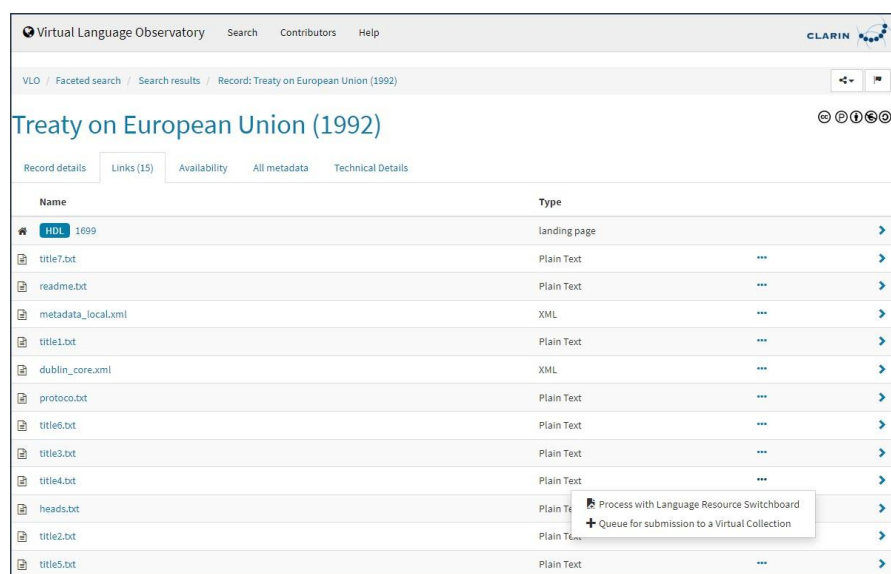


Figure 5: Sending plain text files to Switchboard and Virtual Collection Registry

The metadata quality of language resources and their **suitability for educational settings can be evaluated through** a simple checklist, e.g.

- Does the language resource provide sufficient metadata to help you decide to use it in your research and/or teaching? If used for research purposes, does it sufficiently address the research problems?
- Who created the resource and when?
- Is it reliable? How often has the resource been viewed and/or downloaded by others (researchers, teachers, students)?
- What type of language data has the resource creator collected, and from whom, when and where? How was the data collected and processed?
- Does the resource creator provide any information about how the resource can be used in either research or teaching?

- In what format the data and metadata are available? Are the files in the resource available in a common format compatible with other tools?
- Does the repository provider offer an online service to help you look inside the language resources/ dataset?
- If the infrastructure does not provide any tools to preview or explore the contents of the resource, can you and your students download it and use it for processing and analysis in other tools without restrictions?
- Do you and your students need special software skills to process and analyse the dataset? If you lack the skills to process a large dataset, you can try to contact the data provider or the research data management department at your university and ask for help.

After the students learn to evaluate resources and their metadata properly, teach them how to use them in the tools available through the infrastructures or other preferred linguistic tools.

To summarise, the VLO cannot provide answers to research questions. However, it can assist in identifying existing resources that can aid in answering those questions without duplicating efforts. Through the use of the VLO, both teachers and students can gain knowledge on how research data repositories promote the **FAIR principles** (findable, accessible, interoperable, and reusable) with the help of metadata standards like Dublin Core, Olac, and CMDI, persistent identifiers such as handles, and licensing agreements that specify terms and conditions.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

- For an overview of metadata standards used to describe language resources (e.g. corpora), see 2.2. *Metadata Standards for Language Resources*. This interactive presentation contains a few exercises that can be used to help the students understand the different types of metadata.
- To learn and teach students how to use the VLO to find language data, pick and choose from the following activities:
 - ◆ 1.7. *Applying the FAIR data principles to corpora*
 - ◆ 2.6. *Case Study: Finding Experimental Data in Language Data Acquisition [developed by Andreassen (2019)]*
 - ◆ 3.6. *Finding and Using a Parallel Corpus for a Translation Assignment*
- For an overview of guidelines for citing linguistic data, see *Unit 4. Citing Language and Linguistic Data*.

Moodle learning resource 7: Finding Data in the VLO

4.3.2. Searching for language data across multiple corpora collections

While the VLO allows only metadata searches to locate full language resources and texts, one can use the [Federated Content Search](#) (FCS) to identify specific linguistic patterns (e.g. collocations) across various corpora hosted in different CLARIN centres simultaneously. The corpora stay at the centre where they are hosted; therefore, the underlying technique is called *federated content search*. As of April 2023, there are 207 corpora searchable via FCS in various languages.

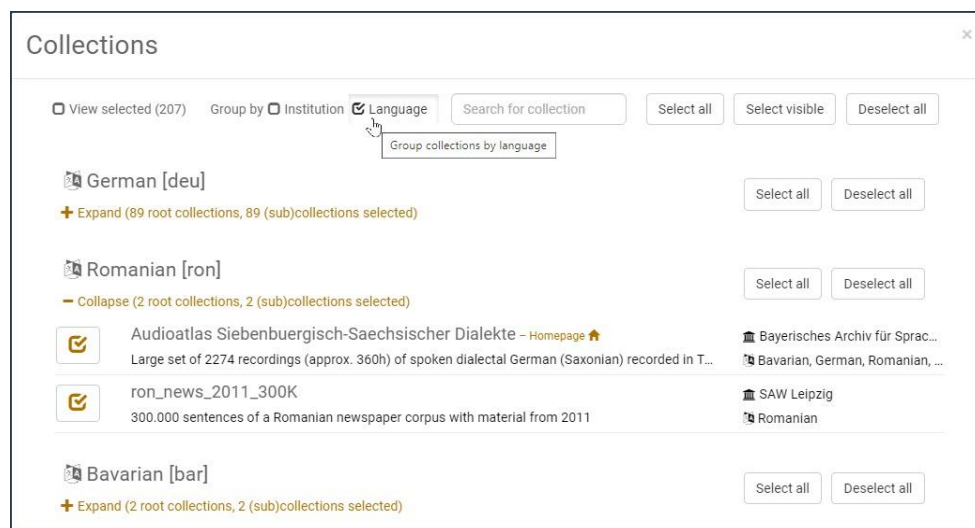


Figure 6: Collections of corpora browsable in Content Search sorted by language

To view the available corpora per language, go to the **Content Search** main interface and click on **Collections**. Then, tick the **Language** box to group the collections per language, see Fig. 6. Finally, click the + sign to expand and view the corpora available for a specific language. Monolingual searches can be performed with the help of integrated **Contextual Query Language (CQL)** queries. The search results can be displayed as **Key Words in Context** and downloaded in various file formats.



Use the steps described in the quick guide in Figure 7 to learn how to use the FCS service. You can, for example, search for collocations in a specific language and export the results.

Searching Corpora Collections

CLARIN Federated Content Search

- 1 Go to FCS and perform a CQL query
- 2 Select language
- 3 FCS will search for the word or phrase in several corpora at the same time and display the results
- 4 Display the results as Key Word in Context
- 5 Download the search results in different formats
- 6 For more advanced searches, view the corpora collections and access the search interfaces on their homepage

Figure 7: Searching corpora collections through Federated Content Search

Although the service offers basic functionalities, it can help teach students basic corpus query searches to investigate how certain words and phrases are used in context. To perform more sophisticated queries, view the collections of corpora available through the FCS service and go to the search interface of the centre hosting the text collection.

4.3.3. Locating and querying corpora in the resource families

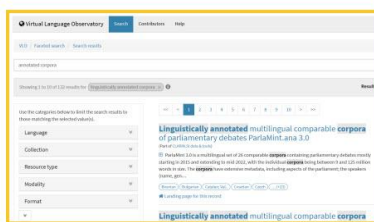
Research has demonstrated that scholars, instructors, and learners have used corpora for "data-driven learning" (Bernardini, 2006) to examine genuine language usage, contextual subtleties, and actual linguistic variations, allowing them to make generalisations about language use. Teachers who already use corpora in their own research and/or teaching methods may already have one or more preferred corpora and corpus analysis tools they are accustomed to using in the classroom.

For teachers considering integrating corpora into their course or programme, the [CLARIN Resource Families](#) can be an invaluable starting point. The families include various corpus types (and tools), each catering to specific linguistic or research needs (see Fig. 8 for examples), and are meant to facilitate comparative research. Additionally, the listings in each

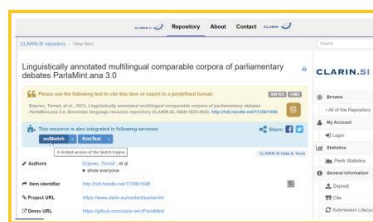
resource family are sorted by language and provide brief overviews about the size of the resource, text sources, data type, time periods, annotation types, standards, formats and licence information. Unlike the Virtual Language Observatory, this overview of corpora collections is much user-friendlier and makes it easier to identify what type of corpora matches the goals of a specific course/project/assignment.

CLARIN in the CLASSROOM

Finding and Querying Corpora



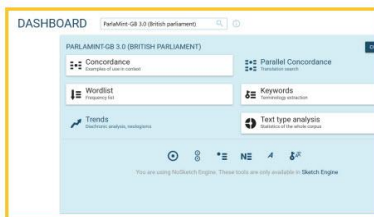
1 Search the VLO for annotated corpora (*language, resource type, modality, format, time period, licence*)



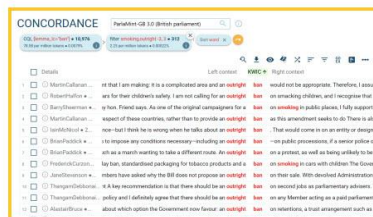
2 Access the landing page
Perform queries in noSketch or KonText
Cite the resource



3 Search for corpora in Resource Families
See list of corpus query tools



4 Query open corpora in NoSketch Engine (*limited open-source version of SketchEngine*)



5 Concordance search results



6 Corpus information (*language, lexicon size, counts, text types*)

Figure 8: Finding and Querying Corpora



If this is your first time working with corpora, see *A Practical Handbook of Corpus Linguistics* by Paquot & Gries (2020) to understand how different corpus types can be designed, analysed and compiled (see chapters 1 to 3). The handbook can also be used as a course book in the classroom or for individual study.

Most importantly, many corpora listed in the resource families are available in open access and can be directly queried in online concordancers, such as [Korp](#), [Corpuscle](#), [KonText](#) and [noSketchEngine](#). These concordancers are integrated into the infrastructures of

the CLARIN national repositories in Finland, Norway, the Czech Republic and Slovenia, giving access to many corpora developed by researchers in their communities⁴⁷. Moreover, some corpora come with detailed tutorials demonstrating how to use them for research. The tutorials are available in open access and can be adapted them for classroom use. All these aspects make the resource families of corpora a great open educational resource, which can be a valuable addition to the teachers' toolbox, especially in the era of hybrid learning and teaching.

Nevertheless, several factors can make locating an appropriate corpus for educational purposes challenging. First, as Deshors (2021) points out, different corpora serve different purposes. While some specialised corpora are more suitable for research (e.g. [Louvain International Database of Spoken English](#)), others are used as teaching/learning resources (e.g. BNC, COCA). On the other hand, while larger corpora may provide more comprehensive coverage, more specialised corpora may be more suitable to answer more specific research questions. Second, the design and accessibility of a corpus may pose challenges for both teachers and learners.

For example, some studies show that [Lextutor](#), a text-based concordancer, may be too technical for learners. In contrast, [AntConc](#) has been found suitable only for adult learners with a high level of English proficiency. Other corpus query platforms, such as [SketchEngine](#), are powerful and user-friendly but available only via subscription, which could be a financial barrier for some universities. Finally, learning how to effectively use corpora in the classroom and design corpus-based materials to fit specific teaching objectives and students' levels of digital literacy can be challenging and time-consuming.

To help with these challenges, we have gathered several samples of CLARIN corpora and tools currently used in teaching and training by UPSKILLS consortium partners and other educators within the CLARIN community. Furthermore, we have provided links to available tutorials that can be repurposed for classroom use.

4.3.3.1. Computer-Mediated Corpora (CMC)

Teachers of language variation or pragmatics, may be interested in using computer-mediated communication corpora (CMC) in the classroom because they include informal writing styles. The resource family of CMC corpora⁴⁸ contains open CMC corpora in Slovenian, Czech, Dutch, Estonian, Finnish, French, German, Italian, and Lithuanian. Most corpora are tagged and available for exploration via integrated concordancers.

As a means of example, go to resource families and search for the following corpora: [sms4science](#) and [What's Up, Switzerland?](#). Both corpora have been compiled by Swiss researchers in the Swiss official languages and their varieties and are freely available for linguistic research through online corpus query platforms. The platforms allows users to

⁴⁷ To access the full functionalities of some of the concordancers available through the resource families, you might need to log in using your institutional credentials, as explained in section 3. *Accessing CLARIN*.

⁴⁸ <https://www.clarin.eu/resource-families/cmc-corpora>

perform simple and advanced queries, produce a frequency list and export the results. The corpora are also a great example of collecting, processing, cleaning, anonymising and annotating CMC corpora.

Another example of a well-documented CMC corpus is the *Twitter corpus Janes-Tweet 1.0*, a corpus of Slovenian tweets collected between 2013 and 2017. This corpus is tokenised and can be queried through [NoSketch Engine](#) and the [KonText](#) concordancers. Moreover, the corpus is published under a CC BY-NC 4.0 licence and [can be downloaded](#) from the CLARIN.SI repository. To learn how to use this corpus for research, see the PARTHENOS tutorial, [Using social media corpora in CLARIN](#).



For those interested in using CMC corpora for research purposes, we recommend the following learning resources:

- Go to PARTHENOS training module: [Digital Humanities Research Questions and Methods](#) (Bunout. et al., 2019) and read the following sections on CMC corpora:
 - ◆ Collections of Computer-Mediated Communication
 - ◆ Working with social media corpora
 - ◆ Boosting digital humanities research with CMC data
- Remember that on the website of the CLARIN K-Centre for CMC, you will find
 - ◆ FAIR best practices⁴⁹ for creating a CMC corpus, which actually can be adapted to any type of corpus, including guidance on the use of standards and formats.

4.3.3.2. L2 Learner Corpora

Learner corpora are a valuable pedagogical tool for the teacher-researcher because they contain spoken and/or written texts produced by language learners. The CLARIN resource family of [L2 learner corpora](#) contain 72 L2 corpora, out of which 11 are multilingual, while the rest are monolingual forms and various modalities (written, spoken, video)⁵⁰.

The [British Academic Written English Corpus](#) (BAWE) is often used to teach linguistic analysis of different genres and academic writing registers, helping students understand academic conventions and develop writing skills. The corpus is hosted in the [Oxford Text Archive Repository](#), and discoverable via the [Virtual Language Observatory](#).

⁴⁹ <https://cmc-corpora.org/ckcmc/docs/fair/>

⁵⁰ As per August 2023

Academic users can use it free of charge for academic purposes⁵¹, and query it in Sketch Engine or Lextutor. To learn how to use it in SketchEngine, see the detailed *Using SketchEngine with BAWE tutorial* on the [website](#), which consists of 7 lessons demonstrating how to make a simple concordance search, analyse collocations, learn how to use corpus query language and extract keywords. To make it easier for teachers and learners to use BAWE in the classroom, a collection of quicklinks to concordances in SketchEngine have been directly integrated as feedback directly in the students' assignments to help them improve their academic writing errors. This is a great example of how corpora and data-driven learning (DDL) are used to improve learner autonomy.

The [Das Digitale Wörterbuch der Deutschen Sprache \(DWDS\)](#)⁵² corpora, developed at the CLARIN-D centre Berlin-Brandenburg Academy of Sciences and Humanities, are used at the University of Kansas⁵³ as [an open educational resource](#) to teach German to English-speaking learners (Vyatkina, 2020b). The resource contains a guide to the main search functionalities of DWDS and a set of interactive corpus-based activities created in h5p⁵⁴. The learning activities have been designed based on guided induction, combining data-driven exploration with linguistic expertise to analyse the DWDS corpus. This pedagogical resource is a great example of how large open-source corpora in languages other than English can be integrated into the classroom (see Vyatkina, N., 2020, for more details).

Another tool worth mentioning here, although it is not a CLARIN tool, is [SkELL](#). This tool is a free online version of SketchEngine specifically developed for English language learners and provides a very simple search interface to look up words and their meanings in a one-billion-word web corpus. The tool includes a concordance tool, word sketches and a thesaurus functionality. The concordance results are presented in a sentence format. Teachers can use SkeLL to create exercises (gap-fill, matching and multiple choice) to teach students how to infer the meaning of words and phrases, analyse synonyms/polysemy, frequency, word association etc. For more examples of learning activities using SketchEngine and SkeLL, see Thomas (2017).

4.3.3.3. Comparable Corpora

Comparable corpora, defined as “a collection of texts composed independently in the respective languages and put together based on similarity of content, domain and communicative function” (Zanettin 1998:614), have extensively been used in educational settings. In translation teaching, for example, they are used as a pedagogical tool to help

⁵¹ Interested academic users need to register with Oxford Text Archive and agree to the terms and conditions before being able to use the corpus.

⁵² Digital Dictionary of the German Language

⁵³ Open Language Resource Center, <http://olrc.ku.edu/>. The centre focuses on the development of OER materials for teaching languages other than English at the secondary and post-secondary level.

⁵⁴ <https://en.wikipedia.org/wiki/H5P>

students enhance their understanding of the source language text and produce fluent translations.

ParlaMint is one of the most well-documented multilingual comparable corpora in CLARIN. It comprises 33 parliamentary corpora that cover most of the EU languages. These corpora are an important multidisciplinary language resource for social sciences and humanities researchers. They can be accessed through online corpus query platforms and are available in many languages. This makes it easy to integrate parliamentary corpora into classroom teaching.

For example, search in Parliamentary Corpora Resource Family for *ParlaMint.ana 3.0*⁵⁵. This collection contains a multilingual 26 comparable corpora of parliamentary debates, with each corpus being between 9 and 125 million words in size. What makes ParlaMint dataset easy to use both for research and teaching is that the corpora are tokenised and syntactically parsed using the Universal Dependencies (UD) framework⁵⁶, and annotated with named entities, which enhances the understanding and analysis of the data. The corpora are open and available through the [noSketch Engine](#)⁵⁷ concordancer, a user-friendly tool to introduce the students to concepts in corpus linguistics analysis. To use the corpus in another concordancer, you can download it (with or without linguistic markup) from the CLARIN.SI repository.

The corpora have been used in several editions of the [Helsinki Digital Humanities Hackathon](#),⁵⁸ various tutorials, university courses and student theses; see the examples below and more information on the [ParlaMint project information page](#). Here, we would like to highlight two tutorials demonstrating the use of the parliament corpora in research, which teachers can integrate into any course or programme involving modern languages, digital humanities, social sciences and corpus linguistics.

→ [Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research](#).

This tutorial uses the siParl 2.0 corpus to teach fundamental corpus linguistic methods to students and scholars of modern languages and learners from other fields, such as digital humanities and social sciences. After a brief introduction to corpora and corpus analysis methods and an introduction to the characteristics of parliamentary debates, the tutorial demonstrates how to use NoSketch Engine and KonText concordancers to analyse the corpora and explore topics female members of the parliament debate in the Slovenian Parliament and contrast their language use with their male counterparts.

⁵⁵ Erjavec, Tomaž; et al., 2023, Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1488>.

⁵⁶ UD is a framework that provides guidelines for consistent annotation of grammar across different languages, more info here: <https://universaldependencies.org/>.

⁵⁷ The corpora are openly available, so no login is required to use the noSketch for linguistic analysis. noSketch is the open-source implementation of the commercial SketchEngine tool. If you are new to these tools, see the SketchEngine Quick Start Guide: <https://www.sketchengine.eu/quick-start-guide/>.

⁵⁸ [Semparl - Cities in parliament](#); [Parliamentary Debates in COVID Times](#)

The tutorial takes about five hours to complete, and it has been included in the curriculum of *Corpus Linguistics and the Use of Language Technologies in Lexicography* course at the Postgraduate School of the Research Centre of the Slovenian Academy of Sciences and Arts.

- [What's on the agenda? Topic modelling parliamentary debates before and during the COVID-19 pandemic](#). This tutorial can be used to introduce students to basic text-mining concepts by applying topic modelling to a comparable corpus of parliamentary debates. The ParlaMint-GB corpora is used in the tutorial, which students can download from the CLARIN.SI repository. All the techniques demonstrated in the tutorial are language-independent, which means that they can also be applied to parliamentary corpora in other languages.

Both tutorials can be reused under a [CC-BY-NC-ND 4.0 licence](#), which means that they can be repurposed for classroom use provided that the authors are given appropriate credit. The licence prohibits the reuse of the tutorials for commercial purposes. If modified, the materials cannot be distributed outside the classroom.

The [ORVELIT](#) corpus is a comparable monolingual corpus of original and translated **Lithuanian** consisting of four sub-corpora of original and translated fiction and popular science. Although the corpus is not available through a concordancer, it can be downloaded from the CLARIN-LT repository (both in a raw and morphologically annotated version). The corpus is being used in the curriculum of the MA course in “Contrastive Stylistics” at Vytautas Magnus University. Students get acquainted with building procedures and characteristics of the ORVELIT corpus, after which they are encouraged to think of possible research questions and registers that would be interesting to explore using the corpus.

In addition, students learn to create a **basic research data management plan**. Other learning activities include downloading the morphologically annotated and raw versions of the corpus to investigate the features of original and translated Lithuanian independently with the help of a corpus analysis tool. Finally, students learn to generate and compare wordlists, keyword lists and concordances and discuss their findings on the similarities and differences between original and translated texts. The ORVELIT corpus has been used to design learning activities about the use of Lithuanian collocations in teaching, learning and translation in this online resource book:

- Kovalevskaitė J., Rimkutė E., Vaičenonienė J. 2022: [Lietuvių kalbos kolokacijos: vartojimas, mokymas\(is\) ir vertimas](#) (Lithuanian Collocations: Usage, Teaching, Learning, and Translation). Kaunas: Vytauto Didžiojo universitetas. ISBN 978-609-467-524-9, <https://doi.org/10.7220/9786094675249>.

4.3.3.4. Parallel Corpora

Parallel corpora are defined as a “collection of texts in language A and of their correspondent translations into language B” (Baker, 1995). They can be bilingual or multilingual and contain published translations. Parallel corpora are often used in corpus-based contrastive linguistics and translation studies (Levshina, 2016). In translation studies, bilingual concordances can be used to analyse cross-linguistic correspondences between lexical items, syntactic patterns or grammatical structures and discuss translation strategies for proper names and culture-specific elements (Lefer, 2020, p.263). In the practical translation classroom, parallel corpora are also used to teach students how to import it into their computer-aided translation tool and use it as a translation memory (e.g. using corpora from [OPUS](#)) or extract bilingual terms using such tools as [Synchrotem](#) to create bilingual glossaries. Furthermore, parallel corpora are used in the machine translation/AI classroom to teach students how to train statistical machine translation engines.

The CLARIN resource family provides access to 87 parallel corpora,⁵⁹ out of which 5 contain language data in more than **50 languages**. Here, we highlight those that can be easily integrated into the classroom. Well-known parallel corpora used in corpus-based translation/contrastive studies and NLP are: *EuroParl*, *DGT-Translation Memory*, *European Central Bank parallel corpus*, *Opus*, *Helsinki*, *ParaCrawl*.

[InterCorp](#), a parallel corpus of more than 40 languages, is often used for comparative research in foreign language teaching, translation studies, theoretical studies, and NLP (Čermák, 2019). The texts come from various sources, such as fiction, EU legal texts, film subtitles, and the Bible. The corpus can be queried via [KonText](#). The Wiki page of the Czech National Corpus provides a nice [tutorial](#) consisting of 11 lessons about using KonText and performing queries in a parallel corpus (InterCorp), spoken corpus, diachronic corpus and syntactically annotated corpus. Each lesson contains clear examples of queries and an exercise at the end.

[Compara](#) is a parallel corpus of English and Portuguese, and it can be queried in a free and user-friendly parallel concordancer. The corpus is used in translation studies for descriptive and empirical research and by lecturers to prepare exercises and discuss translation problems with students.



Unfortunately, few concordancers are available for parallel corpora. In the CLARIN resource family of [corpus query tools](#), we found AntPConc (free desktop-based tool), ParaConc (commercial tool), and SketchEngine (commercial).

⁵⁹ <https://www.clarin.eu/resource-families/parallel-corpora>



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

There are numerous parallel corpora available in the CLARIN repositories. To teach students how to use the Virtual Language Observatory to find a parallel corpus for a translation assignment, see

→ [3.6. Finding and Using a Parallel Corpus for a Translation Assignment](#)

Moodle learning resource 8: Finding a Parallel Corpus

4.3.3.5. Multilingual Web Corpora

In June 2023, a new collection of large web corpora was launched, called CLASSLA-web⁶⁰. This collection is specifically for **Slovenian, Croatian, and Serbian** languages, and can be queried using the [NoSketch Engine](#) concordancer of CLARIN.SI. To learn how to analyse word usage in contexts, collocations, dictionary examples, and more, visit the [tutorial on the CLARIN.SI website](#). The corpora consist of professional and informal texts, such as forums and blog posts, harvested from the web. This provides linguists and corpus linguists with valuable insights into non-standard language use. Additionally, the monolingual datasets have been linguistically annotated to make them more usable. The previous **Bosnian/Croatian/Montenegrin/Serbian** Web corpora (srWaC, hrWaC, bsWaC, meWaC) were used in the MA course, *From computational linguistics via clinical linguistics to forensic linguistics*, at the **University of Graz**. The corpora can be downloaded from the [CLARIN.SI repository](#) and queried in [NoSketch Engine](#) and [KonText](#).

[Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics](#) (Krasselt et al., 2020) contains about 8 million texts in **German, French and Italian** from **selected resources on the web** (i.e. news and specialist publications, governmental opinions, and parliamentary records, web sites of political parties, companies, and universities, statements from industry associations and NGOs). Unlike previous web corpora (e.g. WaCky⁶¹), Swiss-AL promotes **data-based and data-driven research** on societal and political discourses in Switzerland, not for NLP purposes. It can be queried via [Swiss-AL Workbench](#) and via [CQPweb](#).



If you are interested in multilingual corpora in other languages, you can find a curated collection on the [Wiki of the Association for Computational Linguistics](#).

⁶⁰ Each CLASSLA web corpus contains about 2 billion words.

⁶¹ Baroni, M. et al. (2009). The Wac corpora are also available for free via CLARIN.SI repository: <https://www.clarin.si/ske/#open>

4.3.3.6. Spoken Corpora

Spoken corpora are compiled as audio and text transcriptions for linguistic purposes, including hypothesis testing and language teaching and for developing grammars and dictionaries. In research, these tools serve various purposes, such as phonetics, conversation analysis, grammar, pragmatics, dialectology, language acquisition, and the creation of acoustic models for speech technology. Additionally, they are used to compile corpora of endangered languages in language documentation. For a discussion about the challenges in compiling spoken corpora see Gut (2020, p.249).



See the [Language Archive](#) repository for examples of speech corpora from worldwide languages.

As of August 2023, there are 133 spoken corpora⁶² in the CLARIN resource families, mostly monolingual in 15 languages. Some corpora are available for querying directly in KonText or Korp. Here, we would like to highlight a few spoken corpora used in educational settings.

[TalkBank](#) is a CLARIN K-Centre offering the world's largest open-access integrated spoken-language data repository. It provides language corpora and audio resources to support researchers in various fields of study, including Linguistics. Academic users can transcribe sound files using the CHAT⁶³ standard and analyse them in CLAN⁶⁴. A [user-friendly guide](#) is available for teachers and students who want to learn how to use CLAN and Praat to analyse speech data. The Slavic data collection from CHILDES corpora is used in the MA, PhD *Language Acquisition in Slavic: Obtaining, Representing and Analysing Empirical data in Linguistics* course at the University of Graz to help students enhance their research skills in language acquisition, as well as problem-solving and data-analysis skills.

The web-based LABLASS platform and [Bulgarian LabLing Corpus](#) are included in the curriculum of the linguistic disciplines at Konstantin Preslavsky University of Shumen, Bulgaria. The corpus has been published on the CHILDES platform to facilitate cross-linguistic research, including other **Slavic languages**. Students work in groups on specific projects involving recording and transcribing children's speech in the CHILDES universal format and data collection tasks in the LABLASS system.

The [Database for Spoken German \(DGD\)](#) is used in the German Linguistic department at the University of Mannheim to teach BA and MA-level students how to analyse the usage of interaction signs, such as interjections and compare their frequency to written language data, such as postings on Wikipedia talk pages that are available via the Corpus Search, Management and Analysis System (COSMAS II). The latter is often used for

⁶² <https://www.clarin.eu/resource-families/spoken-corpora>

⁶³ The CHAT user manual can be accessed here: <https://talkbank.org/manuals/CHAT.pdf>

⁶⁴ The CLAN user manual can be accessed here: <https://talkbank.org/manuals/CLAN.pdf>

cross-lingual studies. Students also explore collocations and create word development curves and word profiles with the help of corpus tools.

Teachers teaching Dutch as a second language may be interested in the [OpenSoNaR corpus](#), which contains written and spoken corpora of Dutch from the Netherlands and Flanders (500 million words). The corpus is used for linguistic and human language technology research, and the development of NLP applications. It is easily accessible through a user-friendly interface (see Fig. 9) that offers basic and advanced functionalities for corpus queries, including regular expressions. Tutorials in Dutch are available in open access on Surf.nl⁶⁵, and video recordings on Vimeo⁶⁶. You can access the corpus using your university credentials; otherwise, apply for a CLARIN ERIC account.

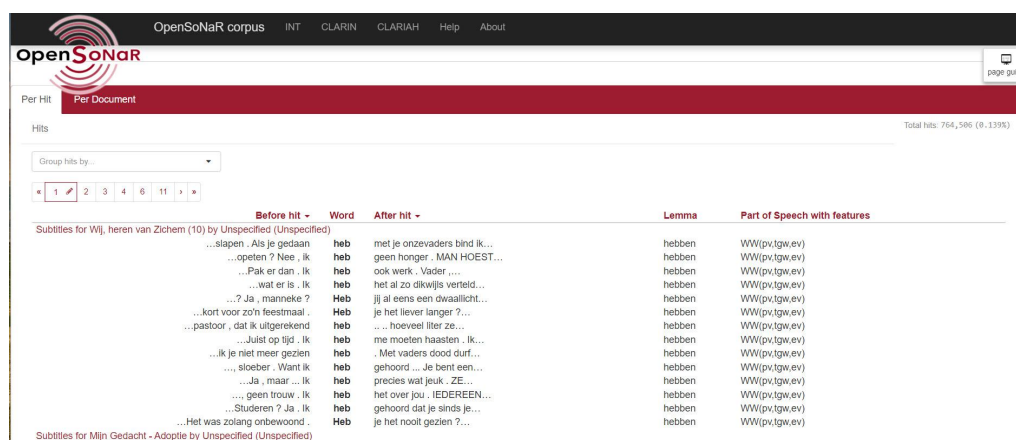



Figure 9: OpenSoNaR Search Interface

To sum up, the resource families provide an excellent opportunity for teachers and educators to enhance their students' learning experience through "data-driven learning" (Perez-Paredes et al., 2022). Many of the corpora listed in the resource families are open and available via concordancers, which makes them an excellent open educational resource. This allows for a broader scope of instruction beyond conventional materials.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

If you want to learn more about speech technologies, we recommend the [Automatic Speech Recognition and Forced Alignment](#) learning block on the UPSKILLS Moodle, designed by Louis ten Bosch and Henk van den Heuvel from Radboud University, Nijmegen.

Moodle learning resource 9: Introduction to Automatic Speech Recognition

⁶⁵ <https://surfdrive.surf.nl/files/index.php/s/HAuOPmTgwnCHtpx>

⁶⁶ <https://vimeo.com/channels/clariahlearn/>

4.4. Compiling, sharing and citing virtual collections

Language resources discovered in the VLO, other research data repositories, or simply on the Web can be added to a virtual collection using the [Virtual Collection Registry](#) (VCR) for later exploration and processing. These virtual collections can be shared with others and cited using persistent identifiers, e.g. handle and/or DOI. Furthermore, plain text files can be easily processed with **Language Resource Switchboard** for other types of linguistic tasks, such as annotation, translation, or visualisation. See the quick guide below, Fig. 10, to learn how to create a virtual collection.

Collecting Data from the VLO

Creating and Citing a Virtual Collection

- 1

Go to the VLO and search for text data in English, text/plain format, public licence
- 2

Select resources, analyse the metadata, availability, licences and queue the text file to VCR
- 3

Submit all the collected files to VCR
- 4

Log in to VCR using the credentials of your home organisation or CLARIN.eu account
- 5

Create a new virtual collection: name, description, keywords, authors
- 6

Save, publish, share and cite the collection

Figure 10: Collecting Data from the VLO and Creating a Virtual Collection

When used in the classroom, the VCR can be useful in the following scenarios:

1. Before the class, the VCR can be used to create a collection around a specific topic that you would like to teach, e.g. links to exemplary datasets, tools and tutorials that you would like to share with the students. For example, a few virtual collections that have been compiled for demonstration purposes:
 - A collection of books, articles, tutorials and guidelines on Research Data Management in linguistics: DOI [10.34733/vc-1078](https://doi.org/10.34733/vc-1078).

- A collection of tutorials and user guides for the NLP tools available via the Language Resource Switchboard: [DOI 10.34733/vc-1079](https://doi.org/10.34733/vc-1079).
 - A collection of Jane Austen's works in plain text format, which can be used for linguistic analysis with Switchboard tools, or downloaded and used to compile a corpus with corpus building tools, such as BootCat or SketchEngine: DOI [10.34733/vc-1080](https://doi.org/10.34733/vc-1080).
2. When students need to collect language data for a specific project, they could use the Virtual Language Observatory or other CLARIN national repositories to search for datasets, digital text collections, and corpora, add them to a virtual collection and save them for later analysis. They can then share the collection with the teachers, who can evaluate the metadata quality and give feedback on the selection of resources. The collections can then be cited with the help of persistent identifiers. Reusing digital texts from repositories and adding them to a virtual collection may save some time because the students do not have to manually search the Web, download, convert files to text format and note down each file's metadata in a spreadsheet.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

Use the following interactive presentation to teach students how to create and cite virtual collections:

- *3.4. Collecting, Citing and Processing Language Resources from Data Catalogues*

Moodle learning resource 10: Creating and Citing Virtual Collections

4.5. Finding tools for data processing and analysis

Nowadays, there is a wide variety of applications to discover, explore, analyse and annotate language data. Selecting a tool suitable for teaching language data science might be challenging. In CLARIN, tools can be discovered via the following paths:

1. The Virtual Language Observatory is a discovery platform not only for language resources and digital text collections but also software and various tools, e.g. the source code for open-source translation tools (e.g. [LINDAT Translation service](#)), language models for training machine translation engines (e.g. <https://hdl.handle.net/11234/1-3732>), [corpus taggers](#), term extraction tools, OCR tools and file conversion tools.
2. The [Language Resource Switchboard](#) connects to several NLP tools developed within the infrastructure, and it can be used to find a matching tool for a specific text type and language. More details about this service are provided in the section below.

3. [The CLARIN Resource Families](#) contain both corpora and corpus query tools and a curated collection of NLP tools for [normalisation](#), [named entity recognition](#), [PoS tagging](#), [lemmatisation](#) and [sentiment analysis](#).
4. Because not all tools developed in the CLARIN member countries are discoverable via the above services, we recommend also checking the websites of the national consortia.

Below, we briefly introduce Language Resource Switchboard and point to other useful NLP tools and applications that are often used in educational settings.

4.5.1. Using the Language Resource Switchboard for text processing

For teachers new to incorporating natural language processing tools in their classroom and uncertain about which tools to choose for text or file processing and analysis, exploring the Language Resource Switchboard could prove beneficial. The service can process digital text collections found via the Virtual Language Observatory or other research data repositories (e.g. TextGrid) with matching NLP tools available through the infrastructure. See Fig. 11 for the list of tools that connect to the Switchboard. Note that some tools require authentication with your institutional credentials. See [Section 3.1. Accessing CLARIN](#) for instructions.

Tool Inventory	
> Constituency Parsing	> Morphological Analysis
> Coreference Resolution	> Named Entity Recognition
> Dependency Parsing	> Named Entity Relation Detection
> Distant Reading	> Part-Of-Speech Tagging
> Extraction of Polish terminology	> Sentiment Analysis
> Inclusion detection	> Shallow Parsing
> Keyword Extractor	> Spatial expression detection
> Lemmatization	> Speech Recognition
> Machine Translation	> Stylometry
> Metadata Processing	> TF, IDF, TF-IDF calculation
> Morpho-syntactic tagger	> Text Analytics
	> Text Enhancement
	> Text Summarization

Figure 11: NLP tools accessible via the Language Resource Switchboard



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

Before using Switchboard, teachers may need to introduce students to basic NLP methods for text analysis.

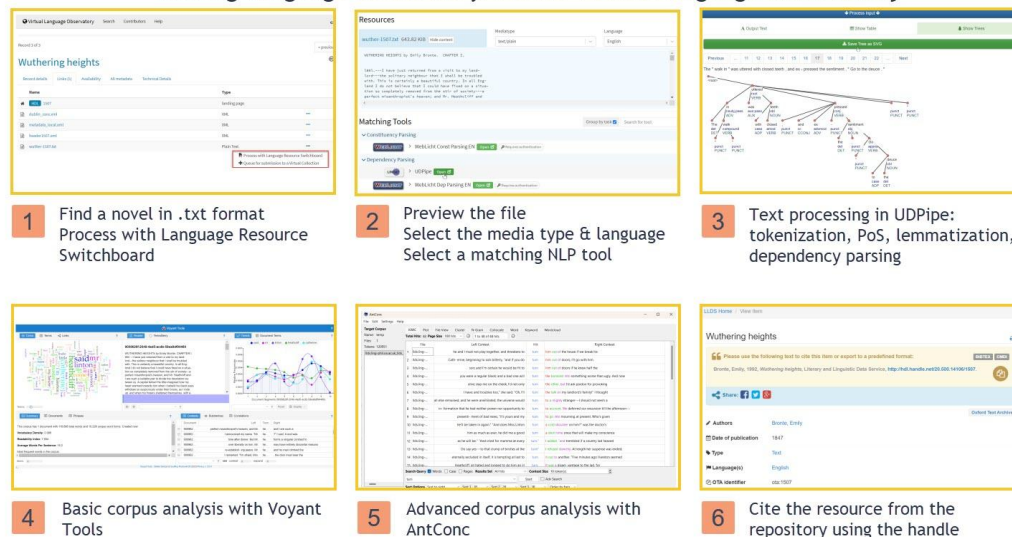
→ 3.2. *Basic Introduction to Natural Language Annotation for Linguists*

Moodle learning resource 11: Basic Introduction to NLP Annotation

The quick guide in [Fig. 12](#) shows how to process a plain text file⁶⁷ from the VLO with Switchboard directly from the search results interface. In Switchboard, depending on the level of the students and the teaching goals, teachers could use [UDPipe](#) to demonstrate automatic NLP tasks or choose Voyant Tools to introduce students to basic text analysis in a visually appealing way. To teach more advanced corpus query functionalities, the file should be downloaded from the repository where it is located and uploaded to AntConc or SketchEngine. Finally, the language resources findable and processed via the VLO and Switchboard tools can be cited using the handle assigned by the repository. UDPipe and Voyant tools are described in more detail in the next paragraphs.

Text Processing & Analysis

Reusing Language Resources from the Virtual Language Observatory



- 1 Find a novel in .txt format
Process with Language Resource Switchboard
- 2 Preview the file
Select the media type & language
Select a matching NLP tool
- 3 Text processing in UDPipe:
tokenization, PoS, lemmatization,
dependency parsing
- 4 Basic corpus analysis with Voyant
Tools
- 5 Advanced corpus analysis with
AntConc
- 6 Cite the resource from the
repository using the handle

Figure 12: Text processing and analysis with Switchboard

⁶⁷ The file used for the demo is available at: <https://hdl.handle.net/20.500.14106/1507>.

Below, we highlight more tools accessible directly from Switchboard that may be suitable for classroom use.

4.5.1.1 Corpus analysis and visualisation

In [Section 4.3.3](#), we presented a few available corpora for browsing and linguistic exploration in tools such as NoSketch Engine, Korp, and KonText. Below, we would also like to mention Voyant Tools, which are directly accessible via the Language Resource Switchboard.

[Voyant Tools](#) is a user-friendly web-based reading and analysis environment for digital texts offered by CLARIN-DK and is suitable for teachers and students with no technical background. The tools can also be downloaded and run on local computers. Teachers and students can use this online environment to analyse automatic text with functionalities such as word frequency lists, frequency distribution plots and KWIC displays.⁶⁸ Students can use the tools to analyse an online collection of journals, blogs, and websites and include the link to their analysis directly in their research reports (if these are published online) so that the readers can view the results.



See the Resource Families of [Corpus Query Tools](#) for more examples of corpus-query tools and platforms. The overview includes desktop and online applications, covered languages, and links to user guides.

4.5.1.2 Automatic annotation of texts

[UDPipe](#) is a language-independent software that produces custom annotation pipelines for tokenisation, tagging, lemmatisation and dependency parsing of CoNLL-U files⁶⁹. The tool is based on the Universal Dependencies (UD) framework,⁷⁰ which produced about 200 treebanks (consistently annotated by human annotators) for over 100 languages. The framework is very popular because it is intuitive and does not adhere to any formal theory. LINDAT CLARIAH-CZ infrastructure provides UDPipe as a free web service for testing services. UD and the CoNLL-U format are also supported by other annotation tools, such as the BRAT rapid annotation tool⁷¹, WebAnno, and SketchEngine. An international team of volunteers maintains the framework, and everyone can join them to start building their own corpus through the [TEITOK platform](#). UDPipe and the TEITOK UD 2.11. corpus are used in the NLP courses at the Institute of Formal and Applied Linguistics at Charles

⁶⁸ [CLARIN-DK presents: Teaching the teachers – an interactive workshop for the Voyant Tools | CLARIN ERIC](#)

⁶⁹ [CoNLL-U Format \(universaldependencies.org\)](#)

⁷⁰ Zeman, Daniel; et al., 2023, Universal Dependencies 2.12, 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/11234/1-5150>

⁷¹ An online environment for collaborative text annotation, available for free at <http://brat.nlplab.org/>.

University to teach MA students morphological analysis.⁷² Students acquire NLP skills by participating in real-life research projects⁷³ related to the development and training of machine translation engines, corpora, lexicons, speech and dialogue systems, and text processing and semantics.



For ready-made training materials to introduce students to UD, check the following learning resources:

- The course [Dependency Grammars and Treebanks](#), offered by Charles University, includes a syllabus, lecture slides and information about the practical lab sessions.
- This MOOC course shared via the CLARIN learning hub: [Applied Language Technology](#) by Tupmo Hiipala (University of Helsinki). Part III briefly introduces NLP concepts and universal dependencies.

[WebLicht](#) (“Web-based Linguistic Chaining Tool”)⁷⁴ is a user-friendly web service that can be used to teach and demonstrate **automatic annotation of texts** in English, German, Dutch, French, and Italian. It is included in the practical lab sessions on Corpus Linguistics at the University of Saarland, Germany, to teach corpus annotation. The service is integrated with the CLARIN infrastructure and accessible from Switchboard, Federated Content Search and the Virtual Collection Registry. Several NLP tools (e.g. sentence splitting, tokenisation, lemmatisation, POS tagging, morphological analysis, named entity recognition, dependency parsing, and constituency parsing) are integrated to help researchers create and visualise custom processing chains. See Fig. 13 for an example of dependency parsing in WebLicht⁷⁵. The workflow is explained in detail in Hinrichs (2022). Users can access the service through their academic institutions.



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

Use this tutorial from Moodle to teach students how to annotate a raw collection of text corpora from the GitHub repository with WebLicht.

- *How to Annotate Text Collections with WebLicht*

Moodle learning resource 12: Annotation with WebLicht

⁷² The course syllabus, lecture slides, assignments and exam questions are free to view online: https://ufal.mff.cuni.cz/courses/npfl124#lectures_practicals. It is being offered both to BA and MA students.

⁷³ <https://ufal.mff.cuni.cz/projects>

⁷⁴ WebLicht is hosted by the [CLARIN centre at the University of Tübingen](#).

⁷⁵ For more WebLicht tutorials, see this virtual collection: <https://doi.org/10.34733/vc-1079>.

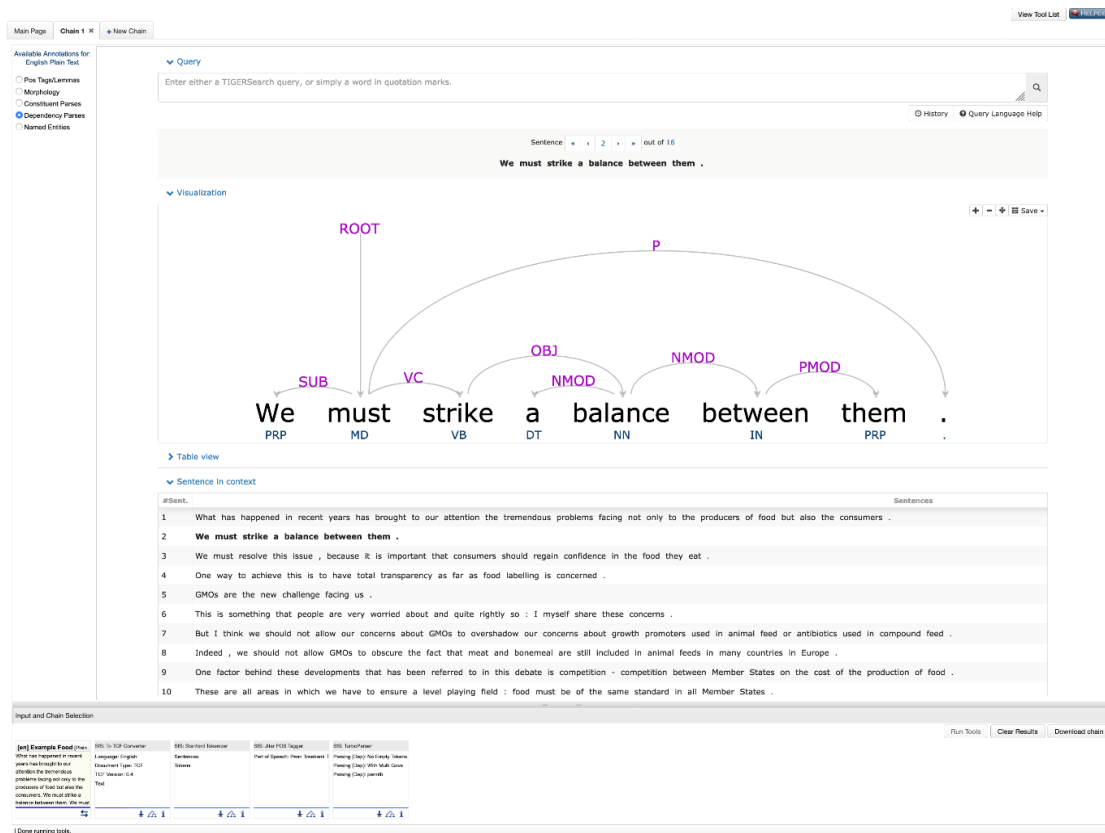


Figure 13: Dependency parsing in WebLight

4.5.1.3 Manual annotation of texts

[WebAnno](#) is a user-friendly open-source web-based annotation tool⁷⁶ for manual linguistic annotation tasks (e.g. morphological, syntactical, and semantic annotations) that can be introduced in teaching at the BA and MA level to introduce the students to annotation. The teacher can conduct multiple annotation projects in parallel and assign students different roles, e.g. annotator, curator, or project manager. Students can annotate in groups following the full annotation workflow, see Fig. 14. Machine learning capabilities are integrated to make the work of the curator/reviewer easier. For example, WebAnno learns from pre-annotated data and makes suggestions. The annotator can accept or reject the suggestions with a single click. The suggestions help improve the training data. The project manager can assign workload to annotators and monitor their projects. The tool is also offered as a service to members of research institutions, e.g. CLARIN-D and Fin-CLARIN.

⁷⁶ It is platform independent Java-based application.

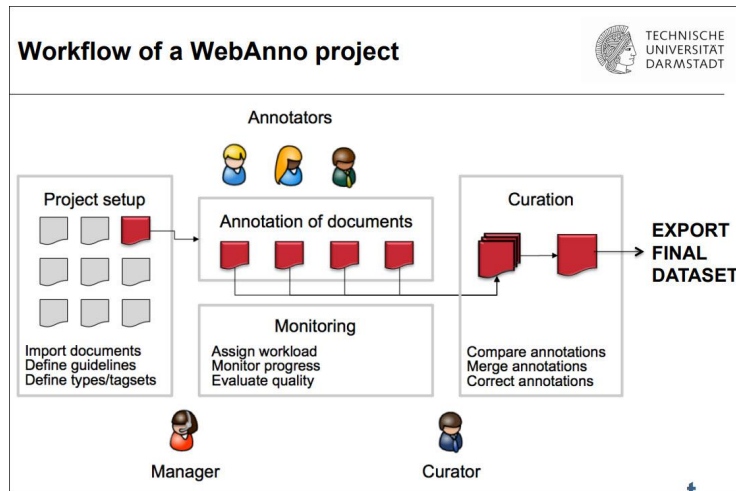


Figure 14: Workflow of a WebAnno project (Castilho, 2014)

[INCEpTION](#) is an upgraded version of WebAnno supported by UKP Lab at TU Darmstadt. This text-annotation platform offers a user-friendly interface for various collaborative semantic annotation tasks on written text, mostly for linguistic and machine-learning purposes. The platform is available both in a desktop version and online, and it includes a [recommender system](#) to assist in creating annotations quickly and easily. Additionally, it enables corpus creation by searching external document repositories and importing documents. Due to its user-friendly interface, user guide and tutorials, it can be used in the classroom to create semantic annotation projects with the students. INCEpTION is available as a [service](#) to CLARIN-EL members. It is also used by the CORLI K-Centre for Corpora⁷⁷ for collaborative annotation.



Figure 15: Examples of annotations in INCEpTION

⁷⁷ <https://corli.huma-num.fr/kcentre/>

4.5.1.4 Annotation of speech

WebMaus is a web service developed by the Bavarian Archive for Speech Signals, and it can be used to **automatically align transcriptions and speech signals**. The results can be used in [Praat](#). The service is part of the suite of [BAS web tools for speech processing](#)⁷⁸ and provides word and phoneme alignment for more than 25 languages. Watch [this short YouTube tutorial](#) to learn more.



Teaching and Learning Resources on Moodle

To learn more about automatic speech recognition, see the UPSKILLS Learning content on Moodle:

→ [Automatic Speech Recognition and Forced Alignment](#) (6 ECTS).

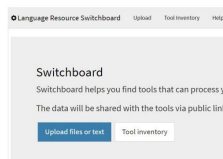
Moodle learning resource 13: Introduction to Automatic Speech Recognition



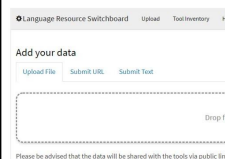
Go to the Language Resource Switchboard and try the service by following the steps in the quick guide in Figure 16. Remember that the resources in plain text format from the VLO can be submitted to Switchboard directly from the VLO interface.

Finding NLP Tools

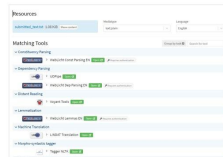
Language Resource Switchboard




1 Go to **Language Resource Switchboard**. Click on the **Tool Inventory** to view all the NLP tools connected to this service.



2 Upload a text file or submit a URL to an online text or a Handle to a text collection.



3 Select a matching tool from the list to process your text. Use your university credentials to access the tools that require authentication.



4 Example of a syntax tree produced with UDPipe.

Created by Ivan der Lek | 06 August 2023 CLARIN in UPSKILLS

Figure 16: Finding NLP Tools in Language Resource Switchboard

⁷⁸ The BAS Web Services page contains a list of tutorials and use cases: [BAS Services Tutorials](#).

1. Language Resource Switchboard automatically recognises the format and language of input files.
2. It suggests a list of data processing tools categorised by task.
3. Select a tool from the list. If authentication is required, use your university credentials.
4. Clicking on a tool provides additional information and opens it in a new browser tab. Your file is processed automatically and can be downloaded for further analysis.

Switchboard can also be accessed from other research data repositories, e.g. CLARIN [VLO](#), CLARIN [Virtual Collection Registry](#), [ARCHE](#), [DARIAH-DE Repository](#), [PARTHENOS VRE](#), and [TextGrid](#).



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

Use the following presentations and tutorials to teach students how to use the Switchboard tools to process digital text collections:

Presentations:

- 3.2. *Basic Introduction to Natural Language Annotation for Linguists*
- 3.3. *Finding Tools in CLARIN To Process Digital Text Collections*

Tutorials:

- 3.7. *How to Annotate Text Collections with WebLicht*
- 3.8. *How to Annotate Text Collections in TEI-format*

Moodle learning resource 14: How to Find and Use NLP Tools

4.5.2. Other NLP tools and applications

To teach students how to compile and analyse large text corpora using programming languages but you do not know where to start, see these UPSKILLS courses on Moodle:

- [The Essence of Machine Learning for Linguists in Tech](#)
- [Start Programming with Python in 10 Steps](#)

[Jupyter Notebooks](#) provide a popular environment for programming, especially in educational contexts, to teach coding and concepts such as topic modelling. To understand how they can be used both for research and teaching, we recommend:

- Quinn Dombrowski, Tassie Gniady, and David Kloster. (2019) “Introduction to Jupyter Notebooks,” *Programming Historian* 8, <https://doi.org/10.46430/phen0087>.

The tutorial demonstrates how to write a Jupyter notebook for data analysis as part of a research project and then adapt it for classroom use.

The CLARIN centres have also implemented notebooks in their infrastructure to support researchers in data analysis as part of their research projects. A collection of notebooks is available [here](#). For example, [Portulan CLARIN](#) has implemented various NLP tools via a [workbench](#) (e.g. quantitative tools for syntax) and includes examples and documentation to help researchers and students design an experiment with Jupyter Notebooks.⁷⁹

CLARIN Center of Estonian Language Resources⁸⁰ has developed the EstNLTK python package for processing Estonian, as well as a series of [tutorials](#) in Jupyter Notebooks on different NLP components, such as text segmentation, morphological processing, syntactic analysis, word embeddings, etc. For educational materials in Estonian, please check the NLP course taught at the University of Tartu, available on [GitHub](#). EstNLTK can be installed in [Google Colab](#), which implements Jupyter Notebooks within the Google suite. This allows learners to create a document containing executable code, which is stored on GDrive and can be shared with peers for editing and commenting. It may be a more user-friendly option than GitHub for classroom use. To learn more about EstNLTK, see this [post](#) in Tour de CLARIN.

Finally, some other helpful teaching and learning resources are:

- [Introduction to Programming for NLP with Python](#), a web-based course taught by Koenraad de Smedt at the University of Bergen⁸¹. This course focuses on text processing and data analysis related to linguistics, language studies, digital humanities and cognitive science. Students learn basic programming skills using Jupyter Notebooks that combine Python code examples with explanatory text. The course is suitable for teaching at the BA level.
- The Natural Language Toolkit (NLTK) and tutorial, [Natural Language Processing with Python](#) (Bird et al., 2009), are often used as teaching material in programming courses both at the undergraduate and graduate level. After students have acquired knowledge of basic text processing and corpora, you could use this tutorial to teach students how to access and analyse large text archives on the web (e.g. Project Gutenberg), various types of corpora, and lexical resources. The tutorial contains a list of exercises of different difficulty levels that can be integrated into the classroom.
- Another platform often used in NLP training and education is the **GATE NLP toolkit** (Maynard et al., 2021). Training materials and other useful resources are available on the [CLARIN Learning Hub](#).

⁷⁹ A great collection of open-source course materials on various NLP topics, including the use of Jupyter notebooks in educational settings, is available on GitHub: <https://github.com/quinnanya/dh-jupyter>.

⁸⁰ <https://www.keeleressursid.ee/en/resources/text-processing-tools>

⁸¹ The course has been shared via the Teaching with CLARIN call 2023. It will be presented at the CLARIN Annual Conference, 16-18 October 2023.

- To teach programming to students and researchers with a strong background in the humanities, the [Humanities Data Analysis: Case Studies with Python](#) (Karsdorp, F. et al., 2021) guide may be a useful resource. It contains a series of case studies of data-intensive humanities research using Python programming language in Jupyter Notebooks to **gather, clean, represent, and transform textual and tabular data**. Exercises and resources are included, which turns the guide into a great educational resource.



After acquiring basic programming skills, learn how to use Jupyter notebooks⁸² and CLARIN NLP tools to process large collections of texts from Virtual Language Observatory with the help of this tutorial:

- *Jupyter Notebooks for Europeana Newspaper Text Resource Processing with CLARIN NLP Tools*. Version 1 Retrieved Aug 8, 2023, from the SSH Open Marketplace:
<https://marketplace.sshopencloud.eu/training-material/duVII1>.

All in all, this section provided an overview of CLARIN infrastructure services, demonstrating how they can be used in educational settings based on best practices in the UPSKILLS and CLARIN communities. Ready-made learning and teaching materials are also available on Moodle, which teachers can use to integrate as tutorials or handouts in the classroom. The VLO, Federated Content Search, and Virtual Collection Registry can be used to search and locate language data that can be analysed with matching tools from the Language Resource Switchboard. The Resource Families give access to many types of open corpora through user-friendly concordancers, which can be used freely by anyone involved in language research and teaching.

👉NB: *If you need training on using the services and tools described in this section, email the CLARIN training officer at training@clarin.eu. We also encourage you to check the [CLARIN learning hub](#) for an overview of training materials and events and to watch the upcoming [CLARIN cafes](#). If you have questions or more specific training needs, please contact the representative of the [User Involvement Committee](#) or the [national consortia](#) in your country.*

⁸² For an introduction to Jupyter notebooks and their added value for teaching, see this tutorial: Quinn Dombrowski, Tassie Gniady, and David Kloster, "Introduction to Jupyter Notebooks," *Programming Historian* 8 (2019), <https://doi.org/10.46430/phen0087>.

5. Teaching Linguistic Research Data Management

This section is relevant for teachers who plan to include essential **FAIR-research data management skills** in their linguistic, translation and other language-related courses or programmes. As demonstrated throughout this guide, if language resources and tools are documented well, curated, published and archived in a discipline-specific research data repository, they are easy to find, validate and reuse for research and educational purposes. Moreover, research data management and responsible data sharing have become mandatory in doctoral programmes and for researchers participating in [Horizon Europe](#) projects. Therefore, various initiatives at the [EU level](#), such as EOSC and FAIRsFAIR recommend that universities include basic skills for open science and research data management across all domains, disciplines and levels. The integration of research data management is generally achieved by following the FAIR guiding principles for data management, which ensures that *research data is findable, accessible, interoperable, and reusable*. Nevertheless, there are still numerous challenges due to the lack of consensus around the notion of “data”, especially in linguistics (Good, 2022).

Traditional arts and humanities programs have yet to emphasise the importance of research data management and computer literacy skills. However, in contemporary education, marked by increasing interdisciplinarity, such skills are becoming indispensable regardless of one’s career path. As many lecturers have already pointed out during UPSKILLS multiplier events and in testimonials, students of language and linguistics often struggle with basic computer skills at the MA level, e.g. organising data in an Excel file, recognising the difference between basic file formats, zipping and unzipping files, and/or understanding technical concepts, such as “annotation”. This lack of basic computer literacy at any level of study can impact the students’ learning experience and the lecturer's initial course design and plan (Simonovic, M. et al., 2023).

To prepare language and linguistics students to work with research data and engage in data-driven projects, they can be recommended to take an introductory ICT course at the start of their BA programme and a general research data management course after they learn what research data is. Usually, RDM courses are offered via the university library, are not discipline-specific and often target students in doctoral programmes.



If your institute does not offer general RDM courses, curriculum designers are referred to the case studies documented in *Good Practices in FAIR Competence Education* (Garbuglia et al., 2021) for inspiration on how to design such programmes. Additionally, check our own [Existing Learning Materials Surveyed Within UPSKILLS](#) and the [SSH Training Discovery Toolkit](#) for training resources in open access.

Moreover, in the first years of their BA studies, students can be introduced to basic data management skills by simply teaching them how to organise their files on their computers in a meaningful way, handle their personal data carefully, share files responsibly via secure cloud-based platforms, perform regular backups and archive their work at the end of a project or semester for possible future reuse.

After students have acquired basic data management skills, teachers can engage them in domain-specific (research-based) data-driven projects, e.g. produce a dataset to answer a research question, add new annotation types to an existing corpus, compile a corpus for an under-resourced language, or develop a language model. Students can work on such projects collaboratively in the classroom, remotely, or as part of their final BA/MA theses under the teacher's supervision. The CLARIN infrastructure can be used to find guidance on standards, formats, and how to handle legal and ethical issues when the projects involve working with personal and sensitive data. At the end of the project, teach students how to find a suitable repository for their data and deposit, publish and share their corpus with the research community.



For guidelines on depositing general research outputs and language resources in a research data repository, see Section 6 in the *Guidelines for Student Projects and Research Reporting Formats* (Simonovic et al., 2021).

Testimonial:

It often takes students time and effort to understand the current interpretations of the legal requirements and to try and fulfil all of them, and they need individual help with this. International deposits can be problematic (if copyrighted or sensitive data is collected from outside the EU and deposited within the EU or when using (possibly sensitive) data from old research projects). If the data can be public and open and the licenses have been cleared, the student can deposit the data independently, and all is well.

Anonymous lecturer
(Questionnaire of Lecturers - Annex B)

To teach students **how to create a research data management plan** in language and linguistics and **deposit, share and archive their language resources**, see the following learning and teaching resources on Moodle:



Teaching and Learning Resources on Moodle

[Introduction to Language Data: Standards and Repositories](#)

UNIT 1: Introduction to the Language Resource Lifecycle and Management Presentation:

- 1.8. *Creating a Research Data Management Plan for Linguistic Research*. This presentation is based on Kung (2022) and other RDM practices shared through CLARIN knowledge infrastructure.

Unit 5: Legal and Ethical Issues in Language Data Collection, Sharing and Archiving

Presentations: 5.3. *Copyright Exceptions for Text and Data Mining*

- 5.4. *Sharing and Archiving Language Resources*

Tutorials and Activities for Self-Study:

- 5.2. *Data Protection in Research Practice*
- 5.8. *Quiz: Sharing and Archiving Language Resources*

Unit 6: Student Project

- *Title: Designing, compiling and archiving a corpus of bank bulletins*

Moodle learning resource 15: Linguistic Research Data Management



For other examples of how RDM practices have been integrated in linguistics research, we recommend the chapters and case studies from [The Open Handbook of Linguistic Data Management](#) (ed. Berez-Kroeker et al. 2022). The handbook shares best practices for managing, archiving, sharing, and citing linguistic research data. Although the case studies may be too advanced for the BA level, they could be discussed at the MA and PhD levels. The [online companion course](#) has interactive quizzes and learning activities, which you can use and adapt freely for non-commercial educational purposes.

5.1. Examples of learning outcomes for FAIR data skills

To target basic RDM skills at BA, MA and PhD levels, we recommend following the general guidelines in *FAIRsFAIR Teaching and Training Handbook for Higher Education Institutions* (2021). This handbook provides a map of FAIR skills and competences for all levels of study, along with practical implementation guidelines, lesson plans, and learning outcomes. Please note that the learning outcomes are domain-independent and must be adapted to a specific discipline.

According to the authors, students at the **Bachelor level** (the level targeted in UPSKILLS) should acquire the following basic competences when working with research data:

- Can paraphrase the concepts of Open Research, Open Access, and Open Data and explain their benefits
- Can paraphrase the FAIR principles in data management and recognise the relationship between FAIR, Open and RDM
- Can define what Research Data Management is and explain its benefits
- Can develop a basic data management plan for their work, and identify different types of data documentation
- Can identify, describe and use different types of metadata, formats and standards
- Can identify, describe and use metadata registries to find data (e.g. Virtual Language Observatory)
- Can explain what a trusted data repository is and how to find it; can compare different certifications for data repositories (e.g. CoreTrustSeal vs CLARIN certification)
- Can explain the importance of data discovery and reuse
- Can recognise, explain and use persistent identifiers to access data and other resources
- Can identify and use different levels of data security, protection and backup
- Can explain basic rules and regulations for handling personal and sensitive data and know how to comply with them
- Can summarise and explain ethical principles and responsible data use (e.g. CARE principles) and identify potential legal issues around data use, management and sharing
- Can explain the research data lifecycle and compare different models
- Can explain the role of ontologies and vocabularies in data discovery, identify and use domain-specific ones

To include RDM learning outcomes in a domain-specific course or programme, pick the ones from the list that match the overall course goals and objectives, and adapt them to the type of research project and language data the students will work with during the course. As examples, see how RDM practices have been integrated into the learning outcomes of the UPSKILLS learning content and guides:

- Unit 1, 2, 4 and 6 of the learning block, [Introduction to Language Data: Standards and Repositories](#) on Moodle
- Part A, 6 of the [Research-Based Teaching: Guidelines and Best Practices](#) Guidelines (Simonovic et al, 2023)
- Section 2 of the *Guidelines for the Formulation of Student Projects* (Simonovic et al., 2021)

5.2. Example of student project

For exemplification, we include the outline of the CLARIN student project designed in collaboration with the University of Bologna for the UPSKILLS learning content block: [Introduction to Language Data: Standards and Repositories](#). The full project description, instructions for students, templates and guidelines are available on **Moodle, Unit 6. Student Project**.

***Title:** Designing, Compiling and Archiving a Corpus of Bank Bulletins*

Research Question: How do linguistic and communicative strategies differ between original and translated economic bulletins within the banking and financial sectors across various EU member states, considering institutional factors such as language diversity, cultural influences, and publication dates?

Imagine you have been commissioned to collect a corpus of quarterly economic bulletins in the framework of an EU-wide effort to analyse and compare institutional communication strategies in the economic, financial, and banking domains. You are part of an international team of linguists who will build corpora from as many national central banks as possible, and the corresponding texts by the European Central Bank.

You can either construct a corpus of ECB bulletins (in English or another of the EU official languages) or national bank bulletins for a country of your choice (in English or the country's official language).

The following metadata is considered relevant by your commissioners:

- Whether the institution authoring the texts is based in a country in which English is one of the official languages
- Whether the texts are original or translated
- Date when the text was produced/last modified

You are encouraged to provide further metadata that you consider relevant.

The corpus should contain at least 50 texts in plain text format. Linguistic annotation is an optional task for extra points.

The commissioner requires that the corpus is deposited in a plain-text format in a domain-specific certified data repository and that it can be shared with the other linguists via a Creative Commons Licence.

Workload: 1-2 ECTS (depending on the tasks that the teacher will choose to include)

Level: BA

Learning Outcomes:

By the end of this project, the students will be able to:

- Apply the FAIR data principles to corpora
- Design and compile a corpus with specialised texts from the web and research data repositories (e.g. VLO) using corpus compilation tools
- Perform a basic annotation and analysis using your preferred tool or find a matching tool on Switchboard
- Find a linguistic data repository in CLARIN to archive and publish the corpus with a CC-BY licence
- Understand the techniques involved in writing in academic popularisation genres

Prerequisites:

To be able to work on this project independently, students first follow the following learning blocks on Moodle:

- Processing Texts and Corpora
- Introduction to Language Data: Standards and Repositories

Recommended Background Knowledge:

- Basic ICT skills
- Academic writing skills
- UPSKILLS Introduction to scientific research

Reporting Format: Blog post (600-800 words) + classroom presentation (5-10 slides)

Assessment:

The project could be evaluated in the following way:

- Corpus design, construction and documentation 60%
- Corpus archiving and sharing 25%
- Blog post 15% (peer review)
- Presentation 10% (peer review)

Templates and Guidelines:

- Blog post template
- FAIR checklist for corpora
- Guidelines for depositing corpora in a CLARIN research data repository

Designers: Iulianna van der Lek (CLARIN ERIC) and Silvia Bernardini (UNIBO)

Reviewers: Novella (UNIBO), Darja Fišer (CLARIN ERIC), Marko Simonovic (UGraz) and Francesca Frontini (CLARIN ERIC)

Moodle learning resource 16: Example of Student Project



For more examples, see the [16 research-based courses](#) we piloted in UPSKILLS, which include specific learning outcomes related to using research infrastructures, repositories, corpora and tools.

6. How to Get Involved in the Community

CLARIN is one of the largest European Research Infrastructures for Language Resources and Technologies, with members in 23 European countries and three observer countries in South Africa, Switzerland and the United Kingdom. Each member country has a national consortium, which may consist of various types of data and knowledge centres. Within CLARIN, a strong community of researchers, educators, students, project managers, trainers, developers, and language enthusiasts ensures a continuous transfer of knowledge and expertise between different countries, universities, libraries, and organisations. Teachers and educators unfamiliar with CLARIN can engage with the infrastructure in the following ways:

1. Join the CLARIN Trainers' Mailing List

We invite teachers, trainers, and curriculum designers to [subscribe](#) to our trainers' mailing list to share and discuss initiatives in education and training, invite guest lecturers at your university, find collaboration opportunities, and exchange knowledge on training topics and language resources and tools.

2. Collaborate with the Trainers' Network

If you are passionate about teaching language technologies and like to travel, join the CLARIN [Trainer Network programme](#). The network consists of experts conducting training events and organising workshops at prominent summer schools, conferences, and COST

Actions in disciplines such as linguistics, digital humanities, language technologies and social sciences. The call is open to instructors with a full-time or part-time academic position and teaching experience for at least one year.

3. Adapt, Create and Share Training Materials

Teachers and trainers who use this guide and adapt any of the UPSKILLS learning content for the classroom are invited to share their experience via this [CLARIN training and education call](#), part of the CLARIN Annual Conference. Each year, a session showcases new educational initiatives and training materials on different language research topics. All training and learning materials are published under a Creative Common Licence on the [CLARIN Learning Hub](#) and included in the [SSH Training Discovery Toolkit](#), which has been set up to improve the discovery of training resources developed in the Social Sciences and Humanities. The toolkit links to various training resources on topics such as Open Science, Research Data Management, and didactics, but also specific topics relevant to multiple disciplines, like text encoding and spatial data.

4. Contribute to the Digital Humanities Course Registry

Teachers and trainers in Digital Humanities are encouraged to register the metadata of their courses and training programmes in the [Digital Humanities Course Registry](#). The DH Course Registry is a joint effort of the CLARIN ERIC and DARIAH-EU research infrastructures. The registry does not contain training materials but it is an online inventory of digital humanities modules, courses and programmes in Europe and beyond. Its goal is to help students, researchers, lecturers, and institutions discover, promote, and connect to DH teaching and training activities. By contributing courses to the DH course registry, educators increase the visibility of their programmes, attract international students and open up new opportunities for collaboration in research and teaching.

Students can search for courses using the advanced filters per country, institution, city, language, discipline, education, and type of course (online, onsite, recurring). For example, BA students from a language-related programme can use the registry to search for MA and PhD programmes, workshops or summer schools in Linguistics and Language Studies or related topics, e.g. Human Language Technologies, Linguistics and Computer Sciences.

5. Engage in Outreach Programs and Events

Participate in the [CLARIN Cafés](#) where lecturers and students meet informally to exchange and discuss specific topics related to language research. Furthermore, the [Impact Stories](#) can be used in the classroom to show how language resources and technologies are used to tackle societal issues. Another way to engage with the community is to participate in user involvement activities in your country and follow the activities of the K-Centres, which are showcased in [Tour de CLARIN](#).

6. Encourage Student Involvement

Involve students in community engagement activities, such as participating in online Cafés, hackathons and summer schools, and presenting their projects and ideas at the CLARIN Annual Conference. The conference includes a [session](#) for PhD students who use the CLARIN infrastructure in their research and would like to receive feedback on their work. Another way to encourage students to engage in research-based activities outside the classroom is by interviewing researchers from the CLARIN network about their research and reporting the findings to their fellow students through a class presentation, blog post or YouTube video. Outreach activities outside the university environment will foster collaboration between students and researchers.

7. Conclusions and how to contribute

This guide aims to assist UPSKILLS consortium partners, as well as teachers and trainers who would like to become more familiar with the CLARIN infrastructure, in effectively using the core services for teaching language and linguistic research, as well as the essential aspects of linguistic research data management. The guide includes references to ready-made teaching and learning content available on Moodle, *Introduction to Language Data: Standards and Repositories*, which teachers can customise to match the overall goals of their current courses. Additionally, the guide provides information on other learning resources and opportunities for continuing professional development to increase awareness and help teachers build up their corpus literacy and technical skills.

The guide is structured into several sections. [Section 2](#) provided an overview of the larger European Research Infrastructure landscape, highlighting the research infrastructures used to collect, manage and disseminate language resources and technologies. [Section 3](#) introduced CLARIN and how teachers and students can benefit from their national networks and knowledge centres. [Section 4](#) was the most detailed part of the guide, and started with general [recommendations \(in 4.1\)](#) on introducing data-driven learning using corpora and corpus-based technologies in the classroom. This section also included information about a research tracking tool developed by our Zurich consortium partners to help teachers and students track the progress made during a research project and give/receive feedback at intermediate steps in the project. [Section 4.2](#) presented an overview of the core CLARIN services and examples of how they benefit teachers. [Sections 4.3](#) and [4.4](#) presented each service in more detail and highlighted those language resources and tools suitable for educational settings. Finally, [Section 5](#) briefly discussed the benefits of introducing students to research data management early in their academic journey. It also provided examples of learning outcomes targeting basic skills in research data management and the use of repositories. Furthermore, this section included an example of a student project we designed

in collaboration with consortium partners from the University of Bologna. In the project, students use research data repositories and integrated concordancers to search and collect language data and find a FAIR repository to deposit their corpus at the end of the project. [The last section](#) gave an overview of different channels through which teachers and students can engage with the community around the infrastructure.

Overall, this guide and accompanying learning content on Moodle demonstrate how the CLARIN central services for language data discovery, processing, analysis, sharing, and archiving can be used in the classroom and provides examples of open corpora and NLP tools that can be used not only for advanced research but also as pedagogical tools. The guide may also benefit other stakeholders, such as librarians and language professionals who want to use the infrastructure for teaching and learning. Teachers and students are encouraged to engage with the research infrastructure and the CLARIN community outside the classroom by participating in collaborative research projects, hackathons, workshops, and the CLARIN Annual Conference. Finally, teachers and trainers using corpora and NLP tools from the CLARIN infrastructure (and not only) are invited to contribute their examples of teaching and learning activities to via training@clarin.eu.

Bibliography

- Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology*, 21, 195-216.
- Agerri, R. et al. (2023). State-of-the-art in language technology and language-centric artificial intelligence. In: Rehm, G., Way, A. (eds) *European Language Equality. Cognitive Technologies*. Cham: Springer. https://doi-org.proxy.library.uu.nl/10.1007/978-3-031-28819-7_2
- Andreassen, H. (2019, March 04). The acquisition of definiteness: Analysis of child language data. [OER Commons](#). Accessed [15 May 2023]
- Armaselu, F. (2021). The digital humanities classroom as a “node”. From toolbox to mindset? *DH Benelux Journal* 3: 103-115
- Arnold, D., Campbell, B., Eckart, T., Fisseni, B., Trippel, T. & Zinn, C.. (2020). The CMDI Explorer. <https://doi.org/10.3384/ecp1802>
- Assimakopoulos, S., Vella, M., van der Plas, L., Milicevic Petrovic, M., Samardžić, T., van der Lek, I, Bernardini, S., Ferraresi, A., & Pallottino, M. (2021). *Graduate skills and employability: Focus interviews with selected job market stakeholders*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.5030913>
- Baker, M. (1995) Corpus linguistics and translation studies: Implications and application. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins.
- Baldrige, J. & Erk, K. (2008). Teaching computational linguistics to a large, diverse student body: Courses, tools, and interdepartmental interaction. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pp. 1–9, Columbus, Ohio. Association for Computational Linguistics.
- Baroni, M., Bernardini, S., Ferraresi, A. et al. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources & Evaluation* 43, 209–226. <https://doi.org/10.1007/s10579-009-9081-4>
- Bassett, S., Wessels, L., Krauwer, S., Maegaard, B., Hollander, H. et al. (2019) *Connecting the Humanities through Research Infrastructures*. 4th Digital Humanities in the Nordic Countries (DHN 2019), Mar 2019, Copenhagen, Denmark.
- Bennett, G. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Michigan: University of Michigan Press ELT. <https://doi.org/10.3998/mpub.371534>
- Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest

- Group. (2018). The Austin Principles of Data Citation in Linguistics (Version 1.0). <https://site.uit.no/linguisticsdatacitation/austinprinciples/> Accessed [22 March 2022].
- Bernardini, S. (2002). Exploring new directions for discovery learning. *Language and Computers* 42(1), 165–182.
- Bernardini, S. (2006). Corpora for translator education and translation practice: achievements and challenges. In *Proceedings of LREC 2006 (5th Language Resources and Evaluation Conference)*. Paris: ELRA, pp. 17 - 22.
- Bezjak, S., Conzett, P., Fernandes, P.L., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F., Ross-Hellauer, T., Schneider, R., Tennant, J., Verbakel, E. & Clyburne, S. (2019). *The Open Science Training Handbook*. Zenodo. <https://doi.org/10.5281/zenodo.2587951>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly. Retrieved July 28, 2023, from <https://www.nltk.org/book/>.
- Blumel, B. (2014). Learning in parallel: Using parallel corpora to enhance written language acquisition at the beginning level. *Dimension* 1: 31–48.
- Borek, L., Dombrowski, Q., Perkins, J., Schöch, C. (2016), TaDiRAH: a Case Study in Pragmatic Classification, *Digital Humanities Quarterly* 10 (1).
- Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, 35(1): 81-106.
- Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, 50(4), 483–506. <https://doi.org/10.1017/S0261444817000167>
- Boulton, A. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (eds.), *Language, Education and Technology: Encyclopedia of Language and Education*. New York: Springer. https://doi.org/10.1007/978-3-319-02328-1_15-1
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393. <https://doi.org/10.1111/lang.12224>
- Broeder, D., Buddenbohm, S. & Elbers, W. (2021). *Use Cases for the Virtual Collection Registry in SSHOC*. Zenodo. <https://doi.org/10.5281/zenodo.5535818>
- Bunout, E., Cooper, S., Düring, M., Fišer, D., Hodgson, H., Jones, C., Kogou, S., Lenardič, J., Liang, H., Middendorf, J., Pickup, C & Thomson, C. (2019). *Digital Humanities Research Questions and Methods*. Version 1.0.0. Edited by M. Annisius, N. Baldszuhn, V. Garnett & U. Wuttke. PARTHENOS Training Suite. [Training module]. <https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/>

- Callies, M. (2019). Integrating corpus literacy into language teacher education. In S. Götz & J. Mukherjee (Eds.), *Learner corpora and language teaching*, pp. 245–263. Amsterdam/Philadelphia: John Benjamins.
- Čermák, P. (2019). InterCorp. A parallel corpus of 40 languages. In I. Doval & M.T. Sánchez Nieto (Eds.), *Parallel Corpora: Creation and Applications*, pp. 93-102. Amsterdam/Philadelphia: John Benjamins.
- CESSDA Training Team (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Retrieved from <https://www.cessda.eu/DMGuide>
- CESSDA Training Team. (2020). *CESSDA Data Management Expert Guide*. CESSDA ERIC. <https://doi.org/10.5281/zenodo.3820473>
- Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching* 52(4): 460–475. <https://doi.org/10.1017/S0261444819000089>
- Chen, M., Flowerdew, J., & Anthony, L. (2019). Introducing in-service English language teachers to data-driven learning for academic writing. *System* 87: 102-148. <https://doi.org/10.1016/j.system.2019.102148>
- Cheng, W., & Lam, P. W. Y. (2022). What can a corpus tell us about language teaching? In A. O’Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (2 edn), pp. 299-312. London: Routledge.
- CLARIN-NL (2022). Referencing and citing data. <https://dev.clarin.nl/node/4238>. [Accessed: 22 March 2022]
- Coto-Solano, R., Nicholas, S. A., Hoback, B., & Tiburcio Cano, G. (2022). Managing data workflows for untrained forced alignment: Examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management*.
- Cox, C. (2022). Managing data in a language documentation corpus. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management*.
- Crosthwaite, P. (2020). Taking DDL online: Designing, implementing and evaluating a SPOC on data-driven learning for tertiary L2 writing. *Australian Review of Applied Linguistics* 43(2): 169–195. <https://doi.org/10.1075/aral.00031.cro>
- David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C. et al. (2020) ‘FAIRness Literacy: The Achilles’ Heel of Applying FAIR Principles’. *Data Science Journal* 19: 32. <https://doi.org/10/gkbnxs>.

- de Jong, F., Maegaard, B., Fišer, D., van Uytvanck, D., & Witt, A. (2020). Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3406-3413). Marseille, France: European Language Resources Association.
- de Smedt, K., Jong, F.D., Maegaard, B., Fišer, D., & Uytvanck, D.V. (2018). Towards an Open Science Infrastructure for the Digital Humanities: The Case of CLARIN. *DHN*.
- Demchenko, Y., Stoy, L., Engelhardt, C. & Gaillard, V.. (2021) *D7.3 FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework)*. Zenodo. <https://doi.org/10.5281/zenodo.5361917>.
- Deshors, S.C. (2021). Corpora in applied linguistics. In H. Mohebbi & C. Coombe (Eds) *Research Questions in Language Education and Applied Linguistics*, pp. 805-809. Cham: Springer.
- Ebrahimi, A., & Faghih, E. (2016). Integrating corpus linguistics into online language teacher education programs. *ReCALL* 29(1): 120. <https://doi.org/10.1017/S0958344016000070>
- Eckart de Castilho, R., Biemann, C., Gurevych, I. & Yimam, S.M. (2014): WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.
- Elbers, W., Stemle, E.W., Moreira, A., König, A., Cattani, L., & Palma, M. (2019). *The CLARIN ERIC deployment infrastructure and its applicability to reproducible research*. Zenodo. <https://doi.org/10.5281/zenodo.3556747>
- Engelhardt, C., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., ... Zhou, B.. (2022). *D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions (V1.2.1)* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.6674301>
- Erjavec, T. & Pančur, A.. (2019, September 19). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. Zenodo. <https://doi.org/10.5281/zenodo.3446164>
- ESFRI. (2021). *The ESFRI Roadmap 2021: Strategy Report on Research Infrastructures*. <https://roadmap2021.esfri.eu/media/1295/esfri-roadmap-2021.pdf> Accessed [23 April 2023].
- Eskevich, M., & Frontini, F.. (2021, July 8). *SSHOC'ing Drama in the Cloud*. Zenodo. <https://doi.org/10.5281/zenodo.5082522>
- Eskevich, M., Jong, F.D., König, A., Fišer, D., Uytvanck, D.V., Aalto, T., Borin, L., Gerassimenko, O., Hajic, J., Heuvel, H.V., Kahusk, N., Liin, K., Matthiesen, M., Piperidis, S., & Vider, K. (2020). CLARIN: Distributed Language Resources and Technology in a European Infrastructure. *IWLTP*.

- European Commission, Directorate-General for Research and Innovation (2016). *European charter of access for research infrastructures: principles and guidelines for access and related services*, Publications Office. <https://data.europa.eu/doi/10.2777/524573>
- European Commission, Directorate-General for Research and Innovation (2017), O'Carroll, C., Hyllseth, B., Berg, R. et al., Providing researchers with the skills and competencies they need to practise Open Science, Publications Office. <https://data.europa.eu/doi/10.2777/121253>
- European Commission, Directorate-General for Research and Innovation (2021) *Digital skills for FAIR and Open Science: report from the EOSC Executive Board Skills and Training Working Group*, Edited by Manola, N., Lazzeri, E., Barker, M., Kuchma, I., Gaillard, V. & Stoy, L., Publications Office. <https://data.europa.eu/doi/10.2777/59065>
- European Commission. Directorate General for Research and Innovation & EOSC Executive Board. (2021) *Digital Skills for FAIR and Open Science: Report from the EOSC Executive Board Skills and Training Working Group.* LU: Publications Office. <https://doi.org/10.2777/59065>.
- European Union General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1. Available at: <https://gdpr-info.eu/> Accessed [10 June 2021]
- Fang, L., Ma, Q., & Yan, J. (2021). The effectiveness of corpus-based training on collocation use in L2 writing for Chinese senior secondary school students. *Journal of China Computer-Assisted Language Learning* 1(1): 80–109.
- Ferraresi, A., Aragrande, G., Barrón-Cedeño, A., Bernardini, S., & Miličević Petrović, M.. (2021). *Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements*. UPSKILLS Task Report. Zenodo. <https://doi.org/10.5281/zenodo.5030879>
- Fišer, D., Lenardic, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. *LREC*.
- FOSTER Consortium. (2018). *Managing & sharing research data*. Zenodo. <https://doi.org/10.5281/zenodo.2630562>
- FOSTER consortium. (2018). What is Open Science? Zenodo. <https://doi.org/10.5281/zenodo.2629946>
- FOSTER Plus Consortium. (2019). FOSTER Plus Taxonomies [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.2860669>
- FOSTER. 'Use Open Data in Teaching'. Online Course. <https://www.fosteropenscience.eu/learning/use-open-data-in-teaching/#/id/5b06ca3ba8216e322da360b6f> Accessed [20 July 2021].

- Gabber, S., Yarbrough, D., Berez-Kroeker A.L., McDonnell, B., Koller, E., Collister, L.B. Conzett, P. & De Smedt, K. (2022). "Lesson 11." *Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management*.
<https://sites.google.com/hawaii.edu/linguisticdatamanagement/course-lessons/11-guidance-for-citing-linguistic-data>
- Gabber, S., Yarbrough, D., Berez-Kroeker, A.L, McDonnell, B., Koller, E. & Collister, L.B. (2022). "Lesson 1." *Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management*.
<https://sites.google.com/hawaii.edu/linguisticdatamanagement/course-lessons/01-data-data-management-and-reproducible-research-in-linguistics-on-the>
- Garbuglia, F., Saenen, B., Gaillard, V. & Engelhardt, C. (2021). *Good Practices in FAIR Competence Education (1.2)*. Zenodo. <https://doi.org/10.5281/zenodo.6657165>
- Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359-370). London: Routledge.
- Gledić, J., Assimakopoulos, S., Buchberger, I., Budimirović, J., Đukanović, M., Kraš, T., Podboj, M., Soldatić, N. & Vella, M.. (2021). *UPSKILLS guidelines for Learning Content Creation*. Zenodo. <https://doi.org/10.5281/zenodo.8302296>
- Gledić, J., Budimirović, J., Đukanović, M., Samardžić, T., Jukić, S., Ferraresi, A., Aragrande, G., van der Plas, L., van der Lek, I., & Soldatić, N. (2021). *Survey of business sectors hiring linguists and language professionals*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.5030891>
- Gledić, J., Đukanović, M., Miličević Petrović, M., van der Lek, I. & Assimakopoulos, S. (2021). *Survey of curricula: Linguistics and language-related degrees in Europe*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.5030861>
- Gledić, J., Đukanović, M., Miličević Petrović, M., van der Lek, I., & Assimakopoulos, S. (2021). *Survey of curricula: Linguistics and language-related degrees in Europe*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.5030861>
- Godfrey, J.J. & Zampolli, A.. (1997). Language resources. In A. Zampolli & G. Battista Varile (Eds) *Survey of the State of the Art in Human Language Technology. Linguistica Computazionale, XII-XIII*, pp. pages 381–384. Pisa: Giardini Editori e Stampatori (also Cambridge University Press)
- Good, J. (2022). The scope of linguistic data. In A.L. Berez-Kroeker, B. McDonnell, E. Koller & L.B. Collister (Eds.) *The Open Handbook of Linguistic Data Management*, pp. 27-48. Cambridge, MA: MIT Press Open.

- Gorgaini, E. (2021). Language Resource Switchboard - SSHOC Service Catalogue's Factsheets Series (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5495696>
- Gut, U. (2020). Spoken Corpora. In M.Paquot, M. & S.T. Gries (Eds) *A Practical Handbook of Corpus Linguistics*. Cham: Springer.
- Heather, J. & Helt, M. (2012). Evaluating corpus literacy training for preservice language teachers: Six case studies. *Journal of Technology and Teacher Education* 20(4): 417.
- van den Heuvel, H. & Draxler, C.. (2020, March 2). *SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data*. Zenodo. <https://doi.org/10.5281/zenodo.3694223>
- Hensel, N. (ed.) (2012), *Characteristics of Excellence in Undergraduate Research*. Washington: The Council on Undergraduate Research (CUR). http://www.cur.org/assets/1/23/COEUR_final.pdf.
- Hinrichs, M. (2022). CLARIN Cafe on Text+: Using a distributed technical infrastructure. <https://youtu.be/ceoeGXRUBpA>. Accessed [24 March 2022]
- Hoorn, E., van den Heuvel, H., Bessel, N., Klessa, K., Lee, A., Salaasti, S., Trilsbeek, P. (2021): Privacy by Design in Research: A DPIA role play with video. Teaching material published at the CLARIN Annual Conference 2021. <https://www.clarin.eu/content/privacy-design-research>
- Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M. & van Gelder, C.W.G. (2021). 'Professionalising Data Stewardship in the Netherlands. Competences, Training and Education. Dutch Roadmap towards National Implementation of FAIR Data Stewardship'. <https://doi.org/10.5281/Zenodo.4320504>
- Johns, T. (1991). "Chapter 2: Should you be persuaded: Two examples of data-driven learning". Classroom Concordancing. Birmingham: ELR.
- Jong, F.D., Maegaard, B., Fišer, D., Uytvanck, D.V., & Witt, A. (2020). Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. *LREC*.
- König, A., Frey, J., & Stemle, E.W. (2021). Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. *Inf.*, 12, 199.
- Krasselt, J., Dreesen, P., Fluor, M., Mahlow, C., Rothenhäusler, K., & Runte, M. (2020). Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 4138-4144.
- Kuebler, S., & Zinsmeister, H. (2014). *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.
- Kung, S.S. (2022). Developing a data management plan. In A.L. Berez-Kroeker, B. McDonnell, E. Koller & L.B. Collister (Eds.) *The Open Handbook of Linguistic Data Management*, pp. 101-115. Cambridge: MIT Press Open.

- Latif, A., Limani, F., & Tochtermann, K. (2021). On the Complexities of Federating Research Data Infrastructures. *Data Intelligence* 3(1): 79–87. https://doi.org/10.1162/dint_a_00080
- Lefter M.-A. (2020). Parallel corpora. In M. Paquot & S. Th. Gries (eds). *A Practical Handbook of Corpus Linguistics*, pp. 257-282. Cham: Springer.
- Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology* 21(3): 217–241.
- Lennes, M. (2021). *Introduction to Speech Analysis (Puheen analyysin perusteet)* [Online course material]. Available at: <http://urn.fi/urn:nbn:fi:lb-2021063021>.
- Lennes, M. (2021, September 14). Puheen analyysin perusteet – Introduction to Speech Analysis. Zenodo. <https://doi.org/10.5281/zenodo.5506969>
- Lessard-Clouston, M. & Chang, T.. (2014). Corpora and English language teaching: Pedagogy and practical applications for data-driven learning. *TESL Reporter* 47: 1-20.
- Levshina, N. (2016). Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50 (2): 507-542.
- Lin, M.H. (2019). Becoming a DDL teacher in English grammar classes: A pilot study. *The Journal of Language Learning and Teaching*, 9(1), 70–82.
- Ljubešić, N.; Erjavec, T. and Fišer, D. (2017). Twitter corpus Janes-Tweet 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1142>.
- Lüngen, H., & Beißwenger, M. (2021). Data formats and standards for interoperable CMC corpora. Presentation at the CLARIN Workshop on Data Management for CMC Corpora, Online Event, 27 October 2021
https://www.clarin.eu/sites/default/files/luengen_beisswenger-standards_and_data_formats.pdf.
Accessed [24 March 2022]
- Lušický, V., & Wissik, T. (2016). Evaluation of CLARIN services, user requirements, usability, VLO, and translation studies. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, France*, pp. 63-75. Linköpings universitet: Linköping University Electronic Press.
- Ma, Q., Lin, S., & Tang, J. (2021). The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning*, 1–30. DOI: <https://doi.org/10.1080/09588221.2021.1895225>
- Ma, Q., Yuan, R., Cheung L.M.E. & Yang, J (2022). Teacher paths for developing corpus-based language pedagogy: a case study. *Computer Assisted Language Learning* 2: 1-32. <https://doi.org/10.1080/09588221.2022.2040537>

- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd edn). Mahwah, NJ: Lawrence Erlbaum Associates
- Mattern, E. (2022). "The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management". In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management*.
- Maynard, D. Bontcheva, K., Roberts, I., Song, X., Greenwood, M.A., Bakir, M. Petrak, J. & Jiang, Y. (2021). GATE Training Course, <https://gate.ac.uk/wiki/TrainingCourseFeb2021/>.
- McCarthy, M., & O'Keefe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). London: Routledge.
- Miličević Petrović, M., Bernardini, S., Ferraresi, A., Aragrande, G., & Barrón-Cedeño, A.. (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.5030929>
- Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art—and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpora and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt: Peter Lang.
- O'Keefe, A., & McCarthy, M.J. (Eds.). (2022). *The Routledge Handbook of Corpus Linguistics* (2nd ed.). Routledge. <https://doi.org/10.4324/978036707639>
- Odiijk, J. 2017. Introduction to the CLARIN Technical Infrastructure. In J. Odiijk & A. van Hessen (Eds.) *CLARIN in the Low Countries*, pp. 33–44. London: Ubiquity Press.
- Osipova, E. S. (2018). Corpus Linguistics Technology In Teaching English As A Foreign Language. In V. Chernyavskaya, & H. Kuße (Eds.), *Professional Culture of the Specialist of the Future, vol 51: European Proceedings of Social and Behavioural Sciences* (pp. 273-283). Future Academy. <https://doi.org/10.15405/epsbs.2018.12.02.30>
- Paquot, M. & Gries, S.T. (Eds). *A Practical Handbook of Corpus Linguistics*. Cham: Springer.
- PARTHENOS, Hollander, H. Morselli, F., Uiterwaal, F., Admiraal, F., Trippel, T. & Di Giorgio, S. (2018). PARTHENOS Guidelines to FAIRify Data Management and Make Data Reusable <https://doi.org/10/gkbnxt>
- Pérez-Paredes, P. (2022). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer-Assisted Language Learning* 35: 36-61
- Petrauskaite, R., Amilevičius, D., Dadurkevičius, V. Krilavičius, T., Raškiniš, G., Utkā, A. & Vaičėnonienė, J.. (forthcoming). CLARIN-LT: Home for Lithuanian Language Resources. Berlin: De Gruyter.

- Quinn D., Gniady, T. & Kloster, D. (2019). Introduction to Jupyter Notebooks. *Programming Historian* 8. <https://doi.org/10.46430/phen0087>
- Ratner, N. B., Brundage, S. B., & Fromm, D. (2022). A Clinician's Complete Guide to CLAN and PRAAT. University of Maryland, George Washington University, & Carnegie Mellon University.
- Rosén, V., De Smedt, K., Meurer, P., & Dyvik, H. (2012). [An open infrastructure for advanced treebanking](#). In J. Hajič, K. De Smedt, M. Tadić, & A. Branco (Eds.), *META-RESEARCH Workshop on Advanced Treebanking at LREC2012* (pp. 22–29). Istanbul, Turkey, S. T.
- Simonović, M, van der Lek, I., Fišer, D. & Arsenijević, B. (2021). *Guidelines for the students' projects and research reporting formats*. UPSKILLS task report. Zenodo. <https://doi.org/10.5281/zenodo.8297430>
- Simonović, M., B. Arsenijević, I. van der Lek, S. Assimakopoulos, L. ten Bosch, D. Fišer, T. Kraš, P. Marty, M. Miličević Petrović, S. Milosavljević, M. Tanti, L. van der Plas, M. Pallottino, G. Puskas & T. Samardžić. (2023a). Research-based teaching: Guidelines and best practices. UPSKILLS task report. Zenodo. DOI: <https://doi.org/10.5281/zenodo.8176220>
- Sripicharn, P. (2010). How can we prepare learners for using language corpora? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*, pp. 371-384. London: Routledge.
- Stark, E., Ueberwasser, S., Göhring, A. (2014-2020). *Corpus "What's up, Switzerland?"*. The University of Zurich. www.whatsup-switzerland.ch
- Stark, E.; Ueberwasser, S.; Ruef, B. (2009-2015). Swiss SMS Corpus. University of Zurich. www.sms4science.ch
- Stoy, L., Saenen, B., Davidson, J., Engelhardt, C., & Gaillard, V.. (2020). *D7.1 FAIR in European Higher Education (1.0)*. Zenodo. <https://doi.org/10.5281/zenodo.5361815>
- Berez-Kroeker, A.L., McDonnell, B., Koller, E & Collister, L.B. (Eds.), *The Open Handbook of Linguistic Data Management*. Cambridge, MA: MIT Press Open. <https://doi.org/10.7551/mitpress/12200.001.0001>
- Thomas, J. (2017). *Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration* (2nd edn updated). Versatile.
- Tribble, C. (2010). What are concordancers and how are they used? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*, pp.167-183. London: Routledge.
- Ueberwasser, S. & Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. <https://bop.unibe.ch/linguistik-online/article/view/3849/5834> Accessed [7 August 2023]

- Vaičėnienė, J., Kovalevskaitė, J. & Boizou, L. (2020). ORVELIT v3, CLARIN-LT digital library in the Republic of Lithuania, <http://hdl.handle.net/20.500.11821/40>
- van der Lek, I. & Woldrich, A. (2022, May 30). Put Yourself on the Map! The DH Course Registry Story & its Actors. DARIAH Annual Event 2022, Athens, Greece. Zenodo. <https://doi.org/10.5281/zenodo.6602988>
- van Dijk, E.E., Tartwijk, J.V., Schaaf, M.F., & Kluijtmans, M. (2020). What makes an expert university teacher? A systematic review and synthesis of frameworks for teacher expertise in higher education. *Educational Research Review*, 31, 100365.
- van Gelder, C. W. G., Psomopoulos, F., & Melo, A. M. P. (2021, June 8). Digital Skills for FAIR and open science - Report of the EOSC Skills & Training Working Group. Zenodo. <https://doi.org/10.5281/zenodo.4911506>
- Visser-Wijnveen, G.J., van der Rijst, R.M. & van Driel, J.H. (2016). A questionnaire to capture students' perceptions of research integration in their courses. *Higher Education* 71: 473–488. <https://doi.org/10.1007/s10734-015-9918-2>
- Vyatkina, N. (Ed.). (2020a). *Incorporating corpora: Using corpora to teach German to English-speaking learners* [Online instructional materials]. University of Kansas Open Language Resource Center. <https://corpora.ku.edu>
- Vyatkina, N. (2020b) Corpora as open educational resources for language teaching. *Foreign Language Annals* 53: 359–370.
- Weisser, M. (2015). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis* (1st edn). Oxford: Wiley-Blackwell. <https://doi.org/10.1002/9781119180180>
- Whyte, A., Leenarts, E., De Vries, J., Huigen, F., Kuehn, E., Sipos, G., Kalaitzi, V., Dijk, E., Jones, S. & Ashley, K.. (2019). *D7.5: Strategy for Sustainable Development of Skills and Capabilities (1.1)*. Zenodo. <https://doi.org/10.5281/zenodo.5095052>
- Wiljes, C., & Cimiano, P. (2019). Teaching Research Data Management for Students. *Data Science Journal* 18(1): 38. <http://doi.org/10.5334/dsj-2019-038>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018–160018. <https://doi.org/10.1038/sdata.2016.18>
- Wissik, T. (2019). *The Role of Research Infrastructures in the Area of DH Education, Training and Competence Building*. In S. Allegrezza (ed.) *AIUCD 2019 - Book of Abstracts, Udine 2019, Italy*, pp. 233-235. Associazione per l'Informatica Umanistica e la Cultura Digitale.

- Wissik, T., Edmond, J., Fischer, F., de Jong, F., Stefania Scagliola, S., Scharnhorst, A., Schmeer, H., Scholger, W. & Leon Wessels, L. (2020). Teaching Digital Humanities Around the World: An Infrastructural Approach to a Community-Driven DH Course Registry. *Library Tribune* 40: 1-27.
- Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. *Meta* 43(4): 616–630. <https://doi.org/10.7202/004638ar>

Annex A: Resources for teachers

The resources below are available for download from the project website: <https://upskillsproject.eu/deliverables/io2/research-infrastructures/>.

A1. Research Tracker Tool

The research tracker mentioned in Section 4.1.7 can be downloaded from the project website: <https://upskillsproject.eu/deliverables/io2/research-infrastructures/>.

A2. Glossary

A glossary with key concepts related to (linguistic) research data repositories and infrastructures has been created in [Moodle](#), which is also relevant for the readers of this guide. The glossary can be downloaded from the website. Please note that the glossary is not exhaustive. Students can be asked to add terms to the glossary in Moodle as they browse through the content.

A3. List of Open Corpora

We compiled a collection of **open corpora** in the languages of the students enrolled in the UPSKILLS Summer School in Petnica, Serbia, in July 2023.

Except for Armenian, Bengali, and Chinese corpora, all other corpora can be directly accessed and queried via [NoSketch Engine \(clarin.si\)](#). Depending on the corpus type (comparable vs. parallel), the following NoSketch Engine features can be used to analyse the corpora:

1. Concordance: searches for examples of use in context.
2. Wordlist: extracts a list of frequent words or phrases.
3. Keywords: extracts a list of terms.
4. Text type analysis: provides statistics of the whole corpus.
5. Parallel concordance: allows searches for translation equivalents in two languages.

The corpora can also be downloaded from the CLARIN.SI repository.

Language (s)	Name of corpora	Nr of words
English	<ul style="list-style-type: none"> ● ParlaMint-GB 3.0 (British parliament) ● EU DGT-UD: English ● ukWaC (British Web) 	<ul style="list-style-type: none"> ● 124,744,599 words ● 123,559,238 words ● 1,802,927,963 words
Catalan, Spanish	<ul style="list-style-type: none"> ● ParlaMint-ES-CT 3.0 (Catalan parliament) 	<ul style="list-style-type: none"> ● 15,831,179 words
Armenian	<ul style="list-style-type: none"> ● W2C - Web to Corpus - Corpora⁸³ 	
Bengali	<ul style="list-style-type: none"> ● W2C - Web to Corpus - Corpora⁸⁴ 	
Croatian	<ul style="list-style-type: none"> ● CLASSLA-web.hr (Croatian Web) ● CLASSLAWiki-hr (Croatian Wikipedia) ● ENGRI (Croatian news portals) ● ParlaMint-HR 3.0 (Croatian parliament) 	<ul style="list-style-type: none"> ● 2,302,589,013 words ● 51,719,524 words ● 591,247,552 words ● 85,925,479 words
Chinese	<ul style="list-style-type: none"> ● Sheffield Corpus of Chinese, http://hdl.handle.net/20.500.14106/2481⁸⁵ ● The Lancaster Corpus of Mandarin Chinese https://hdl.handle.net/20.500.14106/2474⁸⁶ ● Set of parallel texts in the domain of Law and Health, with 1 G per language. Languages: cs-pt, de-pt, en-pt, es-pt, fr-pt, it-pt, and pt-sk. https://hdl.handle.net/21.11129/0000-000D-F938-C 	
French	<ul style="list-style-type: none"> ● EU DGT-UD: French ● frWaC (French Web) ● LeMonde: francosko ● ParlaMint-FR 3.0 (French parliament) 	<ul style="list-style-type: none"> ● 112,107,093 words ● 1,328,628,428 words ● 618,201 words ● 48,841,818 words
German	<ul style="list-style-type: none"> ● deWaC (German Web) ● EU DGT-UD: German ● ParlaMint-AT 3.0 (Austrian parliament) 	<ul style="list-style-type: none"> ● 1,348,199,799 words ● 75,213,001 words ● 58,275,585 words

⁸³ Found via the CLARIN VLO (<https://hdl.handle.net/11858/00-097>); hosted on LINDAT; it can be queried via KonText. Also see the Deltacorporus 1.1. <http://hdl.handle.net/11234/1-1743>.

⁸⁴ Found via the CLARIN VLO (<https://hdl.handle.net/11858/00-097>); hosted on LINDAT; it can be queried via KonText. Also see the Deltacorporus 1.1. <http://hdl.handle.net/11234/1-1743>.

⁸⁵ Found via the VLO; hosted on OTA, it can be downloaded

⁸⁶ Found via the VLO, hosted on OTA, it can be downloaded

Italian	<ul style="list-style-type: none"> ● EU DGT-UD: Italian ● itWaC (Italian Web) ● ParlaMint-IT 3.0 (Italian parliament) 	<ul style="list-style-type: none"> ● 97,088,615 words ● 1,593,977,091 words ● 31,537,893 words
Macedonian	<ul style="list-style-type: none"> ● CLASSLAWiki-mk (Macedonian Wikipedia) 	<ul style="list-style-type: none"> ● 36,008,479 words
Polish	<ul style="list-style-type: none"> ● EU DGT-UD: Polish ● ParlaMint-PL 3.0 (Polish parliament) 	<ul style="list-style-type: none"> ● 79,219,433 words ● 35,327,340 words
Serbian	<ul style="list-style-type: none"> ● CLASSLA-web.sr (Serbian Web) ● CLASSLAWiki-sr (Serbian Wikipedia) ● ParlaMint-RS 3.0 (Serbian parliament) ● SETimes.SR ● ReLDI-sr (manually tagged Serbian tweets) 	<ul style="list-style-type: none"> ● 2,435,080,588 words ● 97,258,485 words ● 83,266,206 words ● 74,585 words ● 77,883 words
Serbian-Croatian	<ul style="list-style-type: none"> ● CLASSLAWiki-sh (Serbo-Croatian Wikipedia) 	<ul style="list-style-type: none"> ● 63,541,996 words
Slovenian	<ul style="list-style-type: none"> ● CLASSLA-web.sl (Slovenian Web) ● CLASSLAWiki-sl (Slovenian Wikipedia) ● EU DGT-UD: Slovenian 	<ul style="list-style-type: none"> ● 1,868,204,625 words ● 42,063,728 words ● 77,865,562 words

Annex B: Questionnaire of lecturers - Current practices of integrating research into teaching

This questionnaire targets researchers who do research and teach in linguistics and other language-related disciplines. It should take you about 10-15 minutes to complete. The questionnaire consists of 5 sections:

- A. Background
- B. Integration of research into teaching
- C. Integration of industry-based research into teaching
- D. Usage of language resources, tools and repositories
- E. End of the survey

The aim of this questionnaire is to collect some insights about the extent and practices in which researchers integrate their ongoing research activities into their teaching. The insights will be used in the UPSKILLS project (an ERASMUS+ project) to develop guidelines about "Best practices for integrating research into teaching."

Further information about the project can be found at <https://upskillsproject.eu/>. If you have questions about the survey, please contact Marko Simonovic from the University of Graz at rkiema@gmail.com.

The information provided by you in this questionnaire will be used for research purposes. It will not be used in a manner which would allow the identification of your individual responses. Anonymised research data will be archived in order to make it available to other researchers in UPSKILLS in line with the current data sharing practices.

A. General questions

A1. In which language-related disciplines do you do RESEARCH? (Check all boxes that apply)

A2. In which language-related disciplines do you TEACH? (Check all boxes that apply)

B. Integration of research into teaching

B1. At what level do you teach? (Check all boxes that apply)

B2. In which country/countries do you teach?

B3. What are the topics that your teaching focuses on? (Check everything that applies)

B4. To what extent is your teaching based on your ongoing research?

Please add any further comment(s) you may have regarding the previous question.

B5. Which methods do you use to integrate your research into your ongoing teaching? (Check everything that applies.)

B6. Have you ever taught a course mostly based on your ongoing research?

B7. To what extent can your students be included in your ongoing research?

Please add any further comment(s) you may have regarding the previous question.

B8. In your experience, to what extent do students appreciate the inclusion of ongoing research into teaching?

Please add any further comment(s) you may have regarding the previous question.

B9. How often does coursework produced by your students (homeworks, final papers, presentations...) include original and valuable scientific contributions that could be published/presented at international conferences etc.?

Please add any further comment(s) you may have regarding the previous question.

B10. To what extent do your courses increase the students' research skills? (Disregard courses entirely focused on research skills, e.g. research seminars etc.)

Please add any further comment(s) you may have regarding the previous question.

B11. How well does your field of expertise lend itself to integrating ongoing research into teaching?

Please add any further comment(s) you may have regarding the previous question.

C. Integration of industry-based research into teaching

C1. Do the courses that are being taught at your department currently include any input from industry actors (e.g. student participation in industry projects, joint supervision of theses, guest lectures by industry representatives, provision of work placements, etc.)?

C2. Do you think that exposure to industry-based research is/would be beneficial for your students?

Please add any further comment(s) or examples you may have about industry-based research into teaching.

D. Usage of language resources, tools and repositories

D1. Do you use any language resources and technologies (see above definition) in teaching the subject (s) you indicated in section B?

D2. If you have answered YES to the previous question, you may skip this question and go to D3. If you have answered NO to the previous question, please indicate the reason (s) why you do not use any language resources and tools in your teaching.

D3. What type of language resources and tools do you usually use in your teaching? Multiple answers are possible.

D4. Which platforms and/or repositories do you use to search for language resources and tools relevant to your teaching? Multiple answers are possible.

D5. What challenges do you and/or your students encounter when trying to use language resources and tools from the platforms and/or repositories you selected in the previous question? Please give some examples.

D6. What challenges have you and/or your students encountered when trying to use digital resources and tools in teaching and learning?

D7. Where do you usually store, archive and deposit the language resources you and/or students create during your course?

D8. Please indicate what problems you have encountered when you and/or students tried to deposit language resources and tools in the repository (ies) you selected in the previous question.

D9. If you would like to share any best practices for using language resources, tools and repositories into your teaching, please do so using the field below. You can also insert a link to case study or article you wrote on this topic.

D10. What kind of support would you need from your institution and digital research infrastructures to facilitate better integration of language resources, tools and repositories into your teaching and research?

E. End of the questionnaire and follow-up

Annex C: Accompanying Moodle learning block

This guide comes with an accompanying learning content block on Moodle, which can be accessed via the UPSKILLS project website.

[Introduction to language data: Standards and repositories](#)

6 ECTS (+ a student project amounting to 1 or 2 extra ECTS)

Block Designers

Iulianna van der Lek and Darja Fišer

(with contributions from Francesca Frontini, Walter Scholger, Esther Hoorn, Pawel Kamocki, Alexander König and Willem Elbers)

Description and scope

The aim of this learning block is to provide lecturers in BA/MA language-related programmes with a pool of learning resources and activities that they can use in the classroom to introduce students to research data repositories and their role in the linguistic research data lifecycle in the context of Open Science and FAIR data principles.

The philosophy behind the block is that students should take an active approach in learning. We provide a discussion of key concepts accompanied by examples and practical activities that will guide students towards their deeper understanding.

The block's units are conceived in a modular way that allows lecturers and learners to take (or adapt) them either in sequence or as self-standing contents, depending on their needs. The units comprise interactive presentations and learning activities, examples of assignments and hands-on tutorials, demonstrating how research data repositories can be used to discover, process, analyse, share, publish and archive language research data.

Block Outline

(the overall workload associated with the first 5 units of this block amounts to 6 ECTS)

1. Introduction to the Language Resource Lifecycle and Management
2. How Research Data Repositories Help Make Language Data FAIR
3. Finding and (Re)using Language Resources in the CLARIN repositories
4. Citing Language and Linguistic Data
5. Legal and Ethical Issues in Language Data Collection, Sharing and Archiving
6. Student project: Designing, compiling and archiving a corpus of bank bulletins
(1 or 2 ECTS)
7. Glossary

Learning Outcomes

Overall, the materials and activities present in this block will allow students to:

- explain the main concepts related to research data repositories and the role they play in the linguistic research data lifecycle in the context of Open Science and FAIR;
- find and use certified research data repositories to discover, share, publish, and archive language and linguistic resources and datasets;
- find and use integrated repository services and tools to process, annotate, and analyse different types of corpora according to standards and formats used by the community;
- identify potential legal and ethical issues when collecting, sharing and reusing language data and resources.

Target Audience

The primary target audience are lecturers who (want to) teach about standards and repositories related to linguistic data. Students can also use the materials autonomously but should be aware that this is not a typical self-study course.