



CLARIN AND ITS SWISS NODE CLARIN-CH: SUPPORTING RESEARCH BASED ON LANGUAGE RESOURCES

INFORMATION SESSION

CRISTINA GRISOT, CLARIN-CH SCIENTIFIC COORDINATOR

UNIVERSITY OF FRIBOURG, 31 MARCH 2023



The research infrastructure for language as social and cultural data

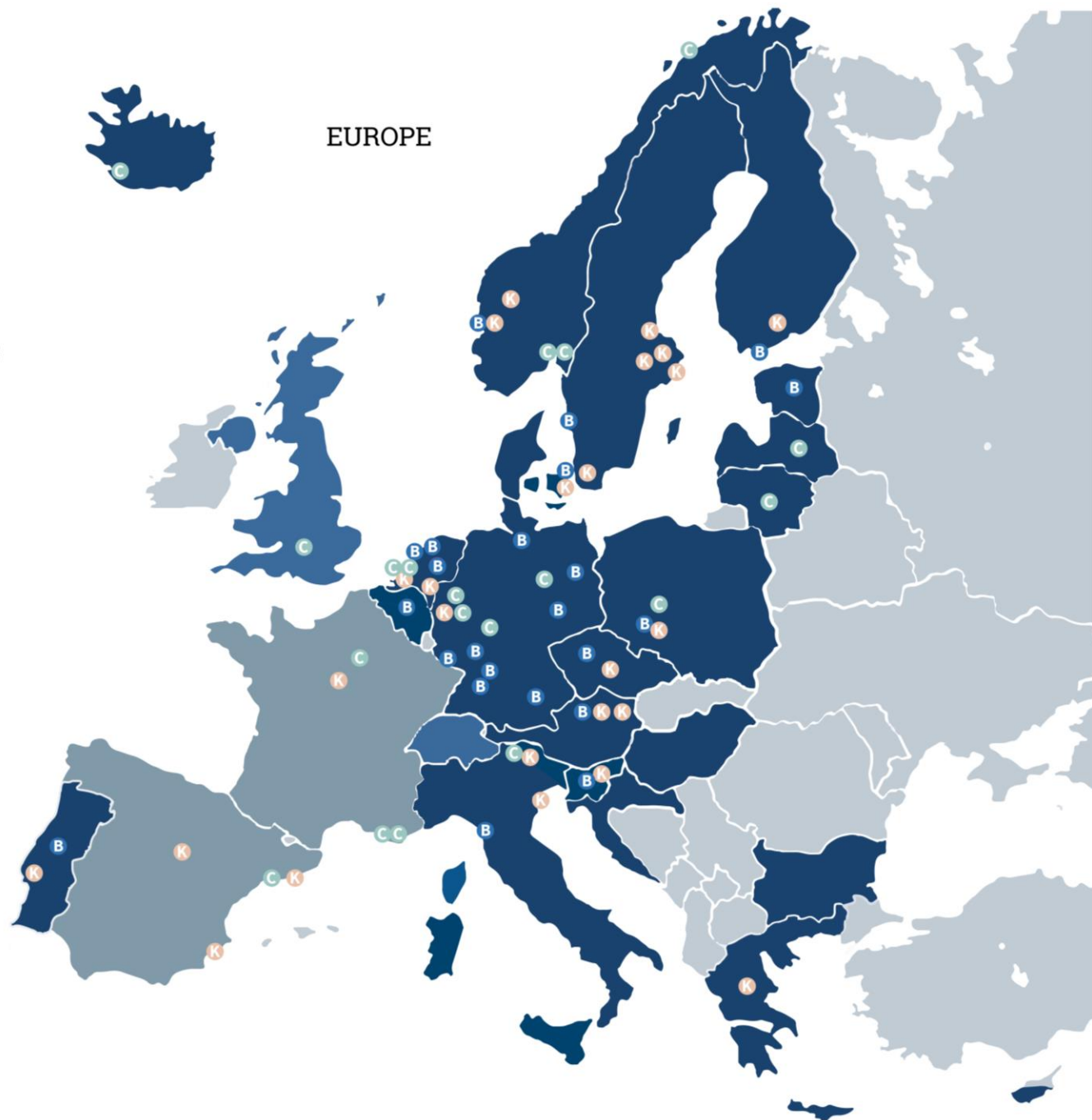
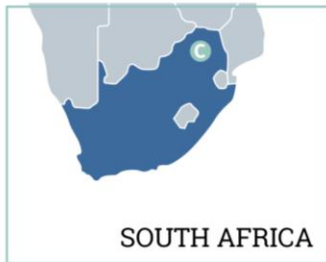
CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

- CLARIN is a pan-European research infrastructure that provides easy and sustainable access to digital language resources to support research in the SSH and beyond
- Fully operational since 2016
- 23 European countries are members of CLARIN
- Switzerland has joined CLARIN ERIC on 1 January 2023 as observer



CLARIN

- ERIC members
- Observers
- Countries with participating centres
- B** Centre Providing Data
- C** Centre Providing Metadata
- K** Knowledge Centre





CLARIN's vision:

Render accessible all digital language resources and tools from all over Europe through a single sign-on online environment.

CLARIN's mission:

Create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the SSH.





**Language
technology**

Services

Expertise

**Language
Resources**

Innovation

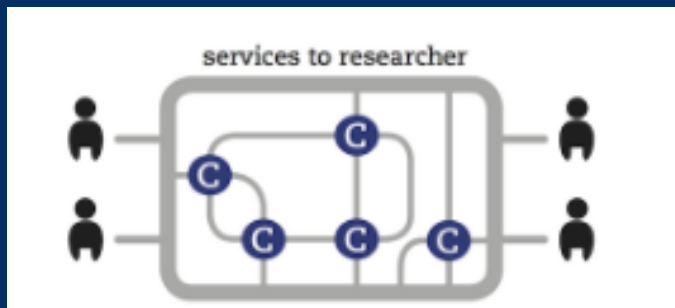
Network

**Research
policy**



CLARIN is a distributed research infrastructure.

Concretely, CLARIN is a networked federation of centers— academic and research institutions that offer the scientific community expertise, resources, tools and services.



Researchers are both service-users and service-providers

**Service Provider
Centers
B-centers**

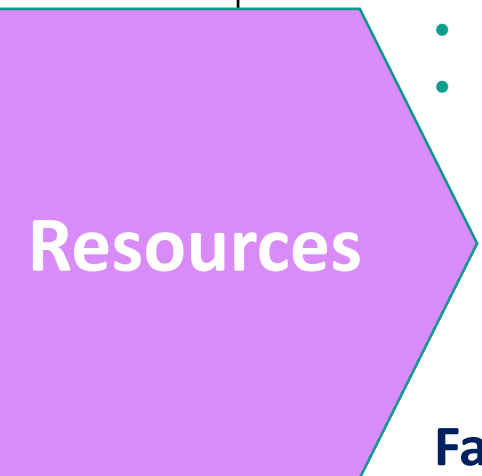
**Knowledge
Centers
K-centers**

**Metaknowledge
centers
C-centers**



Families of corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- **L2 learner corpora**
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora



74
in 2023!

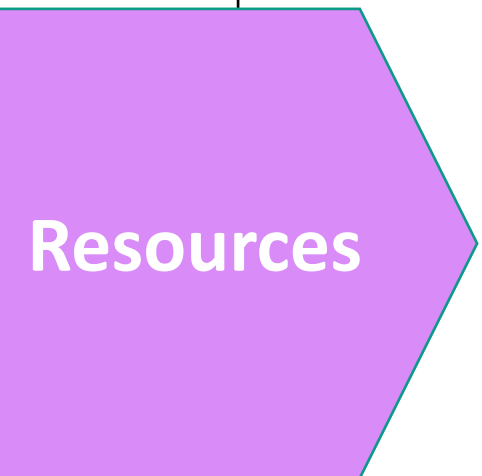
Families of lexical resources

- Lexica
- Dictionaries
- Conceptual resources
- Glossaries
- Wordlists





How to find and access CLARIN resources?



1. Manually, by searching in the Resource Families listing

2. CLARIN's unified catalogue: the Virtual Language Observatory VLO

- An easy-to-use interface
- A powerful query syntax



3. The individual CLARIN data repositories

- Located in the various CLARIN centers

4. The SSH Open Marketplace

- A discovery platform for the SSH field





Tools to discover, explore, exploit, annotate, analyze or combine language data.

- Text normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Sentiment analysis and opinion mining

Text normalization

Tools

PICCL: Philosophical
Integrator of Computational
and Corpus Libraries

Functionality: OCR,
normalisation, tokenisation,
dependency parsing,
shallow parsing,
lemmatisation,
morphological analysis,
NER, PoS-tagging

Domain: independent

Licence: GNU GPL

Dutch, Swedish,
Russian, Spanish,
Portuguese,
English, German,
French, Italian,
Finnish, Modern
Greek, Classical
Greek, Icelandic,
German (Fraktur),
Latin, Romanian

This is a set of workflows for corpus building through
OCR, post-correction, modernisation of historic
language and Natural Language Processing. It
combines [Tesseract Optical Character Recognition](#),
[TICCL](#) and [FROG](#) functionality in a single pipeline.

- **Availability:** [download](#)
- **CLARIN Centre:** CLARIAH-NL
- **Platform:** cross-platform
- **Input format:** images (tiff, vnd.djvu), plain text,
xml
- **Output format:** FoLiA XML
- **Publication:** Reynaert et al. (2015)

<https://www.clarin.eu/content/tools>



CLARIN **Knowledge** Centers (K-centers) share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure.

K-centers have their own specific areas of expertise, which can belong to many different categories, such as:

- ✓ **Individual languages** (e.g. Danish, Czech, Portuguese)
- ✓ **Language families** (e.g. South Slavic)
- ✓ **Groups of languages** (e.g. morphologically rich languages, the languages of Sweden)
- ✓ **Written text and other modalities** (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)

Expertise



Services

1. **Services** provided by K-centers, such as:
 - ✓ Online courses Training materials
 - ✓ Best-practice documents
 - ✓ Guidance in getting access to and using data and tools
 - ✓ Hosting of receivers of CLARIN mobility grants
2. **Learning Hub**
 - ✓ Training Materials
 - ✓ Training Events
 - ✓ Teaching Award
 - ✓ DH Course Registry
3. **CLARIN Trainer Network programme**
4. **Data depositing services**
5. **CLARIN Youtube channel**



Services

6. Technical webservices:

- *Weblicht*: an execution environment for automatic annotation of text corpora, built by CLARIN-D.
- Linguistic tools (tokenizers, speech taggers, parsers) are encapsulated as web services, which can be combined by the user into custom processing chains.
- The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format.
- *Language Resource Switchboard*: a tool that helps to find a matching language processing web application for a set of data

<https://www.clarin.eu/content/language-resource-switchboard>



Network

- 1. Funding opportunities:** CLARIN has developed funding and support instruments for their members to address strategic priorities that require cross-country collaboration, exchange of expertise, training or mobility.
 - ✓ Bridging Gaps Call
 - ✓ Resource Families Project Funding
 - ✓ Workshop Funding
 - ✓ User Involvement Funding
 - ✓ CLARIN Training Calls
 - ✓ CLARIN Seed Grants
 - ✓ Mobility Grants
- 2.** The central CLARIN office and the national consortia can provide **support for preparing EU-funded projects**, such as help for networking and documentation, assistance in writing the proposal, small grants that can be used to cover certain costs coming with the preparation of a project proposal, facilitation services regarding the Open Science and FAIR Data requirements. <https://www.clarin.eu/content/suport-eu-funded-projects>

What is CLARIN-CH?

CLARIN-CH is the Swiss branch of CLARIN

In December 2020, a network of six Swiss academic institutions, supported by the Swiss Academy for SSH, founded the CLARIN-CH Consortium.



Mission

Join the European CLARIN community and build an active and impactful national network.

Objectives:

1. Obtain Switzerland's full membership in CLARIN and connect the Swiss scientific community with the entire CLARIN infrastructure.
2. Foster the sharing of expertise and of resources at the national and European level.
3. Bring together the Swiss community using language resources and create national working groups.
4. Encourage the initiation of national and international collaborations.

National node

TECHNICAL CENTER



University of Zurich ^{UZH}

LiRI - Linguistic Research Infrastructure

LiRI is a national technology platform to serve the needs of quantitative research at UZH, Switzerland and beyond.



NATIONAL REPOSITORY

Library

University of Zurich

SWISS  base

LaRS - Language Repository of Switzerland

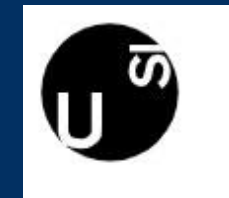
LaRS is a national platform for the publication of linguistic research data. It uses SWISSUbase as its repository system.



TRAINING OFFERS



Doctoral programmes in applied linguistics: Argumentation in Professional Practice



Universität Zürich ^{UZH}

Deutsches Seminar - Zukoko

Sommerschule der Zürcher Korpuslinguistik und Korpuspragmatik

National node

TECHNICAL CENTER



University of Zurich^{UZH}

LiRI - Linguistic Research Infrastructure

LiRI is a national technology platform to serve the needs of quantitative research at UZH, Switzerland and beyond.

NATIONAL REPOSITORY

Library

University of Zurich

SWISS  base

LaRS – Language Repository of Switzerland

LaRS is a national platform for the publication of linguistic research data. It uses SWISSUbase as its repository system.

TRAINING OFFERS



Doctoral programmes in applied linguistics: Argumentation in Professional Practice

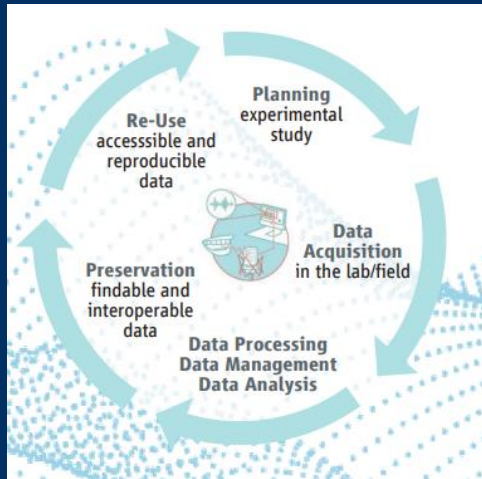


Universität Zürich^{UZH}

Deutsches Seminar – Zukoko

Sommerschule der Zürcher Korpuslinguistik und Korpuspragmatik

LIRI AS CLARIN B-CENTER



- Deals with the entire data life-cycle
- Provides support services, facilities and equipment
- Enables collaborative research
- Deals with big data
- Engages with Open Science
- Enables Data Reproducibility

1. Language Technology

- IT services
- Application development
 - LiRI Corpus Platform (LCP) to handle and query large corpora
- Data processing & NLP
 - Automatic data processing with NLP tools
 - Best practices, coaching, workshops, programming
 - Swissdox@LiRI: a tool that allows retrieving large quantities of Swiss media data for research purposes

2. Lab & Equipment

- On site and portable

3. Statistical consulting

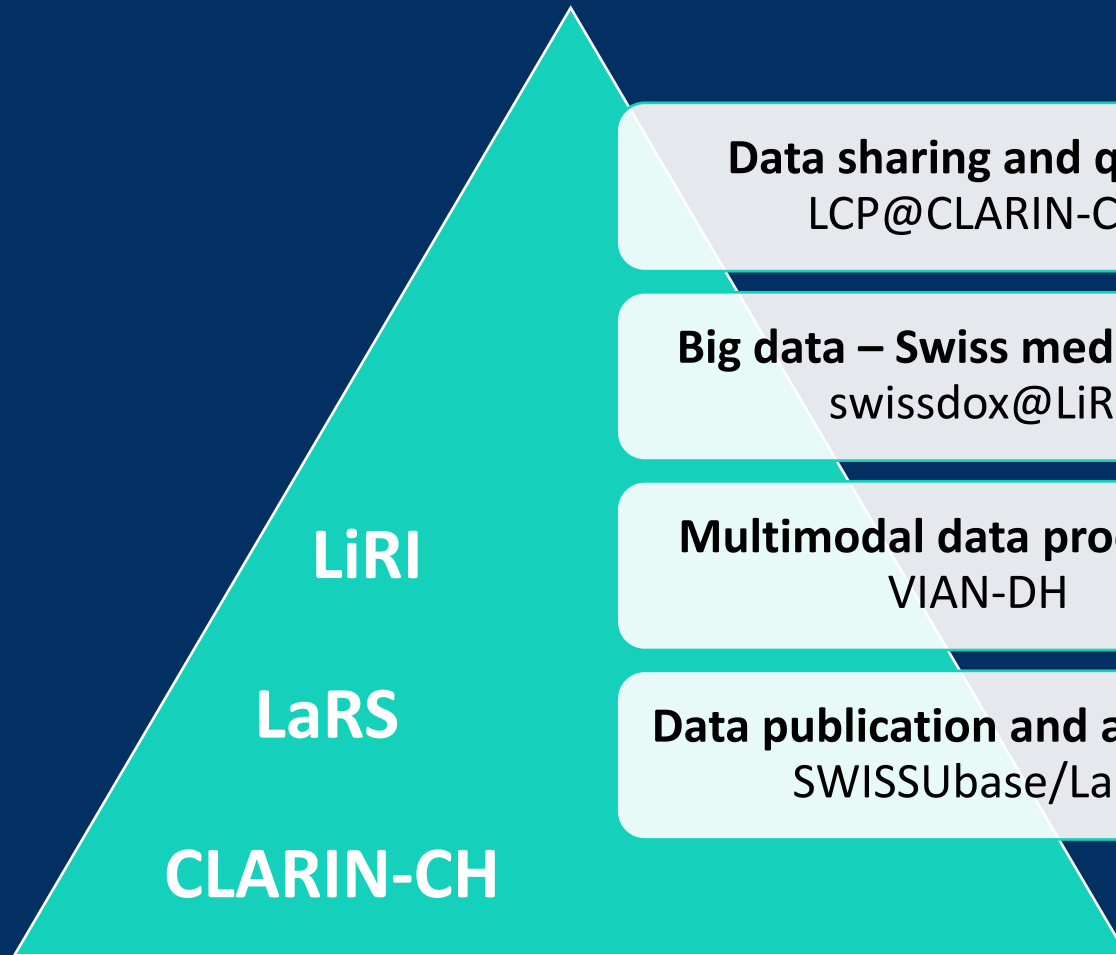
- Advice or feedback on study design, statistical methods
- Data analysis service

SOS Charged services

NEW To be added in the budget of research projects

UPGRADING THE SWISS LINGUISTIC ORD-ECOSYSTEM – *UPLORD* PROJECT

Data Life-cycle



Data sharing and query:
LCP@CLARIN-CH

Big data – Swiss media texts:
swissdox@LiRI

Multimodal data processing:
VIAN-DH

Data publication and archiving:
SWISSUbase/LaRS

FAIR resources
Standardization
Interoperability

Annotation tool for
multimodal data

Sensitive data
Legal issues
Metadata

Training
Documentation
Best practices

Working with us



- In collaboration with Schweizer Mediendatenbank AG, LiRI makes the national Swissdox database easily accessible to **researchers => Swissdox@LiRI**
- **Swissdox@LiRI** includes about **29 million media articles** (press, online) from a wide range of **Swiss media sources** covering many decades and is updated daily with about 5'000 to 6'000 new articles from the German and French speaking parts of Switzerland.
 - Data stock comes from partners such as *CH Media, NZZ media group, Ringier, Ringier Axel Springer Schweiz and TX Group (Tamedia), SRF/SRG, Le Temps, Weltwoche and Wochenzeitung*, overall 260 sources with planned further expansion.
- Usage possibilities:
 - access to Swissdox database with API
 - designed for **big data** analyses
 - data may be enriched, automatically processed and analyzed
- Access upon institutional or project **subscription**

Swissdox@LiRI Corpus query Retrieved datasets Swissdox - T1

Corpus query

Languages *
 German French

Document type *

Source *

Content keywords

Separate multiple entries by comma. (Eg. alps, swiss*, ...)

Date ranges

* leaving the field blank will select all options in this field

NATIONAL REPOSITORY



- LaRS is a FAIR compliant national platform for the publication and long-term archiving of linguistic research data on Swiss servers (SWITCH)
- LaRS collaborates with LiRI to provide **long-term data archiving services** (such as, data processing and conversion, compilation of datasets for archiving, anonymization, metadata) **Services FREE for CLARIN-CH members**
 - info.liri@linguistik.uzh.ch



<https://info.swissubase.ch>

swissubase@ub.uzh.ch

KEY FEATURES

FAIR repository recognised by the
Swiss National Science Foundation

Multi-disciplinary & multilingual

Long-term preservation of your data

Access control options (prior
approval, embargo, closed contract, etc.)

Free persistent identifier (DOI) for
all published datasets

Connected with Swiss and European
key research catalogues

Data curated by archiving experts

Expert support for depositing and
re-using data

Free for researchers

FOSTER SHARING OF RESOURCES

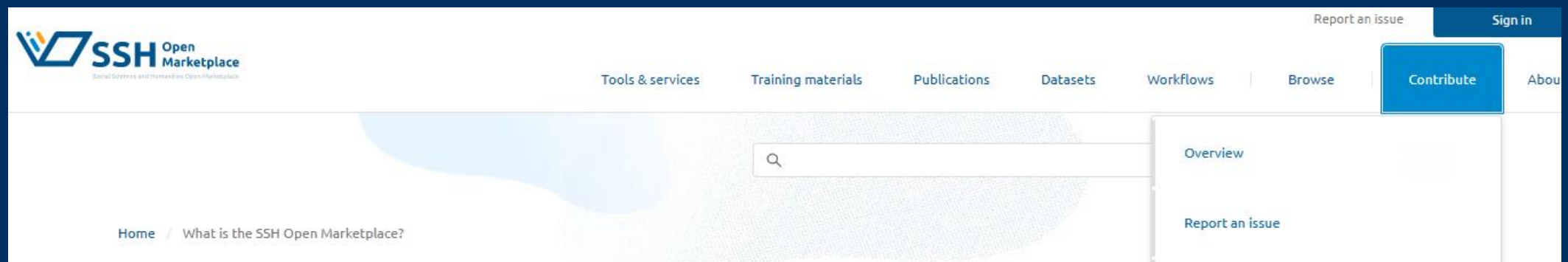
1. Collaborate with LiRI for including corpora in the national **Linguistic Corpus Platform**

- LCP has a Python backend API for querying corpora and a frontend that connects to this API providing a graphical interface for linguistic research
- Release planned: April 2023; learn more [here](#)

2. List them on the clarin-ch.ch website to increase visibility

3. Help with adding some datasets on the **SSH Open Marketplace**

- European discovery platform for resources in the SSH field



CLARIN-CH national working groups



Description: groups of researchers (i.e. CLARIN-CH members, other national and European scholars) that are interested in language-, resource- and infrastructure-related topics.



- Advantage: work together with peers in a formalized and sustainable environment.
- Purpose: bring together expertise on a specific topic, prepare joint research projects and to serve the community, which can be scholarly and technical.
- Long-term goal: build and to extend the CLARIN and CLARIN-CH infrastructure.

CLARIN-CH national working groups



Spring 2023

1. ORD projects for linguistic data
 - i. UpLORD*
 - ii. Swiss-AL: Linguistic ORD Practices for Applied Sciences*
 - iii. Data-sharing skills in corpus-based research on talk-in-interaction*
2. Management of sensitive data and legal aspects for linguistic data in Switzerland
3. Research infrastructures for Argumentation and Rhetoric
4. Multilingual corpora and Second Language Acquisition

OPEN TO
MEMBERS

Other ideas, needs, interests?

Go to <https://www.clarin-ch.ch/> under «CLARIN-CH Working Groups»

Why getting involved? What are the benefits?

At the institutional level

1. **Increased national and international visibility** for Swiss corpus-based projects, corpora and linguistic databases, tools, etc.
2. **Adoption of international standards** to ensure **interoperability** in the construction of national databases and infrastructure.
3. **Involvement in European infrastructure programs and flagship-projects** in Linguistics and its various sub-fields, Natural Language Processing, Machine Translation, etc.
4. **Cost reductions** - by sharing resources nationally and internationally, the effective costs invested for creating new linguistic databases and developing annotation and query tools are diminished.

Why getting involved? What are the benefits?



At the individual level

- 1. Access to all resources, tools and services** available in CLARIN and CLARIN-CH infrastructures.
- 2. Access to the knowledge infrastructure** of CLARIN and CLARIN-CH, which secures a continuous transfer of knowledge and expertise between all members.
- 3. European funding** for training and scientific events for sharing technical expertise and know-how in building the CLARIN infrastructure and to reinforce international collaborations.
- 4. Support** for networking, documentation, and assistance in writing project proposals provided by the central office and the CLARIN members.

Take home messages



1. CLARIN is a distributed research infrastructure for language resources and language technology.
2. CLARIN-CH is a consortium of scholars targeting the development of an active and impactful Swiss CLARIN community.
3. CLARIN and CLARIN-CH are based on networking and sharing of expertise, language resources, and tools.
4. Researchers are both service-users and service-providers.
5. Conducting research with digital language data can be challenging at the individual level, but much easier at the collective level.



Marianne Hundt
National Coordinator
Representative of UZH



Noah Bubenhofer
NC ad interim
Director of LiRI



Anita Auer
President
Representative of UNIL



Cristina Grisot
Scientific Coordinator



Martin Luginbühl
Representative of UNIBAS



Sandrine Zufferey
Representative of UNIBE



Eric Haeberli
Representative of UNIGE



Anita Thomas
Representative of UNIFR



Martin Hilpert
Representative of UNINE



Andrea Rocci
Representative of USI



Cerstin Mahlow
Representative of ZHAW



Beat Immenhauser
Representative of the ASSH

Meet the members of the CLARIN-CH consortium

Home

[National Consortium](#)[Governance and Members](#)[CLARIN-CH resources](#)[CLARIN-CH partnerships](#)[CLARIN-CH co-branding](#)[CLARIN-CH working groups](#)[CLARIN Europe](#)[News](#)

Welcome to CLARIN Switzerland



CLARIN – Common Language Resources and Technology Infrastructure – is a pan-European research infrastructure aiming to render accessible all digital language resources and tools from all over Europe through a single sign-on online environment.



Several Swiss academic institutions have manifested their intention to join CLARIN, first as an Observer member and later as Full member. For this, they have founded the consortium CLARIN-CH on 18 December 2020.

The **mission** of CLARIN-CH: **join the European CLARIN community**, and **build an active and impactful CLARIN-CH community**.

For this, we formulated four **objectives**:

- ✓ Obtain Switzerland's CLARIN membership and give Swiss researchers access to the entire CLARIN infrastructure.
- ✓ Foster the sharing of **resources** and **expertise** among the



EMAIL

cristina.grisot@uzh.ch

WEB

<https://clarin-ch.ch>

SOCIAL MEDIA

<https://www.linkedin.com/company/clarin-ch>

Thank you!