

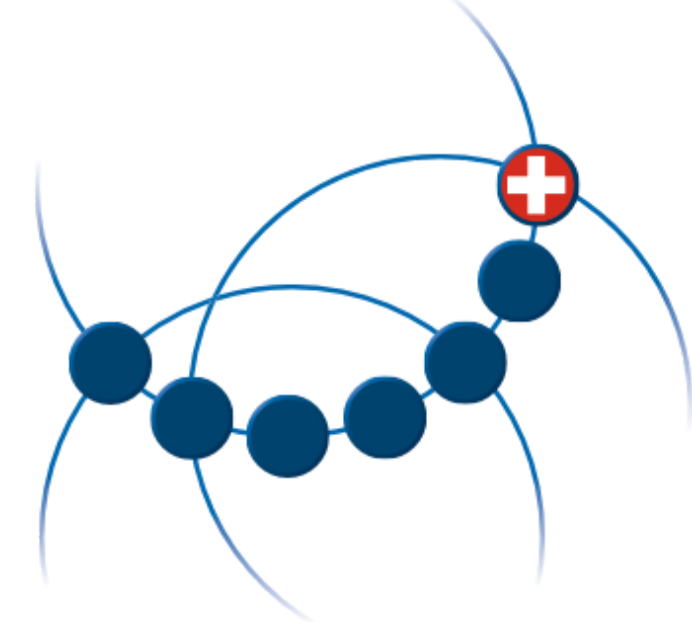
CLARIN

Common Language Resources and
Technology Infrastructure



CLARIN-CH

Common Language Resources and
Technology Infrastructure



INFORMATION SESSION

Tour de Suisse 2022

Outline

Part 1: CLARIN Europe

Part 2: What is CLARIN-CH?

Part 3: Developing CLARIN-CH

Part 4: Benefits

Part 5: Our roadmap



The research infrastructure for language as social and cultural data

CLARIN

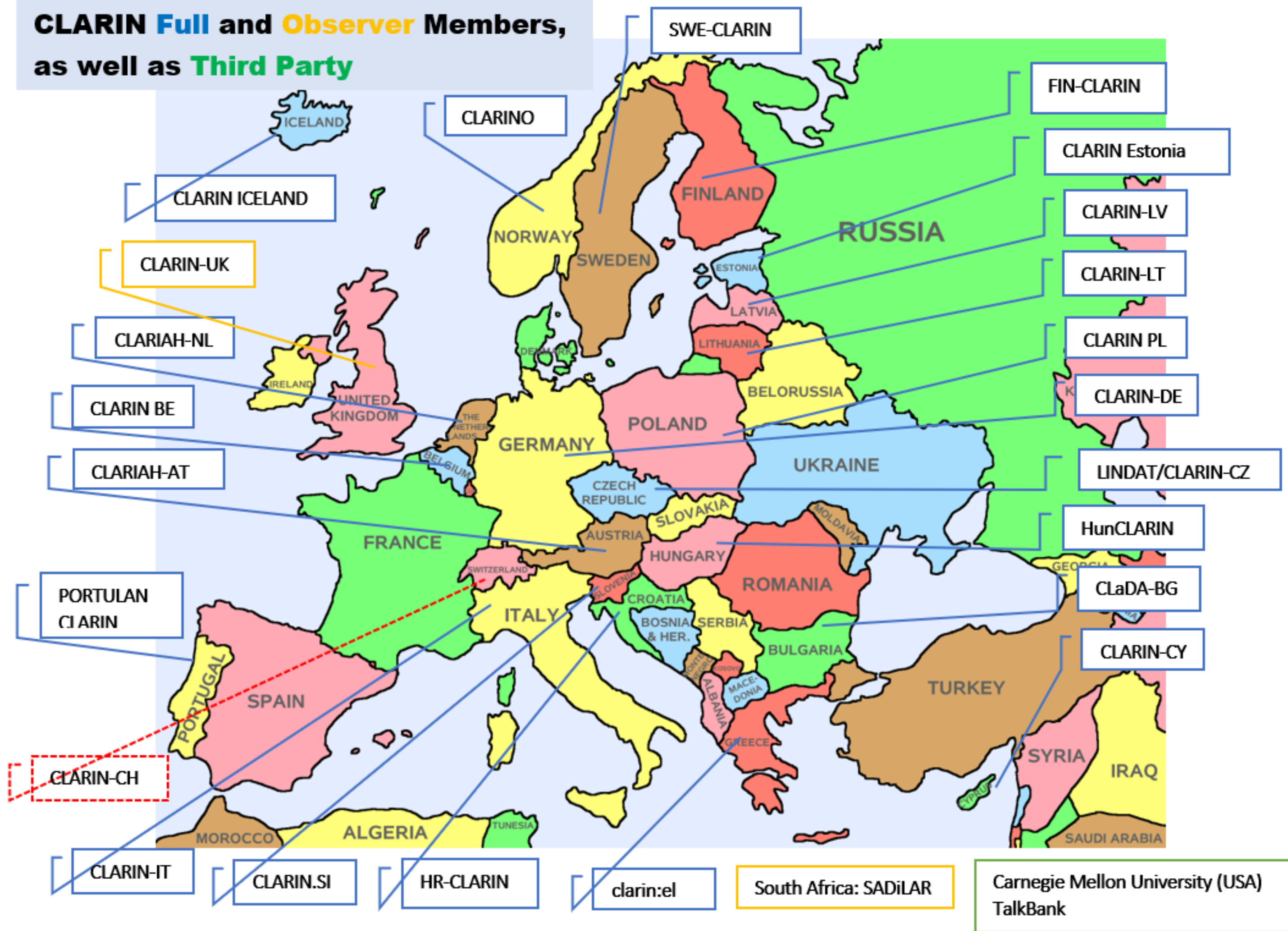


CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

SOME FACTS:

- Designed and implemented between 2008-2012 by 9 founding countries, and fully operational since 2016.
- Included in the *European Strategy Forum on Research Infrastructures* (ESFRI) in 2016.
- The CLARIN infrastructure is governed and coordinated by CLARIN ERIC – a consortium of countries, represented by their research and education ministries.
- Currently, 25 European countries are members of CLARIN. Switzerland is in the process of joining the CLARIN Consortium.

**CLARIN Full and Observer Members,
as well as Third Party**



**CLARIN's vision:**

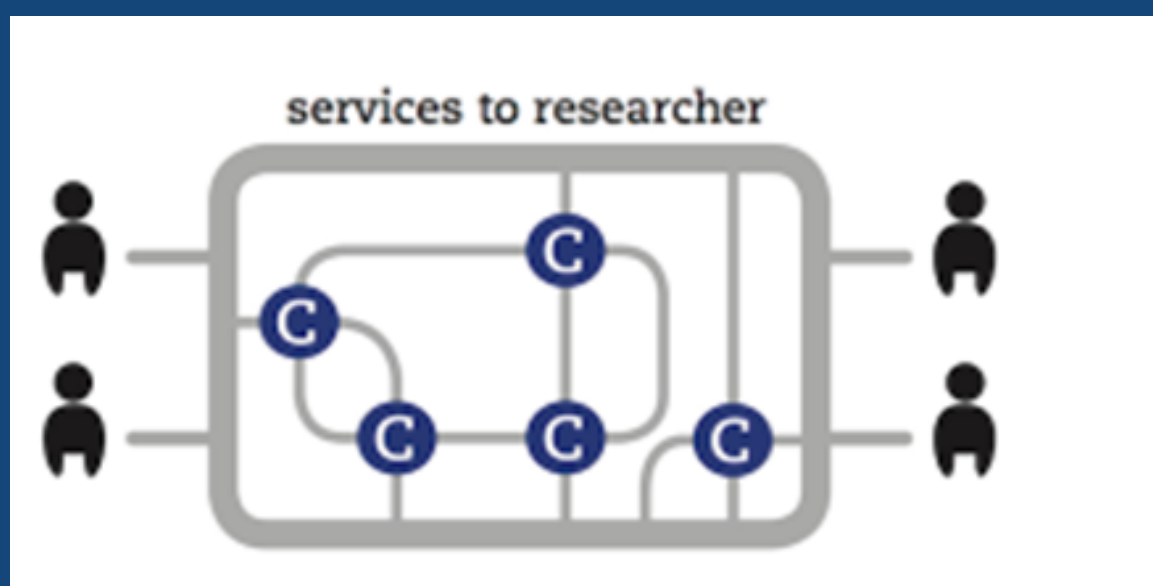
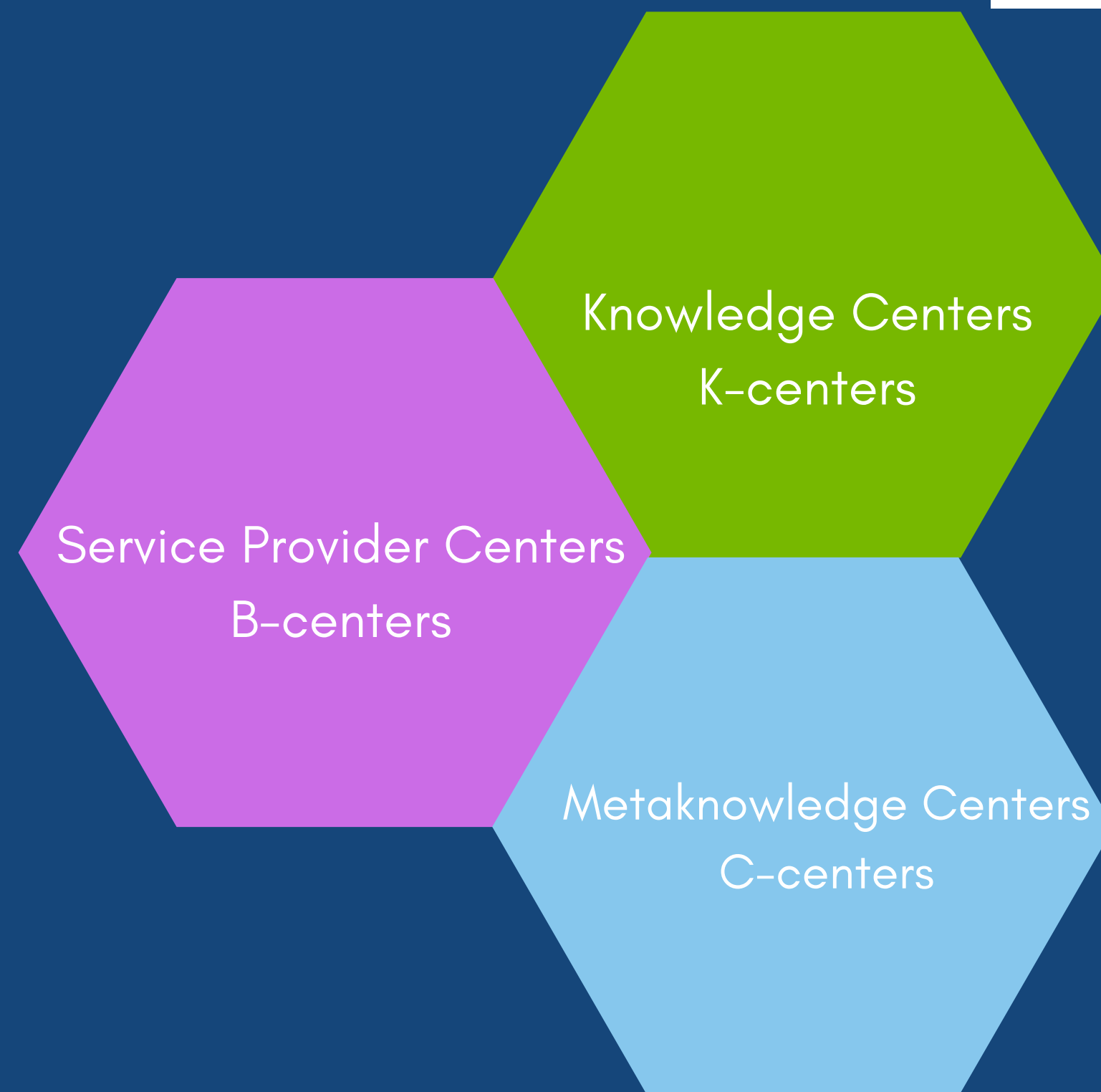
Render accessible all digital language resources and tools from all over Europe through a single sign-on online environment.

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.

Tools**Expertise****Resources****Network****Services**

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



Researchers are service-users and service-providers

Website	http://www.ids-mannheim.de/
Consortium	CLARIN-D
Type(s)	B
Type status	Certified
Description	Providing long-term storage of Germanic language resources.
CoreTrustSeal/DSA	seal
.PID. status	Handle (own server and prefix: 10932).
Repository system	Fedora Commons


Full name	CORLI French CLARIN Knowledge Centre for Corpora, Languages and Interaction
Short name	CORLI-K-centre
URL	https://corli.huma-num.fr/en/kcentre/
Hosted by	(1) Huma-Num / CNRS UMS 3598, Paris, France
City of main hub	Paris
Country of main hub	FR
Date of certification	2020-07-14
Area of competence	Corpus linguistics with a special focus on the French language and the languages of France
Audiences	<ul style="list-style-type: none"> - Linguists - Sociolinguists - Language teachers - Computational linguists - Literature - Language diachrony
Types of services	- How-to documents
Language portal for	- French
Other languages covered	<ul style="list-style-type: none"> - LSF (French Sign Language) - Old French
Modalities covered	<ul style="list-style-type: none"> - Audio: speech - Audio-visual - Text
Linguistic topics	<ul style="list-style-type: none"> - Format and tools for corpus studies - Multilingual, multimodal corpora
Language processing topics	- Corpus format and tools for creating and analysis corpora
Data types	<ul style="list-style-type: none"> - Text - TEI - Spoken language - Video
Resource families	
Generic topics	- Legal issues
Other keywords	- Continuous education



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

Resources



L2 LEARNER CORPORA

GENERAL INFORMATION	AVAILABILITY
<p>34 L2 corpora surveyed</p> <p>10 MULTILINGUAL 24 MONOLINGUAL: 9 LANGUAGES:</p> <p>1 Arabic 2 German 1 Czech 1 Hungarian 10 English 1 Norwegian 4 Finnish 3 Swedish 1 French</p>	<p>5 through a concordancer 15 for download 3 both</p>
ANNOTATION	SIZE
<p>5 PoS-tagged 1 lemmatised</p>	<p>8 small (<10 million tokens) 5 medium (10–100 million tokens) 0 large (>100 million tokens)</p>
LICENCE	
<p>12 CLARIN RES 10 CC-BY 2 ELRA END USER/VAR</p>	



NEWSPAPER CORPORA

GENERAL INFORMATION	AVAILABILITY
<p>27 newspaper corpora surveyed</p> <p>4 MULTILINGUAL 23 MONOLINGUAL: 8 LANGUAGES</p> <p>1 Arabic 2 French 1 Polish 2 Czech 4 German 11 Swedish 1 Finnish 1 Norwegian</p>	<p>5 through a concordancer 10 for download 11 both</p>
ANNOTATION	SIZE
<p>12 PoS-tagged 4 lemmatised</p>	<p>11 small (<10 million tokens) 3 medium (10–100 million tokens) 6 large (>100 million tokens)</p>
LICENCE	
<p>12 CC-BY 4 ELRA END USER/VAR 2 CLARIN PUB 1 CLARIN ACA</p>	

Monolingual L2 learner corpora in the CLARIN infrastructure

Written corpora

Corpus	Language	Description	Availability
CzeSL – Czech as a Second Language Size: 0.9 million words Annotation: tokenised, PoS-tagged, lemmatised, error labels Licence: CC-BY	Czech	This corpus contains essays written in 2013 by learners from 54 L1 backgrounds. The corpus is available for download from LINDAT. For the relevant publication, see Rosen (2016) .	Download
British Academic Written English Corpus Size: 2761 texts Licence: CC-BY	English	This is primarily a L1 corpus although it also contains L2 texts. The corpus is available for download from the University of Oxford Text Archive.	Download

[English Scientific Text Corpus](#) English This corpus contains journal articles in the following disciplines:
 [Concordancer](#)

Size: 35 million tokens
Annotation: PoS-tagged, lemmatised, author/text metadata, document structure
Licence: restricted

- computer science,
- computational linguistics,
- informatics,
- digital construction,
- microelectronics,
- linguistics,
- biology,
- mechanical engineering, and
- electrical engineering.

The articles were published in the 1970s, 1980s and the 2000s.
 The corpus is available for online querying through CQPWeb (CLARIN-D distribution).
 For the relevant publication, see [Degaetano-Ortlieb et al. 2013](#)



Resources

The access to the various resources depends on their type of licence.

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



Lexical resources

- **Lexica**
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

<p>EngVallex</p> <p>Size: 4,337 entries, 7,148 frames</p> <p>Annotation: verb valency</p> <p>Licence: CC-BY-NC-SA 4.0</p>	English	<p>This is a valency lexicon linked to the English side of the PCEDT corpus (WSJ corpus). The resource is available for download from LINDAT and for online browsing.</p>	<p>Browse</p> <p>Download</p>
<p>The Database of Estonian Multi-Word Expressions</p> <p>Size: 12,500 words</p> <p>Licence: proprietary</p>	Estonian	<p>This is a collection of lexica that contain multi-word expressions consisting of a verb and a particle or a verb and its complements. The resource is available for download from META-SHARE (CELR distribution) and for online browsing through a dedicated website.</p>	<p>Browse</p> <p>Download</p>
<p>Démonette</p> <p>Size: 96,027 entries</p> <p>Annotation: MSD-tags (grace format), semantic types</p> <p>Licence: CC-BY 4.0</p>	French	<p>This is a morphological lexicon available for download from ORTOLANG.</p>	<p>Download</p>

- [Easy-to-Use Language Resources](#)
- [Data](#)
- [Tools](#)
- [Resource families](#)**
- [Services](#)

Resource Families

Introduction

The aim of the CLARIN Resource Families initiative is to provide a user-friendly overview per data type of the available language resources in the CLARIN infrastructure for researchers from the digital humanities, social sciences and human language technologies. The overviews are meant to facilitate comparative research and the listings are sorted by language.

The listings for each family include the most important metadata and brief descriptions, such as resource size, text sources, time periods, annotations and licences as well as links to download pages and concordancers. In addition to the resources found in the CLARIN infrastructure, an overview is provided of other existing valuable language resources which have not yet been integrated into the infrastructure.

The listings also provide hyperlinks to other relevant materials, such as the thematic CLARIN workshops and tutorials and their accompanying videolectures, as well as a list of key publications on the resources surveyed.

Currently, overviews are available for 12 corpora families, 5 families of lexical resources, and 4 tool families. See below. For information about applying for funding for small projects that can help to extend the scope of the initiative, see <https://www.clarin.eu/content/clarin-resource-families-project-funding>.

Corpora

- [Computer-mediated communication corpora](#)
- [Corpora of academic texts](#)
- [Historical corpora](#)
- [L2 learner corpora](#)
- [Literary corpora](#)

Lexical Resources

- [Lexica](#)
- [Dictionaries](#)
- [Conceptual Resources](#)
- [Glossaries](#)
- [Wordlists](#)

Tools

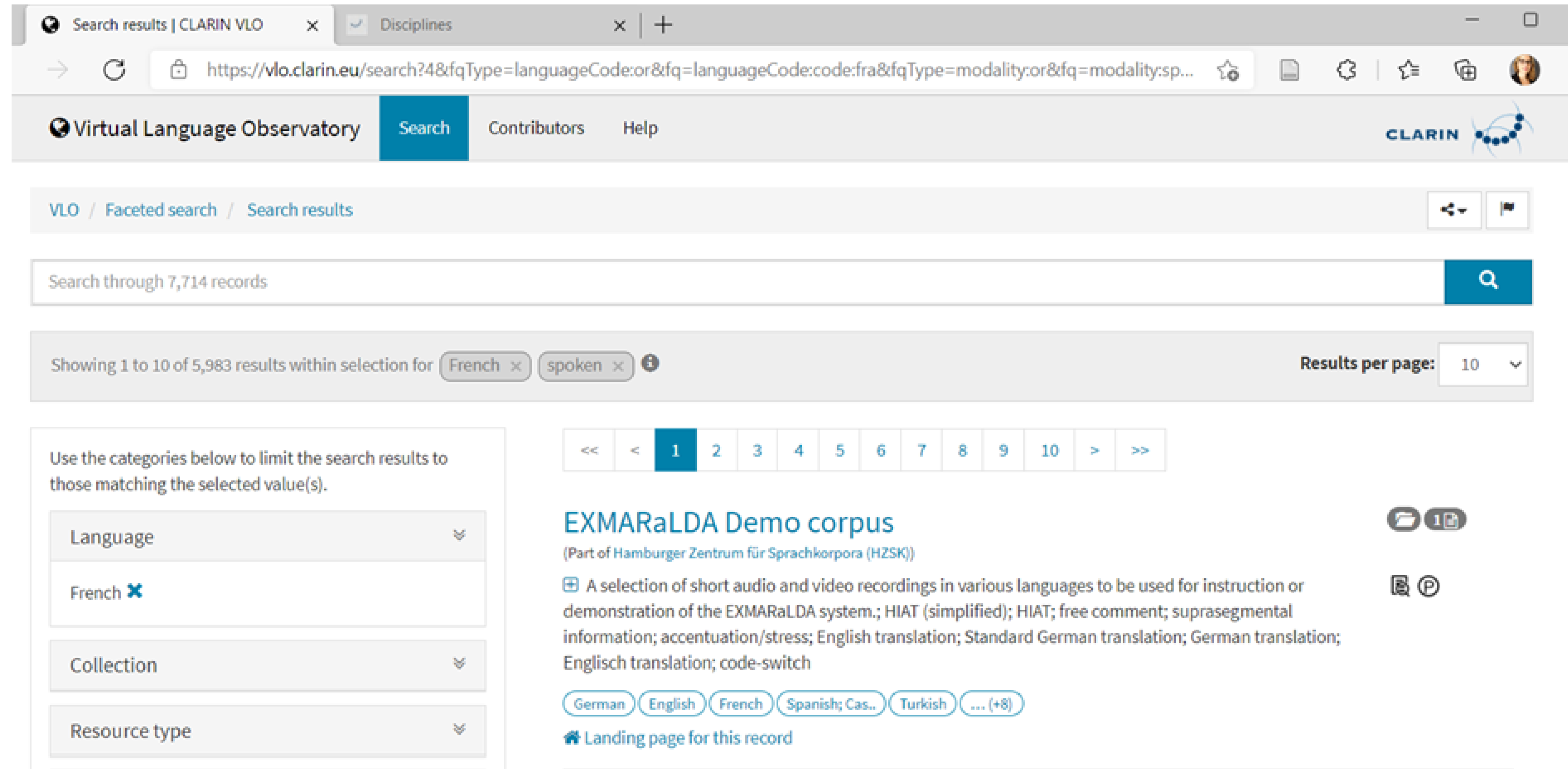
- [Normalization](#)
- [Named entity recognition](#)
- [Part-of-speech tagging and lemmatization](#)
- [Tools for sentiment analysis](#)

Resource families

- [CMC corpora](#)
- [Historical corpora](#)
- [L2 corpora](#)
- [Manually annotated corpora](#)
- [Multimodal corpora](#)
- [Newspaper corpora](#)
- [Parallel corpora](#)
- [Parliamentary corpora](#)
- [Reference corpora](#)
- [Spoken corpora](#)
- [Lexica](#)
- [Dictionaries](#)
- [Conceptual resources](#)
- [Glossaries](#)
- [Wordlists](#)
- [Tools for normalization](#)
- [Tools for named entity recognition](#)

CLARIN language resources can be explored via the individual repositories and via an **unified catalogue**: the **Virtual Language Observatory**

Resources



The screenshot shows a web browser window with the URL <https://vlo.clarin.eu/search?4&fqType=languageCode:or&fq=languageCode:code:fra&fqType=modality:or&fq=modality:sp...>. The page title is 'Virtual Language Observatory' and the search results are for 'French' and 'spoken'. The search results show 5,983 records within the selection. The first result is 'EXMARaLDA Demo corpus' (Part of Hamburger Zentrum für Sprachkorpora (HZSK)). The description of the corpus is: 'A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental information; accentuation/stress; English translation; Standard German translation; German translation; Englisch translation; code-switch'. The language tags are German, English, French, Spanish; Cas..., Turkish, and (+8). There is a link to the landing page for this record.

<https://vlo.clarin.eu/?2#>

Access to **protected resources**

Thanks to a *federated login* – one's institutional credentials – researchers can gain access to protected tools and data sets.

- This functions automatically for academics from all member countries.
- For Switzerland, this will be possible starting with 2023.

Resource or Tool	Description	Provided by
Bavarian Archive for Speech Signals	Mostly German spoken language resources	Bayerisches Archiv für Sprachsignale
CLARIN-DK repository	Danish language resources	The Clarin center at University of Copenhagen
CLARIN.SI repository	Includes a.o. the following corpora: <ul style="list-style-type: none"> • Croatian-English parallel corpus hrenWaC 2.0 • Finnish-English parallel corpus fienWaC 1.0 • Serbian-English parallel corpus srenWaC 1.0 • Slovene-English parallel corpus slenWaC 1.0 	CLARIN.SI Language Technology Centre



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

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

Text normalization

PICCL: Philosophical Integrator of Computational and Corpus Libraries

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

Domain: independent

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)



PoS-tagging and lemmatization

Sparv

Functionality: PoS, MSD, lemma, compound analysis, dictionary lookup

Bulgarian, English, Estonian, Finnish, French, Galician, Italian, Catalan, Latin, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, German

This tool is [Språkbanken](#)'s corpus annotation pipeline infrastructure. The pipeline uses in-house and external tools on the text to segment it into sentences and paragraphs, tokenise, tag parts-of-speech, look up in dictionaries and analyse compounds. The pipeline can also be run using a web API with XML results, and it is run locally to prepare the documents in [Korp](#), which is SWE-LANG's corpus search tool. While the most sophisticated support is for modern Swedish, the pipeline supports additional 19 languages.

Availability: [web application](#), [web API](#)

Input: plain text, XML

Output: plain text, XML

CLARIN Centre: SWE-CLARIN

Related publication: Borin et al. (2016)

Named entity recognition

GATE

Functionality: tokenization, PoS-tagging, NER, semantic and orthographic coreference, pronominal coreference

Platform: cross-platform

Licence: LGPL

English, French, German, Romanian, Russian, Welsh, Danish, Chinese, Arabic

This is a complete NLP platform with modules for named entity recognition.

Availability: [download](#), [online service](#)

CLARIN Centre: CLARIN-UK

NER categories: person, location, organisation, date, percent, money

Publication: Cunningham et al. (2019)



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 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)

<https://www.clarin.eu/content/knowledge-centres>



1. **Services** provided by K-centers can take different shapes:

- 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. **Digital Humanities Course Registry** (CLARIN & DARIAH): an online platform for announcing and finding DH courses across Europe.

3. The **CLARIN Trainer Network Programme**: training events carried out by a group of experts at prominent summer schools, conferences, COST Actions, etc. in disciplines and communities relevant for CLARIN, such as linguistics, digital humanities, language technologies and social sciences.



4. **Depositing Services:** researchers can store language resources in a sustainable repository at a CLARIN centre.

5. The **VideoLectures.net** video channel: an online library of talks (synchronised with their corresponding slides) and tutorials from the CLARIN training and academic events.

6. **Technical webservices:**

- The *Language Resource Switchboard*: a tool that helps to find a matching language processing web application for a set of data.
- *Weblicht*: an execution environment for automatic annotation of text corpora, built by CLARIN-D. Linguistic tools (e.g. 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.
- Other individual web services proposed by certain CLARIN centers.

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

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

Network

2. The central CLARIN office and the national consortia can provide **support for preparing EU-funded**

Projects:

- Help for *networking* and *documentation*
- *Assistance* in writing the proposal
- *Seed grants*: a small grant 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.

Network

Involvement of CLARIN in current European research projects

Since 2015, CLARIN ERIC has been participating as partner in European projects funded by the European Commission and addressing topics related to research infrastructures, mostly in the Horizon 2020 programme.

ELE	European Language Equality for Collaboration	ADAPT Centre	2021-01-2022-06
ENRIITC	European Network of Research Infrastructures & Industry for Collaboration	ESS	2020-01-2022-12
EOSC Future	EOSC-FUTURE aims to combine various initiatives and projects in the field of EOSC (European Open Science Cloud) development.		2021-04-2023-09
ERIC Forum	ERIC Forum Implementation Project	BBMRI	2019-01-2023-12
EUROPEANA-DSI	Develop Europeana as a Digital Service Infrastructure (DSI-4)	Europeana Foundation	2018-09-2020-08
SSHOC	Social Sciences & Humanities Open Cloud	CESSDA	2019-01-2022-04
TRIPLE	Transforming Research through Innovative Practices for Linked interdisciplinary Exploration	CNRS	2019-10-2023-04
UPSKILLS	UPgrading the SKIlls of Linguistics and Language Students	Universita ta Malta	2020-09 - 2023-08



CLARIN-CH

What is CLARIN-CH?



CLARIN-CH is the Swiss node of CLARIN Europe.

Concretely, CLARIN-CH is a network of six Swiss academic institutions, supported by the ASSH, who founded the *CLARIN-CH Consortium* in December 2020.

What is CLARIN-CH?





Our mission

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



1. Obtain Switzerland's CLARIN membership and give Swiss researchers access to the entire CLARIN infrastructure.

2. Foster the sharing of expertise and of resources.

3. Bring together the Swiss community using language resources and create national working groups.

4. Encourage the initiation of national and international collaborations.

In Switzerland, we target the creation of:

one B-center

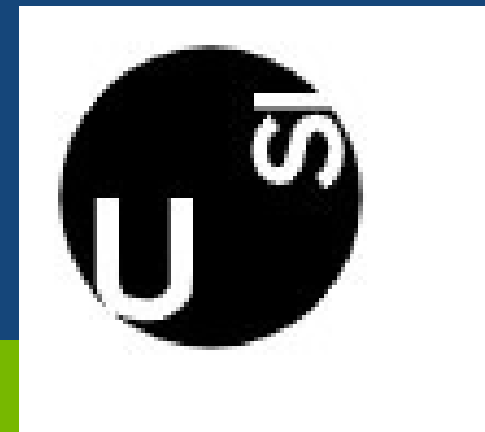


University of Zurich^{UZH}

LiRI - Linguistic Research Infrastructure

several K-centers

discipline specific working groups





University of
Zurich^{UZH}

LiRI - Linguistic Research Infrastructure



LiRI is the national research infrastructure included in SERI's 2021–2024 Roadmap for Research Infrastructures. As a technology platform, LiRI:

- Provides expert support for data management, text processing, and database engineering, as well as statistical consulting services.
 - State-of-art **laboratory facilities and technical infrastructure** for:
 - Linguistic research and the analysis of language and speech
 - Collection of naturalistic data in a methodologically rigorous fashion.
 - Access upon institutional subscription:
<https://www.liri.uzh.ch/en/services/Regulations.html>
- Launched the Swissdox database with press articles.
- Collaborates with SWISSUbase to provide long-term data archiving services.

LiRI as a CLARIN national B-center

↑ Working with us



Unsplash Foto:
Raphael Ferraz

Swissdox@LiRI

- Intro
- Service and conditions
- Subscription
- Roadmap for the year 2022

Direct access for authorized users

Link to → users' guide

Link to → Swissdox@LiRI database

Intro

LiRI cooperates with ↗ **SMD** (Schweizer Mediendatenbank AG) to make the ↗ **Swissdox database** easily accessible to researchers. The database includes about 29 million media articles (press, online) from a wide range of Swiss media sources

Swissdox@LiRI

Members of the following supporter institutions have free access to the Swissdox@LiRI service:

- University of Basel
- University of Bern
- ETH Zurich
- ZHAW Zurich University of Applied Sciences
- University of Zurich

Follow this link → <https://www.li-ri.ch/swissdox@li-ri/>

Accept terms and conditions

- register a project
- make your corpus query
- download the returned datasets in the file requested format (no compression)
- extract the data in order to work with them

Important notice

- for academic use only
- storage is allowed only in one's own IT infrastructure (not an academic institution)
- data may not be shared with third parties
- it is not allowed to remove the entire corpus
- your data have to be deleted six months after project completion

User guide: <https://www.li-ri.ch/swissdox@li-ri/>
More info: <https://www.li-ri.ch/swissdox@li-ri/>

Fact sheet (click to download pdf)

Background

Swissdox@LiRI has been initiated by Prof. Dr. Noah Bubenhofer, Prof. Dr. Fabrizio Gilardi (UZH) and Roberto Nespeca (SMD) and is funded by the University of Zurich UZH (Technology Platform Commission) and the following supporters: **Zurich University of Applied Sciences** (Department of Applied Linguistics), **University Basel / University Library Basel**, **ETHZ Library**, **University Library Bern**.



Access to Swiss Media Database for Academic Use



LiRI cooperates with [SMD](#) (Schweizer Mediendatenbank AG) to make the [Swissdox database](#) easily accessible to **researchers**. The database 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.

List of sources: <https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/core/source>

Access: <https://swissdox.linguistik.uzh.ch/>

User guide: <https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/start>

Swissdox@LiRI has been initiated by Prof. Dr. [Noah Bubenhofer](#), Prof. Dr. [Fabrizio Gilardi](#) (UZH) and [Roberto Nespeca](#) (SMD) and is funded by the University of Zurich UZH (Technology Platform Commission) and the following supporters: **Zurich University of Applied Sciences** (Department of Applied Linguistics), **University Basel / University Library Basel**, **ETHZ Library**, **University Library Bern**.

Free access for all members of the University of Basel (students and staff)!

Query Interface

Swissdox@LiRI

Corpus query Retrieved datasets

Swissdox - T1 ▾

Logout ↗

Corpus query

Languages *

German ×

French ×



Source *

20 minuten (ZWA) ×



Date ranges

2022-02-01 ~ 2022-02-28



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

Reset filters

Next

Document type *

Select document types ▾

Content keywords

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

Submitting Query

Swissdox@LiRI

Corpus query Retrieved datasets

Swissdox - T1 ▾

Logout ↗

Corpus query

Estimated rows: 1

Query name

2022-02-20 01:55

Query comment

Enter a comment for your query

Expiration date

20/02/2022

Send email notification when query finishes

Back

Submit query

YAML

```
query:
  sources:
    - ZWA
  dateRange:
    - 2022-02-01..2022-02-28
  languages:
    - de
    - fr
  result:
    format: TSV
    columns:
      - id
      - pubtime
      - medium_code
      - medium_name
      - rubric
      - regional
      - doctype
      - doctype_description
      - language
```

Retrieve Datasets

Swissdox@LiRI

Corpus query Retrieved datasets

Swissdox - T1 ▾

Logout ↗

Retrieved datasets

#	Name	Submitted	Done	User	Results	Download	Actions
Finished	2022-02-18 12:51	18.02.2022 12:51	18.02.2022 12:52	User	1	↓	i Details ↻ Open query
Finished	2022-02-18 12:41	18.02.2022 12:41	18.02.2022 12:41	User	1	↓	i Details ↻ Open query

Further Services of Language Technology

<https://liri.linguistik.uzh.ch/wiki/langtech/services>

The screenshot shows the LIRI Wiki page for Language Technology services. The header includes the LIRI logo and the text 'LIRI Wiki Linguistic Research Infrastructure - University of Zurich'. A breadcrumb trail indicates the current page: 'You are here: LIRI Wiki > Language Technology > Service description'. A left sidebar contains navigation links for 'Lab', 'Lab Equipment', 'Loan Equipment', 'Language Technology', 'Services', and 'Swissdox@LIRI'. The main content area is titled 'Service description' and lists four categories: 1. Application development, 2. Data processing, 3. Consulting & training, and 4. IT services. Each category has a corresponding sub-section with a brief description and 'Costs' information.

Service description

Services provided by the LIRI Language Technology group (LT) can be subdivided into four categories:

1. Application development
2. Data processing
3. Consulting & training
4. IT services

Application development

The LIRI LT group has vast experience in building interactive web applications, which can be used to collect data and support the development process. LT helps users better understand their requirements and collaborates with them in the development process.

Costs

The effort required for developing an application heavily depends on the required features, e.g. if data is collected from user input need to be implemented or changes are applied automatically, and, obviously, the size and complexity of the project. The effort typically starts in the low three-digits, in terms of number of required development hours.

Data processing

Data processing includes the conversion, cleaning, filtering and automatic processing of linguistic data, such as text-to-speech, speech-to-text, machine learning/deep learning solutions, or purpose-built conversion/extraction pipelines.

Costs

For data archiving via [SWISSUbase](#), LT offers services including extraction and conversion of relevant data.

Consulting & training

The LIRI LT group offers consulting and training in all its areas of expertise. This includes both the processing of large amounts of linguistic data, as well as the development of web applications for working with such data, the preparation of proposals and preparing data management plans (DMPs). Because LT members have backgrounds in linguistics and engineering, LT is able to offer a holistic view of problems that lie at the border of scientific and operational research.

The screenshot shows the SWISSUbase website. The header includes the 'SWISSUbase' logo, a 'Catalogue' link, and language selection options for EN, DE, and FR, along with 'Register' and 'Login' buttons. The main banner features a map of Switzerland with the text 'Find data and projects within Switzerland' and a search bar with the placeholder 'Search the catalogue...'. Below the banner, it states '11832 studies'. The main content area is titled 'SWISSUbase facilitates access to research data and projects across scientific disciplines.' and includes a paragraph describing the platform as a national cross-disciplinary solution for Swiss universities and other research organizations. At the bottom, there are four action buttons: 'Explore the catalogue' (with a magnifying glass icon), 'Get data' (with a download icon), 'Publish your study' (with a box icon), and 'Deposit your data' (with a checkmark icon).

SWISSUbase facilitates access to research data and projects across scientific disciplines.

SWISSUbase is a national cross-disciplinary solution for Swiss universities and other research organisations in need of local institutional data repositories for their researchers. The platform relies on international archiving standards and processes to ensure that data are preserved and accessible in the long-term.

- Explore the catalogue
- Get data
- Publish your study
- Deposit your data

- A platform proposing archiving, data curation and data publication services for data in the SSH field.
- Partners: FORS, UNIL, LiRI/UZH, SWITCH, Swissuniversities
- It draws on the FORSbase platform: an existing national archive solution for Swiss social science research data.
- The module for linguistic data:
 - Will be launched beginning of July for a 2-months phase
 - Need of 10–15 scholars who can submit linguistic data to test the platform and provide feedback
- Access to services: free to the CLARIN-CH community until 2024.

Foster the sharing of expertise



Dialectology
Sociolinguistics
Language ideologies
Pragmatics
Psycholinguistics
Discourse Studies
Prosody and regional variation
Semiology
Corpus Linguistics
Historical Linguistics
Bilingualism
Linguistic Anthropology
Engendered languages
Digital Humanities



Language variation
Language change
Conversational analysis
Linguistic narrativity
Spanish linguistics
Italian language and culture
Epistemology
Dialectology
Slavistics
Stylistics
Media discourse
Language sound structure
Computational linguistics
Language teaching
Digital Humanities



Cognitive Linguistics
Corpus Linguistics
Historical Linguistics
Language teaching
Dialectology
Interactional Linguistics
Discourse Studies
Pragmatics
Semantics
Developmental pragmatics
Cognitive anthropology
Cognitive sociology
Speech therapy
Language pathologies
Corpus statistics



Digital literacy
Discourse Analysis
Language teaching
Political Linguistics
Legal Linguistics
Swiss German Sign Language
Institutional Communication
Social multilingualism
Professional translation
Multimodal communication
Information design
Journalism
Media Linguistics
NLP, Machine Learning
Writing research

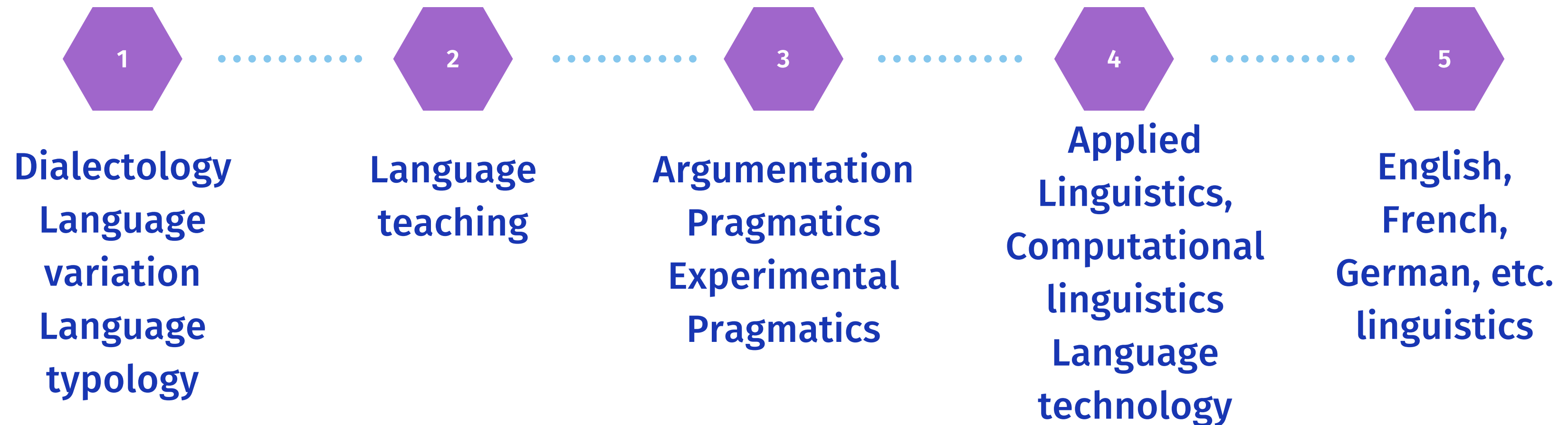


Discourse Studies
Argumentation
Persuasion
Intercultural pragmatics
Financial Communication
Semantics
Multimodality
Bilingualism
Italian Linguistics
Applied Linguistics
Stylistics

Foster the sharing of expertise and encourage collaborations



Discipline specific national working groups: examples



Digital literacy in university contexts (DigLit)

As digitalisation progresses, new technologies can offer valuable support to academic writers. However, research and practice show that students and their lecturers are often not aware of the applicability of available tools and are unable to exploit their full potential. These are the issues addressed by the DigLit project, supported by swissuniversities and the partner institutions.

Leading house: ZHAW School of Applied Linguistics

Project financing: P-8 Digital Skills programme, coordinated by swissuniversities and the partner universities

Overall budget: CHF 2.06 million

Project duration: January 2021 – December 2024



swissuniversities



Support we need to create Working Groups

- Manifest your interest to the CLARIN-CH representant from your institution.
- Discuss your interest with other colleagues.
- Start from already existing collaborations.
- Contact the CLARIN-CH coordination office to help establishing a collaboration at the national level.

In this way, we can work together from the beginning to develop a national network suitable for the needs of every interested discipline and to gather our resources in a sustainable, open and interoperable way.

**Discipline-specific
Working Groups**



**Future
CLARIN K-centers**

Foster the sharing of resources: corpora and databases



- Interrogation Programme & Supersenses Extraction (IT)
- Lyra database (IT)
- ENIAT database (Slavic and South Asian Languages)
- FLORALE database (audio, FR)
- Phonocolor.ch (FR L2)
- COSUIZA (SP)
- COLESfran (SP)
- Dialectos del español
- Fueros medievales
- Mapa del español en Suiza



- ORFEUS databasis (FR)
- Glossaire de la Suisse Romande (FR)
- OFROM (audio and transcribed, FR)



- Swiss Web Corpus for Applied Linguistics (multilingual)



- Ortsnamen.ch/toponymes.ch (Swiss and international toponymy)
- Schweizer Textcorpus (DE)
- Schweizerisches Idiotikon (Swiss German)
- Vocabolario dei dialetti della Svizzera italiana (IT)
- Dicziunari Rumantsch Grischun (Romansh)
- Medieval Latin Dictionary
- Thesaurus linguae Latinae (ancient Latin)
- Historical Dictionary of Switzerland (DE, FR, IT)
- Collocation dictionary (DE)
- OLdPhras (DE)

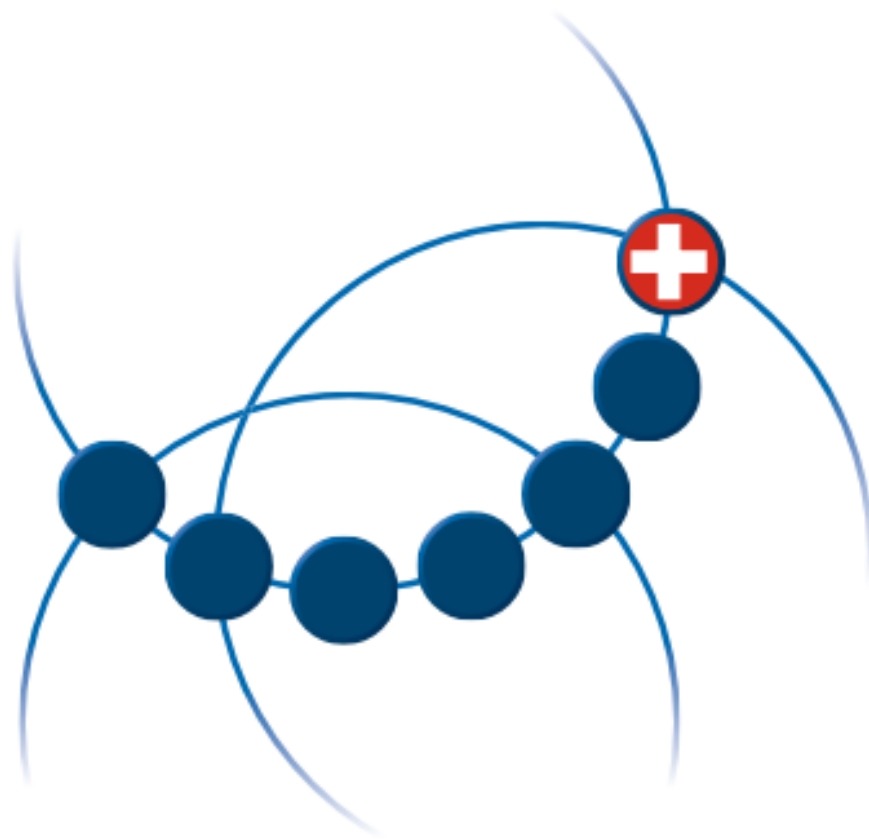
Foster the sharing of resources: corpora and databases



- BNC Dependency Bank (EN)
- Giessen-Long Beach Chaplin Corpus (EN)
- Zurich English Newspaper
- International Corpus of English
- PADLF-Les plus anciens documents linguistiques de la France (FR)
- Tzéro database (FR)
- SenS-Korpus (DE)
- Archimob (DE)
- Picture postcard corpus (DE)
- DAVADS -Digital Audio/Video Archive (DE)
- Metalanguage Discourses (DE)
- Zurich summer 1968 (DE)
- Temporal entity extraction from historical texts (DE)
- JAKOB lexicon (DE)
- SADS-Syntactic Atlas of German-speaking Switzerland
- bulletin4corpus (DE, FR, IT, EN)
- sms4science (DE, FR, IT, Romansh)
- SMULTRON-Stockholm Multilingual Treebank (Swiss German, DE, IT, FR, Romansh)
- eSSRQ-Electronic Collection of Swiss Legal Sources (FR, IT, DE, Romansh, Latin)
- Phonogram Archives (Swiss dialects)
- Text+Berg corpus (Swiss German, DE, IT, FR, EN, Romansh)
- Bullinger Digital corpus (DE, Latin)
- CoNTra: the Federal Gazette (parallel FR-DE)
- PHOIBLE database (more than 1000 languages)
- Chintang Language Corpus
- Nepali National Corpus
- SEAlang corpus (South Asian languages)
- Corporum (Latin)
- European Language Equity resource collection (~60 corpora, ~40 applications, some lexica) (DE, IT, FR, Romansh)
- Genes and Languages Together database
- AUTOTYP database
- Macedonian Spoken Corpus
- Pre-Standardized Balkan Slavic Literature
- Torlak (a language from Timok area)
- Serbian Forms of Address (Serbian)

CLARIN-CH

Common Language Resources and
Technology Infrastructure



**Why get involved?
What are the benefits?**



At the institutional level

- **Increased national and international visibility** for Swiss corpus-based projects, corpora and linguistic databases, tools, etc.

- **Adoption of international standards** to ensure **interoperability** in the construction of national databases and infrastructure.

- **Involvement in European infrastructure programs and flagship-projects** in Linguistics and its various sub-fields, Natural Language Processing, Machine Translation, etc.

- **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.



At the individual level

- Access to all **resources, tools** and **services** available in CLARIN and CLARIN-CH infrastructures.

- Access to the **knowledge infrastructure** of CLARIN and CLARIN-CH, which secures a continuous transfer of knowledge and expertise between all members.

- **European funding** for training and scientific events for sharing technical expertise and know-how in building the CLARIN infrastructure and to reinforce international collaborations.

- **Support** for networking, documentation, and assistance in writing project proposals provided by the central office and the CLARIN members.

Doctoral Programme in Applied Linguistics: *Managing Languages, Arguments and Narratives in the Datafied Society*

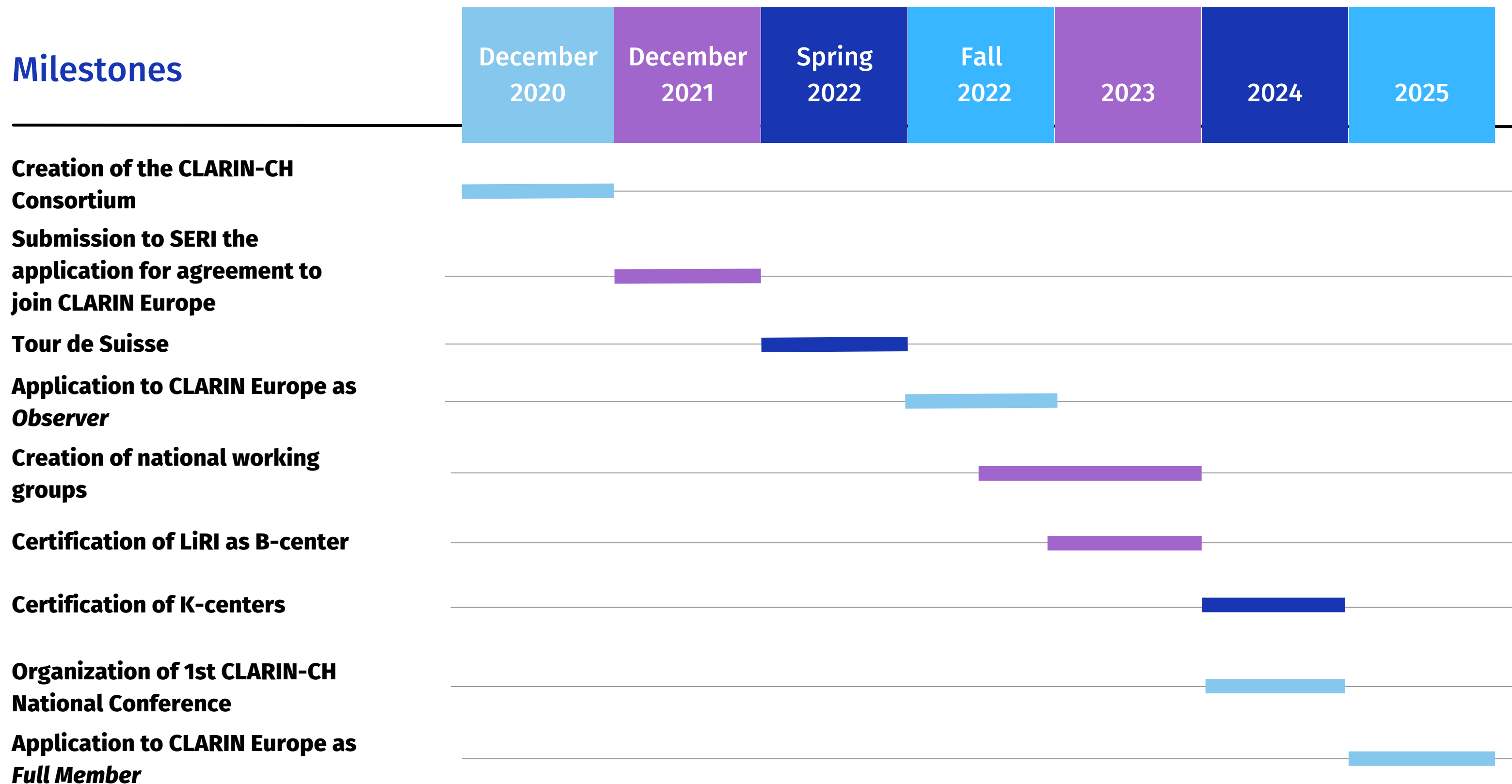


- Co-organized by the ZHAW School of Applied Linguistics and the USI Faculty of Communication, Culture and Society, and funded by Swiss universities.
- The programme is geared towards doctoral students whose research project focuses on languages, data, argumentation and narration.
- All courses are **free for PhD students from Swiss universities that are members of the CLARIN-CH network.**
- Participation is possible in individual courses or in the entire doctoral programme.
- E.g. *Python for Linguists, Workshop Annotation, Designing empirical studies in linguistics, Open Science Publication*

<https://www.zhaw.ch/en/linguistics/study/doctoral-programmes/doctoral-programme-2021-2024/>

Our roadmap until 2025

Milestones





Take home messages:

- CLARIN is a community: a networked federation of centers.
- CLARIN-CH is a consortium of scholars targetting the development of an active and impactful Swiss CLARIN community.
- CLARIN and CLARIN-CH are based on networking and sharing of expertise, resources, and tools.
- Researchers are both service-users and service-providers.
- Conducting research with digital language and text data can be challenging at the individual level, but much easier at the collective level.

Meet the founding members of the CLARIN-CH Consortium



Marianne Hundt

National Coordinator
Representative of UZH



Anita Auer

President
Representative of UNIL



Cristina Grisot

Scientific Coordinator



Sandrine Zufferey

Representative of UNIBE



Martin Hilpert

Representative of UNINE



Andrea Rocci

Representative of USI



Cerstin Mahlow

Representative of ZHAW

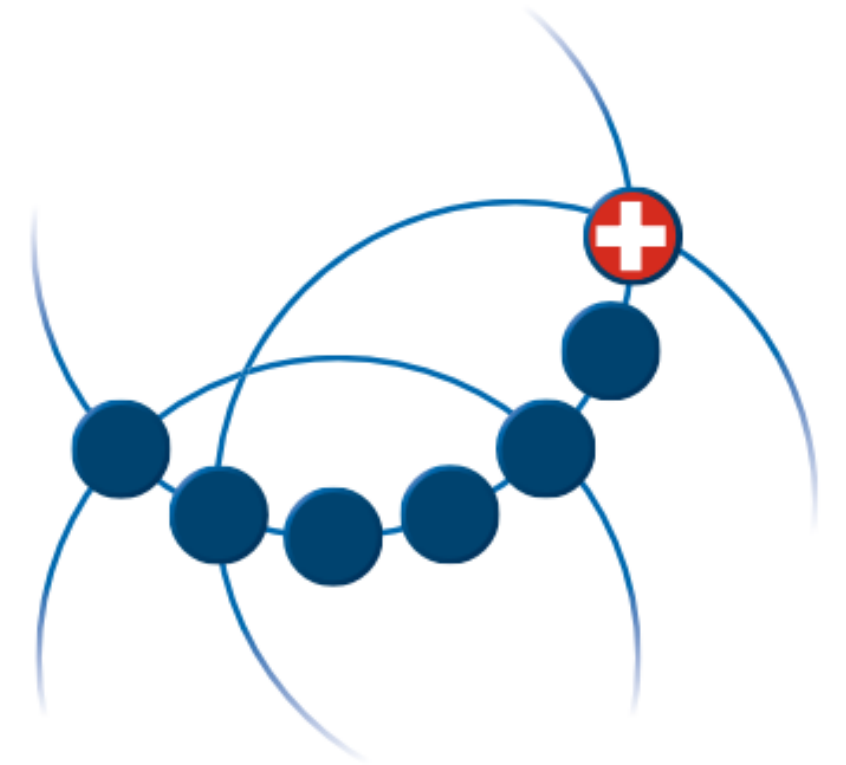


Beat Immenhauser

Representative of SAGW/ASSH

CLARIN-CH

Common Language Resources and
Technology Infrastructure



EMAIL

cristina.grisot@uzh.ch

WEB

<https://clarin.eu>

<https://clarin-ch.ch>

LINKEDIN PAGE

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

Thank you!