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 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.
- 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'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.
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
Leibniz-Institut für Deutsche Sprache

<table>
<thead>
<tr>
<th>Website</th>
<th><a href="http://www.ids-mannheim.de/">http://www.ids-mannheim.de/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortium</td>
<td>CLARIN-D</td>
</tr>
<tr>
<td>Type(s)</td>
<td>B</td>
</tr>
<tr>
<td>Type status</td>
<td>Certified</td>
</tr>
<tr>
<td>Description</td>
<td>Providing long-term storage of Germanic language resources.</td>
</tr>
<tr>
<td>CoreTrustSeal/DSA</td>
<td>seal</td>
</tr>
<tr>
<td>PID status</td>
<td>Handle (own server and prefix: 10932).</td>
</tr>
<tr>
<td>Repository system</td>
<td>Fedora Commons</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Full name</th>
<th>CORLI French CLARIN Knowledge Centre for Corpora, Languages and Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short name</td>
<td>CORLI-K-centre</td>
</tr>
<tr>
<td>Hosted by</td>
<td>(1) Huma-Num / CNRS UMS 3598, Paris, France</td>
</tr>
<tr>
<td>City of main hub</td>
<td>Paris</td>
</tr>
<tr>
<td>Country of main hub</td>
<td>FR</td>
</tr>
<tr>
<td>Date of certification</td>
<td>2020-07-14</td>
</tr>
<tr>
<td>Area of competence</td>
<td>Corpus linguistics with a special focus on the French language and the languages of France</td>
</tr>
</tbody>
</table>

**Audiences**
- Linguists
- Sociolinguists
- Language teachers
- Computational linguists
- Literature
- Language diachrony

**Types of services**
- How-to documents

**Language portal for**
- French

**Other languages covered**
- LSF (French Sign Language)
- Old French

**Modalities covered**
- Audio: speech
- Audio-visual
- Text

**Linguistic topics**
- Format and tools for corpus studies
- Multilingual, multimodal corpora

**Language processing topics**
- Corpus format and tools for creating and analysis corpora

**Data types**
- 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
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

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

https://www.clarin.eu/resource-families
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.
CLARIN language resources can be explored via the individual repositories and via an **unified catalogue**: the Virtual Language Observatory

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.

<table>
<thead>
<tr>
<th>Resource or Tool</th>
<th>Description</th>
<th>Provided by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bavarian Archive for Speech Signals</td>
<td>Mostly German spoken language resources</td>
<td>Bayerisches Archiv für Sprachsignale</td>
</tr>
<tr>
<td>CLARIN-DK repository</td>
<td>Danish language resources</td>
<td>The Clarin center at University of Copenhagen</td>
</tr>
<tr>
<td>CLARIN.SI repository</td>
<td>Includes a.o. the following corpora:</td>
<td>CLARIN.SI Language Technology Centre</td>
</tr>
<tr>
<td></td>
<td>• Croatian-English parallel corpus hrenWaC 2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Finnish-English parallel corpus fienWaC 1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Serbian-English parallel corpus srenWaC 1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Slovene-English parallel corpus slenWaC 1.0</td>
<td></td>
</tr>
</tbody>
</table>

https://www.clarin.eu/content/easy-access-protected-resources
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)

[https://www.clarin.eu/content/tools](https://www.clarin.eu/content/tools)
PoS-tagging and lemmatization

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, tokenize, 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

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)

Tools

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

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 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

[https://www.clarin.eu/funding](https://www.clarin.eu/funding)
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.

https://www.clarin.eu/content/support-eu-funded-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.

<table>
<thead>
<tr>
<th>Project</th>
<th>Description</th>
<th>Partner</th>
<th>Start Date - End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELE</td>
<td>European Language Equality for Collaboration</td>
<td>ADAPT Centre</td>
<td>2021-01-06 - 2022-06</td>
</tr>
<tr>
<td>ENRIITC</td>
<td>European Network of Research Infrastructures &amp; Industry for Collaboration</td>
<td>ESS</td>
<td>2020-01-01 - 2022-12</td>
</tr>
<tr>
<td>EOSC Future</td>
<td>EOSC-FUTURE aims to combine various initiatives and projects in the field of EOSC (European Open Science Cloud) development.</td>
<td></td>
<td>2021-04-01 - 2023-09</td>
</tr>
<tr>
<td>ERIC Forum</td>
<td>ERIC Forum Implementation Project</td>
<td>BBMRI</td>
<td>2019-01-01 - 2023-12</td>
</tr>
<tr>
<td>EUROPEANA-DSI</td>
<td>Develop Europeana as a Digital Service Infrastructure (DSI-4)</td>
<td>Europeana Foundation</td>
<td>2018-09-01 - 2020-08</td>
</tr>
<tr>
<td>SSHOC</td>
<td>Social Sciences &amp; Humanities Open Cloud</td>
<td>CESSDA</td>
<td>2019-01-01 - 2022-04</td>
</tr>
<tr>
<td>TRIPLE</td>
<td>Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration</td>
<td>CNRS</td>
<td>2019-10-01 - 2023-04</td>
</tr>
<tr>
<td>UPSKILLS</td>
<td>UPgrading the SKILLS of Linguistics and Language Students</td>
<td>Universita ta Malta</td>
<td>2020-09-01 - 2023-08</td>
</tr>
</tbody>
</table>
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.
Our mission

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

Objectives

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:

- several K-centers
- discipline specific working groups
- one B-center

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
Swissdox@LiRI

LiRI - Linguistic Research Infrastructure – Swissdox@LiRI

→ Working with us

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 20 million media articles (press, online) from a wide range of Swiss media sources.
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](https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/core/source)
Access: [https://swissdox.linguistik.uzh.ch/](https://swissdox.linguistik.uzh.ch/)
User guide: [https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/start](https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/start)

Swissdox@LiRI has been initiated by Prof. Dr. **Noah Bubenkofer**, 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)!**
Corpus query

Languages *

- German
- French

Source *

- 20 minuten (ZWA)

Date ranges

2022-02-01 ~ 2022-02-28

Document type *

Select document types

Content keywords

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

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

Reset filters  Next
Submitting Query

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

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
### Retrieved datasets

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Submitted</th>
<th>Done</th>
<th>User</th>
<th>Results</th>
<th>Download</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finished</td>
<td>2022-02-18 12:51</td>
<td>18.02.2022</td>
<td>18.02.2022</td>
<td>User</td>
<td>1</td>
<td></td>
<td><img src="#" alt="Details" /> <img src="#" alt="Open query" /></td>
</tr>
<tr>
<td>Finished</td>
<td>2022-02-18 12:41</td>
<td>18.02.2022</td>
<td>18.02.2022</td>
<td>User</td>
<td>1</td>
<td></td>
<td><img src="#" alt="Details" /> <img src="#" alt="Open query" /></td>
</tr>
</tbody>
</table>
Further Services of Language Technology

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

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 user input. LT helps users better understand their requirements and collaborates with them in the definition of the development process. LT helps users better understand their requirements and collaborates with them in the definition of the development process.

Costs

The effort required for developing an application heavily depends on the required features. For example, if data is collected online, the cost of implementing this feature is significantly lower than if the data is collected through interviews.

Data processing

Data processing includes the conversion, cleaning, filtering, and automatic processing of linguistic data. The process involves developing solutions, or purpose-built conversion/routing pipelines.

Costs

Data processing costs can vary depending on the complexity of the task. Generally, costs are estimated based on the number of hours required for the process.

Consulting & training

The LIRI LT group offers consulting and training in all its areas of expertise. This includes both the process of large amounts of linguistic data, as well as the development of web applications. LT provides guidance and support to researchers and professionals in the field of natural language processing.
• 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.

https://info.swissubase.ch/  Florian Steurer: swissubase@ub.uzh.ch
Foster the sharing of expertise
Foster the sharing of expertise and encourage collaborations

Discipline specific national working groups: examples

1. Dialectology
   - Language variation
   - Language typology

2. Language teaching

3. Argumentation Pragmatics
   - Experimental Pragmatics

4. Applied Linguistics, Computational linguistics
   - Language technology

5. English, French, German, etc. linguistics
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
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 medievals
- Mapa del español en Suiza

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

- Swiss Web Corpus for Applied Linguistics (multilingual)

- Ortsnamen.ch/topnymes.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)
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 Swissuniversities.
- 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

## Our roadmap until 2025

<table>
<thead>
<tr>
<th>Milestones</th>
<th>December 2020</th>
<th>December 2021</th>
<th>Spring 2022</th>
<th>Fall 2022</th>
<th>2023</th>
<th>2024</th>
<th>2025</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation of the CLARIN-CH Consortium</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Submission to SERI the application for agreement to join CLARIN Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tour de Suisse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application to CLARIN Europe as Observer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creation of national working groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certification of LiRI as B-center</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certification of K-centers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization of 1st CLARIN-CH National Conference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application to CLARIN Europe as Full Member</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Take home messages:

- CLARIN is a community: a networked federation of centers.
- CLARIN-CH is a consortium of scholars targeting 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
Thank you!

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