CLARIN-CH Day 2024

Open Research Data: Challenges and Opportunities

Book of Abstracts

Université de Neuchâtel, 09. September 2024

CLARIN-CH

Organizing Committee

Anita Auer (UNIL)
Cristina Grisot (UZH, CLARIN-CH national coordinator)
Martin Hilpert (UNINE)
Julia Krasselt (ZHAW)
Martin Luginbühl (UNIBAS)
Johanna Miecznikowski-Fünfschilling (USI)
Seraina Nadig (CLARIN-CH)
Melanie Röthlisberger (UZH)
Simon van Rekum (ZHAW)

The event is organized with the financial support of the CLARIN-CH Consortium, the Swiss Academy for Humanities and Social Sciences, the Zurich University of Applied Sciences and hosted by the University of Neuchâtel.

Content

UpLORD: Upgrading the linguistic ORD-ecosystem

Upgrading the linguistic ORD-ecosystem, short UpLORD is a swissuniversities ORD-funded 2-year project (2023-2024, with an extension until June 2025) hosted by the University of Zurich, with the participation of the Linguistic Research Infrastructure LiRI, the Zurich University Library and the CLARIN-CH Coordination Office.
Since 2018, a consortium of partners has been working on building a national ecosystem of infrastructures, which covers the whole linguistic data lifecycle according to ORD requirements (FAIR principles: Findable, Accessible, Interoperable, Reusable) from data generating, processing and analyzing to data sharing and archiving. This ecosystem includes : (i) the Linguistic Research Infrastructure LiRI which is a national technology platform hosted at the University of Zurich, (ii) the national repository for publishing and archiving linguistic data: LaRS@SWISSUbase (iii) a database of Swiss media texts: Swissdox@LiRI, (iv) a corpus platform for hosting of and searching in large text and audio/video corpora: LCP@LiRI.
The mission of this national ecosystem is to provide the Swiss scientific communities using linguistic data the services and the infrastructure necessary for their data to adhere to FAIR principles, and therefore to adopt sustainable practices that support sustainable linguistic research data, which in turn, will lead to replicable and sustainable research results.
Up to now, the project has made progress on:

- upgrading workflows and interoperability of existing infrastructure services (APIs, harvesting by CLARIN Virtual Language Observatory, national corpus platform for text, video and audio language data LCP@LiRI)
- establishing CLARIN-CH working groups on the national level
- documenting and promoting best practices
- raising awareness and training about ORD practices in the context of teaching, research and publishing
- building a robust practice of data curation.

Until its end in June 2025, the UpLORD project will pursue its work on :

- populating the LCP@LiRI with various corpora
- organising training and hands-on sessions on FAIR-compliant data management, data conversion and upload on the LCP, using the LCP and Swissdox@LiRI for research

Principal investigators:
Noah Bubenhofer (LiRI)
Andrea Malits (Universitätsbibliothek Zürich)
Cristina Grisot (CLARIN-CH)

https://www.liri.uzh.ch/en/projects/UpLORD.html

CHORD-talk-in-interaction: Data-sharing skills in corpus-based research on talk-in-interaction

The project (April 2023-September 2024) is funded by swissuniversities, Università della Svizzera italiana, and the universities of Basel, Lausanne and Neuchâtel. Based on experiences and expertise in the fields of Interactional Linguistics, Conversation Analysis and dialogue-oriented Argumentation Studies, the project members explore practices of sharing and reusing audio-video recorded and transcribed corpora of naturally occurring spoken interaction. They describe and asses existing practices and discuss possibilities of improvement, with a particular focus on researchers' data-sharing skills and the Swiss situation, considering also the issue of digital infrastructure. They conduct a theoretical reflection on the challenges that these data raise and on what this example reveals more generally about ORD practices as socio-technological practices in the scientific domain.

To pursue these goals, the team has organised five workshops with experts from the University of Bologna, the Institut für Deutsche Sprache IDS in Mannheim, the University of Southern Denmark, the ICAR Laboratory in Lyon, and USI. Furthermore, project members have attended a series of ORD event organized in Europe, conducted a survey on data-sharing practices in Switzerland, reviewed relevant literature and successfully participated in funding requests for subsequent projects to develop the Swiss digital infrastructure for spoken and multimedia corpora. The team collaborates with Clarin-CH and Swiss ORD research projects in Linguistics and engages with relevant international communities via various channels. The project results are documented in reports and articles, which are published on the project website and distributed via a newsletter and social media.

Principal investigators:

Johanna Miecznikowski-Fuenfschilling (USI)
Sara Greco (USI)
Andrea Rocci (USI)
Lorenza Mondada (UNIBAS)
Martin Luginbühl (UNIBAS)
Jérôme Jacquin (UNIL)

https://www.chord-talk-in-interaction.usi.ch

Swiss-AL: Linguistic Open Research Data Practices for Applied Sciences

The ZHAW School of Applied Linguistics is developing Swiss-AL (Swiss Applied Linguistics), a platform for language data in the applied sciences. Swiss-AL is currently funded by swissuniversities as part of the Swiss Open Research Data Grants (2023 to June 2025), supported by matching funds from ZHAW.

The Swiss-AL platform contains an extensive collection of text data in all Swiss national languages (e.g. from administrative and political websites, parliamentary debates, print and online newspapers), a linguistic processing pipeline, and a browser-based analysis workbench that allows researchers to explore data on public language use. Swiss-AL is part of the national linguistic ecosystem CLARIN-CH and has become an essential component of the national ecosystem of language resources and linguistic infrastructures. It is hosted by the ZHAW Digital Discourse Lab, where it is used in inter- and transdisciplinary research projects to analyze the use of language in public discourses (e.g. on vaccination, COVID-19, renewable energy, social welfare).

In the swissuniversities project, Swiss-AL will be further developed in order to comply with FAIR and open language data standards and to make it accessible to researchers in the applied sciences. To this end, a scientific panel from various academic disciplines (e.g. social sciences, law and architecture) is evaluating the specific disciplinary requirements for linguistic ORD. The overall goal is to integrate good practices for language-related ORD matching disciplinary research routines as manifested by the target scientific communities of Swiss-AL: applied sciences in general, applied linguistics in particular, and the CLARIN-CH and European CLARIN communities. The project promotes in an innovative way competences in the use of linguistic ORD and explores the value of such data outside linguistic disciplines. This will be achieved by implementing standards for sustainable documentation, legal issues/privacy, data dissemination and research data management according to the FAIR principles.

Principal Investigators

ZHAW Digital Discourse Lab: Julia Krasselt, Philipp Dreesen, Peter Stücheli-Herlach

https://www.zhaw.ch/en/linguistics/research/swiss-al-linguistic-open-research-data-practices-for-applied-sciences/

FAIR-FI-LD: Moving towards a national FAIR-compliant ecosystem of Federated Infrastructure for Language Data

FAIR-FI-LD is a swissuniversities ORD-funded 1-year project (July 2024- June 2025) hosted by the University of Zurich, with the participation of the Linguistic Research Infrastructure LiRI, ZHAW Zurich University of Applied Sciences (Digital Discourse Lab), Università della Svizzera italiana (Istituto di studi italiani, Istituto di argomentazione, linguistica e semiotica, E-Learning Lab) and the CLARIN-CH Coordination Office.
In the last 5-10 years, Swiss higher education institutions (HEIs) have been working on building national services for language data. They include, up to now, the Linguistic Research Infrastructure (LiRI-UZH), the Swiss-AL Platform for Applied Sciences (ZHAW), a national repository for the publication and long-term preservation of language data LaRS@SWISSUbase (UNIL, UZH), and various smaller tools and services. These units however are not all interoperable, which reduces the potential for collaboration and data reuse. In addition, fields such as interactional linguistics or second language acquisition lack adequate infrastructure.
With the foundation of the CLARIN-CH consortium in 2020 (9 HEIs and the SAGW), the HEI's efforts took a new direction: work together to build a FAIR-compliant, sustainable and expandable CLARIN-CH ecosystem of federated infrastructure to answer the needs of researchers and professionals using language data in Switzerland and beyond; an ecosystem that must be interoperable at the national and European levels. The present project aims at realizing important steps towards this mid- and long-term goal, in compliance with the Swiss ORD strategy, by prototyping

- interoperable underlying software using NLP techniques and exploratory AI techniques
- harmonized metadata between the existing Swiss infrastructure components and the European CLARIN infrastructure
- CLARIN federated content search (FCS) to query each component of the infrastructure
- a FCS multilingual landing page hosted on the CLARIN-CH website
- a frontend of the VIAN-DH@LiRI environment to visualize, query and analyze multimodal talk-in-interaction data, hosted at USI,

by producing

- documentation and training to support the use of the infrastructure and inform about legal and ethical issues related to language data in the context of Open Science,
- and by planning the future collaboration with further stakeholders and aggregation of further tools and services.

The FAIR-FI-LD project is building on three previously ORD-funded projects: Upgrading the linguistic ORD-ecosystem Up-LORD, Swiss-AL: Linguistic ORD Practices for Applied Sciences and Data-sharing skills in corpus-based research on talk-in-interaction.

Principal investigators:
Noah Bubenhofer (LiRI)
Cristina Grisot (CLARIN-CH)
Julia Krasselt (ZHAW)
Johanna Miecznikowski (USI)
https://www.liri.uzh.ch/en/projects/FAIR-FI-LD.html

Teodora Vukovic (UZH), Manuel Hendry (ZHDK):

Copyright Challenges in Creating and Publishing a CineMinds YouTube Video Corpus for Academic Research

The fields of digital humanities, media studies or interaction analysis often relies on large-scale video corpora to conduct comprehensive research. Our project, CineMinds, has amassed approximately 1700 hours from 2665 files, including YouTube videos and podcasts, featuring in-depth interviews and discussions from a range of film industry professionals. Initially intended for exploratory research within academic frameworks, the corpus has proven to be a valuable educational and academic resource, prompting us to consider its wider publication and distribution. The dataset could be useful as a public knowledgebase of film, but also for computational interaction analysis, identity recognition models, etc.
However, the primary challenge we face are the copyright policies, as most of the videos are not under the Creative Commons License, which significantly restricts their use outside personal or internal research contexts. We tried reaching out directly to the content creators (BAFTA, Directors Guild of America, etc.), which was not successful.
This presentation aims to initiate a discussion regarding strategies for overcoming these legal barriers, promoting open academic access to these valuable digital resources.

Alessia Battisti (UZH), Lisa Arter (UZH):

De-identifying persons in videos: The case of sign languages

Sign languages use manual components (hands and arms) and non-manual components (shoulders, head, and face) to produce utterances. Techniques common for de-identifying persons in videos in the context of spoken languages are not directly applicable to sign languages, as operations such as blackening or blurring parts of the face effectively remove parts of the linguistic content of a signed utterance. To overcome this, researchers have made available poses (skeletal representations) derived from videos instead of the videos themselves in the past, departing from an anecdotal assumption that poses are de-identified representations of signers. However, recent research, including our own study, has challenged this assumption. We have shown that individuals represented as pose estimates can still be identified by various features, such as upper body movements and signing fluency (Battisti et al., 2024). Hence, more recent work focuses on normalizing these poses and evaluating whether the resulting representations are indeed fully anonymous.

Reference: Battisti, A., van den Bold, E., Göhring, A., Holzknecht, F., and Ebling, S. (2024). Person identification from pose estimates in sign language. In Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J. A., Mesch, J., and Schulder, M., editors, Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, pages 209–221, Torino, Italy.

Johanna Miecznikowski (USI), Elena Battaglia (USI), Christian Geddo (USI), Nina Profazi (USI):

ShareTIGR - Sharing the TIGR corpus of spoken Italian: an ORD case study

ShareTIGR (USI, 1/2/2024 –31/1/2025) makes the TIGR corpus of spoken Italian available for scientific use, respecting data protection and FAIR principles. TIGR was collected within a specific research project (InfinIta, SNSF grant no. 192771), but was also designed to increase the diversity of resources for the study of spoken Italian. It includes 23.5h of video recordings: table conversations, food preparation, tutoring encounters, lessons and practical instruction, interviews. They were transcribed in ELAN adopting the GAT 2 conventions for fine transcription, with some adaptations. The transcripts were pseudonymized. A script-assisted workflow has been implemented to produce TXT transcripts that are optimized for the human eye and preserve a reduced amount of timecode stamps. Later we intend to create tokenized XML transcripts readable by corpus linguistic software. In A/V files we will mask faces and voices, where so required, and replace proper names by noise. For each event, we will edit a single compact, easy-to-use movie file with split screen and mixed audio. The data will be uploaded to the LaRS@SWISSUbase repository. The project team is discussing the various phases of this process as a case study of ORD practices, engaging with potentially interested communities via scientific presentations and publications and via a lab blog and social media.

Martin Luginbühl (UNIBAS), Oliver Spiess (UNIBAS):

From Collection to Publication: Managing and Sharing Children's Argumentation Data

In our data pitch, we will discuss our preparatory work for publishing a dataset of 180 peer discussions of elementary school children. The data have been gathered and annotated in the context of two SNF funded projects on oral argumentation skills.
We will shortly address theoretical and conceptual aspects of this corpus regarding representativity, metadata and authenticity and explain how these align with the relevant research objectives. Again briefly, we then discuss challenges of data gathering before elaborating in more detail the challenges of processing the data in order to publish them on a corpus repository specialized for conversational data. Next to more technical aspects (e.g. transformation into the TEI format), the processes of de-identification had to be balanced between data protection and data richness. In addition, as the repository is in Germany, different data protection practices had to be taken into account. This was e.g. important when it came to the question of whether or not to transform the children's voices – and how to do that irreversibly.
In a brief outlook, we will present an interactive website including visualizations of the aspects annotated in our data that allows certain analyses of the data set without data protection issues.

Jennifer Thorburn (UNIL):

Ethical sharing of legacy data

Data sharing is an ever-growing priority for scholars and funding agencies, but it one of the most delicate elements of the research data life cycle, particularly in the context of legacy data (Bossaller & Million 2023). In this case, the data in question are sociolinguistic interviews conducted in the early 1980s in St. John's, Canada, which I am digitizing as part ofa longitudinal study. The original researcher wishes me to make them publicly available, but the issues are (at minimum) two-fold. First, while participants may have consented to their data being used in future research, data sharing as it is understood today was not a possibility they would have considered. Re-contact is impossible so re-negotiating consent retroactively (cf. Corti et al. 2000, Kuula 2010/2011) cannot be done. As such, is it ethical to share these interviews, knowing that "[t]he protection of the individual takes priority over the scientific interests of society" (swissethics, citing HRA Art. 1)? If so, then de-identification becomes the next hurdle. Due to the original project's research design, there are significant issues of reidentification through linkage (Eliot et al. 2020), adding an extra dimension of complexity to theprocess of de-identifying the materials while maintaining their useability in accordance with FAIR data principles.

References

Bossaller, J., & A. J. Million. 2023. The research data life cycle, legacy data, and dilemmas in research data management. *Journal of the Association for Information Science and Technology* 74.6, 701-706. https://doi.org/10.1002/asi.24645

Corti, L., A. Day, & G. Backhouse. 2000. Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum: Qualitative Social Research*, 1(3). https://doi.org/10.17169/fqs-1.3.1024

Eliot, M., E. Mackey, & K. O'Hara. 2020. *The Anonymization Decision-Making Framework: European Practitioners' Guide*. Manchester: UKAN.

Kuula, A. 2010/2011. Methodological and ethical dilemmas of archiving qualitative data. *IASSIST Quarterly* 34.3-4, 12-17.

Samuel Felder (UNIFR):

Data privacy and anonymization in an L1 learner text project

In the research project “QuaTexD: Qualität von Deutschschweizer Lernertexten” at the University of Fribourg, we are currently compiling a corpus of argumentative texts from high school and university students. Along with the texts, various demographic aspects and information about the language use of the subjects are collected through questionnaires. The data collection among thehigh school students is already well advanced, with care taken during the collection to ensure that the names of the participating students are unknown to the research team. Thus, the data is largely anonymized. However, the data is not completely anonymous in all cases, as it is sometimes possible to infer the identity of the writers based on the information in the questionnaires, for example, if they provide their parents’ workplace when asked about their parents’ profession. In rare cases, the subjects may also reveal information in their argumentative texts that could enable identification. Against this background, we face various questions regarding further data processing: To what extent do we need to consider the data as personalized data and take appropriate security measures? What does this mean for handling the data, as long as only project members and student assistants are working with the data internally? How do we best ensure the complete anonymization of the data so that it can be made publicly accessible later, as planned in the project? We look forward to discussing these questions at the CLARIN-CH Day with the experts present.

Joanna Blochowiak (UCLouvain), Christina Grisot (UZH):

How subjectivity and individual differences shape argument reception: a cross-country project

We present a project proposal submitted to SNSF' MAPS Funding Scheme MAPS – Multilateral Academic Projects – which supports collaborative research projects between researchers in Switzerland and their colleagues in Bulgaria, Croatia, Hungary, Poland and Romania. The project which aims to assess how subjectivity impacts the reception of arguments, i.e. their acceptance or rejection, depending on language users' individual differences in political discourse across three different countries: Switzerland, Romania, and Poland. Working with data and researchers from multiple countries makes adhering to data protection laws, as well as data management more difficult than when a single country is involved.

In this pitch, we will focus on the corpus data and the experimental data collection, which present challenges regarding copyrights, personal and sensitive data protection, compliance to both GDPR and Swiss law on data protection, as IPR shared between multiple countries/institutions.

As corpus data, we plan to collect parliamentary debates and political blogs data, from each the three countries represented in the project. Political blogs are only partly published with CC. licenses, so we will have to manage copyrights. We plan to ask permission or to buy the copyright-protected data.

We will carry out a crowdsourcing experiment to account for individual differences, i.e. cognitive profiles and socio-demographic data. To do so, we will recruit 200 adult participants from each country platforms such as Prolific and social media. Three questionnaires measuring personal individual profiles will be run: (i) a questionnaire to collect socio-demographic characteristics like gender, age and education; (ii) the Autism Spectrum Quotient (AQ); (iii) the Mini-IPIP questionnaire to assess personality traits: extraversion, agreeableness, consciousness, neuroticism and openness. All participants will sign an informed consent form, which explains the purpose of the research, personal data protection (anonymization, data security) and data sharing for research, education, and training purposes only. This point is particularly challenging with respect to being able to collect enough data.

For such a collaborative project, the challenges regarding data storage, preservation and sharing will be solved by making use of the national CLARIN repositories, which adhere to interoperable data and metadata formats, as well as data management best practices.

Anita Auer (UNIL), Rajamathangi Shanmugasundaram (UNIL):

Challenges of data protection and storage in a Swiss-Indian research project

The research project "Mapping Heritage Language Structure through Sociolinguistic Cues: A Case Study of Swiss Tamil" (2023-2026), which was funded as part of the Indo-Swiss Joint Research Programme, aims at testing the so-called simplification hypothesis that claims that the linguistic structure of so-called heritage languages is simpler and less complex than that of the same language spoken in the homeland of migrant communities. To do so, the project collects and compares data from second-generation Swiss Tamil speakers in the German- and French-speaking parts of Switzerland, as well as data collected in the homeland, i.e. South India and Northern Sri Lanka. The project collaboration with the Indian Institute of Technology in Jodhpur and the fact that data is collected in India and Sri Lanka, in addition to Switzerland, has led to some challenges from the point of view of data protection, data storage, and data sharing. In our talk, we will describe the challenges and the solutions found by the involved universities.

Noémie Treichel (UNIFR):

Experimental research with individuals with a rare genetic condition: ethical and legal considerations

This presentation will address a few ethical and legal challenges associated with conducting experimental research on individuals with Williams syndrome (WS), a rare genetic disorder. First, the rarity of WS poses significant difficulties in recruiting a sufficient number of participants within Switzerland. However, variations in data protection laws across different countries further complicates the possibility of recruiting participants internationally. Second, the open-access (OA) sharing of data is particularly sensitive and often restricted when research involves individuals with specific developmental conditions. This presentation aims to open up questions on data protection and data sharing when doing research with individuals with particular (medical) conditions and when doing research internationally.

Ignacio Pérez Prat (Lia Rumantscha):

Challenges with intellectual property: The case of the Lia Rumantscha

The Lia Rumantscha, established in 1919, is a non-profit organization dedicated to promoting and supporting the Romansh language and culture in Switzerland. It coordinates regional initiatives and fulfills a public contract to enhance the preservation and development of Romansh in various cultural and educational contexts.

"We have different challenges with the usage of intellectual property in the field of corpora and currently dictionaries. Especially in the last point we have a specific challenge to handle the data which is has an intellectual property wall, so it can't be used freely but has to be maintained by us. This represent a practical and a legal challenge"

Nathalie Dherbey Chapuis (UNIFR):

Challenges in Utilizing Automatic Speech Recognition for a Corpus of Young L2 French Learners (FrOMi)

This presentation aims to share the difficulties encountered when automatically transcribing the corpus FrOMi of L2 oral productions. The context of the study involves children entering compulsory school in a French-speaking schools, where many pupils have an immigrant background (70-80%). In this study, the participants with diverse linguistic backgrounds begin to learn L2 French when entering school at 5 years old. These participants constitute a population that is largely understudied in L2 French acquisition.
This study is based on a 3-year longitudinal corpus of spontaneous oral productions of twelve L2 learners of French who were recorded at three-month intervals (11 collections). The oral productions were recorded during a morning at school (~3hours).
Several difficulties were encountered for automatic transcription of their speech. First, automatic diarization based on F0 measures was complicated because speaker identity is frequently confounded between children, and because recording background is always noisy. Second, after doing a manual diarization, we've tried to obtain an automatic transcription with the help of several tools (Microsoft, Whisper...) but the rate of accuracy (~50%) obliged us to transcribe manually. Third, words were frequently mispronounced (30%) or code-switched (up to 40%).
We are wishing to train a model of automatic speech recognition, based on the quarter of the recordings that were manually transcribed, to obtain an automatic transcription with a rate of errors inferior to 20% in the other three quarters of the corpus. However, two new difficulties are arising: First, participants are growing during the 3 years of experimentation and second, participants are improving their level of French along the 3 experimental years.

Sandrine Zufferey (UNIBE), Fabio Testa (UNIBE):

Creating a learner corpus from existing texts: ethical and methodological challenges

Corpora have proved to be an invaluable resource for research in second language acquisition. Yet, these resources are still not widely available for French as a second language for intermediate to advanced learners who do not have English as a mother tongue. I will present corpus data that has been gathered from texts produced by German-speaking learners of French during their high school years from 2000 to 2023. These texts have been provided by teachers in paper format and have been transcribed as text files amounting to over 600 000 words. The next steps will be the annotation and the distribution of this data. The main challenges that I would like to discuss concern copyright issues related to the distribution of data that were not produced to be published, as well as technical issues related to lack of complete metadata information about the participants.

Jérôme Jacquin (UNIL):

A video-recorded and time-aligned transcribed corpus associated with a database of epistemic and evidential markers of French-in-interaction

The paper aims to present the outcomes – in terms of Open Data – of an SNF-funded project about epistemic and evidential markers in French-in-interaction (Project POSEPI [100012_188924], 2020-2024, Jacquin, 2019). These markers were studied in a 28-hour video-recorded corpus documenting political debates and work meetings. Recordings were transcribed following the ICOR transcript conventions (Groupe ICOR, 2013) and using ELAN for the alignment of the transcription with the speech signal. These primary data have been made publicly accessible and requestable in CLAPI (Baldauf-Quilliatre et al., 2016; to access the POSEPI corpus, see Jacquin, Etienne, et al., 2024). Using CLAPI as an existing and dedicated infrastructure was a unique opportunity for the project since there is currently no equivalent solution in Switzerland. However, because CLAPI cannot handle annotations, it has been necessary to use anotherplatform for the database consisting of the annotated epistemic and evidential markers. This database is stored, accessible,and requestable in DaSCH (Jacquin, Keck, et al., 2024). Each annotated marker can be studied in context, by following a permalink, which opens a webpage in CLAPI and allows examining the video sequence (20s or 40s) and the time-aligned transcription in which the marker occurs.
The paper addresses the following topics selected for the CLARIN-CH Day (in order of importance)

- Topic 4: strategies for storing and sharing data by complying with Open Data principles
- Topic 3: infrastructure for handling large-scale, complex and standardized data
- Topic 2: workflow and techniques for de-identification

References

Baldauf-Quilliatre, H., Carvajal, I. C. de, Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., & Traverso, V. (2016). CLAPI, une base de données multimodale pour la parole en interaction: Apports et dilemmes. *Corpus*, *15*. http://journals.openedition.org/corpus/2991

Groupe ICOR. (2013). *Convention ICOR*. ENS de Lyon – laboratoire ICAR. http://icar.cnrs.fr/corinte/conventions-de-transcription/

Jacquin, J. (2019). *Prendre une position épistémique dans l'interaction. Les marqueurs du savoir, du non-savoir et du doute en français // Projet soumis au Fonds National Suisse de la recherche [100012_188924]*.

Jacquin, J., Etienne, C., Keck, A., Petitjean, C., Robin, C., Roh, S., & Stern, G. (2024). *Corpus POSEPI* [dataset]. CLAPI, ENS-Lyon (France). http://clapi.icar.cnrs.fr/Posepi

Jacquin, J., Keck, A., Robin, C., Roh, S., & Rivoal, M. (2024). *Marqueurs épistémiques et évidentiels du français identifiés et annotés dans un corpus d'interactions relevant de débats politiques et de réunions d'entreprise*([Projet FNS 188924, POSEPI]) [dataset]. DaSCH. https://ark.dasch.swiss/ark:/72163/1/0120

Julia Krasselt (ZHAW), Philipp Dreesen (ZHAW), Matthias Fluor (ZHAW), Klaus Rothenhäusler (ZHAW), Sooyeon Cho (ZHAW), Dolores Lemmenmeier-Batinić (ZHAW):

Challenges of Representing Large Linguistic Corpora in Linguistic Repositories and Databases

Swiss-AL is a family of linguistic corpora dedicated to the analysis of multilingual public discourse in Switzerland. In Swiss-AL we collect the following types of sources: journalistic sources (e.g. based on Swissdox@LiRi), organizational sources (based on webcrawling and scraping), parliamentary debates (based on the Official Bulletin of the Swiss Parliament) and historical sources (based on e-periodica of ETHZ). These source types result in either source type specific (i.e. homogeneous) corpora, which are usually updated regularly, or they are mixed together for discourse specific corpora, which are mostly static. The resulting heterogeneity of Swiss-AL as a corpus family poses challenges for its representation in linguistic repositories such as SWISSUbase, but also in databases such as the CLARIN Virtual Language Observatory. In the data pitch, we will elaborate on the following challenges in particular:

- adherence to (meta)data schemas predefined by the repository or database
- the development of a suitable metadata model (e.g. based on the Component Metadata Infrastructure CMDI)
- findability of individual Swiss-AL corpora (e.g., based on a specific source contained in one of the corpora)
- hierarchical relationships between individual corpora and corpus versions (e.g. as a result of regular corpus updates).

Therefore, Swiss-AL requires a flexible linguistic repository that is suitable for large linguistic corpora that are constantly updated and extended and have a complex hierarchical structure.

Elisa Squadrito (UniMC), Monica Monachini (UniMC):

Enhancing Terminological Interoperability with Linked Open Data,

The Role of H2IOSC

The growth in quantity, diversity, and complexity of language data accessible online, demands shared standards for efficient storage, connection, and exploitation. Publishing language resources such as *Linguistic Linked Open Data* (LLOD) offers a promising approach to enhancing data accessibility, reusability, and interoperability. Terminologies, which are time-consuming and costly to produce, particularly benefit from being represented in RDF, the data framework for linked data. This method enables the interoperability of terminological data and its integration with resources like lexicons and corpora. However, no widely accepted model exists for converting TBX, the XML-based standard for terminology, into RDF. SKOS, the W3C recommendation for thesauri, offers a general-purpose RDF vocabulary but lacks the complexity needed for terminological representation. Likewise, Ontolex, the W3C standard for publishing lexicons as LLOD, faces challenges due to conceptual mismatches. A proposed solution is to develop an Ontolex module for terminologies. The H2IOSC project, which aims to create a federated cluster of the Italian nodes of four European research infrastructures, including CLARIN-IT, offers an opportunity to address this discussion. Through a pilot focused on publishing and exchanging terminologies as LLOD, the project intends to advocate for Ontolex's extension, while seeking community collaboration to establish a Semantic Web standard for terminologies.

Federico Grasso Toro (UB Bern), Jonas Hässig (UNIBE):

Recommendations for Master Theses as Executable Books

The University of Bern established the Open Research Data group inside the Open Science Team, following the Swiss Universities Mandate. In collaboration with Digital Humanities an internship was coordinated for a student writing his Masters Thesis.
The focus of the internship was Open Science efforts, aimed mainly on drafting a Data Management Plan for 'Shareable Assets', in accordance with the FAIR principles and other relevant Openness metrics, i.e., transparency, replicability and reproducibility of intellectual assets. Additional Open Science efforts included drafted guidelines towards creating curated references, including persistent identifiers (i.e., DOI's, ROR's, RRiD's etc.), as well as data repositories submissions and even workflows, tools, services and training materials, following DARIAH's SSH Open Marketplace.
The overall goal was to allow students to create Master Theses as Executable Books with minimum adoption resilience.
The resulting recommendations lead to a conduct of science in such a way that MA students can start their researchers' careers aiming towards collaborative and contributing efforts, where processes, notes and even resulting research linguistic data are available under properly defined licenses. Tools include literature-organising Zotero, open source scientific-computing environments JupyterLab, scientific publishing infrastructure and writing tools curvenote and open source software development community GitHub.
Finally, we suggest periodical revisions to these recommendations to facilitate future MA students to face challenges like sensitive data management, while redefining responsibilities for researchers, institutions and funders.

François Delafontaine (UNINE):

From tiers to TEI

The reference format for text encoding in the Digital Humanities is TEI (ISO 24624), which has a semantic, nested structure. The OFROM+, DoReCo and other corpora use Praat or Elan with document structures that are neutral and tier-based. A universal conversion procedure can't find semantic clues and often the data itself does not easily fit into the standard categories: OFROM+ has two types of PoS and DoReCo files have tier and segment metadata. The solution found for automatic conversion (such as with TEI-Corpo) has been to use existing TEI tags (spanGrp, note) to reassert a tier-based structure. This only leaves a fully semantic conversion as well as the integration of the standard format into existing annotation tools for its adoption in oral linguistics.

Johanna Miecznikowski-Fuenfschilling (USI), Sara Greco (USI), Andrea Rocci (USI), Nina Profazi (USI), Lorenza Mondada (UNIBAS), Martin Luginbühl (UNIBAS), Jérôme Jacquin (UNIL), Simona Pekarek Doehler (UNINE) :

Data-sharing skills in corpus-based research on talk-in-interaction (CHORD-Talk-in-interaction)

The project (April 2023-September 2024) is funded by swissuniversities, USI, Unibas, Unil and Unine. It explores current socio-technological practices of sharing and reusing audio-video recorded and transcribed corpora of naturally occurring spoken interaction and the possible future development of such practices – with a focus on their epistemological implications, the skills researchers need to develop, and the Swiss situation. To this end, the team organised five workshops with experts, designed a survey on data-sharing practices in Switzerland and reviewed relevant literature. The explored practices, besides their advantages (efficiency, comparative/diachronic investigations, quantitative research), also raise several challenges that open data policies should take into account: i) Spoken corpus data are technically complex (several modalities and data formats) and data owner may lack skills and resources to process them for sharing. (ii) Switzerland still lacks digital infrastructure to host this kind of data as FAIR data. (iii) To make data FAIR while respecting privacy, data owners need to balance permissions granted by declarations of informed consent, de-identification techniques, and access restrictions. (iv) Data pertaining to various genres/settings/groups are not all equally easy to share widely, independently of their scientific interest. (v) Reusing data implies skipping fieldwork and transcription tasks that are potentially relevant stages of qualitative data interpretation.

Chiara Jermini-Martinez Soria (USI):

The 'RefraMe Swiss corpus': dispute mediation sessions

I would like to present a dataset called “Swiss Mediation Corpus”, composed of transcriptions of mediation sessions of interpersonal conflicts of various kinds in French and Italian. All cases have been collected in Switzerland in collaboration with professional dispute mediators in the context of two different research projects, between 2017 and 2023. The video-recordings have been transcribed using transcriptions’ conventions that allow for a (partial) analysis of extra linguistic elements as well (Traverso 1999). This corpus has the potential to be useful for carrying out various types of analysis in different research fields, however there are relevant ethical issues to be addressed, in order to respect participant’s confidentiality.

Anita Auer (UNIL):

Copyright issues linked to Late Modern English pauper letters

The SNSF-funded research project "The Language of the Labouring Poor in Late Modern England" (2020-2025) investigates the language of the lower social class based on a collection of c. 2000 pauper petitions that were written during the period 1795-1834 within the context of the Old Poor Law. At the time, people who were in distress could write letters to their home parish and apply for out-relief, which was often given in the form of money. These letters were thus sent to parishes all over England and are today mostly stored in different county archives, which vary greatly with regard to their regulations, particularly linked to copyright. While some archives allow us to publish facsimiles of the letters for free, other archives are much stricter with the data; in some cases, we had to contact the parishes directly. In this presentation, the challenges linked to collecting this type of historical data will be presented, particularly with regard to copyright issues and thus also storing and sharing of the data.