CLARIN-CH Day 2026
Booklet of Abstracts

SESSION 1. AI AND LANGUAGE TECHNOLOGIES 3

Integrating Generative AI into Corpus-Assisted Analysis of Public Discourse 3

Whose dialect, whose variety? AI transcription as a hidden technology of language planning 3

Towards a Fine-Grained Persian Dataset for Sexualized and Gender-Based Hostile Language Detection 4

SESSION 2. CORPORA AND DISCOURSE ANALYSIS 5

Building a Corpus of Russian Military Telegram Blogs: Methodology, Challenges, and Ethical Considerations 5

A multi-level approach to annotating deniability in political and legal discourse 6

Digital stancetaking towards gender diversity on the platform gutefrage.net 6

SESSION 3. OPEN RESEARCH DATA & FAIR 7

Metamorphosis of a DMP: Navigating Consent and De-identification Across Three Stages of an MA Project 7

Language Data Across Disciplines: Survey Insights on ORD Practices, FAIR Awareness, and Training Needs 8

From Infrastructure to Impact: Understanding Adoption in European Digital Research Infrastructures 9

SESSION 4. LANGUAGE RESOURCES AND INFRASTRUCTURES 9

Making the Ephemeral Durable: The Swiss Sticker Corpus (SSC) as Open Research Data 9

VotingBooklets: A Multilingual Parallel Corpus of Swiss Federal Referendum Texts 10

Database-first lexical infrastructure for Proto-Albanian research: interoperability, FAIR data, and reusable workflows for Albanian–Romanian cognacy and the Proto-Albanian–Thracian substrate segment 10

SESSION 5. MULTILINGUALISM & LINGUISTIC DIVERSITY 11

The importance of multilingual and cross-linguistic corpora for the study of language contact 11

Linguistic data structures and standardization are not theory-neutral 12

At-Issueness Sensitivity in L2 Misinformation Detection: The Role of Cognitive Load 13

SHOWCASES & DEMOS 14

AViRI – Towards a Shared Infrastructure for Audiovisual Data and Metadata 14

Audio-video data processing pipelines 14

MAVA and DAVA: building a cluster of infrastructure for processing multimodal audio-visual data 14

POSTERS 14

RefCo2 from the fields to the archive: multi-perspective testing of a corpus-quality platform for reusable language data 14

The “Person Marking in South-Central Trans-Himalayan” (PMST) database 15

Infrastructural support for data analysis and reuse in Interactional Linguistics: potential of LCP-Videoscope 16

Beyond Clinical Pain Descriptors: Detecting Figurative Pain Descriptions in Patient Narratives 17

FAIRifying a Multilingual European Parliament Corpus for the Study of Gender-Inclusive Lexical Innovation 18

CLARIN – a pan-European distributed digital infrastructure for language data and technology 19

Two communities, one shared practice: presentation of CLARIN-CH RDM Working Group and SRDSN L-RDM Node 19

LiRI and the Lia Rumantscha work together to build new infrastructure and collect data to promote Romansh 19

LiRI and Linguist List 20

Swissdox RAG: Retrieval-First AI for Large Multilingual News Archives 20

Session 1. AI and Language Technologies

Integrating Generative AI into Corpus-Assisted Analysis of Public Discourse

Dolores Lemmenmeier (ZHAW)

Although corpus-linguistic methods have played a central role in text and discourse linguistics over the past decades, the widespread adoption of large language models (LLMs) has led to what some scholars describe as a possible stagnation of traditional machine-assisted text analysis (cf. Bubenhofer & Scharloth, 2025: 96). LLMs have fundamentally changed the way researchers search for information and have raised expectations regarding corpus platforms. Against this background, traditional corpus tools such as collocations, keywords, and n-grams increasingly appear outdated. At the same time, many corpus platforms remain hesitant to integrate LLMs, despite the often convincing quality of their results (cf. Davies, 2025).

This contribution presents the GAIS project (Generative AI for Swiss Public Discourse Analysis, 2026–2028), which investigates how generative AI can be meaningfully integrated into corpus-linguistic workflows for discourse analysis. The project builds on the Swiss-AL corpus (Krasselt et al., 2026), which includes parliamentary debates, news media, and press releases. Its goal is to develop an AI component that complements existing corpus-linguistic methods while enabling new forms of corpus exploration.

The project consists of three main steps:

1. Identifying key research needs through collecting concrete use cases;
2. Developing and integrating new functionalities, including AI-enhanced versions of existing corpus tools (e.g. KWIC and collocations), semantic search methods such as Retrieval-Augmented Generation (RAG; Lewis et al., 2020) as well as AI-assisted text categorisation (cf. Huang & He, 2025).
3. Evaluating the developed approaches and deriving practical guidelines for the use of generative AI in corpus-assisted discourse analysis. In this contribution, I present key methodological decisions and preliminary approaches to integrating AI-assisted binary classification and LLM-guided text clustering (Huang & He, 2025) into corpus-assisted discourse analysis.

References

Bubenhofer, N., & Scharloth, J. (2025). Textanalyse im Data-driven Turn: Von den Anfängen bis zum wahrscheinlichen Ende der datengeleiteten Korpuslinguistik. OBST, 104, 85–108. https://doi.org/10.17192/OBST.2025.104.8807

Davies, M. (2025). Comparing the predictions of large language models to actual corpus data (White Paper). English-Corpora.org. https://www.english-corpora.org/ai-llms/

Huang, C. & He, G. (2025): Text Clustering as Classification with LLMs. Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 374-384.

Krasselt, J., Dreesen, P., Stücheli-Herlach, P., Lemmenmeier-Batinić, D., Geckeler, S., Rothenhäusler, K., & Fluor, M. (2026). Swiss-AL: A platform for language data in the applied sciences. Zeitschrift für germanistische Linguistik, 54(1), 210–229. https://doi.org/10.1515/zgl-2026-2008

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle et al. (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). Curran Associates.

Whose dialect, whose variety? AI transcription as a hidden technology of language planning

Francine González Borrell and Yvette Bürki (UNIBE)

Within Yvette Bürki's research project on New Fatherhoods, we work with multilingual audio recordings, most of which combine Spanish varieties, Swiss German dialects, and Standard German. For the analysis of, for example, child agency, identities, language ideologies and production, it is quintessential that the transcriptions show which Swiss German dialect is spoken, which Spanish variety, and when translanguaging towards Standard German occurs. As

some words are written identically in Swiss German dialects and Standard German, this has to be marked while listening.

This is where we encounter AI's shortcomings. Many AI transcription tools offer the option of transcribing into only one language. If we opt for Spanish, everything in German and Swiss German turns into gibberish. Sonix offers the possibility of choosing two languages at a time, but Swiss German is not an option. Moreover, Sonix may recognize speaker turns correctly in a structured conversation such as an interview, but in spontaneous family conversations speaker turns are mostly identified incorrectly. Töggl, designed specifically for Swiss German, transcribes by translating into Standard German and is therefore unusable for our purposes. Whisper supports Swiss German and segments spontaneous turns more reliably than Sonix, but cannot transcribe two languages at once.

The most consequential limitation, however, is that these tools do not render the actual dialect or variety spoken. AI either defaults to one variety or generates a 'neutral', artificial one, thereby erasing the very dialectal and varietal diversity our analysis depends on. In doing so, AI transcription imposes norms and standards from outside the speech community: it operates as a hidden technology of language planning and policy, a non-state actor whose models drive linguistic homogenization and flatten the diversity they claim to capture.

Additional note: This abstract is based on the previously mentioned project led by Yvette Bürki, in which I am participating as a collaborator. I will be the one giving the presentation.

Towards a Fine-Grained Persian Dataset for Sexualized and Gender-Based Hostile Language Detection

Hanieh Habibi, Amir H. Payberah and Davide Picca (UNIL)

This project focuses on the development and analysis of an annotated dataset for detecting sexual harassment and related forms of gender-based abusive language in Persian (Farsi) social media data. Despite growing interest in abusive language detection, Persian remains a low-resource language with a notable lack of publicly available, fine-grained datasets that support structured analysis of gendered and identity-based harassment. This limitation constrains both NLP research and the development of robust, context-sensitive models for abusive language understanding in multilingual settings.

The dataset is constructed from posts collected on X (formerly Twitter) using targeted Persian-language queries focused on a period of large-scale political unrest and state–society conflict in Iran during 2025–2026, widely reported to involve mass protests, severe repression, and extended internet shutdowns. In this context of disrupted communication and information scarcity, social media became a key space for expressing political anger, collective trauma, and social tensions through highly polarized discourse, often intersecting with gendered and identity-based language. In addition to original posts, reply threads are included where available, enabling the study of harassment in its conversational context.

We developed a structured preprocessing pipeline to clean, normalize, and organize the raw X data into a structured dataset for subsequent analysis. The resulting dataset is designed to support reproducible research practices and align with FAIR (Findable, Accessible, Interoperable, Reusable) data principles by enabling reuse, interoperability, and transparent methodological documentation.

We developed a multi-level annotation framework to characterize sexualized and gender-based hostile language in Persian social media. At Level 1, each tweet is classified according to the type of hostility (e.g., sexist, misogynistic, and/or homophobic). Level 2 identifies whether the tweet

contains targeted or non-targeted hostility. For targeted instances, Level 3 specifies the target identity, distinguishing between individual and group targets, and identifies the target category (e.g., political leaders, military figures, journalists, women activists, LGBTQ+ communities, ethnic or religious groups, political organizations, or national communities). Level 4 captures the communicative function of the hostile expression, such as humiliation, delegitimization, sexual objectification, questioning masculinity or morality, dehumanization, victim blaming, or intimidation. Finally, Level 5 distinguishes between explicit and implicit forms of sexualized hostility. These complementary annotation dimensions enable fine-grained analysis of both the semantic properties and pragmatic functions of abusive language in Persian social media.

The resulting dataset is intended as a benchmark resource for Persian NLP, supporting transformer-based models for abusive language detection in low-resource settings and contributing to ethical NLP, corpus linguistics, and multilingual resource development within the ORD(Open Research Data) paradigm.

Session 2. Corpora and Discourse Analysis

Building a Corpus of Russian Military Telegram Blogs: Methodology, Challenges, and Ethical Considerations

Olga Bikkulova (UNIBE)

Russian military bloggers emerged as a major influence on public perception of the Russian–Ukrainian war, particularly after February 2022. They explicitly position themselves as "information warriors," producing a distinctive hybrid discourse that combines reporting, ideological positioning, and audience construction. Studying this phenomenon through the lens of Corpus-Assisted Discourse Studies (CADS) requires building a corpus from the ground up, and this paper reflects on what that process actually involves.

The study draws on three corpora compiled manually for a PhD dissertation at the University of Bern. The primary corpus consists of ten high-subscriber Russian military Telegram blogs (ranging from 605,000 to 3.1 million subscribers), covering the period 2017–2025 and totalling approximately 6.5 million words. A media corpus of ten pro-government Russian Telegram news channels (4.3 million words) serves as a genre comparison point, and the RuTenTen17 web corpus (c. 9 billion words) functions as a general-language reference, enabling comparative analysis using Sketch Engine.

The paper addresses the practical challenges of corpus construction from Telegram data. While the platform permits channel archive export in plain text, corpus size limitations and Sketch Engine subscription constraints required the development of a randomised diachronic sampling methodology rather than full-archive upload. Equal proportions of posts were selected across the 2017–2025 timeline per channel, with metadata tagging for year, channel identity, and pre-/post-February 2022 period, enabling both synchronic keyness analysis and controlled diachronic comparison. Multimodal content was deliberately excluded.

The paper also addresses ethical questions specific to politically sensitive data: the status of publicly accessible content from channels legally classified as media under Russian law; informed consent in a context where anonymisation is neither standard nor straightforward; and the challenge of maintaining scholarly neutrality when analysing discourse that is itself engaged in legitimising an active armed conflict. The tension between FAIR data principles, which favour openness and reusability, and the ethical constraints on sharing politically sensitive material

from an active conflict zone represents an unresolved challenge for digital language research infrastructure.

A multi-level approach to annotating deniability in political and legal discourse

Bruna Paz Schmid, Steve Oswald and Lou Odermatt (UNIFR)

Consider the following responses (made under oath) during questioning in two related committee hearings:

(1) Mr. Craig Kielburger: If I can start by clarifying, in fact it actually wasn't an administration fee; it was a program implementation fee. (...) The costs here were simply directly for reimbursements on the delivery of the program.

So to answer your question precisely, for the first 20,000 students, WE Charity has provided $14.5 million for all-inclusive aspects.1

(2) Mr. Craig Kielburger: Your statement is not correct. What we said [in the 2020 hearing] was that every program has administration, because it requires administration [including the rental costs of buildings owned by the Kielburgers, co-founders of WE Charity] to deliver it.2

The issue here is whether it is plausible for Kielburger to deny in (2) having excluded the budgeting of administration fees in (1). These statements were made in the context of an investigation into questions of conflict of interest and lobbying in relation to pandemic spending. It involved the implementation of a student grant program and the relationship between the co-founders of WE Charity and the Trudeau government. Accordingly, the stakes in these hearings were high which explains their being conducted under oath – lying under oath may result in a charge of perjury, a criminal offense.

Our SNSF-funded project investigates the deniability of such statements as in the example above in terms of their linguistic, rhetorical and argumentative features. The goal of the project's corpus strand is to produce a dataset of annotated denials to study deniability and deliver a holistic account of this phenomenon as it occurs in natural discourse in sources from the legal and political contexts.

The talk will report on the annotation process of the project's corpus strand and more specifically on how we are adapting Inference Anchoring Theory (IAT)3 - a framework from argument mining that is implemented as an online annotation software4 - to capture not only the argumentative and dialogical structures of denials but also the implications of statements such as (1) and (2) and how commitments are managed in such exchanges. The talk will also address the challenges in annotating these implicit components of denials with the tool's current affordances.

References

Canada, House of Commons, Standing Committee on Finance (FINA), Evidence, 43rd Parl., 1st sess., No. 45 (July 28, 2020), testimony of Craig Kielburger.

Canada, House of Commons, Standing Committee on Access to Information, Privacy and Ethics (ETHI), Evidence, 43rd Parl., 2nd sess., No. 24 (March 15, 2021), testimony of Craig Kielburger.

Budzynska, K., Janier, M., Reed, C., Saint-Dizier, P. 2016. "Theoretical foundations for illocutionary structure parsing." Argument & Computation, 7(1), 91-108. IOS Press.

Janier, M., Lawrence, J., Reed, C. 2014. "OVA+: An argument analysis interface." In Computational Models of Argument: Proceedings of COMMA, Vol. 266, No. 2014, p. 463.

Digital stancetaking towards gender diversity on the platform gutefrage.net

Aline Siegenthaler (UNIFR)

In German-speaking countries, there has been an observable shift in the nature of discourse surrounding gender diversity, with an increasing tendency to take place in digital public spheres. This shift is characterised by a rise in the level of engagement and intensity of these discussions. While linguistic research has thus far primarily examined (anti-)gender discourses in the public

sphere and how they are reflected in discussions about gender-inclusive language (Hergenhan 2020, Coady 2024), there has been less research into how stances towards gender diversity (trans, non-binary, etc.) are produced in digital spheres. This paper therefore examines digital stancetaking practices related to gender diversity, which reveal underlying gender ideologies and are interactively reinforced.

The data set consists of questions, answers, and subsequent comments on the gutefrage.net platform, selected using the platform-specific topic tag “LGBT+” (data collected by Doninic Schmitz, HHU Düsseldorf). This restriction is consistent with the platform's logic and is intended to yield relevant discussions without asserting a claim to representativeness. Methodologically, the study employs a research design that integrates quantitative and qualitative analyses. In the first step, recurring linguistic patterns relative to relevant keywords are automatically identified using collocation, n-grams, and cluster analyses through CorpusExplorer (Rüdiger 2022). In the second step, selected thread sequences are analysed qualitatively in order to reconstruct stancetaking practices in their dialogic context. This analysis is based on the previously identified patterns.

In theoretical terms, this study aligns with research on positioning and stancetaking (see Spitzmüller 2023; Merten 2025). The emphasis is placed on utterances in which stances regarding the socially relevant topic of gender diversity become apparent. These utterances are interpreted as indicators of gender ideologies that emerge in digital comment sections. This paper thus establishes an empirical foundation for further research on gender-related ideologies in the digital sphere.

References

Coady, Ann. 2024. The gender-inclusive language debate in France: A battle to save the soul of the nation? In Falco Pfalzgraf (ed.), Public Attitudes Towards Gender-Inclusive Language. A Multilingual Perspective, 45-72. Berlin, Boston: De Gruyter.

Hergenhan, Jutta. 2020. Langage non sexiste et antiféminisme en Allemagne. In Cahiers du Genre 2(69), 85-107.

Merten, Marie-Luis. 2025. Social Positions – Social Constructions: Stance-Taking in Online Commenting (Formulaic Language Series 7). Berlin, Boston: De Gruyter.

Rüdiger, Jan-Oliver. 2022. CorpusExplorer – Eine Software zur korpuspragmatischen Analyse. University thesis. University of Kassel.

Spitzmüller, Jürgen. 2023. Metapragmatic Positioning: Reflexive Situating Between Interaction and Ideology. In Mark Dang-Anh (ed.), Political Positioning: Linguistic and Social Practices (Akademiekonferenzen, Vol. 33), 39–66. Heidelberg: Universitätsverlag Winter

Session 3. Open Research Data & FAIR

Metamorphosis of a DMP: Navigating Consent and De-identification Across Three Stages of an MA Project

Oscar Jordan (UNIL)

his manually collected corpus of social media posts encountered significant data management challenges through different iterations of the project. One year of posts to two public Telegram channels by Russian influencers, specifically messages featuring anglicisms, were initially collected for an MA project analysing the morphosyntax of content extracted from the dataset. Bloggers, contacted prior to data collection, provided informed consent (D’Arcy and Bender, 2023) developed solely for curricular purpouses. They therefore consented to a de-identified project, in which the annotated data would remain accessible only to the student and supervisors. However, analysis of original terms, some associated with specific (potentially sensitive) communities and used in creative manners also raised questions about de-identification and reasonable expectations of publicity, considering the accessibility of the channels and the participants’ influencer status (NESH, 2019). Beyond the specificity of certain items and the participants’ morphological creativity, occasional quotations were used to identify syntactic functions, further complicating questions of de-identification (Fielser and Proferes,

2018; Mackenzie, 2017). In other words, the research process raised microethical concerns specific to the study's context (Dalmaijer et al., 2024), namely, the nature of the data and the informants.

When the researcher sought to disseminate the results of the MA project and further explore the dataset, an updated, retroactive consent became necessary, not only regarding the due to changes in data storage and publicity of the research. Thus, a second data management plan was created with these issues in mind. Finally, a third data management plan was developed—from this sample data—to expand the project and render it statistically significant: a larger and more diverse corpus (participants and languages). Each iteration's specificities resulted in a three-step metamorphosis in data management planning, illustrating the learning process regarding issues ranging from data storage to microethics, and input from bodies in Swiss academia.

References

Dalmaijer, E., Stommel, W.J.P., Pas, B.R., and Spooren, W.P.M.S. 2025. Ethical challenges in collecting pre-existing digital data for linguistic research. Linguistics, 63.2, 407–27.

D'Arcy, A. and Bender, E. 2023. Ethics in linguistics. Annual Review of Linguistics, 9, 49–69.

Fiesler, C. and Proferes, N. 2018. "Participant" perceptions of Twitter research ethics.

Social Media + Society, 4.1. DOI: doi.org/10.1177/2056305118763.

Mackenzie, J. 2017. Identifying informational norms in Mumsnet talk: A reflexive-linguistic approach to internet research ethics. Applied Linguistics Review, 2.3, 293–314.

NESH [Den nasjonale forskningsetiske komite for samfunnsvitenskap og humaniora]. 2019. A Guide to Internet Research Ethics. Available at https://www.forskningsetikk.no/en/guidelines/social-sciences-and-humanities/a-guide-to-internet-research-ethics/ (accessed 26.06.2026).

Language Data Across Disciplines: Survey Insights on ORD Practices, FAIR Awareness, and Training Needs

Julia Krasselt (ZHAW), Elsa Liste Lamas (ZHAW), Maike Fischer (ZHAW), Cristina Grisot (UZH), Joanna Blochowiak (UZH)

The adoption of Open Research Data (ORD) practices and the FAIR principles has become an important objective across the Swiss research landscape. However, researchers working with language data differ considerably in their experience, competencies, and support needs. Within the framework of the Language Data Across Disciplines (LaDaD) project, embedded in the Swiss CLARIN-CH infrastructure, we conducted a national survey to assess current data management practices, FAIR awareness, and training needs among researchers working with language-related data. At the CLARIN-CH day 2026, we will present the results of the survey and the insights gained from this regarding the development of training materials.

The survey was completed by 51 researchers from linguistics and a range of other disciplines, including education, social sciences, psychology, medicine, theology, history, translation studies, and computer science. This paper presents a comparative analysis of responses from linguists (n = 27) and researchers from other disciplines (n = 24), focusing on their preparedness to implement FAIR and ORD principles.

The results reveal substantial differences between the two groups. Linguists generally report higher familiarity with FAIR principles and greater experience in preparing and publishing datasets. Almost half of the linguists had shared datasets publicly, compared to roughly one third of researchers from other disciplines. Likewise, linguists expressed greater confidence in applying FAIR principles, while researchers from other fields more frequently reported limited familiarity and low confidence.

Despite these differences, both groups face common challenges. Legal and ethical requirements, particularly regarding data protection, informed consent, copyright, and data sharing, remain major concerns. Respondents also report difficulties with repository selection, metadata creation, long-term preservation, and the considerable time investment required to

prepare data for reuse. Across disciplines, there is strong demand for practical guidance on FAIR implementation, management of sensitive data, and legal and ethical aspects of language data stewardship.

By identifying both existing competencies and persistent gaps, the survey provides an evidence base for the development of discipline-sensitive training materials, support services, and community-building initiatives. The findings demonstrate that while linguistics may serve as a valuable source of expertise and best practices, effective capacity-building for ORD and FAIR implementation requires tailored approaches that address the specific needs of different research communities working with language data.

From Infrastructure to Impact: Understanding Adoption in European Digital Research Infrastructures

Beliz Gökmen, Andrew A. Clark, Gorka Fraga Gonzalez and Steven Moran (UZH)

European research infrastructures have invested heavily in repositories, interoperability frameworks, metadata standards, authentication systems, and FAIR data creation. While these investments have succeeded in creating increasingly mature technical ecosystems, evidence suggests that technical availability alone does not guarantee researcher adoption, workflow integration, or demonstrable scientific impact. Across infrastructures including CLARIN, DARIAH, EOSC, and broader FAIR implementation efforts, recurring challenges increasingly concern visibility, discoverability, community engagement, and impact assessment. This paper discusses literature across research infrastructure studies and technology adoption research and proposes a conceptual framework for understanding infrastructure uptake. We argue that digital research infrastructures may be entering a post-engineering phase in which the principal bottlenecks increasingly concern social embedding and research practice rather than technical implementation alone. We further argue that infrastructure evaluation must distinguish technical capacity, institutional prominence, practical adoption, workflow embedding, and scientific impact.

Session 4. Language Resources and Infrastructures

Making the Ephemeral Durable: The Swiss Sticker Corpus (SSC) as Open Research Data

Yvette Buerki, Kellie Gonçalves and Janis Schneider (UNIBE)

Stickers are ephemeral, mobile and categorised “wall” signs (Kallen 2010) whose functions are extremely diverse. Used since the 1960s primarily for commercial purposes (Reershemius & Ziegler 2025), they have become a ubiquitous sign type found in a wide range of places and contexts, including car bumpers (Bloch 2000; Hafez 2020), lampposts (Reershemius 2019) and, more recently, laptop covers (Bock & Busch 2025). Responding to the growing academic interest in this text type, and as part of the project “Stickers: diverse functions and changing geographical emplacement at the ‘online/offline nexus’ – A trans-university and trans-disciplinary dialogue”, funded by the Walter Benjamin Kolleg at the University of Bern, we have developed the Swiss Sticker Corpus (SSC).

The SSC is an openly accessible database designed for use by students, researchers and the wider interested public. Built with the open-source platform Omeka and published as a public website, it provides a corpus view filterable by taxonomies, enabling users to compare stickers according to discourse types, multilingual usage and geosemiotic criteria (Scollon & Scollon 2003). An interactive map displays the locations where the stickers were recorded, and a

contribution form allows users to upload and categorise their own sticker photographs. This crowdsourcing functionality is designed to ensure the continuous expansion and long-term sustainability of the resource through public participation.

In our contribution, we introduce the SSC, present the methodological decisions underlying its taxonomies and metadata, and reflect on the challenges of documenting ephemeral and often anonymous signs in accordance with FAIR and Open Research Data principles. The SSC thereby offers itself as an open database for future projects drawing on stickers as empirical data.

VotingBooklets: A Multilingual Parallel Corpus of Swiss Federal Referendum Texts

Elina Stüssi and Jannis Vamvas (UZH)

Swiss federal voting booklets constitute a unique multilingual resource, presenting equivalent political content in German, French, Italian, and Romansh over nearly five decades. Turning this collection into a usable, machine-readable parallel corpus requires solving two problems, accurate text extraction from heterogeneous scanned and born-digital PDFs, and reliable cross-lingual alignment, both of which remain challenging for historical documents and low-resource languages.

We present two complementary resources. VotingBooklets-Diamond is a manually corrected and verified gold-standard benchmark comprising 11 voting booklets from June 1977, December 1985, and March 2007, providing full OCR transcriptions and paragraph-level alignments across all four national languages. It serves as a reusable benchmark for evaluating document processing systems on real-world multilingual administrative material. We use it to evaluate recent AI-based approaches for OCR and multilingual alignment, comparing classical OCR systems, multimodal LLMs, and hybrid alignment pipelines, as well as fixed pipelines against more autonomous, agent-based ones. The results demonstrate that modern multimodal LLMs substantially outperform traditional OCR engines, while LLM-assisted alignment methods achieve high-quality multilingual alignments, including for the low-resource language Romansh.

Building on the best-performing pipeline, we introduce VotingBooklets, a large-scale four-language parallel corpus comprising 529 voting booklets, more than 19,000 PDF pages, and approximately 4.5 million tokens, spanning 1977 to 2026.

The corpus opens up several directions for future work. For low-resource language technology, it offers one of very few large-scale resources including Romansh, supporting machine translation and cross-lingual transfer. For NLP more broadly, it provides training and evaluation data spanning nearly five decades of consistent administrative language. And for political science and digital humanities, it enables longitudinal study of Swiss referendum discourse across four linguistic communities.

Database-first lexical infrastructure for Proto-Albanian research: interoperability, FAIR data, and reusable workflows for Albanian–Romanian cognacy and the Proto-Albanian–Thracian substrate segment

Edmond Cane (Luarasi University)

This paper presents the design of a database-first lexical infrastructure for research on Proto-Albanian, with particular attention to Albanian–Romanian cognacy and to the problem of the Thracian substrate in Romanian. The project starts from a familiar but methodologically difficult dossier. Romanian preserves a non-Latin lexical layer that has often been compared with Albanian, while Albanian retains inherited and early contact material that may provide the best living access to parts of the Paleo-Balkan record. The main difficulty is that such

correspondences have to be screened first against Greek, Latin/Romance, Slavic, and later areal interference before they can be interpreted as evidence for inheritance, contact, or substrate residue.

The project is situated in the current research landscape of Indo-European studies, where linguistic reconstruction increasingly has to be discussed together with chronology, archaeology, and population history (Heggarty et al., 2023; Lazaridis et al., 2025). It also takes into account the growing role of computational and phylogenetic methods in historical linguistics, especially Bayesian approaches that allow more explicit modeling of inheritance, uncertainty, and contact (Hoffmann et al., 2021; Neureiter et al., 2022).

The proposed infrastructure treats the database as the main research instrument rather than a place where results are deposited at the end. Each lexical observation is recorded with its form, concept, variety, attestation, distribution, and bibliographic trace. Each etymological interpretation is entered as a hypothesis with its diagnostics, confidence level, and review history. The model develops an initial SQL schema into a Lexibank-inspired, CLDF-compatible architecture intended for interoperability, FAIR publication, and reuse (List et al., 2022; Blum et al., 2025). It supports multilingual lexical data, competing etymologies, loan stratigraphy, substrate candidates, toponymic and onomastic evidence, and domain-based archaeological comparison, with export paths to CLDF for comparative datasets and TEI-Lex0 for long-term lexicographic preservation and reuse (Romary & Tasovac, 2018).

One important part of the design is the use of anchoring domains such as pastoralism, agriculture, metallurgy, transport, and kinship. These domains make it possible to compare lexical evidence with archaeological horizons and with structured semantic profiles. In this way, the Albanian–Romanian cognacy set is approached not as a list of isolated matches, but as a stratified research problem that can be reviewed, queried, and expanded. The broader aim is to build a standards-aware lexical resource that supports historical analysis, computational reuse, and later quantitative modeling within a FAIR and open research data framework.

References

Heggarty, P., Anderson, C., Kühnert, D., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. Science.

Hoffmann, K., et al. (2021). Bayesian phylogenetic analysis of linguistic data using BEAST.

Lazaridis, I., Patterson, N., Anthony, D. W., et al. (2025). The genetic origin of the Indo-Europeans. Nature.

List, J.-M., Forkel, R., Greenhill, S. J., et al. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scientific Data.

Blum, F., et al. (2025). Lexibank 2: pre-computed features for large-scale lexical data.

Neureiter, N., et al. (2022). contacTrees: a Bayesian phylogenetic model with horizontal transfer. Humanities and Social Sciences Communications.

Romary, L., & Tasovac, T. (2018). TEI Lex-0.

Session 5. Multilingualism & Linguistic Diversity

The importance of multilingual and cross-linguistic corpora for the study of language contact

Olivier Winistörfer (University of Zurich), Maxim Makartsev (University of Oldenburg), Anastasia Escher (ETH Zurich)

No geographically delimited area has attracted as much attention from linguists studying language contact as the Balkan Peninsula. Research on language contact among the Balkan languages has often focused on the presence or absence of particular linguistic features. However, as Winistörfer (submitted) argues, their quantitative distribution across geographical areas, genres, and L1 and L2 speakers is equally important for understanding contact-induced change. Such analyses have long been hampered by the lack of multilingual corpora and

linguistic resources, while for many Balkan varieties comprehensive corpora are still entirely absent.

To address this gap, we are developing a multilingual corpus comprising the Aromanian, Albanian, and Macedonian varieties spoken in North Macedonia. The corpus is designed to facilitate quantitative research on language contact by bringing together written and spoken data from different linguistic communities.

In this talk, we present the first stage of the project: a corpus of Macedonian currently comprising approximately 7 million tokens. The corpus includes literary texts, spoken language collected during our own fieldwork, the so-called Bombi, and published dialect transcripts (Vidoeski 1998; Guševska & Minova-Ǵurkova 1999). Importantly, it also contains speech produced by L2 speakers, which is explicitly annotated as such. To support a wide range of linguistic analyses, the corpus is annotated with both Multext-East (Kuzman Pungeršek et al. 2026) morphological tags and Universal Dependencies syntactic annotation (Straka 2018). We demonstrate how the combination of these annotation layers enables detailed investigations of language contact phenomena. The corpus is currently being expanded to include Aromanian and Albanian data from North Macedonia, ultimately providing a multilingual resource for comparative research on Balkan language contact.

References

Guševska, Liljana, Minova-Ǵurkova Liljana (2000). Taka se zboruva vo Ohrid, Skopje: Filološki fakultet "Blaže Koneski".

Kuzman Pungeršek, Taja; Rupnik, Peter and Ljubešić, Nikola (2026). South Slavic web corpus collection CLASSLA-web 2.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/2079.

Straka, Milan. "UDPipe 2.0 prototype at CoNLL 2018 UD shared task." Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. 2018.

Vidoeski, Božidar (2000). Tekstovi od dijalektite na makedonskiot jazik [Texts from the dialects of the Macedonian language], Skopje: Institut za makedonski jazik "Krste Misirkov".

Winistörfer, Olivier (submitted). Typology on the smaller scale - the construction of a typological dataset for varieties in and around the Balkans.

Linguistic data structures and standardization are not theory-neutral

Sandra Auderset (UNIBE), Tiago Tresoldi, Independent Researcher

This talk discusses the theoretical implications of data formats and standardization increasingly common in comparative linguistics. The quantitative turn has made the field of linguistics more data-hungry with large databases (e.g., Glottolog, Hammarström et al. 2024; Grambank, Skirgård et al. 2023) used to make broad claims about language, cognition, social structure, and genetics (e.g. Dediu 2011; Hahn 2020; Shcherbakova et al. 2023). 'Big data' sets all involve standardization and multiple levels of abstraction. Such standardization is often treated as neutral and objective, but we argue this is not the case: standards and formats encode decisions, thereby organizing what we know and shaping what questions we ask. These decisions are seldom explicitely built into formats but accumulate in the pipelines building on them. Lexibank's (List et al. 2022) transcription and cognate-coding conventions are a clear instance, with tone dropped from normalized lexemes, even though tone has been shown to carry phylogenetic signal (Auderset 2024). Standardization can also result in obscuring linguistic diversity by applying a one-fits-all approach from one language or family to others, such as basing our cross-linguistic knowledge of tone on Sinitic languages.

Two cases from our own work make these choices explicit. The first pairs a bottom-up database of sound change in Mixtec (Auderset & Campbell 2024) with a web application (Tresoldi & Auderset, in prep), deferring standardization to the analysis stage. The second builds on this with a Bayesian phylogenetic model that takes suprasegmentals into account (Tresoldi, in prep). We will briefly demonstrate both and illustrate how bottom-up database design, late aggregation for standardization, and modeling are feasible and informative, making theoretical implications explicit and open to question. For this, a federated and format-plural infrastructure such as the

one supported by CLARIN proves better suited than centralized curation, broadening the languages and data we focus on as well as the questions we can ask.

References

Auderset, Sandra. 2024. Rates of change and phylogenetic signal in Mixtec tone. Language Dynamics and Change 14(1). https://doi.org/10.1163/22105832-bja10031

Auderset, Sandra & Eric W. Campbell. 2024. A Mixtec sound change database. Journal of Open Humanities Data 10(1). https://doi.org/10.5334/johd.184

Dediu, Dan. 2011. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. Proceedings of the Royal Society B 278(1704): 474–479.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5: 180205.

Hahn, Michael, Dan Jurafsky & Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. Proceedings of the National Academy of Sciences 117(5): 2347–2353.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2024. Glottolog 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9: 316.

Shcherbakova, Olena, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon J. Greenhill, Damián E. Blasi & Hedvig Skirgård. 2023. Societies of strangers do not speak less complex languages. Science Advances 9(33): eadf7704.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. Science Advances 9(16): eadg6175.

Tresoldi, Tiago. In prep. [Bayesian phylogenetic model of suprasegmentals — title/venue to add.]

Tresoldi, Tiago & Sandra Auderset. In prep. [Mixtec web application — title/venue to add.]

At-Issueness Sensitivity in L2 Misinformation Detection: The Role of Cognitive Load

Alan Lombardini (UNIFR), Giulia Giunta (UNINE), Diana Mazzarella (UNINE), Didier Maillat (UNIFR)

This preregistered study examined L2-induced cognitive load on misinformation detection as a function of its at issueness. Previous work showed that false information is detected faster and more accurately when at issue (Giunta et al. 2025). We extended this line of research to L2 processing.

As L2 is typically more effortful to process (Foster-Cohen 2000), the resulting cognitive strain may alter how cognitive resources for misinformation detection are allocated. First, misinformation detection in L2 may rely on an optimization procedure that reallocates limited resources toward at-issue, relevant information. We call this the optimization hypothesis.

Alternatively, difficulty integrating syntactic information in L2 (Sorace 2011) may impair the identification of at issue content, as syntactic structure encodes propositional prominence (Gutzmann 2023). Misinformation detection may thus be less sensitive to the at-issueness of information. We call this the deprecation hypothesis.

To test these hypotheses, we adapted a misinformation detection task from Giunta et al. (2025). Four hundred participants evaluated twelve dialogues between a policeman and an informant. Their task was to judge whether the informant's answer was true. We measured accuracy and response times of misinformation detection.

At-issue content consistently facilitated misinformation detection across the L1 and L2 groups, with higher accuracy despite faster responses in at-issue trials. L2 participants were slightly more accurate overall, aligning with evidence that cognitive strain can enhance analytical thinking (Alter et al., 2007). The results of our correlation models reveal a graded effect of L2 competence on at-issue sensitivity, rather than strict support for either the optimization or deprecation hypothesis. Our broad claim is that language proficiency modulates sensitivity to at-issueness.

Selected references

Alter, Adam L., Daniel M. Oppenheimer, Nicholas Epley, and Rebecca N. Eyre. 2007. "Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning." Journal of Experimental Psychology: General 136, no. 4: 569-76.

Foster-Cohen, S. H. 2000. "Relevance: communication and cognition." Second Language Research 16, no. 1: 77–92.

Giunta, Giulia, Diana Mazzarella, and Filippo Domaneschi. 2025. "Are Presuppositions Really Misleading? Assessing the Impact of Linguistic Encoding, at-Issueness, and Source Reliability on Epistemic Vigilance." Mind & Language: 1-21.

Gutzmann, Daniel. 2023. "Gradient at-Issueness, Minimum Relevance, and Propositional Prominence." Theoretical Linguistics 49, no. 3-4: 239–47.

Sorace, Antonella. 2011. "Pinning down the Concept of 'Interface' in Bilingualism." Linguistic Approaches to Bilingualism 1, no. 1: 1-33.

Showcases & Demos

AViRI – Towards a Shared Infrastructure for Audiovisual Data and Metadata

*Josephine Diecke, Simon Spiegel and Teodora Vukovic (UZH)*Audiovisual (AV) materials such as archival footage, oral history recordings, video corpora, digitized broadcast collections are held and studied across many units at the University of Zurich (UZH), from film studies and various philologies and linguistics departments to ethnology, psychology, as well as life sciences. Until now, these various efforts have largely developed in isolation, with departments digitizing, curating, and analyzing AV data and metadata using different standards, tools, and workflows. AViRI (Audiovisual Research Initiative) was launched to address this fragmentation. Rather than a single research project, AViRI is a cross-departmental initiative that pools AV knowledge at UZH, bringing together researchers, librarians, technicians and developers around a shared digital ecosystem for audiovisual and multimodal data.

The initiative works on two fronts. The first concerns digitization and digital curation: aligning workflows, metadata standards, and infrastructural strategies for AV collections held by UZH and the University Library, so materials become sustainably accessible for research and teaching. The second concerns tools and methods for analyzing digitized and born-digital AV materials, drawing on in-house developments, including Videoscope (querying multimodal corpora), VIAN (computer-assisted video annotation), but also on external partners like TIB-AV-A (semi-automatic video analysis, developed by the Visual Analytics group at TIB Hannover).

This talk presents AViRI as a case study in building shared AV research infrastructure within a university. We will outline how the initiative emerged from recognizing scattered, redundant AV expertise across UZH; how it fosters collaboration between disciplines that rarely interact (film studies, linguistics, library science, visual analytics); and what challenges arise in aligning metadata standards, tool interoperability, and data governance across units.

Audio-video data processing pipelines

Masoumeth Chapariniya, Teodora Vukovic (UZH)

Abstract to follow.

MAVA and DAVA: building a cluster of infrastructure for processing multimodal audio-visual data

Teodora Vukovic (UZH)

Abstract to follow.

Posters

RefCo2 from the fields to the archive: multi-perspective testing of a corpus-quality platform for reusable language data

Jocelyn Aznar (UZH), Miranda Dickerman (UZH) and Jonathan Reich (UNIBE)

Linguistic corpora are our discipline's central, lasting output, but unlike the papers built on them, they are not reviewed (Michaelis, 2014), which weakens their reusability and reproducibility (Berez-Kroeker et al., 2018). RefCo2 answers this with theory-agnostic quality control: instead of imposing conventions (IPA, Leipzig, GOLD), the corpus creator documents their conventions and the platform checks the corpus for consistency (it contains exactly what it should) and coherency (the corpus and its documentation match each other).

Building on the RefCo toolkit, a spreadsheet and command-line checker (Lange & Aznar, 2022), RefCo2 is a web platform with guided documentation tabs and live, actionable feedback (each finding located by file, tier and timecode, with a suggestion), which generates reusable resources: glossary, grapheme chart, morpheme–gloss index. A standalone version can run offline for sensitive corpora. Its workflow runs in several steps: corpus upload, documentation, automatic checks, and an optional review by a certifying body. For this poster, we test the platform from the perspective of different linguistic fields and an archive.

Jonathan Reich, documenting an oral language of Northeast India, tests RefCo2 as a fieldwork companion: does early, automatic feedback make archive-ready, FAIR corpora easier to produce?

Miranda Dickerman, who built a cross-sectional language acquisition corpus of Vietnamese in the Language, Acquisition & Diversity Lab, checks how relevant that feedback is for curating a massive corpus.

Kelsey Neely, who documented a language and is now a Training Officer at ELDP, brings the archive and training perspective: can automatic verification and certification ease deposit curation?

Together, these three perspectives map what a corpus-quality platform must cover to serve corpus creators and archives alike, and where RefCo2 still falls short. We present RefCo2 as a concrete case of infrastructure for reusable language data and an invitation to say where it should go next.

References

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., et al. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1), 1–18.

Lange, H., & Aznar, J. (2022). RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process. *Proceedings of LREC 2022*, 2721–2729.

Michaelis, S. M. (2014). Annotated corpora of small languages as refereed publications: A vision. *Diversity Linguistics Comment*. https://dlc.hypotheses.org/691

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

The “Person Marking in South-Central Trans-Himalayan” (PMST) database

Linda Konnerth, Sandra Auderset, Jonathan Reich, and Susie Kanshouwa (UNIBE)

We present the database of Person Marking in South-Central Trans-Himalayan (PMST) as an open language resource. The PMST database contains paradigms of independent pronouns, intransitive, and transitive verb forms. Each language constitutes a separate, self-contained data set constructed according to the Paralex standard (https://www.paralex-standard.org/). Due to the common design principles underlying the data sets and their modular structure, they can be combined for comparative analyses and expanded in the future. The first two releases of PMST cover 20 languages and varieties (https://zenodo.org/communities/pmst/).

The South-Central (aka “Kuki-Chin”) branch of Trans-Himalayan comprises four dozen languages. Although the languages are closely related (e.g., VanBik 2009, van Driem 2015,

Peterson 2017, and DeLancey 2021), they display rich diversity in verbal person indexation, particularly in transitive paradigms. There we find an array of innovative markers for speech act participant object scenarios and inverse constructions (DeLancey 2017, and Konnerth and DeLancey 2019). Studying such diversity in a group of related languages provides a microtypological window into the diachronic dynamics of person marking (cp. the "language laboratory" idea, De Vogelaer and Seiler 2012).

This diversity also poses challenges for data collection, annotation, and organization, which we approach by combining methods and tools from language documentation and typology. We aim at a dense sample including both languages and lower-level varieties for maximal coverage. To ensure high quality data, comprehensive paradigms are collected including variant forms, in close collaboration with language experts and primary fieldworkers. Data are collected and distributed in accordance with the FAIR (Wilkinson et al. 2016) and CARE (Carroll et al. 2020) principles.

Our poster shows the structure of the database as well as several use cases illustrating the research questions that can be investigated with this resource.

References

Carroll, Stephanie & Garba, Ibrahim & Figueroa-Rodríguez, Oscar & Holbrook, Jarita & Lovett, Raymond & Materechera, Simeon & Parsons, Mark et al. 2020. "The CARE Principles for Indigenous Data Governance". Data Science Journal 19(1). 43. (doi:10.5334/dsj-2020-043).

DeLancey, Scott. 2017. "Hierarchical and accusative alignment of Verbal Person Marking in Trans-Himalayan". Journal of South Asian Languages and Linguistics 4(1). 85–105. (https://doi.org/10.1515/jsall-2017-0003).

DeLancey, Scott. 2021. "Classifying Trans-Himalayan (Sino-Tibetan) languages". In: Sidwell, Paul; and Jenny, Mathias (eds.), The Languages and Linguistics of Mainland Southeast Asia, 207-223. Berlin: Mouton de Gruyter.

De Vogelaer, Gunther, and Guido Seiler, eds. 2012. The Dialect Laboratory: Dialects as a Testing Ground for Theories of Language Change. Amsterdam: John Benjamins.

Driem, George van. 2015. "Tibeto-Burman". In: Wang, William S.-Y. & Sun, Chaofen (eds.), The Oxford Handbook of Chinese Linguistics, 135-148. Oxford: Oxford University Press.

Konnerth, Linda & DeLancey, Scott. 2019. "Introduction to 'Verb agreement in languages of the Eastern Himalayan region'". Himalayan linguistics 18(1). 1–7.

Peterson, David A. 2017. "On Kuki-Chin subgrouping". In Ding, Picus Sizhi & Pelkey, Jamin (eds.), Sociohistorical Linguistics in Southeast Asia: New Horizons for Tibeto-Burman Studies in honor of David Bradley (Languages of the Greater Himalayan Region 20), 189–209. Leiden, Boston: Brill. (doi:10.1163/9789004350519_012).

VanBik, Kenneth. 2009. Proto-Kuki-Chin: A Reconstructed Ancestor of the Kuki-Chin Languages (STEDT Monograph Series 8). Sino-Tibetan Etymological Dictionary and Thesaurus Project, Department of Linguistics research unit in University of California, Berkeley.

Wilkinson, Mark D. & Dumontier, Michel & Aalbersberg, IJsbrand Jan & Appleton, Gabrielle & Axton, Myles & Baak, Arie & Blomberg, Niklas et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data 3. 160018. (doi:10.1038/sdata.2016.18).

Infrastructural support for data analysis and reuse in Interactional Linguistics: potential of LCP-Videoscope

Johanna Miecznikowski (USI), Elena Battaglia (University of Bologna), Jérôme Jacquin (UNIL), Thomas Schmidt (University of Duisburg-Essen / linguisticbits), Teodora Vuković (UZH) and Jeremy Zehr (UZH)

LCP-Videoscope is a corpus platform offered by LiRI@UZH (Graën et al. 2024, Bubenhofer et al., June 23, 2025, Vuković et al., 2025). It supports browsing and querying transcribed and annotated multimodal corpora (see Bubenhofer et al., June 23, 2025, and Miecznikowski et al. under review, section 6.3.). By integrating this type of data, it has the potential of supporting data analysis and reuse in Interactional Linguistics (IL). We reflect on how the platform could be developed to more fully meet the needs of users in this field.

First, we describe recurrent analytical tasks in IL, based on several case studies (e.g. Jacquin et al. 2024 a,b, Miecznikowski & Battaglia under review). Data analysis in IL is typically based on collections of transcribed media snippets. Collections are built either from a corpus query or by manually identifying sequences; they are interpreted using various qualitative and, sometimes, quantitative methods. Several tasks are performed recursively: browsing and viewing, querying, selecting sequences, storing snippets and metadata, grouping.

Second, we identify technical features that support IL tasks across four specialized applications: two online corpus platforms (CLAPI, see Baldauf-Quilliatre et al. 2016, and ZuMult, see Frick & Schmidt 2025) and two offline applications (ELAN and Transana) (cf. also Batinić, Frick & Schmidt 2021).

Third, we evaluate which of those features are lacking in LCP-Videoscope and could be beneficial, especially in view of multi-application workflows. The features include: layout of transcripts in playscript style beside the existing tier-based view; video zooming beside existing timeline zooming; an internal workspace to store and modify collections; bulk export of media and/or transcript snippets in various interoperable and user-friendly formats; permalinks to the platform in exported collections, which allow for continued platform use to support the sequential and multimodal analysis of snippets and of their larger co-text.

References

Baldauf-Quilliatre, H., Colón de Carvajal, I., 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), 1–21. https://doi.org/10.4000/corpus.2991

Batinić, J., Frick, E., & Schmidt, T. (2021). Accessing spoken language corpora: An overview of current approaches. Corpora, 16(3), 417–445. https://doi.org/10.3366/cor.2021.0229

Bubenhofer, N., Vukovic, T., Miecznikowski, J., Zehr, J., Pavic, I.: "Making LCP-Videoscope fit for conversational video data: platform development and conversion pipelines". Presentation at the workshop Digital infrastructure for Language Data in Switzerland: current state and perspectives, USI, June 23, 2025

Frick, E., & Schmidt, T. (2025). Querying spoken language data. In P. Bański, U. Heid, & L. Herzberg (Eds.), Harmonising language data: Standards for linguistic resources (pp. 339–376). De Gruyter. https://doi.org/10.1515/9783112208212-014

Graën, J., Schaber, J., McDonald, D., Mustač, I., Rajović, N., Schneider, G., Vuković, T., Zehr, J., & Bubenhofer, N. (2024). The LiRI Corpus Platform. Linköping Electronic Conference Proceedings, 62–75. https://doi.org/10.3384/ecp210010

Jacquin, J., Etienne, C., Keck, A. et al. (2024a). Corpus POSEPI. CLAPI, ENS-Lyon (France). http://clapi.icar.cnrs.fr/Posepi.

Jacquin, J., Keck, A., Robin, C., Roh, S., & Rivoal, M. (2024b). 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]. DaSCH. https://ark.dasch.swiss/ark:/72163/1/0120.

Miecznikowski, J., Battaglia E. (under review). Usi evidenziali di ah come segnale di acquisizione di conoscenza in situ.

Miecznikowski, J., Battaglia, E., Schmidt, T., Zehr, J. (under review). "The TIGR corpus of spoken Italian". Submitted to: Herzberg, L., Witt, A., Mair, C. (eds.): Corpus Linguistics 2040: Which Data, Which Methods, Which Models? Series Digital Linguistics. De Gruyter Brill.

Vuković, T., Zehr, J., Schaber, J., Mustač, I., Rajović, N., McDonald, D., Graën, J., Bubenhofer, N. (2025) LiRI Corpus Platform: Demonstration of a Web-Based Infrastructure for Multimodal Corpus Analysis. Proc. Interspeech 2025, 302-303

Beyond Clinical Pain Descriptors: Detecting Figurative Pain Descriptions in Patient Narratives

Dolores Lemmenmeier (ZHAW), Nikolai Shurakov (UZH), Yvonne Ilg (UZH), Lucien Baumgartner (UZH) and Sabina Maria Räz (USZ)

Pain is a subjective experience that is inherently difficult to communicate. Figurative language therefore plays a central role in pain narratives, allowing patients to express qualities of pain that are difficult to convey through conventional clinical vocabulary (Semino, 2010), for example by describing pain as „als wenn jemand mit Hammer und Bohrer von innen die Stirn durchbrechen will" ("as if someone were trying to break through the forehead from the inside with a hammer and a drill"). Current diagnostic systems such as the International Classification of Headache Disorders (ICHD-3), however, rely on a limited inventory of pain descriptors (e.g., pulsating, pressing), which does not fully reflect the diversity of patients' pain descriptions (Eicher et al., 2023), let alone the figurative language they employ.

To better understand how patients communicate pain beyond established clinical terminology, the DHDH (Decoding Headaches in Digital Helvetia) project investigates pain descriptions in clinical consultations at the University Hospital Zurich (USZ) and in the online patient forum Headbook.de.

To detect figurative descriptions of pain, we combine rule-based candidate retrieval with manual linguistic annotation and sentence-embedding similarity search (Reimers & Gurevych, 2019). Candidate expressions are manually annotated for figurative form, source domain, and

associated pain quality. These validated examples are then used as seeds for retrieving additional semantically similar expressions.

Our preliminary analyses suggest that figurative language systematically refines clinical pain descriptors by distinguishing phenomenological subtypes within individual pain qualities. For example, pressing pain is conceptualized as a helmet, ring, metal plate, or fog, each highlighting different spatial, tactile, or dynamic properties. In a subsequent step, the annotated inventory will serve as evaluation data for few-shot LLM classification, allowing us to estimate the prevalence of figurative pain descriptions across the corpus.

References

Eicher, E.; Räz, S.; Stucki, P.; Röthlin, C.; Stattmann, M.; Grossenbacher, B.; Neumann, E.; Pohl, H.; Ilg, Y.; Maatz, A.; et al. ComPAIN—Communication of Pain in Patients with Headache. Clin. Transl. Neurosci. 2023, 7, 14. https://doi.org/10.3390/ctn7020014

ICHD-3 – Headache Classification Committee of the International Headache Society (IHS) (2019): Internationale Klassifikation von Kopfschmerzerkrankungen. 3. Auflage (ICHD-3). In: Nervenheilkunde 38, 3–182.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

Semino, E. (2010). Descriptions of Pain, Metaphor, and Embodied Simulation. Metaphor and Symbol, 25(4), 205–226. https://doi.org/10.1080/10926488.2010.510926

FAIRifying a Multilingual European Parliament Corpus for the Study of Gender-Inclusive Lexical Innovation

Nikos Tsourakis, Aurélie Picton, Julie Humbert-Droz, Charlotte Jeanneret and Anne Kosakevitch (UNIGE)

This contribution presents an ongoing effort to build and FAIRify a multilingual corpus of European Parliament committee documents in the context of the SEED4EU+ project IN-GENIOUS: Lexical Innovation for Gender-Inclusive Europe. IN-GENIOUS investigates how gender-inclusive lexical innovations emerge, circulate, become legitimised or contested, and enter institutional, media and digital discourse, with a primary focus on French and comparative perspectives from other European languages. The project combines corpus linguistics, discourse analysis and sociolinguistic approaches to develop a shared framework for studying inclusion across Europe.

The corpus focuses on three European Parliament committees: the Committee on the Environment, Public Health and Food Safety (ENVI), the Committee on Women's Rights and Gender Equality (FEMM), and the Committee on Civil Liberties, Justice and Home Affairs (LIBE). It covers three parliamentary legislatures between 2009 and 2024, offering a diachronic perspective on institutional discourse across successive political cycles. These committees offer complementary institutional contexts for analysing how gender, equality, inclusion, rights, public health, environment and civil liberties are articulated in parliamentary documentation. The source material was collected from the European Parliament Public Register of Documents and is organised by committee, date and language. The corpus is designed to include the 24 official languages of the European Union.

The contribution addresses both a research question and a data-management challenge: how can a locally organised collection of institutional documents be transformed into a reusable, citable and computationally analysable corpus? The proposed structure includes committee-level folders, date-based subfolders, language-specific TXT files, document-level metadata, a root-level file manifest, corpus-level metadata, methodological documentation, stable identifiers, source URLs and checksums. By preparing the corpus for deposit in SWISSUbase, the project aims to make this empirical basis transparent, reproducible and reusable. This corpus thus constitutes a rich multilingual resource that, once publicly available, will open a wide range

of research possibilities in computational linguistics and other fields, well beyond the scope of the original project.

CLARIN – a pan-European distributed digital infrastructure for language data and technology

Cristina Grisot (UZH)

Abstract to follow.

Two communities, one shared practice: presentation of CLARIN-CH RDM Working Group and SRDSN L-RDM Node

Joanna Blochowiak and Cristina Grisot (UZH)

Language data – spanning text, audio, video, and increasingly multilingual and multimodal corpora – has become central not only to linguistics but to law, medicine, history, and the social sciences. Yet it poses distinctive research data management (RDM) challenges: sensitive personal content such as voice recordings, heterogeneous formats, copyright and licensing constraints, and ethical questions around vulnerable communities and endangered languages. Two complementary initiatives have emerged in Switzerland to address these challenges from different angles.

The CLARIN-CH Research Data Management Working Group strengthens RDM practice among linguistic researchers themselves, connecting institutional data stewards and research-support services to build shared guidance, training, and domain-specific workflows. Its 2026 programme centers on producing curated, annotated Data Management Plan (DMP) examples for the Swiss language-sciences community, alongside a two-part workshop introducing SNSF-funded researchers to DMP planning and writing.

The SRDSN Language Research Data Management Node extends this work toward the data-stewardship and research-support community, promoting FAIR-aligned, discipline-spanning approaches to managing language data and raising awareness of AI- and LLM-assisted workflows – automated annotation, transcription, and metadata generation – alongside their risks and responsible use.

Though their audiences differ – researchers and linguists on one side, data professionals on the other – the two groups share common ground: Data Management Plans, FAIR data principles, AI-assisted workflows, and a single national network spanning UZH, ZHAW, SAGW, UNIBAS, UNIFR, USI, UNIGE, UNIL, UNIBE, UNINE, and swissrn.org. Both feed into a longer-term ambition: a nationally coordinated CLARIN-CH Training Programme built on lessons from both communities.

This poster presents the two initiatives side by side, highlighting their distinct objectives and audiences as well as the shared practices – DMPs, FAIR principles, and responsible AI use – that bridge them, illustrating how coordinated support across researcher and data-steward communities can raise the quality and reusability of language data across Swiss institutions.

LiRI and the Lia Rumantscha work together to build new infrastructure and collect data to promote Romansh

Ignacio Pérez Prat (Lia Rumantscha), Jannis Vamvas, Nikolina Rajović and Igor Mustač (UZH)

This work aims to translate computational linguistics research into practical societal impact, with a particular focus on Romansh language resources and technologies. Current and prospective activities include expanding Romansh data in the LCP, supporting educational

applications, developing TTS, and exploring collaborations on resources such as Pledari Grond and an online learning platform.

LiRI and Linguist List

Steven Moran and Valeriia Vyshnevetska (UZH)

LiRI is a strategic technology platform at the University of Zurich (UZH) that supports experimental, corpus-based, and digital research in linguistics, language sciences, and related disciplines across all project stages. It is also the official global host and operating institution for The LINGUIST List, the world's largest online resource and digital hub for the academic discipline of linguistics.

Swissdox RAG: Retrieval-First AI for Large Multilingual News Archives

Igor Mustač and Nikolina Rajović (UZH)

Swissdox RAG is a retrieval-first AI system for exploring large multilingual news archives. It combines lexical and semantic retrieval, reranking, grounded answer generation, citations, and corpus-level analytics to help researchers move from keyword search to evidence-based synthesis.