CLARIN

Selected papers from the
CLARIN Annual Conference 2025
Vienna, Austria

Selected papers from the

CLARIN Annual Conference 2025

Vienna, Austria, 30 September–2 October 2025

edited by Cristina Grisot and Thalassia Kontino

Linköping Electronic Conference Proceedings 222
eISSN 1650-3740 (digital) • ISSN 1650-3686 (print) 2025
ISBN 978-91-8118-672-7 (PDF)

Introduction

Cristina Grisot
University of Zurich
Zurich, Switzerland
cristina.grisot@uzh.ch

Thalassia Kontino
CLARIN ERIC
Utrecht, The Netherlands
thalassia@clarin.eu

The Common Language Resources and Technology Infrastructure (CLARIN) is a European Research Infrastructure Consortium (ERIC) dedicated to providing sustainable access to language resources, tools, and expertise for research in the humanities and social sciences (SSH) and beyond. Over the past decade, CLARIN has become a cornerstone of the European Open Science landscape by enabling the discovery, access, interoperability, and reuse of language data across national and disciplinary boundaries. Through its distributed infrastructure, technical standards, and continued commitment to the FAIR principles, i.e. making data Findable, Accessible, Interoperable, and Reusable, CLARIN supports a wide range of communities, including linguists, digital humanists, social scientists, cultural heritage professionals, educators, language technology developers, and increasingly researchers working with artificial intelligence. Currently there are 27 member countries and around 250 associated research institutions, which jointly promote standardization, multilingualism, and the responsible use of language technologies.

The CLARIN Annual Conference serves as the primary community forum for presenting and discussing developments within the CLARIN ecosystem and the wider field of language resources and technologies. Each year, the conference brings together researchers and infrastructure providers from across Europe and beyond. It offers a unique opportunity to exchange knowledge, showcase innovations, strengthen collaborations, and explore emerging challenges and opportunities related to language data, digital infrastructures, Open Science, and artificial intelligence. In recent years, workshops and panels have fostered the exchange with non-academic stakeholders, such as industry, policy-makers and cultural heritage institutions.

The CLARIN Annual Conference 2025, held in Vienna, Austria, reflected the breadth and vitality of the CLARIN community. The programmeme featured keynote lectures by Andreas Bauman (University of Vienna) and Lonneke van der Plan (Università della Svizerra italiana), thematic sessions, poster presentations, PhD meet-ups, and community discussions covering a wide range of topics. Contributions addressed advances in language resource creation and curation (e.g., Csuros and colleagues, Lenardič and colleagues, Fedchenko and colleagues, Depuydt and colleagues, Pluschkovits and colleagues), corpus and speech technologies (e.g., Schmidt and colleagues, Verdonik and colleagues), interoperability and metadata standards (e.g., Stemle and colleagues, Melaccio and colleagues, Sichera and colleagues), natural language processing, machine learning and large language models (e.g., Vandeghinste and colleagues, Mallia and colleagues, Grattan and colleagues, Kamocki and colleagues, Ahltorp and colleagues), good practices and the implementation of FAIR and Open Science principles (e.g., Lennes and colleagues, Wissik and colleagues, Heinisch and colleagues, Holdt and colleagues), and capacity building (e.g., Blockowiak and Grisot, van der Lek and colleagues). Together, the programme highlighted both the scientific achievements of the community and the continuing evolution of CLARIN as a research infrastructure that supports AI-enabled and data-intensive scientific activities.

The call for extended abstracts for the CLARIN Annual Conference 2025 was launched in early 2025 and invited contributions describing research, infrastructure developments, tools, services, resources, and community initiatives relevant to CLARIN and its stakeholder communities. The

Cristina Grisot and Thalassia Kontino 2025. Introduction. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. i–v. DOI: https://doi.org/10.3384/ecp222

conference attracted 59 submissions from 22 countries, demonstrating the broad international engagement with the infrastructure and its activities. Following a rigorous peer-review process conducted by the programme Committee, 44 contributions were accepted for presentation at the conference as oral presentations (16) or as posters (28) (acceptance rate 0.71). Accepted articles represent a diverse range of disciplines, methodologies and perspectives, reflecting the interdisciplinary character of the CLARIN community and its commitment to excellence in research, infrastructure development, and service provision. The Proceedings with extended abstracts are published on Zenodo.

After the conference all authors of accepted abstracts were given the opportunity to present their work in a full-length paper, for the current volume Selected Papers from the CLARIN Annual Conference 2025. All full-length paper submissions underwent the review cycle again, and 21 papers were accepted:

- The paper by Vandeghinste, V., Vanroy, B., & van Doeslaar, J. *Results of Crowdsourcing Human Evaluation of Synthetic Simplification* describes the creation of a set of crowd-sourced human evaluations of automated simplifications, which were assessed on four dimensions: fluency, clarity, accuracy and complexity.
- The paper by Mallia, M., Khan, F., Calvi, S., & Dankova, K. *Methodology for Converting and Publish Tabular Data into SKOS Resources via Python Notebooks* presents a Python notebook-based methodology for converting tabular data into multilingual SKOS vocabularies and Linked Open Data. The methodology enables reproducible workflows, user training, and efficient linguistic resource publication and management.
- The paper by Verdonik, D., Bizjak, A., & Donaj, G. *Govorjena slovenščina: A Platform for the Structured Collection of Conversational Speech* describes a specific component of the CLARIN.SI infrastructure useful for collecting, transcribing, and archiving spoken Slovene. The infrastructure supports sustainable conversational speech resources through standardized workflows, legal governance, and community participation.
- The paper by Grattan, M., Fried, A., & Jönsson, A. *Analysing Speech Acts in Organisational Communication* presents a speech act analysis pipeline and annotated dataset for Swedish organisational communication. The pipeline combines quantised large language models, retrieval-augmented prompting, and a fine-tuned multilingual classifier to classify discourse related to ISO/IEC 27001 and corporate communication.
- The paper by Wissik, T. et al. *CLARIAH-AT: Back (and) to the Future* presents the CLARIAH-AT consortium, including centres, services, training, and funded projects. The authors describe its evolution, organisational structure, and infrastructure supporting CLARIN and DARIAH activities in Austria.
- The paper by Csuros, K. et al. *Reading LEMI: A Corpus-Based Tool to Support Literacy in an Under-Resourced Language* describes a corpus-based Romanian literacy support platform that combines curated children's texts with an interpretable readability model, and demonstrates its use in education, text analysis, and teacher training across multiple empirical contexts.
- The paper by Schmidt, T., Ferger, A., & Frick, E. *Putting Things on Top of Other Things: The ZuMult Platform for Multimodal Corpora and Its Ecosystem* introduces an open-source platform for multimodal corpora with a flexible three-layer architecture enabling integration of diverse backends and customizable frontends for varied user needs.

- The paper by Lenardič, J. et al. *The CLARIN Resource Families Workflow* presents a new JSON-based workflow for the CLARIN Resource Families (CRF) initiative, enabling collaborative metadata management via a GitHub-hosted backend and aligning CRF with the CLARIN Virtual Language Observatory.

- The paper by Ahltorp, M., & Skeppstedt, M. *Multilingual Word Rains: Text Visualisation Built on Cross-Language Word Similarities* presents an extension of Word Rain for multilingual text visualisation which enables cross-lingual lexical frequency analysis and providing an open tool and web interface demonstrated on ParlaMint parliamentary data.

- The paper by Fedchenko, V., Urieli, A., & Bikard, A. *A Trilingual Parallel Corpus of Yiddish, French, and Russian Literary* presents a trilingual Yiddish–French–Russian parallel corpus integrated into ORTOLANG, detailing its construction, alignment using a custom dictionary, and use for translation studies and multilingual concordance analysis.

- The paper by Heinisch, B. *A Digital Humanism Perspective on Providing Language Resources to CLARIN in an Age of AI Commodification: The Case of UniTermGPT* discusses the challenges associated with data openness, attribution and potential downstream reuse in opaque AI training pipelines. The paper argues for ethical-by-design approaches to language resource provision that make human labor visible, preserve linguistic diversity and address power asymmetries between public research infrastructures and commercial AI actors.

- The paper by Depuydt, K. et al. *An Infrastructure for Historical Dutch Corpus Development* describes a 2025 Dutch historical text annotation infrastructure combining guidelines, gold standards, trained models, platforms and tools for automatic and manual linguistic annotation from the 13th to 19th centuries.

- The paper by Arhar Holdt, Š. et al. *Users' Experience and Development Priorities for CLARIN.SI* reports a CLARIN.SI user survey on infrastructure services, tools, and usage patterns. Despite the high user satisfaction, the survey reveals issues in resource discovery, metadata clarity, and search functionality, thus pointing to the high importance of outreach, onboarding and training.

- The paper by Kamocki, P. et al. *The AI Act and Its Impact on Large Language Models and the CLARIN Infrastructure* analyses the EU AI Act (2024/1689) and its implications for CLARIN, focusing on research exemptions, transparency requirements for AI-generated outputs, GPAI obligations, and dataset documentation in the context of language resource provision.

- The paper by Blochowiak, J., & Grisot, C. *Advancing FAIR Language Data through Training and Capacity Building: The Swiss Contribution to CLARIN* presents the CLARIN-CH training programme supporting Open Science and FAIR-based language data practices in Switzerland through needs assessment, stakeholder collaboration, training events, and development of open educational resources.

- The paper by Pluschkovits, M. et al. *A Resource for Lexical Variation in a Pluri-Areal Context: Introducing LexAT21, the Atlas on Lexical Variation in Austria in the 21st Century* introduces an atlas of lexical variation in Austrian German, outlining its survey dataset, initial analyses, CLARIN integration, and challenges in representing language varieties within standards.

- The paper by Stemle, E. W. et al. *Towards FAIR Metadata for Specialised Corpora: A Community-Informed Empirical Study of Schema Development in Two Communities* examines two case studies regarding domain-specific metadata schemata for FAIR data management. It highlights challenges linked to standardisation and stakeholder-driven design, and discusses CLARIN's role in sustainable metadata infrastructure.

- The paper by Melaccio, D. et al. *Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-Based Mediation* presents an ontology-based mediation approach for integrating CLARIN language resources into the Italian H2IOSC semantic framework, enabling layered semantic interoperability from CMDI metadata to marketplace-level data models for FAIR-aligned discovery and integration.

- The paper by Lennes, M. et al. *Implementing and Promoting Data Citation for CLARIN Resources at FIN-CLARIN* describes automated citation generation for Language Bank of Finland corpora using persistent identifiers and metadata, and discusses challenges and opportunities for harmonising data citation practices across CLARIN repositories.

- The paper by Sichera, P. et al. *Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring Communities and Orchestrating Digital Services* presents the Italian H2IOSC Observatory and the Marketplace. The two services combine community monitoring and service discovery to support SSH infrastructures through mixed-method analysis, semantic normalisation, and CLARIN-IT and OPERAS-IT integration.

- The paper by van der Lek, I. et al. *UPSKILLS Two Years on: Teaching about Language Resources and FAIR Data Principles with CLARIN* presents CLARIN-based educational initiatives building on the UPSKILLS project. These follow-up efforts provide a framework for progressive competence development and infrastructure-based pedagogy to enhance FAIR data literacy, curriculum modernization, and employability in language-related studies.

The editors would like to express their sincere gratitude to all authors who submitted their work to CLARIN 2025 and whose contributions made the conference possible. We thank the members of the programmeme Committee for their careful evaluations and constructive feedback, which ensured the high quality of the scientific programmeme and these proceedings. Our appreciation also goes to the keynote speakers, session chairs, panel contributors, and all participants whose engagement fostered lively discussions and knowledge exchange throughout the conference. Finally, we warmly acknowledge the Local Organising Committee, the CLARIN Office, the CLARIN National Coordinators, and the many individuals whose dedication and professionalism contributed to the successful organisation of the CLARIN Annual Conference 2025 and the publication of this proceedings volume.

programmeme Committee

- Carole Tiberius, Leiden University, the Netherlands
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia, Latvia
- Sara Stymne, Uppsala University, Sweden
- Steinþór Steingrímsson, Árni Magnússon Institute for Icelandic Studies, Iceland
- Marko Tadić, University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Vincent Vandeghinste, Dutch Language Institute, the Netherlands & KU Leuven, Belgium
- Joshua Wilbur, Center of Estonian Language Resources, Estonia
- Andreas Witt, University of Mannheim, Germany
- Tanja Wissik, Austrian Academy of Sciences, Austria
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University, South Africa
- Martin Wynne, University of Oxford, United Kingdom

Contents

Results of Crowdsourcing Human Evaluation of Synthetic Simplification

Vincent Vandeghinste, Bram Vanroy, Job van Doeslaar
Instituut voor de Nederlandse Taal, Netherlands
first.lastname@ivdnt.org

Abstract

This paper describes the creation of a set of crowd-sourced human evaluations of automated simplifications. We created a synthetic simplification dataset which was assessed by the crowd on dimensions of *fluency, clarity, accuracy* and *complexity*. We briefly describe the crowdsourcing environment and some techniques to keep the crowd engaged. This resulted in a dataset of nearly 25,000 responses from 384 respondents on synthetically simplified Dutch data. In order to mitigate the low inter-annotator agreement we investigate the effect of outlier removal techniques on Krippendorf α. The main conclusions stay the same: the synthetic texts are not simpler according to the linguistic proxies investigated, but the crowd judged that in about 80% of the cases the synthetic text was more clear than the original and less complex in about 70% of the cases. The fluency was the same for original sentences versus synthetic sentences, and in the case of accuracy the numbers changed most after outlier removal. A medium accuracy was achieved in about 80%, a high accuracy in 50% before outlier removal and 60% after outlier removal.

Both the synthetic parallel data as well as the crowd assessments are made available to the CLARIN infrastructure.

1 Introduction

Automatic text simplification (ATS) is the somewhat unfortunate term used in Natural Language Processing for the domain that works on automated text difficulty adaptation with the goal to enhance text accessibility for different groups of users. In government communication, there is particular emphasis on *plain language*, as evidenced by national campaigns in Belgium (Vandeghinste et al., 2021) and the Netherlands, and the ISO Standard on Plain Language (ISO, 2023). Clear and comprehensible communication is essential to ensure that all citizens, including those with limited literacy, can access and understand government information.

Recent advancements in artificial intelligence (AI) have enabled machine-based text simplification methods. However, a major challenge remains: the lack of reliable automated evaluation metrics, which depend on high-quality reference data. For most languages, including Dutch, such data is scarce and the development of robust training and evaluation datasets is still a key hurdle.

This work describes the pilot project *Duidelijke Taal* [EN: Clear language] which aligns with the policy of the Nederlandse Taalunie [EN: Dutch Language Union] for plain and understandable language in government and other societal sectors (law, healthcare). The Taalunie has been working closely with the Dutch and Flemish governments and various field organizations in the Netherlands and Flanders since 2016 to achieve this goal.[1]

The project originated from discussions between K-Dutch, the CLARIN Knowledge Centre for Dutch[2], hosted at the Instituut voor de Nederlandse Taal [EN: Dutch Language Institute][3] and the Taalunie, where it was established that in order to develop and properly evaluate a data-driven, state-of-the-art

[1] https://taalunie.org/dossiers/44/begrijpelijke-overheidstaal
[2] https://kdutch.ivdnt.org
[3] https://www.ivdnt.org

Vincent Vandeghinste, Bram Vanroy, and Job van Doeslaar 2025. Results of Crowdsourcing Human Evaluation of Synthetic Simplification. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 1–12. DOI: https://doi.org/10.3384/ecp222

approach for automatic text conversion to *plain language* in Dutch, there is a significant lack of manually validated simplified data for both training and evaluation purposes.

Until now, plain text writing (or text adaptation) of government texts was mainly done manually. However, recent rapid advancements in AI offer the possibility of having machines perform this task as well. This pilot project explores the current feasibility of this and assesses its effectiveness. We primarily target government communication, as it is supposed to be written in plain language so that as many people as possible can understand its content.

We developed a crowdsourcing approach to manually assess and validate AI-generated simplifications. The evaluation focused on four dimensions: *fluency*, which measures grammatical correctness; *complexity*, which assesses ease (or rather difficulty) of comprehension; *clarity*, which evaluates comparative clarity in text pairs; and *accuracy*, which determines how well the meaning is preserved. To enhance engagement of the crowd, gamification elements were incorporated in the crowdsourcing environment, encouraging participants to contribute consistently.

Related work on text simplification, Dutch datasets for text simplification, and human evaluation of simplifications are described in Section 2. The data used for this study and how synthetic simplifications were created are described in Section 3.1. To create a large-scale gold-standard test set, we developed a crowdsourcing application in which users manually evaluated automatic simplifications. This application is described in Section 3.3. The results of the crowd sourcing and analysis of inter-annotator agreement and outlier removal are presented in Section 4, and Section 5 concludes the paper.

2 Related Work

In Section 2.1, we review existing applications for Dutch text simplification. Section 2.2 outlines parallel datasets available for this task, and Section 2.3 discusses the metrics used for the automated evaluation of simplified texts.

2.1 Automatic Text Simplification

Until recently, automatic text simplification was often performed using machine translation techniques (Xu et al., 2016), where the original text was treated as the source and the simplified text as the target language. Wubben (2013) attempted this approach for Dutch using statistical machine translation. More recently, large language models for Dutch, such as (multilingual or Dutch-specific variants of) the T5 model (Raffel et al., 2020), have been fine-tuned with parallel data for text simplification (Seidl & Vandeghinste, 2024).

Several applications for automatic text simplification in Dutch are available on the market. We try to keep track of these at the CLARIN Knowledge Centre for Dutch (K-Dutch).[4]

The most notable is the *Lees Simpel* [EN: Read Simple] mobile app developed by Benedictus (2025), which allows users to take a picture of a complex government letter and provides a simplified version of the letter, i.e. a number of bullets as to what the letter contains, with a special focus on actions expected from the user. The app is completely free and quite popular. It is wrapped around existing commercial generative AI models (currently with extensive prompt engineering). No systematic methodologically stringent evaluations have been performed in order to measure the quality of the app (Benedictus, 2025).

A related app with a similar approach is the *Schrijf Simpel* [EN: Write Simple] app,[5] which is from the same creators and takes the text writer's point of view by providing suggestions for simplifying letters.

2.2 Data for Dutch Text Simplification

At K-Dutch,[6], the CLARIN Knowledge Centre for Dutch, we aim to maintain an overview of existing datasets for Dutch text simplification. We list the most important ones here.

[4]https://kdutch.ivdnt.org/wiki/Text_simplification

[5]https://schrijfsimpel.nl/

[6]https://kdutch.ivdnt.org/wiki/Simplification_Data

Manually Created Datasets So far, we have identified only one manually created dataset for Dutch, compiled by Vlantis et al. (2024), which pertains to municipal texts.

Automatically Translated Datasets Seidl and Vandeghinste (2024) translated the ASSET dataset from Alva-Manchego, Martin, et al. (2020) and the WikiLarge dataset from X. Zhang and Lapata (2017) into Dutch and used them for fine-tuning a language model. Another automatically translated dataset is the SimpleWiki dataset,[7] created by the National Forensic Institute.

Synthetic Datasets With the advent of generative large language models, we can now prompt such models to generate simplified texts. This approach has been used for the UWV Leesplank NL Wikipedia dataset,[8] and for the data employed in Van de Velde et al. (2023). It is also the method used in our research.

Comparable Corpus of Easy and Regular-Level Texts A comparable corpus in this context is a corpus of regular text on the one hand and easy or plain language texts on the other hand, with an indication of how similar their topic and content is.

A comparable corpus is available containing simplified news articles from the Wablieft newspaper and articles on the same topics or events from De Standaard (Vanackere & Vandeghinste, 2022).

2.3 Evaluation of Text Simplification

Metrics for the automatic evaluation of text simplification face several challenges. One of these challenges is that they often require reference simplifications based on a gold standard to compare against automatic simplifications. This applies to metrics such as SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTscore (T. Zhang et al., 2020). For research on Dutch simplification, there are hardly any reference sets available, and if they exist, they often consist of (automated) translations of English test sets, as in Seidl and Vandeghinste (2024).

For manual evaluation, Alva-Manchego, Scarton, and Specia (2020) describe how 3-point or 5-point Likert scales are often used to assess grammatical correctness (also referred to as fluency), meaning preservation (adequacy), and simplicity. Fluency is evaluated solely on the output, while adequacy and simplicity are generally assessed based on the text pair.

This approach is refined in Alva-Manchego, Martin, et al. (2020), where evaluation is conducted on a continuous scale (0-100) for adequacy, fluency, and simplicity. The use of continuous interval scales in crowdsourcing human evaluations is common practice in machine translation (Bojar et al., 2018), as it results in higher levels of inter-annotator consistency compared to ordinal Likert scales. In our approach, we also use continuous interval scales (without numerical indicators).

3 Methodology

3.1 The Dataset

We created a test set of 6,986 sentence pairs, selecting the original sentences from the WRPEI component of the SONAR corpus (Oostdijk et al., 2013),[9] which consists of texts from websites. We chose this component because we assume that it contains language intended for a broad audience and is therefore expected to be clear and accessible.

We selected only sentences containing between 10 and 50 words, including at least one verb, with a Readability Index coefficient (Brouwer, 1963) higher than 60, and consisting of more than one clause. This means that the sentence must include at least one subordinate clause. The selection was performed using the tools described in Vandeghinste and Bulté (2019) in order to only simplify sentences that initially exhibit a reasonable level of complexity.

These sentences were automatically simplified using GPT-4 with the same prompt as used in the UWV/Leesplank project, available on HuggingFace:

[7] https://huggingface.co/datasets/NetherlandsForensicInstitute/simplewiki-translated-nl

[8] https://huggingface.co/datasets/UWV/Leesplank_NL_wikipedia_simplifications

[9] Available through the CLARIN infrastructure at http://hdl.handle.net/10032/tm-a2-h5.

> "Simplify a Dutch paragraph directly into a single, clear, and engaging text suitable for adult readers that speak Dutch as a second language, using words from the 'basiswoordenlijst Amsterdamse kleuters.' Maintain direct quotes, simplify dialogue, explain cultural references, idioms, and technical terms naturally within the text. Adjust the order of information for improved simplicity, engagement, and readability. Attempt to not use any commas or diminutives."

The resulting dataset is made available for download in the CLARIN infrastructure of the Instituut voor de Nederlandse Taal, at https://hdl.handle.net/10032/tm-a2-y7. This parallel dataset, or parts thereof, could be uploaded into the application described in section 3.3, in the form of a CSV file, via a dashboard accessible to administrators.

Characteristics of the Synthetic Dataset

From the total dataset, a random selection of 1,071 parallel text snippets was made. In order to get multiple assessments for as many sentence pairs as possible. and have not too many sentence pairs with only a single user assessment, we started with a limited set of 500 sentence pairs and gradually increased the number of sentence pairs because the most frequent users had consumed nearly all available sentence pairs in the application.

Characteristics of the source material and the synthetic adaptation are provided in Table 1, showing average scores for the source and target material, as well as the significance of differences between the source and target language using the p-value from a t-test. For details on the operationalisation of these characteristics, we refer to Vandeghinste and Bulté (2019).

Characteristic	Source	Target	p	Characteristic	Source	Target	p
Text length	17.25	35.17	$<.01$	Syllables/word	1.51	1.57	$>.05$
Words/sentence	16.53	20.22	$<.01$	Age of acquisition	9.78	8.61	$>.05$
Avg. clause length	10.95	10.56	$<.01$	Content words/word	0.54	0.53	$>.05$
Subordinate clauses	0.81	1.58	$<.01$	Concreteness	2.89	2.85	$>.05$
Subordinate clauses/clause	0.35	0.31	$<.01$	77% ratio	78.57	84.15	$<.01$
Subordinate clause length	7.53	7.25	$>.05$	Long words	0.05	0.05	$>.05$
Words/finite verb	8.92	8.63	$<.05$	Preposition ratio	0.14	0.12	$<.01$
NP length	4.02	3.87	$>.05$	Specificity ratio	0.02	0.02	$>.05$
Dependency/head ratio	1.76	1.79	$<.05$	Pronoun ratio	0.08	0.10	$<.01$
Tree depth	7.17	7.53	$<.01$	Noun ratio	0.21	0.20	$<.01$
Type-token ratio	0.91	0.72	$<.01$	Numeral ratio	0.04	0.04	$>.05$
Guiraud	3.68	3.39	$<.01$	Article ratio	0.11	0.10	$<.05$
Flesch Reading Ease	62.21	53.68	$<.01$	Conjunction ratio	0.05	0.05	$>.05$
Flesch Douma	75.19	67.30	$<.01$	Adverb ratio	0.05	0.05	$>.05$
CLIB	75.47	83.11	$<.01$	Adjective ratio	0.06	0.06	$>.05$
CILT	78.00	79.06	$<.01$	Letter ratio	0.09	0.09	$>.05$
Brouwer	60.77	49.52	$<.01$	Verb ratio	0.16	0.17	$>.05$
Characters/word	4.74	4.78	$>.05$				

Table 1: Characteristics of the evaluated dataset

As we can see, the synthetic simplifications do not align with the standard characteristics expected from the comparison between regular and easy-to-read language, as established in Vandeghinste and Bulté (2019). For example, the average text length of the alleged simplification is longer, and the number of words per sentence is also higher. This is likely due to the fact that the prompt explicitly instructs the model to explain certain terms. However, we do observe that the average clause length is shorter, as is the number of words per finite verb. Differences in NP length and subordinate clause length are not significant. In terms of lexical density measures (type-token ratio, Guiraud), there is a significant indication of simplification.

Surprisingly, for readability metrics (Flesch, CLIB, CILT, Brouwer), the synthetic simplifications are rated as more difficult (lower scores). This may be largely due to the fact that nearly all these metrics use words/sentence as one of their features, and, as Table 1 shows, sentences in our targets are on average nearly 4 words longer.

For part-of-speech ratios, we find that the number of prepositions is significantly lower in the simplified texts, while the number of nouns and pronouns is significantly higher.

The six best-performing features for classifying texts as regular vs. simplified, as identified in Vandeghinste and Bulté (2019), all score counter-intuitively here.[10]

We would like to emphasize that these characteristics were calculated over the entire dataset, without considering the quality (particularly adequacy) of the simplifications. Further research should determine whether these patterns persist when calculations are made only for instances where the simplification was judged accurate.

Experimentation with different prompts also remains future work. Placing less emphasis on explanation may lead to very different values for the characteristics of the synthetic text.

3.2 The Assessors

The crowd that created the evaluation dataset was recruited through social media posting. Apart from an email address for the participants in the competition, no information was collected about the assessors. 384 people have assessed questions, with a maximum of 3756 questions answered by a single user, with an average of 64 and a median of 19, indicating that more than half of the users only answered a single questionnaire of 19 cases.

3.3 The Application

3.3.1 Introduction Screen

The introduction screen invites the user to participate. To keep the application as accessible as possible, users are not required to log in initially, allowing them to start answering questions without any additional steps. However, users have the option to log in immediately, enabling their score to be retrieved and allowing questions to be tailored based on their profile, ensuring they receive questions they have not previously answered.

3.3.2 Evaluation Module

After clicking on PARTICIPATE, the user enters the evaluation module. Manual evaluation of automatically generated simplified sentences is performed across multiple dimensions, such as *complexity*, *accuracy*, and *fluency*. Each session presents twenty questions, five from each of the four tasks. For every evaluation, the user sees a slider on a scale with labelled extremes, as shown in Figures 1a and 2. The scale is an interval scale with 100 units and the slider starts in the *neutral* position with a value of 50. The users do not see the exact value.

For complexity evaluation, we have defined two tasks: assessing a sentence for its *complexity* and determining which of two sentences in a sentence pair is clearer.

Sentence Evaluation for *Complexity*

In this task, sentences are randomly selected from the dataset, and users are asked to evaluate their complexity. These sentences can come from either the original or the automatically simplified set, ensuring a condition-blind evaluation.

Figure 1a shows the slider that users can adjust along a scale with *simple* and *complex* as endpoints.

Pairwise Comparison for *Clarity*

In pairwise evaluation, a sentence pair from the database, consisting of an original sentence and its automatic simplification, is randomly presented to the user for assessment. This ensures that there is no order effect and that the user evaluates in a condition-blind manner.

As shown in Figure 1, the user sees a slider on a scale with *Text A* and *Text B* as endpoints.

[10] These are words/sentence, dependency tree depth, clause length, Brouwer, Flesch and Flesch Douma.

(a) Example of a simplicity evaluation question

(b) Example of a pairwise evaluation question

Figure 1: Screen shots for the two complexity-related dimensions

(a) Example of a fluency evaluation question

(b) Example of an accuracy evaluation question

Figure 2: Screen shots for Fluency and Accuracy

Sentence Evaluation for *Fluency*

For this task, both original and automatically simplified sentences are presented to the user, who is asked to determine how well the sentences conform to proper Dutch. This is again a condition-blind evaluation.

As shown in Figure 2a, the user sees a slider on a scale ranging from *incorrect Dutch* to *correct Dutch*.

Pairwise Evaluation for *Accuracy*

In this task, a sentence pair is presented to the user, who assesses whether the content and meaning of the target sentence match the source sentence.

As shown in Figure 2b, the user sees a slider on a scale ranging from *different meaning* to *same meaning*.

3.3.3 Gamification

To encourage participants to provide as many answers as possible in the evaluation module, various gamification elements have been added, as shown in Figure 3.

Figure 3: Gamification elements

Streak The *streak* is *active* if the user has completed a session within the past 30 hours. When the streak is active, session scores increase by 10%.

Score A scoring system has been added to the application to motivate users. The score is based on a combination of effort, speed, and accuracy, operationalized as follows:

- Effort is measured by the number of words read.
- Answering speed is measured by the time taken per word and adjusted based on a scale. Answering too quickly suggests a lack of thorough evaluation.
- Accuracy is measured for sentences with three or more judgements. If the new rating is outside of 0.5 standard deviations of prior ratings, the accuracy score is gradually reduced.

Scoreboard A scoreboard is maintained on the website, displaying the top three scores at all times. A separate page shows the full ranking.

Level A user's level is displayed at the top of the page and is determined as follows: *Pro*: Top 10%; *Advanced*: Top 66%; *Beginner*: Remaining users.

We do not think Gamification put pressure on the annotators to provide answers closer to previously given answers, as this information was not available to the annotators. We did not provide much detail on the scoring system to the annotators to avoid adapting their annotation behaviour to the operationalisation of the scoring system. The main result of the scoring mechanism was that more answers lead to higher scores, so the main effect was that, for at least some users, this seemed to motivate them to provide many answers.

4 Results

The answers are stored in a database and can be downloaded in CSV format. Both individual responses and averages are made available for download at https://hdl.handle.net/10032/tm-a2-y8.

In total, we received 24,745 responses to the six questions that could be asked per sentence pair. Table 2 shows the distribution of answers across the six questions and the 1,071 sentence pairs. The *Count* column indicates the number of different annotations that were received for a specific question. The columns with the specific question headers show how many of such questions received a certain number of answers. For instance there were 58 *Clarity* questions that received 0 answers, and 523 *Clarity* questions that received 6 answers. The *All* column shows the number of sentence pairs for which all questions were answered at least so many times as indicated in the *Count* column.

It is noteworthy that for questions concerning parallel text pairs (clarity and accuracy), questions that were answered by six different annotators have the highest frequency. The system was set up to first select questions from a limited pool of 500 questions. Once those were exhausted for a user, a new pool of questions was activated.

For the questions where only one side of the parallel pairs was assessed, we see the highest number of responses for texts answered by three people. For sentence pairs where all questions were answered, we observe the highest frequency when two individuals answered all the questions.

4.1 Inter-annotator Agreement

Table 3 presents baseline inter-annotator agreement (IAA) scores, measured by Krippendorff's α, (Krippendorff, 2019) for the six evaluation scores. Overall, agreement levels are low across all conditions, with α values ranging from 0.10 (Clarity) to 0.17 (Complexity B), indicating substantial inconsistency among annotators. Accuracy and Clarity, despite having the highest number of ratings and items (around 1000), show particularly low agreement ($\alpha = 0.12$ and $\alpha = 0.10$ respectively), suggesting that annotators may have interpreted these criteria in varied ways or applied them inconsistently. The other four scales (Complexity A and B, and Fluency A and B) exhibit slightly higher αs (between 0.13 and 0.17) but are based on fewer ratings and items, around 3200–3300 ratings across approximately 740–750 items. These results collectively highlight the subjectivity of judgements on these criteria.

Count	Clarity	Accuracy	Fluency A	Fluency B	Simplicity A	Simplicity B	All
0	58	71	327	321	325	322	333
1	131	104	51	44	45	40	146
2	81	80	123	95	106	93	209
3	43	46	148	139	113	112	155
4	10	18	122	113	107	147	104
5	1	4	92	83	108	106	67
6	523	514	81	231	220	212	57
7	44	46	53	25	16	17	
8	36	42	36	9	18	6	
9	32	28	11	8	9	9	
10	29	32	11	1	3	4	
11	26	29	13	1		1	
12	17	20	3	1	1	1	
13	16	20				1	
14	9	7					
15	6	6					
16	3	3					
17	3	1					
18	2						
20	1						

Table 2: Distribution of responses across the six different questions. The *Count* column represents the number of different answers given to the same question. The *All* column represents the number of sentence pairs for which all questions were answered the number of times indicated in the *Count* column.

		All ratings		Outliers removed					
Condition	α	Ratings	Items	θ	α	Ratings	%Ratings	Items	%Items
Clarity	0.10	5848	1013	0.7	0.72	2619	45	928	92
Accuracy	0.12	5880	1000	0.8	0.70	3344	57	999	100
Fluency A	0.13	3222	744	0.7	0.79	1358	42	614	83
Fluency B	0.13	3256	750	0.7	0.76	1399	43	645	86
Complexity A	0.15	3263	746	0.7	0.76	1482	45	638	86
Complexity B	0.17	3276	749	0.7	0.79	1417	43	647	86

Table 3: Inter-annotator agreement measured in Krippendorff's α. Behind the vertical line, you see the threshold (θ) for the Z-score, which was the best outlier removal method.

To improve the reliability of the human ratings by filtering out inconsistent judgements, we have tested three outlier removal methods: Z-score, Modified Z-score (Z-mod), and Interquartile Range (IQR). *Z-score* (Montgomery & Runger, 2010) identifies outliers by measuring how many standard deviations a value lies from the mean; values beyond a threshold θ are flagged as outliers. *Z-mod* (Iglewicz & Hoaglin, 1993) uses median and median absolute deviation, making it more robust to skewed or non-normal distributions. *IQR* (Tukey, 1977) classifies outliers as values that fall below the first quartile (Q1) minus θ times the IQR, or above the third quartile (Q3) plus θ times the IQR. While each technique balances sensitivity to outliers and robustness differently, experimenting showed that Z-score provided the best outlier filtering method, allowing us to keep the largest number of ratings while achieving an acceptable agreement ($\alpha > 0.667$).[11]

For each of the conditions we have gradually raised the θ value until we reach the acceptability limit of $\alpha > 0.667$. The resulting θ and α values are presented on the right side of Table 3. The table also includes the number of remaining answers after filtering out the outliers. The Ratings and Items columns indicate how many ratings and items remained after removing the outliers.

Table 4 shows the updated distribution of answers after outlier removal. We see that for the pairwise

[11]The attentive reader may notice a difference in the numbers and percentages in the last four columns of Table 3 compared to the numbers presented in Vandeghinste et al. (2025). The previously reported numbers came from a different practical operationalisation and implementation of the outlier-removal procedure. The current numbers are the ones obtained after actually generating the dataset without the outliers.

Count	Clarity	Accuracy	Fluency A	Fluency B	Simplicity A	Simplicity B	All
0	60	72	329	322	326	324	342
1	195	158	156	153	123	134	398
2	234	243	323	307	323	299	291
3	202	176	119	143	150	195	39
4	152	157	78	94	109	84	1
5	133	177	38	45	30	27	
6	39	35	21	5	9	3	
7	18	24	5			4	
8	16	8	2	1	1	1	
9	11	8		1			
10	6	3					
11	3	6					
12		3					
13	1	1					
15	1						

Table 4: Distribution of answers after outlier removal

conditions, having two remaining answers has the highest frequency, and this also holds for the single sentence assessments. Overall, we see that having at least a single answer per condition has the highest frequency.

4.2 Analysis of the Answers

A summary of the answers is shown in Table 5, both for the full dataset as well as the one with outliers removed. If we set the threshold for an accurate simplification at 50 (*Accuracy 50*), we find that 83% of the simplifications are rated as accurate. If we set the threshold at 70, this is true for only 49% of cases. After outlier removal, this does not change much, with nearly 82% of the remaining data points indicating an accurate simplification of at least 50%.

Condition	**% of Texts**	**% of Texts after outlier removal**
Accuracy 50	83.00	81.88
Accuracy 70	48.60	61.16
Complexity A $>$ B	69.44	68.86
Fluency B $>$ A	50.54	51.15
Clarity B $>$ A	79.57	80.42

Table 5: Pairwise comparisons. Text A is the original text. Text B is the synthetic text.

When we raise the bar and only assume accuracy when the average user assessment is higher than 70, we see the accuracy levels drop to 49% for the full dataset. After outlier removal, this increases to 61%, indicating that in the majority of cases (with outliers removed) there is a strong judgement that the accuracy is good. Together with the *Accuracy 50* numbers, we can conclude that the accuracy is considered better than neutral between *different meaning* and *same meaning* in more than 80% of the cases (interpretation: *more ok than not ok*, and scoring above 70/100 accurate (interpretation: *good*) in about 50-60% of the cases.

In terms of complexity (in which source and synthetic simplification were assessed independently), we see that the complexity of the source sentences is assessed to be higher in nearly 70% of the cases. This does not change much with outlier removal. So according to the assessors, simplification took place in about 70% of the cases.

In terms of fluency (which was also assessed independently), there is virtually no difference in assessment between the original sentences and the synthetic versions. This does not change much after outlier removal either.

In terms of clarity, the synthetic versions are judged to be clearer in about 80% of the cases. This does not change much with outlier removal.

Looking at raw averages (Table 6), we see that the accuracy of the simplification is rated at 67, which rises up to 71 after outlier removal. The complexity of the original sentence (A) receives an average score of 50, while the complexity of the synthetic sentence (B) is rated at 41. This difference is statistically significant ($p < .01$), and even increased after outlier removal.

The difference in fluency between sentence A and sentence B is negligible, both before and after outlier removal. In pairwise comparison of clarity, the average assessment was 64% (67% after outlier removal), indicating that on average, synthetic sentences were assessed to be clearer.

	Before		After outlier removal	
Condition	**Mean**	**Stdev**	**Mean**	**Stdev**
Accuracy	67.08	19.59	71.16	23.44
Complexity A	50.18	14.12	51.78	17.52
Complexity B	40.95	15.87	41.22	18.71
Fluency A	62.73	19.50	64.43	23.61
Fluency B	63.65	18.56	64.97	23.15
Clarity	64.19	19.47	66.73	23.21

Table 6: Raw averages of the evaluations, before and after outlier removal

This information can be leveraged to improve future ATS systems by focusing on aspects that most impact readability. Additionally, discrepancies between automatic readability scores and human judgements underscore the need for refined evaluation metrics that align more closely with human perception.

5 Conclusions

In the preceding sections, we described a crowd sourcing evaluation campaign to evaluate state-of-the-art automatic text simplifications for Dutch. Our findings show that there is little to no parallel data where we can reliably assume that the simplified version is indeed a proper simplification. To obtain human evaluations, we established a crowdsourcing infrastructure and gathered opinions from the crowd on a number of sentence pairs.

The responses from the crowd indicate that a distinction should be made between *clarity* and *simplicity*. Sentences in which the most difficult concepts are clarified are not necessarily syntactically or lexically simpler.

It should further be noted that there is a large amount of inter-annotator disagreement in our dataset, and that we also applied pairwise comparisons after extensive outlier removal. As Tables 5 and 6 show, this has only a minor effect on the results, and does not alter our interpretation of these results.

The dataset can assist in building or fine-tuning a rewriting system, as demonstrated in Seidl and Vandeghinste (2024). Additionally, we can sample a subset of this dataset, retaining only those parallel texts with the highest inter-annotator agreement, high accuracy, and substantial simplification or clarification. This subset could then be made available separately to the research community, for example, as an evaluation set.

Based on the available data, work can also be done on developing automatic evaluation metrics for text simplification in Dutch that optimally correlate with human judgement, following the approach of the COMET metric in machine translation (Rei et al., 2020). The data can also be used to build classification systems that assess a given text for fluency, simplicity, and clarity. We can think of integration into tools such as the Schrijfassistent.[12] Furthermore, research can be conducted on the impact of specific linguistic features on aspects such as language clarity.

[12] https://schrijfassistent.be/index.php

References

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., & Specia, L. (2020, July). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4668–4679). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.424

Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135–187.

Benedictus, H. (2025). Lees Simpel App. Personal Communication. https://www.leessimpel.nl

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., & Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 272–303. https://doi.org/10.18653/v1/W18-6401

Brouwer, R. (1963). Onderzoek naar de leesmoeilijkheden van Nederlands proza. *Pedagogische Studiën*, *40*, 454–464.

Iglewicz, B., & Hoaglin, D. (1993). *How to detect and handle outliers*. ASQC Quality Press.

ISO. (2023). *Plain language — part 1: Governing principles and guidelines* (tech. rep. No. ISO 24495-1:2023). International Organization for Standardization. Geneva, Switzerland. https://www.iso.org/standard/78907.html

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th edition). Sage Publications.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

Montgomery, D., & Runger, G. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential speech and language technology for Dutch* (pp. 219–247). Springer.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Seidl, T., & Vandeghinste, V. (2024). Controllable Sentence Simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, *13*, 31–61. https://www.clinjournal.org/clinj/article/view/171

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Van de Velde, C., Vanroy, B., & Vandeghinste, V. (2023). Automatic sentence-level simplification for Dutch. *Computational Linguistics in the Netherlands. Abstracts*.

Vanackere, N., & Vandeghinste, V. (2022). *Building a comparable corpus between easy-to-read Dutch Wablieft and De Standaard* [Master's thesis, KU Leuven. Faculteit Ingenieurswetenschappen].

Vandeghinste, V., & Bulté, B. (2019). Linguistic proxies of readability: Comparing easy-to-read and regular newspaper dutch. *Computational Linguistics in the Netherlands Journal*, *9*, 81–100. https://www.clinjournal.org/clinj/article/view/97

Vandeghinste, V., Muller, A., François, T., & Declercq, O. (2021). Easy Language in Belgium. In C. Lindholm & U. Vanhatalo (Eds.), *Handbook of Easy Languages in Europe* (pp. 53–90). Frank & Timme.

Vandeghinste, V., Vanroy, B., & Van Doeslaar, J. (2025). Human evaluation of automated text simplification through crowdsourcing. *Proceedings of the CLARIN Annual Conference*, 143–147.

Vlantis, D., Gornishka, I., & Wang, S. (2024, May). Benchmarking the simplification of Dutch municipal text. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2217–2226). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.199

Wubben, S. (2013). *Text-to-text generation by monolingual machine translation* [Doctoral dissertation, Tilburg University] [Series: TiCC Ph.D. Series Volume: 26].

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401–415. https://doi.org/10.1162/tacl_a_00107

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Zhang, X., & Lapata, M. (2017). Sentence simplification with deep reinforcement learning. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 584–594. https://doi.org/10.18653/v1/D17-1062

Methodology for Converting and Publish Tabular Data into SKOS Resources via Python Notebooks

Michele Mallia
CNR-ILC
Pisa, Italy
michele.mallia@ilc.cnr.it

Fahad Khan
CNR-ILC
Pisa, Italy
fahad.khan@ilc.cnr.it

Silvia Calvi
Osservatorio di Terminologie
e Politiche Linguistiche
Università Cattolica
del Sacro Cuore, Italy
silvia.calvi1@unicatt.it

Klara Dankova
Osservatorio di Terminologie
e Politiche Linguistiche
Università Cattolica
del Sacro Cuore, Italy
klara.dankova@unicatt.it

Abstract

This paper presents a methodology for creating and converting tabular data into SKOS linguistic resources using Python notebooks. Designed to support users with limited technical skills, the approach offers a structured and reproducible process through interactive notebooks. The methodology covers metadata preparation, data scraping from repositories, information mapping, and data normalization for standardized vocabulary publication. Various specialized multilingual vocabularies, including those related to textile description and smart city terminology, were analyzed to evaluate the approach. Guidelines were also developed to optimize LOD environment deployment, including resource uploading and web application configuration. The methodology supports linguistic data management such as Linked Open Data, offering a platform for hosting SKOS resources and training users to create structured data efficiently. The system was tested through external user engagement, demonstrating its scientific relevance and practical utility.

1 Introduction

The increasing adoption of Linked Open Data (LOD) practices in linguistic resource management highlights the need for efficient and accessible methods to convert tabular data into RDF[1]. While progress has been made in promoting Linked Data principles, transforming non-standardised formats such as CSV or TSV into SKOS[2] remains a challenge. Existing tools (Fiorelli & Stellato, 2021) often require advanced technical skills and are not always user-friendly. OpenRefine[3] stands out as a powerful tool for data transformation, but lacks a training-oriented approach.

To address this gap, we propose a methodology based on Python Jupyter notebooks[4], designed to guide non-expert users step by step through the process of converting tabular data into SKOS. This approach combines user accessibility with data interoperability and quality, serving both as a practical tool and a training resource. Unlike general-purpose tools or language models like ChatGPT, our notebooks offer structured, domain-specific, and reproducible workflows aligned with LOD best practices. Although a systematic user evaluation is still ongoing, the notebooks are designed to support non-expert users through a guided, transparent workflow with a strong didactic focus. Unlike tools such as OpenRefine, our approach emphasizes step-by-step training, reproducibility, and domain-specific alignment with SKOS modelling practices.

The methodology has been tested in a real-world use case as part of a pilot initiative (described in Section 2.1), where external users applied the notebooks in a research context. This evaluation confirmed

[1] https://www.w3.org/RDF/

[2] https://www.w3.org/2004/02/skos/

[3] https://openrefine.org/

[4] https://github.com/cnr-ilc/ilc4clarin-jupyter-skos-notebook

Michele Mallia, Fahad Khan, Silvia Calvi, and Klara Dankova 2025. Methodology for Converting and Publish Tabular Data into SKOS Resources via Python Notebooks. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 13–24. DOI: https://doi.org/10.3384/ecp222

their usability and potential for broader adoption within linguistic data platforms. In addition, this workflow was used to represent and disseminate terminological data from the multilingual lexicons of the Pan-Latin Terminology Network – REALITER[5] in compliance with FAIR principles, thanks to the collaboration between the CNR-ILC's B center of ILC4CLARIN - CLARIN-IT[6] (the Italian national node of CLARIN ERIC[7], the European infrastructure for language resources and technologies) and OTPL (Osservatorio di Terminologie e Politiche Linguistiche, Università Cattolica del Sacro Cuore – member of CLARIN-IT). More specifically, the REALITER – OTPL collection in the CLARIN-IT repository has been created and the lexicons have been published on the ILC4CLARIN - Skosmos Service platform.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work and contextual background. Section 3 presents case studies of REALITER lexicon to illustrate the practical application of this approach; in this section we discuss also the results and challenges encountered. Section 4 details the methodology, including data acquisition, transformation, and deployment, followed by conclusions and suggestions for future work in Section 5.

2 Background and context

The landscape of linguistic data management is increasingly shaped by the principles of Linked Open Data (Berners-Lee, 2006), which aim to enhance the accessibility, interoperability, and usability of structured data across digital platforms. The integration of linguistic data into LOD frameworks is particularly significant, as it fosters the development of interconnected language resources, contributing to a more dynamic and semantically enriched web of data (Khan et al., 2022). In the following article we will focus on resources which are based on the SKOS vocabulary[8], which are part of the Linguistic Linked Open Data Cloud[9] under the category of 'Terminologies, Thesauri and Knowledge Bases'.

2.1 H2IOSC LLOD Pilot

The project described in this paper is part of a broader pilot initiative within the H2IOSC project[10] (Humanities and Heritage Italian Open Science Cloud), which aims to promote open and FAIR (Lamprecht et al., 2020) access to services and data in the domain of humanities and cultural heritage. As part of this initiative, the LLOD pilot[11] (Linguistic Linked Open Data pilot) focuses on advancing LLOD technologies for managing and disseminating linguistic resources in interoperable and semantically rich formats. The LLOD pilot is developing a comprehensive infrastructure to support the entire lifecycle of linguistic data as Linked Data. This includes components for the transformation of tabular or XML-based resources into RDF, tools for browsing and editing language data, and services for publishing and querying linked data via SPARQL endpoints. The platform also integrates with Skosmos for vocabulary browsing, and features a quality assessment dashboard to ensure data consistency, adherence to LOD principles, and compliance with vocabularies such as OntoLex-Lemon[12] and SKOS.

To ensure long-term accessibility and sustainability, resources are ingested into a central CLARIN repository[13], which supports version control, metadata management, and integration into the broader European research infrastructure. Moreover, the pilot places strong emphasis on training and community support, providing tutorials, interactive Jupyter Notebooks for data conversion, and dedicated assistance to facilitate adoption by researchers, educators, and institutions. Through this approach, the LLOD pilot

[5] https://www.realiter.net/

[6] https://www.clarin-it.it/it

[7] https://www.clarin.eu/

[8] https://www.w3.org/TR/swbp-skos-core-spec/

[9] https://linguistic-lod.org/

[10] H2IOSC Project – Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU – National Recovery and Resilience Plan (NRRP) – Mission 4 "Education and Research" Component 2 "From research to business" Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures" Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe" – Project code IR0000029 – CUP B63C22000730005. Implementing Entity CNR. https://www.h2iosc.cnr.it/

[11] https://pllod.clarin-it.it/

[12] https://www.w3.org/2019/09/lexicog/

[13] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/

within H2IOSC contributes to the creation of a national infrastructure for linguistic Linked Data that is open, reusable, and aligned with European Open Science practices.

3 Case Studies: REALITER Pan-Latin lexicons

This methodology has been applied in a use case involving the processing of specialised vocabularies (Grimaldi & Zanola, 2021) developed within the Pan-Latin Terminology Network - REALITER. This association, created in 1993, aims at promoting plurilingual communication in specialised fields. REALITER brings together experts, universities, associations, and institutions. Its institutional members include the Délégation générale à la langue française et aux langues de France, the Office québécois de la langue française, the Centre de Terminologia TERMCAT, the Servizo de Normalización Lingüística of the Universidade de Santiago de Compostela, the Associazione Italiana per la Terminologia (Ass.I.Term) and the Osservatorio di Terminologie e Politiche Linguistiche of the Università Cattolica del Sacro Cuore[14]. REALITER encourages activities aimed at harmonising Romance languages and fostering linguistic diversity. More specifically, it develops and manages multilingual terminological resources to support experts, citizens, and academic communities (Gilardoni, 2011; Zanola, 2014). The Network members participate in collaborative projects to develop terminology products (e.g. lexicons) on domains of current interest. Since its foundation, around 40 lexicons across domains such as economics, environment, medicine, biology, digital technologies, education and fashion have been produced. The lexicons are multilingual since they show equivalents in seven Romance languages (Catalan, Spanish, French, Galician, Italian, Portuguese, Romanian), including some varieties such as Spanish spoken in Argentina, Colombia, Mexico; French spoken in Belgium, Canada; Brazilian Portuguese and Moldovan Romanian, as well as in English.

The development of REALITER multilingual lexicons follows the Common methodological principles[15]. Each member can propose a multilingual terminological project, that needs to be approved at the General Assembly of members. The proposals are evaluated according to the relevance of the chosen domain and its social interest. The member leading the project is responsible for coordinating the work and for the nomenclature proposal. For each concept, a definition in the reference Romance language is provided alongside with one or several terms used to designate it. Furthermore, one or more English equivalents are also identified to facilitate the communication among plurilingual working groups. For each language and variety considered, a team of terminologists and domain experts is gathered to provide the equivalents taking into account terminological variation and cultural aspects of each language and variety (Grimaldi et al., 2022; Lo Presti & Dankova, 2021). The structure and the content of lexicons can be slightly different from one project to another in order to meet specific needs of the domain of interest. For instance, the lexicons dealing with the new Olympic sports published in 2023-2024 [16] – beach volley, surf, trampoline and climbing – are divided in several thematic sections, such as equipment, competition types and athlete roles. Another interesting example concerns the definition of concepts: usually, the definition is given in the source language of the project (Fig. 1), however, in some cases, a definition is missing (Fig. 2) or it can be replaced by an image, as in the *Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume* (Fig. 3).

This case study addresses the conversion of 18 multilingual REALITER lexicons (3,467 entries) into SKOS and their hosting on the LLOD pilot SKOSMOS platform. The vocabularies, mainly in CSV format, contain labels, definitions and notes. Each entry contains a preferred term (`prefLabel`) and often several alternative terms (`altLabel`), with grammatical information following language-specific conventions (e.g. *s. m.* for masculine singular nouns in Romanian, *n. f.* for feminine singular nouns in French). Terms are marked with language and, for geographical variants, with country (e.g. Spanish

[14]For more information, see Gilardoni 2011; Zanola 2014 and the REALITER website: https://www.realiter.net/ (accessed on 9 December 2025).

[15]For more information, please refer to the *Principes methodologiques du travail terminologique*. Available at: https://www.realiter.net/fr/quest-ce-que-realiter (accessed 9 December 2025).

[16]These lexicons are available on the website: https://www.culture.gouv.fr/thematiques/langue-francaise-et-langues-de-france/agir-pour-les-langues/moderniser-et-enrichir-la-langue-francaise/nos-publications/vocabulaires-panlatins-du-sport (accessed 9 December 2025).

27.	*it*	**città intelligente (n.f.)** **Smart City (n.f.)**
	DEF	Città caratterizzata dall'integrazione tra saperi, strutture e mezzi tecnologicamente avanzati, propri della società della comunicazione e dell'informazione, finalizzati a una crescita sostenibile e al miglioramento della qualità della vita.
	ca	ciutat intel·ligent (n.f.)
	es	ciudad inteligente (n.f.)
	es [MEX]	smart city (n.f.)
	fr	ville intelligente (n.f.)
	gl	cidade intelixente (n.f.)
	pt	cidade inteligente (n.f.)
	ro	oraș inteligent (n.n.)
	en	smart city

Figure 1: Città intelligente, *Pan-Latin Smart City Lexicon* (2018), p. 21.

madre biológica (n.f.)

cat	**mare biològica (n.f.)**
fra	**mère biologique (n.f.)**
glg	**nai biolóxica (s.f.)**
ita	**madre biologica (n.f.)**
por	**mãe biológica (s.f.)**
ron	**mamă biologică (s.f.)**
eng	***biological mother***

Figure 2: madre biológica, *Panlatin Bioethics Glossary* (2007), p. 35.

16.	*it*	**manica a sbuffo (s.f.)**
	ca	màniga de fanal (n.f.) màniga de pernil (n.f.)
	es	manga abullonada (s.f.)
	es [MEX]	manga inflada (s.f.)
	fr	manche gonflée (n.f.)
	pt	manga curta bufante (s.f.)
	pt [BR]	manga sopro (s.f.)
	en	puffed sleeve balloon sleeve

Figure 3: Manica a sbuffo, *Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume* (2022), p. 25.

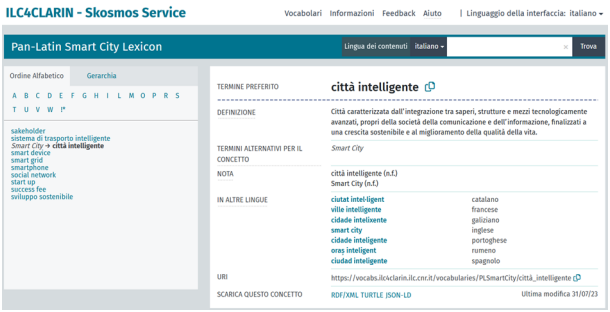

Figure 4: Città intelligente, *Pan-Latin Smart City Lexicon*.

(Mexico), Portuguese (Brazil)). Additional information is provided in notes. The collaborative nature of the data required dedicated parsing algorithms. By analysing the syntax of these taxonomies, an algorithmic foundation was defined to guide the conversion workflow. After conversion, the resources were integrated into an institutional vocabulary platform based on SKOSMOS, enabling browsing, multilingual search, and REST API access.

The Figure 4 represents the lexical entry *città intelligente* (smart city) of the *Pan-Latin Smart City Lexicon* in the SKOSMOS vocabulary platform.

4 Methodology

In this section, we describe a comprehensive pipeline for creating SKOS resources from tabular data, focusing on the transformation and publication processes. The workflow is designed as a modular and reproducible sequence of steps that guides users from the initial acquisition of metadata to the final deployment of the generated RDF vocabulary. Particular attention is given to ensuring that the conversion from tabular formats to SKOS remains transparent, standards-compliant, and suitable for integration into LLOD-oriented infrastructures. The following subsections outline the main components of this pipeline and illustrate how the methodology supports both technical processing and user-oriented training activities.

4.1 Overview of the LLOD Pipeline

This methodology outlines a streamlined approach for creating SKOS-formatted linguistic resources. The workflow begins with the collection and normalisation of metadata retrieved from institutional repositories, ensuring that each vocabulary is associated with consistent descriptive information before conversion. Once metadata has been harmonised, a manual alignment step is carried out to associate the structured data files (e.g. CSV, TSV, XLSX) with their corresponding metadata record. This non-programmatic association allows for precise matching between the conceptual scope of the resource and the structure of the tabular content, enabling the construction of an initial SKOS `ConceptScheme`.

Following this alignment, the cleaned structured data is transformed into RDF triples through a mapping-based approach. The pipeline employs human-readable YARRRML[17] specifications to define how tabular elements should be projected onto SKOS properties, and these specifications are then converted into executable RML mappings. The resulting RML files are processed using the `morph-kgc` engine[18], which generates the final Turtle serialisation while ensuring consistency in language tagging, encoding, and URI formation.

Once the RDF has been produced, the resulting dataset is prepared for deployment on standard knowledge-graph infrastructures. Depending on the needs of the data providers, the output can be loaded into a triplestore — such as GraphDB — and exposed through a SKOS - compliant interface such as

[17] https://rml.io/yarrrml/
[18] https://github.com/morph-kgc/morph-kgc

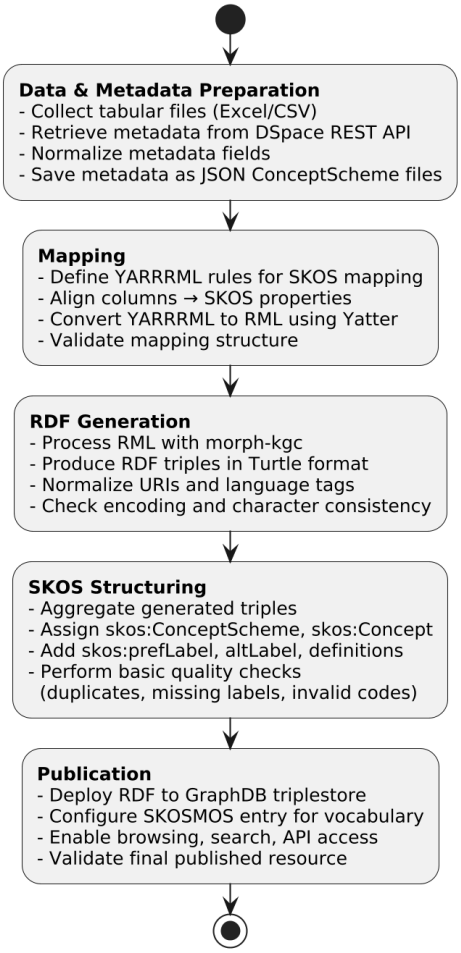

Figure 5: Overview of the SKOS conversion workflow, from data preparation to publication.

SKOSMOS. This final publication stage enables efficient exploration, search, and reuse of the newly created lexical resources, ensuring their integration into broader LLOD ecosystems.

4.2 Data Acquisition and Preparation

This phase of the workflow focuses on gathering, normalising, and structuring the information required to support the creation of SKOS-formatted vocabularies. Data acquisition is carried out in two distinct but complementary steps: (i) the retrieval and preparation of metadata describing the vocabulary resource, and (ii) the ingestion and cleaning of the tabular files containing the actual lexical content. Together, these operations ensure that both descriptive and structural information are consistently represented before the RDF transformation process begins.

4.2.1 Metadata

The first component of the acquisition process concerns the retrieval and preparation of metadata associated with each vocabulary. In our implementation, metadata are sourced from the DSpace repository of ILC4CLARIN through its REST API. The pipeline identifies the target collection (e.g. the REALITER terminology network), retrieves its items, and accesses the corresponding item-level metadata records.

These records include core descriptive elements such as authorship, licensing information, language, date of issuance, subject classification, and persistent identifiers (e.g. Handles). Metadata keys are standardised, multi-valued fields are grouped, and the records are reshaped into structured JSON descriptors.

Particular attention is given to fields such as `dc.language.iso`, which are converted into explicit language objects for later use in the SKOS `ConceptScheme`.

The resulting JSON files act as authoritative descriptors for each vocabulary, ensuring that the subsequent RDF generation is aligned with the metadata deposited in the institutional repository and providing a solid foundation for the remaining phases of the conversion pipeline.

4.2.2 Source Files

The second component of the preparation workflow concerns the ingestion and normalisation of the structured files containing the lexical content of each vocabulary. Unlike metadata — retrieved automatically from the institutional repository — these tabular files are provided separately by data producers and manually uploaded into the Jupyter Notebook environment. Typically distributed in Excel or CSV format, they must be aligned with the previously extracted metadata to ensure a consistent and reproducible transformation into SKOS.

To guarantee a correct association between each tabular dataset and its corresponding metadata record, source files are manually renamed according to the identifier of the `ConceptScheme`. This ensures that metadata JSON files and structured data share a common base name, allowing the notebook to automatically link them during processing. Once harmonised, the structured data undergoes a series of automated normalisation procedures implemented within the notebook.

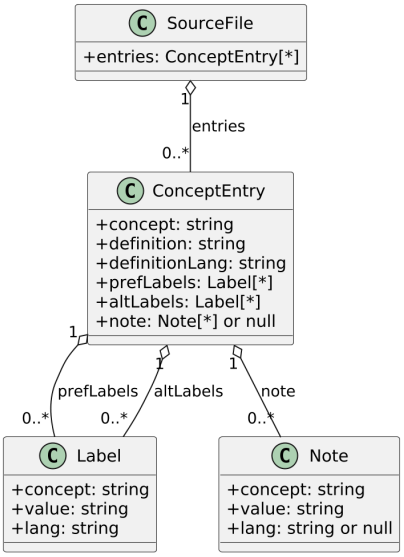

Figure 6: Schema of the JSON used to represent concepts information extracted from the source files.

The notebook includes dedicated routines for cleaning column names, handling language variants, and standardising the dataset structure. Column headers are processed using the `clean_column_name` function, which unifies naming conventions, resolves regional language notations (e.g. `es-MX`, `pt-BR`, `fr-CA`), and removes problematic characters. Each row is then analysed to extract lexical elements such as preferred and alternative labels, definitions, and notes, while concept identifiers are sanitised through `clean_concept_name` to ensure URI-compatible strings.

The processed contents are restructured into a uniform JSON representation including concept identifiers, multilingual labels, definitions, and notes. These intermediate JSON files provide a normalised, machine-readable representation of the vocabulary, enabling consistent handling of heterogeneous datasets and preparing them for the semantic lifting procedures described in the following sections.

4.3 Preparation of the YARRRML Mapping Files

Once the metadata descriptors and the normalised source files have been generated, the next phase of the workflow consists in preparing the YARRRML mapping files that define how input data is translated into RDF triples. In this pipeline, YARRRML acts as an intermediate, human-readable layer that

abstracts the complexity of RML while retaining fine-grained control over the mapping logic. Each mapping file corresponds to a single vocabulary and specifies the rules required to produce a coherent SKOS representation from its metadata and lexical content.

The construction of each mapping file starts by pairing a ConceptScheme descriptor with its corresponding structured data file. This pairing relies on filename alignment: once metadata and source files share a common identifier derived from the repository's persistent URI, they can be reliably processed together. From the metadata, key parameters are extracted, most notably the `dc.identifier.uri`, which is used to derive a custom namespace for the vocabulary. This namespace serves as the base for all generated SKOS concept URIs, ensuring stability and predictability.

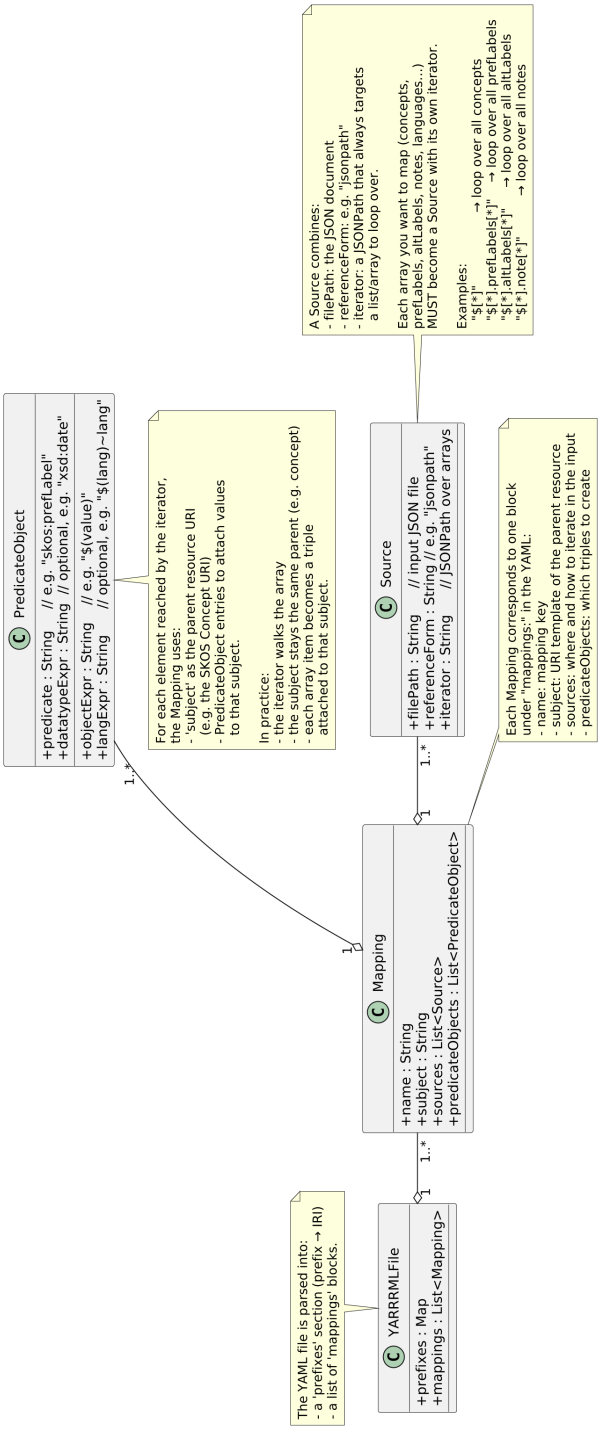

Figure 7: YARRRML Mapping Logic for Concepts.

The mapping file is organised into two main sections. The first defines the ConceptScheme itself by mapping metadata properties — such as title, description, identifier, creation date, authorship, and languages — to Dublin Core and SKOS predicates. This section establishes the structural container for the vocabulary and ensures that provenance, licensing, and linguistic information are correctly represented.

The second section specifies the transformation of the structured data into SKOS concepts. Using JSONPath expressions over the normalised JSON produced in the preprocessing phase, the mapping defines how concepts, labels, definitions, and notes are materialised as RDF triples. For each lexical entry, a URI is generated under the vocabulary namespace and enriched with the appropriate SKOS properties. Language-sensitive fields are handled explicitly, with language tags assigned to labels and notes and inherited by definitions from their source columns.

To ensure interoperability, the mapping includes standard prefixes (e.g. `skos`, `dc`, `dct`, `xsd`) together with a dynamically generated prefix representing the vocabulary namespace. The resulting YARRRML files, saved as individual YAML documents in the `mapping_files` directory, provide a declarative and reproducible specification of the transformation logic. These files are subsequently converted into RML and processed with `morph-kgc` to generate the final RDF representation described in the next section.

4.4 Creation of RML Files

After defining the YARRRML mappings for each vocabulary, the next step of the pipeline consists in translating these human-readable specifications into RML (RDF Mapping Language) files. This phase introduces an intermediate representation that can be directly consumed by RDF generation engines, while preserving the structure and semantics defined in the YARRRML layer.

The conversion is performed using the Yatter library, which takes YARRRML YAML files as input and produces equivalent RML mappings serialised in Turtle syntax. The notebook iterates over all mapping files stored in the `mapping_files` directory, parses each `.yaml` document, and passes it to Yatter for translation. For each vocabulary, a corresponding RML file is generated, mirroring the original mapping rules in terms of sources, iterators, subject templates, and predicate–object definitions.

During this step, the pipeline also normalises the RML namespace by replacing the legacy namespace (`http://semweb.mmlab.be/ns/rml#`) with its updated counterpart (`http://w3id.org/rml/`), ensuring compliance with current best practices and facilitating integration with contemporary tools. The resulting RML documents are stored in the `rml_files` directory, preserving the base filename of the original YARRRML file and adopting the `.rml` extension.

By separating mapping authoring (YARRRML) from execution (RML), this phase increases transparency and modularity. The resulting RML files provide a stable and reusable specification that can be repeatedly applied to the prepared JSON sources to generate RDF triples in a controlled and reproducible manner, as described in the subsequent phase of the pipeline.

4.5 Creation of RDF Files

The final stage of the transformation pipeline consists in generating RDF data from the RML mapping files created in the previous step. To perform this operation, we rely on the `morph-kgc` library, a Python framework capable of executing R2RML, RML, and RML-star mappings and materialising the corresponding RDF triples. This step produces the final, serialised RDF representation of each vocabulary, ready to be published and ingested into standard Linked Data platforms.

4.5.1 RDF Generation Workflow

The creation of RDF files follows a unified workflow that combines configuration, execution, and serialisation into a single processing phase. First, a configuration string is generated for each RML file, specifying the input mapping, the desired output format, and the destination directory for storing the resulting RDF data. This dynamic configuration replaces the need for a static `config.ini` file and ensures consistent processing across all vocabularies.

Once the configuration is prepared, the notebook iterates over the RML files contained in the `rml_files` directory. For each mapping, the `morph-kgc` `materialize()` function is invoked to

execute the RML rules. During this process, iterators defined in the mapping files identify the relevant elements in the structured JSON data, subject templates are expanded to generate IRIs for SKOS concepts, and predicate–object patterns produce the associated labels, definitions, and notes.

After the triples are materialised, the resulting `rdflib` graph is enriched with a predefined set of prefixes (e.g. `dc`, `dct`, `skos`) and a dynamic prefix derived from the mapping filename. This dynamic namespace corresponds to the final vocabulary URI and ensures that all IRIs in the output adhere to the expected namespace pattern. Finally, the enriched graph is serialised in Turtle format and saved to the `rdf_files` directory, yielding the complete and FAIR-compliant RDF dataset for each vocabulary.

4.5.2 Aggregating Concepts into the ConceptScheme

Once the RDF files have been generated from the RML mappings, each Turtle file contains two distinct types of resources: the `skos:ConceptScheme` describing the vocabulary as a whole, and the individual `skos:Concept` entries representing the terms defined within it. In their initial form, however, these resources remain unconnected. Establishing explicit links between the `ConceptScheme` and its concepts is therefore a necessary post-processing step to ensure the semantic completeness and navigability of the resulting SKOS resource.

This aggregation step relies on two complementary SKOS properties:

- **`skos:hasTopConcept`** — associates a `skos:ConceptScheme` with each of its top-level concepts.
- **`skos:topConceptOf`** — establishes the inverse relationship, linking each `skos:Concept` back to its parent `ConceptScheme`.

The procedure operates over the set of Turtle files generated in the previous step. For each file, the script identifies the unique `skos:ConceptScheme` resource it contains, extracts all the instances of `skos:Concept`, and systematically adds the bidirectional relationships `skos:hasTopConcept` and `skos:topConceptOf`. The enriched RDF graph is then serialized back into Turtle format, replacing the original file. This ensures that all vocabularies produced by the pipeline expose the hierarchical structure required for correct visualisation and browsing in SKOS-oriented platforms such as SKOSMOS.

4.5.3 Normalising Character Encoding in Concept URIs

A final refinement step is applied to ensure that all URIs contained in the generated RDF files are human-readable and free from percent-encoded characters. During previous phases, concept identifiers originating from tabular sources may contain accented or non-ASCII characters, which can become percent-encoded in the form of sequences such as `%C3%A9` for `é`. While these encodings are technically valid, they reduce readability and hinder the use of the resulting SKOS resources in user-facing environments.

To address this issue, each Turtle file produced in the RDF generation stage is scanned for URIs containing percent-encoded segments. The normalisation process proceeds as follows:

- The file is parsed into an `rdflib` graph, preserving all existing namespace declarations.
- Each triple is inspected, and both subject and object nodes are checked for the presence of percent-encoded characters.
- When such patterns are detected, the URI is decoded into its UTF-8 representation using standard URL-decoding rules.
- The triple is then reconstructed using the cleaned URI, and added to a new RDF graph that retains the original prefixes.

Once all triples have been processed, the updated graph is serialised back into Turtle format and written to disk, replacing the original file. This ensures that all concept URIs are expressed in a normalised and readable form, improving interoperability and guaranteeing that multilingual labels encoded within identifiers (e.g. concept names derived directly from tabular structures) are faithfully represented in the final SKOS dataset.

4.6 Configuring SKOS Vocabularies in SKOSMOS

Once the SKOS vocabularies have been generated as RDF and loaded into the GraphDB repository, the final step consists in exposing them through the SKOSMOS web interface. This is done by adding one configuration block per vocabulary in the SKOSMOS configuration file (`config.ttl` or its deployment-specific variant). Each block defines a `skosmos:Vocabulary` resource associated with the corresponding named graph in GraphDB, allowing SKOSMOS to provide browsing, search functionalities, and API access for that specific lexicon.

4.6.1 Vocabulary Declarations

SKOSMOS requires each lexicon to be declared as an instance of both `skosmos:Vocabulary` and `void:Dataset`. In the Pan-Latin pipeline, each vocabulary receives an identifier derived from the internal handle-based name. For example, a vocabulary identified internally as `20_500_11752_OPEN_1014` is represented in the configuration as follows:

```
:pl_20_500_11752_OPEN_1014 a skosmos:Vocabulary, void:Dataset ;
    dc:title "Pan-Latin Office Supplies Vocabulary"@en ;
    skosmos:shortName "Pan-Latin Office Supplies Vocabulary" ;
    dc:subject panlatin:LexiconCat ;
    void:uriSpace
        "https://vocabs.ilc4clarin.ilc.cnr.it/vocabularies/pl_20_500_11752_OPEN_1014/" ;
    skosmos:language "it", "en", "fr", "es-AR", "pt-BR" ;
    skosmos:defaultLanguage "it" ;
    skosmos:showTopConcepts "true" ;
    void:sparqlEndpoint
        <https://vocabs.ilc4clarin.ilc.cnr.it/graphdb10/repositories/h2iosc_skosmos> ;
    skosmos:sparqlGraph
        <https://vocabs.ilc4clarin.ilc.cnr.it/vocabularies/pl_20_500_11752_OPEN_1014/> .
```

This template is reused for all vocabularies managed by the pipeline. The configuration block mirrors the URI pattern already adopted for publishing the RDF data: each lexicon is stored in its own named graph (`https://vocabs.ilc4clarin.ilc.cnr.it/vocabularies/pl_{`*vocab_name*`}/`), and the same URI is referenced in `skosmos:sparqlGraph`, ensuring that SKOSMOS directs all SPARQL queries to the appropriate dataset. The descriptive fields then supply the interface with meanin gful labels, indicate the URI namespace for concepts, define the available languages, specify the default language, and connect the vocabulary to the shared SPARQL endpoint.

To group all Pan-Latin lexicons under a dedicated section in the interface, a custom category is introduced:

```
@prefix panlatin: <http://vocabs.ilc4clarin.ilc.cnr.it/panlatinlexicon> .

panlatin:LexiconCat a skos:Concept ;
    skos:prefLabel "REALITER - OTPL Pan-Latin Lexicons"@it,
                   "REALITER - OTPL Pan-Latin Lexicons"@en ;
    skos:definition
        "Category for grouping the Pan-Latin lexicons section"@en .
```

Each vocabulary is linked to this category via the `dc:subject` property, enabling SKOSMOS to display them together as a coherent thematic group.

4.6.2 Automatic Generation of Vocabulary Blocks

To avoid maintaining configuration blocks manually and to keep SKOSMOS aligned with the contents of the repository, all vocabulary declarations are produced automatically from the metadata and intermediate files generated by the pipeline. The script responsible for this task analyses the concept-scheme files and the corresponding structured data in JSON format, derives the vocabulary identifier, retrieves

the title defined in the concept scheme, identifies all languages present in the labels and notes, and assembles a complete configuration block following the structure illustrated above. The resulting Turtle file (for example, `vocabs_data.ttl`) contains one declaration per vocabulary. Since its prefix declarations match those already present in the main SKOSMOS configuration, the operator simply appends the generated vocabulary blocks to `config.ttl`. After SKOSMOS is reloaded, each Pan-Latin lexicon becomes immediately available for browsing, consistently linked to its GraphDB named graph and integrated within the URI design defined earlier in the workflow.

5 Conclusions and Future Work

The methodology proposed in this article provides a structured approach for transforming linguistic data into SKOS resources, using Python notebooks to support data conversion and training. The pipeline integrates data acquisition, transformation, and publication, demonstrating its effectiveness in generating high-quality LLOD resources. The integration of these resources into a SKOSMOS-based platform enhances their accessibility and practical utility. Applying this methodology to Pan-Latin vocabularies has shown its ability to handle complex, multilingual data. Integrating these resources into the institutional vocabulary platform increases their visibility and supports reuse within the Linked Open Data community, both through a web interface and REST APIs.

Future developments will include adopting the Mapheator tool[19] to simplify mapping rule creation in formats like R2RML, RML, and YARRRML. The spreadsheet-based approach will make mapping more accessible for non-expert users and streamline the transformation process. The training materials from this project will be integrated into the H2IOSC training platform[20], which is currently being developed to support users in replicating the workflow and adopting LLOD best practices.

References

Berners-Lee, T. (2006). Linked open data design issues. https://www.w3.org/DesignIssues/LinkedData.html

Fiorelli, M., & Stellato, A. (2021). Lifting tabular data to RDF: A survey. In *Metadata and semantic research* (pp. 85–96). Springer International Publishing.

Gilardoni, S. (2011). I lessici della Rete panlatina di terminologia. In *Terminologie specialistiche e prodotti terminologici* (pp. 101–112). EDUCatt.

Grimaldi, C., Marzi, E., Puccini, P., Zanola, M. T., Zollo, S., et al. (2022). *Terminologia e interculturalità. problematiche e prospettive*. I libri di Emil.

Grimaldi, C., & Zanola, M. T. (2021). *Terminologie e vocabolari: Lessici specialistici e tesauri, glossari e dizionari*. Firenze University Press.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., Pagé-Perron, É., Passarotti, M., Muñoz, S. R., & Truică, C.-O. (2022). When linguistics meets web technologies. recent advances in modelling linguistic linked data (P. Cimiano, J. Bosque-Gil, P. Cimiano, & M. Dojchinovski, Eds.). *Semantic Web*, *13*(6), 987–1050. https://doi.org/10.3233/sw-222859

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P. A., et al. (2020). Towards fair principles for research software. *Data Science*, *3*(1), 37–59.

Lo Presti, M. V., & Dankova, K. (2021). Trattamento della terminologia culturale in una prospettiva multilingue. Il caso del *Lexique panlatin de la mobilité étudiante*. *AIDAinformazioni*, *39*(3-4), 45–66.

Zanola, M. T. (2014). Le réseau Realiter, un acteur du plurilinguisme. *PLAISANCE*, *11*(32), 149–165.

[19] https://github.com/oeg-upm/mapeathor

[20] https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it/

Govorjena slovenščina: A Platform for the Structured Collection of Conversational Speech

Darinka Verdonik, Andreja Bizjak, Gregor Donaj
University of Maribor, Slovenia
{darinka.verdonik,andreja.bizjak1,gregor.donaj}@um.si

Abstract

Conversational spoken language resources remain underrepresented despite their importance for linguistic research and speech technology development. Their collection is constrained by methodological, technical, and legal challenges, particularly in the case of private, everyday interaction. This paper presents Govorjena slovenščina, a CLARIN.SI service developed to support the systematic collection, transcription, and archiving of spoken Slovene. The platform integrates standardized workflows for consent management, metadata collection, data submission, and transcription, and is designed as a reusable research infrastructure rather than a project-specific collection tool. Drawing on insights from citizen-science and crowdsourcing initiatives, the paper discusses key design choices related to participant motivation, data quality, and legal governance, and reports on initial corpora contributions enabled by the platform. The approach demonstrates how a standards-driven, legally grounded infrastructure can support the sustainable development of conversational speech resources, particularly for smaller or under-resourced languages.

1 Introduction

Spoken language resources remain considerably less developed than written ones, especially with regard to data representing everyday human–human interaction. Nevertheless, spoken data are indispensable for achieving comprehensive and empirically grounded linguistic descriptions, as they make it possible to observe variation, spontaneity, and context-sensitive language use that are largely absent from written texts. Analyses of spoken interaction provide crucial theoretical insights into how speakers manage turns, negotiate meaning, coordinate social actions in real time, etc. In addition, spoken corpora play a key role in the advancement of speech technologies, most notably automatic speech recognition and dialogue systems, which rely on authentic spoken input to function robustly in natural communicative settings. The continued scarcity of high-quality spoken resources therefore limits both linguistic theory-building and the development of language technologies capable of handling informal, interactional speech.

Substantial quantities of spoken data can be obtained from readily available sources such as parliamentary proceedings, broadcast media, or online platforms including YouTube. While these resources are valuable, they cannot replace a broad and representative range of private data, as speech practices differ markedly across communicative domains. Previous research has demonstrated that conversational data are in higher demand and used more frequently in linguistic research than context-governed data (Love et al., 2017). Despite this demand, conversational corpora remain scarce as they are among the most difficult resources to compile. Their collection requires considerable organizational effort, including the recruitment and coordination of a wide network of field collaborators and volunteer speakers. Moreover, potential participants may be reluctant to consent to the recording of their speech for personal or privacy-related reasons. The processing and storage of spoken data also present significant challenges, as they necessitate transparent data management policies and strict adherence to personal data protection regulations. Further complications arise from technical factors affecting recording quality, such as background noise, variability in recording equipment, and uncontrolled recording environments. Finally,

Darinka Verdonik, Andreja Bizjak and Gregor Donaj 2025. Govorjena slovenščina: A Platform for the Structured Collection of Conversational Speech. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 25–38. DOI: https://doi.org/10.3384/ecp222

achieving adequate speaker diversity with respect to age, gender, dialectal background, and sociolectal variation is essential for representativeness, yet remains difficult in practice, further constraining the development of balanced conversational language resources.

This paper presents an online platform Govorjena slovenščina (Spoken Slovene)[1], launched in March 2025 as part of the LLM4DH project and financed by the Slovenian Research and Innovation Agency. The platform is designed to address the challenges of conversational data collection by providing scenarios and guidelines for field recording of conversational data, consent forms for speakers to allow the use of recordings, standardized forms for documenting information about recordings and speakers, remote submission of the data, secure and long-term archiving of the data, and unified guidelines for transcribing speech in Slovene. The platform is offered as the CLARIN.SI service for speech data collection, and is provided by the University of Maribor, Faculty of Electrical Engineering and Computer Science as the CLARIN.SI consortium partner.

This paper focuses on challenges related to creating the Govorjena slovenščina platform, which are: (1) Resolving legal aspects of conversational speech data collection and distribution, (2) Attracting participating individuals, (3) Defining scenarios for who, where, what, and how to record, (4) Standardizing the metadata on speakers and recordings, (5) Facilitating recording submission and data management and (6) Standardizing and supporting the transcription processes digitally.

2 Citizen Science and Crowdsourcing in the Collection of Spoken Language Resources

Traditional laboratory and field-based approaches remain essential for producing high-quality spoken language data, yet they are difficult to scale when the research objective is to document broad social and regional variation in everyday speech. In particular, the collection of spontaneous interaction is inevitably affected by the observer's paradox, as formulated by William Labov (Labov, 1972): speakers tend to adjust their language use once they become aware that they are being recorded. This effect persists regardless of whether recordings are made by researchers, by speakers themselves, or through online interfaces, and thus cannot be fully eliminated by changing the recording setting alone.

In response to these constraints, increasing attention has been paid to remote and large-scale data collection strategies based on citizen science and crowdsourcing. Enabled by widespread access to digital technologies and mobile devices, these approaches make it possible to reach large and diverse speaker populations at comparatively low cost. At the same time, crowdsourcing is not a monolithic method: it encompasses paid microtask platforms, volunteer-based citizen science initiatives, and alternative incentive models such as Games-With-A-Purpose (GWAP) and Collect4NLP. Each of these models differs with respect to sustainability, participant motivation, cost structure, and the degree of control over data quality (Lyding et al., 2022).

Moving speech collection online, however, introduces new design challenges (Eskenazi et al., 2013). The potential to reach many contributors must be balanced against risks related to authenticity, representativeness, and technical quality. Task design is particularly critical: insufficient guidance may result in unusable recordings, while overly detailed instructions can unintentionally shape participants' speech and reduce natural variation. Successful initiatives therefore tend to combine simple, user-friendly interfaces with carefully calibrated instructions, basic quality-control mechanisms, and transparent communication about data use.

To identify effective strategies and transferable design principles, a review of 45 crowdsourcing-based speech collection initiatives was conducted (Bizjak, 2025). On the basis of contributor scale, collection goals, targeted speech types, and other, three European projects were selected for closer analysis: CorCenCC (Welsh), Anneta kõnet (Estonian), and Lahjoita puhetta (Finnish). These projects illustrate how technical solutions, legal and ethical frameworks, and motivation strategies can be combined to support large-scale collection of spoken language data, providing valuable reference points for the design of the Govorjena slovenščina platform.

CorCenCC (2016–2020) represents one of the most comprehensive national corpus-building efforts for a lesser-resourced language, collecting over 11 million words in total, including 2.8 million words of spo-

[1] https://govorjena-slovenscina.um.si/

ken language, of which approximately 240,000 words were recorded in private, informal settings (Knight et al., 2020). A distinctive feature of CorCenCC is its "bottom-up" corpus design, aligned with the participatory ethos of citizen science. Participation was supported through a dedicated mobile application for iOS and Android, as well as a web interface, enabling contributors to submit recordings together with consent documentation and sociolinguistically relevant metadata (Knight et al., 2001; Neale et al., 2017). The project's emphasis on detailed speaker profiling supported representativeness and balancing goals, while its governance model foregrounded transparency and participant control over contributions. Notably, the project also faced challenges common to smaller language communities: despite extensive media presence and outreach, private, informal recordings proved difficult to obtain, plausibly due to heightened concerns about identifiability in a relatively small speech community (Knight et al., 2020; Neale et al., 2017). This illustrates an important general lesson for spontaneous speech collection: privacy concerns can be a stronger limiting factor than technical barriers, and trust-building measures (clear anonymization procedures, accessible documentation, and repeated explanation of safeguards) are not optional add-ons but central infrastructure.

Anneta kõnet (2022–2023) adopted a donation-based model and collected over 100 hours of Estonian spontaneous speech. A key insight from this campaign is the value of minimizing friction in participation: contributors could simply select a topic and begin speaking, including the option to speak freely without predefined prompts [2][3]. The campaign also used promotional messaging tailored to different target groups—framing participation for younger speakers through the lens of speech technologies in everyday life, and for older speakers through language preservation and cultural identity[4]. This dual framing aligns with broader citizen-science findings that motivations are heterogeneous and that recruitment strategies must be audience-specific rather than generic. It also demonstrates how "purpose communication"—explaining how data will be used and why it matters—can function as a motivational lever and as a trust mechanism.

Lahjoita puhetta (2020–2021) demonstrates the scalability of large-scale remote collection, gathering more than 4,000 hours of Finnish speech within a few months (Lindén et al., 2022). The campaign relied on both web and mobile channels, but it also made a design decision with methodological consequences: to reduce technical complexity and improve signal quality, it prioritized recordings with a single speaker rather than multi-party conversational interaction. This choice highlights a recurring trade-off: multi-speaker conversation may be more interactionally "natural," but it is also harder to capture with adequate audio quality and clean channel separation in uncontrolled environments. Lahjoita puhetta therefore pursued naturalness through task prompts that elicited spontaneous responses without encouraging read speech, while simplifying the acoustic capture problem. Notably, nearly 90 percent of the 220,000 recordings ranged from 10 seconds to 3 minutes, with an average duration of 30–60 seconds (Lindén et al., 2022). This distribution offers an important operational lesson (Bizjak, 2025): short, low-effort contributions can accumulate into very large datasets, particularly when the interface is simple and the campaign is strongly visible. At the same time, short clips may limit discourse-analytic uses and may not fully mitigate the observer's paradox, reinforcing the need to align recording formats with research goals (e.g., ASR training vs. interactional linguistics).

Across these case studies, several cross-cutting challenges emerge that are directly relevant to designing sustainable systems for collecting spontaneous speech. First, motivation must be treated as a design problem rather than a "nice-to-have" consideration: why would someone participate, and why would they return (Rutten et al., 2017)? A strong argument in the literature (Eskenazi et al., 2013) is that if payment is used, it should be fair and transparent; token payments can backfire, reduce trust, and complicate ethical governance. Volunteer-based approaches, in turn, require credible purpose communication and visible evidence that contributions are used and valued, not merely collected (Rutten et al., 2017).

Second, quality control requires multiple layers (Eskenazi et al., 2013). Remote speech collection is vulnerable to trivial failures (silent recordings, wrong microphone settings), environmental noise, device

[2]https://www.kratid.ee/en/keeletehnoloogia-projektid

[3]https://laecwador.ee/en/projekt/donate-a-speech-campaign

[4]https://laecwador.ee/en/projekt/donate-a-speech-campaign/

variability, and occasional bad-faith contributions (e.g., bots or "fast earners" in paid settings). Best practice therefore includes: (i) clear, concise recording guidance; (ii) lightweight "pretesting" or sample checks; (iii) systematic validation before or after submission (including random sampling and gold-standard comparisons where applicable); and (iv) metadata capture sufficient for diagnosing and filtering problematic recordings. Technical recommendations—such as using lossless formats, adequate sampling rates, and minimizing microphone distance—are important, but they must be translated into user-friendly instructions that do not overwhelm contributors.

Third, legal and ethical governance is foundational for speech resources because speech is a biometric identifier and thus personal data in the GDPR sense. Projects must establish a lawful basis (often consent, sometimes legitimate interest), provide clear privacy policies, and implement anonymization practices (Bizjak, 2025). They must also anticipate downstream access conditions and licensing constraints: permissive licensing supports reuse (including potential commercial applications), but it requires that participant consent and rights management are robust from the outset. Importantly, the case studies show that legal compliance alone is insufficient: participants must understand the governance model in accessible language, and they must feel confident that their contributions will not expose them or others (Knight et al., 2020; Neale et al., 2017).

Taken together, CorCenCC, Anneta kõnet, and Lahjoita puhetta demonstrate how crowdsourcing framework can address traditional bottlenecks in speech data collection—particularly for under-resourced languages—while also revealing the methodological compromises such systems often require. The most successful initiatives treat the platform not merely as a technical upload interface, but as an integrated socio-technical system: it must balance user experience, motivation, trust, legal safeguards, and quality control, while remaining aligned with the linguistic or technological goals of the resulting corpus. These lessons are directly transferable to the Slovene context, where the long-term challenge is not only to collect "more speech," but to collect speech that is sufficiently diverse, spontaneous, well-documented, and ethically governed to support both linguistic research and robust speech technologies.

In contrast to the large-scale initiatives discussed above—each supported by national institutions, dedicated funding, and extensive media campaigns—Govorjena slovenščina is conceived as a comparatively lightweight academic infrastructure. The project does not aim to compete with national donation campaigns in terms of volume or visibility, nor does it have resources for mobile application development or sustained public promotion. Instead, its primary focus lies in establishing robust legal frameworks, developing reusable technical infrastructure, maintaining unified standards for metadata and transcription, and exploring intrinsic motivational factors that may support voluntary participation in smaller-scale research-driven data collection. The portal is designed as an open service that can be used by individual researchers or project teams from different institutions who wish to rely on its workflows and guidelines for their own speech collection efforts. The project team draws not only on the insights gained from the literature review presented above, but also on its extensive practical experience in speech data collection, accumulated over more than two decades of work on Slovene speech corpora. This experience includes, in particular, the development of the reference speech corpus GOS (Verdonik et al., 2013) and the ARTUR corpus (Verdonik, Bizjak, et al., 2024) dedicated to the development of automatic speech recognition for Slovene.

Govorjena slovenščina does not provide a dedicated recording application; participants use their own recording devices and freely available recording software, and upload the resulting files through the web interface. This design choice deliberately shifts complexity away from custom software development toward standardized procedures and documentation. To support data diversity and prevent overrepresentation of individual speakers, contributions are limited to a maximum of two hours per speaker. At the same time, the platform prioritizes longer recordings, with a minimum duration of five minutes and a maximum of ninety minutes, in order to allow conversations to unfold naturally, reduce initial self-monitoring effects, and provide sufficient contextual depth for linguistic analyses that extend beyond acoustic or lexical levels, including semantic, pragmatic, and interactional research. Furthermore, the platform is conceived not as a project-specific solution, but as a reusable research infrastructure that can be adopted by individual researchers or research teams across institutions for a wide range of speech data collection purposes,

including dialectological, sociolinguistic, pragmatic, lexical, phonetic, or interactional studies, as well as research on speech development, speech disorders, second language acquisition, human–computer interaction, and similar.

3 Platform Govorjena slovenščina

This section outlines solutions for key challenges in creating the Govorjena slovenščina portal.

3.1 Legal Aspects

Legal experts in personal data protection from APHAIA company [5] were engaged to design a comprehensive personal data protection protocol in full compliance with European Union legislation, most notably the General Data Protection Regulation (GDPR). Based on their legal assessment, speech recordings of identifiable individuals were classified as personal data, establishing informed consent from the speaker as the most appropriate and transparent legal basis for data processing. The experts further determined that the processing of users' personal data for the purposes of incentives or rewards must rely on a legal basis separate from that governing the collection and processing of speech recordings themselves. In addition, they concluded that the spontaneous speech recordings collected through the portal are unlikely to qualify as copyrighted works, as they typically lack the level of originality required for copyright protection under applicable law.

These legal principles were implemented within the Govorjena slovenščina portal through a set of clearly defined legal and procedural documents. The legal experts prepared a detailed Privacy Policy Statement [6], which transparently informs users about the scope, purpose, storage, and processing of personal data. In parallel, contributors are required to accept a Consent for the Publication of Speech Recordings and Related Personal Data, which explicitly covers the recording, storage, annotation, and research use of the donated speech data. In addition, the Terms of Contribution of Recordings and Contest Participation regulate the conditions under which users may submit recordings and, where applicable, participate in incentive-based activities hosted on the portal. All documents are made accessible directly through the portal interface, ensuring that contributors are fully informed prior to participation. Before publication, all data are anonymised by removing personal identifiers. Speakers' names and surnames are deleted from metadata, ensuring that individuals cannot be easily identified.

3.2 Attracting Participating Individuals

As the Govorjena slovenščina portal lacks the financial resources to support a large-scale marketing campaign, particular emphasis has been placed on identifying and fostering intrinsic motivational factors that encourage sustained participant engagement. The portal homepage therefore foregrounds motivations that appeal to contributors' sense of social value and personal relevance, such as the opportunity to donate speech in order to preserve samples of Slovene and its regional varieties as part of the national linguistic and cultural heritage. Additional emphasis is placed on supporting the development of digital language technologies for Slovene and its dialects, as well as contributing to research on child speech and language development. To complement these intrinsic incentives, the portal also offers modest but tangible rewards, including items such as memory cards or headphones, which are awarded to participants who submit recordings that meet quality criteria and reach a cumulative duration of 50 minutes. This hybrid approach balances voluntary participation with light extrinsic encouragement, while avoiding the pressures associated with monetary compensation.

A second key focus of the portal is the establishment of trust between contributors and the research team, as trust is a prerequisite for individuals to feel comfortable donating personal speech data. The portal therefore provides transparent and accessible information on how recordings will be processed, stored, anonymized, and used for research and technological development. Clear explanations of data protection measures are accompanied by information about the institutions and individuals responsible for maintaining the portal. In addition, the portal publishes popular-science articles and explanatory texts

[5] https://aphaia.co.uk/

[6] https://govorjena-slovenscina.um.si/uporaba-posnetkov/

about spoken language and speech research, which serve both to disseminate knowledge and to reinforce the public-interest orientation of the project.

Third, the portal places strong emphasis on the clarity of task descriptions in order to minimize uncertainty, which is a known barrier to participation in crowdsourced data collection. Contributors are provided with concise and comprehensible instructions that guide them through the recording process step by step. To further support participation, the portal outlines a range of possible recording scenarios and suggests concrete speech topics, helping contributors understand what kinds of recordings are desired while still allowing for spontaneity and naturalness in speech production. These instructions are complemented by an explanatory video with illustrative example that demonstrates the recording procedure in a clear and accessible manner.

Finally, considerable attention has been devoted to the development of a clear, user-friendly, and visually appealing web interface with an engaging and playful aesthetic. The design incorporates illustrative graphics and examples of authentic speech drawn from previously collected conversational data. In addition, the portal features two introductory videos that serve complementary communicative functions: the first video is designed to attract and motivate potential participants by presenting the initiative in an accessible and engaging manner; the second video explains the importance, intended use, and broader societal and scientific value of the collected speech data, while the third video, as outlined above, provides a step-by-step demonstration of how to create and submit a recording. Together, these visual elements support user engagement, enhance transparency, and contribute to a more intuitive and reassuring user experience.

The usability of the platform has been critically evaluated by experts in linguistics and design using a heuristic evaluation based on Nielsen's usability principles (Nielsen, 1994), systematically assessing interface design, feedback mechanisms, consistency, and error prevention. Based on this evaluation, iterative improvements to both the interface and the task workflow were implemented.

3.3 Recording Scenarios and Metadata Standardization

Conversational data are often categorized in a relatively limited and abstract manner. Existing classification schemes typically distinguish between parameters such as direct versus distanced contact, dialogue versus multilogue, and broad domains including face-to-face conversation, telephone interaction, interviews, or business transactions (Oostdijk, 2000). In the Slovene reference corpus GOS (Verdonik et al., 2013), conversational data were further specified through domains such as conversations among friends or within families. While such typologies are useful for corpus design and analysis, they offer limited practical guidance to contributors engaged in data donation, particularly with respect to what constitutes an appropriate conversational form, level of spontaneity, or choice of topics.

To address this limitation, the Govorjena slovenščina platform places strong emphasis on providing concrete, illustrative guidance for participants, both in terms of speech genres and topic selection. Contributors are encouraged to produce recordings that fall within a set of clearly defined and easily recognizable genres, including interviews, discussions, instructions, narrations, explanations, and informal or casual conversations. These genres are described through short, example-based explanations grounded in everyday communicative situations, using non-technical language and avoiding specialist linguistic terminology. They are selected to capture a broad spectrum of communicative functions, such as information exchange, opinion expression, procedural discourse, and narrative structuring. In addition, the platform proposes a wide range of everyday topics that reflect natural spoken interaction, such as hobbies and leisure activities, sports, music, travel experiences, technology use, food and cooking, health and well-being, environmental issues. Importantly, contributors are explicitly advised to select only topics they feel comfortable sharing publicly, thereby reinforcing awareness of the public and research-oriented nature of the recordings and supporting informed participation.

Special attention is also devoted to technical and procedural guidelines aimed at ensuring adequate recording quality without imposing excessive technical demands on contributors. The platform provides step-by-step instructions on how to prepare for recording, including recommendations to choose a quiet environment, reduce background noise (e.g. television, traffic, or other speakers), and position the record-

ing device at an appropriate distance from the speaker's mouth. Guidance is given on using common devices such as smartphones or laptops, adjusting basic recording settings, and avoiding handling noise. To further reduce uncertainty, contributors are offered the possibility to make and submit a short test recording before donating longer samples. Together, these measures are designed to balance recording quality with accessibility, enabling contributors with varying levels of technical expertise to participate while producing speech data suitable for linguistic analysis and technological applications.

Participating citizens are required to collect metadata on both speakers and recorded events alongside the recordings. The categorization of these metadata is based on the Slovene reference speech corpus GOS 2.0 (Verdonik, Dobrovoljc, et al., 2024), with additional metadata fields incorporated to address requirements identified across the various disciplines that make use of such data. The final lists of speaker metadata and recording metadata are provided in Appendices B and C.

3.4 Online interface for data submission

The submission of recordings is facilitated through an online interface, which categorizes users into three distinct roles: participating citizens, administrators, and transcribers. Participating citizens–individuals submitting recordings–are required to register before contributing. For each submission, they must complete metadata forms detailing information about both the speaker and the recorded speech, upload an image of the speaker's signed consent, and submit the recording. The administrators are responsible for overseeing the data and recordings. They have the ability to download and approve submissions, edit the metadata associated with recordings and speakers, and export summarized spreadsheet data pertaining to registered users, recordings, and speakers. Transcribers are tasked with processing the submitted recordings. They can download audio files and upload transcriptions, which may be provided in both literal and standardized forms as separate entries.

The online interface for data submission is implemented as a web application based on the Flask framework and is hosted under the official domain of the University of Maribor,[7] ensuring institutional reliability and long-term maintainability. Flask was selected due to its modular architecture, flexibility, and suitability for developing lightweight yet scalable web services. The application integrates a relational database that stores all user-related data, including user profiles, consent records, metadata associated with individual recordings, and system logs. To ensure transparency and traceability, the system implements comprehensive versioning mechanisms: all edits to metadata or user-submitted information are recorded in change log files, allowing complete backtracking and auditability of modifications over time.

Speech recordings and other uploaded files are stored securely and separately from the main application database on the server, which reduces the risk of data corruption and contributes to overall system robustness and maintainability. This architectural separation also facilitates more fine-grained access control and supports compliance with data protection and security requirements, particularly with respect to personal and sensitive data contained in speech recordings. The platform architecture has been designed with scalability in mind, allowing for the future integration of additional annotation layers, processing modules, and corpus-specific workflows. A graphical diagram of the online interface is provided in the appendix A.

Practical use of the portal during its first year of operation has revealed several limitations of the existing administrative interface. Originally designed primarily for internal use, the interface does not sufficiently support efficient data management in scenarios involving broader participation by external users. In particular, the need has emerged for clearer separation of data, a more flexible access management system across different corpora, users, and research projects, as well as a configurable metadata framework capable of accommodating different file types and corpus designs. To address these challenges, a comprehensive upgrade of the administrative interface, underlying data structures, and core functionalities is planned.

The planned upgrade will significantly enhance system stability, accessibility, and reliability, while enabling administrators to flexibly define file types, metadata schemas, and input forms directly through

[7] https://govorjena-slo-portal.ietk.um.si/

the web interface, without requiring technical intervention at the code level. This functionality will allow corpus administrators without advanced technical expertise to manage and adapt the platform independently. The upgrade includes the implementation of a full-featured system for managing users, projects or corpora, and uploaded files, as well as mechanisms for assigning access rights based on corpus affiliation and user roles. Extended functionalities will support secure uploading, editing, and storage of recordings, alongside reinforced safeguards for the protection of personal data.

In addition, the upgrade process will encompass systematic testing of the platform, including evaluations with external users, and the preparation of clear and comprehensive user documentation to ensure a user-friendly experience and reliable operation. The resulting infrastructure will establish a solid technical foundation for the sustainable development of spoken language resources, their long-term research and pedagogical use, and increased user involvement in data collection and management processes. In line with principles of openness and reproducibility, the full source code of the platform will be made publicly available upon completion of the project through the CLARIN.SI GitHub repository,[8] thereby enabling reuse, inspection, and further development by the wider research community.

3.5 Transcription standardization and digital support

The transcription guidelines applied in earlier Slovene speech collection initiatives (Verdonik, Bizjak, et al., 2024; Verdonik, Dobrovoljc, et al., 2024; Verdonik et al., 2013) have been systematically revised, updated, and consolidated on the Govorjena slovenščina platform.[9] The guidelines are presented in a clearly structured and accessible format, designed both for experienced transcribers and for contributors with limited prior exposure to speech transcription. They are organized into thematically coherent sections covering transcription tools, segmentation principles, literal transcription, standardized transcription, punctuation, transcription-specific conventions, additional annotation marks, and common transcription errors. This modular structure allows users to easily navigate between general principles and more detailed instructions, thereby supporting consistency and transparency in transcription practices.

The platform recommends the use of established transcription tools such as Transcriber (Barras et al., 2001) and EXMARaLDA (Schmidt & Wörner, 2014), both of which support time-aligned transcription and are well suited for spoken corpus compilation. Segmentation into utterances is explicitly advised in order to enhance readability and analytical clarity. Utterances are defined as semantically coherent units, typically corresponding to statements or sentences, rather than being determined solely by prosodic or syntactic criteria. This approach facilitates subsequent linguistic annotation and analysis across different levels. In cases of overlapping speech, transcription is carried out at the word level, and overlapping segments are separated regardless of higher-level utterance boundaries, allowing precise representation of interactional dynamics.

The transcription standard follows a well-established tradition in Slovene spoken language corpora and is based on a dual-mode approach. Literal transcription requires that words be transcribed exactly as pronounced, including reductions, assimilations, and non-standard forms typical of spontaneous speech. In parallel, standardized transcription provides a normalized representation in which each word form is rendered in its standard orthographic form, enabling comparability with written corpora and supporting downstream applications such as lexicographic analysis and language technology development (Verdonik, Bizjak, et al., 2024). Both transcription layers are aligned at the word level, ensuring systematic correspondence between spoken realization and standardized representation.

Punctuation is obligatory and follows conventional written Slovene usage, with sentences beginning with uppercase letters. However, punctuation is applied with sensitivity to spoken discourse structure rather than strict written syntax. The guidelines also define a set of special transcription tags to handle recurrent phenomena in spontaneous speech, including incomprehensible segments, laughter, non-verbal sounds, foreign-language insertions, and hesitation phenomena. In addition, explicit procedures are provided for the anonymization of personal data, such as names, locations, or other identifying information, ensuring compliance with ethical and legal requirements.

[8] https://github.com/clarinsi

[9] https://govorjena-slovenscina.um.si/transkribiranje/

4 Collected Corpora and Use Cases

The Govorjena slovenščina portal has been in active use since its initial launch in March 2025. During its first year of operation, it has successfully facilitated the collection of several small-scale corpora of Slovene everyday spoken interaction, each serving distinct research and language technology needs. These include (1) the training corpus of spoken Slovene ROG-dialog (Verdonik et al., 2025), (2) the corpus of conversational humour Krohot (Krajnc Ivič et al., 2025), and (3) an addition to the Slovene SloBench benchmark suite for automatic speech recognition evaluation, referred to as SloBenchASR [10].

ROG-dialog was conceived as an extension and enhancement of the existing training corpus of spoken Slovene, ROG-Artur (Verdonik, Ljubešić, et al., 2024), which predominantly comprises public, monologic speech. In contrast, ROG-dialog focuses on informal dialogic interactions, sampling a balanced set of speakers with respect to regional residence, age, and gender to better reflect natural conversational variation. In total, the ROG-dialog corpus comprises 12 recordings involving 25 speakers, resulting in more than 5 hours (50,000 words) of recorded speech. This corpus is openly accessible via the CLARIN.SI repository, supporting reuse in both linguistic research and language technology development.

The Krohot corpus represents a specialized collection of conversational humour. It consists of 10 audio recordings of private, spontaneous conversations involving two or three participants, with a total duration of 3 hours and 52 minutes of natural speech collected primarily between May and September 2025. The recordings capture memories and narratives that elicited spontaneous humorous reactions among interlocutors. Humorous segments within these conversations have been manually annotated using a purpose-built tagging scheme encompassing categories such as vocabulary choice, interpersonal relation, content focus, speaker attitude, and stylistic manner of delivery. The Krohot corpus is provided in high-quality WAV format, with aligned transcriptions available in EXMARaLDA, Transcriber, and plain text formats, and is is openly accessible via the CLARIN.SI repository.

Furthermore, Govorjena slovenščina has contributed data to the SloBenchASR component of the SloBENCH evaluation platform, an independent benchmarking framework for Slovenian natural language processing technologies. SloBenchASR provides a dedicated leaderboard for the evaluation of automatic speech recognition systems, using evaluation data that, to the best of current knowledge, are not available in existing Slovene speech databases. The original benchmark dataset consisted of 15 speech recordings with a total duration of 3 hours and 18 minutes, encompassing both public and private speech. This dataset has been expanded through Govorjena slovenščina with the addition of six recordings of private conversational speech, amounting to a total of one hour of additional data. The expanded dataset enables more robust and objective comparison of ASR systems and supports longitudinal tracking of progress in Slovene speech recognition, thereby strengthening the platform's role in language technology evaluation.

In addition to these purpose-driven collections, more than 25 hours of speech had been contributed to the portal through donated recordings by March 2026. While this reflects an initial stage of community engagement, the project has set a concrete target of collecting at least 70 hours of speech by October 2027. Data collection is conceived as a long-term effort that will continue beyond the project period, with the aim of developing a substantially larger and continuously expanding resource.

In parallel with public participation, the platform is designed to attract researchers and research teams seeking to collect speech data for specific research objectives. By using the portal, such teams can rely on established legal, technical, and organisational frameworks, thereby reducing overhead and ensuring compatibility with existing Slovene speech corpora.

5 Discussion and Conclusion

The persistent underrepresentation of conversational spoken language resources, particularly for everyday human–human interaction, remains a central challenge for both linguistic research and language technology development. As demonstrated throughout this paper, this scarcity cannot be attributed to a single factor, but rather to an interplay of methodological, technical, legal, and social constraints. Tradi-

[10] https://slobench.cjvt.si/leaderboard/view/10

tional laboratory and fieldwork-based approaches, while indispensable for producing high-quality data, are difficult to scale and are subject to well-known limitations such as the observer's paradox. At the same time, large-scale online sources provide volume but lack representativeness for private, informal speech. These conditions motivate the exploration of alternative collection paradigms that can balance data quality, diversity, and feasibility.

The analysis of international citizen-science and crowdsourcing initiatives—CorCenCC, Anneta kõnet, and Lahjoita puhetta—illustrates both the potential and the limitations of remote speech collection. These projects demonstrate that large volumes of spoken data can be collected efficiently when technical design, legal governance, and participant motivation are addressed holistically. At the same time, they reveal recurring trade-offs between spontaneity and acoustic control, between short low-effort contributions and longer interactionally rich recordings, and between openness and privacy protection. Crucially, the most successful initiatives treat speech collection platforms as socio-technical systems rather than as purely technical upload tools, foregrounding trust, transparency, and purpose communication alongside usability and quality control.

Within this broader landscape, Govorjena slovenščina occupies a distinct position. Rather than aiming for rapid large-scale data accumulation, the platform is designed as a reusable academic infrastructure that prioritizes legal clarity, methodological consistency, and long-term interoperability of data. Its design choices—such as the absence of a dedicated recording application, the emphasis on longer conversational recordings, contribution limits per speaker, and detailed metadata and transcription standards—reflect a deliberate focus on data depth, contextual richness, and ethical governance. This orientation aligns the platform particularly well with linguistic research that requires extended interactional context, while still supporting language technology applications.

The corpora collected through Govorjena slovenščina—ROG-dialog, Krohot, and the SloBenchASR extension—demonstrate the platform's capacity to support diverse research goals, ranging from conversational analysis and humor studies to speech technology evaluation. Although the total volume of collected speech remains modest compared to national donation campaigns, these resources validate the platform's workflows and standards and provide a foundation for future expansion. Early signs of voluntary participation further suggest that, with targeted outreach and sustained engagement strategies, broader community involvement may be achievable over time.

In conclusion, Govorjena slovenščina shows how a standards-driven, legally grounded, and transparent platform can contribute meaningfully to the sustainable development of spoken language resources, even under constrained financial and organizational conditions. Its primary contribution lies not in maximizing data volume, but in lowering structural barriers for researchers and contributors alike by providing a shared infrastructure that integrates legal compliance, technical robustness, and methodological coherence. As the platform continues to evolve—through planned technical upgrades, expanded dissemination efforts, and increased collaboration with research teams—it has the potential to become a central hub for the collection and reuse of spoken Slovene. More broadly, the approach outlined in this paper offers a transferable model for other smaller or under-resourced languages seeking to balance ethical responsibility, research quality, and practical feasibility in the collection of conversational speech data.

6 Acknolwedgment

The work presented in this paper was conducted within the research project "Large Language Models for Digital Humanities" (GC-0002) and the Research Programme "Advanced Methods of Interaction in Telecommunication" (P2-0069), both funded by the Slovenian Research and Innovation Agency (ARIS).

A Architectural diagram of the online interface

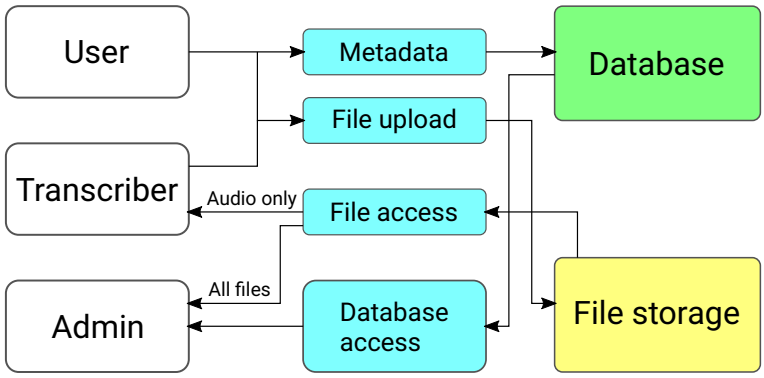

Figure 1: Architectural diagram of the online interface

B Metadata on speakers

- **Transcription ID** – Unique identifier of the transcription linked to the recording.
- **Recording ID (and source ID)** – Identifier of the edited (approved) recording and source recording (in case of renaming).
- **Corpus ID** – Identifier of the corpus to which the recording belongs.
- **Personal ID** – Anonymous identifier assigned to each speaker.
- **Speaker order** – Order in which speakers enter the interaction (e.g. first speaker, second speaker).
- **Gender** – Speaker's gender.
- **Age** – Speaker's age at the time of recording, expressed in years, or in months for children up to 6 years old.
- **First language** – Speaker's primary (native) language.
- **Information on bilingualism** – Whether the speaker has been exposed to and uses more than one language from birth, and in what contexts.
- **Register** – Type of speech variety used (standard, colloquial, dialect).
- **Dialect group** – Broad regional dialect classification.
- **Dialect** – Specific dialect spoken by the speaker.
- **Degree of dialect retention** – Extent to which dialect features are preserved in speech, measured on a scale from 1 to 3.
- **Education** – Speaker's highest completed level of education.
- **Parent or guardian education (in the case of minors)** – Educational level of the speaker's parent or guardian who provides consent.
- **Permanent residence** – Current place of residence (settlement, region, country).
- **Childhood residence** – Place where the speaker spent most of their childhood (settlement, region, country).

- **Other residences** – Other places where the speaker has lived for more than a year (settlement, region, country).
- **Atypical speech** – Whether the speech deviates from typical development; if yes, description of the condition.

C Metadata on speech recordings

- **Transcription ID** – Unique identifier of the transcription linked to the recording.
- **Recording ID (and source ID)** – Identifier of the edited (approved) recording and source recording (in case of renaming).
- **Corpus ID** – Identifier of the corpus to which the recording belongs.
- **Language of speech** – Language used in the recording.
- **Manual description of the recorded situation** – Short description of the communicative situation.
- **Date** – Date when the recording was made (month, year).
- **Source** – Origin of the recording (e.g. own recording, recordings from media sources such as national broadcaster, etc.).
- **Text region** – Name of the settlement (city, town, or village) where the recording was made.
- **Genre** – Communicative genre (e.g. conversation, interview, narration).
- **Speech type** – Type of speech situation classified as public formal, public informal, institutional formal, institutional informal, or private.
- **Channel** – Medium of communication, classified according to the target interlocutors or audience (e.g. in-person, telephone, television, radio, internet audio, internet video).
- **Number of speakers** – Total number of participants in the recording.
- **Keywords** – Descriptive tags indicating topics or content.
- **Recording device** – Description of the device used (e.g. smartphone, recorder).
- **Recording location** – Type of physical setting where the recording was made (e.g. living room, student room, office).
- **Original recording format** – File format of the original recording (e.g. WAV, MP3).
- **Manual rating of recording quality** – Subjective assessment of audio/video quality, measured on a scale from 1 to 5.
- **Length of the recording** – Total duration of the recording.
- **Additional field notes** – Any relevant observations about the recording process, participant interaction and behaviour, or notable linguistic and speech features.

References

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4

Bizjak, A. (2025). Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *13*(1), 58–103. https://doi.org/10.4312/slo2.0.2025.1.58-103

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2001). Developing computational infrastructure for the CorCenCC corpus: The national corpus of contemporary Welsh. *Language Resources and Evaluation*, *55*(3), 789–816. https://doi.org/10.1007/s10579-020-09501-9

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., & Thomas, E. M. (2020). The national corpus of contemporary Welsh: Project report — y corpws cenedlaethol cymraeg cyfoes: Adroddiad y prosiect. https://doi.org/10.48550/arXiv.2010.05542

Krajnc Ivič, M., Mihailović, L., Ivič, D., & Verdonik, D. (2025). Corpus of conversational humor Krohot 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/2065

Labov, W. (1972). *Sociolinguistic patterns (conduct and communication)*. University of Pennsylvania Press.

Lindén, K., Jauhiainen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., & Pitkänen, O. (2022, October). Donate speech. In D. Fišer & A. Witt (Eds.), *Clarin: The infrastructure for language resources* (pp. 481–510). De Gruyter. https://doi.org/10.1515/9783110767377-019

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Lyding, V., Nicolas, L., & König, A. (2022, June). About the applicability of combining implicit crowdsourcing and language learning for the collection of nlp datasets. In C. Callison-Burch, C. Cieri, J. Fiumara, & M. Liberman (Eds.), *Proceedings of the 2nd workshop on novel incentives in data collection from people: Models, implementations, challenges and results within LREC 2022* (pp. 46–57). European Language Resources Association. https://aclanthology.org/2022.nidcp-1.8

Neale, S., Spasic, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., Marshall, L., & Knight, D. (2017). The CorCenCC crowdsourcing app: A bespoke tool for the user-driven creation of the national corpus of contemporary Welsh. *The 9th International Natural Language Generation conference*, 24–28. https://eprints.ncl.ac.uk/242273

Nielsen, J. (1994, April). Enhancing the explanatory power of usability heuristics. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of the sigchi conference on human factors in computing systems* (pp. 152–158). Association for Computing Machinery. https://doi.org/10.1145/191666.191729

Oostdijk, N. (2000, May). The spoken Dutch corpus. Overview and first evaluation. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the second international conference on language resources and evaluation (LREC'00)*. European Language Resources Association (ELRA). https://aclanthology.org/L00-1083/

Rutten, M., Minkman, E., & van der Sanden, M. (2017). How to get and keep citizens involved in mobile crowd sensing for water management? A review of key success factors and motivational aspects. *Wiley Interdisciplinary Reviews: Water*, *4*(4), e1218. https://doi.org/10.1002/wat2.1218

Schmidt, T., & Wörner, K. (2014). "EXMARaLDA". In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *Handbook on corpus phonology* (pp. 402–419). Oxford Academic. https://doi.org/10.1093/oxfordhb/9780199571932.013.030

Verdonik, D., Bizjak, A., Žgank, A., Maučec, M. S., Trojar, M., Gros, J. Ž., Bajec, M., Bajec, I. L., & Dobrišek, S. (2024). Strategies for managing time and costs in speech corpus creation: Insights from the Slovenian ARTUR corpus. *Language Resources and Evaluation*, *59*, 1899–1924. https://doi.org/10.1007/s10579-024-09792-2

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024, May). Gos 2: A new reference corpus of spoken Slovenian. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 7825–7830). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.691/

Verdonik, D., Kosem, I., & Vitez, A. Z. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*(4), 1031–1048. https://doi.org/10.1007/s10579-013-9216-5

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K., & Čibej, J. (2024, September). Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG. In Š. A. Holdt & T. Erjavec (Eds.), *Conference on language technologies and digital humanities (JT-DH-2024)* (pp. 469–484). Inštitut za novejšo zgodovino. https://doi.org/10.5281/zenodo.13936425

Verdonik, D., Rupnik, P., Vidinić, J., & Ljubešić, N. (2025). Corpus of spoken slovenian ROG-dialog 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/2073

Analysing Speech Acts in Organisational Communication

Marcus Grattan
Computer and Information Science
Linköping University, Sweden
margr792@student.liu.se

Andrea Fried
Management and Engineering
Linköping University, Sweden
andrea.fried@liu.se

Arne Jönsson
Computer and Information Science
Linköping University, Linköping, Sweden
arne.jonsson@liu.se

Abstract

The research presented in this paper addresses the challenges of analysing organisational stakeholder communication using speech act theory, supported by a novel analytical model. We present a large, automatically annotated dataset focused on present- and future-oriented speech acts in Swedish organisational communication, specifically with regard to the information security standard ISO/IEC 27001. To construct and evaluate the dataset, we employ a hybrid annotation pipeline combining quantised large language models, retrieval-augmented prompting, and a fine-tuned multilingual XLM-RoBERTa classifier, which is evaluated against a manually annotated gold standard. The classifier is subsequently applied to a large corpus of corporate texts from Swedish companies' websites, enabling comparative analysis between general corporate communication and ISO-focused discourse. The results show that ISO/IEC 27001 communication is dominated by assertive speech acts, with commissives playing a secondary role, reflecting an emphasis on factual compliance rather than aspirational or promotional claims. Beyond its empirical findings, the study contributes a reusable dataset and classifier to the CLARIN infrastructure, demonstrates how CLARIN K-centres can support social science research, and provides insights into the use of quantised large language models for speech act classification in mid-resource language settings.

1 Introduction

In a joint project between the Swedish SMS CLARIN K-centre[1] and management researchers, we investigate stakeholder communication regarding information security on Swedish corporate websites. Rogers (2002) argues that preventive innovations, such as those aimed at ensuring information security, are often adopted slowly because their benefits to organisations are intangible and long-term. Clear communication with stakeholders that emphasises the benefits of information security is therefore essential, particularly given the regulatory context that demands accountability for managing not only social and environmental risks, but also technical risks.

Recent work suggests that stakeholder communication of organisations is influenced by two circumstances. First, it is influenced by the fact that preventive innovations not only generate indirect benefits in the form of social capital, such as reputation and shared understandings, but can also yield direct benefits in the form of economic capital gains (see Bourdieu (1986), Kuhn (2008), and Nahapiet and Ghoshal (1998)). For instance, companies can monetise safety training and consultancy services around information security (Hu et al., 2022; Khando et al., 2021; Siponen, 2000) generating revenue streams and increasing economic capital (Bourdieu, 1991). Yet, such economic interests may weaken stakeholder communication and potentially undermine social capital in the longer run (Kuhn, 2008). The assumption is that the pursuit of direct economic benefits encourages organisations to make more promises, relying heavily on commissive utterances, which can crowd out dialogic, trust-building communication and,

[1] https://sprakbanken-clarin.lingfil.uu.se/centra/sms_en.html

Marcus Grattan, Andrea Fried, and Arne Jönsson 2025. Analysing Speech Acts in Organisational Communication. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 39–51. DOI: https://doi.org/10.3384/ecp222

if unmet, erode credibility, mutual learning, and durable social ties among stakeholders (see Cooren & Taylor, 1997; Searle, 1985).

Secondly, we can assume that organisations with standard certifications (here: ISO/IEC 27001 certification) for information security tend to focus more on their achievements when communicating with stakeholders. Upon certification, they undergo an external audit in which their achievements in information security were recognised. Therefore, they presumably rely more than non-certified organisations do on assertive utterances to demonstrate their credibility, compliance with regulations and leadership in information security. This could potentially result in a tone that is less self-promotional and less promising on their websites (see Cooren and Seidl, 2020; Kuhn, 2008; Schoeneborn et al., 2019; Taylor, 1993).

The ISO/IEC 27001 standard requires organisations to establish, implement and audit the rules and authorities relating to the legal and ethical aspects necessary for the proper functioning and accountability of the entire data and algorithm lifecycle (Janssen et al., 2020, p.3), see also (Culot et al., 2021; Disterer, 2013; Mirtsch et al., 2021). ISO/IEC 27001 certification ensures information security and encourages good corporate behaviour by requiring compliance to be audited by external groups.

We therefore ask: How do Swedish companies communicate information security to stakeholders on their corporate websites? Specifically, we are interested in understanding if the speech acts in these communications are influenced by (a) the pursuit of social versus economic capital gains and (b) ISO/IEC 27001 certification. The objective is to determine if explicit ISO/IEC 27001 communication reflects distinct, more assertive speech acts that emphasize descriptive content and speech act logic, as opposed to commissive speech acts that express future-oriented intentions to ensure information security. Methodically, the paper aims to develop a novel approach to computational text analysis that is sensitive to specific organizational topics by analysing the content of corporate websites to reveal differences in speech act distributions.

2 The corpus and previous analyses

A specialised corpus of corporate texts from Swedish company websites was created in 2020, using a web crawler, to analyse the discourse surrounding information security with a focus on the ISO/IEC 27001 standards (Jönsson et al., 2024). The corpus comprises 441,625 sentences, filtered for duplicates. This method of data collection introduces significant noise, characterised by the presence of links, button labels, dates, and other extraneous elements. In addition, some sentences are incomplete, fragmented, or consist of multiple unrelated sentences, and many web pages are in English, or both English and Swedish. English text, both web pages or paragraphs in English and Swedish web pages, is filtered out. We further filtered out sentences with more than 250 words and paragraphs with more than 500 sentences, see (Jönsson et al., 2024). The final corpus contains data from 291 companies, a total of 5,960,328 tokens, with a vocabulary size of 204,312 and an average sentence length of 24.3 tokens.

Following Mirtsch et al. (2020)'s suggestion, we first identified companies that were certified for ISO/IEC 27001. These companies adopted the standard and underwent a related audit. We call these companies Compliers[2]. Mirtsch et al. (2020) also discovered companies that implemented the requirements without being certified for ISO/IEC 27001, which we refer to as Stewards. Beyond these findings, we identified a third group of companies that do not claim to have implemented the ISO/IEC 27001 standard requirements. Rather, they refer to their suppliers and other stakeholders to inform of their secure handling of information, we calll this group Advocates. Furthermore, our dataset revealed that some companies have integrated ISO/IEC 27001 consulting or training into their business models, creating an opportunity for economic capital gains. Other companies recognize only the preventive nature of adhering to ISO/IEC 27001 requirements and focus on social capital gains.

Based on these categorisations, we partitioned the corpus into subsets. The first denotes the three approaches to standard adherence Compliers (C), Stewards (S), and Advocates (A). The second signifies social capital (S) or economic capital (E) respectively. These categories help to understand the varying approaches of companies to information security communication and are added manually to the dataset.

[2]Compliers were previously referred to as Adopters (Grattan et al., 2025).

We can also form six distinct types of adoption of preventive innovation, referred to as CS, CE, SS, SE, AS, AE[3], in which each company belongs.

Our previous analyses of the corpus included word frequency counts, word clouds, and content analysis along with sentiment analysis to uncover differences in communication about preventive cyber innovations (Jönsson et al., 2024). To capture future-versus-present-oriented communication in relation to InfoSec standards, we conducted a speech acts analysis, which is presented in this paper.

We will compare all paragraphs, Section 5.1, with only those paragraphs containing ISO sentences, Section5.2. This allows for analysis of separate general patterns of organizational website communication from communication that is explicitly focused on ISO/IEC 27001.

Furthermore, looking at all paragraphs captures how information security is embedded in broader corporate self-presentations, where ISO/IEC 27001 is only one of many topics and is often mentioned indirectly or in passing. In contrast, the ISO paragraphs subset focuses on moments where organizations explicitly foreground the standard and therefore articulate more deliberate and institutionally framed meanings of information security. Comparing the two makes it possible to examine whether organizations change what they say and how they position themselves—for example, by emphasizing verified achievements, compliance, or formal responsibilities—when ISO/IEC 27001 becomes the explicit topic, as opposed to being part of general corporate discourse. This sheds light on whether ISO communication reflects a distinct content logic (e.g. legitimizing, evidencing, or distancing responsibility) rather than merely mirroring overall communication styles.

3 Model building

Our working assumption was that organisations express their adherence to standards through intentional statements (Commissives) or descriptive statements (Assertives). This approach aligns with the concept of direction of fit (Searle, 1985), where Commissives represent a world-to-word fit, expressing future intentions and commitments that an organisation plans to fulfil. Conversely, Assertives represent a word-to-world fit, stating current or past conditions within the organisation. For the purposes of streamlined analysis and because of their overlap with Assertives in expressing current realities, Declaratives were grouped under the Assertive class. Furthermore, Expressives and Directives were deemed outside the scope of this analysis and thus grouped, along with unclear and fragmented data, under Other. A summation of the classes used is provided in Table 1.

Class	Speech Act
Commissive	Commissive
Assertive	Assertive & Declarative
Other	Directive & Expressive & Unclear data

Table 1: Classes used in the text classification task of this paper.

The construction of the Speech Act classification model began with selecting a large language model (LLM) for an initial round of annotation. Specifically, only quantised models were considered due to computing restraints. The selection was based on the performance on a small manually curated data set, sampled from the ISO/IES 27001 corpus (Jönsson et al., 2024), containing approximately 100 manually selected instances per class. Initially, two models were considered. That is, Mixtral8x7B-Instruct quantised to an average of 3.5 bits per weight (BPW), and Llama3-70B-Instruct with 2.4BPW. The models were then prompted using the same prompt, which included Role-, Chain-of-Thought-, and Few-Shot prompting strategies. The metrics considered for model selection were Accuracy and per-class F1 score. Llama3 showed superior scores, as per Table 2, across the metrics tested and was therefore selected.

After selecting Llama3, an optimal hyperparameter configuration was sought. The hyperparameters considered were *Top P*, *Top K*, and *Temperature*. Other parameters were, given the exponential increase in computation time, not considered.

[3]In a previous paper Jönsson et al., 2024, these were referred to as 11, 12, 21, 22, 31 and 32, respectively.

Model selection		
Metrics	Mixtral	Llama3
Accuracy	0.695	0.786
Commissive F1	0.785	0.826
Assertive F1	0.645	0.767
Other F1	0.675	0.774

Table 2: Model selection results

Using the dataset mentioned above, each parameter permutation (with a defined range and step-size) was tested in a two round process. The first round tested a broad range of configurations to identify a potential subset of parameters that yielded task-optimal results. This was achieved by defining a wide range and large step sizes for each parameter, with performance assessed on the basis of accuracy and per-class F1 score.

The first round of testing resulted in a set of permutations that indicated increased performance on the dataset, compared to results from the model selection process. The configuration that resulted in the highest accuracy was picked from the first round for the second round of tuning. The second round of testing aimed to fine-tune the model hyperparameters around the most promising configurations identified in the first round. Thus, this round of tests involved narrowing the range and reducing the step size to fine-tune the parameters, seeking possible optimal configurations within a narrower range.

Despite observing a slight decrease in overall accuracy in the second round of testing, the decision to proceed with the parameters from the second round was driven by a significant improvement in the F1 score for the Commissive class, which increased from 0.766 to 0.851. This improvement indicated that the model more accurately identified commissive instances. Furthermore, the improved accuracy from the first round of testing seemed to stem from better performance on the Other class. The results from the chosen hyperparameter configuration from each round during the tuning process are summarised in Table 3.

Hyperparameter tuning: Results		
Metrics	First-round	Second-round
Accuracy	**0.792**	0.789
Commissive F1	0.766	**0.851**
Assertive F1	**0.768**	0.762
Other F1	**0.854**	0.764

Table 3: Hyperparameter Tuning: Results.

After the hyperparameter tuning process, the selected model was deployed to annotate a larger section of the corpus in order to create a demonstration dataset for Clue and Reasoning Prompting (CARP) (Sun et al., 2023). The demonstration dataset is used for few-shot example retrieval in later classification steps and consists of a text, its classification, supporting clues (e.g keywords, phrases, and/or syntactic features), as well as a supporting diagnostic reasoning for the classification given the clues. The supported clues and diagnostic reasoning are generated by the LLM.

Generated annotations for the demonstration dataset were manually validated, and correctly labelled instances were organised into class-specific datasets, each containing the original text and its corresponding Speech Act label. Ultimately, each class contained 550 validated entries. To ensure diversity and mitigate model bias, 50 instances per class were manually curated by reviewing and correcting the model's

misclassifications. Due to data sparsity in the minority class, Commissive, ChatGPT-4 was employed for data augmentation. Forty carefully selected examples were expanded to better capture under-represented syntactic and semantic structures.

To complete the process of constructing a demonstration dataset the LLM was prompted, following the procedure described in Sun et al. (2023), to generate supporting clues and a diagnostic reasoning for each entry. This process yielded a 1,650-entry demonstration dataset, evenly distributed across classes.

To support classification-sensitive demonstration retrieval in the final annotation round, an XLM-RoBERTa model (Conneau et al., 2020) was fine-tuned on the CARP demonstration dataset. Fine-tuning was essential for aligning the model not only with sentence embeddings but also with class annotations, enabling accurate retrieval of semantically and label-relevant examples for unseen sentences. The model was used to compute text embeddings, which were appended to each dataset entry.

In the final annotation step the Llama3 model was provided with task instructions, class definitions, and retrieved demonstrations via the fine-tuned retrieval model. Demonstrations were added up to the model's token limit, sorted in ascending order of cosine distance to the input sentence, as described by Sun et al. (2023).

This resulted in an automatically annotated dataset, comprising a random 56% subset of the full corpus. A final XLM-RoBERTa classifier was fine-tuned on this dataset. The model and dataset are available to the CLARIN community[4]. Training employed a linear learning rate scheduler with warm-up steps corresponding to 10% of batches per epoch and used the AdamW optimiser. To address class imbalance, the loss function incorporated class weights. Early stopping was triggered if validation loss increased for three consecutive evaluations. Weight decay was set to 0.01, and the learning rate was set to a conservative value of 4 x 10^{-5} to prevent premature convergence. A batch size of 16 was chosen to encourage generalisation. An 80/20 train/validation split was used, with evaluation occurring three times per epoch. Model selection was based on the weighted F1 score. The final model was trained for 2.33 epochs. Entries labelled "No match found" were excluded from the training data.

To evaluate the classifier, a test set of 5,000 entries was manually annotated. An average of 1,000 sentences were labelled per day, split across two sessions. The final test set comprised 665 Commissive, 3,228 Assertive, and 1,107 Other instances. Following Abercrombie et al. (2025), intra-annotator agreement was measured on 10% of the test set one week after completion, yielding an agreement score of 0.8, indicating strong consistency.

For more detail and discussion related to the model-building process, see Grattan (2024).

4 Classifier model results

The automatically annotated dataset comprised 138,052 entries excluding instances for which no annotation was found. In total the automatically annotated dataset included 32,458 (24%) commissive entries, 80,377 (58%) assertive entries and, 25,215 (18%) entries annotated as other. The model returned no annotation 301 times. Dataset metrics are presented in Table 4

Automatically Annotated Dataset				
Metric	Commissive	Assertive	Other	Total
Count	32,458	80,377	25,215	138,052
Percentage	24	58	18	

Table 4: Automatically annotated dataset metrics

The classifier model achieved an accuracy of 0.73 and a weighted F1-score of 0.75 on the test dataset, outperforming the most frequent class baseline, which had an accuracy of 0.65 and a weighted F1-score of 0.51. The model's weighted precision and recall were 0.80 and 0.73, respectively. For the Commissive class, the model achieved precision, recall, and F1-scores of 0.34, 0.68, and 0.45, respectively. In the

[4]https://huggingface.co/MarcusGrattan/classifier-model/tree/main

Assertive class, the model performed better with precision of 0.88, recall of 0.74, and an F1 score of 0.80. The Other class achieved a precision of 0.84, recall of 0.74, and F1 score of 0.79. Per class metrics are summarised in Table 5.

Per class model metrics			
Metric	Commissive	Assertive	Other
Precision	0.34	0.88	0.84
Recall	0.68	0.74	0.74
F1	0.45	0.80	0.79

Table 5: Per class results

5 Applying the model

The model was applied to the corpus of corporate websites of Swedish companies that communicate with their stakeholders (e.g. clients, customers, government, non-profit organisations and other) about their ISO/IEC 27001 implementation efforts, see Section 2. As in previous studies between the management researchers and the SMS CLARIN K-centre we looked at various categorisations and subsets of the corpus. In this case, we use the whole corpus as described in Section 2, called All paragraphs, and contrasted that with the subset comprising only paragraphs containing ISO-related terms, i.e. if a paragraph has at least one sentence with an ISO-related term the whole paragraph is included. The datasets are further divided into groups according to the various governance approaches, capital benefit dimensions, and preventive innovation adoption, see Section 2.

We performed statistical analyses on the data, including pairwise comparisons, using Cramér's V, and no difference showed an effect. However, the primary goal of our study is pattern recognition and hypothesis generation rather than hypothesis testing. Traditional significance testing adds rigour to confirmatory research aiming to generalise the findings. We have chosen to focus on effect sizes and qualitative validation through illustrative examples that demonstrate the substantive meaning behind the numerical patterns. This approach acknowledges the exploratory nature of our analyses.

5.1 All paragraphs

The corpus comprising all paragraphs consists in total of 81,555 paragraphs with 441,625 sentences.

The results when data are categorised according to governance approaches are shown in Table 6 and visually in Figure 1, revealing little difference between the analysis categories Compliers, Stewards, and Advocates.

	Assertive	Commissive	Other
Compliers	100,547 (51)	47,805 (24)	48,673 (25)
Stewards	81,372 (52)	33,746 (21)	42,305 (27)
Advocates	51,404 (59)	18,568 (21)	17,205 (20)

Table 6: Speech act distribution for all paragraphs according to governance approaches. Number of sentences with percentage in parenthesis.

Table 7 and Figure 2 show the result when using the social/economic capital dimension. The results show very little difference between the dimensions.

Table 8 and Figure 3 show the result when combining the dimensions into six categories that represent the three governance approaches and the social/economic capital dimension. The results show very little difference between the six dimensions[5].

[5]Here we have the largest value on Cramér's V, 0.16, for CE vs AS, which is considered weak.

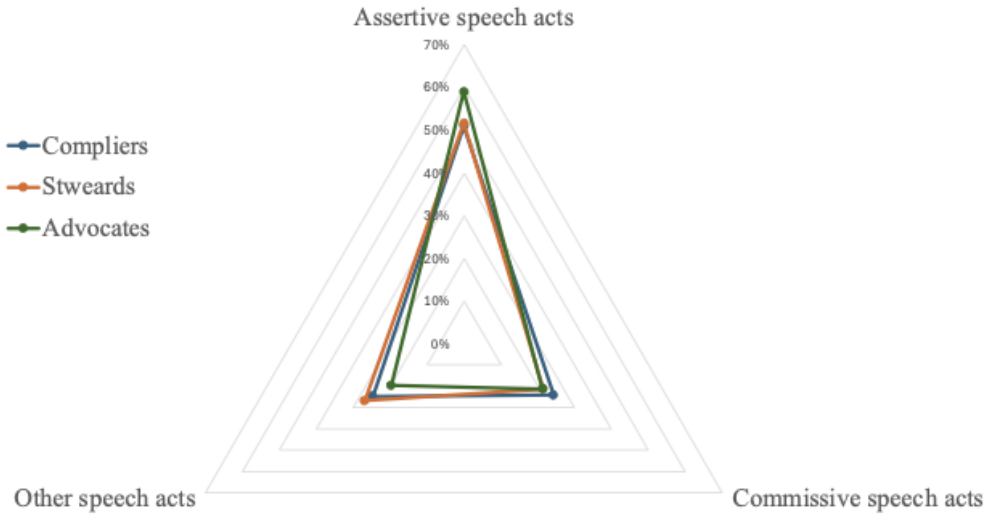

Figure 1: Diagram of the distribution for all paragraphs according to governance approaches, percentage.

	Assertive	Commissive	Other
Social capital	180,096 (52)	77,127 (22)	86,037 (25)
Economic capital	53,227 (54)	22,992 (23)	22,146 (23)

Table 7: Speech act distribution for all paragraphs for the social/economic capital dimension. Number of sentences with percentage in parenthesis.

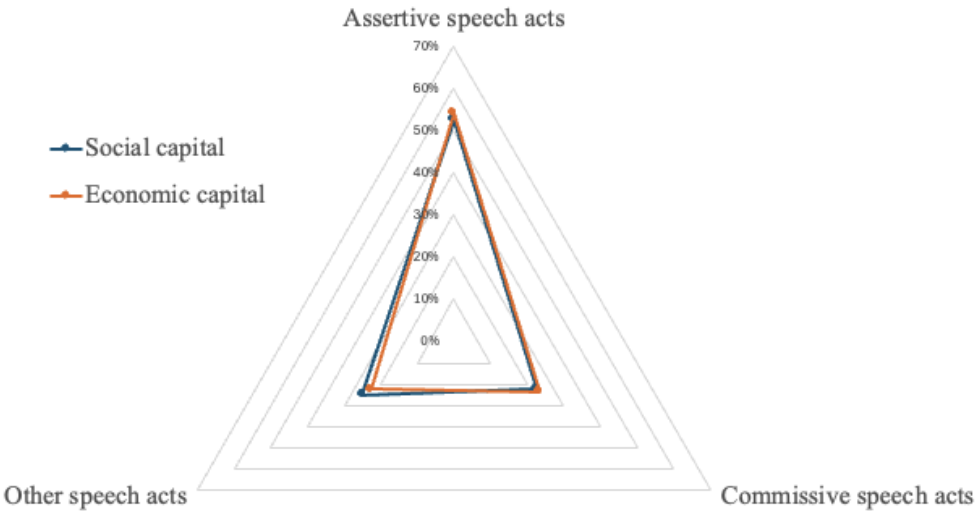

Figure 2: Diagram of the distribution for all paragraphs for the social/economic capital dimension, percentage.

5.2 ISO paragraphs

The corpus comprising ISO-related paragraphs consists in total of 1,492 paragraphs with 31,440 sentences.

The results when data are categorised according to governance approaches are shown in Table 9 and visually in Figure 4, revealing little difference between the analysis categories Compliers, Stewards, and

	Assertive	Commissive	Other
CS	95,015 (51)	44,779 (24)	46,000 (25)
CE	5,532 (49)	3,026 (27)	2,673 (24)
SS	65,454 (51)	27,090 (21)	35,454 (28)
SE	15,918 (54)	6,656 (23)	6,851 (23)
AS	19,627 (67)	5,258 (18)	4,583 (16)
AE	31,777 (55)	13,310 (23)	12,622 (22)

Table 8: Speech act distribution for all paragraphs using the six types of preventive innovation adoption. Number of sentences with percentage in parenthesis.

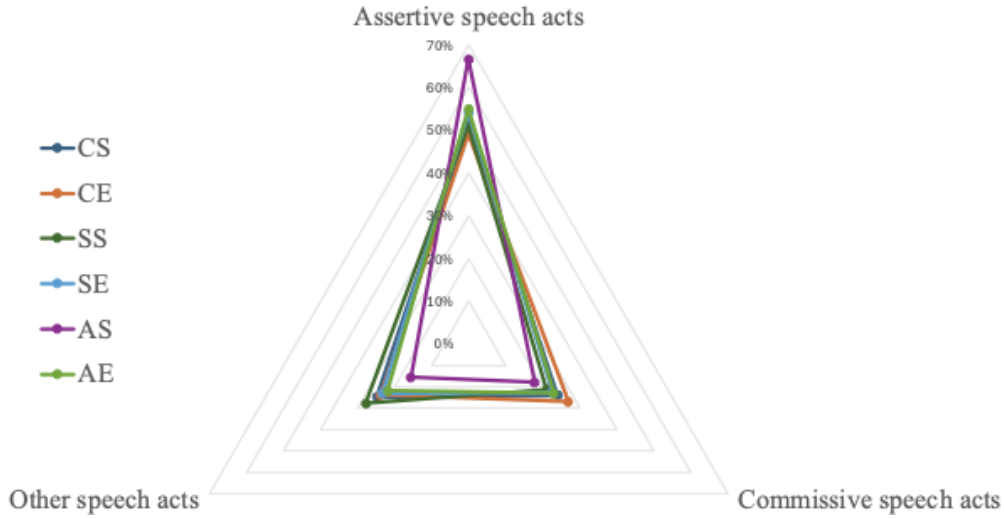

Figure 3: Diagram of the distribution for all paragraphs using the six types of preventive innovation adoption, percentage.

Advocates.

	Assertive	Commissive	Other
Compliers	4,887 (59)	2,199 (27)	1,174 (14)
Stewards	3,868 (58)	1,983 (30)	824 (12)
Advocates	11,430 (69)	3,492 (21)	1,583 (10)

Table 9: Speech act distribution for ISO paragraphs according to governance approaches. Number of sentences with percentage in parenthesis.

Table 10 and Figure 5 show the result when using the social/economic capital benefit dimension. The results show very little difference between the dimensions.

Table 11 and Figure 6 show the result when combining the dimensions into six categories that represent the three governance approaches and the social/economic capital dimension. The results show very little difference between the six dimensions.

5.3 Discussion

The performance of the classification model revealed that it often confused Commissives with Assertives, and to a lesser degree, vice versa. The precision and recall metrics for the Commissive class suggest that

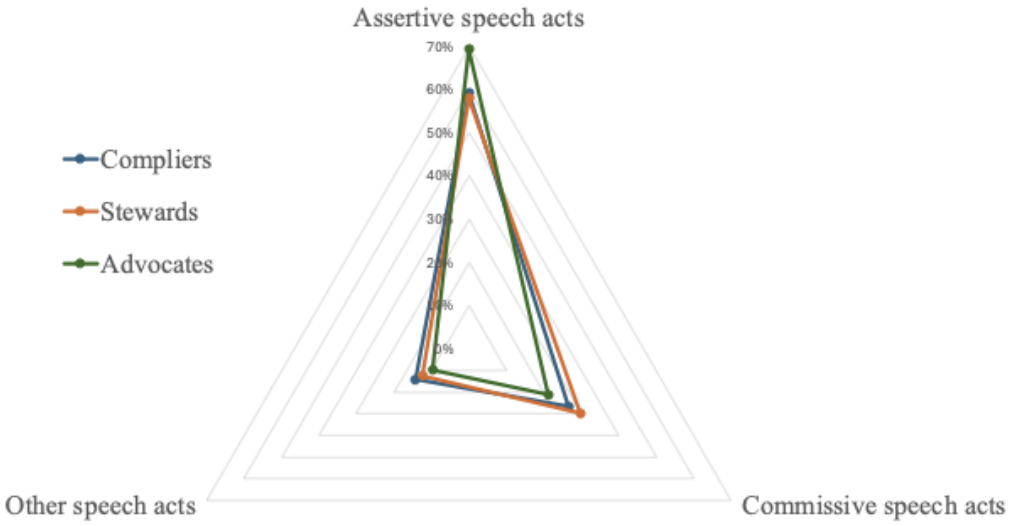

Figure 4: Diagram of the distribution for ISO paragraphs according to governance approaches, percentage.

	Assertive	Commissive	Other
Social capital	7,070 (58)	3,362 (28)	1,663 (14)
Economic capital	13,115 (68)	4,312 (22)	1,918 (10)

Table 10: Speech act distribution for ISO paragraphs for the social/economic capital dimension. Number of sentences with percentage in parenthesis.

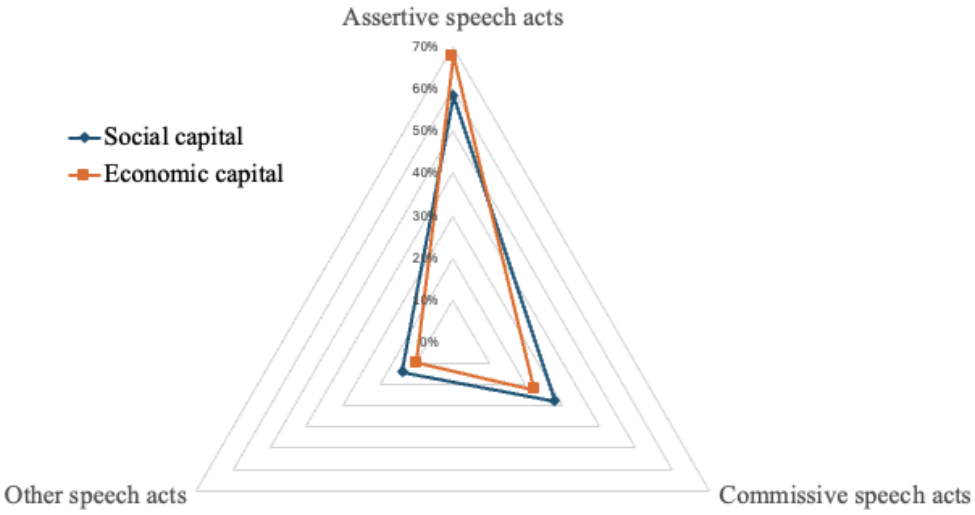

Figure 5: Diagram of the distribution for iso paragraphs for the social/economic capital dimension, percentage.

the model had difficulty accurately identifying Commissive speech acts. Specifically, the precision was low (0.34), indicating that many instances labelled as Commissives were incorrect. On the other hand, the recall for Commissives was higher (0.68), suggesting that the model was relatively effective at identifying true Commissive statements, though not always accurately.

	Assertive	Commissive	Other
CS	4,214 (59)	1,848 (26)	1,045 (15)
CE	673 (58)	351 (30)	129 (11)
SS	2,564 (58)	1,327 (30)	536 (12)
SE	1,304 (58)	656 (29)	288 (13)
AS	292 (52)	187 (33)	82 (15)
AE	11,138 (70)	3,305 (21)	1,501 (9)

Table 11: Speech act distribution for ISO paragraphs using the six types of preventive innovation adoption. Number of sentences with percentage in parenthesis.

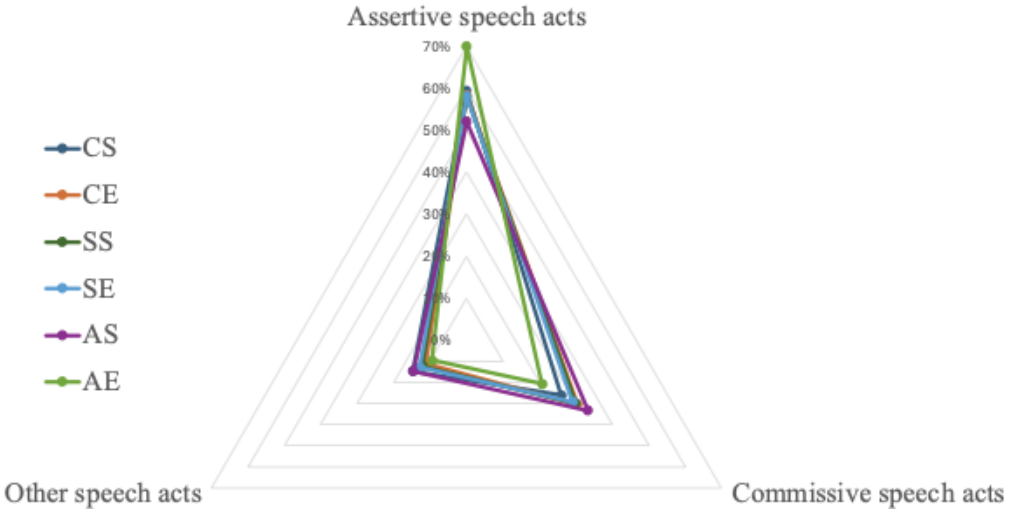

Figure 6: Diagram of the distribution for ISO paragraphs using the six types of preventive innovation adoption, percentage.

During manual annotation, it was noted that few Commissive instances were explicit commitments; instead, many were implicit guarantees or commitments. For instance, *The contracted supplier must be certified according to ISO 27001, which is a globally renowned information security standard.* is formulated as assertive, but not for the speaker itself, but for others, in this case suppliers.

This made it particularly challenging to differentiate between Assertives and Commissives. It is likely that the LLM used for automatic annotation faced similar difficulties in distinguishing between these two categories. Additionally, the indirect nature of Commissives may have contributed to ambiguities in the demonstration dataset, influencing the model's performance.

One general result when applying the model, as shown in the visualisations in Figures 1, 2, and 3, compared with Figures 4, 5, and 6, is that in the ISO paragraph sub-corpus the proportion of utterances classified as Other is much smaller; instead, most utterances are classified as Assertives or Commissives. The dataset comprising all paragraphs includes substantial web clutter and generic content, which inflates the proportion of statements classified as Other and can mask meaningful communicative patterns, whereas the ISO-focused subset concentrates on thematically relevant paragraphs where organisations explicitly discuss information security.

By contrasting the two, it becomes possible to assess how topic filtering affects speech act distributions and to demonstrate that ISO-related communication is more strongly characterised by Assertive and Commissive speech acts. This comparison therefore strengthens the validity of the findings by showing which results are robust across general corporate communication and which emerge only when analysing

communication that is directly tied to the ISO standard.

Contrary to expectations, the various analysis dimensions of stakeholder communication: preventive innovation benefits, Figures 2 and 5, governance approaches, Figures 1 and 4, and both dimensions combined, Figures 3 and 6, show very similar patterns. For instance, contrary to expectations, Advocates exhibit a similar frequency of Assertive speech acts as the other categories, see Figures 1 and 4. Manual analysis of these Assertive acts reveals that they primarily describe the actions of third parties. This poses a challenge for the model, as it struggles to determine whether a statement reflects the organization's own reality or that of others, which impacts classification results.

The ISO paragraphs dataset reflects a formalised, institutionally anchored mode of corporate communication. By explicitly referencing ISO/IEC 27001, these paragraphs focus on compliance-related meanings such as certification status, audits, roles and responsibilities, and the governance of information security practices. As a result, this communication is less promotional and more oriented toward legitimacy, accountability, and alignment with externally defined rules. ISO-focused paragraphs therefore primarily express what is (assertive speech acts) and what is formally required or guaranteed (often implicit commissives), rather than aspirational or marketing-oriented claims. In terms of content, this suggests that, when organisations discuss ISO/IEC 27001 directly, they portray information security as a structured, rule-bound field overseen by standards and third-party verification rather than as an open-ended strategic objective or competitive pledge.

As shown in Section 5.2, ISO/IEC 27001 communication is dominated by assertive speech acts, with commissives playing a secondary but meaningful role, Tables 9 and 10. Surprisingly, this indicates that organisations primarily use ISO-focused paragraphs to state facts about existing structures, certifications, procedures, and third-party requirements. In doing so, they present information security as an established and verifiable organisational reality. At the same time, the presence of commissive speech acts across both the social/economic capital dimension, Table 10, and the governance approaches of Compliers, Stewards, and Advocates, Table 9, suggests that ISO communication also functions to signal ongoing responsibility and future-oriented assurance, even when these commitments are implicit rather than explicit. The relatively stable distribution of Assertives and Commissives across categories implies that ISO/IEC 27001 is framed less as a strategic aspiration and more as a matter of factual compliance. This means, credibility is established by describing what is already in place and reinforcing expectations of continued conformity, regardless of standard adherence or benefit orientation.

6 Conclusions

The research presented in this paper demonstrates how large language models (LLMs) can be effectively leveraged to reduce reliance on manually annotated datasets in the domain of Speech Act classification. Utilising techniques such as Clue and Reasoning Prompting (CARP) (Sun et al., 2023) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2021). The research successfully generated a substantial dataset of 138,052 entries. This dataset facilitated the training and evaluation of a classifier, which achieved an accuracy of 73% and a weighted F1-score of 0.75.

An expanded LLM selection process, incorporating a wider array of models, more data and focused tests specifically aimed at indirect and/or context dependent commissives, could improve the selection process and extend to better Commissive classification results. Additionally, while LLM selection was restricted to quantised models, using un-quantised models while keeping similar model parameter size, would likely improve performance.

The results of applying the model did not exhibit differences in how the adoption of the ISO/IEC 27001 standard was expressed by Swedish companies depending on the various dimensions of the analysis, e.g. the governance approach or the capital dimension. However, for the management researchers, the results revealed other interesting properties of the way companies present themselves, such as the role of assertive speech acts for ISO communication.

This work builds upon existing frameworks and proposes novel approaches for dataset generation in text classification tasks, particularly within the context of organisational communication related to ISO/IEC 27001 standards. Beyond its technical contributions, the study demonstrates how CLARIN K-

centres can support management researchers and add value to the CLARIN infrastructure and to the broader field of organisational communication research. Furthermore, it highlights the potential of quantised LLMs for Speech Act classification, offering valuable insights into their performance, especially in mid-resource language settings.

References

Abercrombie, G., Dinkar, T., Cercas Curry, A., Rieser, V., & Hovy, D. (2025, November). Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. In G. Abercrombie, V. Basile, S. Frenda, S. Tonelli, & S. Dudy (Eds.), *Proceedings of the the 4th workshop on perspectivist approaches to nlp* (pp. 63–74). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlperspectives-1.6

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). Greenwood Press.

Bourdieu, P. (1991). *Language and symbolic power*. Harvard University Press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Cooren, F., & Seidl, D. (2020). Niklas luhmann's radical communication approach and its implications for research on organizational communication. *Academy of Management Review*, *45*(2), 479–497.

Cooren, F., & Taylor, J. R. (1997). Organization as an effect of mediation: Redefining the link between organization and communication. *Communication Theory*, *7*(3), 219–260. https://doi.org/10.1111/j.1468-2885.1997.tb00151.x

Culot, G., Nassimbeni, G., Podrecca, M., & Sartor, M. (2021). The ISO/IEC 27001 information security management standard: Literature review and theory-based research agenda. *The TQM Journal*, *33*(7), 76–105.

Disterer, G. (2013). ISO/IEC 27000, 27001 and 27002 for information security management. *Journal of Information Security*, *4*(2), 92–100.

Grattan, M. (2024). *Now or later: Classifying future- and present-oriented speech acts in organizational communication* [Bachelor's Thesis]. Linköping University.

Grattan, M., Fried, A., & Jönsson, A. (2025). Challenges of analysing speech acts in organisational communication. *Proceedings of the CLARIN Annual Conference 2025*, 148–152.

Hu, S., Hsu, C.-H., & Zhou, Z. (2022). Security education, training, and awareness programs: Literature review. *Journal of Computer Information Systems*, *62*(4), 752–764.

Janssen, M., Brousa, P., Estevez, E., Barbosad, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, *37*. https://doi.org/https://doi.org/10.1016/j.giq.2020.101493

Jönsson, A., Bandyopadhyay, S., Dragisic, S. P., & Fried, A. (2024). Analyses of information security standards on data crawled from company web sites using SweClarin resources. *Selected papers from the 2023 CLARIN Annual Conference*. https://doi.org/https://doi.org/10.3384/ecp210

Khando, K., Gao, S., Islam, S. M., & Salman, A. (2021). Enhancing employees' information security awareness in private and public organisations: A systematic literature review. *Computers & Security*, *106*, 102267.

Kuhn, T. S. (2008). A communicative theory of the firm: Developing an alternative perspective on intra-organizational power and stakeholder relationships. *Organization Studies*, *29*, 1227–1254.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Mirtsch, M., Blind, K., Koch, C., & Dudek, G. (2021). Information security management in ict and non-ict sector companies: A preventive innovation perspective. *Computer & Security*, *109*. https://doi.org/109.https://doi.org/10.1016/j.cose.2021.102383

Mirtsch, M., Kinne, J., & Blind, K. (2020). Exploring the adoption of the international information security management system standard ISO/IEC 27001: A web mining-based analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, *23*(2), 242–266.

Rogers, E. M. (2002). Diffusion of preventive innovations. *Addictive behaviors*, *27*, 989–993.

Schoeneborn, D., Kuhn, T. R., & K"arreman, D. (2019). The communicative constitution of organization, organizing, and organizationality. *Organization Studies*, *40*(4), 475–496.

Searle, J. R. (1985, November). *Expression and meaning*. Cambridge University Press.

Siponen, M. T. (2000). A conceptual foundation for organizational information security awareness. *Information Management & Computer Security*, *8*(1), 31–41.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Taylor, J. R. (1993). *Rethinking the theory of organizational communication: How to read an organization*. Norwood, NJ: Ablex.

CLARIAH-AT: Back (and) to the Future

Tanja Wissik
Austrian Academy of Sciences,
Vienna, Austria
tanja.wissik@oeaw.ac.at

Walter Scholger
University of Graz, Graz,
Austria
walter.scholger@uni-graz.at

Kerstin Klenke
Austrian Academy of Sciences,
Vienna, Austria
kerstin.klenke@oeaw.ac.at

Vesna Lušicky
University of Vienna, Vienna,
Austria
vesna.lusicky@univie.ac.at

Matej Ďurčo
Austrian Academy of Sciences,
Vienna, Austria
matej.durco@oeaw.ac.at

Martina Trognitz
Austrian Academy of Sciences
Vienna, Austria
martina.trognitz@oeaw.ac.at

Seta Štuhec
Austrian Academy of Sciences,
Vienna, Austria
seta.stuhec@oeaw.ac.at

Elisabeth Steiner
University of Graz, Graz,
Austria
elisabeth.steiner@uni-graz.at

Alexandra N. Lenz
Austrian Academy of Sciences,
Vienna, Austria
alexandra.lenz@oeaw.ac.at

Markus Pluschkovits
Austrian Academy of Sciences,
Vienna, Austria
markus.pluschkovits@oeaw.ac.at

Anna Woldrich
Austrian Academy of Sciences,
Vienna, Austria
anna.woldrich@oeaw.ac.at

Abstract

This contribution presents the CLARIAH-AT consortium, which coordinates and drives Austrian activities in the European research infrastructures CLARIN and DARIAH and supports the development of digital humanities in Austria. The paper describes the evolution from the early CLARIN-AT activities to the formal establishment of the CLARIAH-AT consortium in 2019. The paper outlines the current organisational structure and infrastructure, including CLARIN B-centres and K-centres that provide repositories, services, and expertise for researcher in the humanities and beyond. It further highlights community-building measures such as funding schemes, training initiatives, and dissemination activities. Furthermore, some CLARIAH-AT funded projects are described in more detail. The paper concludes with an outlook on future developments.

Tanja Wissik, Walter Scholger, Kerstin Klenke, Vesna Lušicky, Vesna Lušicky, Matej Ďurčo, Martina Trognitz, Seta Štuhec, Elisabeth Steiner, Alexandra N. Lenz, Markus Pluschkovits, and Anna Woldrich. 2025. CLARIAH-AT: Back (and) to the Future. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 52–62. DOI: https://doi.org/10.3384/ecp222https://doi.org/10.3384/ecp222

1 Introduction

The Austrian consortium CLARIAH-AT, comprising members from Austrian universities, research institutions, and GLAM institutions, coordinates and advances Austrian activities within the European research infrastructures CLARIN and DARIAH. This paper describes the CLARIAH-AT consortium, its infrastructure, and the current activities of the consortium. In Section 2 we first revisit the history of CLARIAH-AT, as Austria is one of the founding members of both CLARIN and DARIAH. Then we describe the current consortium and infrastructure including the CLARIN B-centres and CLARIN K-centres in section 3. Furthermore, in Section 4 we discuss initiatives aimed at building and fostering the Austrian community before we give a brief outlook into future activities in Section 5.

2 History of CLARIAH-AT

Austria was already active during the preparatory phase of CLARIN, operating under the umbrella of CLARIN-AT. One of the first initiatives in this phase was a survey of existing language resources and language technologies (Wissik and Budin, 2010), conducted to support the establishment of a network and infrastructure in Austria. At that time, the coordinating institution was the University of Vienna (Mörth and Wissik, 2018). When DARIAH later also became a European Research Infrastructure Consortium (ERIC), the Austrian actors in CLARIN and DARIAH joined forces to exploit synergies and formed the core group of what became the CLARIAH-AT network. This group worked under the coordination of the Austrian Academy of Sciences on behalf of the Austrian Federal Ministry of Science, Research and Economy (Mörth and Wissik, 2018). However, it was not until 2019 that the informal CLARIAH-AT network became an official consortium operating under the name CLARIAH-AT. Prior to its formal establishment, the activities presented to the public were disseminated under the umbrella of the virtual network Digital Humanities Austria. During this period, the first Digital Humanities Austria Strategy was published (Alram et al., 2015).

3 CLARIAH-AT Consortium and Infrastructure

In 2019, a formal CLARIAH-AT[1] consortium was established by signing a consortium agreement. The consortium brings together Austrian institutions with relevant expertise in research, development, and teaching, and with an active role in the development and expansion of technical and social infrastructures for the humanities in general. The current full members include: Austrian Academy of Sciences, Austrian National Library, University for Continuing Education Krems, University of Graz, University of Innsbruck, University of Klagenfurt, University of Salzburg, and University of Vienna. The consortium also includes three observers: University of Music and Performing Arts Vienna, Central European University and Natural History Museum Vienna. The consortium is represented by the speaker (Walter Scholger, University of Graz) and vice-speaker (Tanja Wissik, Austrian Academy of Sciences). They also respectively act as National Coordinators for DARIAH (Walter Scholger) and CLARIN (Tanja Wissik). The CLARIAH-AT Infrastructure includes two CLARIN B-centres and three CLARIN K-centres, described in more detail below.

3.1 CLARIN B-Centre ACDH ARCHE

The CLARIN B-centre ACDH ARCHE, is located in Vienna. It is led by the Austrian Centre for Digital Humanities (ACDH)[2], which hosts the long-term repository ARCHE[3], which stands for "A Resource Centre for the HumanitiEs" (Trognitz 2021). The ACDH is a research institute at the Austrian Academy of Sciences founded to promote digital methods and technologies for humanities research. ARCHE, a trusted repository certified with the CoreTrustSeal (ARCHE, 2025) was launched in 2017 as a successor to the first official Austrian CLARIN centre, the CLARIN Centre Vienna / Language Resources Portal (CCV/LRP), which operated from 2013 until 2017. The CCV/LRP focused on the digital preservation of digital resources matching the fields covered within CLARIN, such as modern and classical languages, linguistics, and literature. ARCHE took over this responsibility and expanded its scope to the broader humanities, including history, jurisprudence, philosophy, archaeology, religion, music, and art history.

[1] https://clariah.at/en/
[2] https://www.oeaw.ac.at/acdh/acdh-home
[3] https://arche.acdh.oeaw.ac.at/

ARCHE's core task, the long-term preservation of digital research data from the Austrian humanities community, is supported by a dedicated preservation strategy based on the migration of formats, and aims at preserving a file's content and functionality for future reuse. Thus, ARCHE accepts a broad range of data types, including images, texts, structured and tabular data, audio recordings, videos, 3D models, and geographic information (Trognitz et al., 2022), but only in a limited and selected number of formats, which are documented in the Format Recommendations by the CLARIN Standards Information System.

As of January 2026, ARCHE holds 52 top-level collections with a total of 516,783 individual resources amounting to 9.7 TB of data and accompanying metadata stored with 19M triples in the Resource Description Framework (RDF). The current technical setup of the repository, an open source set of microservices called the ARCHE Suite, was designed with flexible and fast handling of machine-readable and interoperable metadata at its core, and the principles of Linked Open Data in mind (Żółtak et al., 2022). To search and retrieve data and metadata hosted in ARCHE, a core API is available for creating, reading, updating, and deleting (CRUD) entities. Built on top of the core API, a growing set of bespoke dissemination services allows users to preview, download, serialise, or disseminate the stored data and metadata in various ways, including citation suggestions, IIIF presentation manifests, previewing images via a IIIF endpoint, and a bespoke viewer for 3D content.

In 2024, ARCHE's graphical user interface (GUI) underwent a major overhaul after following a user research study evaluating its usability. The new GUI not only features an updated visual appearance but also includes new search options, and leverages the ARCHE APIs for improved performance. To ensure sustainable preservation of FAIR research data, ARCHE maintains high quality standards, which require not only reliable technology, but also transparent policies, clear procedures, and essential human curation and interaction. To further increase the reusability and findability of the data stored in ARCHE, the respective metadata is shared via a public endpoint to allow external aggregators to harvest ARCHE metadata. This endpoint complies with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, Lagoze et al., 2002) and comes with a templating mechanism to allow for custom metadata mapping (Żółtak et al., 2022) from the ARCHE metadata schema to the ones preferred by the respective aggregators: CMDI for the CLARIN Virtual Language Observatory, DataCite for OpenAIRE, AO-Cat for ARIADNE, schema.org for the Google Dataset Search, and the Europeana Data Model (EDM) for Kulturpool, the Austrian aggregator for Europeana.

3.2 CLARIN B-Centre DH Graz

The Department of Digital Humanities (DDH, previously known as the Centre for information Modelling)[4] at the University of Graz is operating a trusted digital repository "Geisteswissenschaftliches Asset Management System" (GAMS)[5] and has been officially recognised as a CLARIN B-centre since 2020. At the heart of all considerations and developments stand the FAIR principles as well as a strong commitment to the open publication of data. These efforts led to the first recognition as a trusted digital repository with the Data Seal of Approval in 2014, which has since been renewed with the CoreTrustSeal in 2019 and 2024.

GAMS is an OAIS compliant asset management system for the management, publication, and long-term preservation of digital resources from the humanities, with a particular research focus on digital scholarly editions. In operation since 2003, the repository has been continuously developed and adapted to address the specific scholarly needs and technical challenges for research data in the humanities. This includes the application of community-specific modelling languages, vocabularies, and standards like the TEI guidelines as well as the exposure of standardised metadata and facsimiles via relevant interfaces. As an example, resources are delivered to aggregation services like the CLARIN-VLO, Europeana, Kulturpool, CorrespSearch, Nomisma, and others. Persistent identification and permanent citability is ensured by a dedicated Handle server. In contrast to many other data centres, the department's philosophy is to accompany research endeavours from proposal to publication throughout the whole data life cycle, ensuring compliance with legal and ethical norms as well as responsible and high-quality data management from beginning to end.

[4] https://digital-humanities.uni-graz.at/en/
[5] https://gams.uni-graz.at/

As of the beginning of 2026, the infrastructure hosts over 155,000 data objects organised across approximately 100 research projects. Thematically, there is a strong focus on digital scholarly editions and textual resources in general, as well as on digitised cultural heritage collections.

In addition to providing the repository and the publication platform, the department has been developing and sharing tools and workflows, offering project-specific consulting in DH methods and processes, and hosting seasonal training schools, postgraduate workshops, and academic conferences to promote knowledge transfer and digital humanities scholarship. DDH also offers a Master's degree in Digital Humanities, which introduces students to the practical implementation of DH-related skills in their daily research work.

3.3 CLARIN K-Centre Phonogrammarchiv

Since 2015, the *Phonogrammarchiv*[6] of the Austrian Academy of Sciences has been registered as a Knowledge Centre (K-centre) within CLARIN. As an Austrian K-centre the *Phonogrammarchiv* is part of CLARIAH-AT.

The *Phonogrammarchiv* was founded 127 years ago at what was then the Imperial Academy of Sciences with an early version of an open science approach: Right from the start, its goal was to produce recordings for reuse in academic research, with languages and dialects being among its main areas of interest. Today, the *Phonogrammarchiv* remains an active archive. It houses large collections of audio recordings – and, since 2002, also video recordings – from all over the world, registering global developments in language and dialect as well as music and other cultural expressions covering a period of more than a century. Its physical collections encompass about 35,000 audio and video carriers, while its digital collections amount to about 16,000 hours of archived recordings (digitised and born digital). Digitisation is still ongoing. Some of the *Phonogrammarchiv*'s holdings have been registered as UNESCO Memory of the World or Memory of Austria.

As the oldest sound archive in the world, the *Phonogrammarchiv* has a long tradition and rich experience in knowledge sharing and capacity building. As a CLARIN K-centre, it provides access to its collections (data and metadata), subject to legal and ethical clearance, either online, on demand, or onsite. While researchers have always been its main users, the *Phonogrammarchiv*'s aim now is to make its collections as widely available as possible, e.g., to educators, museums, the media, artists, and the general public. Particularly important in the *Phonogrammarchiv*'s activities are collaborations with the descendants of those who were recorded.

In addition to providing access to its resources, the *Phonogrammarchiv* offers advice and training in the following fields: preservation, restoration and digitisation of historical audio and video carriers, recording and transfer equipment; methods and technologies of AV-supported research in the social sciences and humanities with a focus on fieldwork; annotation and contextualization; provenance research and recirculation; legal and ethical issues. In October 2025, the *Phonogrammarchiv* moved to new premises after almost 100 years in its former location. This move has greatly enhanced its possibilities for offering training and workshops and accommodating guest researchers.

3.4 CLARIN K-Centre TRTC

The CLARIN Knowledge Centre for Terminology Resources and Translation Corpora (K-centre TRTC)[7] is one of the 38 knowledge centres within the CLARIN Knowledge Infrastructure[8].

K-centres play a significant role within the CLARIN infrastructure, providing methodological expertise in data collection and data management, supporting users in accessing relevant resources and services, and delivering training tailored to their specialised fields (van den Heuvel et al., 2022: 374).

Established in 2019, the K-centre TRTC is a single-site facility located at the Centre for Translation Studies at the University of Vienna. As one of the transversal knowledge centres within the CLARIN Knowledge Infrastructure (Vaičenonienė et al., 2024), it provides comprehensive coverage of terminology resources and translation corpora from a methodological perspective. This initiative builds

[6] https://www.oeaw.ac.at/en/phonogrammarchiv/certificates-and-awards
[7] See https://trtc.univie.ac.at
[8] As of January 2026; see the K-centre catalogue for full list of K-centres: https://www.clarin.eu/content/knowledge-centres

upon the significant and longstanding expertise in terminology science and translation studies at its host institution (Lušicky and Budin, 2024).

The K-centre offers a variety of resources, best practices, guidelines, and training materials related to translation-related resources, specifically terminology resources and translation corpora. It primarily focuses on providing access to termbases, with notable examples including higher education terminology and Austrian administrative terminology. Additionally, the K-centre organizes courses and training events and hosts guest researchers.

Like all K-centres, the K-centre TRTC operates a helpdesk, providing support to users on the preparation and documentation of terminology and translation corpora. This includes handling inquiries related to methods, tools, data, and guidance in seeking further expert support. The service is language-independent. In cases of particular language-dependent inquiries, the K-centre TRTC forwards the inquiry to one of the other K-centres in the network with that specific expertise. The K-centre TRTC addresses and assists a diverse range of users and stakeholders, both from within the CLARIN network and external entities. In terms of volume, it is comparable to most other K-centres, handling up to 50 human-processed requests per year (Vaičenonienė et al., 2024). The requests received by the K-centre span various application scenarios, reflecting the centre's broad scope. Beyond the academic audience, TRTC has supported professional associations, public bodies, and other organizations within and outside of the CLARIN network. An illustrative example of a wide-ranging request addressed to the K-centre TRTC led to a collaborative project with the Mongolian Academy of Sciences and the ACDH[9]. The primary objective of this project was to support the establishment of a terminology planning strategy and infrastructure for the Mongolian language. This initiative encompassed several key components, including capacity building and the establishment of a virtual terminology centre, thereby transferring the knowledge of building and operating a knowledge centre beyond the scope of the CLARIN network.

3.5 CLARIN K-Centre GermanAT

The CLARIN Knowledge Centre on German in Austria (K-centre GermanAT)[10] is one of the youngest K-centres in the CLARIN Infrastructure. Established in June 2025, this K-centre is also located at the Austrian Academy of Sciences, specifically at the Austrian Centre for Digital Humanities. GermanAT provides expertise in language variation in the languages and varieties spoken in Austria – primarily, but not exclusively, varieties of German. It also hosts a variety of resources on German in Austria, and provides know-how and consulting for research into language variation in general.

GermanAT emerged as a K-centre out of the recognition that many experts on various aspects of German in Austria are based at the ACDH and that the institute itself has a scholarly tradition in this field. Many large-scale projects focusing on language variation in Austria, particularly on German varieties, have been conducted by scholars at this institute or in cooperation with the ACDH, and the institute hosts a diverse array of language resources. GermatAT also possesses a significant expertise in transcription, annotation and handling of language data and corpora, lexicography and other research fields related to German in Austria.

Many user requests addressed to GermanAT concern access to specific datasets and corpora dealing with German and other languages in Austria. These requests frequently come from students and PhD candidates at universities, who wish to work with the institute’s resources for seminar papers or final theses.

Among the most requested resources are the Austrian Media Corpus (amc) and the corpus of the Special Research Programme (SFB) FWF F60: “German in Austria: Variation - Contact - Perception” (SFB DiÖ). The amc is a corpus of print media in Austria that covers almost the entirety of print news media in Austria over the last decades (Ransmayr and Pirker 2023). With almost 13 billion tokens as of the time of writing, it is one of the largest written corpora in German, and users regularly request access to it as a reference for the written standard language of Austria. The corpus of the SFB DiÖ is the other highly requested resource, covering spoken language in Austria. The data in this corpus ranges from

[9] https://trtc.univie.ac.at/news-and-events/terminology-planning-strategy-and-terminology-infrastructure/
[10] https://k-centre-germanat.acdh.oeaw.ac.at/en/

controlled elicitation tasks to long stretches of spontaneous interactions between friends, targeting a range of registers from dialect to spoken standard in Austria (see Lenz 2023). Access to this corpus is often requested by researchers interested specifically in sociolinguistic variation in Austria. Data from this resource has already been used in several research projects, such as the Sound Effects project funded by the Swiss National Science Fund.

The expertise and resources of GermanAT are not limited to contemporary German. The ACDH also hosts the long-term research project *Language Dynamics in Austria* (DynAT), which focuses on linguistic dynamics of German in Austria in the 20th and 21st century. As part of this project, historical language data (especially dialect data) from the 20th century is also being processed. Comparing this historical data with the institute's contemporary language data enables analyses of language change over the course of more than 100 years.

This unique combination of experts and resources on language(s) in Austria served as the catalyst for the recent establishment of the K-centre GermanAT, with the aim of sharing these resources and skills with the broader community.

4 Building and Fostering a Community: Funding, Dissemination and Outreach

Based on the CLARIAH-AT consortium's experience and an open consultation of the Austrian professional DH community, the consortium developed a vision for the further advancement of Austrian Digital Humanities, along four pillars: (i) research infrastructures and networks, (ii) research data and repositories, (iii) methods, services and tools, and (iv) education, training and knowledge transfer (CLARIAH-AT Consortium, 2021). Closely tied to these focal points, the consortium has engaged in several activities to put this vision into action, such as issuing project-funding calls, providing funding opportunities for early-stage researchers, and hosting training events and workshops. Members of the CLARIAH-AT consortium are also involved in teaching activities, where they integrate resources and tools from the infrastructure. Furthermore, CLARIAH-AT members actively participate in European research projects, including CLARIN flagship initiatives such as *ParlaMint* (Erjavec et al. 2024) and *PressMint*[11]. Within these projects, Austrian partners contribute annotated Austrian parliamentary and historical newspaper datasets to interoperable and comparable corpora, thereby strengthening Austria's role and increasing its visibility within European research infrastructures.

Another tangible outcome of CLARIAH-AT's reputation and community-building efforts is its close cooperation with Austrian cultural heritage institutions. This is exemplified by the Natural History Museum Vienna, which recently became an observer member of the consortium (Scholger, 2024). The Natural History Museum operates *Kulturpool*, a central search portal for digitised cultural heritage collections from Austrian GLAM institutions. *Kulturpool* also serves as the Austrian national aggregator for *Europeana*.

In addition to these activities, members of the consortium are actively involved in various bodies and initiatives at the European level (e.g., CLARIN and DARIAH committees, DARIAH Working Groups), which enhances opportunities for collaboration, as well as dissemination and outreach.

4.1 Funding Schemes

The CLARIAH-AT consortium provides two types of funding: project funding via dedicated calls and early career scholarships.

4.1.1 Project Funding

Since 2022, four calls have been issued with a total of 51 projects funded. There were funding calls for large-scale projects and small-scale projects. The project proposals for the calls 2024 and 2025 were evaluated by three reviewers from the CLARIAH-AT consortium based on the following criteria:

- Relevance to the call: The extent to which the proposed project falls within the scope of the call, including its alignment with the DHA2021+ strategy and the anticipated benefits for the Austrian DH community.

[11] https://www.clarin.eu/pressmint

- Feasibility of planning: The clarity, coherence, and feasibility of the proposed project schedule and financial plan.
- Definition and sustainability of outputs: The degree to which the project defines clear and feasible outputs, as well as sustainable strategies for publishing these outputs and ensuring their long-term availability and reusability
- Relevance to CLARIN and/or DARIAH: The relevance and potential usefulness of the project's outputs in relation to the activities and focus areas of CLARIN and/or DARIAH.
- Overall quality: The overall scientific and methodological quality of the project proposal.

Funded project topics range from the implementation of services to the development of new resources, enhancement of existing resources, formulation of citation guidelines, and organisation of workshops. Figure 1 provides an overview of the number of projects funded per call and per institution. Selected projects are discussed in more detail in the following section.

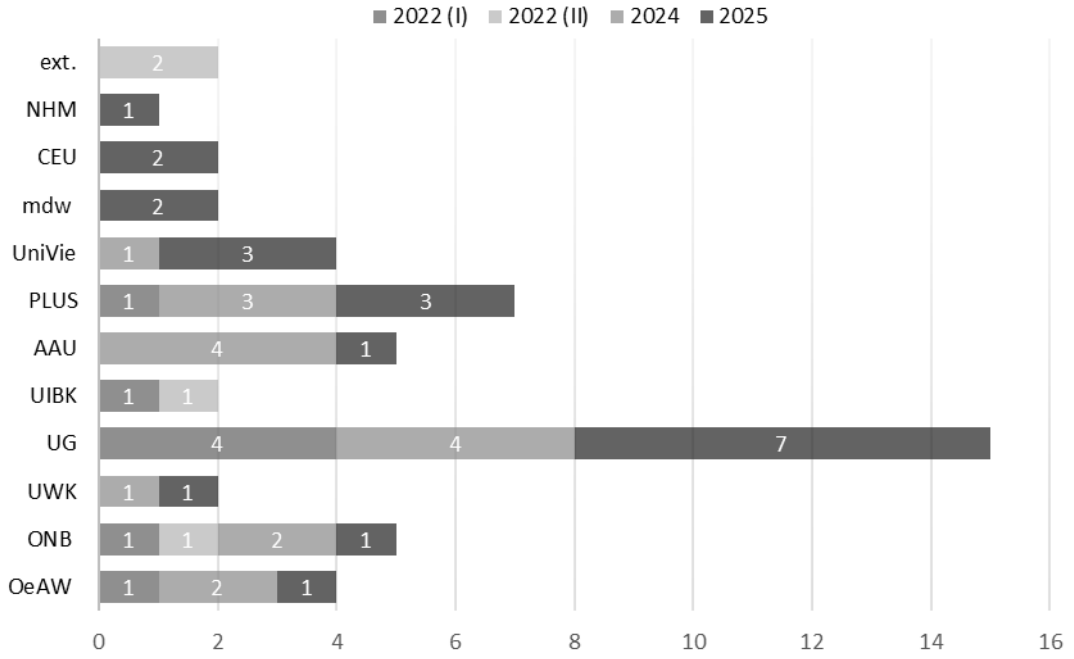

Figure 1: Overview funded projects per consortium partner for 2022 - 2025[12], (adapted, Woldrich, 2025)

4.1.2 Early Career Scholarships

In addition to project funding, CLARIAH-AT offers early career scholarships for researchers based at an Austrian institution, to carry out their research in the field of Digital Humanities, Cultural Heritage, and related fields. These scholarships are available in the form of Transkribus Scholarships, a licence pool for software, and travel bursaries[13]. With a CLARIAH-AT Transkribus scholarship, early career researchers can receive free Transkribus credit packages for the (handwritten) text recognition of their sources within the framework of their thesis, dissertation, paper, or individual project, provided these are not funded via third party funding. Through the licence pool for software, early career researchers can obtain bursaries for software licences needed for their research.

The third type of scholarship consists of travel bursaries. The bursaries aim at students or young scholars who want to present a part of their current research in the fields of Digital Humanities, Cultural Heritage, or related fields in the form of a paper or poster at a conference, or who want to acquire research-relevant skills by participating in a workshop.

[12] Abbreviations in the figure: OeAW - Austrian Academy of Sciences, ONB - Austrian National Library, UWK - University for continuing education Krems, UG - University of Graz, UIBK - University of Innsbruck, AAU - University of Klagenfurt, PLUS - University of Salzburg, UniVie - University of Vienna, mdw - University of music and performing arts Vienna, CEU - Central European University, NHM - natural history museum vienna, ext - externals.

[13] More information is provided at: https://clariah.at/en/funding-opportunities-for-junior-researcher/

Between 2024 and 2025, 23 early career scholarships were awarded, most of them travel bursaries: 5 in 2024 and 16 in 2026. In addition, one software licence scholarship was awarded in 2025 and one Transkribus scholarship in 2026. The increasing number of travel bursaries in 2025 may also indicate that CLARIAH-AT and its Early Career Scholarships are now better known among early career researchers in Austria.

4.2 Selected Projects

In the following we describe two funded CLARIAH-AT projects in more detail: The *CLARIAH-AT Platform* project, because it is a core project for the whole CLARIAH-AT consortium, and the project *Unlocking the full potential of textual resources – Application of FAIR data principles through CLARIN Federated Content Search Implementation* because it is directly linked to a CLARIN core service.

4.2.1 CLARIAH-AT Platform

At the core of *CLARIAH-AT Platform* project[14] is the redesign and expansion of the existing website clariah.at into a comprehensive platform. This platform will serve as a central information and presentation hub for projects, datasets, tools, services, training materials, and publications. Furthermore, the platform will improve the discoverability and visibility of Austrian research resources, projects, and services, while ensuring seamless integration with the European Research Infrastructures CLARIN and DARIAH-EU and the European Open Science Cloud (EOSC).

The technical set up for the platform uses a git-based system with static HTML generation, combined with CMS support via the Keystatic Framework. This enables flexible, field-specific content structuring (e.g., for events or metadata) and long-term maintainability. A prototype is already running, and the final version will be launched at the end of 2026.

4.2.2 Unlocking the full potential of textual resources

Within the project *Unlocking the full potential of textual resources – Application of FAIR data principles through CLARIN Federated Content Search Implementation*[15], textual resources already available at the ACDH will be made available through the CLARIN Federated Content Search (FCS). The FCS[16] is developed, hosted and maintained by CLARIN and allows users to search various text resources hosted at different locations from a central access point – both for keywords and more complex patterns (e.g., collocations). It supports different data formats and linguistic annotations. To make the ACDH resources searchable via FCS, an FCS endpoint needed to be implemented. An instance of the mquery-sru was deployed and registered in the FCS aggregator instance. Furthermore, an automated pipeline for the creation of NoSketch Engine index files for the mquery-sru instance was created. At present, 23 textual resources, including corpora and digital editions, are integrated and searchable via the CLARIN FCS. The FCS will be integrated into the new CLARIAH-AT Platform.

[14] https://clariah.at/en/projects/clariah-at-platform/
[15] https://clariah.at/en/projects/clarin-fcs-implementation/
[16] https://www.clarin.eu/content/content-search

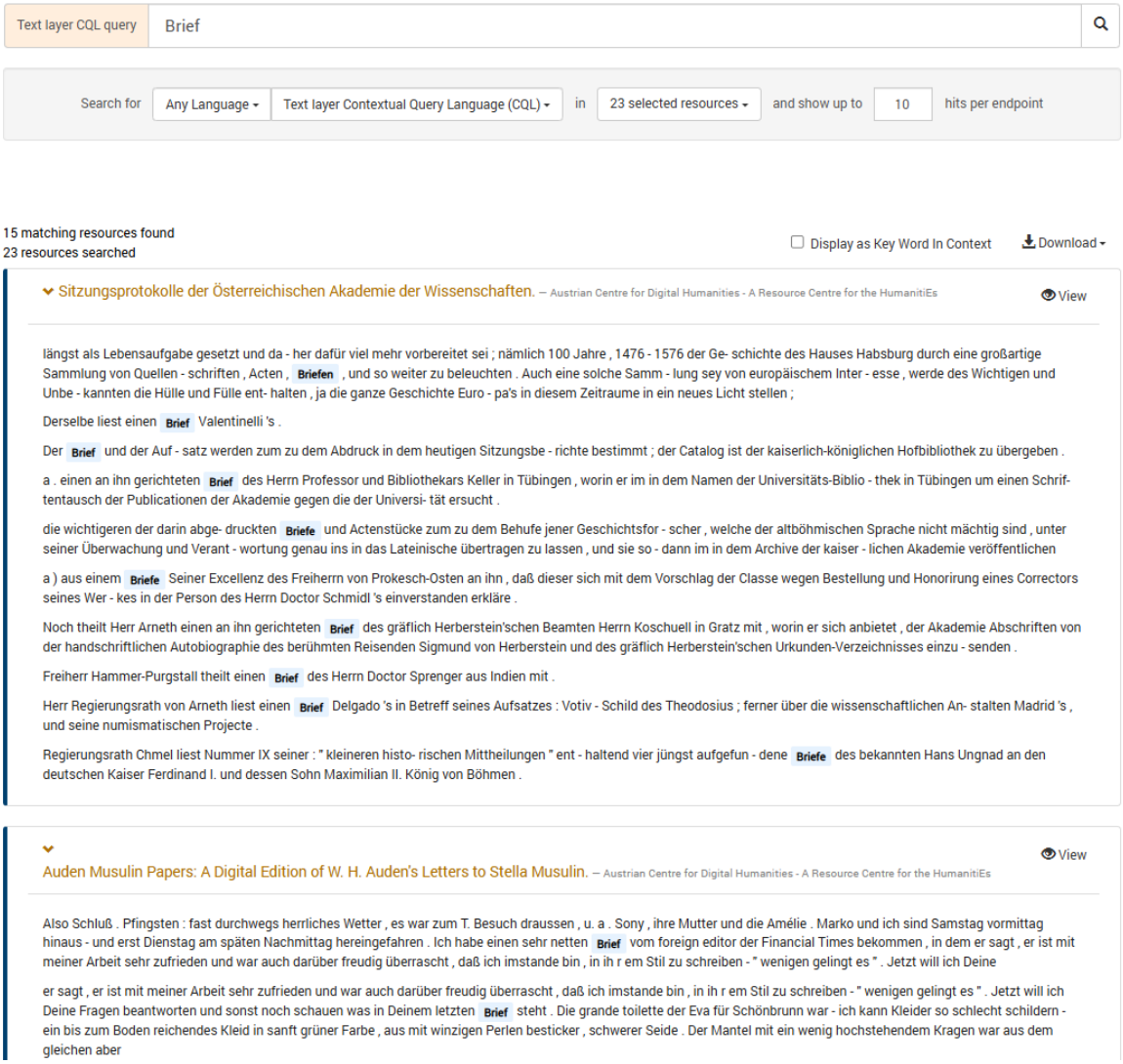

Figure 2: FCS search within Austrian resources

4.3 Dissemination and Outreach

A central objective of CLARIAH-AT is active engagement with the research community, including the dissemination of news, events, and funding opportunities. To this end, the consortium uses several dissemination channels: the CLARIAH-AT website, the CLARIAH-AT LinkedIn account, and the Digital Humanities Austria (DHA) mailing list. The CLARIAH-AT website is the central communication hub providing an overview of past and upcoming events, the latest news, and funding opportunities. At the time of writing, the website is undergoing revision. The consortium plans to roll out a new, improved version by the end of 2026, the so-called *CLARIAH-AT Platform*, discussed in more detail in the project section above. Since January 2025, CLARIAH-AT has also maintained a LinkedIn account and has, as of January 2026, over one thousand followers. The Digital Humanities Austria mailing list functions as a bidirectional communication channel, enabling the dissemination of information on events, open calls and other announcements, as well as the sharing of upcoming activities from the Austrian digital humanities landscape. It currently has 369 subscribers.

To foster skill development and to engage with the community, CLARIAH-AT has established a variety of event series: The *CLARIAH-AT Summer Schools*, a training event series targeting students and researchers in DH. The CLARIAH-AT Summer Schools were established in 2023 and are held each year in a different location in Austria with a different thematic focus. The first edition, titled *Explore Salon*, was held in Vienna in 2023 (Charvat et al, 2025), followed by the Summer School *Vom Archiv*

zum Algorithmus. Digital Humanities in der Praxis organised at the University of Salzburg in 2024 (Döring, 2024) and by the Summer School on Machine Learning for Digital Scholarly Editions organised at the University of Graz in 2025. Additionally, each year the *CLARIAH-AT Roadshow* invites project investigators to showcase (outcomes of) their CLARIAH-AT funded projects. In 2025, the latest roadshow-edition was co-organised and co-located with the first *Digital Humanities Festival* at the Central European University (CEU). CLARIAH-AT also initiated the workshop series "Austrian Meeting on Digital Linguistics" in 2023, held each year in conjunction with the *Österreichische Linguistiktagung (ÖLT)*.

Furthermore, CLARIAH-AT is present at international conferences, especially at the CLARIN and DARIAH Annual Events. In 2025, CLARIAH-AT served as the local host for the CLARIN Annual Conference in Vienna[17].

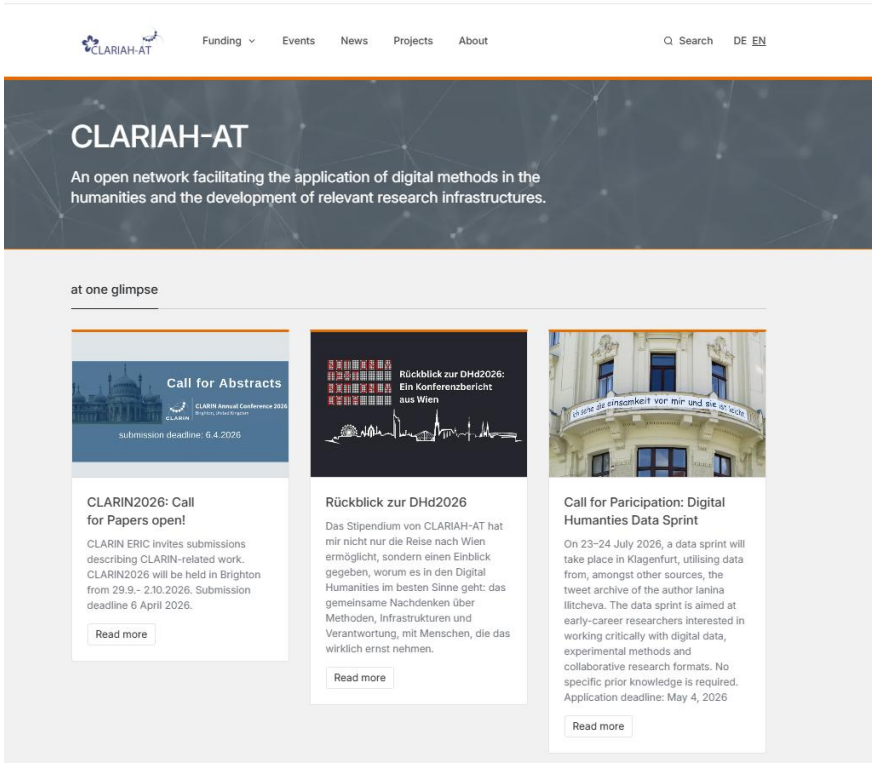

Figure 3: Screenshot of the CLARIAH-AT website

5 Concluding Remarks and Outlook

This contribution has presented the CLARIAH-AT consortium, its infrastructure, and its wide range of activities. A new CLARIAH-AT website is scheduled to be launched by the end of 2026, with the aim of improving the discoverability and visibility of Austrian contributions to European research infrastructures, including via the SSH Open Marketplace and the CLARIN Federated Content Search. Furthermore, it is planned to connect other Austrian textual resources, hosted at other CLARIAH-AT member institutions to the CLARIN Federated Content Search. In addition, the Digital Humanities Austria Strategy 2021+ will be evaluated and updated to reflect developments in the research landscape, including the increasing relevance of artificial intelligence.

Acknowledgments

This work was partly supported by CLARIAH-AT.

[17] https://www.clarin.eu/event/2025/clarin-annual-conference-2025#about-the-clarin-annual-conference

References

Alram, M., Benda, C., Ďurčo, M., Mörth, K., Wentker, S., Wissik, T., Budin, G., Kaiser, M., Majewski, K., Palme, B., Scholger, W., Stigler, H., Thaller, M. (2015). *DH-Austria-Strategie*. https://epub.oeaw.ac.at/DH-AUSTRIA-STRATEGIE-2015

ARCHE (2025). "2028-12-03 - ARCHE - CoreTrustSeal Requirements 2023-2025." DataverseNL. https://doi.org/10.34894/SNEQ7T

CLARIAH-AT Consortium (2021). *Digital Humanities Austria Strategy 2021+. Four Guidelines for Digital Humanities in Austria*. https://gams.uni-graz.at/o:clariah.dha-strategie-2021-en

Döring, K. (2024). CfA: CLARIAH-AT Summer School Digital Humanities „Vom Archiv zum Algorithmus. Digital Humanities in der Praxis", 9–13 September 2024. Digital Humanities an der Universität Salzburg. https://doi.org/10.58079/w84s.

Erjavec, T., Kopp, M., Ljubešić, N. et al. (2024). ParlaMint II: advancing comparable parliamentary corpora across Europe. Language Resources & Evaluation 59, 2071–2102. https://doi.org/10.1007/s10579-024-09798-w.

Mörth, K. and Wissik, T. (2018). 4. Digitale Sprachressourcen in Österreich. *Digitale Infrastrukturen für die germanistische Forschung*, edited by H. Lobin, R. Schneider and A. Witt, Berlin, Boston: De Gruyter, 2018, 73–88. https://doi.org/10.1515/9783110538663-005

Lenz, A. N. (2023). Korpora zur deutschen Sprache in Österreich. *Korpora in der germanistischen Sprachwissenschaft: Mündlich, Schriftlich, Multimedial*, edited by A. Deppermann, C. Fandrych, M. Kupietz, and T. Schmidt, Vol. 2022, pp. 53–69. Walter de Gruyter GmbH. https://doi.org/10.1515/9783111085708-004.

Lušicky, V. and Budin, G. (2024). Terminologie und Sprachdaten im Lichte der Sprachenpolitik. *Sprachenpolitik in Österreich: Bestandsaufnahme 2021*, edited by R. de Cillia, M. Reisigl and E. Vetter, Berlin: De Gruyter, pp. 357–376. https://doi.org/10.1515/9783111329130

Ransmayr, J., and Pirker, H. (2023). Das Austrian Media Corpus (AMC): Inhalte, Zugang und Möglichkeiten. *Zeitschrift für germanistische Linguistik*, 51(1), 203–212. https://doi.org/10.1515/zgl-2023-2007.

Scholger, W. (2024). *Sustaining a Digital Humanities Infrastructure and Network for Austria*. https://www.dariah.eu/activities/impact-case-studies/sustaining-a-digital-humanities-infrastructure-and-network-for-austria/

Trognitz, M. (2021). ARCHE, the Austrian B-Centre for Digital Humanities and Cultural Heritage. *Tour de CLARIN Volume Four*, edited by J. Lenardič, F. Frontini, and D. Fišer, 78–83. https://doi.org/10.5281/zenodo.7019259.

Trognitz, M., Ďurčo, M. and Mörth, K. (2022). Text Technology for the Digital Humanities. *CLARIN: The Infrastructure for Language Resources*, edited by D. Fišer and A. Witt, Berlin, Boston: De Gruyter, 2022, pp. 223-248. https://doi.org/10.1515/9783110767377-009.

Van Den Heuvel, H., Oostdijk, N., Rowland, C. and Trilsbeek, P. (2022). The CLARIN Knowledge Centre for Atypical Communication Expertise. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, ELRA, 3312–3316.

Vaičenonienė, J., Kren, M., Lušicky, V. and Vandeghinste, V. (2024). Transnational Research Infrastructure: A Journey Through CLARIN Knowledge Centres. *Digital Humanities in the Nordic and Baltic Countries Publications*, 6(1). https://journals.uio.no/dhnbpub/article/view/11520/9554

Woldrich, A. (2025). CLARIAH-AT Project Overview. CLARIN Bazaar. CLARIN Annual Conference 2025 (CLARIN2025), Vienna, Austria. Zenodo. https://doi.org/10.5281/zenodo.18415463

Żółtak, M., Trognitz, M. and Ďurčo, M. (2022). ARCHE Suite: A Flexible Approach to Repository Metadata Management. *Selected Papers from CLARIN Annual Conference 2021*, Vol. 189, Linköping University Electronic Press, 190–199. https://doi.org/10.3384/ecp18917.

Reading LEMI: A Corpus-Based Tool to Support Literacy in an Under-Resourced Language

Karla Csuros
West University of Timisoara
karla.csuros@e-uvt.ro

Madalina Chitez
West University of Timisoara
madalina.chitez@e-uvt.ro

Aura Cristina Udrea
University POLITEHNICA Bucharest
aura.cojocaru@upb.ro

Mihai Dascalu
University POLITEHNICA Bucharest
mihai.dascalu@upb.ro

Roxana Rogobete
West University of Timisoara
roxana.rogobete@e-uvt.ro

Andreea Dinca
West University of Timisoara
andreea.dinca@e-uvt.ro

Abstract

LEMI (**Le**ctură pentru **mi**ne, 'Reading for Me') is a corpus-based literacy support platform for Romanian, an under-resourced language in which nearly 42% of students aged 6–15 are classified as functionally illiterate despite a general literacy rate of 99%. The platform combines a curated repository of 250 children's literary texts with an automatic readability analysis tool based on four interpretable linguistic parameters: Average Sentence Length (ASL), Percentage of Complex Words (PCW), Unique Content Word Density (UCWD), and Word Diversity (WD), combined through a weighted scoring formula designed to avoid the syllable-based biases of English readability metrics and better suited to the morphological complexity of Romanian. This paper presents LEMI's design and documents its application across five empirical contexts: classroom-based validation with primary school students, linguistic complexity assessment of school textbooks, contrastive analysis of readability in translated versus original children's literature, lexical modernization of canonical literary texts, and in-service teacher training for corpus-informed text adaptation. Across these contexts, LEMI has functioned as a diagnostic, benchmarking, and decision-support tool, producing interpretable outputs that complement rather than replace professional judgment. The platform adheres to FAIR data principles and implements a metadata schema supporting federated content search and reuse across European research infrastructures. By making both tools and data openly available, LEMI contributes a transparent, reusable resource for readability research and literacy support in an under-resourced language.

1 Introduction

Literacy development in early education relies heavily on access to age-appropriate, linguistically accessible texts. For under-resourced languages like Romanian, tools to evaluate or adapt children's literature for readability are scarce. Romania faces significant literacy challenges: nearly 42% of students aged 6-15 are classified as functionally illiterate, despite a general literacy rate of 99% (Iliescu & Airinei, 2022; Stanciu et al., 2024). This disparity highlights the need for innovative, language-specific solutions tailored to Romanian learners (Balea et al., 2024; Hurjui, 2024).

Romanian, the sixth most spoken language in the European Union, is considered under-resourced from a natural language processing (NLP) perspective due to the limited availability of digitized linguistic data, language technologies, and computational tools. Compared to other Romance languages such as French (d'Hoffschmidt et al., 2020; Étienne et al., 2019; Sohail et al., 2022) and Spanish (Badenes-Olmedo et al., 2026; Gutiérrez-Fandiño et al., 2021), Romanian has historically lacked robust linguistic datasets and NLP infrastructure. Initiatives such as the *Representative Corpus of the Romanian Language* (Midrigan-Ciochina et al., 2020), *CoRoLa* (Mititelu et al., 2018; Tufiș et al., 2016) and *ROMBAC* (Ion

Karla Csuros, Madalina Chitez, Aura Cristina Udrea, Mihai Dascalu, Roxana Rogobete, and Andreea Dinca 2025. Reading LEMI: A Corpus-Based Tool to Support Literacy in an Under-Resourced Language. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 63–74. DOI: https://doi.org/10.3384/ecp222

et al., 2012) have begun to address this gap; however, these resources remain limited in scope and accessibility. Furthermore, access to educational materials, particularly textbooks and digital tools for both native and second-language learners of Romanian, remains uneven, especially outside Romania. These challenges underscore the need for targeted efforts.

In this study, we introduce LEMI (based on the Romanian phrase ”**Le**ctură pentru **mi**ne”, meaning ”Reading for Me”), a corpus-based literacy support platform designed to help educators and researchers assess and deliver Romanian texts tailored to the reading abilities of children aged 7-11. LEMI combines corpus linguistics with NLP techniques to evaluate and suggest suitable literary content. It is built to serve both pedagogical and research purposes, and adheres to FAIR data principles, implementing a structured metadata schema compatible with federated content search, and is designed for reuse across European research infrastructures. The LEMI platform is publicly available at https://www.lemi.ro/.

2 Tool Description

2.1 Overview

LEMI is a cross-platform, web-based application (accessible at https://www.lemi.ro/) that offers two main functionalities: (1) access to a curated repository of Romanian children’s literature, and (2) an interface for the automatic readability analysis of user-submitted texts. The platform is designed to help educators, parents, and researchers select and evaluate children’s literature for ages 7-11.

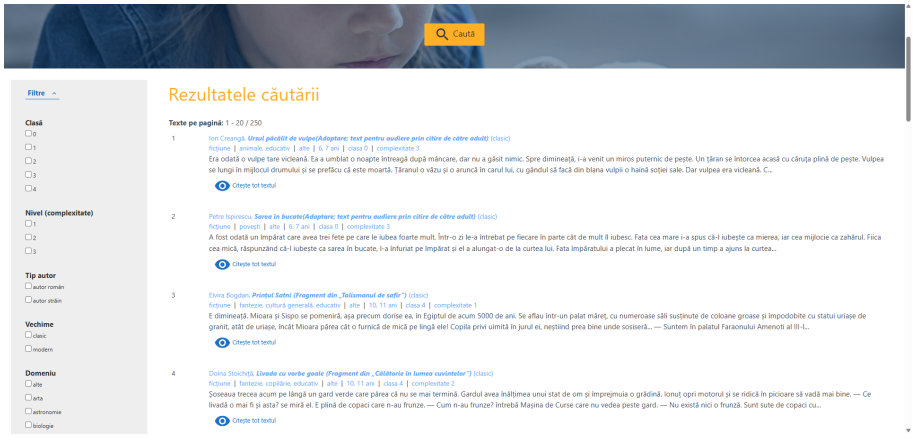

Figure 1: **LEMI search interface displaying filtered results from the repository**. The left panel shows available filter options, including grade level (Clasă), complexity tier (Nivel complexitate), author type (Tip autor), period (Vechime), and thematic domain (Domeniu). The main panel lists matching texts with metadata tags indicating author, age range, grade, and complexity level, alongside a short excerpt and a link to the full text. This interface allows educators and researchers to retrieve texts matching specific pedagogical or research criteria.

Users can search the LEMI Repository (accessible at https://www.lemi.ro/search.php/), which contains annotated texts tagged by author, genre, theme, age group, and estimated reading difficulty. The collection includes works by both Romanian and international authors, spanning classic and contemporary literature. Each entry provides metadata such as domain (e.g., adventure, nature), grade-level targeting, and complexity tier. Authenticated users can filter texts, view highlighted keywords, and download full-text PDFs customized to their search (see Figure 1). PDF was chosen as the primary export format to allow educators and parents to print materials directly without requiring technical expertise; texts are also fully selectable within the platform prior to export, enabling copy-paste retrieval for users who prefer plain text.

In parallel, users can upload their own texts to receive a readability report (via https://www.lemi.ro/text_analysis.php). The analysis is powered by a custom readability formula that integrates surface-level and linguistic features (e.g., sentence length, lexical diversity). The platform returns a predicted reading level and provides metadata tags to help users align content with curricular goals (see Figure 2).

LEMI supports three user roles: visitors, registered users, and administrators, each with dedicated

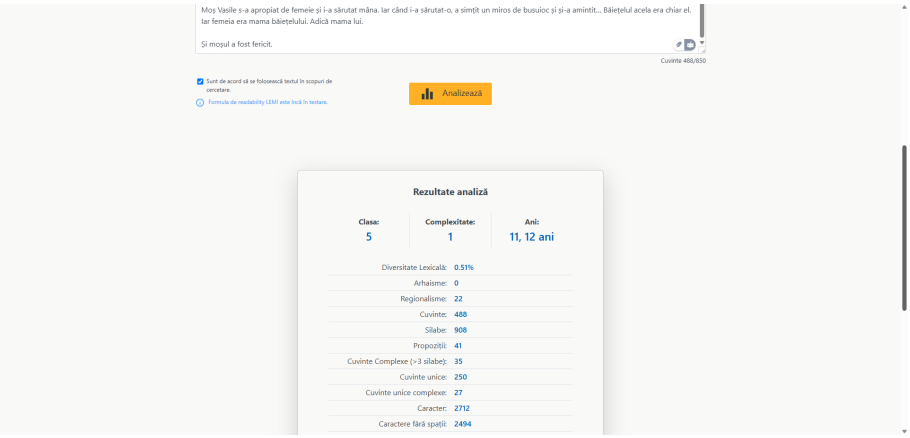

Figure 2: **LEMI readability analysis output for a user-submitted text**. The results panel displays the predicted grade level (Clasă), complexity tier (Complexitate), and estimated reader age range (Ani). Additional metrics include lexical diversity (Diversitate Lexicală), archaisms (Arhaisme), regionalisms (Regionalisme), word count (Cuvinte), syllable count (Silabe), sentence count (Propoziții), complex words of three or more syllables (Cuvinte Complexe), unique words (Cuvinte unice), unique complex words (Cuvinte unice complexe), character count with and without spaces (Caractere; Caractere fără spații). These indicators collectively form the basis of LEMI's weighted readability score.

permissions. Administrators manage corpus entries and metadata through a backend content management system (CMS), which supports full CRUD (Create, Read, Update, Delete) functionality. The platform is designed with usability and educational applicability in mind, adhering to key principles of accessibility and scalability.

2.2 Corpus and Readability Assessment

At the core of the LEMI platform is a curated corpus of Romanian children's literature, compiled from school textbooks, educator recommendations, and publicly available sources. In line with Romanian copyright legislation (Romanian Parliament, 1996), the repository does not reproduce works in full; rather, it provides representative text excerpts used exclusively for non-commercial educational and research purposes. The corpus is also publicly available on *GitHub*[1] under a non-commercial Creative Commons license (CC BY-NC).

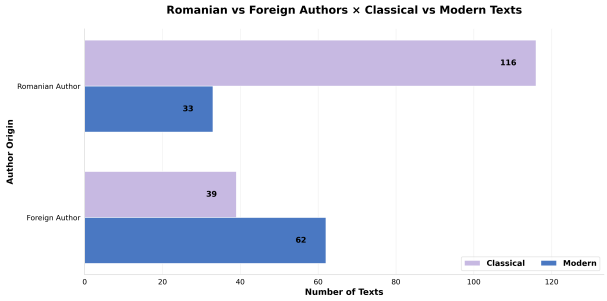

Figure 3: **Distribution of texts in the LEMI corpus by author origin and literary period**. The grouped bar chart compares the number of classical (lighter bars) and modern (darker bars) texts written by Romanian authors (top) versus foreign authors (bottom). Romanian authors contribute a larger share of modern texts (n=116) relative to classical ones (n=33), while foreign authors are more evenly distributed across periods (classical: n=39; modern: n=62). This distribution reflects the corpus design decision to balance canonical and contemporary literature across national and international sources.

The corpus currently comprises over 92,000 words across 250 texts, stored in UTF-8-encoded plain text. The texts cover four primary grade levels (grades 1-4) and include both original Romanian works and translated literature from classical and contemporary authors (see Figure 3). Each text is enriched

[1] https://github.com/chia-AR/LEMI-Romanian-children-literature-corpus

with structured metadata, including author, genre, thematic domain, intended age group, and estimated reading difficulty. Texts are further classified into three readability tiers (easy, medium, challenging) within each grade level. The metadata schema is structured to support Virtual Language Observatory (VLO)-compatible discovery, enabling integration with federated content search and reuse across CLARIN research infrastructures (de Jong et al., 2018).

Readability assessment within LEMI is performed using a custom, linguistically motivated scoring formula. As the formula is part of an in-development proprietary system developed by the authors which is integrated into the publicly available platform, the exact parameter weightings cannot be fully disclosed at this time due to intellectual property considerations; detailed implementation aspects can however be provided upon reasonable request for research purposes. The approach combines four interpretable features: Average Sentence Length (ASL), Percentage of Complex Words (PCW), Unique Content Word Density (UCWD), and Word Diversity (WD), each computed using standard linguistic preprocessing and aggregated through a calibrated weighted scoring scheme normalised on a three-point scale (easy, medium, challenging), allowing consistent readability classification across texts. These features were selected based on their relevance to early literacy development and their robustness across different text types. Unlike the traditional Flesch (Flesch, 1948) and Flesch–Kincaid formulas (Kincaid et al., 1975) for English, which rely only on sentence length and syllable counts, the LEMI formula integrates elements of lexical sophistication and diversity. This design avoids syllable-based biases that overestimate difficulty in morphologically rich languages like Romanian.

The readability pipeline is designed to be modular and extensible, supporting the integration of additional linguistic features or external resources. Beyond formula-based scoring, LEMI supports advanced corpus analysis through interoperability with external NLP frameworks, enabling further exploration of syntactic and lexical complexity patterns in children's literature (Figure 4).

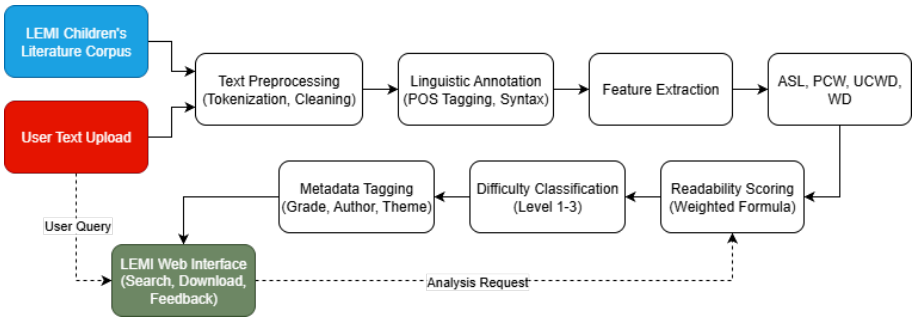

Figure 4: **LEMI processing pipeline for readability assessment**. Input enters the system from two sources: the LEMI Children's Literature Corpus and user-uploaded texts. Texts undergo text preprocessing (tokenization and cleaning), followed by linguistic annotation (POS tagging and syntactic parsing) and feature extraction, yielding four interpretable parameters: Average Sentence Length (ASL), Percentage of Complex Words (PCW), Unique Content Word Density (UCWD), and Word Diversity (WD). These features feed into a weighted readability scoring formula, which produces a difficulty classification on a three-point scale (Level 1–3) and metadata tagging by grade, author, and theme. User queries submitted via the LEMI web interface (search, download, feedback) trigger an analysis request that follows the same pipeline.

3 Applications

LEMI serves multiple applications across education, policy and research domains. In educational settings, teachers use the platform to filter texts by linguistic complexity level or generate input-text readability reports tailored to classroom needs. By aligning content with curricular goals through metadata tagging and automated analysis tools, LEMI supports personalized lesson planning while reducing reliance on printed materials in rural schools. The platform also contributes to the development of curricula in Digital Humanities by providing educators and researchers with tools to analyze and adapt Romanian children's literature for diverse educational contexts. For example, LEMI can be integrated into university courses focused on computational linguistics or language pedagogy, offering students hands-on experience with corpus-based analysis and readability metrics. From a policy perspective, LEMI informs revisions to Romania's National Assessment framework by advocating for readability-adjusted

exams that better reflect students' capabilities across diverse regions. In research contexts, LEMI enables cross-linguistic studies on readability metrics for under-resourced languages by providing a reusable and reliable dataset.

This section illustrates how LEMI has been employed across multiple empirical studies as a reusable, corpus-based platform for readability research and educational practice in Romanian. We summarize concrete research contexts in which LEMI has already been applied, demonstrating its versatility across educational, linguistic, and societal domains.

3.1 LEMI as a Framework for Classroom-based Validation with Primary School Students

Beyond its multifaceted potential for corpus analysis, LEMI has been empirically validated in authentic educational settings. A classroom-based study involving 256 primary school students (3rd and 4th graders; 187 urban, 69 rural) was conducted to evaluate the alignment between LEMI's automated readability predictions and human reader judgments (Oravitan et al., 2023). Four texts were selected from the LEMI repository (one Romanian classic, one Romanian contemporary, one foreign classic, and one foreign contemporary), each previously assigned a complexity score by the platform. Students were asked to rate the perceived difficulty of each text on a five-step scale: 1 (easy), 1.5 (fairly easy), 2 (so-and-so), 2.5 (fairly challenging), and 3 (challenging), and to indicate whether they would recommend the text to a peer. The results indicated a strong correspondence between automated scores and reader evaluations, with a favorability rate of 96.48%, reflecting the proportion of individual student ratings consistent with the LEMI-assigned complexity tier across all four texts. This high level of alignment provides empirical support for the pedagogical reliability of LEMI and confirms its suitability for classroom use. This validation study highlights LEMI's capacity to bridge computational readability modeling and real-world educational practice, reinforcing its value as a trustworthy decision-support tool for teachers working with heterogeneous reading abilities.

3.2 LEMI as a Textbook Assessment and Validation Tool

LEMI has been employed as a textbook assessment tool in corpus-informed research investigating linguistic overload in Romanian lower secondary education (Chitez et al., 2024a). In this study, LEMI was used to evaluate the linguistic complexity of 6th-grade textbooks in both language-based and non-language-based subjects, including Romanian Language and Literature and Mathematics, to determine whether approved textbooks align with students' cognitive and linguistic development levels. Within this research context, LEMI supported the systematic assessment of textbook readability by providing grade-level estimates for reading texts and instructional components. Across multiple textbook samples, the platform found that most analyzed materials exceeded the intended 6th-grade level, with LEMI scores frequently indicating suitability for 7th- or 8th-grade readers. This pattern was observed not only in literary reading texts, but also, and more consistently, in instructions and task formulations, which substantially increased overall text difficulty. LEMI was further used to compare textbooks across publishers, enabling quantitative differentiation between materials with moderate linguistic density and those exhibiting pronounced overload. For Romanian Language and Literature textbooks, LEMI-based assessment showed considerable variation: some textbooks approached age-appropriate readability, while others consistently showed elevated scores. In Mathematics textbooks, LEMI indicated uniformly high linguistic complexity, reflecting dense explanations, abstract terminology, and multi-step task formulations that place significant demands on reading comprehension. At the applied level, LEMI served as a diagnostic tool within a broader framework for analyzing linguistic textbooks. By identifying cases where textbook readability exceeded curricular expectations, the platform supported evidence-based evaluation of textbook suitability. LEMI highlighted specific sources of difficulty, such as excessive word density, complex terminology, and linguistically demanding task design. The use of LEMI made it possible to compare textbooks not only within a single discipline but also across disciplines, revealing that linguistic overload is not limited to language-focused subjects.

3.3 LEMI in Conjunction with Linguistic Complexity Tools and Frameworks

One of the first studies employing LEMI for computational linguistics research investigated how a newly developed readability platform for Romanian children's literature can be used alongside existing linguistic complexity frameworks to support readability research in an under-resourced language (Chitez et al., 2024b). LEMI was used alongside the ReaderBench framework (Dascalu et al., 2017) to explore deeper syntactic and lexical dimensions of text complexity that extend beyond surface-level readability indicators. Through this parallel analysis, textual features such as syntactic tree depth, bigram entropy, and lexical variation were examined and found to systematically correlate with grade-level distinctions previously identified using LEMI's readability tiers (Chitez et al., 2024b). One of the ReaderBench textual indices that best discriminated between the four grades was the maximum number of unique nouns per sentence. As shown in Figure 5, this metric shows a gradual increase from a low median of 3 in the first grade to around 8 in grades 3 and 4. The increasing prevalence of outliers in higher grades suggests greater variability in vocabulary complexity as students advance. To further assess the relationship between LEMI's interpretable readability formula and more data-driven approaches, preliminary machine learning experiments were conducted using Random Forest regression. These experiments demonstrated that LEMI-based readability scores can approximate perceived text difficulty with moderate predictive power ($R^2 = 0.465$), supporting their use as a reliable, linguistically grounded baseline for readability research.

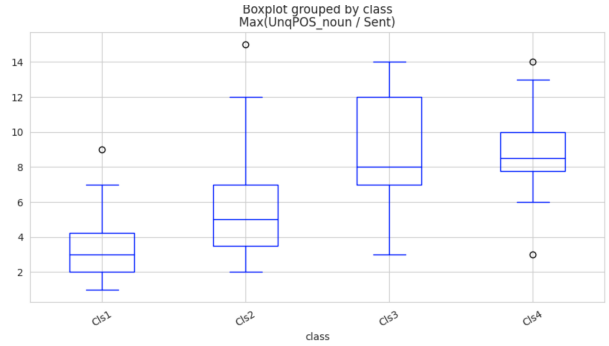

Figure 5: **Boxplot of the maximum number of unique nouns per sentence grouped by school grade, derived from ReaderBench analysis of the LEMI corpus**. Each box represents the interquartile range for that grade, with the horizontal line indicating the median and whiskers extending to 1.5 times the IQR; individual points beyond the whiskers are outliers. The median increases progressively from approximately 3 unique nouns per sentence in Grade 1 to around 8 in Grades 3 and 4, reflecting increasing lexical density across grade levels. The growing spread of values and greater prevalence of outliers in higher grades indicate greater variability in vocabulary complexity as texts become more advanced, corroborating LEMI's grade-level distinctions with an independent syntactic complexity metric.

Rather than replacing machine learning models, LEMI served as an explanatory layer that facilitated the interpretation of model predictions in an educational context, aligning with prior NLP research that emphasizes the importance of interpretable linguistic features in readability modeling alongside data-driven approaches (Collins-Thompson, 2014). The applicability of this combined analytical approach had also contributed to the previously-mentioned classroom-based validation study involving primary school students (Oravitan et al., 2023).

Further linguistic complexity investigations into the LEMI corpus reveal that Romanian translations of English texts tend to score higher on the readability scale than the original versions, indicating increased complexity (Csuros et al., 2024). This shift poses challenges for young Romanian readers, as texts originally intended for grades 2-4 in English are rendered suitable only for grades 5-9 in Romanian. The findings underscore two key factors contributing to this disparity: the prevalence of low-frequency vocabulary items (e.g., plurisyllabic lexical units, archaisms, regionalisms), and complex syntactic structures (e.g., embedded dependent clauses).

3.4 LEMI for Lexical Modernization of Canonical Romanian Texts

LEMI has been used in applied research addressing the readability of canonical Romanian literary texts (Chitez et al., 2025), which are widely used in school curricula but frequently contain archaic and regional vocabulary that poses comprehension challenges for contemporary learners. LEMI was used as a baseline readability assessment platform to quantify the impact of lexical modernization strategies aimed at improving access to historically and culturally significant texts, in line with prior NLP research showing that targeted lexical substitution can significantly reduce text difficulty while preserving semantic and discourse structure (Siddharthan, 2014). More precisely, lexical items were replaced using a semi-automated pipeline: a curated database of approximately 11,800 archaic and 13,000 regional terms compiled from a specialized dictionary of archaisms and regionalisms (Bucă, 2008) was first filtered against a common-use lexicon to exclude everyday vocabulary. Llama 3.3 70B (Grattafiori et al., 2024) was then applied in two steps: first to confirm whether each flagged term was used with its archaic or regional meaning in context, and subsequently to substitute confirmed items with accessible modern equivalents while preserving tone and meaning. The analysis covered literary texts extracted from Romanian school textbooks as well as canonical works by authors commonly studied in lower secondary education. By applying the same readability pipeline to both the original and modified texts, LEMI enabled a direct comparison of predicted reading age and readability category shifts.

The results showed that lexical modernization had a measurable effect primarily on canonical texts. Replacing archaic vocabulary led to an average reduction in predicted reading age of 0.5 years, while replacing regional terms resulted in a mean reduction of 0.59 years; both effects were statistically significant under the Wilcoxon signed-rank test ($p < .001$). In contrast, textbook texts exhibited only marginal changes in predicted reading age (–0.03 to –0.10 years) and maintained high readability category stability (above 95%), indicating limited impact of lexical intervention for texts already aligned with curricular standards, as shown in Figure 6, which illustrates that lexical modernization in canonical texts resulted in minimal shifts in estimated reader age and the distributions remain nearly identical before and after modification, with medians stable at grade 7 (approximately 12 years old) despite statistically significant reductions in predicted reading age. In total, the modernization pipeline replaced 3,025 archaic and 2,640 regional lexical items across the canonical writings. At the lexical level, modernization led to a substantial shift in readability-related indicators in canonical texts. The proportion of complex or rare lexical items decreased, while the number of frequent and easily recognizable words increased, without altering the overall narrative structure. These changes were consistently reflected in LEMI's readability estimates, enabling the platform to serve as a monitoring tool for evaluating the effects of targeted lexical interventions. In practical terms, the study demonstrates how LEMI can support the adaptation of canonical literary texts by identifying lexical sources of difficulty, estimating the impact of proposed modifications, and comparing alternative versions within a unified readability framework (Chitez et al., 2025).

3.5 LEMI in Teacher Training for Corpus-informed and AI-assisted Text Adaptation

LEMI has also been applied in an in-service teacher training context to develop e-literacy competencies through corpus-informed and AI-supported text adaptation (Chitez et al., 2026). The study involved 56 teachers from primary, lower secondary, and upper secondary education, working within a postgraduate professional development course focused on how digital tools can support the evaluation and adaptation of instructional texts across disciplines. Within the training program, LEMI was introduced as a Romanian-specific readability platform and used alongside other corpus-based and AI tools, including Text Inspector, ARTE, ChatGPT, and Perplexity. Teachers used LEMI primarily to assess the readability of authentic educational texts, including textbook excerpts and literary passages, before engaging in adaptation tasks. When texts were classified as exceeding the intended reading level, participants used AI tools to perform lexical or structural simplifications, then returned to LEMI to validate the readability of the adapted versions.

Analysis of the teachers' assignments, transformed into two corpus datasets and examined through keyword-based NLP, indicated that LEMI was most often used as a validation and benchmarking tool,

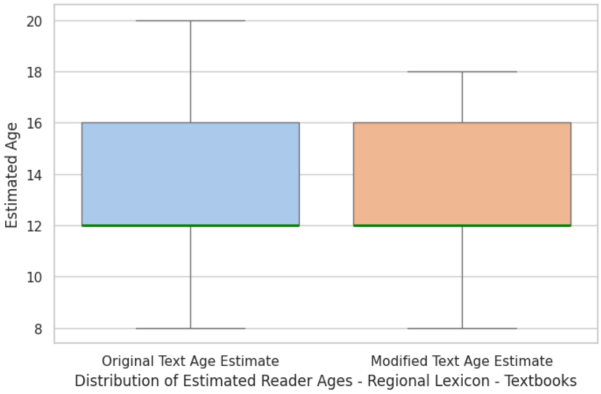

Figure 6: **Boxplot comparing the distribution of estimated reader ages before (Original Text Age Estimate) and after (Modified Text Age Estimate) regional lexicon replacement, for textbook texts**. Each box represents the interquartile range (IQR) of predicted reading ages across the textbook subcorpus, with the median indicated by the horizontal line. Despite statistically significant reductions in predicted reading age following lexical modernization ($p < .046$), the distributions remain nearly identical, with medians stable at approximately grade 7 (around 12 years old) and IQR largely overlapping. This pattern indicates that while regional lexical substitution produces measurable statistical effects, its practical impact on the readability classification of textbook texts is minimal, consistent with the high category stability (above 95%) reported for this subcorpus. Note that the corresponding distribution for canonical texts following archaic lexicon replacement, which showed more substantial readability shifts (mean reduction of 0.50 years; category stability 71.3%), is presented in Chitez et al. (2025, Figure 2) and is not reproduced here.

supporting professional judgment rather than replacing it. Participants consistently referred to the platform to confirm whether a text was appropriate for a given learner group and to compare original and modified versions within a unified readability framework. The most frequently flagged linguistic challenges, identified from teachers' own reflective reports, were complex vocabulary (36 mentions) and high cognitive load (31 mentions), where the latter term was used by teachers themselves to describe texts featuring long sentences, abstract terminology, and dense informational structure independently of LEMI's readability output. Archaisms and regionalisms were noted far less frequently (6 mentions each), playing a secondary role primarily in older literary texts. These patterns suggest that LEMI-guided readability assessment directed teachers' attention toward lexical density and processing demands, particularly in STEM texts. Across educational levels, LEMI supported differentiated pedagogical workflows: in primary education it was mainly used for text simplification and readability checks; in lower secondary education it was combined with AI tools for instructional design; and in upper secondary education it primarily served as a control mechanism, ensuring that adapted texts remained accessible without compromising conceptual content.

4 Reuse Scenarios across Research and Professional Communities

The versatility of the LEMI platform, both as an open-source, language-related platform and as a user-friendly literacy support tool, facilitates its reuse across a wide range of research and professional practice scenarios. By offering immediate access to texts through an online interface and allowing their retrieval in multiple formats, LEMI lowers technical barriers and enables engagement from diverse user communities with different methodological backgrounds and needs.

a. Linguistic research: LEMI can be used as a reference platform for corpus-based investigations of lexical, syntactic, and discourse-level complexity in Romanian texts addressed to young readers. Linguists can access and export texts for quantitative analysis, test readability or complexity measures, and conduct contrastive studies across grades, genres, or original and translated materials.

b. Literature studies: For researchers in literary studies, LEMI provides immediate access to curated children's and canonical texts that support comparative and stylistic analysis. The platform enables scholars to examine how literary language is adapted for educational contexts and to explore differences be-

tween canonical literature and contemporary children's writing from a linguistic perspective.

c. Digital Humanities research: For Digital Humanities research, LEMI functions as a curated digital repository of rarely used literary texts, many of which originate from older printed books or from authors that are underrepresented in contemporary educational canons. By digitizing, structuring, and providing immediate access to these materials, the platform enables scholars to work with texts that are otherwise difficult to obtain, fragmented across archives, or absent from standard digital collections. This supports DH research focused on diachronic language use, literary marginality, and the recovery of overlooked voices in children's literature. The availability of these texts in reusable digital formats allows researchers to combine interpretive analysis with corpus-informed exploration, while preserving access to historically and culturally significant content that would otherwise remain inaccessible.

d. Educational research: In educational research, the platform can be used to study curriculum alignment, progression of reading difficulty, and the relationship between text complexity and learner outcomes. Instant access to texts allows researchers to design experiments, select materials for classroom studies, and replicate analyses across educational settings.

e. Computational linguistics research: For computational linguistics, LEMI offers a readily accessible dataset for experimenting with readability modeling, feature extraction, and interpretability-focused approaches in an under-resourced language. The availability of texts in reusable formats supports benchmarking and methodological exploration without the overhead of corpus construction.

f. Teaching Romanian as a second or foreign language: LEMI can support research and practice in Romanian as a second or foreign language by enabling the selection and reuse of texts suited to different proficiency levels. Researchers and instructors can analyze linguistic features that affect accessibility for non-native readers and develop graded reading materials based on empirically grounded text selection.

g. Teaching practice in literature and language education: For teachers, the platform functions as a practical access point to reading materials that can be directly reused in classroom activities. Texts can be retrieved, adapted, and integrated into teaching materials, supporting informed text selection and reflective practice without requiring advanced technical skills.

h. Editors and educational decision makers: Editors, curriculum designers, and decision makers can use LEMI as a benchmarking resource for evaluating the linguistic suitability of reading materials. The platform enables rapid comparisons of texts and supports evidence-based decisions on textbook selection, reading lists, and assessment design.

i. Interdisciplinary research: By providing shared, easily accessible textual resources, LEMI supports interdisciplinary research at the intersection of linguistics, education, psychology, and literary studies. The platform enables collaborative work grounded in a common empirical basis, facilitating studies that address reading development, comprehension, and educational accessibility from multiple perspectives.

5 Discussion

LEMI exemplifies how digital platforms can address societal challenges while adhering to the technical standards set by the CLARIN infrastructure. By combining a curated corpus, interpretable readability metrics, and interoperable metadata, the platform helps reduce long-standing gaps in Romanian-language resources for early literacy. This area has remained underdeveloped despite persistent educational needs.

The range of applications presented in this paper demonstrates LEMI's relevance as a reusable resource that supports educational practice, linguistic research, and policy-oriented analysis, with its open data and interoperable metadata supporting integration into CLARIN-supported discovery services. Classroom-based validation, textbook assessment, lexical modernization, and teacher training studies collectively show that LEMI can be integrated into diverse workflows while maintaining consistent analytical assumptions and outputs. This versatility reflects the platform's design as an infrastructure component rather than a single-purpose tool, enabling reuse across institutional and disciplinary boundaries. A central aspect of LEMI's design is its commitment to openness and interoperability. Both the platform and the underlying linguistic resources are made openly available, with the full corpus published on GitHub under a Creative Commons licence. This open-access approach facilitates transparency, replication, and extension of the analyses reported here, allowing researchers to reuse the data and methodology in com-

parative studies. The metadata schema follows FAIR principles (data are Findable, Accessible, Interoperable, and Reusable) and is structured to support integration with federated search services such as the CLARIN Virtual Language Observatory.

The applied studies also highlight LEMI's capacity to produce interpretable, explainable outputs as readable scores grounded in linguistically motivated parameters, rather than opaque predictions whose internal logic is inaccessible to educators. In educational contexts, this transparency is consequential: teachers and curriculum designers can understand not only what readability level a text receives, but why, and can use that reasoning to inform adaptation decisions. In research contexts, LEMI can be used alongside more complex linguistic complexity frameworks, serving as a transparent baseline that supports comparative analysis. At the policy level, LEMI-based textbook assessments reveal systematic mismatches between intended grade levels and the actual linguistic demands, offering empirically grounded insights for curriculum design and assessment practices. At the same time, the discussion of applications highlights several limitations that constrain the platform's current scope. First, corpus size and genre diversity remain ongoing challenges: the current repository of 250 texts and 92,000 words covers primarily narrative and literary genres, leaving non-fiction and informational texts, which are central to formal education across STEM and social science subjects, substantially underrepresented. This limits LEMI's applicability beyond literary readability contexts. Second, the readability formula's scope is currently restricted to four surface-level and lexical parameters (ASL, PCW, UCWD, and WD), which, while interpretable and robust across text types, do not capture deeper dimensions of textual complexity such as syntactic embedding, discourse coherence, or inferential demand. Third, the classroom-based validation study (Oravitan et al., 2023) was conducted in a single educational context with primary school students in grades 3 and 4, which means generalizability across other grade levels, school types, and regional settings remains to be established through broader replication. Fourth, the absence of comprehensive, digitally available core vocabulary lists for Romanian limits the precision of word-level analysis and constrains the platform's ability to account for lexical variation across grade levels fully, particularly regarding neologisms, archaisms, and regional forms. These limitations reflect broader infrastructural gaps affecting under-resourced languages and point to areas requiring sustained attention and further resource development.

Within this context, LEMI should be perceived as an evolving platform operating in accordance with CLARIN ERIC's technical and open-access standards, one that both addresses immediate educational needs and exposes structural resource gaps that require sustained attention. By documenting multiple applications and making both tools and data openly accessible, LEMI contributes to a transparent, reusable infrastructure for readability research and literacy support in Romanian.

At present, Romania does not yet host an officially established CLARIN center. However, LEMI is conceptually aligned with the objectives and technical standards of the CLARIN ERIC infrastructure. Ongoing national efforts, coordinated through the CLARIAH-RO initiative, aim to support Romania's formal accession to CLARIN and the subsequent establishment of a national center. These developments are currently under discussion with the relevant national authorities, and LEMI is intended to contribute to and potentially integrate within this emerging framework once the affiliation process is formalized.

6 Conclusions and Future Work

LEMI responds to pressing needs in literacy support for Romanian by offering an innovative solution that integrates corpus-based methods with NLP technologies tailored to societal and educational challenges. Its impact extends beyond classroom instruction, contributing to curriculum design and digital humanities research, exemplifying how FAIR-aligned, openly licensed language resources can promote equitable access to educational tools in under-resourced linguistic contexts.

While LEMI provides a strong foundation for supporting literacy development, it also highlights persistent gaps in Romanian language resources, particularly the absence of comprehensive digital tools for vocabulary profiling. The platform's current inability to fully account for lexical nuances such as neologisms, archaisms, and regionalisms limits its capacity to deliver more fine-grained readability feedback. Future developments will therefore focus not only on broadening genre coverage and integrating adaptive

NLP features, but also on enhancing lexical analysis by supporting efforts to build core vocabulary lists and other linguistic benchmarks. By offering scalable, linguistically grounded, and openly accessible solutions, LEMI serves as a model for addressing educational inequities through digital innovation, while simultaneously identifying areas for further investment in under-resourced language infrastructure.

A priority area for future development is creating grade-specific vocabulary lists for Romanian, thus enabling the classification of texts by grade level based on vocabulary. Currently, the platform's word-level analysis relies on general frequency measures and complexity indicators but lacks validated lists of core vocabulary for each educational level. Work is underway to address this gap by developing reference vocabulary inventories that capture the lexical knowledge expected at each grade level in primary education. These lists would enable LEMI to classify texts based on the proportion of grade-appropriate vocabulary they contain, providing a complementary dimension to the platform's existing readability metrics.

References

Badenes-Olmedo, C., Eyzaguirre-Barreda, P., Chu-Artzt, N., & Gayoso-Cabada, J. (2026). ESQAD: A curriculum-aligned dataset for question answer generation in Spanish. *Knowledge-Based Systems*, 115255.

Balea, B., Kovacs, M., & Pogacean, S. (2024). Development of preparatory grade children's reading skills in Romania. *Journal of Educational Sciences*, *25*, 25–46.

Bucă, M. (2008). *Dictionar de arhaisme si regionalisme (DAR)*. Vox Cart.

Chitez, M., et al. (2024a). Linguistic overload in secondary school textbooks: A corpus-informed case study of Romanian 6th grade textbooks. *Conference Proceedings. Innovation in Language Learning 2024*.

Chitez, M., Csürös, K., & Rogobete, R. (2026). Upgraded literacy: Teacher training approaches to integrating corpus data and AI tools for school text readability adaptation. *Applied Corpus Linguistics*, *6*(1), 100181. https://doi.org/https://doi.org/10.1016/j.acorp.2025.100181

Chitez, M., Dascalu, M., Udrea, A. C., Striletchi, C., Csuros, K., Rogobete, R., & Oravitan, A. (2024b). Towards building the LEMI readability platform for children's literature in the Romanian language. *Proc. LREC-COLING 2024*, 16450–16456.

Chitez, M., Rogobete, R., Udrea, C. A., Csürös, K., Bucur, A.-M., & Dascalu, M. (2025, September). Integrating archaic and regional lexicons to improve the readability of Old Romanian texts. In G. Angelova, M. Kunilovskaya, M. Escribe, & R. Mitkov (Eds.), *Proceedings of the 15th international conference on recent advances in natural language processing - natural language processing in the generative ai era* (pp. 240–246). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2025.ranlp-1.29/

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, *165*(2), 97–135.

Csuros, K., Chitez, M., Oravitan, A., Rogobete, R., & Bucur, A.-M. (2024). Inter-readability: A corpus based contrastive analysis of children's literature readability in English versus translation into Romanian. *AACL 2024*.

Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2017). ReaderBench: A multi-lingual framework for analyzing text complexity. *Proc. EC-TEL 2017*, 495–499.

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., & Van Uytvanck, D. (2018, May). CLARIN: Towards FAIR and responsible data science using language resources. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1515/

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., & Vidal, M. (2020). FQuAD: French question answering dataset. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1193–1208.

Étienne, A., Battistelli, D., & Lecorvé, G. (2019). *Compréhension de textes par les enfants et émotions: Point (s) de vue psycholinguistique (s) et leur mise en œuvre en TAL* [Doctoral dissertation, MoDyCo; IRISA].

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The Llama 3 herd of models. https://arxiv.org/abs/2407.21783

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2021). Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Hurjui, B. A. (2024). Recovery of literacy skills in primary education in Romania. potential systemic solutions. *Revista de Științe ale Educației*, *50*(2), 47–57.

Iliescu, D., & Airinei, M. (2022). Raport privind nivelul de literație al elevilor din România. *BRIO*.

Ion, R., Irimia, E., Ştefănescu, D., & Tufiș, D. (2012). ROMBAC: The Romanian balanced annotated corpus. *Proc. LREC 2012*, 339–344.

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). Flesch-Kincaid Grade Level. *United States Navy*.

Midrigan-Ciochina, L., Boyd, V., Sanchez-Ortega, L., Malancea_Malac, D., Midrigan, D., & Corina, D. P. (2020). Resources in underrepresented languages: Building a representative Romanian corpus. *Proc. LREC 2020*, 3291–3296.

Mititelu, V. B., Tufis, D., & Irimia, E. (2018). The reference corpus of the contemporary Romanian language (CoRoLa). *Proc. LREC 2018*.

Oravitan, A., Chitez, M., Rogobete, R., Csuros, K., & Hagiu, S. (2023). Validating LEMI: A new readability tool for children's literature in Romanian. *eLADDA/ ELN 2023*.

Romanian Parliament. (1996). Legea nr. 8 din 14 martie 1996 [Capitolul V, Art. 27-34, Capitolul VI, Art. 35]. https://legislatie.just.ro/Public/DetaliiDocument/201472

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, *165*(2), 259–298.

Sohail, J., Hipfner-Boucher, K., Deacon, H., & Chen, X. (2022). Reading comprehension in French L2/L3 learners: Does syntactic awareness matter? *Languages*, *7*(3), 211.

Stanciu, F., Beteringhe, A., & Stanciu, L. (2024). Functional literacy - a strategic priority in education in Romania? *EIRP Proceedings*, *19*(1).

Tufiș, D., Mititelu, V. B., Irimia, E., Dumitrescu, Ș. D., & Boroș, T. (2016). The IPR-cleared corpus of contemporary written and spoken Romanian language. *Proc. LREC 2016*, 2516–2521.

Putting things on top of other things: The ZuMult platform for multimodal corpora and its ecosystem

Thomas Schmidt
University of Duisburg-Essen
thomas@linguisticbits.de

Anne Ferger
University of Duisburg-Essen
anne.ferger@uni-due.de

Elena Frick
Leibniz-Institute for the German Language
frick@ids-mannheim.de

Abstract

We present ZuMult, an open-source platform for multimodal corpora, and its ecosystem. ZuMult has a flexible three-layer architecture that enables integration with different types of backend, and development of frontend applications tailored to the needs of highly diverse user groups. The paper puts a special emphasis on the way that ZuMult is built upon existing best practices, standards and technologies arguing that such an "ecosystem" makes it easier to integrate the technology with different workflows and technical environments.

1 Introduction

ZuMult (*Zugänge zu multimodalen Korpora gesprochener Sprache* 'Access to multimodal corpora of spoken language', Fandrych et al., 2023a, https://zumult.org) is an open-source[1] corpus platform for multimodal corpora. By 'multimodal corpora', we understand collections containing (at least) audio and/or video recordings of spoken language and corresponding transcriptions and fulfilling the "data maturity" conditions as suggested in Wamprechtshammer et al., (2022).[2]

As surveyed in more detail in Batinić et al. (2021) and Frick & Schmidt (2025), several multimodal corpus platforms have, in the last fifteen years, either been developed as new applications (e.g. DGD, Schmidt 2014, Spokes, Pęzik 2015, LiRI Corpus platform, Grän et al. 2023, LAPIS, Pluschkovits et al. 2026) or as an extension or adaptation of platforms originally developed for written language (e.g ANNIS, Krause & Zeldes 2016, CQPWeb, Love et al. 2017, KonText, Kren 2020), and a growing amount and diversity of audiovisual corpora are becoming available as FAIR compliant language resources in CLARIN (cf., for example, the CLARIN resource families "Spoken corpora", "Multimodal corpora", "L2 learner corpora", "Oral History corpora").

The most important challenge (and opportunity) for this kind of research data lies in its diversity, with respect to both characteristics of the data themselves (such as types of primary data, granularity and depth of annotation), and the multitude of ways in which they are used in empirical and applied research (ranging from interactional linguistics over dialectology to oral history studies or qualitative social sciences). The user study reported in Fandrych et al. (2016) has revealed that any sustainable technical solution in this area will therefore require considerable flexibility in backend (data) and frontend (applications) if it aims to be adopted more widely.

The ZuMult platform was designed with exactly this flexibility as a guiding principle, putting a special emphasis on support for approaches which are not predominantly quantitative in nature, but rely on fine-grained qualitative analysis of primary and secondary data.

ZuMult was developed in a project funded by the German Science Foundation between 2018 and 2022 as a flexible, versatile three-layer architecture allowing different types of data sources (e.g. file system, relational database) in the backend and different types of user-group specific adaptations (e.g. for language teaching, for conversation analysis, for variational linguistics) in the frontend.

In ZuMult's project phase, the focus lay not only on the development of the architecture itself, but also on the development of frontend applications built upon it to address the needs of language teachers and learners (see the short descriptions of ZuMal, ZuViel and ZuRecht in Section 2). Currently, public instances of ZuMult exist at the Archive for Spoken German (AGD,

[1] The source code is available from https://github.com/zumult-org/zumultapi under a GPL v3 license.

[2] There seems to be no widely established terminology to denote this kind of corpus. "Oral corpus" or "Audiovisual corpus" are also in use. In the CLARIN resource family "taxonomy" (https://www.clarin.eu/resource-families), we find "Spoken Corpora" and "Multimodal Corpora" as the tightest fits, but "Oral History Corpora" and "Disordered Speech Corpora" will usually also have audiovisual primary data. This may also be true of some "Parliamentary Corpora" and "L2 Learner Corpora". See also Schmidt (2026) for an attempt at a classification of "non-textual language data". More recently, ZuMult has also been applied for non-canonical written language data, more specifically: learner texts (Frick et al., 2025).

https://zumult.ids-mannheim.de), where the platform provides access to the largest corpora of spoken German (such as the Research and Teaching Corpus of Spoken German, FOLK, Schmidt, 2016, the corpus *Deutsch Heute* 'German today', documenting regional variation of spoken German, Brinckmann et al., 2008, and the corpus GeWiss, a comparable corpus of academic speech, Fandrych et al., 2012), and at the University of Texas in Austin (https://tgdp-zumult.la.utexas.edu/), where the data of the Texas German Dialect Project (TGDP) can be accessed via ZuMult (Boas et al., 2026, Schmidt et al. 2026). Further public instances are under construction for the French ESLO corpus (Enquêtes Sociolinguistiques à Orléans, archived at HumaNum, Schmidt, 2025), a number of multimodal and multisensorial corpora at the University of Duisburg-Essen (see Section 4) and for the EXMARaLDA demo corpus at the University of Hamburg.

ZuMult can be viewed as a piece of technology that both contributes to an infrastructure such as CLARIN, and profits from other technologies already integrated into or developed alongside the infrastructure. In the words of Monty Python (1970), ZuMult is thus a "thing that is put on top of other things", and other things can be put on top of it (see Figure 9 in the conclusion).

The remainder of the paper will explain in more detail what this means: Section 2 briefly sketches the most important features of ZuMult from a user's point of view. Section 3 then goes into more technical detail, emphasizing the interplay of different tools, standards, and services that underlies the architecture. Section 4, finally, gives an outlook on the current and future development of the platform.

2 Platform functionality

Generally speaking, a multimodal (spoken, audiovisual) corpus platform like ZuMult transfers functionality of written corpus platforms and extends it with functionality that is needed for and specific to the work with audiovisual language data and their transcriptions and annotations (Schmidt & Frick, 2025). Most importantly, this means that such platforms need to enable users:

a) to browse through (read, listen to, view) the primary (audio/video) and secondary (transcribed/annotated) data,
b) to select those (and only those) datasets that are relevant to their research questions,
c) to visualise selected aspects of selected data for qualitative analysis, and
d) to carry out, and possibly quantify, queries on transcriptions and their annotations with the option to re-contextualise query results (i.e. view them in the context of the underlying transcript and/or recording).

Metadata (on speech events and participating speakers) can guide or supplement all these processes (see Frick & Schmidt, 2025 for a comprehensive overview).

ZuMult's architecture is designed in such a way that the required information (e.g.: What corpora are in the platform? What's in corpus XY? Can I get transcript #123? What is the result of query Z?) can be delivered from one and the same backend to different frontends ("applications"), which are optimised for a specific usage scenario or a specific user group. The following subsections illustrate three such applications.

2.1 Data selection

The application ZuMal (*Zugang zu Merkmalsauswahl von Gesprächen*, 'Access to the selection of conversational features') supports the selection of suitable material for language teaching through an interface displaying selectors for relevant metadata and so-called "measures" reflecting the difficulty or degree of orality of a given dataset. Figure 1 illustrates how dialectal region ("Sprachregion"), vocabulary coverage ("Wortschatz") and articulation speed ("Sprechgeschwindigkeit") are used as selectors in the column on the left-hand side. Matching speech events are listed in tabular form and in a graphical visualisation on the right-hand side.

Thomas Schmidt, Anne Ferger, and Elena Frick. 2025. Putting things on top of other things: The ZuMult platform for multimodal corpora and its ecosystem. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 76–88. DOI: https://doi.org/10.3384/ecp222https://doi.org/10.3384/ecp222

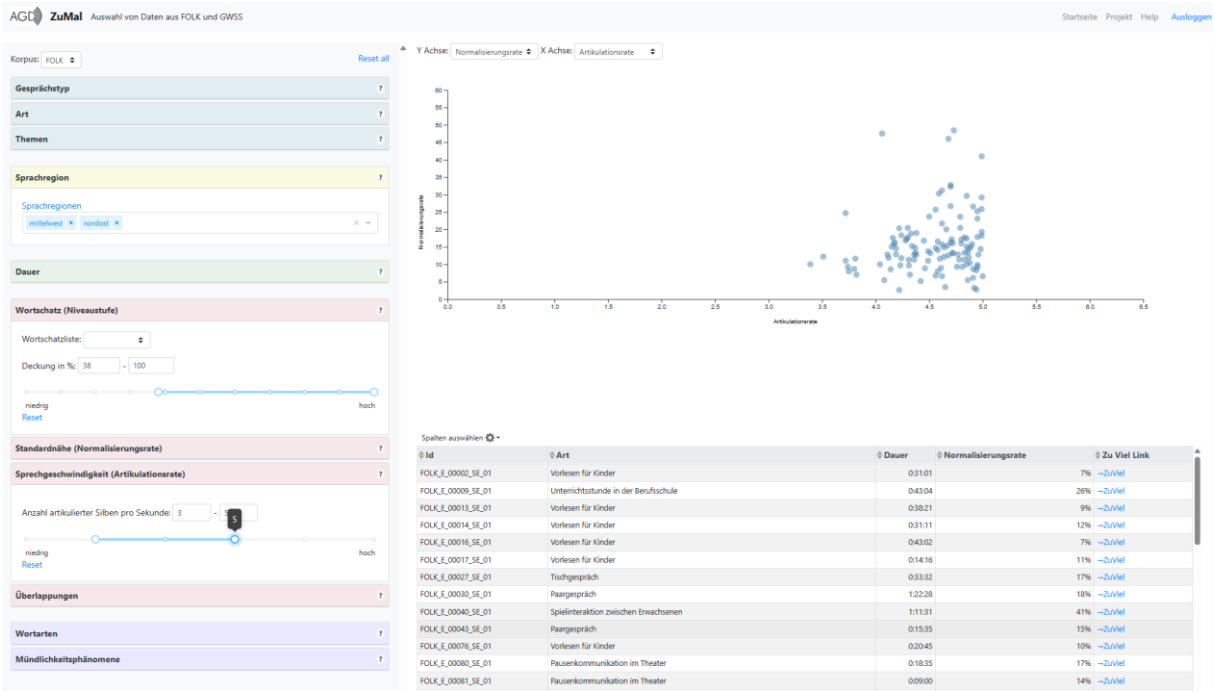

Figure 1: Data selection in ZuMal

Relevant metadata in that sense are, for example, the domain (private, institutional, public) and type (e.g. telephone conversation, workplace meeting, TV talk show) of interaction, and the (dialectal) region where it was recorded. Didactically relevant measures that can be automatically derived from the transcribed and annotated data are, among others, proximity to the standard (calculated as the ratio of tokens where the transcribed form deviates from standard orthography), articulation speed (syllables per second) or lexical coverage with respect to a reference wordlist (such as the Goethe wordlists corresponding to different levels of the Common European Framework of Reference for Languages CEFR).

The idea of the ZuMal tool is that language teachers can use it to find example material that is suitable to a given teaching situation, such as “Introducing B1 speakers of German to a private conversation in Bavarian that is sufficiently slow for them to understand”. Fandrych et al. (2023b) discuss such scenarios and ZuMal’s role in them in greater detail. Other means of data selection enabled by ZuMult include tabular overviews of speech event and speaker metadata that can be filtered and sorted, or the use of metadata inside search queries (see below).

2.2 Data visualization / Qualitative analysis

A crucial characteristic of many methodologies dealing with spoken and/or multimodal data is that, in contrast to classical corpus-linguistic approaches to canonical written data, they systematically include steps of close manual (i.e., intellectual) inspection of individual sections of primary and secondary data. Micro-sequential analysis, as practiced as a core method of conversation analysis (see, for instance, Sidnell, 2012), or the aforementioned use of excerpts in language teaching are prototypical cases in point. The application ZuViel (*Zugang zu Visualisierungs-Elementen für Transkripte* ‘Access to transcript visualisation features’, Schmidt et al., 2023, see Figure 2) is one among several tools developed in the ZuMult architecture to support this kind of approach.

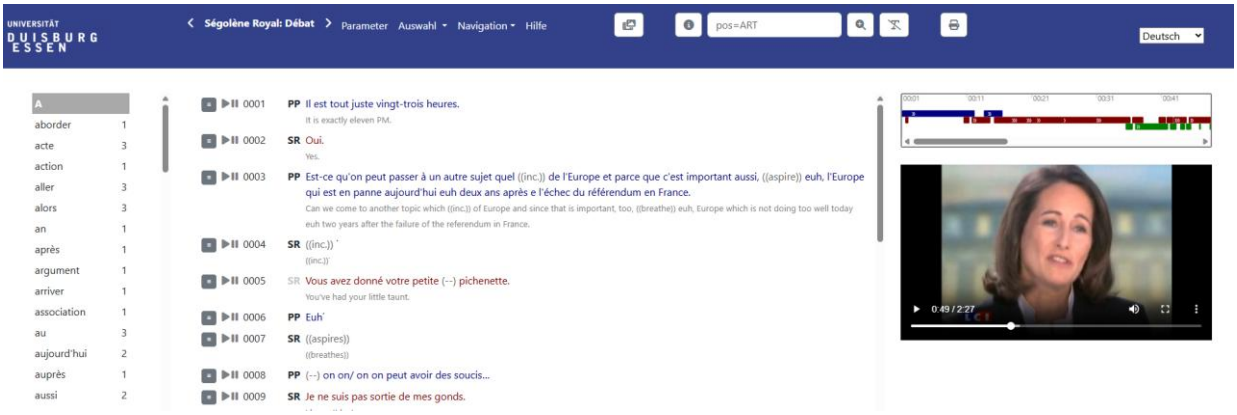

Figure 2: Interactive transcript display in ZuViel

ZuViel is an interactive transcript viewer which offers different, synchronised views (e.g. line transcripts, musical score transcript, lemma frequency list, density viewer, video with subtitles) of a transcript in which parameters (such as: display of transcribed vs. normalized forms, visualisation of speech rate) can be set by users according to their preferred perspective on the data. Interactive inspection and analysis of a piece of data is thus supported through functionality for foregrounding specific information (for example: highlighting all occurrences of tokens with a part-of-speech (POS) tag for pronoun) and for viewing relations between different types of data (e.g., synchronised media playback or a popup for speaker metadata). Whenever the functionality of the platform is not sufficient or appropriate for the analysis at hand, excerpts can be marked and downloaded in the formats for the most widely used desktop tools (such as Praat for phonetic analysis).

2.3 Query

With the application ZuRecht (*Zugang zu Recherchen auf Transkripten* 'Access to transcript search', Frick et al., 2023), the platform also provides an interface for querying audio and video transcriptions using the syntax of CQP (Corpus Query Processor Language, Evert et al., 2022). A notable aspect of this application is that it allows typical characteristics of spoken language to be included in searches. These can be alternative pronunciation variants, deviations from standardized forms, hesitations, pauses, and other non-verbal phenomena (e.g. laughter or blinking). Furthermore, time-aligned annotations, speaker overlaps and metadata can also be taken into account when querying interaction corpora. All these features can be easily integrated into ZuRecht CQP queries in form of token specifications or structural constraints, as illustrated by the query example in Figure 3. This CQP query will search in the German FOLK corpus for all pronunciation variants of 'du kannst' (you can), allowing for a possible hesitation (POS tag 'NGHES') or a pause in between. Both words should occur outside of speaker overlaps but within user action sequences ("as") annotated with 'Vorschlag/Angebot' (suggestion/offer) and during vocational lessons ('Unterrichtsstunde in der Berufsschule').

```
([norm = "Du"] ([pos="NGHES"] | <pause/>)? [norm = "kannst"]
!within <speaker-overlap/>) within <as="Vorschlag/Angebot"/>)
 within <e_se_art="Unterrichtsstunde in der Berufsschule"/>
```

Figure 3: Complex CQP query

Thanks to the expressive power and flexibility of the CQP query language, the search via ZuRecht offers a wide range of possibilities, but it also requires some prior knowledge of the CQP syntax as well as the metadata and annotations available in the corpora. To make it easier for users to get started, a Query Builder has been integrated into the search input field of the AGD ZuMult instance. When the search field is clicked, the Query Builder automatically suggests syntactic elements that are valid at the current position in the query and allows users to select from the annotation and metadata categories as well as their possible specification values available in the selected corpus (Figure 4).

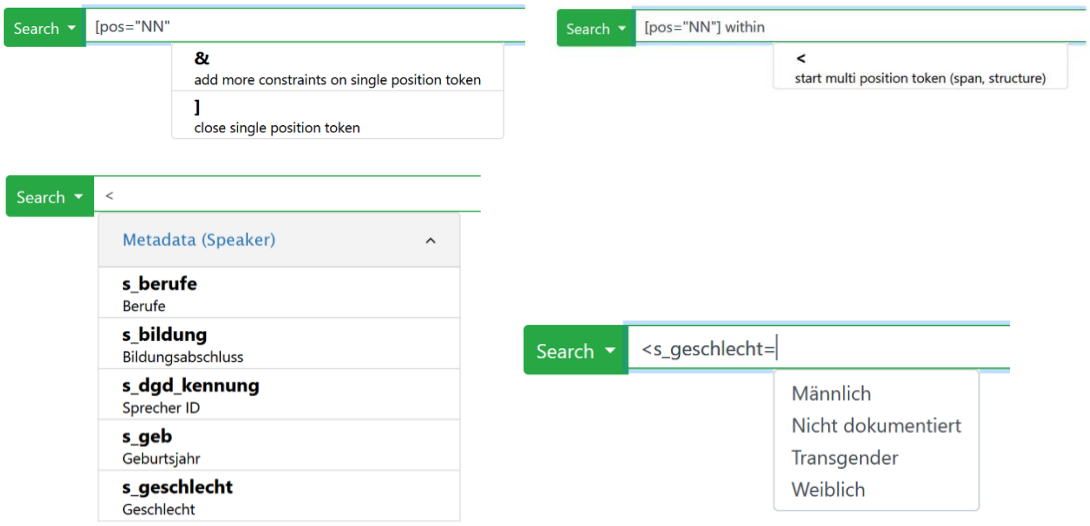

Figure 4: The query builder displaying help for CQP syntax (top left, top right), a suggestion box with the available categories of speaker metadata (bottom left) and a suggestion box with the available values for the metadata category 'Gender' (bottom right).

ZuRecht is primarily a CQP-based KWIC concordancer with special features (e.g. audio playback for KWIC lines, inclusion of video stills, see Schmidt & Frick 2025 and Figure 5) for audiovisual data. Going beyond that basic functionality, it was further extended to enable searches for repetitions (Frick et al., 2024) and n-grams that can be executed directly on the search index without the need of extra annotations or n-gram-lists. To this end, dedicated search forms have been implemented in ZuRecht that allow users not only to search for repetitions and n-grams, but also to formulate queries considering the specific properties of spoken language. For example, users can specify whether hesitations should be ignored or taken into account when computing n-grams (e.g. whether *good erm afternoon* should yield one or two bigrams). As another example, when searching for repetitions, users can define whether a speaker's utterance is repeated by another speaker or if they repeat their own utterance, and whether such self-repetitions occur within the same turn or only after a speaker change – possibly triggered by an intervention from another speaker, such as a question or a request for clarification.

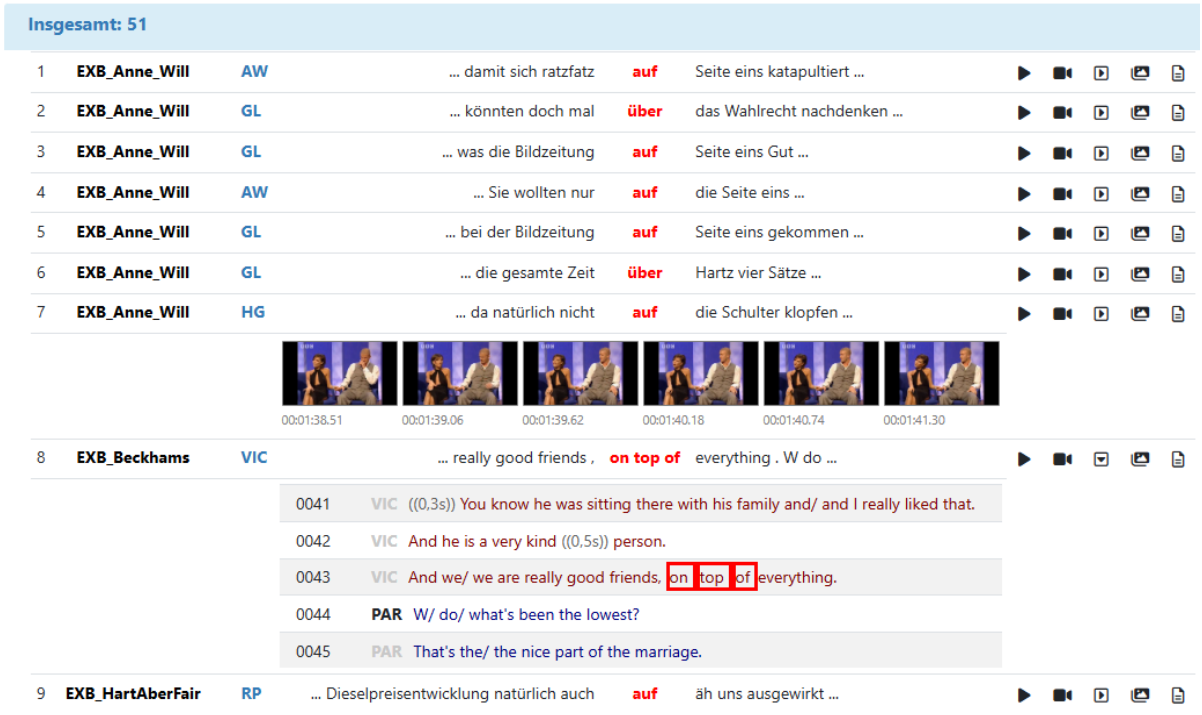

Figure 5: A KWIC concordance in ZuRecht displaying video stills and the transcript excerpt corresponding to line 8.

3 ZuMult and its ecosystem: Things on top of other things

The starting point for ZuMult's design and architecture was experience gained in the development of EXMARaLDA (Schmidt & Wörner, 2014) and the Database for Spoken German (Schmidt, 2014) as well as in the compilation of the corpora FOLK and GeWiss (see introduction). Those tools and corpora, in turn, have greatly profited from (and partly contributed to) standards and best practices developed in the last 30 years, both on a larger scale in text and media technology (Unicode, XML, the MPEG-family of standards), and in more specific linguistic contexts (e.g. Schmidt et al., 2009; Schmidt et al., 2010; Westpfahl & Schmidt, 2016). ZuMult reuses ideas and concepts that have proven fruitful in long-standing application of these resources and attempts to add and improve things that have turned out to be missing, deficient or hard to use.

3.1 Data model / Data interoperability

ZuMult is built on top of an abstract data model which relates corpora, events, speech events and speakers as conceptual objects with binary media objects (audio and video files) and textual transcription objects as sketched in Figure 6.

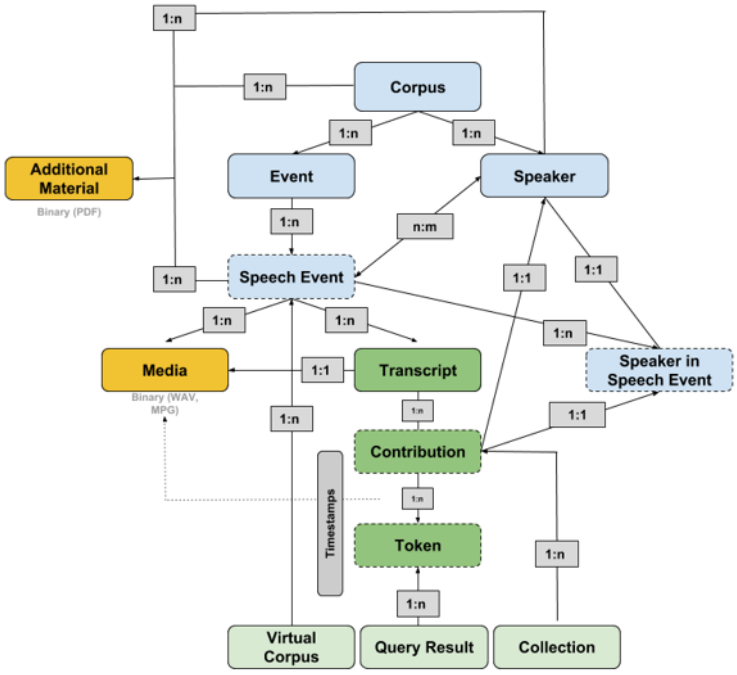

Figure 6: ZuMult data model for corpora

On a more concrete level, ZuMult reuses EXMARaLDA's data model and API with its extensive interoperability, meaning that, from the outset, ZuMult is built on top of a technology interoperable with the most widely used tools (ELAN, Praat, EXMARaLDA, FOLKER, CLAN, Transcriber and others) and formats for multi-layer annotation of audiovisual data (Schmidt et al., 2009).

More recently, this interoperability has been extended to include output formats of ASR technology (e.g. the JSON format of OpenAI Whisper). On top of this level of exchangeability of tool formats sits a standardisation via "ISO 24624:2016 Language resource management – Transcription of spoken language" which itself rests on top of the long-standing efforts of the Text Encoding Initiative. This standard is currently the best option to represent both the multi-tier time-based "macro" structure of tools like ELAN or EXMARaLDA and the hierarchical token-based "micro" structure of annotated transcription text in a single document (Schmidt, 2011). It has been established as a standard in the CLARIN infrastructure (Schmidt et al., 2017; Fisseni & Schmidt, 2020; Hedeland & Schmidt, 2022), and its documentation and tool support are currently improved in a dedicated cooperation project ("Transcription+") in the Text+ consortium of the German NFDI[3].

Concerning metadata (on speech events, on speakers, on the corpus as a whole), no comparably stable or established basis for interoperability and standardisation exists. ZuMult's data model in this

[3] See https://linguisticbits.de/transcription-plus.html for a preliminary website

respect therefore remains on a relatively abstract level which only requires metadata to be organised as attribute-value pairs underneath a 'corpus > event > speech event' hierarchy and a speaker list with a n:m assignment between speakers and speech events (see Figure 6 and Schmidt, 2014b). On the implementation level, ZuMult so far has been made to work with the metadata schema of the Archive for Spoken German (Dickgießer, 2017) and with the metadata schema of the EXMARaLDA Corpus Manager (COMA, Wörner, 2012). If the need arises, it could also be made to work with IMDI, individual CMDI schemas or metadata encoded in TEI headers. We found, however, that these solutions are not frequently used in practice to encode the actual corpus metadata (the single most common solution for this being spreadsheet tables that we convert to the COMA format for integration into ZuMult). Rather, they are in more wide-spread use only for the purpose of documenting selected metadata for archiving purposes.

3.2 Workflows / Processing Pipelines

Beyond the level of formats, ZuMult is compatible with (and thus can be made to sit on top of) workflows developed for larger corpus projects. A ZuMult integration has been established as a final step of the workflows for the ongoing compilation of FOLK (Schmidt, 2016) and the TGDA (Boas et al., 2026), and, most importantly, is now also an element in the git-based continuous integration workflows described in Ferger et al. (2023). In the latter form, using the ZuMult platform also becomes an option in more dynamic setups, i.e. contexts in which corpora do not follow a strict release schedule, but are built up in smaller research projects in a fluid iteration of data collection, curation and analysis. This allows smaller and ongoing corpus projects to leverage ZuMult's functionality.

In the git-based continuous integration workflows, several automated steps support the integration into ZuMult by checking, enriching, converting and indexing the elements of a corpus. In addition to the steps of harmonizing and repairing inconsistencies, which were already established in the workflows of the INEL project (Hedeland & Ferger, 2020), recently added workflow steps include:

- a conversion of transcription files (i.e. from ELAN or EXMARaLDA files, or results from Whisper ASR) into the standardised TEI format required by ZuMult,
- tokenisation and the addition of annotations on the token level, such as part-of-speech, lemmatization, orthographic normalisation, language detection,
- phonetic annotation (with the G2P web service, Reichel & Kisler 2014) and speech rate annotation (syllables per second, calculated from time-alignment and G2Ps syllabification),
- automated translation using the DeepL API (https://developers.deepl.com/),
- compilation of metadata from different sources, e.g. spreadsheets or automatic extraction of media metadata via FFPROBE (https://ffmpeg.org/ffprobe.html),

A final step of the workflows is the (Lucene) indexing of the transcription and metadata files for ZuMult. A schematic depiction of a workflow is shown in Figure 7.

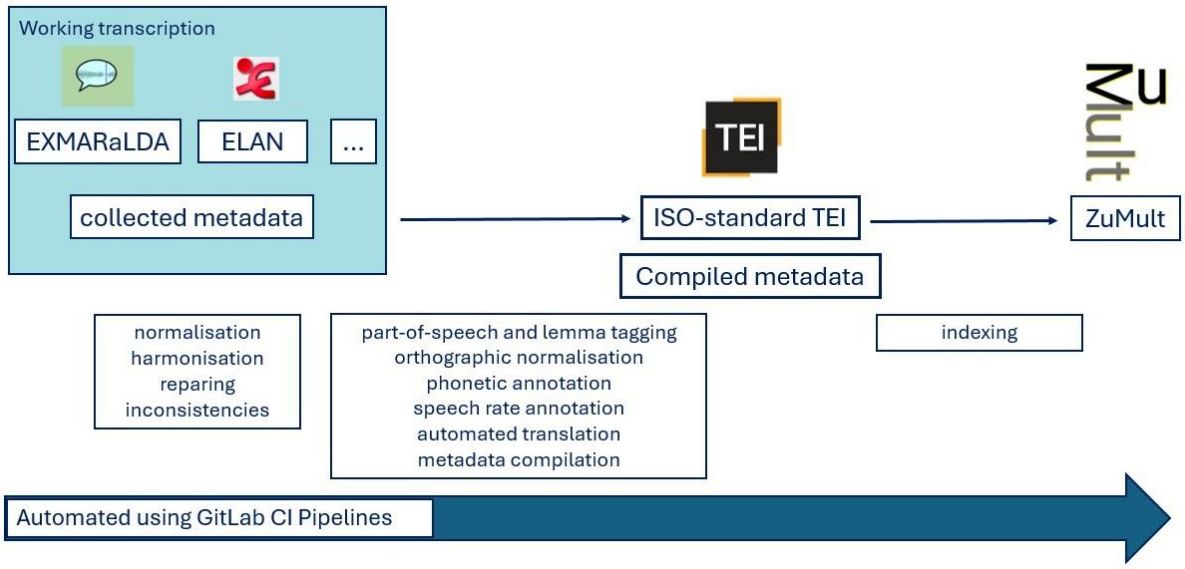

Figure 7: Continuous workflows for ZuMult integration.

Similarly to the integration of corpora into ZuMult, the build processes and deployment of the platform itself are automated using continuous integration workflows alongside Maven and Docker as distribution methods for the code, thus facilitating integration into different environments as well as improving software sustainability.

3.3 Query implementation

ZuMult's query capabilities (see Figure 9 in the conclusion) are developed on top of the Multi-Tier Annotation Framework (MTAS, Brouwer et al., 2016) which itself is built on top of the industry level search software Lucene. MTAS can read and index transcription files adhering to ISO 24624:2016, and make them queryable via the CQP syntax, which has been identified (Fandrych et al., 2016) as the most promising candidate for a query language that is both widely known and suitable for the complexity of spoken language data.

MTAS is an open-source framework originally developed within the Nederlab project for working with large collections of digitized texts with multi-layer annotations. In the ZuMult project, this framework was applied for the first time to index and search audiovisual interaction corpora (cf. Frick & Schmidt, 2020). In what follows, several aspects of the integration of MTAS into the ZuMult architecture are described.

Adjusting the configuration file of the MTAS document parser: The configuration file of the MTAS document parser defines a set of mapping rules that determine which elements of ISO 24624:2016 transcripts should be stored in the search index and how they can be addressed later in CQP queries. These rules can specify, for example, whether annotated pauses and other non-verbal elements of spoken transcriptions should be ignored, treated as properties of words, or indexed as independent tokens. In addition, the mapping rules allow for the use of various filters and constraints to select only specific elements within a group or to modify them dynamically. This functionality can be used, for instance, to store word forms both in their original form and in lowercase, that would enable case-insensitive searches. Another example is the reduction of storage requirements in the search index by excluding speech rate annotations with the tag "0.0" labeling contributions that contain only non-verbal elements. Furthermore, integrating speaker and speech event metadata into transcripts (in the same way as other span-based annotations) and storing them in the search index together with corpus transcriptions and annotations enables metadata to be used as constraints directly in CQP queries (as demonstrated in the example in Section 2.3).

Indexing of transcript documents: Indexing of transcript documents is generally dependent on the specific ZuMult instance and varies according to the type of corpus data and the underlying user requirements. For example, all data may be stored in a single search index, or multiple complementary or alternative search indices may be created and used depending on the system goals. Because the corpora in the AGD collection vary considerably with respect to transcription conventions, the cross-corpus queries are rarely required, therefore a separate index file was created for each corpus. This approach keeps index sizes manageable and improves retrieval performance.

Corpora whose transcripts include punctuation marks are indexed in two different ways: once with punctuation and once without. This allows users to choose between two search modes and to specify token distances in their queries either with or without taking punctuation into account. In addition, two index types are created for each corpus: a transcript-based index, which indexes tokens according to their sequential position in the transcript document, and a speaker-based index, which processes tokens separately for each speaker while treating the tokenizations of other speakers as annotations of the current speaker. This approach helps to resolve conflicting tokenizations (cf. Sauer & Lüdeling, 2016) in spoken corpora that arise from speaker overlaps (for more details see Section 3.2 in Frick et al., 2022; Section 4.2. in Frick et al., 2023 and Section 5.4. in Frick & Schmidt, 2025).

Implementation of the search component: The search itself is performed via Java API for RESTful Web Services (JAX-RS), which establishes a connection to the ZuMult backend. Within the ZuMult backend, an instance of the *Searcher* class is created for each request. This class is responsible for preparing the search request before it is forwarded to the search engine and for post-processing the search results. For example, it determines which indices are to be queried and enriches the search results with various types of metadata, such as the time at which the search was executed.

The search engine then handles the parsing of the CQP query string, executes the query against the search index, and retrieves the required information from the index, such as the token IDs of the

matches or associated metadata that may later be needed for further sorting of the results. In addition, the search engine of the AGD ZuMult instance has been extended to support searches for repetitions and n-grams. For this purpose, the search engine performs several checks directly on the search index during query execution, such as whether both elements of a repetition originate from the same contribution or whether a speaker change occurs between the respective token positions.

Presentation of the search results: The ZuMult backend can return three types of search results. Each result type is modelled as an object within the ZuMult architecture and is implemented by a dedicated class with a specific interface. The *SearchResult* class contains only the number of hits and the number of transcripts in which they occur, along with selected metadata about the search process. *SearchResultPlus* additionally includes the token IDs of the matches and further metadata for each hit. The *KWIC* class is responsible for transforming search results into a KWIC object and for retrieving the relevant contextual information for each hit from the transcript files.

The subsequent serialization of the KWIC object into XML format is handled by the web services layer. Depending on the intended use of the search results – for example, whether they are requested by a ZuRecht instance (Section 2.3) or by the CLARIN Federated Content Search (FCS, Körner et al., 2024) – different XML serializations of the results may be produced.

Although MTAS is currently the search engine used in all active ZuMult instances, the ZuMult architecture is designed in such a way that the Search API can be extended at any time to include additional search engines. The use of multiple search engines and the flexible switching between them, should allow users to leverage the strengths of different query languages. Currently, it is being evaluated whether the BlackLab search engine (Does et al., 2017) is a suitable candidate to extend the ZuMult Query component. Like MTAS, BlackLab is capable of handling data in the ISO 24624:2016 format and can be adopted for new environments via a configuration file. Both MTAS and BlackLab are implemented in Java and built on the Apache Lucene information retrieval library, which have a large user and developer community as well as a strong long-term maintainability, providing a sustainable technical base for the ZuMult architecture.

3.4 Other components

Other components of ZuMult also rest on top of existing technology: For instance, for fine-grained multimodal analysis, transcript excerpts are force-aligned with their audio on the phoneme level using the CLARIN service WebMaus (Kisler et al., 2017) and still images extracted from the corresponding video file via the FFMPEG suite (https://ffmpeg.org/). This is supplemented with an F0 pitch contour extracted from the audio using Praat scripts and visualised as an SVG drawing (see Figure 8).

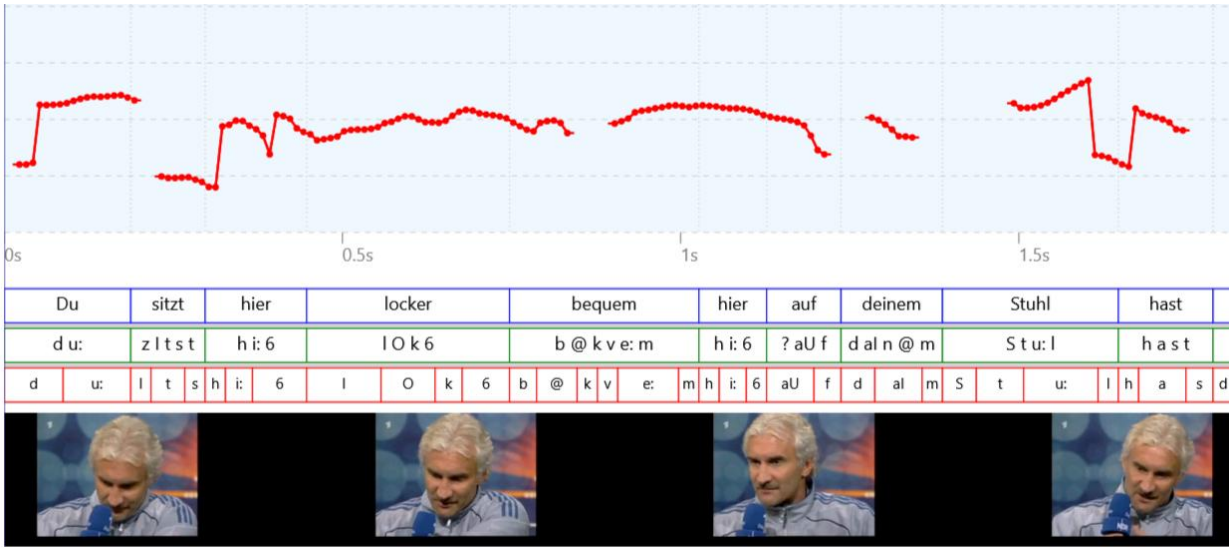

Figure 8: A micro-view of a transcript excerpt in ZuMult:
phoneme-level alignment achieved via MAUS web service, pitch contour calculated via Praat.

3.5 Interoperability with CLARIN

Last but not least, data published on an instance of the ZuMult platform can be directly integrated into the CLARIN infrastructure. This is the case, for example, when the platform obtains its actual data (transcripts, audio, video, metadata) from a CLARIN repository, as is the case in the ongoing

development of a ZuMult instance for the ESLO corpus (Schmidt, 2025) which is tied to the Cocoon platform of the French HumaNum infrastructure (https://cocoon.huma-num.fr/).

Additionally, ZuMult is currently extended to serve CLARIN protocols for FCS and become an endpoint for an FCS aggregator. The Transcription+ project (see above) will demonstrate this by setting up an FCS endpoint for the EXMARaLDA demo corpus (http://doi.org/10.25592/uhhfdm.8364) in collaboration with the CLARIN/Text+ centres in Hamburg.

4 Current and future developments

Current development of ZuMult is carried out at the University of Duisburg-Essen in two contexts:

- The ZAT project (https://zat.nrw/) deals with assistive technologies. Corpus data in this context comprise studies of human-machine interaction (e.g. people interacting with a multimodal chatbot or a remote instruction robot), which are recorded, not only on video from a number of perspectives, but also with sensor technology (eye trackers, biometrical sensors) and through machine log files.
- The research group on personal reference (https://forschungsgruppe-pronomen.de/en/) studies the use of personal, indefinite, and demonstrative pronouns used in German to refer to present and absent persons. One corpus in this context consists of data collected at emergency medical drills ("mass casualty incidents") in the form of videos from many perspectives, eye tracking data from main actors, and (perspectively) GPS data on actors' positions in the scene.

Challenges for ZuMult in both contexts include the handling of very large amounts of video data, the integration and synchronized visualisation of multi-modal annotations and additional time-series data (often with spatial coordinates), and the dynamic (often opportunistic) ways these data are processed, reflecting an orientation towards immediate needs of data exploration and analysis in the research process, rather than towards systematic long-term corpus building. The tighter integration of ZuMult with CI pipelines and the measures aimed at an optimization of the distribution of the platform described in Section 3.2, as well as the tools for fine-grained multimodal analysis illustrated in Section 3.4 are largely guided by these requirements. Additionally, we are exploring ways of interfacing ZuMult with the R programming language to exploit this framework's strengths in statistical analysis and data visualisation.

Prospectively, we are envisaging a second phase for the project to further develop the basic architecture of the platform in three different ways:

- an extension and generalization of the data model to (a) allow for more fine-grained processing (visualisation, navigation) of data, (b) cater also for non-canonical written data (e.g. learner texts) and (c) enable the integration of any kind of time-series data alongside audio, video and transcriptions.
- an extension of functionality for persistent collaborative work, enabling users to create, store and share individual collections of data inside the platform, and to enrich corpora with user-generated comments, references and annotations
- an exploration of LLM-based techniques for automatically enriching data and supporting queries

In the mid-term, we thus hope to extend ZuMult's user base and further establish it as a technology in different research communities dealing with audiovisual language data.

5 Conclusion

As Figure 9 illustrates, ZuMult, as a piece of technology, is built on top of lots of other pieces of technology which, in turn, are also built on top of each other, and it can be integrated in and with other solutions. The fact that the spread of best practices and the establishment of standards for the relevant technology is advancing, and that ties between such solutions are being strengthened, makes an undertaking such as ZuMult easier to implement, easier to maintain and, hopefully, relevant to a larger community of users.

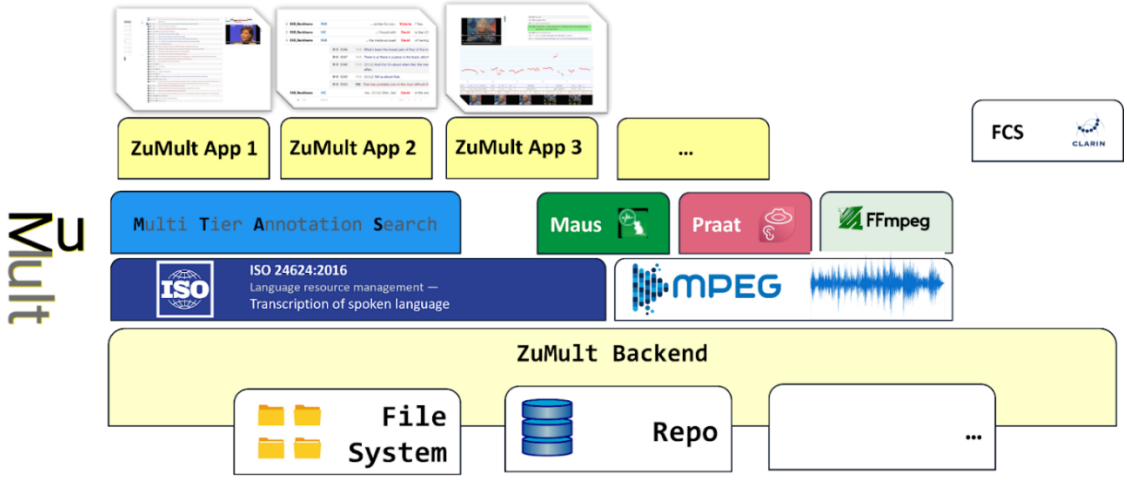

Figure 9: ZuMult and the things on top of which it is built

In our current work, we observe especially that, compared to large "monolithic" solutions, building a platform with such a modular approach makes it more manageable also for smaller projects to adapt and run locally. In the medium-term, this may have a positive effect on the community's ability to contribute to an infrastructure like CLARIN.

Acknowledgments

The original ZuMult project was funded by the German Research Council (DFG) from 01/12/2018 to 31/03/2022 (project number 359880971) as a collaborative project of the Leibniz-Institute for the German Language (program area 'Oral corpora', Thomas Schmidt), the University of Leipzig (Herder-Institute, Christian Fandrych) and the University of Hamburg (Hamburg Centre for Language Corpora, Kai Wörner).

Current work at the University of Duisburg-Essen is carried out with funding from the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen for the project „Zentrum Assistive Technologien Rhein-Ruhr, ZAT Rhein-Ruhr" (PB22-076D) and from the DFG for the project "Multimodalität von Personenreferenz: Pronominale Personenreferenz in Notfallübungen von Medizinern und Feuerwehr" as part of the research group "Praktiken der Personenreferenz: Personal-, Indefinit- und Demonstrativpronomen im Gebrauch" (FOR 5317), project number 457855466. Karola Pitsch is the principal investigator on both projects.

The Transcription+ project is a cooperation project of the Text+ consortium, funded by the DFG (project number 460033370) as part of the German National Research Data Infrastructure (NFDI). Thomas Schmidt is the principal investigator. The Academy of Sciences and Humanities in Hamburg and the University of Hamburg are supporting the project in Text+ integration.

We would also like to gratefully acknowledge the ongoing collaborations with the University of Texas in Austin (Hans C. Boas, Margaret Blevins) and with the Laboratoire Ligérien de Linguistique at the University of Orléans (Flora Badin, Lotfi Abouda, Céline Dugua, Marie Skrovec).

References

Batinić, J., Frick, E. & Schmidt, T. (2021): *Accessing spoken language corpora: an overview of current approaches.* In: Corpora Vol. 16, Issue 3. Edinburgh: Edinburgh University Press, 2021. 417-445. https://doi.org/10.3366/cor.2021.0229

Boas, H., Schmidt, T. & Blevins, M. (2026, accepted). *TGDA 2.0 - A corpus platform for the Texas German Dialect Archive.* Language Resources & Evaluation. Springer.

Brinckmann, C., Kleiner, S., Knöbl, R. & Berend, N. (2008). *German Today: an areally extensive corpus of spoken Standard German.* Proceedings of LREC 2008, Marrakech, Morocco. European Language Resources Association (ELRA), 2008. S. 3185-3191.

Brouwer, M., Brugman, H. & Kemps-Snijders, M. (2016). *MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution.* Selected papers from the CLARIN Annual Conference 2016. Aix-en-Provence, 19-37.

Dickgießer, S. (2017). *Metadaten im Programmbereich „Mündliche Korpora" des IDS.* Mannheim: Leibniz-Institut für Deutsche Sprache. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62525

Does, J., Niestadt, J. & Depuydt, K. (2017). *Creating research environments with BlackLab.* Odijk, J. & van Hessen, A. (eds.) CLARIN in the Low Countries, 151-165. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi

Evert, S. & CWB Development Team (2022). *The IMS Open Corpus Workbench (CWB) - CQP Interface and Query Language Manual CWB Version 3.5.* https://cwb.sourceforge.io/files/CQP_Manual.pdf

Fandrych, C., Meißner & C., Slavcheva, A. (2012). *The GeWiss corpus: Comparing spoken academic German, English and Polish.* Schmidt, T. & Wörner, K. (eds.): Multilingual Corpora and Multilingual Corpus Analysis. Hamburg Studies in Multilingualism 14. Amsterdam: John Benjamins, 319-338.

Fandrych, C., Frick, E., Hedeland, H., Iliash, A., Jettka, D., Meißner, C., Schmidt, T., Wallner, F., Weigert, K. & Westpfahl, S. (2016). *User, who art thou? User profiling for oral corpus platforms.* Proceedings of LREC 2016, Portorož, Slovenia. Paris: ELRA, 280-287.

Fandrych, C., Schmidt, T., Wallner, F. & Wörner, K. (eds.) (2023a). *Zugänge zu multimodalen Korpora gesprochener Sprache.* Special issue of Korpora Deutsch als Fremdsprache (3) 1. https://kordaf.tujournals.ulb.tu-darmstadt.de/issue/92/info/

Fandrych, C., Meißner, C., Schwendemann, M. & Wallner, F. (2023b). *ZuMal: Zielgruppenspezifische Gesprächsauswahl aus Korpora gesprochener Sprache.* In: Fandrych et al. (2023a), 13-43.

Ferger, A., Krause, A. & Pitsch, K. (2023). *A Continuous Integration (CI) Workflow for Quality Assurance Checks for Corpora of Multimodal Interaction.* Lindén, K. / Niemi, J. / Kontino, T. (eds.), CLARIN Annual Conference Proceedings 2023. Leuven, 106-110.

Fisseni, B. & Schmidt, T. (2020). *CLARIN Web Services for TEI-annotated Transcripts of Spoken Language.* Simov, K. & Eskevich, M. (eds.): Selected Papers from the CLARIN Annual Conference 2019. (= Linköping Electronic Conference Proceedings 172). Linköping: Linköping University Electronic Press, 2020. 12-22.

Frick, E., Helmer, H. & Wallner, F. (2023). *ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache.* Fandrych et al. (2023a), 44-71.

Frick, E. & Schmidt, T. (2020). *Using full text indices for querying spoken language data.* Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora, Marseille, France, May. European Language Resources Association, 40–46.

Frick, E. & Schmidt, T. (2025). *Querying spoken language data.* Bański, P., Heid, U. & Herzberg, L. (eds.): Standards for language data and infrastructures. Series 'Digital Linguistics'. De Gruyter.

Frick, E., Schmidt, T. & Helmer, H. (2022). *Querying Interaction Structure: Approaches to Overlap in Spoken Language Corpora.* Proceedings of LREC 2022. Marseille, France. European Language Resources Association, 715–722.

Frick, E., Lösel, A., Messner, A., Schmidt, T., Schwendemann, M. & Wallner, F. (2025). *ExpoKo – eine Korpusressource für das wissenschaftliche Schreiben.* Große Lernerkorpora - Möglichkeiten und Grenzen, Leipzig. Zenodo. https://doi.org/10.5281/zenodo.17207106

Grän, J., Schaber, J., McDonald, D., Mustač, I., Rajović, N., Schneider, G., Vuković, T., Zehr, J. & Bubenhofer, N. (2024): *The LiRI Corpus Platform (LCP): A software system and infrastructure for querying a vast array of corpora of different kinds.* CLARIN Annual Conference 2023. https://doi.org/10.3384/ecp210010

Hedeland, H. & Ferger, A. (2020). *Towards Continuous Quality Control for Spoken Language Corpora.* International Journal of Digital Curation 15. https://doi.org/10.2218/IJDC.V15I1.601

Hedeland, H. & Schmidt, T. (2022). *The TEI-based ISO Standard 'Transcription of spoken language' as an Exchange Format within CLARIN and beyond.* Monachini, M. & Eskevich, M. (eds.): Selected Papers from the CLARIN Annual Conference 2021. Virtual Event, 2021, 27–29 September. (= Linköping Electronic Conference Proceedings 189). Linköping: Linköping University Electronic Press, 2022, 34-45.

Kisler,T., Reichel, U. & Schiel, F. (2017). *Multilingual processing of speech via web services*. Computer Speech & Language, Volume 45, September 2017, pages 326–347.

Körner, E., Eckart, T., Helfer, F. & Kretschmer, U. (2024). *An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences.* CLARIN Annual Conference, 49–59. https://doi.org/10.3384/ecp216.05

Krause, T. & Zeldes, A. (2016): ANNIS3: A new architecture for generic corpus query and

visualization. In: *Digital Scholarship in the Humanities*, 31(1). 118-139. https://doi.org/10.1093/llc/fqu057

Kren, M. (2020): *Czech National Corpus in 2020: Recent Developments and Future Outlook.* In Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora, Marseille, France. European Language Ressources Association, 52–57.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017): *The Spoken BNC2014: designing and building a spoken corpus of everyday conversations.* In: International Journal of Corpus Linguistics, 22:3. 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

Monty Python (1970): *Royal Society For Putting Things On Top of Other Things.* https://www.youtube.com/watch?v=LFrdqQZ8FFc

Pęzik, P. (2015): *Spokes – a Search and Exploration Service for Conversational Corpus Data.* In Selected Papers from the CLARIN 2014 Conference. Linköping Electronic Conference Proceedings. Linköping University Electronic, 99–109.

Pluschkovits, M., Wittibschlager, A., Schopper, D., Bal, J., Kukelka, K., Reichl, O., Rohler, M. & Lenz, A. (2026): *LAPIS: ein Portal zu Sprachressourcen in Österreich.* Contribution to the methods fair of the annual conference of the Institute for the German Language. https://www.ids-mannheim.de/fileadmin/aktuell/Jahrestagungen/2026/Methodenmesse/02_LAPIS.pdf

Reichel, U.D., Kisler, T. (2014): *Language-independent grapheme-phoneme conversion and word stress assignment as a web service.* In: Hoffmann, R. (Ed.): Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation 71, pp 42-49, TUDpress, Dresden.

Sauer, S. & Lüdeling, A. (2016). *Flexible Multi-Layer Spoken Dialogue Corpora.* International Journal of Corpus Linguistics 21(3), Special Issue on Spoken Corpora, 419-438.

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T. & Sloetjes, H. (2009): *An exchange format for multimodal annotations.* In Kipp, M., Martin, J., Paggio, P. & Heylen, D. (eds.): Multimodal corpora: from models of natural interaction to systems and applications. Berlin/Heidelberg: Springer, 2009, 207-221.

Schmidt, T., Elenius, K. & Trilsbeek, P. (2010). *Multimedia Corpora (Media encoding and annotation).* Zenodo. https://doi.org/10.5281/zenodo.18269160.

Schmidt, T. (2011). *A TEI-based approach to standardising spoken language transcription.* Journal of the Text Encoding Initiative 1. 2011. https://doi.org/10.4000/jtei.142

Schmidt, Thomas (2014a). *The Database for Spoken German – DGD2.* Proceedings of LREC 2014, Reykjavik, Iceland. European Language Resources Association (ELRA). 1451-1457.

Schmidt, Thomas (2014b). *(More) Common Ground for Processing Spoken Language Corpora?* In: Ruhi, Ş., Haugh, M., Schmidt, T. & Wörner, K. (eds.): Best Practices for Spoken Corpora in Linguistic Research. Newcastle: Cambridge Scholars Publishing, 2014. S. 249-265.

Schmidt, T. & Wörner, K. (2014). *EXMARaLDA.* Durand, J., Gut, U. & Kristoffersen, G. (eds.): The Oxford Handbook o f Corpus Phonology, Oxford: OUP, 402-419.

Schmidt, T. (2016). *Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project.* Kupietz, M. & Geyken, A. (eds.): Corpus Linguistic Software Tools. Journal for language technology and computational linguistics (JLCL) 31 (1). Berlin: GSCL, 2016. 127-154. https://doi.org/10.21248/jlcl.31.2016.205

Schmidt, T., Hedeland, H. & Jettka, D. (2017). *Conversion and annotation web services for spoken language data in CLARIN*. In: Borin, L. (ed.): Linköping Electronic Conference Proceedings; Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure. Linköping: Linköping University Electronic Press, 2017. S. 113-130.

Schmidt, T., Schwendemann, M. & Wallner, F. (2023). *ZuViel: Transkriptvisualisierung und Arbeiten mit Transkripten.* Fandrych et al. (2023a), 72-91.

Schmidt, T. & Frick, E. (2025). *Reading, listening to and watching concordances of audiovisual interaction corpora.* Poster presented at "Symposium: Reading Concordances in the 21st Century", University of Erlangen, 19-21 March 2025. https://doi.org/10.5281/zenodo.15094134

Schmidt, T. (2025). *Représenter et accéder à la parole dans les corpus oraux : diversification et adaptation des méthodes et technologies.*Kanaan-Caillol,L., Dugua, C. & Gerstenberg, A. (eds.), Représenter la parole. De Gruyter, 185-208.

Schmidt, T. (2026, accepted). *Chapter 3: Typology of Non-Textual Language Data.* Witt, A., Lenz, A., Kamocki, P. (eds.): Routledge Handbook of Digital Linguistics. London: Routledge.

Schmidt, T., Blevins, M., Boas, H. & Gilbert, G. (2026, accepted). *The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact.* Proceedings of LREC Workshop "Dialects in NLP: A Resource Perspective".

Sidnell, J. (2012). *Basic Conversation Analytic Methods.* In: Sidnell, S. & Stivers, T. (eds.): The Handbook of Conversation Analysis. Oxford: Blackwell, 77-99.

Wamprechtshammer, A., Arestau, E., Aznar, J., Hedeland, H., Isard, A., Khait, I., Lange, H., Majka, N. & Rau, F. (2022). *QUEST. Guidelines and Specifications for the Assessment of Audiovisual, Annotatated Language Data.* Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology. Vol. 8 (2022). https://ojs.bibl.u-szeged.hu/index.php/wpcl/issue/view/2474/382.

Westpfahl, S. & Schmidt, T. (2016). *FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German.* Proceedings of LREC 2016, Portoroz, Croatia. European Language Resources Association (ELRA), 2016. 1493-1499.

Wörner, K. (2012). *Finding the balance between strict defaults and total openness - Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager.* In: Schmidt, T. & Wörner K. (eds.): Multilingual Corpora and Multilingual Corpus Analysis: 383–400, 18 S., Vol. 14, Amsterdam; Philadelphia: John Benjamins Pub. Co (Hamburg Studies on Multilingualism).

The CLARIN Resource Families Workflow

Jakob Lenardič
Institute of Contemporary History
Ljubljana, Slovenia
jakob.lenardic@inz.si

Alexander König
CLARIN ERIC
alex@clarin.eu

Kristina Pahor de Maiti Tekavčič
Institute of Contemporary History
Ljubljana, Slovenia &
Faculty of Arts, University of Ljubljana
kristina.pahordemaiti@ff.uni-lj.si

Dieter Van Uytvanck
CLARIN ERIC
dieter@clarin.eu

Abstract

This paper presents the new workflow for the CLARIN Resource Families (CRF) initiative. After introducing the CRF initiative and laying out its primary aim, we present the functioning of the new CRF backend, which uses a JSON-based workflow publicly accessible on GitHub. We then present the guidelines for working with the JSON files and describing the metadata. We also discuss how the new workflow enables a collaborative approach to maintaining the families. Finally, we discuss how the CRF complements and is complemented by CLARIN ERIC's Virtual Language Observatory.

1 Introduction

A key aim of the CLARIN Resource Families (CRF) initiative (Lenardič & Fišer, 2022b) is the promotion of mature language resources and technologies (LRTs), such as language corpora, lexical resources, and tools, that are provided by national CLARIN consortia. The way in which the CRF initiative specifically contributes towards this goal is by offering access to manually curated overviews of LRTs, which are grouped together along thematic lines into 'families', such as *parliamentary corpora*, *tools of sentiment analysis*, *language models*, and so on. Importantly, all other key CLARIN services that automatically collate LRTs dispersed throughout the CLARIN infrastructure, such as the Virtual Language Observatory (VLO; Goosen and Eckart, 2014; Van Uytvanck et al., 2010, 2012), do not provide search facets that would allow researchers to pick out tools or resources belonging just to a particular family to the exclusion of everything else. The lack of such search options also follows in part from the fact that taxonomic categories like parliamentary corpora do not correspond to metadata components in CMDI (Windhouwer & Goosen, 2022). By providing the manually curated thematic groupings, the CRF initiative fills this gap in resource and tool discovery, and thereby facilitates the increased findability of CLARIN LRTs from the perspective of the FAIR principles (Wilkinson et al., 2016).

When the CRF initiative launched in 2018, it consisted of just 4 corpus families, whose creation had a project-internal motivation at the time (Fišer et al., 2018).[1] However, the initiative soon proved popular especially amongst non-CLARIN-affiliated researchers in the (digital) humanities and social sciences, who often first encounter the CLARIN infrastructure through a CRF webpage.[2] Consequently, the CRF overviews have been iteratively expanded throughout the years with the addition of new families and LRTs so that on 1 April 2026 there are 16 corpus families, 6 families of lexical resources, and 6 tool families, which together amount to around 1600 manually curated LRTs.

Because of this expansion, there is a growing danger that the individual families will sooner or later become deprecated, especially since the upkeep of the families has thus far been largely a CLARIN ERIC–internal affair. We have therefore developed a new backend for the CRF overviews that is more

[1] Specifically, it was tied to workshops organised in the context of the CLARIN Plus project: https://www.clarin.eu/content/factsheet-clarin-plus.

[2] CRF webpages are typically the first or second result on Google for search terms like 'parliamentary corpora' or 'parallel corpora'.

Jakob Lenardič, Alexander König, Kristina Pahor de Maiti Tekavčič, and Dieter Van Uytvanck 2025. The CLARIN Resource Families Workflow. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 89–99. DOI: https://doi.org/10.3384/ecp222

geared towards collaboration than what was previously the case. In this paper, we present the functioning of this new backend and the ways in which involvement has now become possible.

The rest of the paper is organised as follows. In Section 2, we first introduce the technical aspects of the new workflow and guidelines for using it. In Section 3, we present the ways in which collaborators can become involved in the upkeep of the CRF overviews as well as our own ongoing efforts for facilitating a greater degree of external involvement. In Section 4, we discuss how the CRF complements and is in turn complemented by the VLO both technologically and conceptually. Section 5 concludes the paper.

2 The GitHub-based workflow for creating CRF overviews

2.1 Main characteristics

```
{
    "Name": "Slovenian parliamentary corpus siParl 4.0",
    "URL": "http://hdl.handle.net/11356/1936",
    "Family": "Parliamentary corpora",
    "Description": "The corpus contains Slovenian parliamentary debates from 1990 to
        2022.\nThe corpus is encoded according to the <a href=\"
       https://github.com/clarin-eric/parla-clarin/\">Parla-CLARIN</a> schema. In terms
       of linguistic annotation, the part-of-speech tags, morphological analyses,
       and syntactic parses in the corpus follow the <a href=\"
       https://universaldependencies.org/sl/\">Universal Dependencies</a> framework.\
       nThe corpus is available for download from the CLARIN.SI repository and for
       online querying through the noSketch Engine concordancer.",
    "Language": ["slv"],
    "Licence": "CC BY 4.0",
    "Size": ["13,233 texts", "1,223,250 utterances", "239,224,540 words",
       "273,023,945 tokens"],
    "Annotation": ["tokenised", "PoS-tagged", "lemmatised", "syntactically parsed",
       "annotated for named entities"],
    "Infrastructure": "CLARIN",
    "Access": {
        "Concordancer": "https://www.clarin.si/ske/#dashboard?corpname=siparl40",
        "Download": "http://hdl.handle.net/11356/1936"
            },
    "Publication": "https://www.zotero.org/groups/562080/clarin/items/K24SGCXC"
}
```

Figure 1: A CRF JSON file.

The new CRF backend is using a Python-based workflow on GitHub.[3] The workflow takes as input a set of JSON files located in three folders (*corpora*, *lexical-resources* and *tools*) and then, with the help of Drupal, converts them to tabular HTML code that is periodically and automatically fed into the online overviews hosted at https://www.clarin.eu/resource-families. Each JSON file corresponds to an entry in the tabular online overviews.

The built-in Drupal features ensure that the CRF overviews have started behaving in a similar way to a generic search engine. The overviews are therefore no longer entirely static – it has become for instance possible to filter the CRF entries so that only resources in a specific language or with a specific annotation layer are shown. It is also possible to use a single filter across all of the CRF overviews to generate a list of high-quality resources for a specific language or even annotation layer, which is a feature that CRF users have long been asking for.

For instance, the workflow maps the contents of the JSON file in Figure 1, which is for the Slovenian parliamentary corpus *siParl 4.0* (Pančur et al., 2024), to HTML code that is then included as part of the Parliamentary corpora resource family.[4] The organisation into subsections seen on the CRF webpages, such as the distinction between CLARIN-affiliated and CLARIN-external LRTs (a distinction that is being phased out; see Section 2.3), along with the division between monolingual or multilingual LRTs, is derived

[3]https://github.com/clarin-eric/clarin-resource-families/

[4]https://www.clarin.eu/resource-families/parliamentary-corpora

from name/value pairs such as the one for *Infrastructure* (line 10 in the figure; here concretely specified as CLARIN-internal) and the number of values in the *Language* array (specified just for Slovenian in line 6; hence, the JSON CRF entry for *siParl 4.0* is mapped to the Monolingual subgroup).

The name/value pairs shown in the figure mostly recur throughout the corpus and lexical-resource families, but not the tool families, so potential users of the workflow are instructed to re-use an existing file in the repository whenever they begin to work on a particular family.

2.2 The guidelines for the workflow

For working with the GitHub-based workflow, a set of guidelines has been prepared by Lenardič and König (2025). The guidelines are based on the existing recommendations of the CLARIN.SI consortium for the documentation of language resources,[5] and on proposals by Odijk (2019, pp. 122–123) for the documentation of language-processing tools.

In the first part, the guidelines present a beginner-friendly tutorial for editing and creating the JSON files. This tutorial assumes no prior knowledge of JSON syntax, so it also highlights rather trivial but often recurring JSON requirements, such as the obligatory use of $\backslash n$ operators for line or paragraph breaks and the usage and function of quotation marks.[6]

The second part of the guidelines, however, is more descriptive in nature, and is crucial for researchers both with or without experience in working with JSON syntax. This part lays out the requirements for the JSON values, some of which have been semi-standardised following recommendations in the related literature (e.g., Bański et al., 2025).

In what follows, we discuss the guidelines for each name–value pair in turn while highlighting some existing non-trivial issues as well.

2.2.1 Language codes

The values for the *Language* array in line 6 in Figure 1 must be specified using the ISO 639-3 language set, which is in line with the recommendation of the CLARIN Standards and Interoperability Committee (Bański & Hedeland, 2022).

In terms of natural-language coverage, the ISO 639 set is quite comprehensive, specifying three-letter codes for 'spoken, written and signed languages, including extinct or historical languages' (Romary, 2025, p. 40). For instance, extinct languages in CRF include corpora in Latin and Old Norse, such as the *LatinISE corpus (version 5)* (McGillivray, 2025) and *The Saga Corpus (Fornritin)* (Helgadóttir & Barkarson, 2018), whose languages are specified with the ISO 639-3 codes *lat* and *non* respectively.

We note, though, that the ISO 639 codes sometimes do not provide the greatest coverage when it comes to varieties within a national language, even those widely spoken in Europe and the Americas. For instance, the Portuguese CLARIN node PORTULAN hosts several corpora dedicated to variants of Portuguese spoken not only in Portugal, but across the world as well. One such resource is *CorPop: a corpus of popular Brazilian Portuguese* (Finatto & Grilo, 2020). However, there is no ISO 639-3 code that would distinguish Brazilian Portuguese from its European variant.[7]

There is a language catalogue whose granularity is finer than that of ISO 639; namely, Glottolog (Nordhoff, 2012).[8] If CRF were to adopt Glottolog for reasons of better language coverage, then *CorPop: a corpus of popular Brazilian Portuguese* (Finatto & Grilo, 2020) could employ two of the so-called glottocodes (see Forkel & Hammarström, 2022) – namely, *port1283* for the Portuguese macro variant,[9] whose corresponding ISO 639-3 code is *por*,[10] as well as *braz1246* for the Brazilian subvariety,[11] which

[5] https://www.clarin.si/repository/xmlui/page/deposit

[6] The paragraph breaks in the CRF descriptions do not necessarily correspond to the breaks in the 'free-text' description in the original metadata. Sometimes, the latter descriptions do not use paragraph breaks at all, or any other textual formatting either, in contrast to the CRF ones, which strive to be consistent about text placement and layout. Note that CRF descriptions are not meant to be copies of the descriptions in the repository entries, but rather serve as partially unified descriptions briefly recapping the main research-domain-specific characteristics of the data.

[7] https://tinyurl.com/ezdxh9bj

[8] We thank Joshua Wilbur for bringing this database to our attention.

[9] http://glottolog.org/resource/languoid/id/port1283

[10] https://iso639-3.sil.org/code/por

[11] http://glottolog.org/resource/languoid/id/braz1246

lacks the corresponding ISO 639-3 code.

Another macro language that could benefit from a move to Glottolog is English, which is still the language of the majority of CRF corpora. In the 5.3 version of Glottolog (Hammarström et al., 2026), there are 34 glottocodes for variants of English under 'Macro English',[12] whereas there are 23 ISO 639-3 codes for the same macro language.[13] For the purposes of the CRF initiative, a switch to Glottolog would then mean that a corpus like *Corpus of Historical American English – Kielipankki Korp version 2017H1* (Davies, 2017), which belongs to the Historical corpora family,[14] could be tagged with the glottocode *nort3314*,[15] which denotes the North American variant and does not have an ISO 639-3 equivalent.

While the coverage of Glottolog is greater than that of ISO 639, there is a downside from the perspective of manual curation, which is the complexity of the glottocodes themselves; compare the ISO 639-3 code for Portuguese – that is, *por* – with the corresponding and lengthier glottocode – that is, *port1283*. This could be problematic in case curators type out the JSON values by hand instead of copying them from somewhere, which they often do in fact. Thus, CRF is for the time being sticking with the ISO 639-3 set, although a switch to Glottolog might be considered in the future due to its greater breadth of linguistic coverage.

2.2.2 Licence and size

Licence (line 7 in Figure 1) is written using the canonical abbreviation (i.e., CC BY 4.0 in this case), although in the future another name–value pair might be added for licence URL, which would be especially helpful for lesser known licences such as CLARIN PUB.[16]

It is worth mentioning, however, that not all CLARIN B- or C-centres consistently use the canonical abbreviations – for instance, the licence of the lexical resource *ViPER verb lexical database* (Mamede et al., 2020) is spelt as 'CC - BY - NC - CD' rather than 'CC BY-NC-CD' (notice also the missing licence version) in the original metadata records. Even in such cases, the CRF overviews stick to the canonical forms, which is in line with the idea from Section 2.2.4 and Footnote 6 that CRF entries are not meant to be copies of metadata records from the repositories, especially if there are issues with the latter.

Size (line 8) is provided in as many sensible categories as possible, where the categories should be relevant for cross-resource comparison (e.g., number of tokens vs. number of words). Just like in the case of licence, the CRF overviews deviate from the metadata records when the latter provide size only in bytes or the number of files included in the deposit, as such information – although common – is not really valuable from a comparative perspective.

2.2.3 Annotation

Annotation (line 9 in Figure 1) uses past-participle forms, so that linguistic annotation layers are described with terms like *tokenised*, *lemmatised*, *PoS-tagged*, and *syntactically parsed*. Sometimes, we use *MSD-tagged* for part-of-speech tags that define complex morphosyntactic features,[17] as for instance those used by Slavic corpora.

An issue with this past-participle approach, however, is that certain other annotation processes such as 'named entity recognition' do not trivially admit such forms. The reason why we are currently sticking to the flawed past-participle approach is that there does not seem to exist a consolidated vocabulary of annotation layers that could be re-used in the overviews; see Lenardič (2023) for further discussion about the diversity of annotation descriptors within CLARIN. We note that a possible solution would be to switch to simple nominal forms, so that the Annotation layers could simply be listed with descriptors like 'tokens', 'lemmas', and 'PoS tags' instead of the past participles.

Importantly, the *Annotation* JSON name–value pair is meant as an array for listing the annotation layers, not for describing them. For further specifying any additional characteristics of the annotation

[12] https://glottolog.org/resource/languoid/id/stan1293

[13] https://tinyurl.com/rfr4vrnp

[14] https://www.clarin.eu/resource-families/historical-corpora

[15] https://glottolog.org/resource/languoid/id/nort3314

[16] https://www.clarin.eu/content/licenses-and-clarin-categories#1-pub-language-resources

[17] The acronym MSD refers to 'a morphosyntactic description of a token, such as the Universal Dependencies morphological features' (Ljubešić & Erjavec, 2025, p. 81).

process, curators are encouraged to use the *Description* name–value pair in line 5 of the figure, to which we know turn.

2.2.4 The 'free-text' description

A problem that commonly recurs across the CLARIN infrastructure is that the deposited LRTs are described in various under-informative ways. Consequently, it is often difficult to determine what kind of data the LRT actually consists of or what the functionality is in the case of tools, especially if the name of the LRT in question is opaque (see Lenardič & Fišer, 2022a, for concrete examples).

The guidelines therefore pay particular attention to the value of the *Description* field (line 5 in Figure 1), providing a list of suggestions for describing each type of LRT separately. We now overview these recommendations for each type in turn.

For describing corpora, curators are advised to provide at least some of the following information: modality (spoken, written, visual, etc.), time period (based on publication date), geographic coverage, data sampling (text types and their ratios; text sources and their ratios), envisaged research domains, and any important and unique domain-specific characteristics (e.g., the range of participants' ages and L1s in L2 learner corpora).[18] Curators are also encouraged to use this name–value pair for further specifying the characteristics of linguistic annotation, including which morphosyntactic tagset (e.g., MULTEXT-East; Erjavec, 2012, 2017) or syntactic framework (e.g., Universal Dependencies; De Marneffe et al., 2021) has been used for tagging or parsing the corpus.

For lexical resources, curators are advised to tailor the description to the subtype of the resource. For training lexica (e.g., *MorfFlex CZ 2.1*; Hajič et al., 2024), they should describe the key features of the annotation process (e.g., tagset, manual mark-up) and other structural information (e.g., lemma frequencies). For dictionaries (e.g., *Words of the 16th-Century Slovenian Literary Language*; Ahačič et al., 2011), curators should briefly overview how the entries are structured and what kind of grammatical (e.g., morphological information, collocation properties, phonemic transcription) and non-grammatical information (examples of use, contextual features) is included.

Lastly, for software, curators are encouraged to provide information on the tool's applicability in terms of the relevant research domain(s), its distribution and installation requirements, as well as its output and input characteristics, which includes MIME-types, annotation schemata, and tagsets. Furthermore, they should provide information that is unique to the functionality, such as the types and granularity of categories recognised by a named entity recogniser (e.g., *NameTag*; Straka and Straková, 2014; Straková and Straka, 2025) or the types (e.g., sentence-level, document-level) and levels (binary, ternary, quaternary, etc.) of sentiment recognised by a sentiment analyser (e.g., *OptaHopper*; Żak and Skuczyńska, 2018).

The following is a CRF description of the Slovenian *Gigafida 2.0* (Krek et al., 2019) reference corpus. The first paragraph provides brief information about the corpus by summarising the modality, data sources and time period; the second paragraph specifies encoding and non-linguistic annotation; lastly, the third paragraph specifies the availability.

> This corpus includes representative Slovenian texts (newspapers, magazines, computer-mediated communication, fiction and non-fiction) published between 1990 and 2018.
>
> The corpus is encoded in TEI. Non-linguistic metadata include information on source, year of publication, text type, title, and author.
>
> The corpus is available for online browsing through the noSketch Engine concordancer (CLARIN.SI distribution), as well as through a dedicated search engine.

What is important is that the description is not the same as that of the original metadata record (see Krek et al., 2019), as the latter is quite lengthier in fact. These sort of succinct three-to-four paragraph descriptions are consistently used in CRF overviews. Ensuring such presentational consistency is particularly important now that external collaborators have become involved in the initiative (Section 3).

[18] https://www.clarin.eu/resource-families/L2-corpora

2.2.5 Publication

Publication values (line 15 in Figure 1) correspond to hyperlinks to records within the CLARIN Resource Families subfolder of the CLARIN Zotero Library.[19] In the figure, the hyperlink is to the Zotero record of the paper by Meden et al. (2025), which describes the *siParl 4.0* (Pančur et al., 2024) corpus.

2.3 Inclusion criteria

CRF stopped accepting submissions of CLARIN-external LRTs in 2025; however, the existing 'CLARIN external' subsections are for now being kept as legacy pages. If these are removed at one point (though the JSONs will not be lost due to GitHub's Version Control), then the 'Infrastructure' name–value pair (line 10 in Figure 1) might acquire another function, perhaps to specify the CLARIN centre associated with the LRT.

We now briefly address what is 'internal' to CLARIN and can therefore be considered as a candidate for inclusion in the CRF overviews. The main criterion for inclusion is that corpora and lexical resources must be deposited in a CLARIN C- or B-centre. Such resources are consequently also findable through the VLO. This criterion is less stringent for tools, which are often found in repositories that are better tailored to software, such as GitHub.

Importantly, CRF also no longer hosts resources or tools that are wholly inaccessible – that is, a resource or tool must either be deposited with a C- or B-centre or be accessible online; in Figure 1, for instance, the online accessibility of *siParl 4.0* (Pančur et al., 2024) is concretely defined as being through CLARIN.SI's noSketch Engine concordancer. In other words, this means that the *Access* array beginning in line 11 of Figure 1 should not be empty. The motivation behind this decision is tied to CRF's mission to provide listings of *reusable and mature* LRTs important for research within the (digital) humanities and social sciences. This is also a point in which the CRF overviews differ from the VLO (see Section 4), as the latter also provides listings of LRTs that are only accompanied by the metadata, but do not include the data themselves.

Another difference from the VLO is that there should generally be a consistent level of granularity in the CRF – only entire resources (e.g., an entire corpus) are typically included in CRF, but not their possible subsets, which sometimes correspond to their own repository entries. Similarly, only one (i.e., typically the latest one) version of a resource or tool is included in a CRF overview.

2.4 Metadata curation

As already stated, CRF entries are rarely just repeats of the metadata records in the CLARIN repositories; aside from producing concise descriptions comparable across the resources and tools (Section 2.2.4), they try to fill in potential gaps in metadata. The latter happens whenever any metadata are missing from the CMDI record of the deposited entry but turn out to be actually available elsewhere – typically in presentational materials related to the LRT or directly from the author(s).

As a by-product of this, the CRF initiative also contributes to ongoing curation efforts at the CLARIN ERIC level. This is done by maintaining a list of GitHub issues that spell out all the missing or unclear metadata identified whenever a new LRT is added to the CRF overviews or even when a new family is created. These GitHub tickets are assigned via the labelling system to national consortia whose representatives are then expected to resolve them.[20] While these issues were until recently collated in a dedicated repository,[21] they are now going to be listed as part of the new workflow (Section 2.1) and linked to the JSON files themselves.

Centre	CRF Family
CKCMC	CMC Corpora
ACE	Corpora of Disordered Speech
CKL2CORPORA	L2 Learner Corpora
CLARIN-MULTISENS	Multimodal Corpora
K-OAr	Oral History Corpora
LLMs4SSH	Language Models
DR-LIB	Corpus Query Tools
CLARIN ELEXIS	Dictionaries
ACE	Sign Language Resources

Table 1: K-centres as curators of particular CRF overviews

3 Involvement

3.1 Using the GitHub-based workflow

In addition to having a more robust technical backend in general, the new setup for the CRF also provides a much easier way for collaboration. With each item of each family being contained in its own JSON file in the GitHub repository, collaborators can simply use the normal git-based way of collaboration. They can fork the repository, edit the specific items that they want to contribute to, or suggest adding new ones or deleting deprecated ones. Afterwards, they can make a pull request, which notifies the maintainers of the CRF working at the central hub, who in turn review the suggested changes. The accepted changes are immediately picked up by the CI pipeline and are reflected on the CLARIN website as well. As described in Section 2.4, the repository also stores open issues in GitHub tickets, so less tech-savvy users can simply create tickets with their suggested changes.

3.2 Domain expertise via CLARIN K-centres

Having the possibility for easy collaboration also means that it makes sense now to involve domain experts in helping with the curation of the CRF. In collaboration with the Knowledge Infrastructure Committee,[22] we have started reaching out to various CLARIN Knowledge Centres that seem to have an obvious connection to one of the families As of 29 August 2025, nine K-centres have agreed to become maintainers, each of their own resource family – they are presented in Table 1.[23]

Such collaboration is beginning to ensure that the work of maintaining the CRF is shared over many people, which will help make the effort more sustainable in the long term.

4 CLARIN Resource Families in comparison to the VLO

4.1 Conceptual and technological differences

We lastly address how the CRF overviews crucially differ from but ultimately complement the VLO (Goosen & Eckart, 2014; Van Uytvanck et al., 2010, 2012), which is CLARIN ERIC's flagship LRT discovery service.

The fundamental difference between the CRF and the VLO boils down to a conceptual as well as technological one. The listings of LRTs in the VLO originate from the periodic and crucially automatic

[19] https://www.zotero.org/groups/562080/clarin/collections/GDJJGIBW/collection

[20] Thus far approximately 70% of the tickets have been closed, with many existing metadata records having also been updated directly because of this. This curation effort has also led to the systematic overhaul of deprecated metadata records of corpora belonging to the legacy LRT collection. An example of such metadata correction pertains to the *Croatian National Corpus* (Tadić, 2014), which initially had a non-working hyperlink to an access page where it had once been browsable.

[21] https://github.com/clarin-eric/resource-families-issues/issues

[22] https://www.clarin.eu/governance/knowledge-infrastructure-committee

[23] See https://www.clarin.eu/k-centre-catalogue for the CLARIN K-centre catalogue.

harvesting of metadata records from resource providers (Van Uytvanck et al., 2012, pp. 1029–1031). These resource providers are usually but not exclusively the repositories of CLARIN B- and C- centres. By contrast, the CRF overviews employ no such automated harvesting – the listings of LRTs in the CRF overviews are wholly manually created and then also curated by hand. While there are fewer than 2000 LRTs listed in the CRF overviews, there are almost 930,000 records in the VLO as of 1 April 2026. Further, the majority of the fewer-than-2000 LRTs listed in the CRF also constitutes a subset of the approximately 930,000 VLO records, which means that what can be found in the CRF overviews can in almost all cases be found in the VLO as well (see also Section 2.3 about how this overlap dovetails with the inclusion criteria).

The CRF overviews might thus seem redundant in view of VLO's considerably larger offer. The existence of the CRF overviews, however, may be justified precisely on the grounds of manual curation, which remains important to this day for working with heterogenous collections of big data (Yoon et al., 2025). In the case of CLARIN, manual curation is valuable in relation to the aforementioned taxonomic division into categories such as 'parliamentary corpora,' 'corpora of academic texts,' and so on, which are not yet included in the VLO (but see Section 4.2). What is important is that such taxonomic categories are defined and demarcated on grounds that are inherently fuzzy and informal, and thus require a large degree of qualitative consideration. For instance, the Spoken corpora family does not correspond to just all corpora of audio recordings,[24] but on the CRF definition refers to those corpora that are primarily used by researchers for the (not necessarily computational) study of the linguistic or discursive features of spoken language production (Adolphs & Knight, 2010). What is excluded in this way is the fairly large set of smaller datasets of speech recordings that are primarily intended to be training models for developing speech technologies; such exclusion is in line with the fact that the CRF initiative is primarily aimed at researchers working in the (digital) humanities and social sciences rather than computer scientists and resource developers.

4.2 Including CRF taxonomies in the VLO

The VLO is strictly agnostic in terms of the metadata values that it harvests. The non-existence of the CRF categories in the VLO has thus far reflected the fact that the service providers (i.e., the B- and C- centres) do not specify that a given corpus or tool belongs to this or that taxonomic category in a principled fashion; for instance, as a metadatum that would end up as a dedicated CMDI component when harvested by the VLO (Windhouwer & Goosen, 2022). They likely do not do so because of the aforementioned fuzziness of the categories, and the related lack of consensus on what constitutes a particular family.

In spite of this, such taxonomic divisions remain important for directing researchers to quality datasets. Because the VLO also comes with the goal of high recall (anything relevant within the CLARIN ecosystem should be findable through it), it is the intention to let the VLO index every item that is featured in the CRF overviews. This might come at the cost of duplication, as certain resources will be shown twice in the VLO, once based on the original repository's metadata and once based on on the CRF entry; however, this is deemed better than not being able to find entries (or at least not easily) that are included in a CRF overview. Additionally, the VLO will be able to group such duplicated entries together and therefore make it clear to the user that these are different representations of the same dataset.

4.3 CRF supporting VLO findability

The manual CRF listings contribute quite considerably to visibility. In a 2018 survey of VLO findability, Fišer and Lenardič (2018) report that by using simple queries such as *parliament* corpus*, only about 75% of the corpora that were at the time included as part of the *Parliamentary corpora* CRF could straightforwardly be found in the VLO. The difficult findability of the rest generally had to do with the fact that the relevant corpora were ranked rather low in the VLO results, sometimes even lower than non–parliamentary corpora, for reasons related to the technical ranking algorithm employed by the VLO. An advantage of the CRF overviews here is that they focus on a consistent level of LRT granularity (e.g., CRF overviews list just one version of a single corpus rather than all the versions or possible subsets, in

[24] https://www.clarin.eu/resource-families/spoken-corpora

contrast to the VLO), and from 2025 onwards, only list available LRTs (see Section 2.3).

Such difficulties in VLO recall persist to this day. For instance, with the same search string as above – that is, *parliament* corpus* –, the Czech corpus *ParCzech 4.0* (Kopp, 2024) ends up only on only page 10 of the search results, and crucially gets outranked by corpora such as the Slovenian *metaFida 1.0* (Erjavec, 2023) corpus and the *Balanced Corpus of Modern Latvian (LVK2013)* (Levāne-Petrova et al., 2013). The latter are not parliamentary corpora proper but rather reference/balanced corpora that contain parliamentary proceedings as a relatively small subset. The CRF initiative thus boosts the findability of a corpus like *ParCzech 4.0* (Kopp, 2024) by including it in a rather small but visible online overview that consists of parliamentary corpora exclusively.

5 Conclusion

We have first presented the new backend of the CLARIN Resource Families (CRF) initiative, which corresponds to a Python-based workflow that takes JSON files as inputs. This backend has been developed to help open up the CRF initiative to external collaborators. We have then at some length discussed the qualitative guidelines for describing the language resources and tools constituting the CRF overviews, while highlighting certain non-trivial issues that pertain – also in a general sense – to the documentation of descriptive metadata like language and annotation. We have then moved on to discussing how potential collaborators can become involved in the initiative and the ongoing process of facilitating curatorial help via CLARIN K-centres, which have taken over management of some of the families. Finally, we have discussed how CRF relates to CLARIN ERIC's flagship resource discovery service – that is, the Virtual Language Observatory –, from both the technological and conceptual perspectives.

Work in the near future will mostly focus on streamlining the collaboration workflow, especially in relation to the K-centres. For K-centre representatives, possible training sessions for using the workflow will be organised as needed. If new overviews are to be added to the CRF (e.g., medical corpora), a gap analysis will first be performed in collaboration with the K-centres to identify options for potential expansions.

Acknowledgments

The work in this paper has been funded by the CLARIN Resource Families project for all authors, as well as by the Slovenian Research Agency programme *P6-0436: Digital Humanities: resources, tools and methods* (2022–2027) for authors 1 and 3.

References

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38–52). London: Routledge. https://doi.org/10.4324/9780367076399-3

Ahačič, K., Legan Ravnikar, A., Merše, M., Narat, J., & Novak, F. (2011). *Words of the 16th-century slovenian literary language*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1127

Bański, P., & Hedeland, H. (2022). Standards in CLARIN. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 307–340). De Gruyter. https://doi.org/10.1515/9783110767377-012

Bański, P., Heid, U., & Herzberg, L. (Eds.). (2025). *Harmonizing language data: Standards for linguistic resources*. De Gruyter. https://doi.org/10.1515/9783112208212

Davies, M. (2017). *Corpus of Historical American English – Kielipankki Korp version 2017H1*. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2017061925

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Erjavec, T. (2012). Multext-east: Morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, *46*(1), 131–142. https://doi.org/10.1007/s10579-011-9174-8

Erjavec, T. (2017). Multext-east. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 441–462). Springer. https://doi.org/10.1007/978-94-024-0881-2_17

Erjavec, T. (2023). *Corpus of combined Slovenian corpora metaFida 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1775

Finatto, M., & Grilo, S. (2020). *CorPop: A corpus of popular Brazilian Portuguese*. PORTULAN. http://hdl.handle.net/21.11129/0000-000D-FE56-5

Fišer, D., & Lenardič, J. (2018). Parliamentary corpora in the CLARIN infrastructure. In M. Piasecki (Ed.), *Selected papers from the CLARIN annual conference 2017* (pp. 75–85). https://doi.org/10.3384/ecp147

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's key resource families. *Eleventh international conference on Language Resources and Evaluation [LREC 2018]*. https://aclanthology.org/L18-1210.pdf

Forkel, R., & Hammarström, H. (2022). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, *13*(6), 917–924. https://doi.org/10.3233/SW-212843

Goosen, T., & Eckart, T. (2014). Virtual language observatory 3.0: What's new. *CLARIN annual conference 2014*. Retrieved January 12, 2026, from http://www.clarin.eu/sites/default/files/cac2014_submission_2_0.pdf

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., & Štěpánková, B. (2024). *MorfFlex CZ 2.1 (2024-12-23)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). http://hdl.handle.net/11234/1-5833

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2026). *Glottolog/glottolog: Glottolog database 5.3* (Version v5.3). Zenodo. https://doi.org/10.5281/zenodo.18840935

Helgadóttir, S., & Barkarson, S. (2018). *The saga corpus (Fornritin)*. CLARIN-IS. http://hdl.handle.net/20.500.12537/32

Kopp, M. (2024). *ParCzech 4.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). http://hdl.handle.net/11234/1-5360

Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., Kosem, I., Robnik-Šikonja, M., Ljubešić, N., Dobrovoljc, K., Laskowski, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M., & Logar, N. (2019). *Corpus of written standard Slovene Gigafida 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320

Lenardič, J. (2023). Documenting corpus annotation in CMDI: State of affairs. In K. Lindén, J. Niemi, & T. Kontino (Eds.), *CLARIN annual conference proceedings* (pp. 48–52). Retrieved January 23, 2026, from https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf#page=56

Lenardič, J., & Fišer, D. (2022a). CLARIN depositing guidelines: State of affairs and proposals for improvement. *CLARIN Annual Conference Proceedings*, 48–52. https://office.clarin.eu/v/CE-2022-2118-CLARIN2022_ConferenceProceedings.pdf#page=54

Lenardič, J., & Fišer, D. (2022b). The CLARIN resource and tool families. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 343–372). De Gruyter. https://doi.org/10.1515/9783110767377-013

Lenardič, J., & König, A. (2025). *Guidelines for preparing CLARIN resource families*. CLARIN Office Archive. https://doi.org/10.34733/doc-187

Levāne-Petrova, K., Pokratniece, K. ı., Vēvere, D., Poikāns, I., & Andronova, E. (2013). *Balanced corpus of modern Latvian (LVK2013)*. CLARIN-LV digital library at IMCS, University of Latvia. http://hdl.handle.net/20.500.12574/44

Ljubešić, N., & Erjavec, T. (2025). Part-of-speech tagging and related annotation. In P. Bański, U. Heid, & L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources* (pp. 61–88). De Gruyter. https://doi.org/10.1515/9783112208212-004

Mamede, N., Baptista, J., & Mamede, N. (2020). *ViPER verb lexical database*. PORTULAN. http://hdl.handle.net/21.11129/0000-000D-F91E-A

McGillivray, B. (2025). *LatinISE corpus (version 5)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). http://hdl.handle.net/11372/LRT-5870

Meden, K., Erjavec, T., & Pančur, A. (2025). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*, *59*(2), 891–911. https://doi.org/10.1007/s10579-024-09746-8

Nordhoff, S. (2012). Linked data for linguistic diversity research: Glottolog/Lnangdoc and ASJP online. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked data in linguistics: Representing and connecting language data and language metadata* (pp. 191–200). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28249-2_18

Odijk, J. (2019). Discovering software resources in CLARIN. *Selected Papers of the CLARIN Conference 2018*, *159*, 121–132. http://hdl.handle.net/1874/391802

Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990–2022) siParl 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1936

Romary, L. (2025). International standards for the identification and the description of languages and their varieties. In P. Bański, U. Heid, & L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources* (pp. 35–60). De Gruyter. https://doi.org/10.1515/9783112208212-003

Straka, M., & Straková, J. (2014). *NameTag*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E

Straková, J., & Straka, M. (2025, July). NameTag 3: A tool and a service for multilingual/multitagset NER. In P. Mishra, S. Muresan, & T. Yu (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 3: System demonstrations)* (pp. 31–39). Association for Computational Linguistics. https://aclanthology.org/2025.acl-demo.4/

Tadić, M. (2014). *Croatian national corpus*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL), Faculty of Mathematics; Physics, Charles University. http://hdl.handle.net/11372/LRT-233

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The virtual language observatory. *Eight international conference on Language Resources and Evaluation [LREC 2012]*, 1029–1034. https://doi.org/11858/00-001M-0000-000F-85EE-8

Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelleni, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, 900–903. https://doi.org/11858/00-001M-0000-0012-B734-9

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Windhouwer, M., & Goosen, T. (2022). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 191–222). De Gruyter. https://doi.org/10.1515/9783110767377-008

Yoon, A., Kim, J., & Donaldson, D. R. (2025). Big data curation framework: Curation actions and challenges. *Journal of Information Science*, *51*(1), 205–223. https://doi.org/10.1177/01655515221133528

Żak, P., & Skuczyńska, B. (2018). *OptaHopper: Phrase-level sentiment with opinion targets*. CLARIN-PL digital repository. http://hdl.handle.net/11321/584

Multilingual Word Rains: Text Visualisation Built on Cross-Language Word Similarities

Magnus Ahltorp
Nakajima Koen Research Institute
Stockholm, Sweden
magnus@nakajimakoen.org

Maria Skeppstedt
Department of Linguistics
Stockholm University, Sweden
maria.skeppstedt@ling.su.se

Abstract

In this study, the text visualisation technique Word Rain is extended with functionality for generating multilingual word rains. Using multilingual pre-trained word embeddings provided by the MUSE library, we show that cross-language studies on lexical frequency with Word Rain are possible in the same way as has previously been possible for monolingual corpora. We make the programming code for generating multilingual word rains available, and we also provide a web page where visualisations can be produced by simply uploading text files that the user wants to visualise. To illustrate the technique, a tentative study of common word groups and cross-language frequency similarities and differences is performed on debates from three parliaments included in the CLARIN ParlaMint corpus.

1 Introduction

The standard word cloud can provide a quick overview of a text or corpus, in the form of a graph that displays the most frequent words in a font size that corresponds to their frequency (Cao & Cui, 2016, pp. 57–59). However, the random (or sometimes alphabetical) word positioning of the standard word clouds makes them unpractical for a deeper exploration of frequent words, and even more so for a comparison of word frequencies between different word cloud graphs.

Despite its limitations, the standard word cloud is often used for analytical purposes, e.g., when studying and comparing texts within digital humanities (Hicke et al., 2022). We have therefore previously developed the Word Rain visualisation technique (Skeppstedt et al., 2024), within the CLARIN and Dariah infrastructures. The aim of the development was to create a visualisation technique that would retain the simplicity of the standard word cloud, while also making the visualisation suitable for exploring word frequencies and for comparing frequencies between texts. Similar to the classic word cloud, the Word Rain technique extracts the most frequent words from a text/corpus, and displays them with a font size corresponding to their frequency. In contrast to a word cloud, the Word Rain technique automatically orders the words along the x-axis in correspondence to their meaning. This causes words with a similar meaning to be positioned horizontally close to each other in the word rain graph.

The semantically motivated word positioning of the Word Rain technique helps the user to identify semantic categories among the frequent words, as well as to locate areas in the graph that might be relevant to explore more closely. The semantic positioning also simplifies comparison between two texts, since several word rain graphs can be generated using the same semantic word positioning along the horizontal axis. Word rains can either be generated by employing the Python programming code in the Word Rain code repository[1], or by uploading texts to the Word Rain Web Service at wordrain.org (Ahltorp & Skeppstedt, 2025). When generating word rains using the programming code, there is a wider range of configuration possibilities available, such as the possibility to change colour scheme or to employ user-defined stop word lists[2]. The Word Rain Web Service, on the other hand, does not require the effort of downloading (and knowledge of how to use) Python code.

[1] https://github.com/CDHUppsala/word-rain

[2] We refer to Skeppstedt et al. (2025) for a detailed description of the customizability of the Word Rain programming code.

Magnus Ahltorp, and Maria Skeppstedt 2025. Multilingual Word Rains: Text Visualisation Built on Cross-Language Word Similarities. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 100–108. DOI: https://doi.org/10.3384/ecp222

We have, previously, aimed at providing solutions to use cases that include the exploration of monolingual corpora. However, for many text analysis tasks, it is also relevant to be able to visualise, study and compare the content of corpora containing texts written in several different languages. The main aim here is therefore to provide and describe our resources for making it possible to create multilingual word rains that can be used in such studies. More specifically, we have extended the Word Rain Web Service, to make it possible to upload corpora containing texts in several languages, and also provided the Web Service with a few pre-defined language options for such a multilingual text exploration and comparison. In addition, we are making a new version of the Python code for generating word rains available. It is now able to handle multilingual corpora for all combinations of languages, as long as aligned pre-trained language models are available for the language combinations. Finally, we have applied the updated code in a proof-of-concept study, where the Word Rain technique is used for visualising lexical similarities and differences in the CLARIN ParlaMint corpora from the United Kingdom, France and Austria.

2 Background

Visualisation of corpora across languages can be approached in several different ways. One method is to create visualisations and conduct the analysis for texts in each language separately and then compare the analyses. This approach was, for instance, taken for analysing Hungarian and Slovenian parliamentary corpora (Sárossy & Ligeti-Nagy, 2024). The authors divided each corpus into two subcorpora (coalition MPs and opposition MPs), and then compared frequent keywords, as well as graphs with collocations containing these keywords. A similar approach was taken by Söderfeldt et al. (2025), who created, analysed and compared four separate topic modelling-based visualisations of patient organisation publications written in English, French, German and Swedish.

When it comes to approaches for including texts written in different languages within the same visualisation, there are a number of solutions for how to visualise and compare an original text with its translation into another language (Yousef & Janicke, 2021). However, for creating visualisations of text collections containing non-parallel texts in several different languages, less work has been carried out.

A possible method for handling visualisations of multilingual text collections is to first automatically translate all texts into a single language, before creating a visualisation. However, such an approach would risk hiding differences between languages, i.e., differences that might be one of the aims of the analysis to find. For instance, there might be closely related concepts where each one of them is described by a separate term in one of the languages, whereas a translation into a language with less linguistic specificity for these concepts could hide the variation in the original language. We, therefore, aimed for a solution to a multilingual Word Rain visualisation, where the words would be retained in their original language, but where the word rain graphs would still be easily comparable across languages.

The semantic word positioning along the x-axis in a word rain graph is based on word embeddings. Multi-dimensional word embeddings, which have been trained on word co-occurrence patterns in a monolingual text corpus, are projected down to one dimension, to determine a word's x position in the graph. Thereby, words that occur in similar contexts in the training corpora are positioned close to each other on the x-axis in a word rain graph.

A commonly used method for handling multilinguality when working with word embedding models is to use aligned models. Then, one embedding model per language is still used, but the models are aligned with each other, so that similar words across languages have similar embeddings. For instance, in two standard word embedding models for German and French, the words *Präsident* in German and *président* in French (both meaning 'president') have word embeddings that are totally unrelated. In contrast, if the two embedding models are properly aligned, they have similar embeddings. Therefore, if using the embedding in the German model for the word *Präsident* when querying for similar words in the French model, the word *président* should be among the retrieved ones.

Word embedding models can be aligned between languages in a supervised fashion, by using content-aligned multilingual text corpora (such as Wikipedia) (Søgaard et al., 2015) or by using multilingual dictionaries (Lample et al., 2018) as data for training. It is also possible to achieve aligned word embeddings in a fully unsupervised fashion (Lample et al., 2018).

3 Method

Most of the updates required for handling multilinguality was carried out in the code for generating word rains. Some updates were also needed for the Word Rain Web Service code repository.

3.1 Updating the functionality for creating word rain visualisations

The previous implementation of the Word Rain visualisation technique needed to be extended in two main ways to handle multiple languages. First, we added the ability to use different stop word lists for each sub-corpus, instead of just using one stop word list for the entire corpus. Second, we added the ability to use different word embedding models for each sub-corpus. These changes enable us to use embedding models that match the language of each sub-corpus included in the visualisation. For instance, for the proof-of-concept study described below, we could use English embeddings for the UK data, French embeddings for the French data, and German embeddings for the Austrian data[3].

In addition, we added a name format requirement. For generating word rains, the user needs to provide a file path to a main folder, containing the corpus that is to be visualised. This main corpus-folder should contain subfolders, which, in turn, contain the actual texts. The subfolders represent a sub-division of the main corpus into sub-corpora, and one word rain graph is generated for each such sub-corpus. For instance, when we previously have generated word rains for analysing changes over time in a corpus, we have created subfolders based on the publication year of the texts, and thus generated one word rain graph per year (Skeppstedt et al., 2025). As a means to indicate the language of the texts in a subfolder, we added the requirement that each such subfolder must be prefixed by a language code, followed by a hyphen. Name of subfolders could, thus, for instance be `en-house_of_lords` and `en-house_of_commons`.

3.2 Updating the Word Rain Web Service

A number of updates were also required for the Word Rain Web Service, in addition to connecting it to the new version of the code for generating word rains.

As described above, to create graphs with a word positioning aligned between languages, we need word embeddings that are themselves aligned. We therefore added aligned word embeddings provided by the MUSE library (Lample et al., 2018). These pre-trained embeddings have been aligned in a supervised fashion, using multilingual dictionaries. As a first step, we added three aligned embedding models to the web service, for French, English, and German.

We created a separate user interface for multilingual aligned word rains, and made it available at: https://wordrain.org/multilingual (see Figure 1). As for the monolingual word rains, the user selects texts to visualise by uploading one or several texts to the web page, (i.e., not folders, as when using the code directly). In contrast to the monolingual service, there is, however, no drop-down list for selecting language. Instead, all uploaded files must be prefixed with a language code matching one of the languages of the aligned embedding models, followed by a hyphen. That is, currently with either `de-`, `en-` or `fr-`.

Similar to the monolingual service, the user can either upload plain text files, or OOXML (.docx) files. After the text files have been uploaded and processed, the user can adjust two parameters in the graphical presentation: i) the height ratio between the part of the graph containing the words and the part containing the vertical bars (see section 4.1), and ii) the rate with which the font size should decrease with decreased word frequency. (More specifically, the decrease rate is governed by a power function, for which the exponent is configured by the user.) Finally, a word rain graph is generated for each uploaded file, with an aligned semantic axis between the graphs.

4 Proof-of-concept study

To illustrate the functionality of the multilingual Word Rain technique we performed a proof-of-concept study.

[3]The extended code can be found at https://codeberg.org/ahltorp/wordrain/src/branch/multilingual

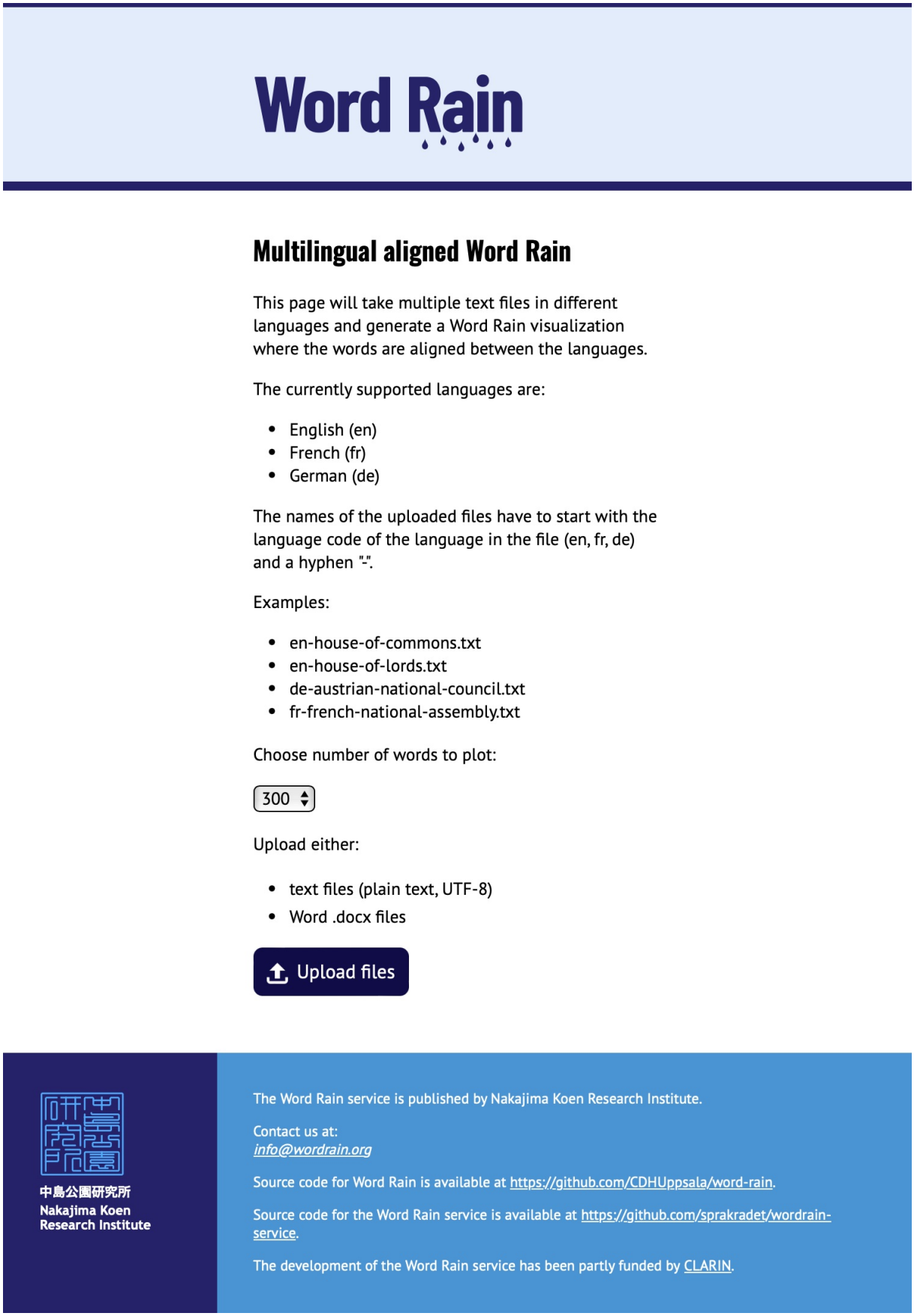

Figure 1: The multilingual aligned Word Rain page at wordrain.org/multilingual.

For this study, we used the ParlaMint corpus. ParlaMint is a corpus collection of debates from 29 parliaments across Europe. We used the ParlaMint (Erjavec et al., 2024) corpora ParlaMint-GB, ParlaMint-FR and ParlaMint-AT from the resource repository CLARIN.SI. We extracted the data for the 2017 parliamentary debates in .txt format. For 2017, ParlaMint-AT had 1 699 623 tokens, ParlaMint-FR had 4 680 332 tokens, and ParlaMint-GB had 14 573 789 tokens when processed through Word Rain's default tokeniser. We treated all texts from each of the languages as one separate sub-corpus, and thus generated one word rain graph for each language.

The graphs were generated using the code in the Word Rain Python code repository. We configured the graph generation to transform all words into lower case before computing word frequencies. The number of words per word rain was set to the 500 most frequent words, and we used the standard NLTK stop word lists for each of the three languages. The functionality to automatically identify frequent n-grams was not enabled.

Font size (and thereby also bar height, see below) was configured to be normalised between graphs so that the font size for the most frequent word was the same in all three graphs. Frequencies, as indicated by font size/bar height, are therefore comparable within graphs but not across graphs, while the frequency relations between words are comparable across graphs. As mentioned above, the font size/bar height decreases with decreased word frequency at a rate governed by a configurable power function. Here, it was configured to decrease linearly.

4.1 Resulting Word Rain graphs

The resulting word rains can be seen in Figure 2. In the same way as in a traditional word cloud, the words have a larger font size if they are more frequent. Since the words are semantically ordered along the horizontal axis according to their positions in the word embedding models, words with similar meaning are placed together horizontally with a high probability. With the modifications carried out in this study, this is now also the case for semantically related words in different languages.

The bar emanating from the faint circle in the upper left corner of each word forms an additional indication of relative word frequency, as the height of the non-transparent part of a bar directly corresponds to the font size of the word. However, since it might be difficult to determine exactly which bar corresponds to which word in areas with many words, the main purpose of the bars is to provide an overview of the frequencies present in a given area, rather than to indicate the frequency of each individual word. Finally, the colours illustrate the horizontal position of the words, and are chosen to make it easier to compare horizontal word positions between graphs.

As an example of similar words being placed closely, the words *Präsident* 'president' in German and *président* in French are very close to each other on the horizontal axis (to the left, in orange). The same is true for *wissen* in German, *savoir* in French and *know* in English (to the right, in magenta). This indicates that the word embeddings are indeed fairly aligned.

When exploring and comparing the three word rain graphs, we can first turn our attention to the green part, where there is an area with a high frequency in all graphs around German *Herr*, French *monsieur* and English *hon*. The words in that area do not mean the same thing, but if we study the common words more closely, we can see a pattern. In Austria, they use several forms of *Herr* 'Mr', *Frau* 'Ms', *Kollege* 'colleague' and *geehrte* 'honourable'. In France, they use *monsieur* 'Mr', *Mme* 'Ms', *madame* 'Ms' and *collègue* 'colleague'. In the UK we find *hon*, *noble*, *lord*, *lady*, *gentleman*, *Mr*, *baroness*, *colleagues* and *friend*. These are ways of addressing other members of parliament and other persons, for example government ministers, and we can see the influences of the House of Lords in the UK, where the titles baroness and lord are used.

In the orange part of the graphs, there is another cluster of words denoting persons. These are not persons being addressed, but instead roles in the parliament, such as German *Präsident*/*Präsidentin* 'President, masculine and feminine form', *Abgeordneter* 'Member of Parliament', French *ministre* 'minister', *rapporteur* 'manager of a bill', and English *minister*, *secretary*, *chancellor*.

In the part of the graph where green turns into yellow, there is yet a cluster denoting persons. However, this cluster seems to denote persons in general (not specific to the parliament). For all three languages,

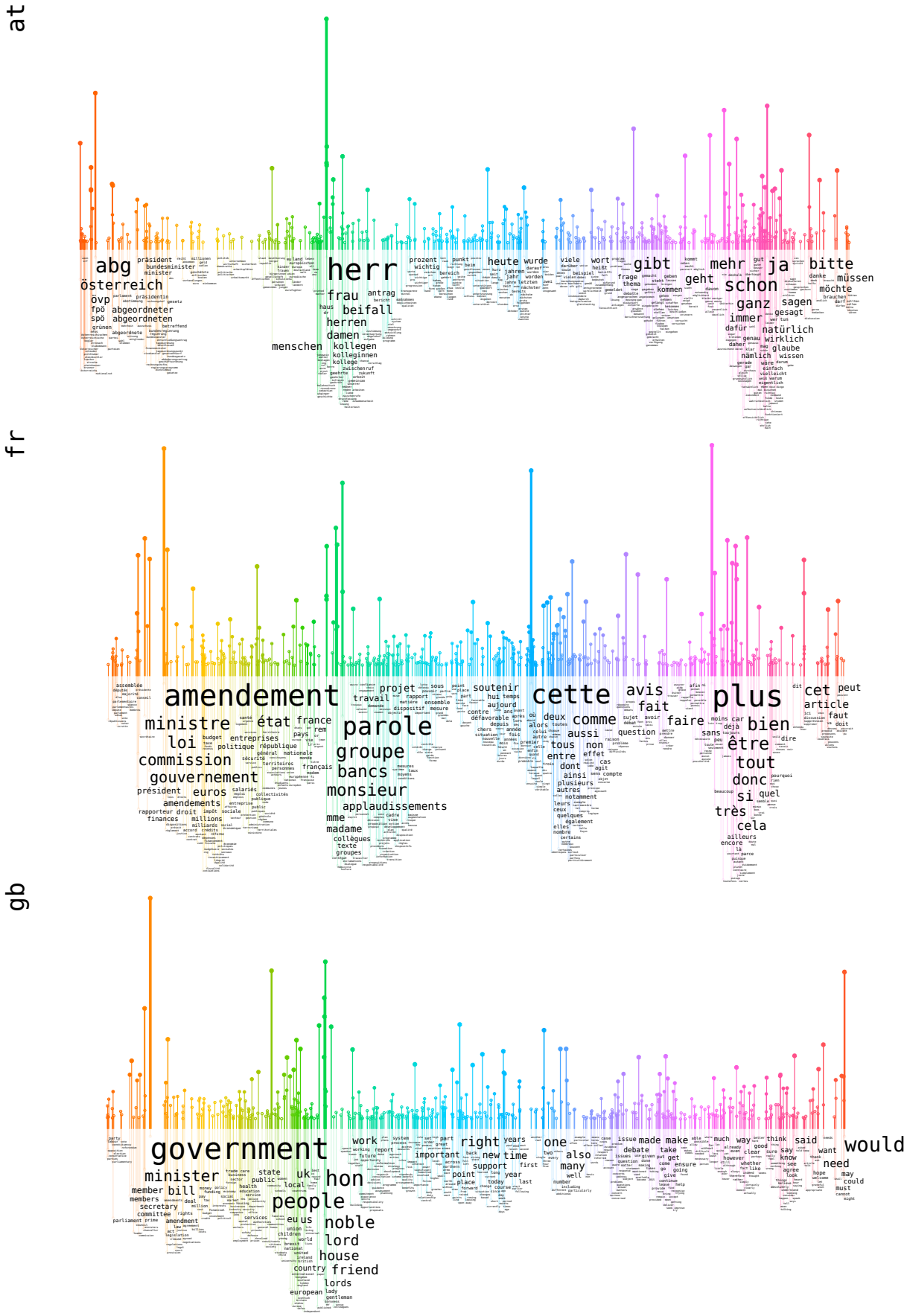

Figure 2: Word rains comparing 2017 parlamentary debates from Austria (at), France (fr) and United Kingdom (gb).

there is a frequently occurring word meaning 'persons in general', i.e. *Menschen*, *personnes* and *people*. These frequent words are then accompanied with closely positioned, less frequent ones, that denote more specific categories of persons. Some of these categories occur among the top 500 most frequent words for all three languages, such as words for 'children' (*Kinder*, *enfants*, *children*) and 'young' (*jungen*, *jeunes*, *young*). Other examples of words in this cluster include: German *Bürger/Bürgerinnen* 'citizen', *Bevölkerung* 'population/people', *Frauen* 'women', French *concitoyens* 'fellow-citizen', *étudiants* 'students', and English *citizens*, *families*, *students*, and *women*.

All three clusters denoting persons provide examples of the advantages of keeping the original languages instead of using an automatic translations. Examples of nuances that might have been lost in the translation are the feminine and masculine forms of words, the exact English titles, as well as the exact terms used for the professional roles within the three parliaments.

We have so far focused on similarities in the form of semantic categories that are frequent in all three graphs, (and on differences within these semantic categories). There are, however, also differences between the three corpora, when it comes to which words and semantic categories that are frequent. For instance, the very high frequency French word *parole* 'speech' is less frequent in the other languages. The word has corresponding words on the same position on the horizontal axis in English (*speech*) and in German (*reden/Reden* 'speak, speeches'), but they are not nearly as frequent.

We can also observe a larger difference, if we instead look at the far left of the word rain graphs, where there is a high frequency area (in orange) that is only present in the Austrian data, with one of the most prominent words being *Österreich* 'Austria'. This word, and the various forms of it, is – perhaps naturally – not among the 500 most frequent words in either France or the UK. There are also names of political parties in Austria: *ÖVP*, *FPÖ*, *SPÖ*, *Grünen* and *NEOS*.

A difference where the words are present in the other languages but with very different frequencies is *cette* 'this' in French and *would* in English. The first is probably explained by it not being in the stop word list for French, even though it is the feminine form of the masculine *ce*, which is present in the stop word list. The second can perhaps be a real difference in language use, since constructions with *would* like *I would disagree* and *I would urge* seem to be very common, and maybe even used as something close to a discourse particle rather than having its usual syntactic function, but further studies would be needed to draw any solid conclusions.

As a last example, *amendement* in French has its counterpart in English *amendment* at the same horizontal position, but is used much more frequently in French. This is probably because *amendement* is a very important term in French lawmaking.

4.2 Discussion

A general problem when comparing the most frequent words between texts, is that more specific terms might be used for related concepts in one text, while another text might use a single term for all these related concepts. An incorrect conclusion might then be drawn that this concept is an important one in the second text (as it appears among the most frequent terms), whereas it is not important in the first. As mentioned in the background, this difference in linguistic specificity is also present across languages. A term used in one language might correspond to several terms in another language even when the terminological domain is the same. When using a word rain instead of a standard frequency list, however, vertical clusters containing related concepts are created. Entire clusters of related words can thereby be compared between texts, instead of comparing frequencies for individual words. For instance, in a standard word frequency list, it would look like the use of formal titles to address people would be more frequent in the German texts than in the English ones, since *Herr* 'Mr' is very common (and *Frau* 'Mrs' less, but still rather, common). However, when looking at the vertical title cluster in the English word rain (i.e., containing *hon*, *noble*, *lord*, *lady* ...), it is clear that titles are very common also in the English texts, but that there is a larger diversity among the titles used.

The standard NLTK stop word lists used were not ideal for the task, and should be amended to include common words that are not relevant to the analysis, such as *cette* 'this'. One word of caution, however, is that some commonly used words, such as *would*, might be relevant for some types of analysis. The

resulting word rain could probably also have been improved by using specifically Austrian stop word lists and word embedding models instead of the generic German used in this study.

The pre-existing aligned word embedding models used for this study make no distinction between upper and lower case characters (Lample et al., 2018), forcing us to transform all words into lower case. This causes pairs of words having intial upper case (proper nouns) or lower case (other words) to be conflated in all of the three languages. This is, however, especially problematic for German, where all nouns use initial upper case in the standard orthography. An example of two conflated German words that appear in our visualisation is the verb *reden* 'speak' and the noun *Reden* 'speeches' (both forms have frequency > 20 in the corpus).

5 Conclusions and future work

We have been able to successfully extend Word Rain to produce multilingual word rains, to support comparison of word frequencies in corpora containing texts in different languages. For this functionality, we have also made programming code, as well as the Word Rain Web Service, available, which makes it possible for everyone to generate multilingual word rains. That is, we have provided a multilingual text visualisation tool both for those who need the flexibility of the programming code as well as for those who want to be able to upload texts to produce visualisations with pre-configured settings. Finally, we have used the CLARIN ParlaMint resource for a proof-of-concept study that tentatively investigates what words are commonly used in the parliaments of different countries based not only on raw frequency lists, but grouped according to cross-language semantic information.

In the proof-of-concept study, we only used three of the 29 corpora in ParlaMint, partly due to limited time and space, but also because it was necessary for the authors to understand the languages to be able to validate the quality of the resulting alignments. The study has also been limited by not having co-authors that are experts in parliamentary debates and political science. We therefore hope future studies using this technique can be made together with authors from various parts of Europe and from different scientific fields. We also plan to apply the Word Rain technique on multilingual corpora from other domains, and in collaboration with domain experts analyse differences between texts written in different languages.

Finally, for the monolingual use cases of the Word Rain technique, many word rains have been produced using a word2vec model trained on the corpus that is visualised, instead of using pre-trained word embedding models. The words are then not positioned along the x-axis based on distributional similarities in a large general corpus, but instead based on distributional similarities in the specific corpus visualised. Custom word embeddings would also enable other types of equivalence criteria for words, such as lemmatisation or distinguishing upper-case-initial words from their lower-case-initial counterparts. Future studies on custom word embedding models in multilingual Word Rain visualisations should also include quality evaluations of different kinds of aligned models.

While there are easy-to-use libraries for training standard word embedding models on a specific corpus, e.g. the Gensim library (Řehůřek & Sojka, 2010), we are not aware of any programming library that is equally easy to use for aligning word embedding models trained on different languages. To simplify the alignment process – and thereby also simplify the generation of multilingual word rains where the word positions reflect the specific distributional similarities in the corpora visualised – is therefore yet a project for future work.

Acknowledgments

Part of the work presented here was conducted within Swe-CLARIN, which is funded by the Swedish Research Council (dnr 2023-00161).

References

Ahltorp, M., & Skeppstedt, M. (2025). Word Rain as a Service: Making semantically structured word clouds available to everyone. In V. Vandeghinste & T. Kontino (Eds.), *Selected papers from the CLARIN Annual Conference 2024*. https://doi.org/https://doi.org/10.3384/ecp216.03

Cao, N., & Cui, W. (2016). *Introduction to text visualization* (Vol. 1). Atlantis Press. https://doi.org/10.2991/978-94-6239-186-4

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., … Fišer, D. (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1912

Hicke, R. M. M., Goenka, M., & Alexander, E. (2022). Word clouds in the wild. *Proceedings of the IEEE Workshop on Visualization for the Digital Humanities*, 43–48. https://doi.org/10.1109/VIS4DH57440.2022.00015

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. *International Conference on Learning Representations*. https://openreview.net/forum?id=H196sainb

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Sárossy, B., & Ligeti-Nagy, N. (2024). How to talk the talk? A comparative overview of keyword usage in Hungarian and Slovenian parliamentary corpora. In V. Vandeghinste & T. Kontino. (Eds.), *Clarin annual conference proceedings 2024*.

Skeppstedt, M., Ahltorp, M., Kucher, K., Aangenendt, G., Lindström, M., & Söderfeldt, Y. (2025). The word rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In G. Bouma, D. Dannélls, D. Kokkinakis, & E. Volodina (Eds.), *Huminfra handbook: Empowering digital and experimental humanities*. University of Tartu Library. https://doi.org/https://doi.org/10.58009/aere-perennius0175

Skeppstedt, M., Ahltorp, M., Kucher, K., & Lindström, M. (2024). From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, *23*(3), 217–238. https://doi.org/10.1177/14738716241236188

Söderfeldt, Y., Burchell, A., Reed, J., & Skeppstedt, M. (2025). Topic timelines for enabling close and distant reading of discursive shifts. a pilot case using periodicals of european diabetes organizations. *Journal of Open Humanities Data*. https://doi.org/10.5334/johd.286

Søgaard, A., Agić, Ž., Martínez Alonso, H., Plank, B., Bohnet, B., & Johannsen, A. (2015, July). Inverted indexing for cross-lingual NLP. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1713–1722). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1165

Yousef, T., & Janicke, S. (2021). A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1149–1159. https://doi.org/10.1109/TVCG.2020.3028975

A Trilingual Parallel Corpus of Yiddish, French, and Russian Literary Texts

Valentina Fedchenko
INALCO
valentina.fedchenko@inalco.fr

Assaf Urieli
Joliciel Informatique
assaf@joli-ciel.com

Arnaud Bikard
INALCO
arnaud.bikard@inalco.fr

Abstract

This article presents a trilingual Yiddish–French–Russian parallel corpus of literary translations, which will be made available through the CLARIN B-centre in France, ORTOLANG. Parallel corpora are fundamental resources for translation studies, contrastive linguistics, and natural language processing. Despite advances in neural machine translation, such corpora remain essential for improving automatic translation tools, particularly for less-resourced languages such as Yiddish. This article describes the phenomenon of translation in Yiddish and the construction of the corpus, including text alignment with the integration of a custom bilingual dictionary. The corpus is integrated into a multilingual concordancer, facilitating data exploration and analysis.

1 Introduction

In this article, we present an aligned trilingual Yiddish-French-Russian corpus of literary translations,[1] created and being developed within the framework of the project "Translation as a strategic issue for the survival of post-vernacular Jewish languages" (LJTRAD) coordinated by Arnaud Bikard at National Institute for Oriental Languages and Civilizations (INALCO)[2]. We have initiated the process of integrating the corpus into the French repository ORTOLANG (CLARIN B-centre)[3]. The corpus will be listed among the corpus resources and made available through the ORTOLANG platform. This integration will ensure visibility through CLARIN's extensive network of repositories, as the project aligns closely with CLARIN's mission to advance language resource accessibility and multilingual research infrastructure. The corpus particularly enhances CLARIN's efforts to preserve and digitize resources for less-digitized languages like Yiddish, while creating new opportunities for comparative literary analysis and translation studies.

However, our contribution goes beyond a purely technical corpus presentation. The resource is situated within a broader historical and sociolinguistic framework. The corpus does not merely align texts across languages; it documents a specific cultural process in which translation functions as a mode of survival for post-vernacular literatures.

The article is structured as follows. First, we introduce the object of the corpus: the phenomenon of translation in Yiddish and, more broadly, in Jewish languages, emphasizing its specificity, historical significance, and practices. We then examine the role and usefulness of aligned parallel corpora given existing resources for Yiddish, and in the context of recent advances in language models and machine translation. Then, we describe the methodology used for corpus construction, followed by a brief description of corpus exploration tools we provide. We also address the specific challenges associated with aligning three languages with distinct syntactic structures and writing systems. Finally, we provide a presentation of the corpus content, including both quantitative and qualitative analyses that highlight the diversity of literary genres and historical periods represented in the current version of the resource.

[1] The corpus is hosted on the Joliciel company domain: https://dev-ljtrad.joli-ciel.com/corpus

[2] https://anr.fr/Projet-ANR-18-CE27-0014

[3] https://www.ortolang.fr/

Valentina Fedchenko, Assaf Urieli, and Arnaud Bikard 2025. A Trilingual Parallel Corpus of Yiddish, French, and Russian Literary. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 109–119. DOI: https://doi.org/10.3384/ecp222

2 Phenomenon of translation in post-vernacular languages

Historically, Jewish languages have been minor, diasporic languages, lacking territory and national organisation. Their cultural heritage has been built up in a position of relative weakness, which has prevented the stable and lasting establishment of institutions that, in a state configuration, play the role of both legislator and conservator of cultural assets (Fishman, 1985; Lowenstein, 2000; Spolsky, 2014; Weinreich, 2008). Despite the many social movements and organisations that have attempted to halt the gradual disappearance of speakers (due to migration, wars, and assimilation), the decline of Jewish languages has been so inexorable that, by the beginning of the 21st century, they have come to be described as post-vernacular (heritage) languages. This term reflects not only their residual presence as mother tongues, but also the emergence of a wide range of new practices and uses accommodating varying degrees of language proficiency (Kahn & Rubin, 2015; Shandler, 2006; Weinstock et al., 1997).

Yiddish is among the twenty or so Jewish languages referenced that experienced, in the 19th and especially the 20th century, the richest cultural development, seeing not only the emergence of an abundant written output ranging from simple newspaper articles to the most ambitious literary works, but also the recording of oral and folk traditions through dedicated research, thus prolonging their fertility (Avishur, 1979; Borovaya, 2011; Hary, 1992; Romero, 1992; Stillman & Stillman, 1999; Zafrani, 1980).

The translation of Jewish languages remains relatively under-theorized, despite the long history and cultural importance of such practices. Yet the specific linguistic and sociocultural configurations of these languages suggest that translation from them raises issues that deserve to be considered in their own right. While this perspective builds on major contributions in translation studies — from foundational reflections by Walter Benjamin and Jacques Derrida to analytical frameworks developed by Antoine Berman, Jean-René Ladmiral, and the polysystem theory of the Tel Aviv school — it also calls for closer attention to the particular constraints and asymmetries involved in translating texts written in Jewish languages (Berman, 2008; Ladmiral, 2014).

A key aspect lies in the internally composite nature of these languages. All languages are heterogeneous to some degree, but Jewish languages are marked by a historically salient form of linguistic fusion. Their lexicon draws on several language families, and speakers — especially writers — often make deliberate use of this internal stratification. Hebrew-Aramaic elements, for instance, may be preferred over other lexical resources for stylistic, pragmatic, or symbolic reasons, even when near-equivalent alternatives exist. Such choices can index register, cultural affiliation, intertextual resonance, or degrees of in-group address. This kind of marked internal multilingualism has no straightforward counterpart in most target languages and is rooted in the specific historical experience of diaspora and the strong embedding of religious and cultural reference in everyday language use. As a result, translation frequently encounters layers of meaning that are linguistically and culturally dense, placing Jewish-language texts at the heart of broader debates on the "untranslatable."

Another difficulty concerns the historical transmission of these texts through translation. Existing translations are unevenly documented and often dispersed across journals, small editions, or archival publications. Moreover, many translations maintain a flexible relationship to the source text: the exact source version may not be specified, and translators sometimes introduce adaptations, omissions, or additions. The material conditions and intellectual contexts in which translations were produced are frequently only partially documented. This situation makes the study of translations inseparable from questions of textual history, mediation, and cultural positioning.

Translation from Jewish languages was very early on considered crucial to their recognition as languages of culture. At the same time, however, it has always been seen as an eminently problematic task, due to the very nature of these languages, whose expressive orality and hybrid composition pose particular challenges for translators (Norich, 2014). Added to this linguistic obstacle is the existence of a cultural barrier linked to the minor status attributed to these languages, which has direct consequences on translation practices. Even today, there is no hesitation in translating a Yiddish novel from an English or Russian version, which would be inconceivable for a work composed in a "major" language.

Historically, translation practices from Jewish languages have been shaped by asymmetries of prestige and power. Early translation initiatives were often isolated efforts marked by ideological conde-

scension, presenting the works as curiosities from marginal "jargons." Later, more systematic projects emerged—frequently at moments when the languages themselves were already in decline—turning translation into a form of cultural preservation (de Graaf, 1996). In this context, translation ceases to be merely a bridge between two living linguistic communities and becomes an act of conservation and continuation, tied to questions of memory, legitimacy, and ethical responsibility (Bikard, Fedchenko, Furci, & Rousselet, 2025).

The presented resource is situated within this historical and sociolinguistic configuration. The corpus does not simply align texts across languages; it captures a specific cultural process in which translation functions as a mode of survival for post-vernacular literatures. Each aligned pair embodies multiple layers: the linguistic structure of the source text, the translator's interpretive choices, the norms of the target literary system, and the historical moment of mediation. Consequently, the corpus should be understood not only as a multilingual textual resource, but as a stratified record of cultural transfer, where linguistic, stylistic, and ideological signals coexist.

These characteristics frame the corpus not simply as a collection of aligned texts, but as a record of historically situated translation practices. At the same time, the linguistic complexity of Jewish languages and the variability of translation strategies highlight the importance of computational tools adapted to these languages. While some natural language processing resources exist[4] — particularly for Yiddish — their coverage and reliability remain limited, and alignment and analysis tasks must contend with orthographic variation, multilingual lexical layers, and non-standardized usage. In this sense, the corpus also contributes to the development of digital infrastructures capable of accommodating the structural and historical particularities of Jewish-language materials.

More broadly, this framework contributes to rethinking the role of digital corpora in contexts where languages are no longer sustained primarily through everyday vernacular use. For post-vernacular languages, corpora are not only descriptive tools, but also instruments of cultural continuity. By enabling circulation between original texts and translations, and by preserving the visibility of the source language alongside its mediations, such resources participate in shaping new forms of reading and engagement adapted to heritage and scholarly communities.

3 The role of parallel corpora with respect to existing Yiddish resources

Yiddish has a long lexicographical tradition that includes several bilingual dictionaries and an ambitious project to create a monolingual dictionary (Moskovich, 2010). However, existing bilingual dictionaries are generally lacking in contextual examples, limiting their usefulness for in-depth analyses or nuanced language learning. This is the gap which parallel corpora attempt to fill.

Parallel corpora occupy an important position in the fields of translation studies and contrastive linguistics. An aligned parallel corpus constitutes a large collection of texts in digital format, establishing precise correspondences between versions of the same content in two or more languages, generally aligned at the sentence or paragraph level (Doval & Nieto, 2019; Lefer, 2021; Mikhailov & Cooper, 2016).

Parallel corpora exist for several languages, many of which are listed on the CLARIN website[5] (De Jong et al., 2018). Within the typology of parallel corpora, our corpus falls into the type offering texts in the original language and their literary translations into the target language, with sentence-by-sentence alignment. It is therefore the first aligned multilingual corpus for Yiddish literature and its translations. Since literary texts are not translated word-for-word, the translations included in the parallel corpus should be understood as literary or interpretive translations, rather than exact equivalents. While this limits some types of linguistic analysis, it opens rich possibilities for comparative literary, stylistic, and cultural studies.[6] All alignments have been done at the excerpt level, on phrases or on clauses between commas, to reflect meaningful correspondences without implying strict literal equivalence.

[4]For example: Yiddish Drama Corpus - YiDraCor: https://dracor.org/yi/; The Online Treasury of Yiddish Poetry - OTYP: https://lider.yiddish.nu/

[5]https://www.clarin.eu/resource-families/parallel-corpora

[6]Several examples of this type of analysis, conducted using our corpus, are available at Bikard, Fedchenko, and Urieli (2025)

Compared to other less-resourced languages, Yiddish benefits from several important resources available online. The largest is the corpus of Yiddish literature at the Yiddish Book Center in Amherst. This exceptional resource offers a sophisticated search engine on the OCR content and various filtering options[7]. Its main value lies in the breadth of its collection, which provides researchers with access to a vast set of digitized texts representative of Yiddish literature. Another smaller resource developed by the Russian Academy of Sciences is the Corpus of Modern Yiddish (CMY)[8], as described by Birzer (2014). This corpus is distinguished by its sophisticated linguistic analysis capabilities, including part-of-speech tagging and morphosyntactic analysis. Both of these corpora were of particular use in the design and construction of our parallel corpus.

In the current context of considerable advances in language models and machine translation, aligned multilingual corpora remain a valuable resource. When manually aligned, they offer a quality that automatic models cannot guarantee, particularly for less-resourced languages such as Yiddish. Indeed, the only encoder model currently available for Yiddish is the multilingual model XLM-RoBERTa (Conneau et al., 2019), where Yiddish is one of 100 languages and represents far less than 1% of the source data. Although Kulick et al. (2023) have shown it is possible to fine-tune this model to improve Yiddish POS-tagging, considerable work is needed to provide a general-purpose high-quality Yiddish encoder model. We are not aware of any decoder models for Yiddish. We hope that our parallel corpus can thus be used as a reference for both the improvement and evaluation of automatic systems based on new Yiddish language models.

4 Corpus construction methodology and research tools

The development of our corpus required several steps: the creation of parallel data in Yiddish, French and Russian; the alignment of Yiddish-French and Yiddish-Russian texts; as well as the establishment of a platform integrating a search engine and concondancer interface. To compile the parallel data, we adopted a mixed strategy. We integrated openly accessible French and Russian translations into our corpus, and OCRed the Yiddish originals using OCR system, Jochre, developed by Urieli et al. (2025) for the Yiddish Book Center (Amherst). Before alignment, the Yiddish texts were manually proof-read, and then semi-automatically converted into a standardized orthography developed by the Yiddish Language Institute (YIVO) to ensure the proper functioning of the bilingual lexicon during alignment and to unify the entries within the corpus for information retrieval with the search engine.

Each text and translation were encoded in TEI format to enable their reuse in future scientific projects, such as possible research on machine translation.

The alignment of the Yiddish-French parallel texts was carried out using the Hunalign tool (Varga et al., 2008). This bilingual alignment system is based on a hybrid approach combining statistical and lexical methods. It operates in three main stages: a pre-alignment based on segment length; a lexical alignment based on a bilingual dictionary; and an iterative refinement of the correspondences. Hunalign can work even in the absence of pre-existing lexical resources, by automatically generating a probabilistic dictionary from the aligned texts. However, better results are obtained with a wide-coverage lexicon. We thus integrated an electronic version of Niborski and Vaisbrot (2002), a Yiddish-to-French dictionary, which significantly improved the accuracy of the Yiddish-French alignment. The addition of this lexical resource compensated for some of the difficulties inherent in aligning a non-Latin script language, such as Yiddish, with a Romance language such as French, particularly with regard to morphosyntactic divergences and cultural specificities of the lexicon. For the Yiddish-Russian alignment, we built a Yiddish-Russian electronic dictionary by automatically translating the Yiddish-French dictionary entries into Russian using the Python package deep-translator.

To maximize alignment accuracy despite morphosyntactic divergences, both the original texts and the dictionary were stemmed prior to alignment using the Snowball stemmer[9], including the Yiddish stemmer contributed by Urieli in November 2020, and the French and Russian stemmers available in the

[7] https://ocr.yiddishbookcenter.org/

[8] http://web-corpora.net/YNC

[9] https://snowballstem.org/

original Snowball. Manual evaluation of a sample of alignments revealed a satisfactory accuracy rate, thus ensuring the reliability of the subsequent translation analyses planned in our study.

The alignment results were stored in a CSV file to simplify manual correction, including the original sentence with its TEI reference and the translated sentence with its TEI reference. In the case of one-to-many alignments, the source or target sentence is repeated.

Once the TEI files in all languages were ready, as well as the CSV alignment files, we uploaded the batch to our custom built search engine and concordancer. The search engine is based on Apache Lucene[10], and allows phrase and wildcard searches. The interface allows users to select the source and target languages, and indicate the number of sentences to display before and after the match (from 1 to 4). The user can also decide whether to match on stems (using the Snowball stemmers), or on exact terms without stemming.

The concordancer is accessible through an authentication system using the Keycloak identity provider. Authenticated users are mapped to internal roles (e.g., "registered user"), which determines their access rights. This system allows the platform to identify users and grant access to the concordancer in a controlled and restricted manner.

5 Corpus content

The construction of a corpus of translations for Yiddish was made possible by the fact that French and Russian possess sophisticated schools of translation from Yiddish. These traditions have existed since the beginning of the 20th century and were particularly enriched after the Second World War. After the Holocaust and the drastic decline in the number of Yiddish speakers, a cultural preservation movement developed in several countries, including France and the former Soviet Union, albeit in very different political contexts.

In France, the historical presence of a large Ashkenazi Jewish community, with a significant number of translators and activists, has stimulated interest in Yiddish culture among French intellectuals. This led to the creation of specialized publishing houses and collections dedicated to Yiddish literature, including bilingual Yiddish-French editions pursuing educational goals (Chinski, 2024).

Our parallel corpus only partially reflects this rich literary output for a less-resourced language, but we plan to expand it to cover it as comprehensively as possible. Currently, the corpus is unidirectional, meaning it only includes translations from Yiddish into French and Russian. The number of translations into Yiddish from French or Russian is also significant, and the inclusion of these translations in the corpus is planned for a later stage. Corpus statistics are shown in Table 1.

	Yiddish	**French**	**Russian**
Token count[11]	1070K	1024K	178K
Document count	19	19	4
Author count	15	15	3
Translator count	15	15	4
Time span of original texts	1873-1972	1873-1972	1911-1972
Time span of translations	1933-2022	1933-2022	1928-2010

Table 1: Corpus statistics for Yiddish, French and Russian

In our bibliographic research on translations of Yiddish literature into French, we relied on the list prepared by the librarians of the Medem Library[12], which includes 228 items (excluding anthologies). Our corpus currently covers only 8% of the entire dataset of French translations. The Yiddish-French corpus was designed from the outset with particular attention to chronological diversity: the Yiddish works cover a period of over a century, from Mendele Moykher-Sforim to Avrom Sutzkever (Fig. 1).

The corpus comprises mostly recent translations (more accessible to automatic processing), but also contains some older translations to ensure a more meaningful diachronic representation. The high num-

[10]https://lucene.apache.org/

[12]https://www.yiddishweb.com/traductions-du-yiddish-en-francais/

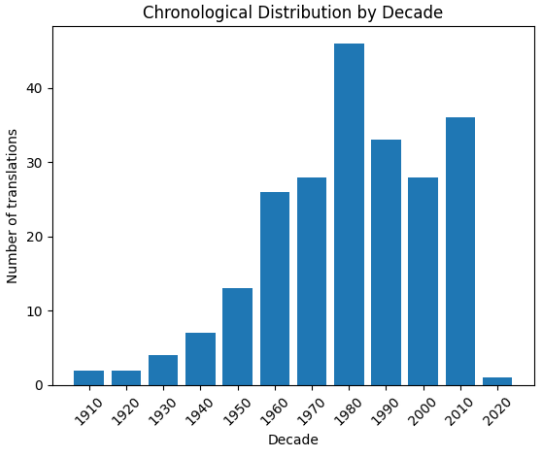

Figure 1: Chronological distribution of the works represented in the corpus.

ber of translators for this relatively modest corpus allows for a comparative analysis of their translation strategies. The diversity of authors (Fig. 2) and translators (Fig. 3) ensures, from the first stage of development, a certain variety in the representation of Yiddish linguistic and stylistic variants within the corpus.

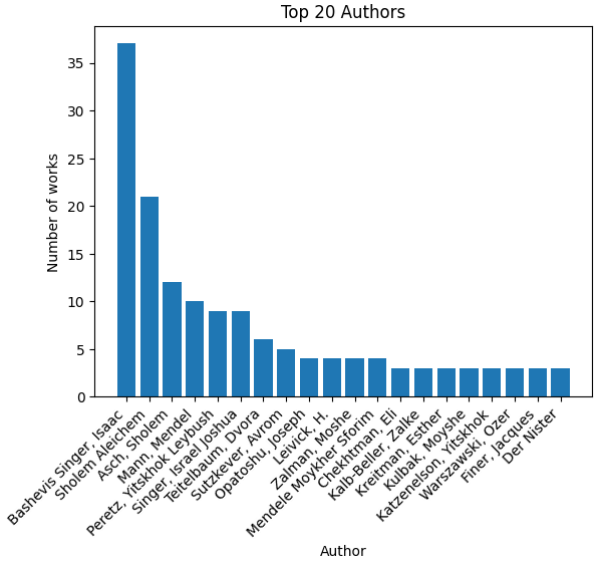

Figure 2: Author distribution in the Yiddish-French corpus.

Our corpus essentially comprises two literary forms: prose works (novels and short stories) and poetry (in much smaller quantities). This generic distribution reflects general trends in the publishing market for translations of Yiddish works, where novels and prose stories are favored over poetry and drama.

The corpus metadata include information about the author, original and translated titles, translator, editor and date of translation publication, and the direction of the translation.

Copyright issues are important to our team, which is why we took care to address them properly by building a registered-user, limited-access parallel corpus. All original Yiddish works included in the corpus are in the public domain, with their copyright statuses individually verified. The French translations are mostly under copyright, while the Russian translations have a mixed copyright status.

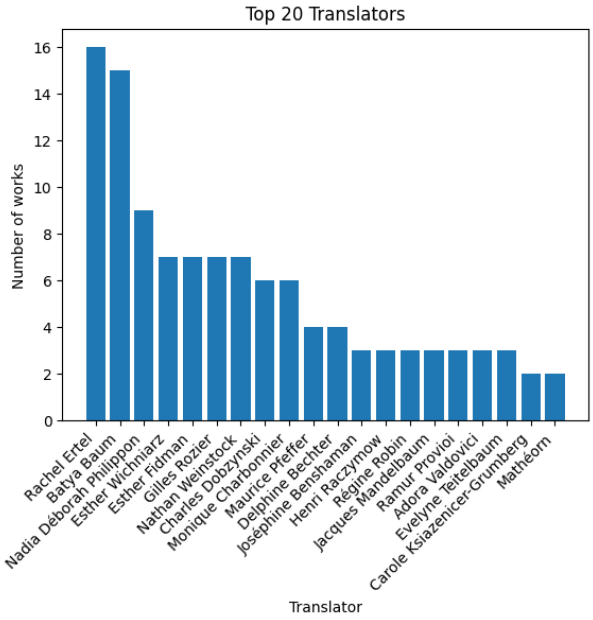

Figure 3: Translator distribution in the Yiddish-French corpus.

For our parallel corpus, we adopted a snippet + access control model: the corpus stores the full texts of all works—regardless of copyright status—but only displays short aligned excerpts (2 to 5 lines of text) in response to each query. These snippets are minimal and necessary for research purposes. Full texts are not available for download. To reduce the risk of data scraping, a limit is placed on the number of queries allowed per day. Access to the concordancer is restricted to registered users only.

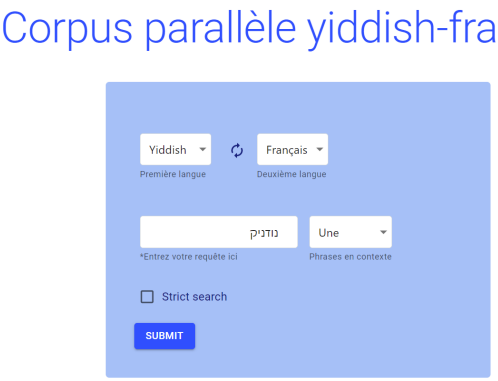

Figure 4: Corpus interface.

The corpus is accessed through a web-based bilingual search interface (Fig. 4 and 5) designed to support both linguistic research and translation-oriented exploration. From the user's point of view, the interface foregrounds parallel reading and contrastive analysis between Yiddish and French or Yiddish and Russian.

At the top of the interface, users select the direction of consultation by choosing a source language ("Première langue") and a target language ("Deuxième langue") from drop-down menus. The direction can be reversed instantly using a swap icon, which makes the tool equally suitable for Yiddish→French and French→Yiddish exploration. This bidirectionality reflects the corpus design, where neither language is treated as secondary.

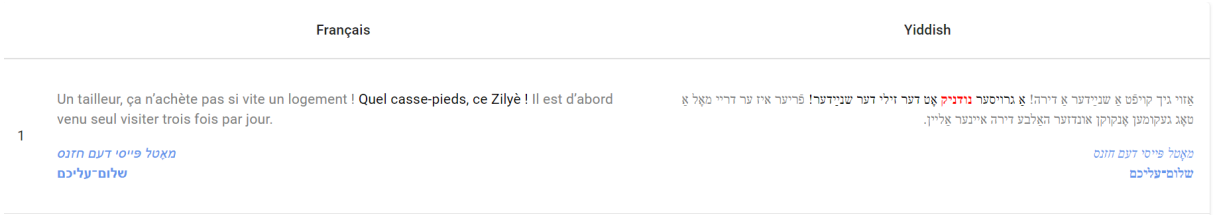

Figure 5: An example of corpus alignment.

The central search area allows the user to enter a text query in the selected source language. Queries can be single words or longer expressions. A complementary option lets users specify the display mode, such as viewing isolated matches or phrases in context, thus supporting both lexical lookup and contextual, discourse-level investigation.

A "strict search" option enables more controlled queries. When activated, it restricts results to exact matches, which is particularly useful given orthographic variation and the multilingual lexical layers characteristic of Yiddish texts. Without this constraint, the system allows more flexible retrieval, which benefits exploratory searches and literary analysis.

Once the query is submitted, results are displayed as aligned bilingual segments. Each Yiddish segment is presented in direct correspondence with its French translation, allowing users to observe how specific lexical items, stylistic choices, or culturally marked expressions are rendered across languages. This alignment makes visible not only equivalence relations, but also translation shifts, omissions, expansions, and stylistic adaptations.

Beyond simple retrieval, the interface encourages a mode of reading that moves fluidly between original and translation. Rather than replacing the source text, the translated segment functions as an interpretive layer, enabling users with different levels of Yiddish proficiency to engage with the material. In this sense, the interface operationalizes the project's broader goal: facilitating access to post-vernacular Jewish-language literature while preserving the visibility of the original text.

Overall, from the user's perspective, the interface serves simultaneously as:

1. a parallel concordancer for linguistic analysis,
2. a translation comparison tool, and
3. a mediated reading environment that supports access to culturally and linguistically complex texts.

6 User Feedback on Corpus Utility and Functionality

To evaluate the practical value of the parallel corpus (Yiddish–French), a questionnaire was administered to three professional translators working from Yiddish into French, Russian, and Italian. Respondents were asked to assess the corpus in terms of usefulness, accessibility, use cases, and potential improvements.

Regarding general usage, translators report consulting the corpus primarily for literary texts and newspaper articles, positioning it as a complementary resource consulted when dictionaries prove insufficient or when confirmation of a translation choice is needed. The tool is deemed most valuable for the translation of idioms, where it serves to illustrate how other translators have resolved specific interpretative challenges.

In terms of interface and data quality, users find the platform intuitive and the presentation of source and target texts with associated metadata satisfactory. While the users acknowledges the inherent limitations of a relatively small corpus—noting that larger databases inevitably offer broader lexical coverage—the examples retrieved are consistently judged as relevant to the translation task at hand. The users'

neutral responses regarding result variation suggest that evaluation criteria for diversity in concordance output may not be immediately self-evident to all users.

The impact of the corpus on the translator's practice is described as selective but meaningful. It has occasionally contributed to terminological precision and enhanced the user's comprehension of Yiddish, though it has not notably influenced stylistic fluency. Functionally, the corpus operates as a "last resource": dictionaries remain the primary consultation tool, with the corpus serving as a final arbiter in cases of residual doubt or when dictionary-based solutions prove inadequate. This positioning is reinforced by the user's concurrent reliance on several Yiddish dictionaries and lexical databases, underscoring the corpus's role as a welcome but secondary integration within a diversified toolchain.

A key area for improvement concerns the search functionality. The users report difficulty in performing exact multi-word searches, such as for compound expressions (e.g., *arbeter-tog*), where the current system returns separate results for each constituent term without prioritizing the precise combination. This limitation, while not impeding single-word or highly specific expression searches, significantly reduces the utility of the corpus when common terms are combined, generating result sets that are too extensive for efficient manual filtering. Addressing this through more advanced morphological or phrase-search capabilities would considerably enhance the resource's practical value for professional translation tasks.

7 Conclusion

We presented a parallel corpus constructed within the LJTRAD project. The corpus not only complements and reinforces the information provided by bilingual dictionaries by offering a wealth of contextual data, but also enables a critical and reasoned analysis of translation practices from Yiddish into French and Russian over time and across individual translators. By presenting source texts and their translations side by side, the corpus tools make it possible to reconstruct what Berman terms the "translation contract" of each translator (Berman, 1995), revealing whether their approach prioritizes the sensitivity of the French or Russian reader or, on the contrary, seeks to decenter the reader's perspective through the introduction of foreign cultural elements.

The resource will ultimately support detailed critiques of individual translations and provide a more informed assessment of how Yiddish literature has been transmitted into French and Russian. It will also help situate translators' work within a broader historical and cultural context, contributing to a more consistent handling of culture-specific elements while opening the way for proposing alternative, and potentially more adequate, translation strategies.

Finally, the aligned trilingual literary corpus represents a valuable resource for the improvement and evaluation of automatic systems based on neural language models in complex literary settings. Unlike conventional corpora used for training machine translation systems, it captures human decision-making processes, stylistic variation, and aesthetic criteria. This raises a broader and still open question: to what extent can machine translation models be trained on literary data without losing the richness, ambiguity, and interpretative depth that characterize literary translation? Addressing this question will be essential for future research at the intersection of computational linguistics and literary studies.

8 Acknowledgments

We are grateful to the translators for their thoughtful feedback.

The work was supported by the French National Research Agency (ANR) and the French Ministry of Higher Education and Research (MESR).

References

Avishur, Y. (1979). The folk literature of the jews of iraq in judeo-arabic. *Pe'amim*, *3*, 83–90.

Berman, A. (1995). *Critique des traductions: John Donne*. Gallimard.

Berman, A. (2008). *L'âge de la traduction. « la tâche du traducteur », de walter benjamin. un commentaire*. Saint Denis, PUV.

Bikard, A., Fedchenko, V., Furci, G., & Rousselet, C. (2025). Introduction à "Les Langues juives en partage" / "Sharing Jewish Languages: Translation and Identity". *Trans : Revue de Littérature Générale et Comparée*. https://doi.org/10.4000/14807

Bikard, A., Fedchenko, V., & Urieli, A. (2025). Corpus parallèle de littérature yiddish et de traductions en français et en russe : Un outil pour l'analyse des défis du transfert culturel. entre les dictionnaires bilingues et les traducteurs automatiques. *Trans: Révue de littérature générale et comparée*, *31*. https://journals.openedition.org/trans/14294

Birzer, S. (2014). Yiddish passive constructions: A case study based on the new corpus of modern yiddish. *Yiddish language structures*, 290–321.

Borovaya, O. (2011). *Modern ladino culture: Press, belles lettres, and theater in the late ottoman empire*. Bloomington: Indiana University Press.

Chinski, M. (2024). Un siècle de traduction du yiddish en france. trajectoires, mémoires et projets culturels. *Tsafon. Revue d'études juives du Nord*, (63), 35–52.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

De Jong, F., Maegaard, B., De Smedt, K., Fišer, D., & Van Uytvanck, D. (2018). Clarin: Towards fair and responsible data science using language resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

de Graaf, T. (1996). Joshua a. fishman. 1991. yiddish: Turning to life. *Studies in Language*, *20*(1), 191–195. https://doi.org/10.1075/sl.20.1.09gra

Doval, I., & Nieto, M. T. S. (2019). *Parallel corpora for contrastive and translation studies: New resources and applications* (Vol. 90). John Benjamins Publishing Company.

Fishman, J. (Ed.). (1985). *Readings in the sociology of jewish languages*. Leiden: Brill.

Hary, B. (1992). *Multiglossia in judeo-arabic - with an edition, translation and grammatical study of the cairene purim scroll*. Leiden: Brill.

Kahn, L., & Rubin, A. D. (Eds.). (2015). *Handbook of jewish languages*. Leiden: Brill.

Kulick, S., Ryant, N., Santorini, B., Wallenberg, J., & Urieli, A. (2023). A part-of-speech tagger for Yiddish. *arXiv preprint arXiv:2204.01175*.

Ladmiral, J.-R. (2014). *Sourcier ou cibliste. les profondeurs de la traduction*. Les Belles Lettres.

Lefer, M.-A. (2021). Parallel corpora. In *A practical handbook of corpus linguistics* (pp. 257–282). Springer.

Lowenstein, S. M. (2000). *The jewish cultural tapestry: International jewish folk traditions*. New York: Oxford University Press.

Mikhailov, M., & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: A guide for research*. Routledge.

Moskovich, W. (2010). Directions in modern yiddish lexicography. *Jews and Slavs*, *22*.

Niborski, Y., & Vaisbrot, B. (2002). *Dictionnaire yiddish-français*. Bibliothèque Medem.

Norich, A. (2014). *Writing in tongues. translating yiddish in the 20th century*. Washington: University of Washington Press.

Romero, E. (1992). *La creación literaria en lengua sefardí*. Madrid: Mapfre.

Shandler, J. (2006). *Adventures in yiddishland: Postvernacular language and culture*. Berkeley: University of California Press.

Spolsky, B. (2014). *The languages of the jews: A sociolinguistic history*. Cambridge, UK: Cambridge University Press.

Stillman, Y. K., & Stillman, N. A. (Eds.). (1999). *From iberia to diaspora: Studies in sephardic history and culture*. Leiden: Brill.

Urieli, A., Clooney, A., Sigiel, M., & Leyfer, G. (2025). Jochre 3 and the yiddish ocr corpus. https://arxiv.org/abs/2501.08442

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2008). Parallel corpora for medium density languages. In *Recent advances in natural language processing iv: Selected papers from ranlp 2005* (pp. 247–258). John Benjamins Publishing Company.

Weinreich, M. (2008). *History of the yiddish language*. New Haven: Yale University Press.

Weinstock, N., Vidal Sephiha, H., & Barrera-Schoonheere, A. (1997). *Yiddish and judeo-spanish, a european heritage*. Brussels: European Bureau for Lesser Used Languages.

Zafrani, H. (1980). *Littératures juives en occident musulman*. Paris: Geuthner.

A Digital Humanism perspective on providing language resources to CLARIN in an age of AI commodification: The case of UniTermGPT

Barbara Heinisch
Institute for Applied Linguistics
Eurac Research, Bolzano, Italy
barbara.heinisch@eurac.edu

Abstract

The accelerated uptake of large language models (LLMs) has intensified the demand for high-quality language resources, while simultaneously reshaping the political economy of language data. Research infrastructures such as CLARIN play a central role in advancing open science by providing FAIR-compliant access to language resources, yet they also operate within a context increasingly characterized by AI-driven commodification and extractive data practices. This paper adopts a Digital Humanism perspective to examine the ethical implications of providing language resources to CLARIN in this context, using the UniTermGPT project as an example. UniTermGPT compiles and annotates German-language university corpora and terminology across Austrian, German, Swiss and South Tyrolean varieties to address limitations of large language models in handling language-variety-specific terminology. By discussing the types of data produced by UniTermGPT, their integration into CLARIN, this paper discusses the challenges associated with openness, attribution and potential downstream reuse in opaque AI training pipelines. By situating FAIR and CARE principles within a broader Digital Humanism framework, the paper argues for ethical-by-design approaches to language resource provision that make human labor visible, preserve linguistic diversity and address power asymmetries between public research infrastructures and commercial AI actors.

1 Introduction

As large language models (LLMs) such as ChatGPT become embedded in everyday professional practice, including (specialized) translation (He, 2024; Gao et al., 2023) and terminology work (Reineke, 2023; Heinisch, 2025), they challenge not only technical standards but also the ethical, social and epistemological foundations of digital infrastructures. Under the accelerating pressure of AI (artificial intelligence) commodification, (open) language resources risk being appropriated by commercial actors in ways that undermine transparency, human agency and academic values. Against this backdrop, Digital Humanism offers a useful lens for rethinking how we create and share language resources in infrastructures such as CLARIN (Common Language Resources and Technology Infrastructure). This contribution presents UniTermGPT, a research project that operates at the intersection of translation studies, terminology work and language resource development. Its dual goals are to assess how ChatGPT handles university terminology across German language varieties (Austrian, German, South Tyrolean and Swiss) and to produce a set of ethically grounded, FAIR-compliant language resources to be shared with the CLARIN community. Central to this work is a commitment to the principles of Digital Humanism, placing human expertise, context and control at the center of digital systems.

2 Development and publication of language resources (as part of UniTermGPT)

Language resources comprise machine-readable language data and their descriptions designed to support and enhance language processing applications. They include, for example, written and spoken corpora, lexical resources, grammars, terminologies and other domain-specific databases, dictionaries, ontologies and multimedia collections. Moreover, language resources encompass the foundational software tools required for their collection, preprocessing, annotation, management, adaptation and

Barbara Heinisch 2025. A Digital Humanism perspective on providing language resources to CLARIN in an age of AI commodification: The case of UniTermGPT. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 120–135. DOI: https://doi.org/10.3384/ecp222

practical use (Language Resources and Evaluation, 2026; Piperidis, 2012).

The research project *University terminology in German in the age of ChatGPT* (UniTermGPT) has the creation and publication of language resources at its core. The background of this study is based on the fact that specialized translation and terminology work are impacted by the acceleration in AI (artificial intelligence) breakthroughs, such as ChatGPT[1]. However, in the field of specialized translation ChatGPT currently does not take terminological variation sufficiently into account (Heinisch, 2025). Since higher education systems differ between countries, there can be significant terminological variation, even within the same language (Terras et al., 2016), such as between the Austrian, German or South Tyrolean variety of the German language. Therefore, the question arises how ChatGPT can cope with this variation in translation in the field of German university terminology. The overall objective is to analyze how ChatGPT treats university terminology of German language varieties (depending on the prompt). For this, a corpus of Austrian, German, South Tyrolean and Swiss university texts will be compiled, from which terminology will be extracted and compared with existing terminology resources, such as university glossaries and terminology databases. Following prompt design, selected texts are translated in ChatGPT, which are annotated by experts to assess the outcome of translations from these language varieties into English and vice versa.

In addition to open research data and the integration of the needs of translators and terminologists, the results of UniTermGPT can be transferred to other areas characterized by the use of system-bound terminology (for translation purposes), such as law. UniTermGPT supplements existing studies on specialized translation and terminology work by means of ChatGPT and addresses the overlooked topic of language-variety-specific terminology (in specialized translation and in large language models). The resulting recommendations for practitioners and a policy brief highlight the societal relevance of terminology in an age characterized by large language models.

2.1 UniTermGPT resources to be provided to CLARIN

As part of UniTermGPT, several language resources will be provided to CLARIN. The project's data pipeline includes:

- Monolingual and bilingual corpora of higher education texts from German-speaking regions, annotated for language variety and domain-specific terminology;
- Terminology resources containing university terminology in different language varieties. This resource may also be used for the retrieval-augmented generation or terminology-augmented generation of translations generated by ChatGPT into different German language varieties in the field of university texts.
- ChatGPT translations of selected documents, creating a basis for comparative analysis;
- Expert annotations assessing ChatGPT's output's terminological accuracy and consistency across varieties. Together with the annotations, annotation guidelines that will be developed together with the experts will be provided.

These resources will be contributed to CLARIN (and other repositories) with open documentation, open usage licenses and a view to long-term reusability. They are intended not only to advance future research by providing language resources, but also to encourage ethical reflection on the provision of language resources to infrastructures like CLARIN, questioning what counts as "valuable" language resource and ensuring that the benefits of such resources are shared fairly.

2.2 AI commodification of language resources

The act of sharing language resources through an open infrastructure in the current AI landscape invites several risks. Openness, compliance and ethics are closely connected yet often conflicting dimensions of open datasets for LLM training (Baack et al., 2025): openness promotes free access and

[1] https://chatgpt.com/

collaborative sharing of data and code, compliance ensures adherence to evolving legal frameworks, and ethics addresses broader issues of justice, equality and responsibility beyond legal obligations (all three of which must be continuously balanced in the face of rapid technological change).

Commercial LLM providers increasingly scrape open repositories, including academic and governmental corpora, without license or copyright compliance (Baack et al., 2025). In doing so, they appropriate public intellectual labor and undermine the collaborative landscapes that infrastructures like CLARIN are built upon.

This leads to questions related to the commodification of knowledge work (Popowich, 2024), which in turn form the basis of several fundamental inquiries raised within the UniTermGPT project:

- How can a researcher requesting language resources from experts, including terminology resources and translation memories, ensure that these resources are shared willingly under an open license? So, how can language resources provided under CLARIN, and other repositories, remain open while preventing extractive AI practices?
- What ethical obligations do researchers have when contributing to research infrastructures in a commodified digital landscape?
- How can CLARIN and similar platforms lead the way in ensuring reciprocity, transparency and accountability in AI data flows?

These are central questions, intertwined with questions also raised within Digital Humanism.

2.3 Digital Humanism

Digital Humanism is an interdisciplinary approach that places human values, rights and agency at the center of technological development and digital innovation (Werthner, 2022). It critically examines how digital technologies shape society and seeks to ensure that they serve democratic, inclusive and ethical goals rather than purely commercial or exploitative interests. The aim of Digital Humanism is to align digital transformation with the public good, thereby promoting transparency, accountability and human dignity in the design, deployment and governance of digital systems (Mayer & Strassnig, 2020; Nida-Rümelin, 2021). In the context of language technologies and infrastructures like CLARIN, it calls for responsible data practices that respect contributors and resist the unchecked commodification of knowledge.

2.4 Digital Humanism and its relation to FAIR and CARE principles

The FAIR and CARE principles provide complementary frameworks for research data management: while FAIR emphasizes making data Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016), CARE highlights the need for Collective Benefit, Authority to Control, Responsibility and Ethics (Carroll et al., 2020), ensuring that openness is balanced with respect for communities.

2.4.1 All about data: FAIR and CARE principles

While the FAIR principles are data-centric, technical and accessibility-oriented (Wilkinson et al., 2016), the CARE principles are people- and rights-centric, especially for indigenous data (Carroll et al., 2020). FAIR is rather technology-neutral and does not address ethical or societal issues. These principles guide how data should be structured and shared to ensure they can be easily located, understood and reused by humans and machines. CARE on the other hand, emphasizes equity, governance and the rights of data subjects, particularly indigenous communities. CARE ensures that data use aligns with ethical standards and governance and community interests (Table 1).

2.4.2 Humane digital technologies through Digital Humanism

The added value of Digital Humanism to the FAIR and CARE principles are human-centric values in digital technologies. Digital Humanism is not just about data and the use of data (as the FAIR and CARE principles are) but also critically examines the diverse impacts of (data use and) technologies. Digital Humanism thus extends beyond research data management and brings in broader human values

and rights into digital technologies often overlooked by FAIR and CARE (Table 1). Applied to research data management, Digital Humanism can add a normative layer to FAIR's technical focus, asking not only *can* we use this data, but *should* we? Digital Humanism promotes human rights and autonomy in the digital age, as well as the design and governance of digital systems that respect and support human agency, rather than prioritizing data efficiency alone. It pushes for accountability, democratic values, inclusiveness and resistance to power imbalances created by opaque data practices and digital systems.

Topic	**FAIR**	**CARE**	**Digital Humanism**
Overall focus	Research data management	Research data management	Digital technologies
Ethics	Not addressed	Indigenous-focused	Universal human rights and dignity
Governance	Technical	Community-led	Democratic (oversight), transparent, accountable
Power and societal impact	Not addressed	Unequal power, community control	Broader justice, algorithmic bias and surveillance
Human agency	Data-machine interoperability	Community-centric control	Agency, autonomy, informed consent, freedom
Inclusivity	Not addressed	Indigenous rights	Cultural pluralism and epistemic diversity
Purpose	Data reuse	Ethical aspects	Societal benefit, autonomy
Awareness	Machine-readability	Indigenous data sovereignty	Critical thinking and digital literacy

Table 1: Comparison of FAIR, CARE and Digital Humanism.

3 Making language resources openly accessible by taking Digital Humanism into account

The web operates on an extractive capitalist model that commodifies user data and labor, perpetuating historical inequalities while offering users (minimal) benefit beyond basic access to services (Lindgren et al., 2025). When applied to the domain of language resource creation and dissemination, the Digital Humanism perspective thus raises fundamental questions: Who creates language data? Who benefits from it? And how can we ensure that open access does not become a gateway to exploitation?

The UniTermGPT project provides a case study in navigating these tensions, focused on the creation of annotated and FAIR-compliant language resources for integration into the CLARIN infrastructure and other repositories. While the consideration of FAIR and CARE principles is by no means unique to the UniTermGPT project (being adopted across digital humanities initiatives and among CLARIN practitioners), UniTermGPT is presented here as an illustrative case study through which these principles can be examined in practice.

However, opening language resources is not a neutral act. This project unfolds in a context where commercial AI companies often scrape open resources (Baack et al., 2025) for use in proprietary models. This practice, frequently done without consent or licensing compliance, has led to growing

concern and rising reluctance among experts and researchers about contributing to open repositories.

From a Digital Humanism perspective, this reluctance must be taken seriously. The aim is not to restrict access, but to redefine what responsible openness means. In UniTermGPT, this means involving experts (translators, terminologists and domain experts) not just as data providers, but as key stakeholders of the resource creation process. Their insights inform not only the content but also the ethical boundaries of data sharing.

3.1 Language data and open science

Language data pose particular challenges for open science. Most contemporary language data are protected by copyright, as linguistic expressions typically qualify as intellectual creations, limiting their unrestricted dissemination (Jong et al., 2022). In addition, language data often contain personal data, making them subject to data protection regulations such as the EU General Data Protection Regulation (GDPR). These legal and ethical constraints complicate the implementation of full openness. CLARIN addresses these challenges by adopting an "as open as possible, as closed as necessary" approach (Jong et al., 2022), applying standardized licensing schemes and providing legal and ethical guidance through its dedicated committee. Where open access to raw data is not feasible, CLARIN also explores alternative methods that allow researchers to benefit from language data through derived outputs or restricted-access services, balancing openness with legal and ethical responsibility (Jong et al., 2022).

CLARIN has been guided from its inception by the core values of the open science agenda, placing openness, sharing and reuse of language resources at the center of its mission (Jong et al., 2022). The infrastructure supports open access to language data wherever possible and promotes the use of open standards, open-source software and interoperable metadata to ensure that research based on these resources is transparent, reproducible and replicable (Jong et al., 2022). By enabling proper documentation and citation of language resources, CLARIN aligns openness with established academic norms and supports sustainable research practices across disciplines (Jong et al., 2022).

Within CLARIN, open science is understood as the promotion of transparent, sustainable and inclusive access to language data and technologies "in support of research in the humanities and social sciences and beyond" (Jong et al., 2022). Central to this vision are the FAIR principles, which aim to facilitate the discovery, access and meaningful reuse of research data by both academic and non-academic communities (Eskevich et al., 2020). By encouraging the use of widely accepted standards for data and metadata, FAIR supports reproducibility, interoperability and long-term usability of language resources (Eskevich et al., 2020).

CLARIN operationalizes open science through a distributed, federated infrastructure that provides sustainable access to FAIR-compliant language data and tools across Europe (Jong et al., 2022). However, CLARIN focuses only on the open access aspect of open science (open data, open standards, open source code) as well as open science infrastructures while open science more broadly encompasses additional pillars, including the active engagement of societal actors through participatory and citizen science, among others, and open dialogue with other knowledge systems (including those of indigenous peoples, marginalized scholars and local communities), as articulated in the UNESCO Recommendations on Open Science (UNESCO, 2021).

Moreover, while CLARIN is strongly aligned with the FAIR principles, it does not explicitly incorporate the CARE principles, and the aspect of ethical data stewardship remains only loosely defined within the infrastructure. In particular, it is not always transparent which concrete criteria or ethical frameworks guide decisions on data provision, reuse and governance beyond legal compliance. Although the work of the CLARIN Committee for Legal and Ethical Issues (CLIC) (Kamocki et al., 2022) plays an important role in addressing copyright, data protection and regulatory developments, ethical considerations are largely framed in procedural or legal terms rather than articulated as a coherent set of normative principles. As a result, broader questions concerning responsibility, power asymmetries and the societal consequences of data reuse (especially in the context of AI-driven commercialization) remain under-specified within CLARIN's current open science framework.

CLARIN already engages with industry through a range of regional collaborations, particularly in areas such as machine translation and speech technology, which provide a foundation for a more coordinated innovation strategy. Through these partnerships, CLARIN positions itself as an important contributor to broader digital transformation processes, extending its impact beyond academia. Many of the tools, datasets and services developed within CLARIN constitute valuable components for commercial software development, as language technologies underpin a wide spectrum of AI applications. As demand for AI-driven solutions continues to grow, interest from commercial actors in CLARIN's technologies and language resources is expected to increase accordingly (Jong et al., 2022).

3.2 AI and language data

"[T]reating language as data has always been problematic" (Erdocia et al., 2024) because it abstracts language from the social, historical and interactional contexts in which meaning is produced. Digital language data are often assumed to represent stable linguistic systems or social groups, yet language use is fundamentally dynamic, situational and indexical. Meanings emerge locally through interaction and cannot be fully captured by decontextualized datasets or quantitative correlations (Erdocia et al., 2024). Linguistic categories, practices and identities are fluid and shaped by power relations, social positioning and context, rather than fixed attributes that can be straightforwardly mapped onto data points (Erdocia et al., 2024).

These epistemological issues are historically rooted. According to Erdocia et al. (2024), dominant Western approaches to linguistics were shaped during European colonialism, when language was often treated as an object to be collected, classified and standardized, frequently ignoring speakers' own understandings and lived experiences. Contemporary data-driven approaches risk reproducing these extractive logics by privileging large-scale digital traces over situated knowledge. As LLM developers (largely controlled by a small number of US-based companies) collect and monetize vast amounts of language data for AI systems, the gap between language as lived practice and language as data commodity widens. This makes it increasingly necessary to develop methods and conceptual tools that take seriously the social, political and technological conditions under which language data are produced, circulated and transformed, rather than relying solely on digital datasets or AI outputs as proxies for language and community.

Western linguistic epistemologies and data-driven approaches to language were, as mentioned before, shaped during European colonialism, where language documentation often served extractive, administrative or missionary goals. Today's AI-driven language technologies echo these dynamics through what critical scholars describe as data colonialism: the large-scale extraction, commodification and surveillance of language data by a small number of powerful technology companies (Erdocia et al., 2024; Zuboff, 2019). Language data has become central to AI infrastructures that not only generate profit but also shape language ideologies, communication practices and even community formation worldwide.

The role of AI companies is therefore pivotal. Usually operating within opaque, profit-driven contexts, they transform linguistic practices into commercial assets while externalizing social costs (Birhane, 2020). Governments, meanwhile, have largely adopted techno-solutionist approaches (Morozov, 2013), delegating responsibility for language technologies to the private sector. This has contributed to profound digital inequalities between languages, privileging those with large, standardized datasets and marginalizing others. As public institutions lose normative authority in language matters, dependency on private AI infrastructures deepens (Erdocia et al., 2024).

In response, scholars (Erdocia et al., 2024) argue for renewed public responsibility and regulation, including ethical public-private partnerships, stronger governance frameworks, and a reassertion of the state's role in protecting fundamental rights. Ultimately, addressing the consequences of language datafication requires not only better data practices, but a holistic, critical understanding of how language, technology, power and profit intersect in the age of AI.

To address the opacity and profit-driven dynamics of the current AI landscape, several European

initiatives (often supported by European Union funding) seek to develop alternative models grounded in principles of transparency, openness and legal accountability. Projects such as OpenEuroLLM[2] and EuroLLM[3] aim to build 'European LLMs' that are open source, multilingual and explicitly aligned with European data governance standards, including clear documentation of data provenance and adherence to licensing requirements. Similarly, the Swiss-developed multilingual model Apertus[4] reflects this broader movement toward more transparent and publicly accountable AI development.

A central feature of these initiatives is their emphasis on responsible data sourcing. The OpenEuroLLM consortium (2025) prioritizes datasets that are openly accessible, accompanied by clear terms of use and free from restrictions on use in model training. This approach seeks to minimize legal uncertainty while promoting reproducibility and trust in AI systems. At the same time, these projects aim to counter the strong English-centric bias of many existing models by foregrounding multilingualism and supporting a wider range of European languages, thus being more aligned with Digital Humanism.

However, not all open models adhere to these principles to the same extent. While some models, such as DeepSeek (DeepSeek-AI, 2024), are presented as open, they provide limited transparency regarding their training data and remain largely focused on a narrow set of languages (Chinese and English). This highlights that openness alone may not be sufficient. Transparency and humane technology development are essential criteria for aligning AI development with the principles of Digital Humanism.

3.3 Who creates language data?

Language data are often perceived as a freely available resource, yet in reality it is the result of substantial human labor, expertise and institutional support. In infrastructures like CLARIN, language data are typically created by researchers, including linguists or digital humanities scholars (Wynne, 2015), often within publicly funded projects. This work involves careful selection, annotation, contextualization and documentation of texts. In addition, these processes require scholarly judgment, linguistic and cultural knowledge and ethical and legal responsibility into the data.

Crucially, much of this labor remains invisible once the data are deposited in a repository. When language resources are treated merely as "datasets," the human expertise and social context behind their creation risk being erased. A Digital Humanism perspective insists on making this labor visible and recognized, acknowledging contributors not as passive data suppliers but as knowledge producers with legitimate interests in how their work is used. This concern is also highlighted by Ousidhoum et al. (2025), whose survey of the NLP community identifies "problematic incentivisation and the lack of recognition for data workers' labor" (Ousidhoum et al., 2025) as a recurring issue.

3.4 Who benefits from language data?

Ideally, language data (in the form of corpora, dictionaries or annotated datasets, for example) shared through infrastructures like CLARIN do not only benefit researchers in the social sciences and humanities, enabling reproducible research and supporting linguistic diversity but also technology providers in achieving technological innovation. However, in the current AI landscape, the benefits are increasingly asymmetrically distributed. Large technology companies can extract enormous economic value by incorporating open language resources into proprietary models.

From a Digital Humanism perspective, the reuse of open language resources for technological innovation is not inherently problematic. On the contrary, the development of language technologies is both desirable and socially valuable. Infrastructures such as CLARIN were created precisely to support research and innovation in general. The ethical concern arises not from the use itself, but from the asymmetrical distribution of benefits that currently characterizes the global AI ecosystem.

[2] https://openeurollm.eu/
[3] https://eurollm.io/
[4] https://apertvs.ai/

At present, it is predominantly large, non-European (most notably US-based) technology companies that are able to extract enormous economic value from publicly funded, openly accessible (European) language resources by incorporating them into proprietary AI models. Meanwhile, the original data creators, expert communities and public institutions that curate and maintain these resources often receive little recognition and virtually no influence over downstream use. This imbalance results in a form of value extraction in which public investment and human expertise in Europe are transformed into private profit elsewhere.

In this sense, Digital Humanism reframes the debate: the key question is not whether language resources should be used for AI development, but who benefits from this use and under what conditions. Ensuring that technological innovation strengthens, rather than undermines, public infrastructures and European knowledge ecosystems is essential for a sustainable and ethically grounded digital future.

3.5 Open access as a gateway to exploitation?

Ensuring that open access remains aligned with human-centered values requires moving beyond a purely technical understanding of openness. Digital Humanism thus advocates for responsible openness, where access is coupled with ethical safeguards and governance structures.

When applied to language resources, this can include the following:

- Transparent licensing that explicitly addresses AI training and commercial reuse;
- Metadata practices that preserve provenance, attribution and context;
- Infrastructure-level policies that articulate ethical norms and discourage extractive practices;
- Community engagement that allows for human agency.

In this sense, open access should not mean unrestricted extraction, but accountable participation in a shared knowledge ecosystem. Infrastructures like CLARIN are uniquely positioned to embody this balance, ensuring that openness fosters collaboration and innovation without sacrificing the rights and recognition of those who create language data.

3.6 Applying Digital Humanism to UniTermGPT

In the context of UniTermGPT, the project does not treat university terminology in German merely as structured data to be made findable or reusable. Instead, it recognizes such terminology as the product of institutional histories, national higher education systems and expert knowledge developed by translators, terminologists and domain experts. From a Digital Humanism perspective, the central question is not only whether these terminology resources can be integrated into CLARIN, but whether their reuse (particularly by commercial AI providers) respects the intent and values of the communities that created them.

Digital Humanism thus extends FAIR's technical focus by introducing a "should we?" alongside the "can we?" In UniTermGPT, this is reflected in decisions about which texts to include, how to document provenance, how to license annotated corpora and how to communicate the intended scope of reuse. The UniTermGPT project explicitly acknowledges that the reuse of its language resources for technology development is both legitimate and, in principle, desirable. However, it critically reflects on the risk that once these publicly funded and openly licensed resources are released, they may be incorporated into opaque, proprietary AI training pipelines. In such cases, the value generated from expert-curated terminology data may flow almost exclusively to commercial actors, while the original contributors and public infrastructures remain invisible and excluded from downstream benefits, decision-making and recognition. This is an imbalance that UniTermGPT seeks to make visible and address through transparent documentation and ethical reflection. While the UniTermGPT project itself does not develop language technologies, it seeks to integrate Digital Humanism principles throughout its research design and implementation. Central to this approach is a critical engagement with language technologies, particularly regarding their development and the distribution of benefits.

Such reflexivity is understood as a fundamental component of responsible academic research. Moreover, the project does not conceptualize translators merely as data annotators or providers, but as key stakeholders whose expertise, expectations and experiences can inform and shape the evaluation and responsible use of language technologies.

Moreover, UniTermGPT aligns with Digital Humanism by reflecting on power asymmetries inherent in the current AI ecosystem and in research related to NLP (natural language processing). By embedding expert annotation, transparent metadata and ethical reflection into the corpus design, UniTermGPT seeks to preserve human agency. Although UniTermGPT cannot directly influence the development of systems such as ChatGPT or other large language models, it contributes to practice-oriented knowledge by formulating recommendations for their use in specialized translation contexts. A further output aligned with Digital Humanism is a policy brief that highlights the importance of linguistic diversity in LLMs, with particular attention to specialized communication and the role of language varieties, which remain insufficiently represented in current models.

At first glance, the use of a proprietary system such as ChatGPT in the UniTermGPT project may appear at odds with Digital Humanism principles. However, this choice reflects the empirical reality of professional translation practice, where such tools are already widely used. At the time the project commenced, institutionally hosted or open-source alternatives were not readily available in many academic contexts (a situation that has since evolved). Importantly, the project emphasizes informed participation: data providers retain agency over their contributions and are explicitly informed that excerpts of their texts may be processed using ChatGPT, allowing them to make conscious decisions about data sharing.

4 AI ethics for language technologies

Closely aligned with Digital Humanism is the field of AI ethics, which is concerned with the responsible design, development and use of artificial intelligence systems in ways that respect human values and morality (Siau & Wang, 2020). AI ethics seeks to ensure that technological innovation serves human and societal interests rather than undermining them. According to Kazim & Koshiyama (2021), AI ethics encompasses a set of core themes that address the societal and human implications of artificial intelligence. Central to these are human well-being, which especially includes human agency. Safety is another key concern, including robustness, reliability, reproducibility, protection against malicious use and preparedness for unknown risks. AI ethics further emphasizes privacy and responsible data stewardship through principles such as data minimization. Transparency, including explainability and clear communication, is essential for building trust, alongside fairness, which addresses bias, accessibility and meaningful participation. Finally, accountability ensures that responsibility for AI systems remains with humans, supported by mechanisms such as impact assessments (Kazim & Koshiyama, 2021).

AI ethics for language technologies starts from the recognition that language is not a neutral or purely technical resource, but a form of symbolic capital that shapes identities, social relations and access to power (Ousidhoum et al., 2025). Treating language data as mere collections of tokens risks ignoring their cultural, political and social embeddedness. As language technologies are fundamentally data-driven, ethical concerns are closely tied to how language data are selected, collected, annotated and represented. Rigorous and transparent data practices are therefore essential for developing human-centered and socially aware NLP systems (Ousidhoum et al., 2025; Kotnis et al., 2022).

Recent research has emphasized the need to move beyond technically optimized models toward approaches that incorporate diverse stakeholders (Bird & Yibarbuk, 2024), particularly those of native speakers and language communities (Ousidhoum et al., 2025). This is especially critical for low-resource languages, where data scarcity often leads researchers to rely on whatever material is available, with limited scrutiny of quality, representativeness or cultural relevance: “Due to the limited availability of online data for some languages, there is a tendency to use any accessible source to build resources – often without considering the ethical implications or the appropriateness of the content” (Ousidhoum et al., 2025).

As a result, NLP artefacts for low-resource languages frequently reproduce standardized or dominant varieties, overlook linguistic variation and lack grounding in local contexts (Ousidhoum et al., 2025). However, the expansion of multilingual language models has created new opportunities for the inclusion of low-resource languages in digital environments, as patterns learned from dominant languages are transferred to support generation in less-resourced ones. In this sense, extractive data practices may have the unintended effect of increasing the visibility and functional usability of marginalized languages, potentially contributing to their presence and, to some extent, their sustainability in digital spaces. However, this dynamic remains deeply ambivalent (Chonka et al., 2026). The largely unsupervised application of such models can produce outputs that are culturally inappropriate, reinforce biases or misrepresent local linguistic norms (Chonka et al., 2026), thereby shaping language use in ways that may be misaligned with community values. Moreover, when the incorporation of these languages into AI systems occurs within frameworks of data colonialism and surveillance capitalism (Couldry & Mejias, 2019), this is entangled with global power asymmetries.

Furthermore, ethical challenges extend to the conditions under which language resources are created. Limited funding, reliance on informal networks to find annotators and the absence of clear standards for crediting and compensating contributors (Ousidhoum et al., 2025) often render human labor invisible and undervalued. This raises concerns about potential exploitation, particularly when community members contribute knowledge that later underpins widely used technologies. At the same time, this dynamic highlights a structural imbalance: the creators of language data, often individuals or small research teams, typically lack the legal and financial resources to enforce licensing terms or challenge misuse. This tension underscores the need for approaches that both support the inclusion of low-resource languages in AI development and ensure that such inclusion does not come at the cost of the erosion of trust in open research infrastructures. Addressing these issues requires promoting community agency and embedding ethical reflection into the entire lifecycle of language technologies, from data creation to deployment, “especially considering that such datasets are likely to persist and influence future research” (Ousidhoum et al., 2025).

Extractivist practices in NLP are visible in the uncritical embrace of large language models that prioritize scale and performance over epistemic depth, social responsibility and linguistic diversity (Bird, 2024). Framing progress as ever-larger datasets and models raises questions about whether empiricist approaches to language genuinely capture meaning or merely rearrange surface forms at great computational and environmental cost (Bird, 2024). The reliance on extractive data collection (often detached from the communities and contexts that produce language) risks reinforcing homogenization rather than responding ethically to the world’s linguistic diversity. The field should address questions like whether current trajectories of extraction, energy consumption and abstraction truly align with the values of human-centered, socially grounded language technologies (Bird, 2024).

4.1 The datafication of language and its consequences

From a Digital Humanism perspective, the datafication of language raises fundamental ethical, epistemological and political concerns that go beyond questions of data management or technological efficiency. While treating language as data has always been analytically reductive (Erdocia et al., 2024), the rise of AI-driven language technologies has amplified its consequences. Large language models increasingly rely on vast quantities of decontextualized digital language data, reinforcing the false assumption that linguistic practices can be meaningfully represented through scalable datasets alone (Erdocia et al., 2024). Digital Humanism challenges this reductive view by insisting that digital technologies (and indirectly also the infrastructures that support them) must be grounded in human values, rights and social realities. Applied to language infrastructures like CLARIN, this perspective reframes language resources not as neutral inputs for computational systems, but as human-centered knowledge artifacts shaped by historical, cultural and institutional conditions.

Moreover, Digital Humanism foregrounds the power asymmetries inherent in the contemporary AI landscape. A small number of technology companies extract and commodify language data at scale, often without transparency, consent or reciprocity. Language infrastructures such as CLARIN occupy a critical position in this landscape: as publicly funded, non-commercial repositories, they can act as

counterweights to extractive AI practices by embedding ethical governance, provenance and accountability into the lifecycle of language resources. Digital Humanism thus provides a conceptual framework for reasserting human agency and public responsibility in an age where language data has become a central resource for AI commodification.

Aligning CLARIN with Digital Humanism therefore means moving beyond technical frameworks such as FAIR alone and incorporating normative questions into infrastructure design and policy: not only how language data can be reused, but by whom, for what purposes and with what consequences. In this sense, CLARIN's role extends from enabling access to language data to stewarding a digital public good. By supporting reflexive, context-aware research and fostering ethical standards for data provision and reuse, language infrastructures can help ensure that language technologies serve human-centered goals.

4.2 Ethical limits of open access under AI commodification

Ethical limits of open access under AI commodification refer to the point at which openness (originally intended to promote knowledge sharing, academic progress and the public good) enables extractive practices that undermine these very goals. In an AI-driven economy, openly accessible language resources can be absorbed into proprietary systems. The ethical issue is, however, not openness itself, but the absence of safeguards that ensure fair benefit-sharing, accountability and respect for the human labor behind the data. From a Digital Humanism perspective, recognizing ethical limits to open access means complementing openness with governance mechanisms that protect contributors, preserve context and prevent the unbalanced commodification of publicly funded language resources.

4.3 Language-variety-specific resources as resistance to AI homogenization

Language-variety-specific resources function as a form of resistance to AI-driven homogenization of language, which arises when large-scale models privilege dominant, standardized or economically valuable varieties at the expense of non-dominant forms of language use. AI systems trained primarily on aggregated and decontextualized data tend to flatten variation and do not treat it as meaningful social practice. This is particularly problematic in specialized domains, such as university terminology, where language varieties encode distinct institutional structures, legal frameworks and cultural norms.

From a Digital Humanism perspective, developing and maintaining language-variety-specific resources affirm linguistic diversity as a value rather than a technical inconvenience. Projects like UniTermGPT, which document university terminology, counteract the implicit normativity embedded in AI systems that often default to one "standard" variety.

4.4 Making human labor visible in language technologies

Making human labor visible in language technologies means explicitly recognizing and documenting the extensive human expertise, intellectual effort and contextual knowledge that underpin language resources and AI systems. Language technologies are often presented as autonomous or self-learning, obscuring the fact that they depend on the work of, among others, linguists, annotators and researchers who select texts, define categories, create annotations and evaluate outputs.

From a Digital Humanism perspective, rendering this labor visible is almost an ethical imperative. It challenges extractive narratives that treat language resources as raw material and counters power imbalances in which commercial AI systems profit from publicly funded and/or expert-created data. Visibility can be achieved through detailed metadata on provenance, authorship and annotation practices; explicit citation and attribution norms and documentation that situates datasets within their relevant contexts.

Within infrastructures like CLARIN, making human labor visible reinforces trust and accountability. It affirms contributors as knowledge producers rather than mere data providers and helps ensure that language resources are reused with an understanding of their limitations, assumptions and intended purposes. Ultimately, acknowledging human labor fosters more responsible AI development by

reminding users and developers alike that language technologies are not neutral tools, but socio-technical systems built upon human knowledge and values.

4.5 Ethical-by-design approaches to language resource provision

Ethical-by-design approaches to language resource provision integrate ethical reflection and human-centered values directly into the creation, annotation, documentation and dissemination of language data, rather than treating ethics as an afterthought or external constraint. Instead of focusing solely on technical standards such as interoperability or scalability, ethical-by-design frameworks ask how language resources can be developed and used in ways that respect human agency, linguistic diversity and social responsibility throughout their entire lifecycle.

In practice, this means embedding ethical considerations at every stage of resource provision: from decisions about which texts are collected and whose language practices are represented, to how provenance, authorship and context are documented. It involves making licensing choices transparent, clearly articulating intended and unintended forms of reuse and explicitly addressing risks related to AI training, commercial exploitation and data misappropriation. Ethical-by-design also prioritizes the involvement of domain experts, language communities and contributors as active stakeholders rather than passive data sources or data providers.

Within infrastructures like CLARIN, ethical-by-design approaches can serve as a counterbalance to extractive AI practices. They support openness while resisting unchecked commodification, ensuring that language resources remain aligned with public-interest goals and Digital Humanism principles. Ultimately, ethical-by-design language resource provision seeks to ensure that technological innovation is guided not only by what is technically possible, but by what is socially responsible and beneficial to the communities that create and use language data.

5 Conclusion

Providing language resources to infrastructures such as CLARIN cannot be understood solely as a technical or organizational task but must be situated within the broader socio-technical transformations driven by AI commodification. While CLARIN's strong commitment to open science and FAIR data principles has enabled access to and reuse of language resources, these achievements also expose publicly funded data to extractive reuse by commercial AI actors operating beyond (European) regulatory and ethical frameworks. The UniTermGPT project illustrates both the value and the vulnerability of carefully curated, expert-driven language resources in this environment.

From a Digital Humanism standpoint, the provision of language resources to infrastructures like CLARIN must extend beyond technical compliance with the FAIR principles to embrace a broader commitment to accountability and ethics. Language data (particularly when derived from public institutions and intellectual labor) carries cultural, social and political value that cannot be abstracted or repurposed without consent, attribution and respect for the (language) communities that produce it. From a Digital Humanism perspective, the central challenge is not the use of language resources for technology development per se, but the asymmetries in benefit, control and accountability that characterize current data flows. CLARIN plays a vital role in safeguarding these values by providing a trusted, non-commercial space where contributors retain agency and users could be held to clear ethical standards.

The UniTermGPT project aligns with this vision by demonstrating that open language resources can be designed to serve both technological advancement and the public interest without compromising the rights and values of those who create them. By involving translators and language experts in the compilation and annotation of corpora, and by emphasizing the importance of language-variety-specific terminology, the project underscores the importance of linguistic diversity in global digital systems. This is not merely a matter of accuracy but an ethical commitment to epistemic justice: ensuring that different ways of expressing knowledge are represented and respected in multilingual large language models.

Addressing these tensions requires complementing FAIR and CARE principles with a normative layer that explicitly foregrounds human-centered values, ethical reflection and power relations. Ethical-by-design approaches to language resource provision, clearer governance mechanisms for commercial reuse and greater recognition of contributors' expertise and values are essential to sustaining trust in open science infrastructures.

Acknowledgements

This paper was funded by the EC-MCSA Seal of Excellence Programme of the Autonomous Province of Bozen/Bolzano – Department for Innovation, Research and University, project *University terminology in German in the age of ChatGPT* (UniTermGPT).

References

Baack, S., Biderman, S., Odrozek, K., Skowron, A., Bdeir, A., Bommarito, J., Ding, J., Gahntz, M., Keller, P., Langlais, P.-C., Lindahl, G., Majstorovic, S., Marda, N., Penedo, G., van Segbroeck, M., Wang, J., Werra, L. von, Baker, M., Belião, J., Chmielinski, K., Fadaee, M., Gutermuth, L., Kydlíček, H., Leppert, G., Lewis-Jong, E. M., Larsen, S., Longpre, S., Lungati, A. O., Miller, C., Miller, V., Ryabinin, M., Siminyu, K., Strait, A., Surman, M., Tumadóttir, A., Weber, M., Weiss, R., White, L., & Wolf, T. (2025). *Towards Best Practices for Open Datasets for LLM Training*: *arXiv e-prints*.

Bird, S. (2024). Must NLP be Extractive? In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14915–14929). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bird, S., & Yibarbuk, D. (2024). Centering the Speech Community. In Y. Graham, & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 826–839). Stroudsburg, PA, USA: Association for Computational Linguistics.

Birhane, A. (2020). Algorithmic Colonization of Africa. *SCRIPT-ed*, *17*(2), 389–409.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*.

Chonka, P., Diepeveen, S., & Haile, Y. (2026). "No Language Left Behind?" Predictive Text, Generative AI and Dilemmas of Digital Inclusion for Marginalized Languages. *Modern Languages Open*, *2026*(1).

Couldry, N., & Mejias, U. A. (2019). *The Costs of Connection How Data Is Colonizing Human Life and Appropriating It for Capitalism.* Culture and Economic Life. Stanford, California: Stanford University Press.

DeepSeek-AI (2024). DeepSeek LLM: Scaling Open-Source Language Models. https://github.com/deepseek-ai/DeepSeek-LLM.

Erdocia, I., Migge, B., & Schneider, B. (2024). Language is not a data set—Why overcoming ideologies of dataism is more important than ever in the age of AI. *Journal of Sociolinguistics*, *28*(5), 20–25.

Eskevich, M., Jong, F. de, König, A., Fišer, D., van Uytvanck, D., Aalto, T., Borin, L., Gerassimenko, O., Hajic, J., van den Heuvel, H., Kahusk, N., Liin, K., Matthiesen, M., Piperidis, S., & Vider, K. (2020). CLARIN: Distributed Language Resources and Technology in a European Infrastructure. In G. Rehm, K. Bontcheva, K. Choukri, J. Hajič, S. Piperidis, & A. Vasiljevs (Eds.), *Proceedings of the 1st International Workshop on Language Technology Platforms* (pp. 28–34). Marseille, France: European Language Resources Association.

Gao, Y., Wang, R., & Hou, F. (2023). How to Design Translation Prompts for ChatGPT: An Empirical Study. *arXiv e-prints*.

He, S. (2024). Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. *arXiv e-prints*.

Heinisch, B. (2025). Large language models for terminology work: A question of the right prompt? *Journal for Language Technology and Computational Linguistics*, *38*(2), 13–30.

Jong, F. de, van Uytvanck, D., Frontini, F., van den Bosch, A., Fišer, D., & Witt, A. (2022). Language Matters. In D. Fišer, & A. Witt (Eds.), *CLARIN: The infrastructure for language resources* (pp. 31–58): De Gruyter.

Kamocki, P., Kelli, A., & Lindén, K. (2022). The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN Infrastructure. In D. Fišer, & A. Witt (Eds.), *CLARIN: The infrastructure for language resources* (pp. 457–480): De Gruyter.

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns (New York, N.Y.)*, *2*(9), 1–12.

Kotnis, B., Gashteovski, K., Gastinger, J., Serra, G., Alesiani, F., Sztyler, T., Shaker, A., Gong, N., Lawrence, C., & Xu, Z. (2022). Human-Centric Research for NLP: Towards a Definition and Guiding Questions.

Language Resources and Evaluation (2026). Language Resources and Evaluation. https://link.springer.com/journal/10579.

Lindgren, C. A., Yunes, E., & Itchuaqiyaq, C. U. (2025). Technical Communication's Fight Against Extractive Large Language Modeling by Applying FAIR and CARE Principles of Data. *Journal of Business and Technical Communication*, *39*(1), 11–25.

Mayer, K., & Strassnig, M. (2020). The Digital Humanism Initiative in Vienna: A Report based on our Exploratory Study Commissioned by the City of Vienna. In J. Fritz, & T. Tomaschek (Eds.), *Digitaler Humanismus: Menschliche Werte in der virtuellen Welt* (pp. 27–39). [S.l.]: Waxmann Verlag GmbH.

Morozov, E. (2013). *To save everything, click here: Technology, solutionism and the urge to fix problems that don't exist*. London: Allen Lane.

Nida-Rümelin, J. (2021). Digitaler Humanismus. In U. Hauck-Thum, & J. Noller (Eds.), *Was ist Digitalität?: Philosophische und pädagogische Perspektiven* (pp. 35–38). Berlin: J.B. Metzler ein Teil von Springer Nature.

OpenEuroLLM consortium (2025). Initial Catalogue and Analytics Reports for Existing Training Datasets: Deliverable number: 3.1. https://openeurollm.eu/reports/D3_1__Initial_Training_Data_Catalogue.pdf.

Ousidhoum, N., Beloucif, M., & Mohammad, S. M. (2025). Building Better: Avoiding Pitfalls in Developing Language Resources when Data is Scarce. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8881–8894). Stroudsburg, PA, USA: Association for Computational Linguistics.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *LREC Proceedings 2012*.

Popowich, S. (2024). ChatGPT and the Commodification of Knowledge Work. *CAUT Journal*, *Special Issue: The labour activism of post-secondary education information workers*, 1–20.

Reineke, D. (2023). Terminologiearbeit mit ChatGPT & Co. *Fachzeitschrift für Terminologie*, *19*(1), 25–28.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, *31*(2), 74–87.

Terras, M., Nyhan, J., & Vanhoutte, E. (2016). *Defining digital humanities: A reader*. London and New York: Routledge.

UNESCO (2021). *UNESCO Recommendation on Open Science*. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en.

Werthner, H. (Ed.) (2022). *Perspectives on Digital Humanism*. Cham: Springer International Publishing AG.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*, 160018.

Wynne, M. (2015). Users of CLARIN - who are they? https://www.clarin.eu/blog/users-clarin-who-are-they.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power, edn. *PublicAffairs, New York*.

An infrastructure for Historical Dutch Corpus Development

Katrien Depuydt
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
katrien.depuydt@ivdnt.org

Jesse de Does
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
jesse.dedoes@ivdnt.org

Vincent Prins
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
vincent.prins@ivdnt.org

Mathieu Fannee
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
mathieu.fannee@ivdnt.org

Roland de Bonth
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
roland.debonth@ivdnt.org

Thomas Haga
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
thomas.haga@ivdnt.org

Abstract

We describe an infrastructure for linguistic annotation of historical Dutch texts, consisting of tagging and lemmatisation guidelines, gold standard data, trained tagging models, the GaLAHaD platform for automatic linguistic annotation and evaluation and the LAnCeLoT manual annotation tool. Users can upload unannotated materials to the GaLAHaD platform, where trained PoS taggers and lemmatizers (including modern deep learning models) annotate the data. Results can be evaluated against newly expanded gold standard corpora (13th to 19th centuries) with various metrics, then exported for further analysis. The LAnCeLoT tool supports manual annotation correction at both type and token level, thus supporting the development of in-domain gold standard material. The infrastructure was developed at the Instituut voor de Nederlandse Taal (INT, Dutch Language Institute) and first released in 2025.

1 Introduction

Historical texts are essential source material for both linguistic and digital humanities research. Adding linguistic annotation layers (e.g. part of speech and modern Dutch lemma) to historical text helps to make the data more accessible. These layers support linguistic analysis, and general users need not be concerned with historical spelling variation when querying and analysing the data.

Linguistic annotation of historical Dutch is far from trivial. Consider the following variants of the lemma *wereld* (”world”), taken from the central database of the INT and covering 15 centuries of Dutch language (the modern Dutch forms are in italics):

> uuerildis, uue roldi, uueroldi, uuerildi, uuerolt, uuerolde, werelde, vuerolti, werelt, weerelt, *wereld*, weerelds, wereldt, *werelden*, weereld, werrelts, waerelds, weerlyt, wereldts, vveerelts, waereld, weerelden, waerelden, weerlt, werlt, werelds, sweerels, zwerlys, swarels, swerelts, werelts, swerrels, weirelts, tsweerelds, werret, vverelt, werlts, werrelt, worreld, werlden, wareld, weirelt, weireld, waerelt, werreld, werld, vvereld, weerelts, werlde, tswerels, werreldts, weereldt, *wereldje*, waereldje, weurlt, wald, weëled

The formal variation (spelling and inflection) is abundant and is not easy to capture in patterns. An additional complication is the irregularity of word separation. What would be considered a single word from the point of view of modern Dutch can be split into several tokens (*te rugh* for modern *terug*) or vice versa, what would be written as separate words is realized as a single token (*sprackse*, *ensagh*).

Providing the tools to produce high quality annotation of historical Dutch language is therefore challenging, and it is equally challenging to do so in a such a manner that the tools are also accessible for non-technical users, and we can support metholodologically sound tool use. This involves much more

Katrien Depuydt, Jesse de Does, Vincent Prins, Mathieu Fannee, Roland de Bonth, and Thomas Haga 2025. An infrastructure for Historical Dutch Corpus Development. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 136–147. DOI: https://doi.org/10.3384/ecp222

than merely providing a processing facility. First, linguistic principles (tagset and tagging and lemmatization guidelines) need to be explicit and suitable for historical Dutch. The training and evaluation material adhering to these guidelines should be publicly available and should cover all historical periods. Moreover, without awareness of the noise and bias that the tools inevitable introduce, a methodologically sound use of these tools is hardly possible. Therefore, instead of treating the tools as black boxes (with at best a generic accuracy score on a benchmark set), users should be assisted in making informed processing tool choices and in analyzing tool performance in detail and inspecting relevant results.

Supported by additional funding of both CLARIAH$^+$ and currently SSHOC-NL, we have been developing an infrastructure, officially released in April 2025, to tackle the above-mentioned issues.

The infrastructure consists of:

- A diachronic tagset (TDN) and lemmatization guidelines for historical and contemporary Dutch
- Gold standard training and evaluation material (13th–19th century)
- The GiGaNT-Hilex historical lexicon[1]
- Trained tagger-lemmatizers (currently PIE framework and Hugging Face transformers framework)
- Detailed evaluation of PoS tagging and lemmatization to allow for a methodologically sound choice of the appropriate annotation tool
- A user-friendly platform, GaLAHaD for corpus annotation and evaluation
- LAnCeLoT, a service (web application) for creating gold standard corpus data

2 Related work

2.1 Linguistic annotation of Historical Dutch

Prior to the development of the infrastructure presented here, several tools were already available for the linguistic annotation of historical Dutch. However, efforts were fragmented, training material was scarce (even absent for some periods) and diverse in adopted tagset and annotation guidelines. A common strategy was lacking to deal with issues such as nontrivial word segmentation and the substantial degree of formal variation in spelling and inflection which is hard to classify. This became very clear in the Nederlab project[2], where an attempt was made to annotate a large diachronic corpus of Dutch from the 6th - 21st century. The quality of the linguistic annotation was insufficient, and the tagset and tagging principles, originally developed for modern Dutch, turned out to be unsuitable for historical Dutch.

We briefly discuss some of the systems that have been developed.

- The well-known Frog[3] tagger (van den Bosch et al., 2007) has been used in two ways: using the standard Dutch models after normalizing spelling to modern Dutch, cf. (Tjong Kim Sang et al., 2017), and by training specific models: a model for Middle Dutch (trained on CRM and Gysseling) has been developed.
- Adelheid[4], developed by Hans van Halteren. Adelheid is a tagger-lemmatizer for fourteenth century Dutch charters (as found in the corpus van Reenen/Mulder), with a special focus on the treatment of the extensive orthographic variation in this material, achieving about 95% accuracy in a tenfold cross-validation experiment for both tagging and lemmatization. For variation in spelling, expanded lexicon with predicted spelling variants is used. For those forms which can also not be found in this expanded lexicon, the word class is derived first, and subsequently a search for the most similar lexicon word is performed.
- The web service and application INL labs incorporated a supervised SVM-based tagger trained on the Letters as Loot corpus[5]. The GiGaNT-Hilex historical lexicon is used for lemmatization, in conjunction with a noisy-channel spelling variation model.
- PIE[6], developed by Enrique Manjavacas, Ákos Kádár, and Mike Kestemont, (Manjavacas et al.,

[1] A morphosyntactic lexicon of historical Dutch, derived from the scholarly dictionaries of historical Dutch, ONW, VMNW, MNW and WNT, cf. https://ivdnt.org/corpora-lexica/gigant/, https://gtb.ivdnt.org

[2] https://www.nederlab.nl/, a project funded by NWO (Dutch research council) that ran from 2013-2017

[3] http://languagemachines.github.io/frog/

[4] http://adelheid.ruhosting.nl/

[5] http://inl-labs.inl.nl/, no longer available

[6] https://github.com/emanjavacas/pie

2019). A deep learning approach is used. Lemmatization is approached as a string transduction task with an encoder-decoder architecture enriched with sentence context information using a hierarchical sentence encoder. They report significant improvements over the state-of-the-art when training the sentence encoder jointly for lemmatization and language modeling.

- RNNTagger[7], (Schmid, 2019). Part of speech tagging is based on bidirectional long short-term memory networks (LSTMs), with character-based word representations. Lemmatization uses an encoder-decoder system with attention.

2.2 NLP Processing services

A substantial amount of work has been devoted to developing web services and web-based interfaces for NLP tasks. Prominent examples include Stanford CoreNLP and the UDPipe web service. Some systems are centered around a specific core task, such as tagging or syntactic parsing, while also integrating auxiliary functionality, including format conversions and preliminary tasks like tokenization. Other platforms support the chaining of multiple web services, as exemplified by Weblicht[8], or facilitate tool selection based on task requirements, as with the CLARIN Language Resource Switchboard[9]. At a larger infrastructural level, the European Language Grid[10] provides a unified environment for accessing, combining, and deploying NLP services across languages. For Dutch, we mention the CLAM[11]-based Language and Speech Tools[12] hosted by the Centre for Language and Speech Technology at Radboud University.

None of the discussed platforms directly supports Historical Dutch, or supports an infrastructure integrating processing, gold-standard data creation, corpus building, and evaluation.

2.3 Tools for manual annotation of corpus data

There are various environments for manual corpus annotation available. We mention the INCEpTION platform[13] (Klie et al., 2018), the FoLiA Linguistic Annotation Tool FLAT[14], and the Arborator Grew[15] system, with which we have in common that it is part of an effort to create an infrastructure for a research community.

The main reason we have developed a separate tool for manual annotation tasks is that we found the purely token-based workflow in the available tools lacking in efficiency. The main features of LAnCeLoT - quickly reviewing all occurrences of a word, and bulk assignment of annotations turned out to be indispensable for efficient corpus processing.

3 Requirements

3.1 Support the research community

As mentioned in the introduction, the availability of processing tools is just a part of an ecosystem that supports historical corpus research. We aim to enable researchers from the Humanities and Social Sciences to make use of corpus data and linguistic annotation tools in a methodologically sound manner, supporting data formats prevalent in these communities. More specifically, we need to support several scenario's in which users are enabled to:

- choose the right tagger for a dataset, with or without having relevant gold standard data available.
- develop gold standard data for evaluation and/or training.
- evaluate taggers on gold standard data.
- inspect the tagging results.
- annotate data while preserving document structure annotation.

[7] https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/
[8] https://weblicht.sfs.uni-tuebingen.de/weblicht/
[9] https://www.clarin.eu/content/language-resource-switchboard
[10] https://live.european-language-grid.eu/
[11] https://proycon.github.io/clam/
[12] https://webservices.cls.ru.nl/portal/
[13] https://inception-project.github.io/
[14] https://github.com/proycon/flat
[15] https://arborator.grew.fr/

This translates to the following requirements:

- Ensure that high quality tools are available.
- Develop clear and documented annotation guidelines, tested on a substantial amount of historical corpus data.
- Support a methodologically sound approach to historical corpus processing and analysis by providing extensive evaluation functionality.
- Support the development of gold standard data.
- Make benchmark data available to support tagger choice.
- Provide training data for tagger development for all stages of historical Dutch.
- Support digital humanities research by keeping the XML-encoding (e.g. TEI) of the uploaded text intact.

Finally , in terms of technical requirements, the following should be met: 1. Simplicity: do not implement a more complex pipelines than strictly necessary. In our experience, complex pipelines are sensitive to errors and results are not transparent. For this reason, we have not included BlackLab (de Does et al., 2017)[16] corpus creation as a step in the GaLAHaD application. 2. Robustness for larger uploads. 3. Separation of backend (web service with a documented API) and user interface.

4 Tagset, tagging guidelines and lemmatisation principles

There are several tagsets for linguistic annotation of text corpora containing historical and contemporary Dutch, such as the tagset of GiGaNT (Groot Geïntegreerd lexicon van de Nederlandse Taal, Ruitenberg et al., 2012), the corpus tagset CGN/D-Coi (Van Eynde, 2005), the tagset used in both the Corpus Gysseling (van Dalen Oskam and Depuydt, 2000) and the Corpus Van Reenen-Mulder, and the tagset for the Corpus Oudnederlands[17]. These tagsets differ in the method of tagging (lexical or functional), in naming (what is called A in one tagset is called B in another) and in degree of detail (which features are tagged). Due to these differences, it is complicated to compare linguistically annotated corpora with one other.

In order to make it possible to study language usage throughout the centuries, we have created a new tagset. This Tagset for Diachronic corpus data of Dutch (Tagset voor Diachroon corpusmateriaal van het Nederlands: TDN, Haga et al., 2024) was developed as a result of the re-evaluation of the GiGaNT tagset and after careful analysis of the above-mentioned corpus tagsets.

A core requirement for its development was that the TDN would be mappable to the existing tagsets, that TDN should be applicable to all stages of Dutch language - from the sixth century to the present day - and that a core tagset should be defined to allow production of large amounts of training and evaluation data. The proposed tagset was submitted to a large number of (historical) linguists who critically read along and provided valuable comments. The core tagset is applied in the gold standard corpora, cf. section 5.

The tagset was compiled on the basis of a number of design principles.

- TDN should enable both coarse tagging (main PoS and lemma) and detailed annotation, including the annotation of difficult features such as case. By accommodating different levels of linguistic detail, we ensure that projects with distinct objectives can nonetheless benefit from one another's efforts through a shared common core.
- TDN is based on words, not tokens. A token is defined here as a string of characters separated from other tokens by a space (and possibly a punctuation mark). Some tagsets apply the principle that one token corresponds to one part of speech and one lemma. This principle cannot be applied to historical language material. In older language stages, words sometimes include whitespace, while tokens can consist of multiple words (as in the case of clitic combinations).
- TDN is based on *functional tagging* (as opposed to *lexical tagging*). Lexical tagging aims to assign the same part of speech to what is considered the same word across different contexts, typically using a dictionary or lexicon as a reference. Functional tagging, by contrast, assigns part of speech based on a word's syntactic role in a specific context. Its main advantage is that a word form can

[16] https://blacklab.ivdnt.org/

[17] https://corpusoudnederlands.ivdnt.org

be tagged according to actual usage, without making claims about how lexicalized that usage is. For historical language, decisons on lexicalisation can only be made after extensive analysis of annotated corpus data.

- TDN should enable mapping to other tagsets. Where tagsets differ in approach (e.g. lexical versus functional), it should in principle be possible to use the information available from TDN, with or without the aid of lexicon information, to translate it into other tagsets, such as Corpus Gysseling/CRM tagset, CGN/D-COI, Universal Dependencies.

The assignment of a modern Dutch equivalent to historical Dutch language data is based on etymological criteria. Words that do not occur in Modern Dutch any more are also assigned a Modern Dutch equivalent. The lemmatisation principles are described in Depuydt et al., 2024.

5 Gold standard corpora

The main gold standard corpora of historical Dutch available before the project were: Corpus of Old Dutch[18], 500–1200; Corpus Gysseling[19], 1200–1300; Corpus van Reenen-Mulder[20], 1300-1400; Corpus Letters as Loot[21], 1661-1783. Obviously, there were significant gaps in the coverage of these corpora. An additional complication was the fact that three different tagsets were used with different leves of detail. The annotation of the Corpus of Old Dutch is the most detailed and provides fine-grained features like case for nouns and mood for verbs, which were ommitted for reasons of feasibility in Gysseling and CRM. The Letters as Loot Corpus provides only the part of speech label. We have taken samples from Gysseling, CRM, and Letters as Loot corpora and have adapted the lemmatisation and part of speech tagging to the principles that were defined. We list the training corpora developed in table 1, and the distribution of the material per period in table 2.

Corpus	**Tokens**	**Tokens (%)**	**Documents**
total	*863,033*	*100.00%*	*13,788*
crm[22]	123,913	14.36%	597
gentse-spelen[23]	87,913	10.19%	1
gysseling-literair[24]	73,195	8.48%	29
gysseling-ambtelijk	66,853	7.75%	195
kranten-19[25]	63,279	7.33%	27
letters-as-loot[26]	63,227	7.33%	112
dictionary-quotations-14[27]	53,058	6.15%	3,191
dictionary-quotations-15[28]	42,537	4.93%	2,316
dictionary-quotations-16[29]	46,034	5.33%	1,838
dictionary-quotations-17[30]	46,287	5.36%	1,919
dictionary-quotations-18[31]	47,485	5.50%	1,794
dictionary-quotations-19[32]	35,006	4.06%	1,550
couranten[33]	29,674	3.44%	119
clvn[34]	27,204	3.15%	88
dbnl-excerpts-15[35]	15,460	1.79%	3
dbnl-excerpts-16	11,714	1.36%	3
dbnl-excerpts-17	10,126	1.17%	2
dbnl-excerpts-18	10,067	1.17%	2
dbnl-excerpts-19	10,001	1.16%	2

Table 1: Distribution of tokens per corpus

[18] https://taalmaterialen.ivdnt.org/download/corpus-oudnederlands-online/
[19] https://taalmaterialen.ivdnt.org/download/tstc-corpus-gysseling/
[20] https://www.middelnederlands.nl/corpora/crm14/
[21] https://taalmaterialen.ivdnt.org/download/tstc-bab-gouden-standaard/

Period	Tokens	Tokens (%)	Documents
total	863,033	100.00%	13,788
1200–1300	140,048	16.23%	224
1300–1400	176,971	20.51%	3,788
1400–1500	52,538	6.09%	2,318
1500–1600	171,277	19.85%	1,929
1600–1700	87,675	10.16%	2,041
1600–1800	63,227	7.33%	112
1700–1800	57,552	6.67%	1,796
1800–1900	113,745	13.18%	1,580

Table 2: Overview of gold standard material by period

In order to fill the gaps, we have extended the available material. Our aim was to ensure that at least 100,000 tokens per century are available, tagged according to the defined lemmatisation principles and the "core" level of the TDN tagset. This has been achieved for all centuries[36] but the fifteenth. Cf. table 2.

To the material that was already annotated in some form we added a small set of excerpts from the DBNL digital library, several per-century sets of dictionary quotations from the scholarly historical dictionaries of Dutch, a sample set from the corpus of Late Middle and early Modern Dutch (CLVN) and the Couranten Corpus, the completely tagged Gentse Spelen collection, and a set of 19th century newspaper issues based on KB (National Library of the Netherlands) data.

6 Linguistic annotation: the taggers/lemmatizers

Currently, GaLAHaD provides tagging and lemmatization models for the PIE framework (slightly adapted for the GaLAHaD platform) and for the transformer-based taggers discussed in the next subsection.

Transformer-based tagging with pretrained language models

In order to benefit from advances in deep learning, we have developed a tagger-lemmatizer based on the BERT token classifier, implemented using the Hugging Face framework[38], and the GysBERT historical Dutch language model[39]. The lemmatizer uses the GiGaNT-Hilex historical lexicon, and a ByT5 (Xue

[22]Corpus van Reenen-Mulder, selection

[23]Cf. https://www.dbnl.org/tekst/_gen001gent01_01/_gen001gent01_01_0004.php

[24]Corpus gysseling, selection

[25]19th century newspaper issues

[26]The Letters as Loot / Brieven als Buit-corpus. Leiden University. Compiled by Marijke van der Wal (Programme leader), Gijsbert Rutten, Judith Nobels and Tanja Simons, with the assistance of volunteers of the Leiden-based Wikiscripta Neerlandica transcription project, and lemmatised, tagged and provided with search facilities by the Institute for Dutch Lexicology (INL). 3rd release januari 2021. http://hdl.handle.net/10032/tm-a2-s4

[27]Dictionary quotations (MNW: Dictionary of Middle Dutch, https://ivdnt.org/woordenboeken/historische-woordenboeken/middelnederlandsch-woordenboek/), 14th century

[28]Dictionary quotations (MNW)

[29]Dictionary quotations (MNW)

[30]Dictionary quotations (WNT: Dictionary of the Dutch Language, https://ivdnt.org/woordenboeken/historische-woordenboeken/woordenboek-der-nederlandsche-taal/)

[31]Dictionary quotations (WNT)

[32]Dictionary quotations (WNT)

[33]Couranten Corpus (version 2.0) (July 2025) [Online Service]. Available at the Dutch Language Institute: https://hdl.handle.net/10032/tm-a3-c2.

[34](Corpus Laatmiddel- en Vroegnieuwnederlands, van der Sijs et al., 2018

[35]DBNL excerpts, 19th century

[36]All available data for the Old Dutch stage (before 1200) is included in the Corpus Of Old Dutch[37], which is tagged manually. Currently, there is no Old Dutch material for which automatic tagging would make sense.

[37]https://corpusoudnederlands.ivdnt.org/

[38]https://github.com/INL/int-huggingface-tagger

[39]https://huggingface.co/emanjavacas/GysBERT

et al., 2022) pre-trained byte-to-byte model, finetuned on data from this lexicon, as a fallback for out-of-vocabulary words.

Annotation results

We list the evaluation results of some of the subcorpora below. Results are taken from the GaLAHaD online application version of January 30, 2026. The *pie-* label refers to PIE trained on different datasets (*all* for the complete training set, *1640-1600* and *1600-1900* refer to the combination of all training sets belonging to these periods); likewise *hug-* refers to the transformer-based taggers. The label *-enhanced* refers to hidden tag enhancements used internally, which are used to connect tokens which are considered parts of the same word by the TDN guidelines.

CLVN

PoS

tagger	macro f1	micro accuracy
hug-tdn-1400-1600	0.52	0.93
hug-tdn-all-enhanced	0.53	0.92

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.64	0.86
pie-tdn-all	0.55	0.86

Couranten Corpus

PoS

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.70	0.96
hug-tdn-1600-1900	0.70	0.95

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.72	0.92
pie-tdn-all	0.60	0.89

Letters as Loot

PoS

tagger	macro f1	micro accuracy
hug-tdn-all	0.72	0.91
hug-tdn-1600-1900	0.67	0.91

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.56	0.85
pie-tdn-all	0.48	0.82

WNT Dictionary quotations, 19th century

PoS

tagger	macro f1	micro accuracy
hug-tdn-all	0.64	0.97
hug-tdn-1600-1900	0.65	0.97

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.85	0.96
hug-tdn-1600-1900	0.81	0.94

7 GaLAHaD: Generating Linguistic Annotations for Historical Dutch

The platform serves two purposes. One is to make annotation and tool evaluation easily accessible to researchers, the other to make it easy for developers to contribute their tools and models in the platform, and thus compare them to other tools with gold standard material included in the platform.

Apart from the basic task of uploading and annotating corpus material, *GaLAHaD*[40] is designed to enable end users to choose the optimal path for their material. The platform provides options to inspect and evaluate the result of the annotation process, in order to raise the awareness of typical errors and biases in the tools. The functionality of comparing annotation layers enables users to assess the accuracy of different tools on their data. It can be used both to evaluate a layer added by an automatic tagger with respect to a gold standard reference layer, or to compare layers added by different taggers. Disagreement between layers is not only represented by global statistics, but also illustrated by examples which are immediately visible in the tool. Different metrics and targeted evaluations for certain token types may lead to different conclusions as to the path of choice, cf. fig. 7, where the evaluation of clitic combinations with macro F1 (on the right) contradicts the rosy picture of the standard micro accuracy metric on the left. The resulting annotated material can be uploaded to the *Autosearch* corpus exploration environment and to the *LAnCeLoT* tool for manual correction of linguistic annotation.

For tool developers, the docker-based application architecture[41] ensures easy contribution of tools to the platform. The application and taggers are hosted by the INT and are accessible with any CLARIN account. There is also the option to self-host an instance using the publicly available docker images from the INT docker hub or the open source code available on GitHub.

[40] https://portal.clarin.ivdnt.org/galahad/, https://github.com/INL/galahad

[41] https://github.com/INL/galahad-taggers-dockerized

Annotation: PoS | Group by: PoS | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.73	0.69	0.70	0.96	Details
hug-tdn-1600-1900	0.72	0.69	0.70	0.95	Details
hug-tdn-all	0.71	0.69	0.70	0.95	Details
hug-tdn-1400-1600	0.52	0.51	0.51	0.92	Details
pie-tdn-all	0.72	0.70	0.70	0.92	Details

Annotation: PoS | Group by: PoS | Single/multiple analysis: Multiple

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
pie-tdn-all	0.44	0.43	0.44	0.66	Details
hug-tdn-1600-1900	0.42	0.37	0.39	0.66	Details
hug-tdn-all-enhanced	0.43	0.37	0.38	0.64	Details
hug-tdn-all	0.40	0.37	0.38	0.64	Details
hug-tdn-1400-1600	0.23	0.22	0.22	0.57	Details

Figure 1: "Best" *Couranten corpus*[42] tagger depending on evaluation type and token type

8 LAnCeLoT: Linguistic Annotation Corpus Laundry Tool

LAnCeLoT is an online tool[43] (service) to manually verify and correct corpora that are linguistically annotated with part of speech and lemma. The core characteristic of the tool is that manual verification is not done token by token in running text, but at both the type and token level. Users are presented with a list of types with the accompanying annotations occurring in the corpus. For each type, the corresponding concordances can be listed to enable verification of the analyses and their correction at token level. When working at token level, one can make a subselection of tokens (concordances) and assign a specific annotation to every instance in this subselection all at once. An example of this bulk annotation can be found in figure 3: the six occurrences type of the *accoord* can be simultaneously tagged and verified in a single action. As additional support to the annotator, suggestions for possible analyses are provided for each type using the historical computational lexicon GiGaNT-HILEX[44], which covers Dutch from ca. 1250 to 1976. The current version of GiGaNT-HILEX contains the lexicon modules based on the Dictionary of the Dutch Language (Woordenboek der Nederlandsche Taal, WNT) and the Dictionary of Middle Dutch (Middelnederlandsch Woordenboek, MNW) . GiGaNT Hilex follows the linguistic principles for part of speech tagging (TDN) and lemmatisation as discussed in section 4.

LAnCeLoT works as a standoff annotation tool on data indexed by BlackLab, the api of which is used both to construct the PostgreSQL database in which the corrected annotations are stored and to retrieve the concordance for display. Therefore, in order to use LAnCeLoT, the corpus that needs manual verification has to be uploaded in LAnCeLoT Search, a customised version of the AutoSearch application which allows nontechnical linguistic researchers to index and search their own data with BlackLab as a corpus search engine. One can navigate from a concordance to LAnCeLoT Search to present more context when this is necessary to assign the correct annotation to a certain token. LAnCeLoT Search can also be used to search within the automatically annotated corpus. For both import and export, LAnCeLoT currently supports the TEI p5 format used by GaLAHaD.

[42] https://couranten.ivdnt.org
[43] It is a successor of CoBaLT (Kenter et al., 2012)
[44] https://ivdnt.org/corpora-lexica/gigant/

9 Technical infrastructure architecture

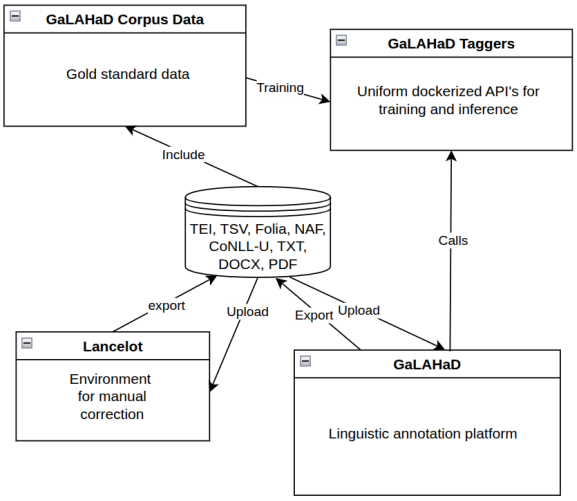

Figure 2: Infrastructure architecture

GaLAHaD consists of a Kotlin Spring Boot backend, a REST backend with Swagger UI API documentation, and a Vue 3 frontend. The backend supports the parsing of various file formats via Aalto XML[45] and other parsers, enabling the extraction of plain text, lemmas, and POS annotations. Parsed data and corpus metadata is stored on disk as JSON, together with the original files. The backend communicates with various dockerized tagger-lemmatizers that turn plaintext into annotated TSV. This protocol has been standardised as *GaLAHaD Taggers Dockerized*[46]. Currently, all taggers are implemented in Python and use Bottle for their web service, but any web service wrapper implementing the simple communication protocol suffices.

The backend is then capable of either exporting these newly produced annotations to various formats, or merging it with the original encoding of the document. As mentioned in section 7, annotation layers of various taggers and the source document can be compared and evaluated. As token-level evaluation can be computationally expensive on large corpora, GaLAHaD uses Ben Manes' Caffeine cache[47] and stores intermediate document results as JSON to disk. As documents are processed independently, multiprocessing is used to improve performance.

LAnCeLoT is built on the Lex'it (Lexicon Interactive Tool)[48] database application development platform, the BlackLab[49] corpus search engine, a Java Jakarta RESTful Web Service, and the PostgreSQL database. File uploading and corpus querying is handled by BlackLab. TEI files exported by GaLAHaD can be uploaded in the LAnCeLoT Search environment, a reskin of the BlackLab frontend. Annotations are corrected in the Lex'it frontend, which is extended for LAnCeLoT using JavaScript/jQuery. Via its Jakarta web service, LAnCeLoT keeps track of lemma and POS annotations that have been corrected or validated in its work table in the PostgreSQL database, and can export these by merging them with the original TEI encoding.

For creating our GaLAHaD gold standard corpus data, we use a couple of python scripts to detect duplicates, split data into machine learning sets, and generate statistics. Once a corpus is exported from LAnCeLoT, it is converted to TSV using the GaLAHaD webservice. We provide both TEI and TSV, as TSV is the preferred data format for machine learning.

Next we check for files with textual overlap. These may be the A and B version of a manuscript, or two news articles in different newspapers based on the same correspondence source. Overlaps are detected

[45] https://github.com/FasterXML/aalto-xml
[46] https://github.com/instituutnederlandsetaal/galahad-taggers-dockerized
[47] https://github.com/ben-manes/caffeine
[48] https://github.com/instituutnederlandsetaal/Lexit
[49] https://blacklab.ivdnt.org/

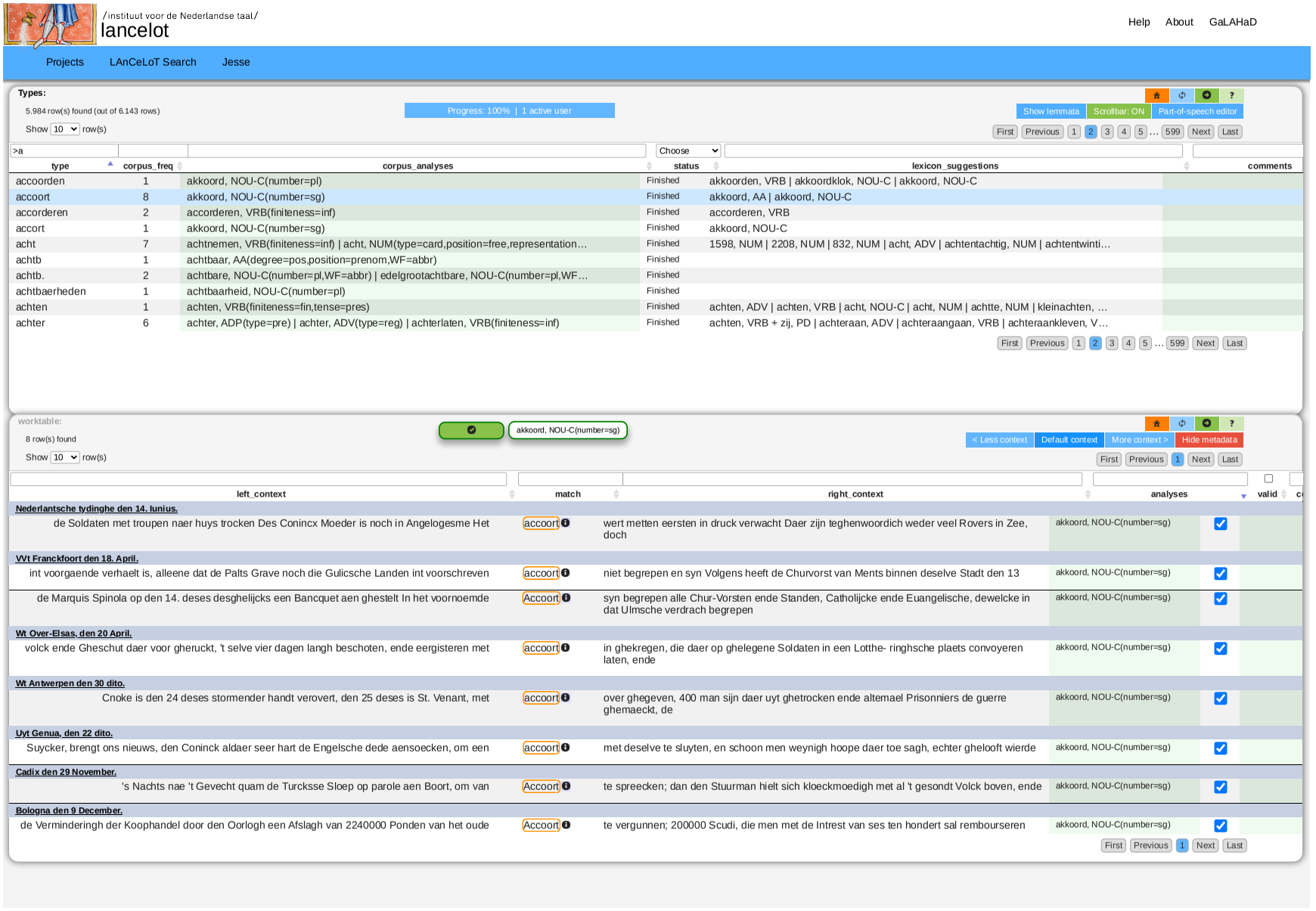

Figure 3: Lancelot: bulk ta ; Couranten training corpus

based on edit distance using the Edlib[50] library. A text is first converted to a single string of alphabetical characters (that includes removing spaces). The string is then split into chunks of 20 characters by default. For each chunk, we then check if it occurs in any other document, in any position, within an edit distance of 5. If a high percentage (e.g. 90%) of all chunks in a document occurs in another document as well, the two are considered duplicates.

The corpus is then split into a train, test, and validation set such that any duplicates are grouped into the same set to avoid bias. Provenance of the files per split is stored in a JSON file. Finally, we calculate some statistics about the corpus, such as token size per split, century, and subcorpus, as well as the frequencies of each lemma and POS annotation. We also try to detect suspicious annotations: annotations that are unlikely to occur. If a certain token is assigned a POS of NOU-C a thousand times, and VRB only once, chances are it was tagged incorrectly. This feedback is then processed by the annotators.

10 Future work

In the SSHOC-NL project, we are working on extending the platform with more linguistic annotation tools and evaluation methods.

We are extending the annotation tools to modern Dutch by integrating spaCy and Stanza. Simultaneously, we are adding support for annotations other than lemma and PoS, namely named entities and dependency relations as they are produced by spaCy and Stanza models.

The existing evaluation methods will be extended to support the new annotation types. Additionally, we are working on new evaluation methods for span-based annotations (e.g. named entities), evaluation metrics grouped by frequency (e.g. hapax accuracy), and document-level evaluation (e.g. text-inline visualisation of errors).

We will extend the gold standard data with data from the Corpus of Middle Dutch[51] and contemporary corpus data.

Of course, we hope that the state of the art in historical language processing will continue to improve to yield better taggers, lemmatizers and parsers to be integrated in the infrastructure. Working with the spaCy, Stanza and UDPipe models for modern Dutch, we found that the lemmatization accuracy of these models is often below par for less frequent lemmata, and can be improved by using the GiGaNT-Molex[52] morphosyntactic lexicon of modern Dutch. We will develop a lemmatizer that incorporates the lexicon.

11 Acknowledgments

The team behind the infrastructure consists, besides the authors, of Vincent Prins, Roland de Bonth, Tim Brouwer, Mathieu Fannee and Thomas Haga, all INT. Since 2024 Bram Vanroy (INT), Eleanor Smith and Antske Fokkens from the VU have also been involved in the further development of GaLAHaD. The work is carried out with funding from the Netherlands Organisation for Scientific Research (NWO) in the CLARIAH-PLUS project (Grant 184.034.023) and the SSHOC-NL project (Grant 184.036.020) , with further support from the Dutch Language Union.

References

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. Ubiquity Press.

Depuydt, K., Haga, T., & Mooijaart, M. (2024). *Lemmatiseerprincipes voor GiGaNT, het centrale lexicon van het INT*. https://ivdnt.org/wp-content/uploads/2024/11/lemmatiseerprincipesV2_combi.pdf

Haga, T., Depuydt, K., de Does, J., de Bonth, R., & Geirnaert, D. (2024). *Tagset voor Diachroon corpusmateriaal van het Nederlands (TDN)*. https://ivdnt.org/wp-content/uploads/2024/11/TDNV2_combi.pdf

[50] https://github.com/Martinsos/edlib

[51] https://corpusmiddelnederlands.ivdnt.org/

[52] https://taalmaterialen.ivdnt.org/download/gigant-molex2-0/

Kenter, T., Erjavec, T., Žorga Dulmin, M., & Fišer, D. (2012, April). Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In K. Zervanou & A. van den Bosch (Eds.), *Proceedings of the 6th workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 1–6). Association for Computational Linguistics. https://aclanthology.org/W12-1001/

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation [Event Title: The 27th International Conference on Computational Linguistics (COLING 2018)]. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.

Ruitenberg, T., van pellicom K., de Does, J., & Depuydt, K. (2012). De morfosyntactische module van het GiGaNT-lexicon. *INL Working Papers - Taalbank Nederlands*.

Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 133–137.

Tjong Kim Sang, E., Bollmann, M., Boschker, R., Casacuberta, F., Dietz, F., Dipper, S., Domingo, M., van der Goot, R., van Koppen, M., Ljubešić, N., et al. (2017). The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, *7*, 53–64.

van Dalen Oskam, K., & Depuydt, K. (2000). Lemmatisering en codering in het VMNW-corpus II Codering in het VMNW-woordenboek. *nstituut voor Nederlandse Lexicologie, Internal report*.

van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series*, *7*, 191–206.

van der Sijs, N., van Kemenade, A., & Rem, M. (2018). Corpus laatmiddel- en vroegnieuwnederlands (clvn) (onderdeel Nederlab).

Van Eynde, F. (2005). Part of Speech tagging en lemmatisering van het D-COI corpus. *Centrum voor Computerlinguïstiek K.U.Leuven. Report.*

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models (B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, *10*, 291–306. https://doi.org/10.1162/tacl_a_00461

Users' Experience and Development Priorities for CLARIN.SI

Špela Arhar Holdt
Faculty of Computer and Information Science & Faculty of Arts, University of Ljubljana, Slovenia
arharhs@ff.uni-lj.si

Katja Meden and **Taja Kuzman Pungeršek**
Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
{katja.meden,taja.kuzman}@ijs.si

Nikola Ljubešić
Jožef Stefan Institute; Faculty of Computer and Information Science, University of Ljubljana; Institute of Contemporary History, Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Tomaž Erjavec
Jožef Stefan Institute, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

This paper reports the results of a user survey evaluating the services, tools, and broader functioning of the CLARIN.SI research infrastructure. The survey was designed to identify areas for improvement and to support development planning and strategic decision-making for the coming years. The questionnaire included 50 questions across eight thematic sections, covering the full range of user engagement with the infrastructure, from depositing and downloading resources to using tools and attending events. It was distributed to existing and potential users in Slovenia and abroad. Based on 228 responses, the results show that CLARIN.SI is generally highly valued by its community, especially for the reliability of its data, its support for open science, and the usefulness of its resources. The findings also point to a highly interdisciplinary user base and to a substantial group of potential users who have not yet directly engaged with the infrastructure. Although repository services and language tools are widely used and positively evaluated, respondents consistently identified challenges related to resource discovery, metadata clarity, and search functionality. More broadly, the results suggest that the future impact of CLARIN.SI will depend not only on technical development, but also on effective outreach, onboarding, and training, including the provision of concrete use cases and reusable educational materials.

1 Introduction

CLARIN.SI, the Slovenian node of the CLARIN infrastructure, recently celebrated its 10^{th} anniversary. Since its establishment, the infrastructure has engaged in a range of activities, both national and international. An overview of its organisation, main areas of activity, and impact on the research community is presented in Erjavec et al. (2025).

To mark this milestone and support planning for the next decade, we set out to gather user feedback that would highlight what works well, what needs improvement, and how we might reach new users who are not yet familiar with the infrastructure. With these goals in mind, we designed a survey, focusing on the quality of the tools, resources, and services offered by CLARIN.SI. To engage both current and potential users, we shared the survey not only within our usual networks but also more broadly, aiming to reach as wide an audience as possible, both in Slovenia and abroad.

Beyond its immediate role in informing the future development of CLARIN.SI, the survey can also be situated within broader practices of research infrastructure evaluation. In distributed infrastructures such as CLARIN, user feedback is important not only for improving services, but also for setting priorities,

Špela Arhar Holdt, Katja Meden, Taja Kuzman Pungeršek, Nikola Ljubešić, and Tomaž Erjavec 2025. Users' Experience and Development Priorities for CLARIN.SI. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 148–160. DOI: https://doi.org/10.3384/ecp222

planning sustainable development, and aligning infrastructure provision with the needs of its research communities. From this perspective, the present study offers a structured account of user views on the accessibility, usefulness, and future development of CLARIN.SI services. It may therefore be of interest not only in the Slovenian context, but also to other CLARIN centres and related digital research infrastructures seeking to support planning and evaluation through direct user consultation.

It is noteworthy that comparable studies, at the moment, seem relatively scarce. Italy reported on a survey (Nicolas et al., 2018) carried out only two years after the establishment of CLARIN-IT, which was restricted to a sample of Italian digital humanists interested in ancient Greek philology. While the survey is not presented in detail, the main conclusions were that most of the available resources did not correspond to users' requirements which are centred around authoring, editing, indexing and presenting a digital edition, and the ability to link the available resources and improve them, i.e. they would need tools to integrating textual data and bibliography links. This survey thus targeted a very specific user community, which is well outside the focus of CLARIN.SI; in Slovenia, publishing digital editions is, by agreement, left to the DARIAH-SI infrastructure, while CLARIN.SI concentrates on corpus download and analysis. The other survey was carried out in Norway (De Smedt, 2020), and, while the questions were somewhat more limited than our survey, the ones given were similar to ours, except for services offered by CLARINO rather than CLARIN.SI. As will be seen in the conclusions, their summary is quite similar to ours: CLARINO had, at the time, satisfied to very satisfied users, although some had unrealistic expectations from the infrastructure. It also appeared that some did not realize what CLARINO is, with the lesson that better information to the target user group is at least as important as maintaining and improving the infrastructure services.

A related landscape survey was conducted by the Humanities and Cultural Heritage Open Science Cloud (H2IOSC), featuring the collaborative involvement of Italian ESFRI research infrastructures (CLARIN, DARIAH, E-RIHS, and OPERAS). The survey questionnaires represent only part of their overall monitoring activities, which also include interviews, focus groups, and landscape mapping, among others (Sichera et al., 2025). While the questionnaire design consists of at least somewhat comparable sections, the results pertain to the overall state of the H2IOSC federation, targeting a more diverse and heterogeneous community and providing an overview of community behaviours and needs, rather than direct feedback. The results (Luzietti et al., 2025; Monachini et al., 2025) indicate stable awareness and involvement among users from diverse disciplinary fields, with a wide selection of tools and services used in their work and strong interest in access to archives and repositories. The results are therefore less comparable with our own survey; however, the work and activities accompanying the survey could be adapted for CLARIN.SI and other Slovenian research infrastructures for in-depth landscape analysis.

2 Method

User opinions were collected through an online survey written in Slovenian and English language, designed and administered using the 1KA survey platform. The questionnaire consisted of 50 questions divided into 8 thematic sections. To structure the survey, we first catalogued the full range of of CLARIN.SI's existing tools and services, and organized the sections around key anticipated user activities, such as depositing or downloading resources, using tools, and attending events. An additional section addressed potential future activities, focusing on areas not yet covered by the infrastructure. To allow for a better understanding of the structure and rationale behind the survey, the complete questionnaire and anonymized survey data are provided at Arhar Holdt et al. (2026).

The survey began by collecting basic participant information and identifying which CLARIN.SI tools and services, if any, they had experience with. Based on their responses, participants were directed to relevant subsections where they could evaluate specific aspects of the infrastructure. Each section also included an open-ended question that invited participants to provide more information on what they felt was missing, allowing for broader, unanticipated input.

The Slovenian survey ran from 28 February to 1 April 2025, while its English counterpart was open from 10 November to 10 December 2025. The surveys were disseminated through academic, educational, and professional networks in Slovenia and Europe judged to have members who are current or

potential CLARIN.SI users. According to the 1KA platform statistics, 895 individuals entered the survey introduction, 296 proceeded to the first page, 235 started responding, 228 submitted at least a partial response, and 169 completed the questionnaire. We therefore analyse 228 valid partial or complete responses. The majority of these are from the Slovenian version (170 respondents), and 58 (25%) from the English version of the survey. The average completion time was 11 minutes and 2 seconds. Although the survey included prompts encouraging participants to answer all questions, they were allowed to skip items, so the number of responses (denoted below by N) varies by question.

The survey data were analysed using descriptive statistics, with a focus on aggregate distributions rather than inferential comparisons. This reflects the exploratory aim of the study, namely to map user profiles, experiences, and development priorities relevant to CLARIN.SI rather than to test predefined hypotheses. The findings below are therefore presented holistically, with occasional identification of differences between the Slovenian- and English-language surveys; these differences should be interpreted as indicative rather than statistically confirmed (see also Limitations). For the open-ended questions, we applied content analysis using a bottom-up, inductive approach (Krippendorff, 2018). Responses were coded using generalized words or phrases reflecting the key concerns, observations, and development priorities identified by participants. These results are summarised in the corresponding sections.

3 Results

3.1 The Participants

The affiliations (N=226) of participants was mainly with academic and research institutions (84%). 43% worked or studied at a higher education institution, and 41% reported working in a research institution. Some were employed in private or public companies (6%), public administration (3%), primary or secondary schools (2%), cultural institutions (1%), and non-profit organizations (1%). An additional 2% selected "Other," for instance, media houses and libraries. Interestingly, the distribution of affiliations differs significantly between Slovenian and international survey participants. Slovenian participants are primarily affiliated with research institutions (47%), followed by higher education institutions (35%). In contrast, most international participants come from higher education institutions (67%), with fewer from research institutions (24%). The results also point at a more diverse profile of Slovenian users in comparison to foreign users.

In terms of professional roles (N=226), over half of the respondents (54%) were engaged in research, including doctoral and postdoctoral work. Others were students (16%) or educators (11%) at different education levels, or involved with technical and development work (8%), administrative tasks (4%), and management (3%). The remaining 4% specified other roles, such as translation or language editing. Disciplinary backgrounds (N=227) were mainly in the humanities (59%), engineering and technology (18%), social sciences (15%), followed by natural sciences (7%), and medicine (1%).

Most participants (N=225) reported they conduct language-related work, however, 10% reported no language-related activities. Their main fields of work (multiple answers possible) were language technologies (38%), linguistic description (36%), natural language processing (28%), translation (20%), theoretical linguistics (20%), and language teaching (19%). Other fields, such as language standardization (13%), history (5%), literary studies (4%), journalism (4%), sociology (2%), library and information science (2%), law and criminology (2%) and political science (1%), were also represented. More Slovenian than international survey participants are active in language standardization (14% vs. 9%). In contrast, international participants are significantly more involved in linguistics: 35% are active in theoretical linguistics (compared to 14% in Slovenia), 55% in language description (30% in Slovenia), and 28% in language teaching (16% in Slovenia). Additionally, 5% of international participants come from political science, a field not represented among Slovenian respondents.

Figure 1 presents the distribution of disciplinary backgrounds among survey participants working in the fields of natural language processing (NLP) and language technologies (LT). Contrary to the common assumption that these areas are primarily rooted in machine learning and therefore dominated by engineering and computer science, the survey results reveal a strong presence of researchers from the humanities. In the case of natural language processing, the share of participants with a humanities back-

ground is equal to that of participants from engineering and technology disciplines. For language technologies, this trend is even more pronounced, with more than half of the survey participants identifying a humanities background. Overall, the figure highlights the strongly interdisciplinary nature of both fields and underscores the substantial role that researchers with a background in humanities play within the CLARIN.SI user community.

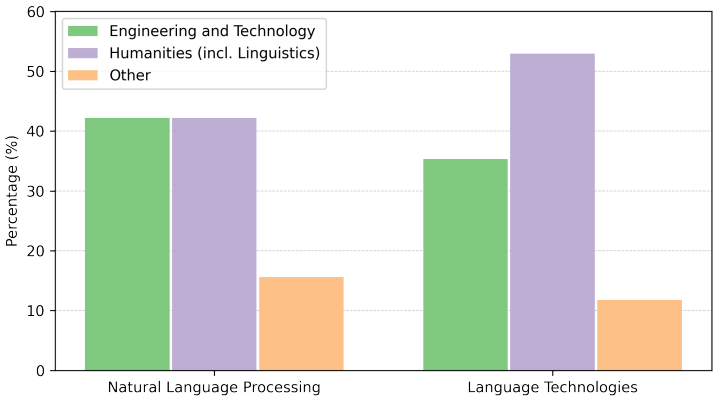

Figure 1: The distribution of disciplinary backgrounds among survey participants working in the fields of natural language processing (NLP) and language technologies (LT).

The participants (N=227) were diverse in age, with the largest groups aged 45–54 (28%), 35–44 (25%), and 25–34 (21%). Smaller segments included respondents aged 55–64 (13%), 18–24 (9%), 65 or older (4%), and two respondents under 18. 60% of the participants were women, 34% were men, 3% another gender, and 3% preferred not to disclose.

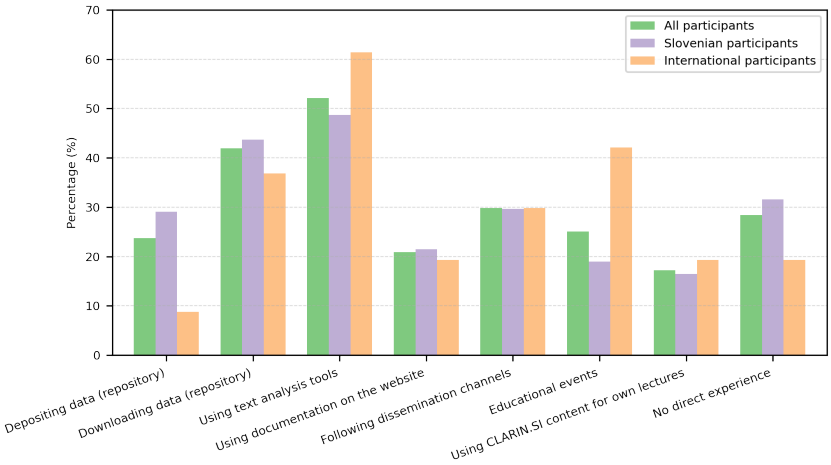

Figure 2: Reported experience of survey participants with the CLARIN.SI activities and services (multiple responses allowed).

The participants (N=215) reported their experience with various components of the CLARIN.SI infrastructure (multiple answers were possible). As shown in Figure 2, 24% had deposited resources, and 42% had downloaded resources from the CLARIN.SI repository. 52% had experience with the CLARIN.SI concordancers and text analysis tools. 21% used documentation on the CLARIN.SI website, 30% followed CLARIN.SI news through website or social media, 25% had attended training or events about CLARIN.SI, and 17% had integrated CLARIN.SI into their own teaching or presentations. Nearly one-third of respondents (28%) reported no direct experience with the infrastructure.

When comparing Slovenian and international participants, the former have more often deposited resources in the CLARIN.SI repository (29% vs. 9%), as well as downloaded and used data from the

repository (44% vs. 37%). In contrast, a greater share of international participants used the CLARIN.SI concordancers and text analysis tools (61% vs. 49%) and have attended training sessions or events promoted on the CLARIN.SI website (42% vs. 19%). These differences may reflect the influence of the CLASSLA-Express international workshop series (Ljubešić et al., 2024), which, while it included a few events in Slovenia, was mostly held in neighbouring countries.

3.2 Depositing Resources in CLARIN.SI

The core service offered by CLARIN.SI is its repository of language resources and tools. The repository, set up in 2016, runs on the open-source CLARIN-DSpace platform and now archives around 700 entries.

The CLARIN.SI repository is widely used by the survey participants – a quarter of them had already deposited their own resources there, and almost half of them have downloaded and used resources from the repository. Among respondents who reported depositing resources in the CLARIN.SI repository, those who completed the follow-up questions (N=51) reported an overall positive experience. The highest-rated items on a 5-point scale included trust in data storage (average with standard deviation: 4.69$\pm$0.87), the importance of deposition for open science (4.76$\pm$0.85), the belief that deposition of resources supports their reuse in the community (4.41$\pm$1.08) and increases resource visibility (4.21$\pm$1.17). The efficiency of the editorial review process was rated highly (4.49$\pm$0.92), but the ease of different technical aspects slightly lower: the ease of license selection (4.04$\pm$0.95), submission workflow (3.87$\pm$0.92), technical file preparation (3.61$\pm$1.14), and metadata entry (3.67$\pm$0.87).

The majority of respondents who deposit resources in the repository (63%) reported consulting CLARIN.SI support, including website instructions and helpdesk assistance. Of these respondents, 30 completed the follow-up questions, and they rated the helpdesk highly: communication clarity and friendliness averaged 4.45$\pm$1.09, and effectiveness and timeliness of help averaged 4.62$\pm$1.05. The availability (4.0$\pm$1.0) and clarity (4.04$\pm$1.02) of instructions were rated slightly lower.

When asking for optional open-ended comments, we received 33 valid responses, 17 of which provided concrete suggestions.[1] Consistent with the quantitative ratings above, several thematic clusters were identified. The largest indicated need for simpler depositing process, including intuitive and automated deposition workflow and less demanding editorial control. A second cluster concerned repository entry management, highlighting simplified versioning, clearer metadata management, automated processing of large datasets, and practical support materials (e.g. video tutorials). The final cluster related to post-deposit visibility, with calls for better resource promotion and search functionality, as well as enhanced resource monitoring and authorship visibility.

3.3 Finding and Downloading Resources from CLARIN.SI

Almost half of the survey participants (N=90; 42%) reported using the CLARIN.SI repository to find and download language resources. The subset of users who responded to the follow-up questions (N=81) reported an overall positive experience. The most highly rated aspects were download reliability and speed (avg. with std = 4.61$\pm$0.68), followed by diversity (4.36$\pm$0.75) and the quality (4.28$\pm$0.86) of available resources. Participants also emphasized that CLARIN.SI resources are important for their work (4.11$\pm$0.97) and help save time in preparing their own data (4.28$\pm$0.85). Licensing information was considered clear and understandable (4.25$\pm$0.8). Slightly lower, but still positive ratings were given to the recency of resources (4.16$\pm$0.85) and the suitability of formats for further use (4.13$\pm$0.88). The lowest scores were given to the metadata clarity (3.97$\pm$0.74), and the ease of search (3.87$\pm$0.99).

Of the 45 valid open-ended responses in this section of the survey, 28 included at least one improvement suggestion. In line with the quantitative ratings above, several thematic clusters were identified, which relate to specific functionalities of the repository: a) search functionalities and downloads (e.g. clearer filtering, better indexing in external search engines, more visible download options); b) more consistent metadata management (easier versioning, addition of Slovenian metadata translations); c) resources and formats (stronger resource quality control, broader format support, more lightweight formats,

[1] In all open-ended questions, some participants indicated they had no specific suggestions, e.g. "/", "Nothing to add", "Keep up the good work", etc.

and more resources overall); and d) user help (inclusion of user ratings and forums, simpler solutions for non-experts, and search video tutorials).

3.4 Using CLARIN.SI Concordancers and Language Tools

CLARIN.SI hosts several concordancers, in particular noSketch Engine (Kilgarriff et al., 2014) and KonText (Machálek, 2020). Furthermore, we have three installation of noSketch Engine: the old "Bonito"[2] and the new "Crystal" version, as well as Crystal that requires log-in but allows more functions. Two further Slovenian-only on-line tools are included among the CLARIN.SI services: Korpusnik (Kosem et al., 2023), gives an overview of word usage based on querying five Slovenian corpora through the concordancer, while SENTA aims to simplify complex Slovenian sentences while preserving the original meaning (Žagar et al., 2024).

About half (52%) of survey participants reported using the CLARIN.SI concordancers and text analysis tools. Among those who responded to the follow-up questions (N=88), levels of user engagement varied across tools. Crystal (57%) and Bonito (46%) are the most widely used, followed by the KonText concordancer (33%), and NoSketch Engine with login (29%). Slovenian-only tools Korpusnik and Senta are used by 35% and 16% of Slovenian survey participants, respectively.

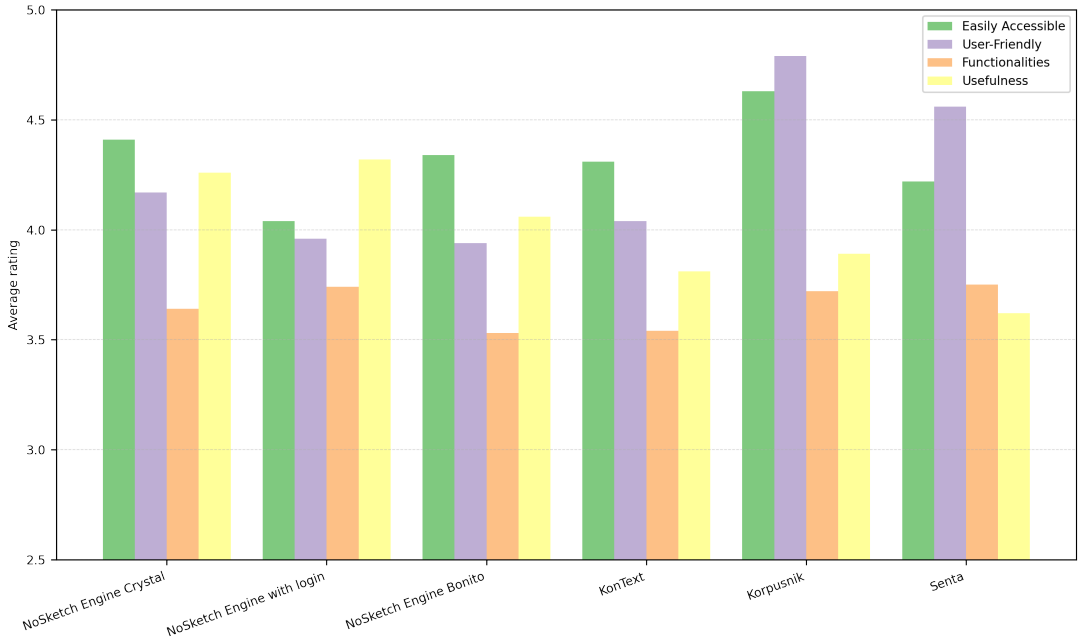

Figure 3: Average ratings of the CLARIN.SI text analysis tools in terms of their accessibility, user-friendliness, provided functionalities, and their usefulness for the participants' work.

Figure 3 compares the CLARIN.SI concordancers and other text analysis tools in terms of accessibility, user-friendliness, available functionalities, and perceived usefulness for participants' work. All tools were rated as easily accessible, with Korpusnik (avg. with std = 4.63±0.60) and NoSketch Engine Crystal (4.41±0.91) rated the highest, and the NoSketch Engine concordancer with login rated the lowest (4.04±1.15). User-friendliness varied notably: the Slovenian tools Korpusnik (4.79±0.42) and Senta (4.56±0.53) were considered significantly more user-friendly than the concordancers, with NoSketch Engine Crystal (4.17±0.93) rated the most user-friendly among the latter. According to participants, the NoSketch Engine concordancer with login offers the most functionalities (3.74±1.10), while NoSketch Engine Bonito (3.53±0.98) and KonText (3.54±1.10) offer the least. In terms of overall usefulness, the NoSketch Engine concordancer with login (4.32±0.99) and NoSketch Engine Crystal (4.26±1.03) were rated as the most helpful for the participants' work.

[2]Note, however, that we have recently shut down Bonito, as it was continuously and very aggressively harvested by robots, leading to outages of the infrastructure.

Although improving these third-party tools is outside our scope, we still collected open-ended user suggestions to gain better insight into perceived weaknesses of each tool. Among the concordancers, Crystal received the most detailed feedback (29 responses; 18 with suggestions), with users highlighting usability and learnability issues, especially the unintuitive interface with functions hidden behind icons. Bonito (21 responses; 10 with suggestions) drew mixed feedback, with some valuing its minimalist interface and stable features, while others noted concerns about maintenance, accessibility, and outdated interface elements. NoSketch Engine with login (11 responses; 6 with suggestions) received modest feedback focused on clearer settings, better onboarding for non-technical users, and alignment with Sketch Engine functionality. Regarding KonText (16 responses; 4 with suggestions), some users noted the interface's complexity and a lack of tooltips and reminders on how to use CQL query language. On the other hand, for Korpusnik (13 responses; 3 with suggestions), comments pointed to the need for more flexible display options and support for multiword lexeme search, while feedback on Senta (5 responses; 3 with suggestions) focused on improving summarisation quality and extending the tool to spoken-language inputs with LLM-supported explanations.

3.5 Staying Informed about CLARIN.SI

A third of survey participants regularly or occasionally follow CLARIN.SI news through its website or social media channels. Among respondents who completed the follow-up questions focused on communication and updates from CLARIN.SI (N=51), most reported receiving information via the CLARIN.SI website (54%) and the CLASSLA mailing list (43%), while fewer used channels such as LinkedIn (26%), X/Twitter (18%), and Discord (14%). The distribution of Slovenian and international users across these channels varies notably. Nearly twice as many international participants follow the CLASSLA mailing list (64%) compared to Slovenian participants (35%), while the opposite is true for news on the CLARIN.SI website, which is followed by 62% of Slovenian and 36% of international users. The proportions of users following CLARIN.SI on X and Discord are similar across both groups, but significantly more international participants follow CLARIN.SI on LinkedIn (43% vs. 19% of Slovenian participants).

Participants found the news clear and engaging (4.13$\pm$0.74), and the frequency of updates generally appropriate (4.09$\pm$0.82). There was less agreement with statements that the updates motivated them to use CLARIN.SI services or resources (3.85$\pm$1.02), that they were useful for their work (3.83$\pm$0.99), and that they were sufficiently visible (3.44$\pm$1.09). The idea of a dedicated CLARIN.SI newsletter received mixed support (3.85$\pm$1.09).

In this section, we received 27 valid open-ended responses, 8 of which included improvement suggestions. Several called for wider dissemination and promotion, including a newsletter (2 participants), stronger promotion of resources (3), especially new/updated resources (3), guidance on using tools and resources (2), plus more visible information about events (1) and a communication style that better serves both new and experienced users (1) and comprises shorter, more focused messages (1).

3.6 Attending CLARIN.SI Related Presentations and Events

A quarter of all survey participants reported attending CLARIN.SI training or events, including lectures, workshops, and other activities. Among respondents who completed the follow-up questions (N=39), most referenced participation in the Language Technologies and Digital Humanities conferences, held biennially in Ljubljana, and related events, including the JOTA lectures and the CLASSLA-Express (Ljubešić et al., 2024) workshops.

The overall experience with CLARIN.SI events (N=40) was rated highly: the average rating of usefulness for their work or research was 4.41$\pm$0.68, with even higher scores for clarity (4.76$\pm$0.43), appropriate level of expertise (4.82$\pm$0.39), and event organization (4.67$\pm$0.53). Participants also appreciated the opportunities for interaction (4.64$\pm$0.63) and the accessibility of materials after the event (4.75$\pm$0.6).

Out of 22 valid open-ended responses in this section, 10 included at least one improvement suggestion. The most common requests were for more events focused on practical usage and usability of tools/resources (3 participants), longer events to allow more time and tasks (2), and better promotion/communication about events (2); two participants also wanted a wider thematic scope (e.g., more

digital humanities, beyond language technologies). In addition, two participants asked for more content on corpus data analyses (including methods/statistics and corpus-building), while single participants mentioned events for community building/networking, sharing good-practice examples, supporting crowdsourcing-oriented resource building initiatives, and offering separate tracks for beginners vs. advanced users.

3.7 Including CLARIN.SI in Presentations and Lectures

Incorporating CLARIN.SI into participants' lectures or presentations (N=37, 17%) was done in various ways (multiple responses were possible). The most common type of integration included citing resources from the CLARIN.SI repository (76%), demonstrating the usability of tools or data through their own practical use (69%), and showcasing tools or resources through live demonstrations (59%). Around half of them referenced CLARIN.SI in discussions of their research methodology (55%) or explicitly directed audiences to the CLARIN.SI infrastructure for further research or learning (52%). A smaller proportion of survey participants incorporated CLARIN.SI into their presentations when explaining concepts such as open science or language resources (38%) or provided comprehensive overviews of CLARIN.SI's structure and purpose (17%; 24% of Slovenian survey participants).

The participants who answered follow-up questions on this topic (N=29) believe that including the CLARIN.SI infrastructure in their lectures or presentations encouraged the attendees to use these resources or tools (avg. with std = 4.05$\pm$0.97). However, they did not find CLARIN.SI content to be particularly well-suited or easily accessible for use in presentations. The availability of materials for citation (3.96$\pm$1.12), the usefulness of tools and resources for clear demonstration (3.92$\pm$0.91), and the accessibility of CLARIN.SI presentation materials providing general information about the infrastructure (3.83$\pm$0.98) were rated as less satisfactory than other CLARIN.SI services.

Out of 16 valid open-ended responses in this section, 7 included improvement suggestions, most often asking for ready-to-use presentation materials (3 participants), such as pre-made slides or short "bullet-point" handouts that can be easily reused in talks. Related to this, one participant asked for better access to visual identity elements (e.g., a clearly findable high-resolution logo), while other single-participant suggestions covered better usage-oriented resource presentations with examples, materials tailored for school use/teaching, improved language accessibility for non-experts (Slovenian aside from English), and easier general access to information (through a more appealing UI/more precise explanations).

3.8 Learning about CLARIN.SI

Almost a third of respondents initially stated that they had not (yet) directly used CLARIN.SI in their work. In response to the question "What would encourage you to use the CLARIN.SI infrastructure?", they identified several key motivators, as shown in Figure 4, including better awareness of the infrastructure's structure and purpose (61% of respondents) and more examples of good practice and concrete applications of resources (49%). Many desired more accessible information about updates, events, and opportunities (27%), as well as targeted training or workshops (27%). Less common were requests for increased technical support, a wider range of language resources, and customisation options (10-18%). 12% (6 respondents) indicated that CLARIN.SI was in no context relevant to their work.

In this section, we also received 24 valid open-ended responses, 11 of which included at least one improvement suggestion. Participants' comments were dominated by the desire for better outreach (7 participants) — more promotion and advertising (including social media), clearer introductory information about what CLARIN.SI is, and more information aimed at a wider professional public (including better information about workshops). A smaller set asked for a clearer website (2 participants), especially clearer navigation, and one participant suggested integration of CLARIN.SI-related information into formal education.

3.9 Priorities for Further Development

The respondents (N=121) were asked to rate the importance of various future development priorities for CLARIN.SI on a scale from 1 (not a priority) to 5 (top priority). As shown in Figure 5, the highest-rated priorities were the development of new language tools (avg. with std = 4.06$\pm$1.03), integration

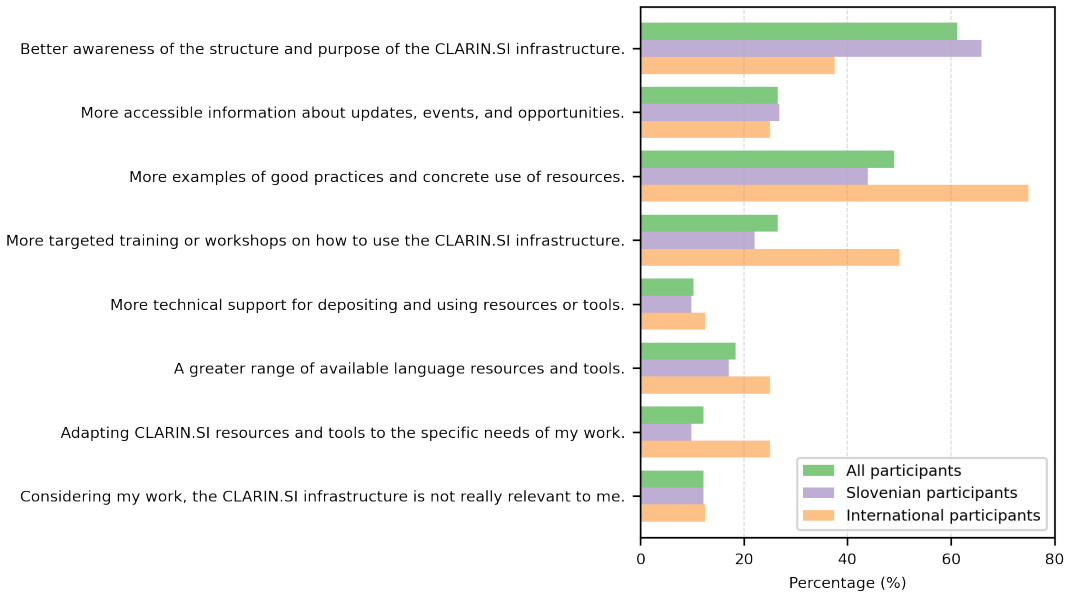

Figure 4: Responses from participants without prior direct experience with CLARIN.SI to the question: "What would encourage you to use the CLARIN.SI infrastructure?" (multiple responses allowed).

of CLARIN.SI with other services (3.94±1.04), and better maintenance and updates of existing tools (3.82±1.1). Improving data search and download processes (3.78±1.15) and increasing the diversity of language resources (3.78±1.16) followed closely.

Slightly lower priorities included improved data quality control (3.71±1.1), increased and targeted training opportunities (3.7±1.18), improved user support and documentation (3.56±1.11), and more frequent and tailored updates on news and events (3.47±1.17). Lower still were clearer presentation of the infrastructure's composition and accessibility (3.41±1.2), improving the data deposition and storage process (3.4±1.24), and better presentation of CLARIN.SI organization (3.09±1.33). Importantly, a relatively high proportion of respondents selected "I cannot assess" for several items (17–42% respondents), suggesting a need for further clarification of certain aspects of the infrastructure.

In this section of the survey, two open-ended questions were provided. For the first, "If you wish, you can further explain your choices in the previous question," we received 56 valid answers; 22 contained substantive explanations. In terms of content, respondents most often called for stronger communication and onboarding (4 participants), with clearer basic explanations of what CLARIN.SI offers and how it benefits users, and for more systematic training and education (3) — especially for students and newcomers — covering not only tool use but also methods and examples of good research practices. A second major thread concerned resources and discoverability: four participants stressed expanding and improving corpora and other resources (including filling specific gaps such as 20th-century coverage), while three participants emphasised need for improved repository search capabilities; this was also linked to desires for better web visibility and improved search functionalities. Additionally, suggestions included more regular tool maintenance and updates (3 participants)[3], more concrete demonstrations of how resources can be used in practice (2), and several single-participant points: providing clearer information on resource currency, improving integration with other platforms and workflows, preparing reusable presentation materials, clearer site structure, clearer information on how to join the consortium, better quality control of content (not just of the technical aspect), and better corpus annotation.

The second open-ended question, "Is there an area or challenge that we haven't addressed in the

[3]For this and similar suggestions, it is not possible to determine whether such observations reflect perceived lower quality, shortcomings in the infrastructure offering, or general observations. This comment, for example, was interpreted as indicating a need for more frequent updates and maintenance of existing tools. All user comments were categorised, although some are more abstract and less suitable for priority development than others.

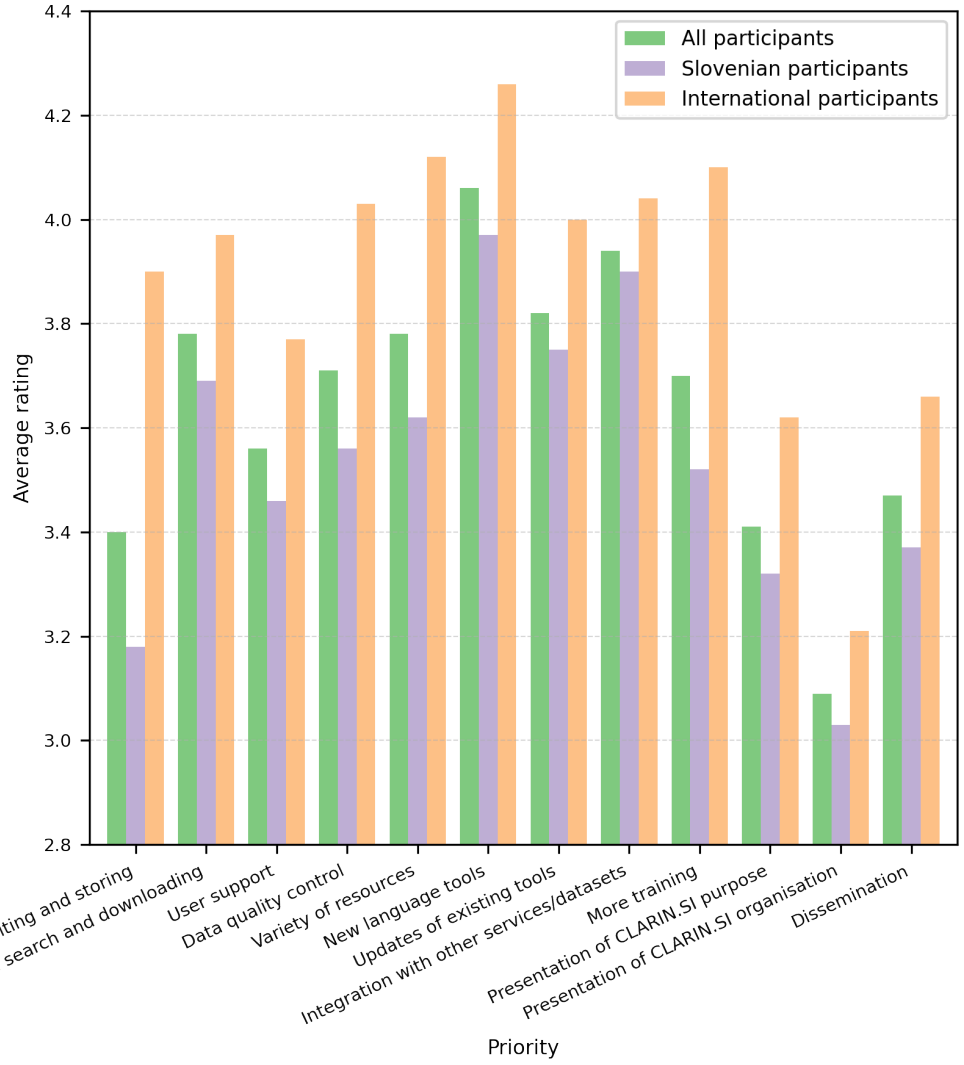

Figure 5: Responses from participants to the question: "How high of a priority for future development are the activities below?" (multiple responses allowed, where 1 represents lowest priority and 5 highest priority).

questions, but which you consider important for the future of CLARIN.SI?" provided 55 valid answers; 16 included at least one substantive comment. The most recurrent themes were better integration with other platforms (2 participants, e.g., Hugging Face and evaluation sites like SloBench), clearer information on consortium membership/participation (2), further development of new resources (2, especially non-standard and spoken/dialect corpora), and more training (2). Other priorities were mentioned by one participant each: clearer guidance on data ownership and personal data protection; better linking of resources across languages; improving visibility/communication to the wider public; improving the concordancers; increasing the value of CLARIN.SI outputs in research evaluation/CVs; more content about AI; clearer quality-control expectations for deposited resources; research support; and improved presentation of available resources and tools that can support educational contexts. As shown, most of these comments align with areas and challenges already addressed in the survey, suggesting that the design effectively anticipated users' main types of engagement and potential concerns.

4 Conclusion

The survey results provide a comprehensive insight into the types of our users, their experiences, needs, and expectations regarding CLARIN.SI. First, almost a quarter of our users are not NLP specialists nor

linguists but come from other fields of the humanities. While we already have a long-standing collaboration with the Slovenian DARIAH, this indicates that collaboration of the two research infrastructure should be further strengthened to improve our visibility and use among historians, political scientists etc., even more so as we offer a rich set of resources for that could be of use to them, e.g. the ParlaMint corpora of parliamentary debates (Erjavec et al., 2024) and related resources. Such collaboration should also be reciprocal, as DARIAH(-SI) also offers resources and services to support users from primarily linguistic fields.

As far as the experiences of the users go, infrastructure is, overall, well-regarded by its users, particularly in terms of data reliability, open science support, and perceived usefulness of deposited resources. The repository is widely used by participants and provides — overall — a positive experience both with deposition and resource retrieval. Users also value the availability of concordancers and tools that support their work, and were rated highly for accessibility. The communication channels are considered clear and engaging for disseminating information about CLARIN.SI and related activities, although Slovenian and international participants use quite different channels. CLARIN.SI-related events received high ratings for professional quality, clarity, expertise, and the accessibility of materials after the event.

The aim of the survey was to gather feedback from existing users as well as potential new users, with almost a third of participants reporting no previous experience with CLARIN.SI. They highlighted the need to raise awareness of the infrastructure's existence, improve access to information about its structure, and provide more concrete examples of resource use. In contrast, participants with previous experience with the infrastructure identified several priorities for further development. The highest-rated priorities concern the services and tools offered, with the development of new language tools considered the highest priority, followed closely by the integration of CLARIN.SI with other services and by improving existing tools.

Responses to the open-ended questions provided further details on specific issues and participant suggestions. The first major cluster, which includes the most frequent and specific requests, concerns the structure and operation of the CLARIN.SI repository. Responses highlighted search-related issues, such as irrelevant results, limited browsing support, and missing filters, as well as the need to optimise the data deposition workflow (e.g., simplifying submission, editorial review, and entry updates). Open-ended responses included concrete suggestions to address these shortcomings, such as search improvements (autocomplete and quick data previews) and deposition enhancements (clearer, centrally visible instructions and reduced metadata requirements), among others. The technical suggestions are mostly out of our control, as we use the LINDAT DSpace repository software, i.e. it would be difficult to improve its functionality at our end. On the other hand, we plan to install a new version of the repository platform shortly, which may address some of these problems. However, we will take on board suggestion related to documentation, where we should improve its accessibility. But while we understand that depositors would like less work with the metadata entry, we will, however, still insist on clear, complete and harmonised metadata for the repository entries.

In parallel, users repeatedly indicate that CLARIN.SI's impact depends on outreach, onboarding, training, and reusable communication assets, not solely on infrastructure. Suggestions to improve outreach and communication include options such as more concise, action-oriented communications and the preparation of a "CLARIN.SI presentation kit" containing ready-to-use materials. For events, suggestions include more hands-on workshops for different skill levels, longer multi-day formats, and training that goes beyond tool use to cover data processing and research workflows.

Lastly, a separate but still prominent theme identified concerns expanding and modernising the resource offer and formats. Suggestions specifically mention the need for more diverse corpora, including non-standard language and spoken corpora, addressing gaps in diachronic coverage (notably the 1918–1990 period), as well as providing resources in additional formats to support downstream work. However, CLARIN.SI does not, as a rule, produce corpora, and offered resources and corpora are mostly third party resources, so its ability to fill the gaps in the corpora offered is limited. On the other hand, the suggestion to expand the range of deposited formats (in particular, to JSON) could, in certain cases, be acted upon.

For certain suggestions it should also be noted that the conclusions of the survey are somewhat contradictory. For example, it was requested to add the option to monitor downloads of deposited resources, however, this option already exists. Similarly, a simpler approaches to updating resources with new versions was suggested, even though it is already very simple. Such suggestions show that users do not always read the on-line documentation, which is a known and persistent problem, as our reviewing of deposited entries often shows. This again shows that the documentation should be made more prominent and accessible.

The survey results highlighted key areas for improvement, which are also aligned with some specific goals in CLARIN.SI strategy 2024-2030 priorities. The first priority identified in the survey concerns providing more corpus analysis tools, which aligns with the second Pillar of the Strategy (Our offer), specifically *Goal 3.2*: Improve the coverage and quality of data, models, and tools offered through the CLARIN.SI infrastructure. The survey results indicate that there is room for improvement in the provision of high-quality language data sets. Additionally, some comments highlighted the need for better authorship visibility, which is addressed in *Strategy 3.2.6*: Act to improve existing citation guidelines and practices in Slovenia.

More importantly, the survey identified the need to focus on user engagement, which is directly linked to *Pillar 1 (Our users)*, addressing both academic and non-academic user bases. The need for more tutorials and presentations of the infrastructure is also linked to *Strategy 1.1.2* (Improve the communication channels to better reflect the mission, network, and offer of CLARIN.SI) and *Strategy 1.1.3* (Improve services through newly created training, workshops, and online tutorials). The highlighted priorities are relevant to the key activities of CLARIN nodes research infrastructures, as well as research infrastructures in general, i.e. user base and infrastructure offering. For CLARIN.SI the survey results and actionable user-identified priorities will serve as guidance for further development.

5 Limitations

The study relies on descriptive statistics, in line with its exploratory aim of collecting structured feedback on the quality of CLARIN.SI services, tools, and resources, and of identifying areas for improvement relevant to future development and strategic planning. No predefined hypotheses were posed, and the analysis was therefore not designed as a basis for inferential testing. Inferential comparisons between respondent subgroups, such as Slovenian and international participants, were consequently not conducted. This was also due to the structure of the questionnaire, which used branching and allowed item non-response, resulting in varying and, for some questions, relatively small numbers of responses, thereby limiting the robustness of such comparisons. The differences observed between groups should therefore be understood as indicative rather than statistically confirmed. A further limitation concerns representativeness. Participation was voluntary, and the survey was distributed through academic, educational, and professional networks considered likely to reach existing or potential CLARIN.SI users. The sample cannot be considered representative of any broader population, and the findings may also reflect self-selection bias, including the possible overrepresentation of respondents who are more aware of the infrastructure or more motivated to provide feedback. At the same time, the inclusion of participants without prior direct experience with CLARIN.SI broadens the range of perspectives captured by the survey and offers insight into factors that may affect future uptake. These limitations should be taken into account when interpreting the findings.

6 Acknowledgements

The authors would like to thank the anonymous reviewers for their userful comments. The work was supported by the Slovenian Research and Innovation Agency (ARIS) via the research infrastructure CLARIN.SI (I0-E004) and the core research programmes Language Resources and Technologies for Slovene (P6-0411) and Knowledge Technologies (P2-0103).

References

Arhar Holdt, Š., Meden, K., Kuzman Pungeršek, T., Ljubešić, N., & Erjavec, T. (2026, February). CLARIN.SI user survey 2025. https://doi.org/10.5281/zenodo.18486462

De Smedt, K. (2020). *Clarino - user survey report* (tech. rep.). CLARINO. https://clarin.w.uib.no/files/2020/08/CLARINO-User-Survey-Report.pdf

Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, Starkadur, Bartolini, R., Bel, N., Pérez, C. M., … Fišer, D. (2024). ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09798-w

Erjavec, T., Ljubešić, N., Meden, K., Kuzman, T., Laskowski, C., Javoršek, J. J., Krek, S., Jemec Tomazin, M., Dobrovoljc, K., Holdt, Š. A., Lenardič, J., & Fišer, D. (2025). CLARIN.SI, the Slovenian node of CLARIN: Ten years on. *Selected papers from the CLARIN Annual Conference 2024, Linköping Electronic Conference Proceedings (ECP)*. https://doi.org/10.3384/ecp216.07

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36. https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf

Kosem, I., Čibej, J., Krek, S., & Dobrovoljc, K. (2023). Korpusnik: a corpus summarizing tool for Slovene. *CLARIN Annual Conference Proceedings*, 129. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Ljubešić, N., Kuzman, T., Filipović Petrović, I., Parizoska, J., & Osenova, P. (2024). CLASSLA-Express: a Train of CLARIN. SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings 2024*, 31–35.

Luzietti, R. B., Spadi, A., Giampietro, N., Mancuso, G., Caravale, A., D'Eredità, A., Caradonna, M., Moscati, P., Quochi, V., Monachini, M., et al. (2025). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an open science cloud. *Umanistica Digitale*, (20), 419–439.

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865

Monachini, M., Quochi, V., Luzietti, R. B., Spadi, A., Mancuso, G., D'Eredità, A., Caravale, A., Giampietro, N., Caradonna, M., & Melaccio, D. (2025, October). D2.2 updated report on the h2iosc landscapes. https://doi.org/10.5281/zenodo.17737101

Nicolas, L., König, A., Monachini, M., Del Gratta, R., Calamai, S., Abel, A., Enea, A., Biliotti, F., Quochi, V., & Stella, F. V. (2018). CLARIN-IT: State of affairs, challenges and opportunities. *Selected papers from the CLARIN Annual Conference 2017*. https://hdl.handle.net/10863/8194

Sichera, P., Monachini, M., Quochi, V., Giampietro, N., Fabiani, V., Ottaviani, R., & Luzietti, R. B. (2025). Synergies between clarin-it and operas-it within h2iosc: Monitoring communities and orchestrating digital services. *CLARIN Annual Conference Proceedings*, 26.

Žagar, A., Klemen, M., Robnik-Šikonja, M., & Kosem, I. (2024). SENTA: Sentence simplification system for Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14687–14692. https://aclanthology.org/2024.lrec-main.1279/

The AI Act and its impact on Large Language Models and the CLARIN Infrastructure

Paweł Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Joanna Blochowiak
University of Zurich, Switzerland
joanna.blochowiak@uzh.ch

Anna Gosławska
Wroclaw University of Science and Technology, Poland
anna.goslawska@pwr.edu.pl

Henk van den Heuvel
Radboud University, Nijmegen, the Netherlands
h.vandenheuvel@let.ru.nl

Erik Ketzan
King's College London, UK
erik.ketzan@kcl.ac.uk

Krister Linden
University of Helsinki, Finland
krister.linden@helsinki.fi

Costanza Navarretta
University of Copenhagen, Denmark
costanza@hum.ku.dk

Andrius Puksas
Vytautas Magnus University, Lithuania
andrius.puksas@vdu.lt

German Rigau
University of the Basque Country, Spain
german.rigau@ehu.eus

Abstract

The EU Artificial Intelligence Act (Regulation 2024/1689), which entered into force in August 2024, represents the world's first comprehensive regulatory framework for AI. While it explicitly excludes AI systems developed solely for scientific research and development, its implications for the CLARIN community are far-reaching, especially given the growing reliance on large language models (LLMs) and general-purpose AI (GPAI). This paper discusses the scope of research exemptions in the AI Act (Section 2), and provides an overview of the obligations related to transparency of AI-generated text outputs (Section 3), as well as those imposed on providers of General-Purpose AI models, including those with systemic risk (Section 4). Given CLARIN's role as a data provider, special attention is paid to the requirement for dataset documentation.

1 Introduction

EU Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, better known as the AI Act, is a recent addition to the legal framework affecting language data and language technology. The Act was officially proposed in 2021; the proposal was then substantially modified to better address the issues related to General-Purpose AI (Bertuzzi, 2023), which delayed its adoption. The AI Act entered into force on 1 August 2024. While most of its provisions will apply from 2 August 2026 (i.e. 24 months after entry into force), the Act provides for a phased application with several earlier deadlines. In particular, the obligations relating to general-purpose AI apply from 2 August 2025.

The AI Act is, globally, the first comprehensive attempt to legally regulate AI (see e.g. Ruschemeier 2022, Wörsdörfer 2025). As such, it cannot go unnoticed by the CLARIN community. For the typical academic researcher employed by a university or research institute, the impact of the AI Act may, at first glance, appear to be negligible, as the Act "exclude[s] from its scope AI systems and models specifically developed and put into service for the sole purpose of scientific research and development"[1]. But even with a clear exemption for scientific research, the AI Act has the potential to impact the CLARIN community's tools and methods, funding, collaborations with industry, corporations that may be hiring academics and students in the

[1] Recital 25 and Article 2(6) of the AI Act.

Paweł Kamocki, Joanna Blochowiak, Anna Gosławska, Henk van den Heuvel, Erik Ketzan, Krister Linden, Costanza Navarretta, Andrius Puksas, and German Rigau 2025. The AI Act and its impact on Large Language Models and the CLARIN Infrastructure. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 161–171. DOI: https://doi.org/10.3384/ecp222

future, stakeholders beyond the typical academic researcher, and a myriad of end-users of CLARIN community members' research. An exploration of the impact of the AI Act on CLARIN-related activities is thus important.

This paper will discuss the general exclusion of research from the scope of the AI Act (Section 2), before analysing the rules on Transparency (Section 3) General Purpose AI (a category in which LLMs fall) in greater detail (Section 4).

2 General Exclusion of Research and Development Activities from the AI Act

One of the objectives of the AI Act, expressly stated in its Recital 25, is to promote innovation, safeguard freedom of science, and avoid hindering research and development activities. One of the provisions aimed at this objective, particularly important for the CLARIN community (albeit not included in the draft – see Smuha et al. 2021), can be found in Article 2(6) of the AI Act, which contains a general scientific research exemption. According to this provision, the AI Act "does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development". In addition to this, Article 2(8) provides that the AI Act "does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service". This indicates that models and systems specifically developed for "scientific research" (as well as their outputs) are fully exempted from the GDPR, whereas commercial "research and development" activities are exempted only until the system or model concerned is "placed on the market" (i.e., first made available on the EU market – cf. Article 3(9) of the AI Act) or "put into service" (i.e. first supplied for use on the EU market – cf. Article 3(11) of the AI Act), i.e. exploited commercially. One can easily imagine that this two-speed solution may be difficult to apply in practice, especially in public-private partnerships, or within research departments of commercial companies (Finck, 2025).

In practice, AI models and systems that are primarily designed for other purposes but are employed within research and development contexts still fall under the scope of the AI Act. Given that research does not operate in isolation from broader societal and technological developments, it is reasonable to expect that the AI Act will influence researchers – particularly when they develop or deploy models and systems intended for general use, or when they test such technologies in real-world conditions. Since CLARIN serves not only the academic community but also a broader range of stakeholders, it is essential for the CLARIN community to remain informed about the evolving regulatory landscape. The rapidly advancing capabilities of AI in solving complex tasks (see, e.g., Kaur et al., 2024) suggest that general-purpose or fine-tuned large language models (LLMs) may increasingly replace many of the specialized NLP tools and packages traditionally used by CLARIN researchers (see, e.g., Min et al., 2021). Therefore, understanding and analysing the implications of the AI Act for CLARIN-related activities is both timely and crucial.

3 Transparency of AI-generated Text Outputs under the AI Act

The rules of transparency of AI outputs are among the most interesting provisions of the AI Act, both for theoreticians and for practitioners. The language community should pay particular attention to Articles 50(2) and 50(4), as they both involve text outputs. According to Article 50(2), providers of AI systems (including general-purpose AI systems) generating synthetic audio, image, video or *text* content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Article 50(4) provides *inter alia* that deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest (hereinafter: "news") shall disclose that the text has been artificially generated or manipulated. The latter obligation does not apply when the text has undergone human review or editorial control, and is published under the editorial responsibility of a legal or natural person.

First of all, it should be stressed that the transparency obligation concerning all AI-generated content weighs on the providers, whereas the more specific obligation concerning "news" weighs on deployers. It is therefore crucial to explain the difference between the two categories of stakeholders: providers, according to Article 3(3) of the AI Act, are the entities who develop an AI system or have it developed and who exploit it under their own name (e.g., OpenAI

is the provider of ChatGPT). Deployers, on the other hand, are the persons or entities who use an AI system in the course of their activity, except when this activity is both personal and non-commercial. In other words, a deployer can be described as a "professional user", e.g., a university or a publishing house using ChatGPT are deployers, and so is an individual researcher using ChatGPT for purposes other than strictly personal.

It is therefore the provider that has to make sure, under Article 50(2), that the outputs generated by an AI system are marked in a machine-readable format; in addition, under Article 50(4), the user (excluding personal non-commercial users) that uses an AI system to generate or manipulate "news" shall disclose this information to the public.

Both provisions enter into application in August 2026, raising significant practical challenges. In anticipation of these difficulties, in September 2025 the AI Office (a part of the European Commission) launched a process of drafting a Code of Practice on Transparency of AI-Generated Content. The first draft of the Code was published on 17 December 2025[2]. Members of two dedicated working groups admit that at this stage the Code remains "high-level and broad", but it proposes some interesting measures. Once approved by the Commission, the Code will serve as a voluntary tool for providers and deployers of generative AI systems to demonstrate compliance with their respective obligations regarding transparency.

The draft Code obliges AI providers who adhere to it to ensure that AI-generated or manipulated text is marked with an "imperceptible watermark", directly interwoven within the content in a manner that is difficult for it to be separated from the content, and that withstands typical processing steps that may be applied to the content (Section 1, Sub-measure 1.1.2.). The watermark can be embedded during model training, model inference, or within the output of an AI model or system. To address the deficiencies of this method, AI providers will also establish and maintain logging facilities that allow for checking whether a text output has been generated or manipulated by their generative AI system (Section 1, Sub-measure 1.1.3.). "Logging" is defined in the draft Code as "verbatim recording and indexing of content" which "can be used for a fast lookup of known content". As an additional measure, AI providers who sign the Code will be obliged to implement a digitally signed manifest, allowing deployers to obtain a certified version of the AI output generated or manipulated by their AI system or model to formally guarantee the origin of content that does not allow secure embedding of metadata. This provenance certificate will enable deployers and end-users to provide third parties with guarantees that the content is AI-generated or manipulated, linking it back to the specific generative AI system or model" (Section 1, Sub-measure 1.2.1.).

Regarding the transparency obligations of deployers, they are also included in the draft Code. In particular, the deployers should agree on a common taxonomy, including clear definitions of what constitutes AI-generated and AI-assisted "news". According to the draft Code (Section 2, Measure 1.1), AI-generated "news" are "fully and autonomously generated by the AI system without human authored authentic content" (in other words, generated from a prompt), whereas AI-assisted "news" are those with "mixed human and AI involvement (...), where the AI system substantially impacts the content" of the "news". AI involvement may consist of, *inter alia*, "adding AI-generated or manipulated text to existing human-authored text", or "AI-enabled alterations to description of events, actors, arguments or interpretation". Furthermore, signatories of the Code will set up internal procedures to identify AI-generated and AI-manipulated "news" (Section 2, Measure 5.1), and documentation of the human review process (Section 2, Measure 5.2). For "news" concerned by the Article 50(4), the signatories will disclose the use of AI by placing a standardised icon in a "fixed, clear and distinguishable position", which could include but is not limited to placing the icon "at the top of the text, beside the text, in the footer or after the closing sentence of the text".

4 General-Purpose AI Models under the AI Act

The rules on General-Purpose AI Models (GPAIM), which potentially include many language models, are found in Chapter V of the AI Act.

[2] https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content

In recent months, the rather abstractly formulated lists of obligations of GPAIM providers have gradually been filled up with practical meaning – some elements are already in place. To a large extent, this is due to the adoption of the GPAI Code of Practice, whose final version was published on July 10, 2025[3]. It was complemented by Commission guidelines on key concepts related to GPAIM[4]. The Code contains, among other interesting elements, a Model Documentation Form for GPAIM, which may seem surprisingly simple and straightforward, totaling only 3 pages.

The rules regarding GPAIM became applicable on 2 August 2025. However, according to Article 111 of the AI Act, providers of GPAIMs that have been placed on the market before 2 August 2025 shall take the necessary steps in order to comply with the obligations laid down in the AI Act by 2 August 2027.

4.1. Concept and typology of General-Purpose AI Models

A general-purpose AI model (GPAIM) is defined as: "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market" (Article 3, (63) of the AI Act).

The key characteristic of a GPAIM is its capability to perform a wide range of distinct tasks. Typically, such models are trained on large amounts of data using machine learning methods and can be made available for direct download, as physical copies, or through libraries or APIs. LLMs, and other large generative models, are considered GPAIMs (Recital 99 of the AI Act). In case of doubt, the generality of a model can also be determined by the number of its parameters – models with a billion or more parameters are presumed to be GPAIMs (Recital 98 of the AI Act). Most modern Large Language Models, at least those released after 2019, meet this criterion; remarkably, BERT, originally released in 2018 is still significantly smaller (despite being considered an LLM)[5].

According to the Commission's guidelines (para. 17), an indicative criterion for a model to be considered a general-purpose AI model is that its training compute is greater than 10^{23} FLOP and it can generate language (whether in the form of text or audio), text-to-image or text-to-video.

In the spirit of the general exclusion of research and development activities from the AI Act (see above), models used for research, development and prototyping and not available on the market are not considered GPAIMs.

Within the category of GPAIM, the AI Act distinguishes two additional sub-categories: GPAIM with systemic risks (which are subject to a stricter framework) and "free open source" GPAIM.

4.2. Obligations of General-Purpose AI providers

The providers of all GPAIM need to comply with the four rather generally phrased obligations listed in Article 53 of the AI Act, i.e.:

1. Prepare and keep up to date the technical documentation of the model, including its training and testing process and the results of its evaluation. (see below in 4.2.1)

2. Prepare and keep up to date information and documentation that should be made available to the systems providers who want to integrate the GPAIM in their AI systems. (see below in 4.2.2)

3. Put in place a policy to comply with copyright and related rights (see below in 4.2.3)

[3] https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[4] https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act

[5] https://en.wikipedia.org/wiki/List_of_large_language_models -- last visit 31.03.2025.

4. Prepare and make publicly available a sufficiently detailed summary about the datasets used for training the model. (see below in 4.2.4).

Obligations 1. and 2. do not apply to GPAIM released under free and open source licenses, and whose parameters (including the weights), information about their architecture, and the information on model usage are publicly available (Article 53(2) of the AI Act), unless these models qualify as GPAIMs with systemic risks.

A GPAIM with systemic risks is a GPAIM that meets one of the following conditions:

a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies OR

b) it has been officially recognised as such in a Commission decision.

The criteria mentioned in a) will be fine-tuned by the Commission in a delegated act. The AI Act only mentions one criterion: a GPAIM is presumed as carrying a systemic risk "when the cumulative amount of computation used for its training measured in floating point operations (FLOPs) is greater than 10^{25}". Over 25 modern GPAIMs, including GPT-4, are believed to meet this threshold.

Providers of GPAIM with systemic risk need to comply with additional obligations under Article 55 of the AI Act:

1. Perform model evaluations, including adversarial testing to identify and mitigate systemic risk.

2. Assess and mitigate possible systemic risks, including their sources.

3. Track, document and report serious incidents and possible corrective measures to the AI Office and national competent authorities;

4. Ensure an adequate level of cybersecurity protection.

One can assume that currently only very large LLMs will be considered as GPAIMs with systemic risks; to what extent this framework will be relevant for CLARIN remains to be seen.

4.2.1. Technical documentation of General-Purpose AI Models

The documentation should at the very least contain the information listed in Annex XI of the AI Act, and upon request be provided to the competent national authorities and the AI Office. The Annex XI lists the following items to be included in the documentation:

1. A general description of the general-purpose AI model including:

   (a) the tasks that the model is intended to perform and the type and nature of AI systems in which it can be integrated;

   (b) the acceptable use policies applicable;

   (c) the date of release and methods of distribution;

   (d) the architecture and number of parameters;

   (e) the modality (e.g. text, image) and format of inputs and outputs;

   (f) the licence.

2. A detailed description of the elements of the model referred to in point 1, and relevant information of the process for the development, including the following elements:

(a) the technical means (e.g. instructions of use, infrastructure, tools) required for the general-purpose AI model to be integrated in AI systems;

(b) the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters, as applicable;

(c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies (e.g. cleaning, filtering, etc.), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases, where applicable;

(d) the computational resources used to train the model (e.g. number of floating point operations), training time, and other relevant details related to the training;

(e) known or estimated energy consumption of the model.

4.2.2. Documentation for AI system providers

The documentation should give said providers a "good understanding of the capabilities and limitations" of the model, and enable them to comply with their obligations under the AI Act (e.g., for providers of high-risk AI systems, the obligation to risk management, transparency, and their own documentation obligations). At the very least, the documentation should contain the information listed in Annex XII of the AI Act, i.e.:

1. A general description of the general-purpose AI model including:

the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated;

the acceptable use policies applicable;

the date of release and methods of distribution;

how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable;

the versions of relevant software related to the use of the general-purpose AI model, where applicable;

the architecture and number of parameters;

the modality (e.g. text, image) and format of inputs and outputs;

the licence for the model.

2. A description of the elements of the model and of the process for its development, including:

(a) the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;

(b) the modality (e.g. text, image, etc.) and format of the inputs and outputs and their maximum size (e.g. context window length, etc.);

(c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.

4.2.3. Copyright policy

GPAIM providers are obliged to put in place a policy to comply with Union law on copyright and related rights, and in particular to respect the opt-out (reservation of rights) under the "general" text and data mining exception (Article 4(3) of the DSM Directive). According to the GPAI Code of Conduct, the policy should include a commitment to reproduce and extract only lawfully accessible copyright-protected content when crawling the World Wide Web, to identify and comply with rights reservations, and to designate a point of contact and enable the lodging of complaints.

4.2.4. Summary of the training data

CLARIN, as a provider of data for GPAIM (LLM) training, should also pay special attention to the details concerning the obligation laid down in Article 53(1)(d) of the AI Act, i.e. to provide a sufficiently detailed summary of the training data. In July 2025, the European Commission published an Explanatory Notice and a standardized Template for the Public Summary of Training Content for general-purpose AI models to support providers in meeting their transparency obligations[6].

As a part of the detailed summary of the training data, the model provider must provide, among other things, the following information (below are described only selected elements that are considered particularly relevant):

1) Modalities, overall training data size and other characteristics

This section requires general information about the overall training data after pre-processing and before the training of the model. The following information must be presented: (i) modalities present in the training data (text, image, audio, video, other), (ii) training data size - the provider must select the range within which the estimated total training data size for that modality falls (tokens for text, hours for audio and video, number for images), (iii) types of content (description of the type of content that has been included in the training data)

In this section, it is also necessary to provide inter alia the description of the linguistic characteristics of the overall training data (for example the languages covered by the training data focusing in particular on EU official languages and other relevant characteristics of the overall training data (for example national/regional or demographic specificities of the training data where readily available).

2) List of data sources

This section requires information on the specific data sources used to train the general-purpose AI model. For the purposes of this section, a "dataset" refers to a single, pre-packaged collection of data.

The data sources should be presented according to the following categories: (i) publicly available datasets, (ii) private non-publicly available datasets obtained from third parties, (iii) data crawled and scraped from online sources, (iv) user data, (v) synthetic data, and (vi) other sources of data.

With regard to publicly available datasets, the GPAI model provider is required to list only those datasets classified as large. A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. Dataset size must be based on the post-processed dataset (for example after filtering) and not on artificially split subsets. For other publicly

[6] https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models?utm

available datasets that are not listed as large ones, model provider should provide a general description of their content.

3) Data processing aspects

a) Respect of reservation of rights from text and data mining exception or limitation

This section requires the model provider to describe the measures implemented to respect reservations of rights under the TDM exception or limitation before and during data collection. In particular, the provider should explain the opt-out protocols and technical or organisational solutions honoured, whether applied directly by the provider or, where applicable, by third parties from whom the datasets were obtained.

In this context, it should be noted that certain dataset providers collect data using identifiable web crawlers and publish clear, practical instructions for preventing such collection (e.g., by specifying how the crawler can be blocked through standard access-control mechanisms such as the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309). See for example https://commoncrawl.org/ccbot

b) Removal of illegal content

This section concerns measures taken to avoid or remove illegal content under Union law from the training data (e.g., blacklists, keywords, and model-based classifiers), without requiring disclosure of internal business practices or trade secrets. These measures are advisable where training data is likely to include illegal or unlawful content—particularly child sexual abuse material, terrorist content, and non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices to increase the model's capability. In the General-Purpose AI Code of Practice published by the European Commission, the Copyright chapter envisages publishing, on an EU website, a list of links to websites that have been determined by EU or EEA courts or authorities to be "persistently and repeatedly infringing copyright and related rights on a commercial scale." As of now such a list has not been published yet.

If a dataset has already been filtered by its provider to remove illegal or unlawful content before training, it would be particularly valuable for AI model providers because it reduces legal and compliance risks and increases the reliability of the data used.

4.3. Concept of Downstream Modifiers of General-Purpose AI Models

In the context of GPAIMs, downstream modifiers constitute an interesting case, especially in scientific setups. Downstream modifiers are entities or actors that take an existing general-purpose AI model or system and modify it for a specific downstream use or application. In this context, 'upstream' refers to the original developer or provider of a GPAIM, such as a foundation model, while 'downstream' describes the later stages of the value chain, where the model is adapted, integrated, or deployed for specific, concrete tasks. Typical tasks of downstream modifiers include (i) fine-tunning or further training of a GPAIM on new or domain-specific data; (ii) adapting the model through techniques such as prompting, reinforcement learning, or parameter adjustments; (iii) integrating the model into products, services, or workflows; (iv) changing or extending functionality, performance characteristics, or intended use.

4.3.1. When a downstream modifier becomes a GPAIM

From a technical standpoint, an indicative criterion for when a downstream modifier is considered a provider of a GPAIM relates to the training compute used to perform the modification: when this modification requires more than one-third of the training compute used for the original model, the downstream modifier is considered to be a GPAIM provider. Where original compute metrics are unavailable, if the original model presents a systemic risk, the threshold should be one-third of 10^{25} FLOPs; otherwise, it should be one-third of 10^{23} FLOPs (cf. Commission Guidelines on the scope of the obligations for providers of general-purpose AI models).

From a regulatory perspective, two articles of the EU AI Act are the most relevant for downstream modifiers and for GPAI models in general – namely, Articles 2(6) and 2(8), listed below:

1) **Article 2(6) – Scientific research exemption**
   The AI Act does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development.
2) **Article 2(8) - Pre-Market research exemption**
   The AI Act does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service.

To illustrate how these articles apply to downstream modifiers that become GPAIM providers, consider a hypothetical scenario where a team of researchers from the Swiss national consortium CLARIN-CH undertakes the development of a comprehensive multilingual model for Swiss language varieties. Building upon an existing foundation model such as Llama 3.1 70B, the researchers conduct extensive continued pre-training using their curated collections of Swiss German dialects, Romansh language corpora, and historical Swiss multilingual parliamentary proceedings. The training process incorporates approximately 2 billion tokens drawn from digitized archives, contemporary media sources, and linguistic fieldwork repositories. This computational effort requires resources equivalent to 40% of the original model's training compute, exceeding the one-third threshold established in the Commission Guidelines. Therefore, under the EU AI Act, the researchers responsible for the development of such a model no longer fall under the category of downstream modifiers, but under that of providers of a new GPAIM.

Following standard academic practice, the researchers publish their findings in a peer-reviewed venue and, in the interest of scientific reproducibility, upload the fine-tuned model weights to Hugging Face, which is a widely-used repository for machine learning models. The act of uploading to repositories such as Hugging Face or Github is not clear from a regulatory perspective because this may be interpreted either as research dissemination or as placing on the market. If the model release is interpreted as market placement, the research would no longer qualify for the Article 2(6) research exemption, despite the fact that the model was developed within an academic context and served primarily scientific objectives. This classification would imply obligations stated in Article 53, including technical documentation requirements, implementation of copyright compliance policies, and provision of training content summaries. At present, there is uncertainty surrounding this kind of case, which demonstrates the need for Commission guidance on how open science practices (particularly model sharing for reproducibility purposes) should be reconciled with the AI Act concerning the market placement of the models.

4.3.2. When a downstream modifier remains a downstream modifier

A downstream modifier is not automatically a GPAIM provider. The Commission makes clear that not every modification of a GPAIM will lead to the downstream modifier being considered a provider of GPAIM. They only become the provider of a (new) GPAIM under specific conditions regarding training compute, i.e., when the modification requires more than one-third of the training compute used for the original model (see section above).

For purposes of illustration, consider one more hypothetical scenario in which the CLARIN-CH undertakes a domain-specific adaptation of an existing large language model for humanities research purposes. The consortium research team initiates a project to develop specialized capabilities for processing historical Swiss German texts from the 17th-19th centuries, drawing upon their extensive digitized archival collections. Rather than training a model from scratch, the research team uses Llama 3.1 70B as a foundation and conducts targeted fine-tuning using approximately 500 million tokens of historical texts curated from their repositories. The computational resources required for this fine-tuning operation represent roughly 20% of the original model's training compute. Since this modification falls below the critical one-third threshold established in the Commission Guidelines, CLARIN-CH does not cross the indicative criterion that would designate them as the provider of a new general-purpose

AI model. Consequently, the consortium maintains its status as a downstream modifier rather than becoming a GPAI model provider under the EU AI Act.

As stated previously, Article 2(6) of the AI Act explicitly excludes AI models developed and put into service for the sole purpose of scientific research and development, provided they are not placed on the market. In this scenario, these conditions are satisfied, since the modified model serves exclusively academic research in humanities, distribution occurs only within the authenticated CLARIN research infrastructure, and no commercial offering or market placement takes place. Documentation clearly establishes the research-only purpose of the development. As in the previous scenario, potential complications arise if CLARIN-CH considers broader dissemination strategies, like uploading the model to repositories, such as Hugging Face. Even though the decision of uploading to such a repository is consistent with open science principles, it may still be interpreted either as research dissemination or as placing on the market (see Article 2(8) of the AI act).

Despite the lack of consensus on the legal requirements concerning the fine boundary between placing a model on the market and putting a model into service for research purposes only, CLARIN can already implement several best practices. These include maintaining comprehensive documentation of the modification's research purpose and implementing robust access controls through existing federated authentication infrastructure. Additionally, CLARIN can use its established metadata infrastructure to document training data sources, licenses, and model capabilities, as stated in the transparency objectives of Article 53 (see above). These proactive measures align with CLARIN's longstanding commitment to the FAIR principles and put the consortium in a position to smoothly adapt to regulatory requirements once they are clarified by the Commission.

5 Conclusion

While the AI Act formally excludes AI systems developed solely for scientific research from its scope, this exemption should not be read as insulating the CLARIN community from the AI Act's practical and strategic consequences. As this paper has shown, the Act's transparency obligations and its regulatory framework for general-purpose AI will shape the environment in which language data and language technologies are developed, shared, funded, and used. For CLARIN-related activities, the relevance of the AI Act lies less in immediate compliance duties for academic researchers than in its indirect effects on research infrastructures, collaborations with industry, downstream uses of research outputs, and the expectations placed on future professionals trained within the community. Engaging proactively with the AI Act is therefore not merely a matter of legal awareness, but a necessary step in safeguarding the sustainability, openness, and societal legitimacy of language resources and technologies in an increasingly regulated AI ecosystem.

References

Bertuzzi, Luca. 2023. "Leading EU lawmakers propose obligations for General Purpose AI". Euractiv.com, https://www.euractiv.com/section/artificial-intelligence/news/leading-eu-lawmakers-propose-obligations-for-general-purpose-ai/ (last visit: 11.04.2025).

Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091.

Finck, Michèle, In Search of the Lost Research Exemption: Reflections on the AI Act, GRUR International, Volume 74, Issue 10, October 2025, Pages 903–904, https://doi.org/10.1093/grurint/ikaf100

Kaur, Pravneet, Gautam Siddharth Kashyap, Ankit Kumar, Md Tabrez Nafis, Sandeep Kumar, and Vikrant Shokeen. 2024. "From Text to Transformation: A Comprehensive Review

of Large Language Models' Versatility." arXiv. https://doi.org/10.48550/ARXIV.2402.16142.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey." arXiv. https://doi.org/10.48550/ARXIV.2111.01243.

Ruschemeier, Hannah. 2023. "AI as a Challenge for Legal Regulation – the Scope of Application of the Artificial Intelligence Act Proposal." *ERA Forum* 23 (3): 361–76. https://doi.org/10.1007/s12027-022-00725-6.

Smuha, Nathalie A. and Rengers, Emma and Harkens, Adam and Li, Wenlong and MacLaren, James and Piselli, Riccardo and Yeung, Karen, How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act (August 5, 2021). Available at SSRN: https://ssrn.com/abstract=3899991 or http://dx.doi.org/10.2139/ssrn.3899991

Wörsdörfer, Manuel. 2025. "The E.U.'s Artificial Intelligence Act: An Ordoliberal Assessment." *AI and Ethics* 5 (1): 263–78. https://doi.org/10.1007/s43681-023-00337-x.

Advancing FAIR Language Data through Training and Capacity Building: the Swiss contribution to CLARIN

Joanna Blochowiak
CLARIN-CH
University of Zurich, Switzerland
joanna.blochowiak@uzh.ch

Cristina Grisot
CLARIN-CH
University of Zurich, Switzerland
cristina.grisot@uzh.ch

Abstract

The CLARIN-CH ecosystem is unique in its approach to language resources, as it brings together a diverse range of research infrastructures distributed across the various language regions of Switzerland. In a larger European context, we present in this paper the CLARIN-CH training program, which plays a central role in Switzerland's efforts to advance sustainable Open Research Data practices within the evolving landscape of Open Science. The program is structured around five principal axes of action: (i) assessing the needs of researchers across disciplines who work with language data, both within linguistics and beyond; (ii) collaborating with stakeholders, including researchers and training providers, to develop offerings tailored to these specific needs; (iii) organizing diverse training events using a variety of educational formats; (iv) co-developing Open Educational Resources aligned with the FAIR-by-design methodology; and (v) ensuring the publication and long-term accessibility of these resources via Zenodo for the CLARIN-CH community. In this contribution, we outline these core components of the training program and showcase three training initiatives delivered through multiple educational formats. The long-term goal of the CLARIN-CH training program is to contribute to capacity building among researchers working with language data, by promoting a spirit of lifelong learning and interdisciplinary collaboration. As such, Switzerland significantly contributes to CLARIN's efforts put into disseminating knowledge and building capacity about language resources and promote best practices aligned with FAIR and Open Research Data principles.

1 Introduction

The CLARIN-CH training program is a central component of Switzerland's efforts to promote sustainable Open Research Data (ORD) practices and to strengthen interdisciplinary collaboration among researchers working with language data. Rooted in the CLARIN-CH ecosystem, the program supports the development of a robust training infrastructure that addresses the evolving needs of a multilingual and multidisciplinary research community. In addition to researchers, the program explicitly encompasses data stewards and data professionals, who play a crucial role as both invaluable sources of expertise and as key beneficiaries of CLARIN-CH training activities. By fostering dialogue and knowledge exchange between researchers and data professionals, the program contributes to the development of shared, sustainable data practices. Its overarching goal is to enhance the capacity of all stakeholders involved in the research data lifecycle to engage with language data in alignment with FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016), thereby supporting a culture of open science across diverse fields including linguistics, medicine, law, journalism, theology, and beyond.

In this contribution, we describe the Swiss approach to training and capacity building and situate it in the larger context of CLARIN. The paper is structured as follows. Section 2 discusses the importance of training and capacity building for advancing FAIR compliance of language resources, as well for the sustainable development of research infrastructures. Section 3 describes existing training and capacity-building initiatives in CLARIN, both at the central level and at the country level. Section 4 presents on the Swiss approach and its five axes of action. Section 5 showcases three training initiatives provided in the framework of working groups, training sessions targeting linguists and activities targeting non-linguists. Section 6 concludes this contribution.

Joanna Blochowiak and Cristina Grisot 2025. Advancing FAIR Language Data through Training and Capacity Building: the Swiss contribution to CLARIN. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 172–181. DOI: https://doi.org/10.3384/ecp222

2 Why is training and capacity building important

Training and capacity building may target on the one had FAIR and ORD data management practices, and the use of tools and research infrastructure components, on the other hand. Investing into training and capacity building is not a peripheral support activity, but a strategic prerequisite for their sustainable development of research infrastructures at both national and pan-European levels, as well as for empowering students, researchers, and data stewards to bridge the gap between theoretical knowledge and practical aspects of linguistic research data management, and ultimately to do better Open Science and data-driven research (see examples of training opportunities provided by CLARIN countries in section 3). For the rest of this section, we will focus in the importance of training and capacity building for the future of research infrastructures.

First, as research infrastructures increasingly operate as federated ecosystems rather than isolated centres or repositories, their effectiveness depends less on central technical capacity and more on the ability of research communities to produce data that are interoperable, reusable, and legally robust from the outset. In this context, training and capacity building enable technical infrastructures to scale and to federate effectively across countries and disciplines. While technical interoperability can be addressed through shared platforms and standards, the harmonisation of data management practices requires sustained capacity building, which enables the convergence of legal, semantic, and organisational practices supporting the vision of a distributed yet interoperable research ecosystem. In the domain of language resources and language technology, training is a decisive factor in ensuring that European research infrastructures remain interoperable, trustworthy, and strategically relevant in an increasingly AI-driven research landscape.

Second, language data are structurally complex, highly heterogeneous, multilingual, often multimodal, and frequently involve human subjects, which introduces significant legal and ethical constraints. The compliance to FAIR principles at onset, i.e. data creation and documentation, has a positive impact on making them usable for cross-project, cross-border, or long-term reuse. For this reason, systematic training and capacity building shifts FAIR and ORD practices upstream – at the level of researchers themselves, where they are most effective and impactful.

Finally, training and capacity building are also a key governance and sustainability mechanism. National and pan-European research infrastructures, such as CLARIN, but also other thematic infrastructure, such as DARIAH, SSHOC, as well as EOSC, rely on data quality, standardised metadata, and clear access conditions to operate efficient ingestion and service pipelines. Training empowers researchers for their scientific endeavours, but also data stewards and other data management professionals to act as co-producers of FAIR and ORD compliant resources. Consequently, this strengthens trust between national and European research infrastructures and their communities, and reinforces credibility within European research infrastructure consortia.

3 Training and capacity-building initiatives in CLARIN: centrally and in member countries

3.1. CLARIN Central Hub

The creation of the CLARIN Learning Hub[1] is the culmination of many efforts put by CLARIN into training for the last fifteen years. For CLARIN, training and capacity building is one of its core services rather than an auxiliary activity, which has been gradually developed as summer schools, tutorials, and workshops, national training initiatives, knowledge centers and competence networks, including coordination of training materials across member countries. In addition, CLARIN's participation in various EU-funded projects (such as Upskills[2]), its alignment with EOSC capacity-building agendas (see project skills4EOSC[3]), and increased emphasis on skills for FAIR data, reproducibility, and digital methods led to the expansion of the CLARIN Learning Hub, as it is today. The Hub aggregates training materials, tutorials, courses, and events produced across the CLARIN network and makes them discoverable and reusable for researchers, educators, and infrastructure professionals. Its underlying logic is federated

[1] https://www.clarin.eu/content/learning-hub
[2] https://upskillsproject.eu/
[3] https://www.skills4eosc.eu/

and community-based: training content is created by distributed experts in national consortia and knowledge centers, then curated centrally to ensure visibility, reuse, and continuity across countries and disciplines.

In terms of thematic scope, the CLARIN training offer covers the full lifecycle of language data and technology use in research. Topics range from applied language technology, corpus linguistics, and computational methods to practical skills such as programming for NLP, data annotation, and standards and repositories (see van der Lenk & Fišer, 2023 *Introduction to Language Data: Standards and Repositories*[4]). There is also a strong emphasis on Open Science, Open Research Data, and FAIR practices, especially in sessions that explain how infrastructures enable reproducibility, sharing, and linking of language datasets to publications. This positioning shows that training is not only about technical tool adoption, but also about embedding good research data management practices into everyday research workflows.

A distinctive feature of CLARIN's approach is its integration of training with infrastructure access, see for example the *CLARIN101 tutorials*[5]. Learning resources in these tutorials are tied directly to operational services such as federated search systems (the CLARIN Virtual Language Observatory and the CLARIN Federated Content Search), repositories maintained by CLARIN B-centers, virtual collections, and language processing tools such as the CLARIN Switchboard. Training therefore serves as an onboarding mechanism into the infrastructure ecosystem, helping users move from awareness to hands-on use.

The Learning Hub also supports methodological and pedagogical innovation. It promotes self-study materials, reusable tutorials designed following the FAIR-by-Design principle (Filiposka et al., 2025), ensuring that educational resources are findable, accessible, interoperable, and reusable. Training content thus can be embedded in university curricula, reflecting a long-term strategy to institutionalize infrastructure literacy within higher education. Recent initiatives around FAIR-by-design training materials further show that CLARIN is not only teaching FAIR data practices, but applying them to the training ecosystem itself, such as the *CLARIN Trainers' Toolkit* (van der Lek, 2025). To facilitate the findability of training resources, *CLARIN's Learning Resources Catalogue*[6] is a non-exhaustive inventory of learning resources developed in the CLARIN community and allows users to search by topic, skill level, resource types, media type, license, publication and modification date. Each entry is described with consistent metadata, allowing learners and educators to find the materials corresponding to their needs, but also to understand how and by whom the materials were created.

Overall, CLARIN's training model is infrastructure-centric, distributed, and lifecycle-oriented. It addresses technical competencies (tools, NLP, corpora), data stewardship practices (FAIR, ORD, metadata), and research integration (linking data, services, and publications). This makes training a strategic instrument for strengthening the adoption, sustainability, and cross-border interoperability of language resources and technologies across the European research landscape.

3.2. Overview of CLARIN member countries

In parallel to the centralized efforts made by CLARIN ERIC, CLARIN Knowledge Centers and national initiatives provide domain-specific guidance, educational materials, and expert support, reinforcing the link between skills development and sustainable infrastructure uptake. The CLARIN Learning Hub[7] provides a centralized entry point for the training and capacity building initiatives from CLARIN countries, allowing users to find their way (see non-exhaustive overview in Table 1). Training materials may be available in English, but also in national languages.

National consortia	Open Learning and Training Resource
CLARIN-IT[8]	materials for the Italian community and beyond; H2IOSC Training Environment

[4] https://www.clarin.eu/content/introduction-language-data-standards-and-repositories-0
[5] https://www.clarin.eu/content/intro-clarin-tutorial
[6] https://www.clarin.eu/learning-resources-catalogue
[7] https://www.clarin.eu/content/guidelines-and-best-practices
[8] https://clarin-it.it/

CLARIAH-NL[9]	materials for Linguistics and Media Studies
CLARIAH-VL[10]	materials on national services, tools and resources, Digital Humanities
CLaDA-BG[11]	materials for Bulgarian Language Technology for Digital Humanities
CLARIN-DK[12]	reports, presentations, tutorials related to the language technologies hosted in the Danish infrastructure, and topics such as FAIR and data management topics
CLARIN:EL[13]	user manual and video tutorials to help users explore and use the CLARIN:EL research infrastructure
CLARIN-LV[14]	videos and tutorials
CLARIN-PL[15]	practical guides, video tutorials, and workshop materials on how Polish applications and infrastructure services are used in research practice
CLARINO[16]	extensive documentation and tutorials on tools and services integrated in the Norwegian infrastructure.
CLARIN-SI[17]	training events and workshops aimed to teach university students, linguists, language teachers and DH scholars how to use the CLASSLA web corpora and NoSketch Engine concordance in South Slavic language research
FIN-CLARIN[18]	guidelines, introduction videos, courses and training materials related to topics related to text and speech processing and analysis, language technology and data management topics
LINDAT/CLARIAH-CZ[19]	slides, videos and tutorials on how to use the LINDAT repositories, concordancers and other tools and services, including best practice guidelines
Portulan CLARIN[20]	workbench of tools and services with embedded documentation
CLARIN-CH[21]	education and training sessions, webinars, workshops
CLARIN-UK[22]	workshops and tutorials on tools and services, and maintains a collection of datasets used in teaching

Table 1 Overview of Open Learning and Training Resources within CLARIN consortia (from CLARIN Learning Hub, made by Iulianna van der Lek, CLARIN Training Officer)

An outstanding example for its complexity, modularity and capacity to serve several languages is the CLASSLA-Express, a workshop series organized under the CLASSLA Knowledge Centre[23] within the Slovenian CLARIN.SI infrastructure. CLASSLA focuses on South Slavic languages and serves as a regional hub for corpus resources, tools, and expertise. CLASSLA-Express was conceived as a flexible and mobile training format that delivers *hands-on, practice-oriented instruction* directly to researchers, students, and other stakeholders working with language data across several South Slavic countries, including Croatia, Serbia, North Macedonia, Bulgaria, and Slovenia. The training program covers a wide methodological spectrum, ranging from introductory corpus-linguistic techniques, such as concordancing, frequency analysis, and corpus exploration using the CLASSLA-web platform, to more advanced topics, including the application, evaluation, and critical assessment of large language models in corpus-

[9] https://clariah.nl/
[10] https://clariahvl.hypotheses.org/training-materials
[11] https://clada-bg.eu/en/dissemination/videos.html
[12] https://info.clarin.dk/en/publications/presentations-posters-etc./
[13] https://www.clarin.gr/en/support/user-manuals
[14] https://www.clarin.lv/en-us/publications
[15] https://clarin-pl.eu/
[16] https://clarin.w.uib.no/
[17] https://www.clarin.si/info/about/
[18] https://www.kielipankki.fi/language-bank/
[19] https://www.lindat.cz/#education
[20] https://portulanclarin.net/
[21] https://clarin-ch.ch/services/education-training/
[22] https://www.clarin.ac.uk/oxford-text-archive
[23] https://www.clarin.si/info/k-centre/

based research. In addition to technical skills, the workshops place emphasis on methodological transparency and the informed reuse of shared language resources, aligning with broader CLARIN goals related to open and reusable research data. A defining feature of CLASSLA-Express is its deliberate decentralized design. Rather than concentrating training activities in a single institutional setting, the program is structured to travel to different locations and host institutions, which lowers participation barriers and extending its reach to geographically dispersed and linguistically diverse research communities. This approach enables close engagement with local research contexts while fostering cross-border collaboration and contributing to the development of a shared regional community of practice around language data.

Taken together, these initiatives illustrate a continual emphasis on bridging technical tool literacy with domain-specific research needs, promoting FAIR and ORD awareness. This broader CLARIN context situates the CLARIN-CH training program within a rich ecosystem of capacity-building efforts that address language data challenges in multifaceted research environments.

4 Axes of action of the CLARIN-CH training approach

The CLARIN-CH training approach is created within the collaborative and community-driven approach and is structured around five principal axes of action, each contributing to the development of a sustainable training infrastructure that supports researchers working with language data across Switzerland's multilingual and interdisciplinary research landscape.

First, a key priority is the systematic **evaluation of the needs of researchers** who work with language data across various disciplines, both within linguistics (see Grisot, 2024) and beyond (see Blochowiak et al., under review). This includes researchers in fields such as medicine, law, journalism, theology, and business, where language data plays an increasingly central role. By conducting targeted needs assessments, such as interviews and surveys, the program aims to gain a comprehensive understanding of the existing competencies, practices, and gaps in knowledge related to ORD and the use of Natural Language Processing techniques. For instance, among the main gaps pointed out by linguists in the 2023 CLARIN-CH Survey (see Grisot, 2024) were numerous, such as data management and legal support, access to large, diverse and multimodal corpora, access to NLP tools and reusable components, access to functional annotation tools, and training in corpus building, coding and data analysis and statistical methods. Outside linguistics, for instance in the area of Film Studies, the main problem is that prohibitive copyright restrictions on audiovisual material establishes a culture of data silos, where datasets are stored on private drives because legal pathways for sharing remain too opaque (see Blochowiak et al., under review). This evidence-based approach ensures that the training offer is built on a clear understanding of actual research needs, grounded in the day-to-day challenges faced by researchers.

Second, the program places strong emphasis on **collaboration with key stakeholders** across the Swiss research ecosystem. This includes engaging both researchers and training providers in co-developing educational content and formats that are tailored to specific disciplinary and linguistic contexts. Through these partnerships, CLARIN-CH fosters a participatory and community-driven model of training design, ensuring that the resulting materials and activities are relevant, adaptable, and aligned with institutional and national priorities related to Open Science and FAIR data practices.

Third, the training program is committed to **organizing a diverse portfolio of training events**, delivered in a variety of educational formats to maximize accessibility and impact. These include interactive working groups, thematic webinars, and practical workshops held both onsite and online. This multimodal approach allows the program to reach a broad and varied audience, from early-career researchers to experienced data stewards, while accommodating different learning preferences and institutional contexts. Importantly, many of these events are designed to foster interdisciplinary exchange and knowledge-sharing, creating spaces where participants can learn from one another and explore how tools and methods can be applied to their specific datasets, but also across disciplinary boundaries.

Fourth, CLARIN-CH works closely with training providers and domain experts to **develop high-quality Open Access Educational Resources** that are aligned with the FAIR-by-design methodology developed in the skills4EOSC project. These resources cover each stage of the language data life cycle from data collection, processing, and annotation to ethical considerations, legal compliance, sharing, and long-term archiving. Special attention is given to discipline-specific needs and to demonstrating the

added value of tools such as the LiRI Corpus Platform[24] (University of Zurich) and Swiss-AL[25] (Zurich University of Applied Sciences ZHAW) for researchers working outside linguistics. The materials are designed to be modular, reusable, and adaptable for both self-directed learning and formal teaching contexts.

Fifth and finally, the training program ensures the **publication and long-term accessibility of all educational resources** via Zenodo[26]. By using this open and trusted repository, CLARIN-CH guarantees that training materials are freely available to the broader research community and can be easily cited, shared, and integrated into other teaching or training initiatives. This axis of action reflects the program's strong commitment to sustainability and transparency, and its role in advancing open science practices not only in Switzerland, but also within the wider CLARIN and European research infrastructures.

5 Educational program

Guided by these core principles, CLARIN-CH is developing a training infrastructure designed to meet the needs of researchers working with language data within Switzerland's linguistically and disciplinarily diverse academic environment. In this section, we highlight three initiatives led by the CLARIN-CH team, each utilizing distinct educational formats and addressing the needs of different target audiences involved in language data research.

5.1 Working Groups

As part of the CLARIN-CH training program, several Working Groups (WG) have been created. In what follows we describe two of them: the first is dedicated to sensitive and personal data management (Section 5.1.1) and the second focuses on management of language research data (Section 5.1.2).

5.1.1 The CLARIN-CH sensitive and personal data management Working Group

The first CLARIN-CH Working Group is dedicated to sensitive and personal data management: *Working Group on the management of sensitive and personal data, ethical and legal issues for linguistic data*[27]. Launched in September 2023, this WG aims to serve as a national platform for disseminating information, promoting best practices, and offering tailored guidance to researchers across disciplines who engage with language data, especially in multilingual and cross-disciplinary settings. The WG began its activities with a kick-off event that brought together experts and researchers to discuss critical issues related to data collection, protection, and long-term preservation. Particular attention was given to the specificities of various types of linguistic data, such as multimodal, historical, experimental, sociolinguistic, and social media-based data, as well as data collected from diverse demographic groups. The event featured three keynote presentations (recorded and made available on the CLARIN-CH Documentation Platform[28]) alongside an open call for questions from the community. These questions, often discipline-specific, highlighted widespread concerns around the management of sensitive and personal data, as well as copyright and licensing issues, particularly in relation to digital and online sources.

In response to these expressed needs, the WG designed and launched a monthly webinar series for the 2024 spring and autumn semesters. These lunchtime talks feature a 30-minute expert presentation followed by a 30-minute open discussion. The topics covered include both legal and technical aspects of working with sensitive data, as well as practical case studies, such as anonymization techniques for textual, video, and audio data, and legal considerations for cross-border research or social media data use. Each session is documented through recordings and presentation slides, which are made publicly available via the CLARIN-CH Documentation Platform. In addition to the webinars, the WG produces concrete outputs designed to support researchers in practice. These include: (i) a growing FAQ section compiling answers to community questions; (ii) slides and video recordings from the keynote talks and webinars; (iii) personalized advice (when needed) and guidance provided by the WG members or through collaboration with legal and institutional experts.

[24] https://www.liri.uzh.ch/en/services/LiRI-Corpus-Platform-LCP.html
[25] https://www.zhaw.ch/en/linguistics/research/swiss-al
[26] https://zenodo.org/communities/clarin-ch/records?q=&l=list&p=1&s=10
[27] https://clarin-ch.ch/about/working-groups/sensitive-personal-data/
[28] https://clarin-ch.ch/documentation-platform/

5.1.2 The CLARIN-CH Working Group on research data management for language data

The second CLARIN-CH Working Group focuses on research data management practices for linguistic research: *Working Group on Research Data Management (RDM) for language data*[29]. Launched in October 2025, this WG aims to strengthen RDM practices across the Swiss linguistic research community by leveraging the expertise of institutional data stewards and research support services. The WG strives to serve as a collaborative platform for developing resources, training, and support that improve awareness, capacity, and consistency in managing language data, while laying the foundation for a future CLARIN-CH Training Program aligned with national and international RDM standards. The WG began its activities with a kick-off workshop collocated with the CLARIN-CH Day 2025 at the University of Lausanne, bringing together data stewards, research support professionals, and linguists to discuss the challenges and opportunities of implementing FAIR principles for language data. Particular attention was given to the specificities of language data management across various research contexts, including multilingual corpora, endangered language documentation, experimental linguistic data, and computational linguistics resources. The discussions highlighted critical needs around Data Management Plans (DMPs), institutional alignment on RDM practices, and the integration of domain-specific requirements within broader Open Research Data frameworks.

To address these needs, the WG designed an action plan targeting specifically the DMPs, as they stood out being of particular interest to researchers as well as to data professionals. As a first step, the WG convenors started to work on a two-part DMP workshop series: Part I addresses the general principles and importance of DMPs for SNSF-funded projects, emphasizing early-stage planning and integration into the research lifecycle; Part II provides hands-on guidance with concrete examples tailored to language data management scenarios. Each session will be documented, and materials will be made available through the CLARIN-CH Documentation Platform to support broader adoption across institutions. In addition to the training activities, the aim of the WG is to produce concrete outputs designed to support researchers and data professionals in practice. These include: (i) best practice guidelines for FAIR-compliant language data management; (ii) DMP templates and examples specific to linguistic research; (iii) a comparative review of DMP tools and templates used across Swiss higher education institutions; (iv) training materials and resources accessible through a centralized repository; (v) personalized consultation and guidance provided by WG members in collaboration with institutional data stewardship teams. The WG fosters strategic collaboration with institutional and national partners, including major Swiss universities, the Swiss Academy of Humanities and Social Sciences, swissrn.org, and doctoral schools, ensuring that outputs are broadly relevant and sustainable. This collaborative approach positions the WG as a key contributor to professionalizing RDM support for language data across Switzerland and establishing CLARIN-CH as the national reference point for FAIR and ORD-compliant practices in linguistic research.

5.2 Training sessions for linguists

As part of its commitment to fostering sustainable and inclusive practices in working with language data, the CLARIN-CH training program organized a series of online training sessions in Spring 2025, titled *Exploring Swiss Language Resources and Tools*[30]. This initiative aimed to introduce researchers, both newcomers and experienced users, to the rich ecosystem of infrastructures, tools, and services developed within CLARIN-CH with a precise aim to offer practical, hands-on guidance on how to effectively leverage these resources for research purposes. The training series was designed to support the growing community of researchers relying on language data across a wide range of disciplines. Participants were introduced to a variety of topics, including corpus querying and analysis, media data processing, multimodal data modeling, and best practices for data sharing and reuse. The sessions were developed with the support of two nationally co-funded projects: *Upgrading the Linguistic ORD Ecosystem* (UpLORD[31]) and *Moving towards a FAIR-Compliant Ecosystem of Federated Infrastructure for Language Data* (FAIR-FI-LD[32]).

[29] https://clarin-ch.ch/about/working-groups/wg-research-data-management/
[30] https://clarin-ch.ch/training-sessions-2025/
[31] https://www.liri.uzh.ch/en/projects/Completed-Projects/UpLORD.html
[32] https://www.liri.uzh.ch/en/projects/Completed-Projects/FAIR-FI-LD.html

The series was delivered collaboratively by key institutions and stakeholders in the CLARIN-CH network, including the CLARIN-CH Coordination Office, the Linguistic Research Infrastructure (LiRI) at the University of Zurich, the ZHAW Digital Discourse Lab, the Language Repository of Switzerland (LaRS@SWISSUbase), and Università della Svizzera italiana (USI). Each session featured expert contributions from representatives of these institutions which enabled the dissemination of up-to-date knowledge and the showcasing of current best practices. The sessions were held biweekly and offered participants a structured journey through the CLARIN-CH ecosystem and its tools for working with language data. A total of 98 participants from Switzerland and beyond took part in the online training sessions. For lack of resources, the training sessions were carried out only in English. In the future, depending on the available financial resources, we plan to create training materials in all Swiss national languages to ensure equitable access.

In line with FAIR principles and the Open-by-design principles, the materials produced for these sessions were developed as Open Educational Resources (OER) and published on the CLARIN-CH Zenodo Community, ensuring long-term accessibility and reusability, as shown in Table 2.

Training session	KPIs
CLARIN-CH Training Session#1: Introduction to the CLARIN-CH ecosystem of infrastructures in the era of Open Research Data	67 views, 69 downloads
CLARIN-CH Training Session#2: LiRI Corpus Platform: Data Model and Query	64 views, 76 downloads
CLARIN-CH Training Session#3 Querying and analysing Swiss media data with Swissdox@LiRI	44 views, 53 downloads
CLARIN-CH Training Session#4: LiRI Corpus Platform: Data Analysis Using Text, Audio, and Video Corpora	54 views, 54 downloads
CLARIN-CH Training Session#5: Modelling of Multi-Modal Data in LiRI Corpus Platform and Beyond	54 views, 57 downloads
CLARIN-CH Training Session#6: Data analysis with Swiss-AL for textual data	47 views, 23 downloads
CLARIN-CH Training Session#7: Publishing Data on the Language Repository of Switzerland (LaRS)	60 views, 53 downloads
CLARIN-CH Training Session#8: Reusing data for teaching and secondary analyses	49 views, 14 downloads

Table 2 CLARIN-CH Training Series and Zenodo KPIs (status on March 31, 2026)

Collectively, the series provided comprehensive, practice-oriented training designed to strengthen the capacity of researchers working with linguistic data in line with FAIR principles and ORD standards.

5.3 Emerging initiative for non-linguists

As part of its ongoing efforts to enrich its educational offering, CLARIN-CH has begun laying the groundwork for developing training programs tailored to non-linguistic disciplines. As language data continues to gain prominence across a broad spectrum of research disciplines beyond linguistics, including medicine, history, theology, business communication, translation and interpretation, and language education, new challenges have emerged regarding the management of such data in line with ORD principles. Researchers in these fields often use language data in ways that diverge significantly from traditional linguistic approaches. For instance, while a linguist may examine syntactic structures, a medical researcher might analyze communication strategies to improve patient care. More specifically, corpus-based studies have become a valuable tool for analyzing the specialized language of these domains, since they are crucial for addressing such domain-related research questions effectively. In recent years, however, conducting corpus-based research has become increasingly regulated, driven by growing positive pressure to define requirements and implement ORD practices, as pointed out by disciplinary experts interviewed in Blochowiak et al. (under review).

A survey conducted by CLARIN-CH within the Swiss community in 2023 highlighted an acute need for structured training in ORD practices (Grisot, 2024). Researchers expressed a strong desire for guidance in technical, legal, and ethical aspects of working with language data. While past initiatives, such

as webinars, workshops, and training programs, have made strides toward addressing these needs, they have often been limited in scope, sporadic in implementation, and focused primarily on linguistics. Researchers from other disciplines who use language data have been left underserved, due to their unique approaches and specific research objectives. To address this gap, the CLARIN-CH ecosystem has launched a new initiative called *Language Data Across Disciplines* aimed at strengthening ORD competencies across the broader research community in Switzerland. The LaDaD initiative is currently in its conceptualization phase and includes the following fundamental disciplines: medicine, history and political sciences, theology, film studies, financial and business communication, translation, terminology and interpretation, and language education. It has three main objectives. The first objective is to develop a comprehensive understanding of the competencies needed to manage language data in these fields at each step of the data life cycle. A needs assessment, conducted through a series of guided interviews as well as a national survey, will identify specific challenges faced by researchers in adhering to ORD practices. The second objective is to promote the development of a collaborative research community gathering researchers and data stewards from different fields working with language. The community will seek to exchange knowledge, address common challenges, and establish a network of researchers and data stewards skilled in managing language data in compliance with FAIR principles. As a result, the community will provide actionable recommendations for composing a comprehensive compendium of ORD best practices. The third objective is the development of training sessions to address gaps in ORD knowledge and skills. These resources will include tutorials and practical guides that showcase best practices for working with language data tailored to the disciplinary needs of researchers working with language data. By promoting a deeper understanding of ORD practices and creating reusable training resources, the project aims to contribute to the sustainable advancement of research data management of language data in the spirit of open science beyond linguistics.

6 Conclusion

In this paper, we discussed how CLARIN ERIC and other national CLARIN national consortia invest into training and capacity building with the aim to empower users, such as students and researchers, to do better Open Science and data-driven research by bridging technical tool literacy with domain-specific research needs and raising FAIR- and ORD-compliance awareness. In alignment with these efforts, the CLARIN-CH training program stands as a key pillar in fostering sustainable and inclusive ORD practices for researchers and data stewards working with language data in Switzerland. Through its multidimensional approach based on five axes of action, ranging from targeted needs assessment to the development of FAIR-aligned educational resources, it supports a growing, interdisciplinary community engaged in Open Science. By continuing to expand its initiatives and reach beyond linguistics, CLARIN-CH is laying the foundation for a robust and collaborative research environment across disciplines.

When viewed in the broader context of training initiatives developed across the CLARIN network, such as decentralized workshop series or federated learning hubs in other countries, the CLARIN-CH program distinguishes itself through its strong emphasis on systematic needs assessment, close integration of data stewards and data professionals, and its explicit focus on disciplines beyond linguistics. By combining international best practices with solutions tailored to Switzerland's multilingual and interdisciplinary research landscape, CLARIN-CH contributes to the diversification and strengthening of the European CLARIN training ecosystem. The CLARIN-CH training program is currently largely dependent on project-based funding, which can limit the continuity and long-term planning of training activities. As a result, ensuring a comprehensive and consistently updated training offer across disciplines and also across Switzerland's linguistic regions remains challenging, particularly beyond the duration and scope of specific funded initiatives.

By continuing to expand its initiatives and fostering collaboration across disciplinary, institutional, and professional boundaries, CLARIN-CH is laying the foundation for a robust and sustainable research environment in which language data can be managed, shared, and reused responsibly across domains. In doing so, the program not only addresses current training needs, but also supports long-term capacity building aligned with FAIR principles and the evolving practices of Open Science.

References

Blochowiak, J., Fischer, M., Krasselt, J., Liste Lamas, E., Vukovic, T. & Grisot, C. (under review). Going beyond linguistics: assessing the status quo and exploring the needs of disciplines using language data in the Swiss context. Submitted to CLARIN2026.

Grisot, C. (2024). Report on 2023 CLARIN-CH User Survey. Zenodo. https://doi.org/10.5281/zenodo.19458087

Grisot, C., Körner, E., Eckart, T., Osenova, P., Piasecki, M., Lennes, M., & van der Lek, I. (2025, October 17). CLARIN101 - Introduction to CLARIN. CLARIN Annual Conference, Vienna. Zenodo. https://doi.org/10.5281/zenodo.17379981

Filiposka, S., Green, D., Mishev, A., Kjorveziroski, V., Corleto, A., Napolitano, E., Paolini, G., Di Giorgio, S., Janik, J., Schirru, L., Gingold, A., Hadrossek, C., Souyioultzoglou, I., Leister, C., & Pavone, G. (2025). FAIR-by-Design Methodology for learning materials. Zenodo. https://doi.org/10.5281/zenodo.14711525

van der Lek, I., Fišer, D. (2023). Introduction to Language Data: Standards and Repositories. In UPSKILLS Learning. Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.

van der Lek, I. (2025). CLARIN Trainers' Toolkit (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.15029514

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18.

A resource for lexical variation in a pluri-areal context: Introducing LexAT21, the atlas on lexical variation in Austria in the 21st century

Markus Pluschkovits
Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Austria
markus.pluschkovits@oeaw.ac.at

Daniel Schopper
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria
daniel.schopper@oeaw.ac.at

Anja Wittibschlager
Institute for German Studies
University of Vienna, Austria
anja.wittibschlager@univie.ac.at

Jakob Bal
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria
jakob.bal@oeaw.ac.at

Kilian Kukelka
Austrian Centre for Digital Humanities
Austrian Academy of Science
Vienna, Austria
kilian.kukelka@oeaw.ac.at

Olivia Reichl
Austrian Centre for Digital Humanities
Austrian Academy of Science
Vienna, Austria
olivia.reichl@oeaw.ac.at

Abstract

The following contribution introduces LexAT21, the atlas on lexical variation in Austria in the 21st century. The linguistic context of this project is briefly covered in section one, suggesting that areal and social/vertical variation is characteristic for the German language in Austria. This is then followed by a brief description of the dataset of the first two survey rounds of LexAT21 and a short sample analysis, showing the potential of the dataset for linguistic research. We then turn to integration into the CLARIN context. The contribution closes by considering the issue of representing different varieties of languages in standards.

1 Introduction and state of research

The field of language variation is a wide and burgeoning one and has evolved considerably during the last hundred years. However, it is far from easy to capture the complex array of variation within the German language empirically. German within Austria is strongly characterized by pluriareal variation (see e.g., Auer 2021 for a discussion) on all linguistic levels, and shows significant variation not just on dialectal levels, but also in its standard varieties (see e.g., Ammon et al. 2016 or Dürscheid et al. 2020). For the longest time, language variation in German has been

Markus Pluschkovits et al. 2025. Introducing LexAT21. Linköping Electronic Conference Proceedings 222, pp. 182–192. DOI: https://doi.org/10.3384/ecp222

considered mostly in nonstandard varieties in the tradition of German dialectology.

Especially in German dialectology, the study of variation on the lexical level in dialects has a rich tradition, involving large-scale projects in the past such as the Deutsche Wortatlas (Mitzka & Schmitt1951 ff).[2] However, many of these projects typically consider only parts of the lexical spectrum (such as purely dialectal variation or, less commonly, variation in standard varieties, notably for example in the Variantenwörterbuch des Deutschen by Ammon et al. 2016) and differ in their methodology so much that the comparison of the data between different projects would be academically negligent (see Schmidt et al. 2019: 43–44). Thereby, what is still missing is a systematic, large-scale investigation of lexical variation across varieties in a comparable methodology, ideally also considering variation within individual speakers (cf. Schmidt et al. 2019: 44). LexAT21, the atlas on lexical variation in Austria in the 21st century (see Lenz 2025) aims to fill this research gap for German varieties in Austria.

The aim of this paper is to introduce the project and dataset and showcase the potential of the data for further use. We briefly lay out the project context and elaborate on data collection in Section 2. This is then followed by short sample analysis presented in Section 3. Finally, in Section 4, we close with the integration of the data into CLARIN, with a reflection on glottonyms and their role for this data set.

2 The LexAT21 project and data

LexAT21 is the result of a close collaboration between the Special Research Program "German in Austria: Variation – Contact – Perception", funded by the Austrian Science Fund FWF (Funding Agreement FWF F 0 60, see Budin et al. 2019) and the Austrian Centre for Digital Humanities. The aim of LexAT21 is to investigate variation in lexis in contemporary German in Austria, focusing on both variation between speakers (interindividual variation, controlled by e.g., areal factors such as dialects) and within a speaker (intraindividual variation across repertoires on the dialect-standard axis). To achieve this aim, LexAT21 considers both vertical-social and horizontal-areal variation in the data acquisition and analysis.

The data used for LexAT21 is acquired in semi-regular rounds of online-surveys, designed in the open-source tool LimeSurvey (LimeSurvey 2026). As of writing, the third survey round is being conducted, and data from the second survey round is being analyzed, with the first survey round already published on the LexAT21 platform. All surveys are comprised of roughly 40 linguistic variables[3], which are presented as picture stimuli accompanied by some variation of the written question "Wie nennen Sie den abgebildeten Gegenstand?" ('What do you call this pictured object?'), or, in the case of verbs, "Was macht die Person auf dem Bild? Antworten Sie in einem vollständigen Satz." ('What is the person in the picture doing? Answer in a full sentence.'). Informants then have to type their variants in dedicated boxes, differentiating between their registers. At the beginning of the survey, informants had to declare themselves as either speakers of dialect (presented to them with the term "Dialekt (Mundart)") or speakers of colloquial German ("Umgangssprache oder Alltagssprache"), a salient distinction in the context of German in Austria – from survey round two on, speakers were also able to declare themselves as speaking a standard variety exclusively. Their choice influenced the answer boxes for each given stimulus – if speakers identified with speaking dialect, the label displayed for their nonstandard answer was "Dialekt" (respective "Umgangssprache" for those identifying as speakers of more colloquial varieties), and if speakers stated that they speak neither when speaking German, there was no answer field for answers from the

[2] For a recent overview on lexical variation in other languages, see e.g. Sandow et al. (2025) or Durkin (2012).
[3] Meaning in this context a concept which can be named with more than one lemma (i.e, variant). See Tagliamonte (2025: 65–89) for a recent overview of the term in linguistics.

nonstandard register. The answer box for their standard register, which was present in every case, was labelled with one random glottonym for standard, drawn from a pool of three ("Ihr bestes Hochdeutsch", "Ihr österreichisches Hochdeutsch", "Ihr Hochdeutsch", 'Your best High German', 'Your Austrian High German', 'Your High German', see e.g., Koppensteiner 2023 for a discussion of these terms). The term for the standard variety stayed consistent throughout the survey. Previous analysis of the data shows that the choice of term influences the answers given by the informants, see Lenz et al. (in print). There was an additional field for further terms the informants use but not associate with any of their registers, simply asking "Verwenden Sie noch weitere Bezeichnungen?" ('Do you use any further terms?').

Figure 1 illustrates how such a stimulus looks like for the variable HAT. The informant in this example identifies as dialect speaker and had the glottonym "bestes Hochdeutsch" ('your best High German') assigned randomly.

Figure 1: Written stimulus on top asking 'How do you call the garment pictured?', with answer boxes for 'dialect', 'best High German', and 'further answers'.

The mode pictured in Figure 1 (i.e. a picture stimulus in combination with a written task) was used to collect the entire data, with variations on the formulation of the verbal stimuli. All pictures were chosen with a permissive CC-0 license. Overall, 45 different stimuli were tested in survey round one, and 40 stimuli were / are being tested in survey round two and three respectively. The choice of variables for survey round one was made on the basis of a sample of variables from the Deutsche Wortatlas, which allows for diachronic comparison. Variable choice for the following survey rounds was made based on introspection and reference to secondary materials within the LexAT21 team. In addition to the language data, metadata on the informants was collected: brief biographical details such as age, gender and employment with additional, more detailed information on residency history. Furthermore, each speaker was asked to declare themselves as an informant for one specific location (in Austria). Each survey round was considered finished as soon as roughly 1,900–2,000 valid responses were collected (valid here meaning that the reference locality is located in Austria).

At the end of a collection period, the raw answer data is exported from the LimeSurvey instance. We then import the data into an instance of baserow (baserow 2026), an open-source online data platform. In baserow, we are able to view the data in grid views, which we use for the first manual processing the data – among these is the exclusion of informants that are not located in Austria and a brief review of the provided residency history vis-à-vis the locality chosen by the informant to be representative for them (as sometimes, this question is misunderstood and the selected localities are not fully consistent with the informants' residency history). After a basic check of the informant data, we annotate the linguistic data. This is done on a variable-by-variable basis. The data is annotated by

lexical type, usually corresponding to a lemma, but sometimes to a collection of different lemmas (in the case of e.g., compound nouns the analysis might only focus on the lemma of the head, thereby subsuming different modifiers to one lemma). Annotation is done manually instead of programmatically, for two key reasons: first, the effort is – for many variables – quite trivial, and bespoke automatic solutions would simply require more effort than manual annotation. In many cases, filtering in baserow allows to quite easily identify all data points for one specific annotation based on substrings characteristic for a variant (e.g., for TABAK RAUCHEN 'to smoke', all variants subsumed under the annotation *tschicken* will – in contrast to others – include some variation of the grapheme <s>). The second reason for manual annotation is that some of the lemmas simply are only found in nonstandard varieties, and therefore have no normalized spelling version. Often, these variants will also only emerge while the data is being annotated, meaning that dictionary-based solutions would not work. For many variables, the concrete taxonomy of annotations is not known before the data is annotated, which also does not lend itself to an automated solution.

Typically, the complexity of the annotation scheme was highly dependent on the individual variable. We employed some guiding principles to ensure that the different variables remained comparable: variants not designating the variable were annotated as "irrelevant" ('irrelevant') across all variables (e.g., "Zigarette" 'cigarette' as answer for the variable TABAK RAUCHEN 'to smoke' – this was categorized as irrelevant as 'cigarette' is not a variant that refers to the act of smoking). These irrelevant variants were typically rare, suggesting that the choice of stimuli was – in general – appropriate, although some uptick of irrelevant variants can be detected with designations for flora and fauna. A further principle was the subsuming of hapax variants under the label "sonstige" ('other') – some informants reported idiosyncratic variants (e.g., names of family members for the variable OMA 'grandmother'), which were not considered as autonomous variants. We made exceptions for those variants with historic relevance (because they e.g. appear in older atlases) and usually did not subsume them as "sonstige" even if they did not meet the frequency we established for autonomous variants (in most cases five independent appearances).

Once all variables for one survey round are annotated and the annotations are reviewed, the data is transferred from baserow to a postgreSQL relational database. This is done utilizing the baserow API and transforming the data to fit the data model before populating the database with the new round. This workflow is replicable, not just to account for the continuous survey rounds, but also to accommodate changes in the data, such as fixing minor mistakes (e.g., typos in annotation names), whereby – depending on the amount of data that is changed – the data of the survey round may be wiped and imported new.

The database acts as the backend from which the LexAT21 frontend (available online, see Lenz 2025 ff.) pulls the actual data via an API. The data is displayed both on a customizable map and in a table view, offering many options for display and filtering of data. The maps and data tables are easily shareable, as the filter parameters are written directly into the URL. Furthermore, users can switch between the data grid view and the map with the filter parameters staying intact, allowing users to easily visualize data on the map. The map can be customized not just with regard to data but also with regard to visualization parameters, such as base maps and colours of the legend, and can easily be exported in a variety of formats for use in publications and the likes. The datasets and maps are one of the main outputs of LexAT21, and are complimented by the map commentaries.

There are two different types of map commentaries: short comments and full map commentaries. The short commentaries are, as the name implies, simply a brief overview over the dominant variants and typically do not exceed a paragraph in length. Their aim is to concisely inform readers on surface-level variation and typical variants in German as spoken in Austria. The full map commentaries in contrast go into more depth. The full map commentaries, additionally to brief

overviews of the variables in other data sources and atlases, elaborate on the variation in different registers (both standard and nonstandard), and offer curated maps on the individual variable – which are linked back to the mapping tool so readers can explore and customize them. These map commentaries are, just like the data, part of the output of LexAT21 and their integration into the CLARIN infrastructure will be discussed in Section 4. Before turning to this integration, a small sample analysis taken from the LexAT21 dataset aims to lay out the relevance of the data to the community.

3 Sample analysis

This brief sample analysis is based on the variable ZUSAMMENKEHREN ('to sweep') taken from LexAT21 survey round one. For further quantitative analyses see e.g., Stöckle & Pluschkovits (forthcoming) or Ziegler et al. (forthcoming). This present analysis is based on the 1,925 informants of this survey round, out of which 996 identified as speakers of colloquial language ('Umgangssprache') and 929 identified as speakers of dialect. The average informant age of this round was 39, with roughly one third of the informants being below 30, one third between 30 and 50, and one third over 50 years. 1,205 informants identified as female, 703 as male and 17 as diverse. Based on these informants, there were 3,408 relevant individual answers for the variable ZUSAMMENKEHREN (excluding irrelevant and miscellaneous variants, see also Pluschkovits 2025).

The variable ZUSAMMENKEHREN is interesting not just because it was also surveyed in the Deutsche Wortatlas (Mitzka & Schmitt 1953) but also because it has been in the focus of variationist analyses before: as early as Kretschmer (1918: 196), who assured that the variant *kehren* is completely dominant in Austria, but that Viennese speakers consider *fegen* more prestigious (in contrast to Berlin speakers, where *fegen* is dominant but *kehren* is more prestigious). Since then, much has been written about the variation between *kehren* and *fegen*, e.g. in Huesmann (1998) or Muhr (2001), and we find this variable in atlases of colloquial language such as the Atlas deutscher Alltagssprache (AdA 2005) or the Wortatlas der deutschen Umgangssprachen (1978), with more complex variation than what was asserted at the beginning of the 20th century.

In the LexAT21 dataset, we can observe complex variation in all registers. Considering the standard register (Figure 2), we find considerable diversity. It is especially noteworthy that the variant *fegen*, presumed to not be relevant in Austria, actually makes up 12 % of all answers in the standard register[4] with seemingly no areal preference. Furthermore, the simple verb form *kehren* is more common than any individual form with a particle, such as *zusammenkehren* or *aufkehren*.

[4] For brevity, the three different standard registers are considered together.

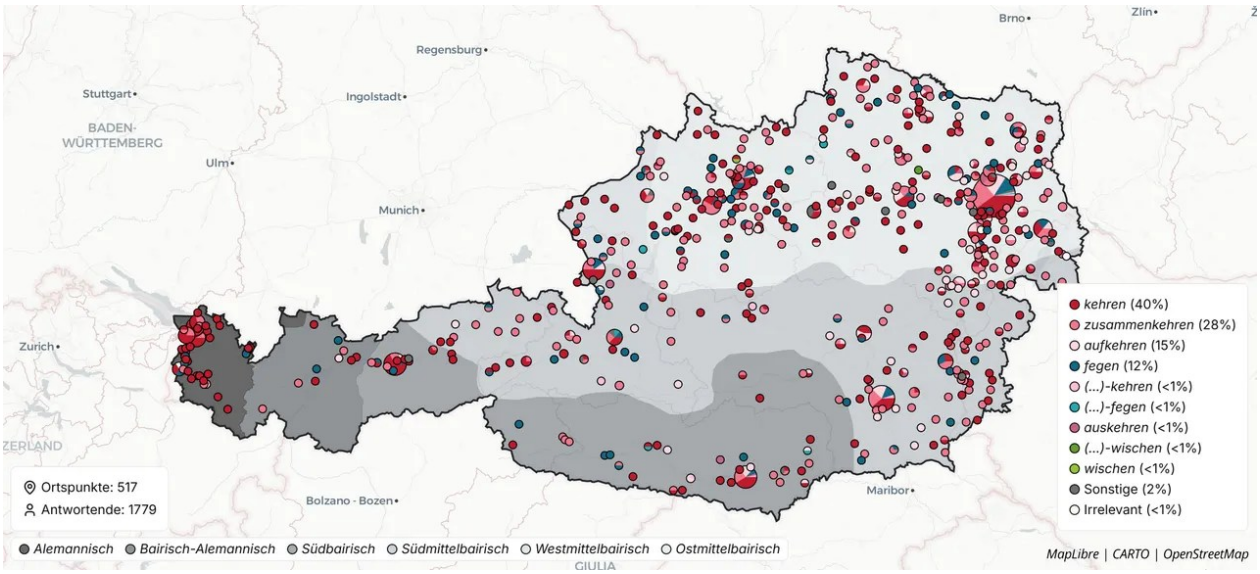

Figure 2: variation in standard registers for ZUSAMMENKEHREN, taken from Pluschkovits 2025

In colloquial language (Figure 3), we can observe this pattern reversed: here, the particle form *zusammenkehren* is clearly more widespread than the base form *kehren*, making up almost half of all data points. Additionally, the variant *fegen* is even more marginal here, with roughly 5 % of all data being *fegen* or a particle verb with *fegen*. We can furthermore observe that colloquial speakers of German tendentially reside more in the eastern than in the western half of Austria (see also Lenz 2025).

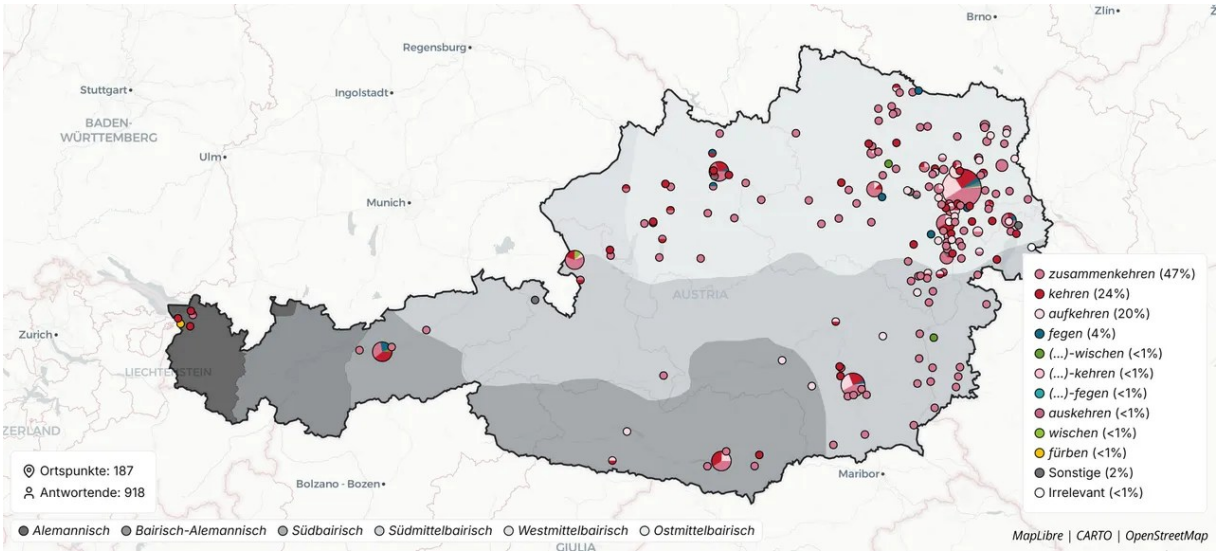

Figure 3: variation in colloquial German for ZUSAMMENKEHREN, taken from Pluschkovits 2025

Figure 4 details the variation of ZUSAMMENKEHREN in dialect. Immediately, one can spot clear differences to the answers of the speakers of colloquial German: *zusammenkehren* is even more dominant than in the colloquial register, and *fegen* even more marginal, accounting for less than 1 % of the data. The variant *fürben*, not present in standard registers and barely present in colloquial German, now accounts for 5 % of all data and is the dominant variant in Alemannic dialects in the west of Austria. We can furthermore observe a different regional spread of data in contrast to colloquial speakers, which are only marginally relevant in the west of Austria – dialect-identified speakers appear all through Austria and constitute the majority of speakers in the west.

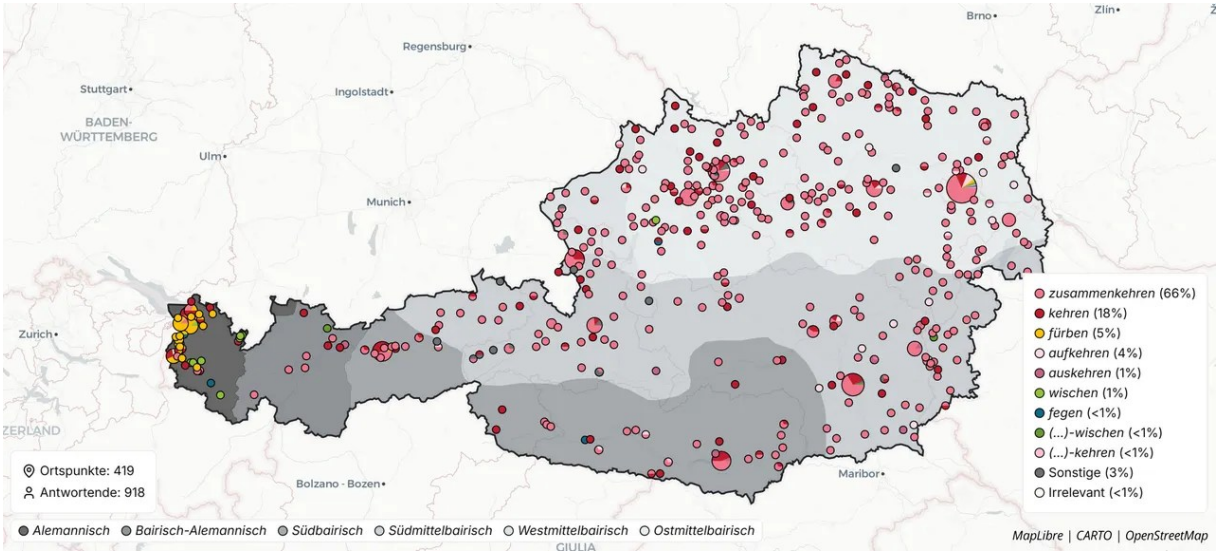

Figure 4: variation in the dialect register for ZUSAMMENKEHREN, taken from Pluschkovits 2025

As can be seen from this brief analysis, language variation in German in Austria is complex and partially dependent on the targeted register(s). We can observe that there is also heterogeneity in standard registers, and that the differentiation in nonstandard registers between colloquial and dialectal speakers can be relevant (see also Lenz et al. 2021 or Lenz et al. 2023). This, of course, has implications for how the data can be reused and shared with the CLARIN community. The final Section 4 is dedicated to circumnavigating this issue and detail how we integrate LexAT21 into the broader CLARIN context.

4. Integration into CLARIN

Both the map commentary and research data are outcomes we consider relevant for the CLARIN context. We will provide both to the community with a permissive license to the Virtual Language Observatory via the repository of the Austrian Centre for Digital Humanities, ARCHE (A Resource Centre for the HumanitiEs, see Żółtak et al. 2021). The data will be provided in two formats, as a csv dump as well as in TEI based XML.

We automated the creation of the TEI versions of both the map commentary as well as the entire corpus. Basis for this creation is a bash-script which pulls the data out of the relational database. We utilize the data of the responses, the corresponding annotations, informant data and metadata as well as the location data and information on the survey round for the construction of the corpus and map commentaries. Each map commentary is considered its own document. We construct the corpus and commentaries via different partials, in an effort to create a replicable export strategy. This is done for two reasons: first, because data is still being re-categorized by researchers by the time it is first imported into the database – the researchers utilize the mapping tool of LexAT21 to uncover possible areal patterns, which might lead to recategorization of variants or renaming of individual lemma (e.g., if a nonstandard lemma is mostly dominant in the Alemannic area, the spelling of the lemma might be changed to reflect this fact). Having an easily replicable workflow means that the production of different versions of the data is trivial. Secondly, as LexAT21 is an ongoing project with multiple survey rounds, the data will expand, which is another argument to make this workflow easily retriggerable. The import step of the data itself is deployed via GitHub Actions.

The first chunk of the export into TEI is the creation of the documents containing the map

comments. In addition to the text of the map commentary itself (i.e., the readable content) with basic markup, the TEI versions of the map commentary includes all necessary information on the provenance of the data – this includes not just the responses of the informants for each map commentary packaged in a `cit` element with an (among others) `ana` attribute but also exact formulation of stimulus as well as a link to the stimulus picture, ensuring that the individual analysis is reproducible and the data generation itself replicable. All possible stimuli (i.e., stimuli for all possible registers that informants could encounter with the exact verbal stimulus) are included as a feature structure for every map commentary.

The map commentaries are accompanied by a feature structure declaration (`fsDecl`), which includes the prose description of how the individual stimuli were constructed (thereby – much like a shorter version of the present paper – contextualizing the dataset).

The corpus document includes the metadata on all speakers as an initial `particDesc` element, including age and sex for every informant. A following `listPlace` details all the localities covered in the corpus. These, just as the informants, will necessarily differ between the individual survey rounds, which is why we will release the data of each survey round as its own (sub)corpus, albeit in the same manner. In an effort to reduce unnecessary redundancy, the linguistic data itself is not included in the corpus element. Instead, we reference the data from the individual map commentaries by utilizing `xi:include` elements with an Xpointer, pointing towards the data in the individual map commentaries.

The corpus object is also an interesting challenge in data modelling. Quite clearly, LexAT21 is sensitive to variation in language, especially across the areal dimension, but also in the social dimension. The project explicitly investigates the difference in registers, i.e., when informants are prompted to give their designations in standard or nonstandard registers. And out of this social variation, the question arises on how these different registers can effectively be identified in the corpus data. There has been a recent uptick in understanding language data from a variationist standpoint and trying to standardize language designations while remaining sensitive to variation and different varieties (see e.g., Romary 2025). However, how can the present possibilities be utilized to account for social variation?

As a first possible solution, the `<langUsage>` element in the header could be used to describe the "languages, sublanguages, registers, dialects" (TEI Consortium: 58). This, however, would only be used to identify the languages (or in this case, registers) of the corpus and has, for the present case, two considerable shortcomings: the first shortcoming is that the concrete tag for language identification should follow the BCP 47 (Philipps & Davis 2009), which in turn necessitates that the individual subtags are registered at IANA. There are currently no subtags that would cover the distinction between *Dialekt* und *Umgangssprache*, especially as those are considered from the perspective of them being glottonyms instead of concrete dialects (e.g., an informant might respond in Alemanic and another one in South-Bavarian, but for this analysis, both of their language use is simply considered *Dialekt*, see also e.g., Kim & Breuer 2017 on subtags for varieties of languages). This could be circumvented by utilizing the private use subtag `x` (i.e., `de-x-`) which would allow to e.g., define a `de-AT-x-dialekt` and `de-AT-x-umgangssprache`. This, however, is not beneficial to keeping the data as interoperable as possible.

The second shortcoming is that the application of such a label has a normative character – if e.g., a subtag for dialect was used, this would imply a certain level of curation on the data that simply is not present (e.g., excluding non-dialectal material). Essentially, we run into the issue of circularity: if we proclaim something to be dialect, we necessarily must have a notion of what constitutes such a

dialect, and by excluding material that does not fit this notion, we have constructed a concept which cannot be disproven. Alternatively, simply all responses elicited for the label *Dialekt* (or *Umgangssprache*) could be declared as such, which does beg the question on what to make of those responses were nonstandard and standard forms correspond perfectly. If nonstandard is defined as being (maximally) distant from standard, the fact that both forms coincide could imply that either the nonstandard or standard form is faulty, caused by either dialectal leveling or someone not being competent in standard language. Furthermore, exclusively labelling nonstandard varieties with a subtag implies that standard varieties of a language are the default, which is a questionable assertion for German, especially as spoken in Austria, at best.

With all this in mind, we consider the current version to represent our data (via a `usg type` attribute of the respective stimulus in the `cit` element) as a workable compromise, that hopefully does not do too much injustice to the data itself. There are, as mentioned above, exciting developments about more-dimensional representation of varieties of languages in TEI (as exemplified by e.g., Romary 2025). This could help make the dataset – which we consider of interest for everyone in the CLARIN context working on German, language variation generally or lexical variation specifically – more easily comprehensible. We hope that our use-case with regard to representing intended varieties based on their glottonyms can be a point of departure for further conversations, specific as this use-case might be.

5. Concluding remarks

The LexAT21 addresses a gap in research and aims to investigate lexical variation throughout Austria along the dialect-standard axis, from dialect through colloquial language to standard, using a consistent survey methodology. The LexAT21 data is collected in a controlled survey-process using multiple rounds of online questionnaires. Exemplary analyses of the linguistic variable "ZUSAMMENKEHREN" ('to sweep') do not only demonstrate a broad spectrum of lexical variants but also highlight significant spatial effects within Austria as well as the effect of the register (dialect, colloquial language, standard).

The technical pipeline begins with the data collection in LimeSurvey, proceeds through the linguistic annotation in Baserow and storage in PostgreSQL, and ultimately results in the presentation of the data in the web frontend. This replicable approach is essential within the project, since it is an ongoing project involving multiple rounds of data collection. Furthermore, integration into CLARIN and ARCHE, as the provision of the raw data in CSV and TEI formats, is intended not only to ensure the long-term preservation but also to enable cross-project reusability and interoperability. Finally, the LexAT21 dataset also shows that labelling language data is not necessarily a trivial act. While all data might be German, the differences in the data depending on the targeted registers demonstrates that no natural language is a monolith and that variation is inherent to natural language use. This, of course, has consequences on the reuse of the data – as the targeted register is essential context of the dataset, it needs to be taken into account for any given reuse of the dataset. This, however, is not necessarily a flaw of the data or limits the reusability – instead, it allows for more control over which part of the data (by e.g. register) is actually reused.

References

Ammon, U., Bickel, H., & Lenz, A. N. (Eds.). (2016). Variantenwörterbuch des Deutschen (2., völlig neu beab. u. erw. Aufl.). De Gruyter. https://doi.org/10.1515/9783110245448 (Original work published 2011)

Auer, P. (2021). Reflections on linguistic pluricentricity. Sociolinguistica, 35(1), 29–47. https://doi.org/10.1515/soci-2021-0003

Baserow. (2026). [Repository]. Github. https://github.com/baserow/baserow

Bülow, L., Wittibschlager, A., & Lenz, A. N. (2023). Variation and change of relativizers in Austria's German varieties. SPRACHWISSENSCHAFT, 48(3), 243–280.

Durkin, P. (2012). Variation in the lexicon: The 'Cinderella' of sociolinguistics?: Why does variation in word forms and word meanings present such challenges for empirical research? English Today, 28(4), 3–9. Cambridge Core. https://doi.org/10.1017/S0266078412000375

Dürscheid, C., Elspaß, S., & Ziegler, A. (2018). Variantengrammatik des Standarddeutschen. Leibniz-Institut für Deutsche Sprache. https://mediawiki.ids-mannheim.de/VarGra/index.php/Start

Eichhoff, J. (1978). Wortatlas der deutschen Umgangssprachen (Vol. 2). A. Francke AG Verlag.

Elspaß, S., & Möller, R. (n.d.). Atlas zur deutschen Alltagssprache. Retrieved https://www.atlas-alltagssprache.de/

Huesmann, A. (1998). Zwischen Dialekt und Standard: Empirische Untersuchung zur Soziolinguistik des Varietätenspektrums im Deutschen. Max Niemeyer Verlag.

Kim, A., & Breuer, L. M. (2017). On the Development of an Interdisciplinary Annotation and Classification System for Language Varieties – Challenges and Solutions. Jazykovedný Časopis, 68, 191–207. https://doi.org/10.1515/jazcas-2017-0029

Koppensteiner, W. (2023). Standard(s) in Österreich: Attitudinal-perzeptive Analysen [University of Vienna]. https://ubdata.univie.ac.at/AC16945911

Kretschmer, P. (1918). Wortgeographie der hochdeutschen Umgangssprache. Vandenhoeck & Ruprecht.

Lenz, A., Höll, J., & Ziegler, T. (2023). Lexikalische Variation in Österreich. Ausgewählte Austriazismen im Stadt-Land-Vergleich. Muttersprache: Vierteljahresschrift Für Deutsche Sprache, 133(1–2), 82–115.

Lenz, A. N. (2025). Atlas on Lexical Variation in Austria in the 21st Century. Austrian Centre for Digital Humanities. https://lexat21.lapis-online.at/en

Lenz, A. N. (2025ff.). LexAT21—Atlas zur lexikalischen Variation in Österreich im 21. Jahrhundert. Konzipiert und entwickelt von Jakob Bal, Kilian Kukelka, Markus Pluschkovits, Marie Rohler, Daniel Schopper und Anja Wittibschlager. Unter Mitarbeit von Amelie Dorn, Jan Höll, Katharina Korecky-Kröll, Wolfgang Koppensteiner, Claudia Mattes, Markus Pluschkovits, Rita Stiglbauer, Florian David Tavernier, Anja Wittibschlager, Theresa Ziegler, Kerstin Lorenz und Eric Schirl. https://lexat21.lapis-online.at/de

Lenz, A. N., Dorn, A., & Ziegler, T. (2021). Lexik aus areal-horizontaler und vertikal-sozialer Perspektive – Erhebungsmethoden zur inter- und intraindividuellen Variation. Sprachwissenschaft, 46(4), 387–431.

Lenz, A. N., Ziegler, T., Höll, J., Kunzmann, M., & Dorn, A. (in print). Lexikalische Dynamik in österreichischen Varietäten aus panchronischer Perspektive. 200 Jahre Dialektologie. Forschungsstand Und Auftrag. Akten Der 15. Bayerisch-Österreichischen Dialektologentagung 2023. Bayerisch-Österreichischen Dialektologentagung 2023.

LimeSurvey Project Team, & Schmitz, C. (2026). LimeSurvey [Repository]. Github. https://github.com/LimeSurvey

Mitzka, W., & Schmitt, L. E. (1951). Deutscher Wortatlas (Vol. 1). Wilhelm Schmitz Verlag.

Mitzka, W., & Schmitt, L. E. (1953). Deutscher Wortatlas (Vol. 2). Wilhelm Schmitz Verlag.

Muhr, R. (2001). Varietäten des Österreichischen Deutsch. Revue Belge de Philologie et d'histoire, 79(3), 779–803.

Phillips, A., & Davis, M. (Eds.). (2009). Tags for Identifying Languages (No. RFC5646; p. RFC5646). RFC Editor. https://doi.org/10.17487/rfc5646

Pluschkovits, M. (2025). ZUSAMMENKEHREN/KEHREN. Atlas Zur Lexikalischen Variation in

Österreich Im 21. Jahrhundert. URL: https://lexat21.lapis-online.at/de/articles/zusammenkehren-kehren
Romary, L. (2025). International standards for the identification and the description of languages and their varieties. In Harmonizing language data Standards for linguistic resources (pp. 35–60). de Gruyter. https://doi.org/10.1515/9783112208212-003
Sandow, R., Grieve, J., Stamp, R., & Schembri, A. (2025). Lexical variation and change: Integrating lexis into variationist sociolinguistics. Language Variation and Change, 37(3), 293–318. Cambridge Core. https://doi.org/10.1017/S0954394525100628
Schmidt, J. E., Dammel, A., Girnth, H., & Lenz, A. N. (2019). Sprache und Raum im Deutschen: Aktuelle Entwicklungen und Forschungsdesiderate. In J. Herrgen & J. E. Schmidt (Eds.), Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Band 4: Deutsch. Unter Mitarbeit von Fischer, Hanna und Ganswindt, Brigitte. (Vols. 1–4, pp. 28–60). https://doi.org/10.1515/9783110261295-002
Stöckle, P., & Pluschkovits, M. (forthcoming). The Structure of Lexical Variation in Austrian Varieties: A Quantitative Analysis. Languages, (Special Issue: Advances in Variationist Linguistics on German—Focus on Lexis and Pragmatics).
Tagliamonte, S. (2025). Analysing sociolinguistic variation (Second edition). Cambridge University Press. https://doi.org/10.1017/9781009403092
TEI Consortium (Ed.). (last updated on 4th september 2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange P5 Version 4.10.2. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html
Ziegler, T., Koppensteiner, W., Gellan, A., Ebner, J., & Lenz, A. N. (forthcoming). Standard Language in Austria: Lexical Variation. In A. N. Lenz, S. Elspaß, & S. M. Newerkla (Eds.), Handbook on German in Austria. Variation – Contact – Perception. Language Science Press.
Żółtak, M., Trognitz, M., & Ďurčo, M. (2022). ARCHE Suite: A Flexible Approach to Repository Metadata Management. 190–199. https://doi.org/10.3384/ecp18917

Towards FAIR Metadata for Specialised Corpora: A Community-Informed Empirical Study of Schema Development in Two Communities

Egon W. Stemle[†]
Institute for Applied Linguistics
Eurac Research
Bozen-Bolzano, Italy
egon.stemle@eurac.edu

Alexander König
CLARIN ERIC
The Netherlands
alex@clarin.eu

Nannan Liu
SUNY at Binghamton
United States
nliu9@binghamton.edu

Jennifer-Carmen Frey[†]
Institute for Applied Linguistics
Eurac Research
Bozen-Bolzano, Italy
jennifer.frey@eurac.edu

Hubert Naets[‡]
CKL2CORPORA
UCLouvain
Belgium
hubert.naets@uclouvain.be

Magali Paquot[‡]
CKL2CORPORA
F.R.S.-FNRS
UCLouvain
Belgium
magali.paquot@uclouvain.be

Mariachiara Russo
University of Bologna
Italy
mariachiara.russo@unibo.it

Abstract

This paper investigates the development of domain-specific metadata schemata to support FAIR data management within specialised research communities. We report on two case studies, the Core Metadata Schema for Learner Corpora (LC-meta) and the metadata schema developed for interpreting corpora within the Unified Interpreting Corpus (UNIC) project. We address how metadata are described and standardised in these projects, the principles that informed the schema design, and the infrastructural challenges that arise when semantically rich, community-specific metadata are to be represented within generic discovery frameworks. Our findings underline the importance of domain-oriented approaches to metadata standardisation, iterative design and stakeholder involvement, and suggest a more active role for CLARIN in supporting long-term sustainability, technical mediation, and shared stewardship.

1 Introduction

The *FAIR Guiding Principles for scientific data management and stewardship* (Wilkinson et al., 2016) have become a cornerstone of contemporary data management, promoting Open Science by ensuring that research outputs are Findable, Accessible, Interoperable, and Reusable. Existing infrastructures, such as CLARIN B-centres (Hinrichs & Krauwer, 2014) and data repositories like Zenodo[1] and Figshare[2], facilitate *F*indability and *A*ccessibility to a certain extent, but achieving true *I*nteroperability and *R*eusability remains a challenge (Bhardwaj et al., 2025; Frey et al., 2020). One important reason is that the generic metadata standards adopted by these repositories are designed for a broad range of content types

[†]Eurac Research CLARIN Centre (ERCC), K-centre for CMC and Social Media Corpora (CKCMC)

[‡]C-centre & K-centre for Learner Corpora (CKL2CORPORA)

[1]https://zenodo.org/

[2]https://figshare.com/

Egon W. Stemle, Alexander König, Nannan Liu, Jennifer-Carmen Frey, Hubert Naets, Magali Paquot, and Mariachiara Russo 2025. Towards FAIR Metadata for Specialised Corpora: A Community-Informed Empirical Study of Schema Development in Two Communities. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 193–206. DOI: https://doi.org/10.3384/ecp222

and therefore often lack the granularity required to capture the methodological, linguistic and contextual specificities of specialised corpora.

These limitations become particularly visible in federated metadata environments, whose primary objective is to aggregate heterogeneous resources from different repositories or infrastructures through a shared discovery layer, as, for example, in CLARIN's Virtual Language Observatory (VLO) (Van Uytvanck et al., 2012) and related cross-repository discovery services. Federation relies on a common descriptive baseline that enables cross-domain indexing and querying, which typically requires metadata to be normalised and reduced to generic elements. While effective for high-level discovery, this approach introduces a structural tension with the needs of specialised research communities, whose corpora depend on fine-grained, domain-specific metadata to support particular methodological paradigms and reuse scenarios. Such distinctions are often difficult to express within generic schemata and risk being flattened or rendered inaccessible in federated settings. In practice, this creates a recurring dilemma, as participation in infrastructures such as CLARIN is essential for visibility and sustainability. However, reliance on generic metadata representations can undermine Interoperability and Reusability by leaving crucial contextual information undocumented or inaccessible.

Domain-specific metadata standards have already been developed for some CLARIN communities, for example, for parliamentary corpora (Erjavec & Pančur, 2021) and cultural heritage data (Riley, 2024). However, many research communities still lack such standards, even though effective Interoperability and Reusability require the metadata schemata to be tailored to the specific characteristics of the data and their intended reuse contexts. Developing such schemata is itself a demanding and resource-intensive endeavour, requiring expertise that spans both the research domain and technical infrastructures, as well as careful negotiation between community-specific needs and more general, cross-domain approaches.

To address some of these challenges, illustrate possible approaches to solutions and highlight persistent problems, this study examines two community efforts to develop domain-specific metadata standards. The first is the Core Metadata Schema for Learner Corpora (LC-meta) (Paquot et al., 2024), developed collaboratively by researchers and infrastructure experts through a community-informed, iterative process. The second is the metadata schema for interpreting corpora (Liu & Russo, 2025) developed within the Unified Interpreting Corpus (UNIC) project and implemented on the UNIC platform.[3] Methodologically, UNIC builds on the path opened by LC-meta and extends it under different empirical, technical and ethical conditions.

By analysing these examples, the paper aims to highlight strategies, challenges and lessons learned in designing metadata standards that balance domain-specific needs with broader Interoperability goals. In addition, it aims to open a broader discussion on the role of CLARIN in supporting research communities, in particular by facilitating the sharing of expertise, promoting methodological convergence, and providing infrastructure and guidance for the development and maintenance of domain-specific metadata standards.

2 Metadata Standards and the Case for Domain-Specific Adaptation

Metadata, often described as 'data about data' (Greenberg, 2005), initially referred to simple descriptive information, such as titles, authors, or dates, that helped identify and organise datasets. Over time, the concept has expanded to encompass a rich system of structured information that documents not only the content of a dataset, but also its provenance, internal structure, semantics, and conditions of creation and use (Gilliland, 2008). Modern metadata captures definitions, rules, contextual information, and the methods used to generate derived data, enabling users to understand, interpret, and evaluate the significance and reliability of digital resources. In this sense, metadata is no longer a peripheral administrative tool but a critical component of research infrastructure, underpinning data transparency, Interoperability, and meaningful reuse across disciplines.

The development of metadata schemata has been central to data stewardship initiatives in digital hu-

[3] https://unic.dipintra.it/ – Code available under the Apache 2.0 License at https://github.com/F-technology-srl/unic

manities and linguistics. Frameworks such as the Text Encoding Initiative (TEI),[4] Dublin Core[5], and the Component MetaData Infrastructure (CMDI)[6] provide formalised structures for describing digital resources, offering Interoperability across disciplines and data types. These general frameworks establish a common foundation for resource description, enabling sharing and integration beyond the original context of creation. At the same time, they often allow some level of domain-specific adaptation, so that communities can extend schemata to capture features relevant to their own research.

However, in practice, these adaptations are often limited, and the general nature of these frameworks can make it difficult to fully represent the specific characteristics and needs of particular research domains. In the CLARIN ecosystem, metadata quality is often uneven across corpora and resource types (Fišer et al., 2018), and some user groups report dissatisfaction with the searchability and transparency of metadata in CLARIN's VLO (Lušicky & Wissik, 2017). These findings highlight the persistent gap between the availability of metadata schemata and their actual usability in supporting resource discovery and reuse.

The effectiveness of metadata schemata depends not only on adherence to formal standards but also on the degree to which they enable accurate, context-sensitive interpretation and reuse of resources (Greenberg, 2005). In other words, metadata must be both technically correct and conceptually aligned with the practices, expectations, and values of the relevant research community. As Velterop and Schultes (2021) argue, 'metadata need to comply with standards that are relevant for the domains and disciplines involved …[and] they need to be properly applied drawing on the controlled vocabularies created by or in close consultation with the practising domain experts'.

Domain-specific metadata schemata are therefore essential. By capturing the unique characteristics, requirements, and interpretive frameworks of resources within a particular field, they help ensure that data can be discovered, understood, and reused accurately. Beyond supporting Interoperability within a research community, such schemata contribute to broader Open Science objectives, for example, by enhancing transparency, enabling replication, and facilitating comparative or meta-analytic studies, thereby improving overall study quality (e.g., Paquot (2024)).

3 Case Study I - Core Metadata Schema for Learner Corpora (LC-meta)

In the following, we present our first case study (see Section 4 for the second) that has developed a domain-specific metadata schema and show how metadata is described and standardised. We discuss design principles and underlying values, as well as the technical implementation of the metadata framework. This will provide valuable insights and feed into the development of a blueprint in Section 5 for community-informed, domain-specific metadata.

LC-meta (Paquot et al., 2024) is a domain-specific metadata schema designed for the Learner Corpus Research community. Its primary goals are to support FAIR resource creation, ensure comparability across resources, facilitate cumulative research, and promote best practices in the field.

The schema includes 168 elements that were identified as broadly relevant across learner corpus research. These elements are either obligatory, to ensure FAIR compliance and comparability, or optional, to support best practices and facilitate future research. The schema was developed through a collaborative, community-informed, iterative process, consisting of three phases: a draft version based on reviewing existing resources to prompt community feedback, iterative refinements based on testing and community feedback, all leading to the publication of a FAIR resource and dissemination and community entrustment via the creation of a dedicated working group with experts from various institutions.

3.1 General Structure of the Schema: Eight Interrelated Components

LC-meta includes 168 partly obligatory, partly optional elements, organised into eight interrelated main components:

(1) Administrative metadata,
(2) Corpus design metadata,

[4]https://tei-c.org/guidelines/
[5]https://www.dublincore.org/specifications/dublin-core/dces/
[6]https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

(3) Learner metadata,
(4) Text metadata,
(5) Situational and task-related metadata,
(6) Annotation metadata,
(7) Annotator metadata,
(8) Transcriber metadata.

The component-based structure is intended to support detailed and structured metadata descriptions while remaining compatible with existing language resource infrastructures, in particular component-based approaches such as CMDI.

Rather than describing all information in a single, flat record, LC-meta separates metadata components by function and scope. This enables clearer documentation, avoids unnecessary duplication, and facilitates linking between different aspects of a corpus, such as participants, texts, annotations, and contributors to transcription and annotation. The components are designed to be combined flexibly, allowing the schema to accommodate a wide range of learner corpus designs.

The eight components are explicitly interrelated through references and persistent identifiers. Except for administrative and corpus design metadata, all components are repeatable. This makes it possible to represent multiple learners, texts, situational contexts or annotation layers within a single corpus. Administrative, corpus design, learner, text, situational and task-related metadata form the core descriptive backbone of learner corpora and are therefore obligatory. By contrast, the annotation, annotator and transcriber components are optional. These latter components focus on data provenance and promote best practices in well-documented data reporting.

As the aim of LC-meta was to offer a comprehensive set of metadata fields that apply to all types of corpora in the field, elements specific to certain research contexts (e.g. corpora for sign languages or corpora collected in educational settings) are not included in the core metadata schema but should be described within separate components yet to be developed.

Below, we briefly describe the eight components of the core metadata schema, provide examples of the elements contained in each and explain the rationale for the proposed components.

(1) Administrative metadata provides high-level information about the learner corpus as a resource. It includes elements required for identification (corpus name, acronym and persistent identifier), citation, versioning, documentation, and access conditions. These metadata support core infrastructure functions such as persistent identification, discovery in repositories, and long-term maintenance. As such, this component, which consists mainly of obligatory elements, aligns closely with metadata typically required for depositing resources in repositories and is crucial for ensuring compliance with FAIR principles (cf.König et al., 2021; Lindström Tiedemann et al., 2018). By explicitly modelling administrative information as a separate component, LC-meta supports consistent data management practices and facilitates integration with existing repository and catalogue systems.

(2) Corpus design metadata documents the conceptual and methodological framework underlying a learner corpus. Elements in this category concern, for example, the language(s) represented in the corpus, whether texts are single- or multi-authored, the mode(s) of communication and other characteristics such as longitudinality of the data as a whole. While more specific metadata might be encoded document-by-document at the text level, this component describes the overall scope of the resource, including the types of data collected, the linguistic focus, and key design decisions made during corpus compilation. This component plays a crucial role in interpretability and comparability across learner corpora. Making design assumptions explicit helps users find relevant resources in a register, understand the structure of a resource as a whole and assess whether a given corpus is comparable to other resources and/or (re-)usable for other research questions.

(3) Learner metadata describes the individuals who contributed language data to the corpus. The component contains basic socio-demographic variables such as the participant's age, gender, and first language, as well as more domain-specific items, including proficiency levels and more detailed information on a learner's language background, multilingual profiles, motivation, or the presence of learning

difficulties. Instead of treating participant information as attributes of texts, LC-meta models learners as independent entities that can be linked to one or more texts. This approach supports common use cases in learner corpus research, such as longitudinal designs or corpora in which multiple texts are associated with the same learner. It also avoids redundancy by allowing learner information to be recorded once and reused across multiple data objects. At the same time, this component also provides the possibility to include information on a participant's experience with and exposure to one or various other languages by filling out sub-components that are themselves repeatable, enabling the modelling of multilingual profiles with exposure to various languages and by leveraging the structural ideas behind CMDI (Broeder et al., 2012).

(4) Text metadata describes individual units of learner language production. As such, the term text is broadly defined to include written, spoken, and multimodal data. This repeatable component contains, for each text, a unique and anonymous identifier for the text itself, as well as descriptive information about the Administrative metadata: it provides high-level information about the learner corpus as a resource. text. It also contains references to associated files (e.g., handwritten text in PDF format, sound files, and videos), as well as links to other relevant components, that are linked through their identifiers. Within the schema, text metadata thus functions as a central linking point: Each text is associated with learner metadata, situational and task-related metadata, and — where applicable — annotation and processing metadata. This design supports fine-grained descriptions at the text level while maintaining coherence across the corpus and enabling selective reuse of metadata components.

(5) Situational and task-related variables describe the communicative context in which learner language data were produced. This includes both general situational characteristics and, in instructed learning settings, details about the specific tasks used to elicit language samples. The variables proposed in this section are based on Biber and Conrad (2019, p. 40) and include, for example, the number of and relationship to addressees, the mode and medium of communication, the communicative purpose and the broad topic domain and register. By modelling situational and task-related information as a separate component, LC-meta allows multiple texts to be linked to the same contextual description where appropriate. This is particularly useful for avoiding the workload of entering metadata for corpora that include repeated tasks or standardised elicitation settings. The component thus supports both detailed contextualisation and efficient metadata management.

(6) Annotation metadata describes the linguistic or other annotations applied to the corpus data. Each annotation type is treated as a separate entity, allowing different annotation layers to be documented independently. This component records information such as the nature of the annotation (manual or automatic), the tools or methods used (including distinct version numbers), and the extent of quality control or evaluation. While it is not yet common practice in LCR to provide this information as machine-actionable metadata, the explicit modelling of annotation layers supports transparency, reproducibility, and reuse, and aligns with the increasing recognition of annotations as valuable research outputs in their own right.

(7, 8) Annotator and transcriber metadata describe the human agents involved in the transcription and annotation of learner language data. Although optional, these components enable provenance tracking and support documentation of annotation practices, which can be relevant for assessing data quality and reliability. Making annotators and transcribers explicit also supports attribution and acknowledgement of the work involved in preparing language resources for reuse.

3.2 Metadata Elements: Design Principles and Key Characteristics

While LC-meta is organised into components at a higher level, all individual metadata elements follow a common internal structure and set of organising principles. These principles were chosen to balance standardisation, usability, and flexibility, while remaining compatible with FAIR requirements and existing practices in corpus and language resource infrastructures.

Main element-structure and documentation: each metadata element is represented by a clearly defined field name, accompanied by a short, human-readable description. In addition, each element is documented by additional columns specifying conditionality, recommended fixed values, examples and further guidance, obligatoriness and repeatability for the element. Together, this makes explicit the information that can be recorded, and how and under which circumstances it should be provided.

Conditional metadata and guided data collection: some metadata elements are only relevant if another element has been answered in a particular way. Beyond its descriptive function, this design choice anticipates future technical implementations. In particular, conditionality could be used to configure adaptive metadata collection interfaces, such as web-based questionnaires, that dynamically guide respondents based on their previous answers. This reduces users' cognitive load and helps prevent inconsistent or irrelevant metadata entries.

Fixed attributes, controlled vocabularies and openness: wherever feasible, LC-meta recommends the use of fixed attributes, that is, closed lists of predefined values. These are intended to support Interoperability, comparability and machine-actionability. At the same time, the schema deliberately avoids enforcing controlled vocabularies when no widely accepted standard exists. For many community-relevant variables, consensus is still emerging. In such cases, metadata fields are left open-ended, while the schema provides recommended attributes and illustrative examples in a separate *further details and examples* column.

Use of examples and dual representations: the *further details and examples* column provides clarifications, usage notes, and concrete examples without making them normative. In some cases, LC-meta adopts a dual-representation strategy: a required field with a controlled vocabulary captures broad, comparable categories, while an optional open field allows users to provide more fine-grained or locally meaningful distinctions. For example, register is represented both by an obligatory field with a small set of broad categories and by an optional free-text field in which corpus compilers can specify more precise register types. This approach allows the schema to support both cross-corpus comparison and detailed documentation of individual resources.

Obligatory and optional metadata elements: similar to the main components, not all metadata elements are mandatory. Obligatory elements are intentionally limited in number and focus on information that is essential for basic interpretability, comparability, and reuse. Many other elements are optional and need only be recorded once at the corpus or design level, reducing the overall documentation burden. Inevitably, some (already existing) learner corpora might not be able to provide all obligatory elements and will have to fill in missing values. Nevertheless, we believe that establishing such a shared minimal core is particularly important for cumulative research and for integrating infrastructure. It provides guidance to newcomers, supports harmonisation across projects and increases the long-term value of corpora; even when primary data cannot be shared and only metadata can be made available. This combination of obligatory and optional fields also responds to desiderata identified by the community and the research literature (e.g. Tracy-Ventura et al. (2020); Wulff (2023)), thereby providing guidance on best practices for corpus compilation.

Repeatability and representation of complex realities: finally, metadata elements are characterised as non-repeatable or repeatable, allowing multiple values to be recorded where appropriate. This allows for capturing complex research designs while avoiding formal restrictions, yielding both flexibility and extensibility of the schema.

3.3 Technical Implementation

LC-meta was initially released as a structured spreadsheet with the explicit goal of later conversion into machine-actionable formats. For successful community uptake of the schema, the working group has identified the need for additional technical tools that enable user-friendly metadata administration, as well as tools for searching corpora based on metadata and survey-templates to collect metadata from

study participants. However, due to the workload and the financial limitations for outsourcing such a task, their technical implementation remains future work.

3.4 Community-Informed Development Process

In order to assure domain relevance and community uptake, the development of the LC-meta went through three main phases: In 2014, a comprehensive review of the metadata used in existing learner corpora[7] and other resources, such as the Talkbank website[8], the TEI and the Corpus Encoding Standard[9] resulted in a first draft proposal for core metadata for learner corpus research by Granger and Paquot[10]. The draft was then presented at the CLARIN *Workshop on Interoperability of Second Language Resources and Tools*[11] to collect community feedback and direct input. The draft proposal was then tested by a group of interested domain experts, including one of the initial authors and other researchers with experience in language research infrastructures and corpus creation from different institutions. The group evaluated the applicability of the initial draft to learner corpora with different corpus designs, developing a revised version which was again presented at two domain-specific conferences (Frey et al., 2023; König et al., 2022) for further community feedback. Additionally, feedback was gathered via mailing lists, online platforms and online questionnaires. After clarifying, evaluating and incorporating – where appropriate – all gathered feedback, LC-meta was published as a versioned, findable and openly accessible resource[12] with a detailed description in Paquot et al., 2024. After this, the schema was entrusted to the community by establishing a working group within the domain-specific Learner Corpus Association (LCA)[13], which collaborates closely with the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA)[14], and took over responsibility for further developing, maintaining and disseminating the metadata schema.

4 Case Study II - Metadata Schema for Interpreting Corpora within the Unified Interpreting Corpus (UNIC) Project

In the following, we present our second case study of a domain-specific metadata schema and show how metadata is described and standardised. We discuss design principles and underlying values, as well as the technical implementation of the metadata framework. This will provide valuable insights and feed into the development of a blueprint in Section 5 for community-informed, domain-specific metadata.

The metadata schema for interpreting corpora within the Unified Interpreting Corpus (UNIC) project (Liu & Russo, 2025) is based on the value-sensitive design paradigm (Friedman & Hendry, 2019), an approach in which conceptual design features and technical implementations embed social and moral values from the start, alongside usability and performance. While inspired by the component structure of LC-meta, the UNIC metadata schema consists of general and domain-specific elements within ten components to meet the needs of the translation and interpreting research community:

(1) Corpus metadata,
(2) Event metadata,
(3) Actor (as an umbrella term for speaker and signer) metadata,
(4) Source text metadata,
(5) Interpreter metadata,
(6) Target text metadata,
(7) Transcriber metadata,
(8) Annotator metadata,
(9) Annotation metadata,
(10) Alignment metadata.

[7] See the Learner Corpora around the World webpage: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[8] https://talkbank.org/
[9] https://www.cs.vassar.edu/CES/
[10] http://hdl.handle.net/20.500.12124/61
[11] https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools
[12] https://doi.org/10.14428/DVN/AAUEM2
[13] https://www.learnercorpusassociation.org/working-group-on-metadata-in-learner-corpus-research/
[14] https://uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora.html

4.1 Design Principles

Aside from FAIR principles, the UNIC schema is also explicitly aligned with CARE principles (collective benefit, authority to control, responsibility and ethics; Carroll et al., 2020) because interpreting corpora often involve sensitive communicative settings, such as healthcare, asylum procedures, or institutional decision-making and may include audiovisual recordings and personal data. The CARE Principles were originally a framework for Indigenous Data Governance, ensuring that data practices respect the rights, interests and well-being of Indigenous peoples, not just technical openness or efficiency. The Principles are often extended to apply more broadly to vulnerable groups. The ethical considerations led to design decisions that affect both metadata content and exposure, including the introduction of opt-out mechanisms, fine-grained access levels and the separation of publicly visible descriptive metadata from access-controlled data download. The UNIC schema remains closely aligned with CLARIN infrastructures, as the majority of metadata elements reuse concepts already defined within CLARIN registries, and the CMDI profile is available in the Component Registry.[15]

4.2 Technical Implementation

A further methodological distinction between LC-meta and UNIC is the technical implementation of the schema. UNIC implemented its schema directly within an open-source platform (see footnote 3), which supports metadata and data registration, validation, filtering and search. The platform also acts as a communication channel between corpus creators and (re)users. The authors conducted surveys with computational and corpus linguists as well as interpreting students, and user feedback indicates greater perceived usefulness of the UNIC filters than those on the VLO. The authors are also collecting information on platform registrations and data use to further refine UI features and schema usability. In this respect, UNIC operationalises the value-sensitive design principle, in which technical quality and usability are closely linked, and schema development should be informed not only by conceptual and empirical considerations but also by interaction with real users.

4.3 Schema Development Process

While addressing a different research domain, the UNIC schema adopts a similar community-informed, empirically grounded, and iterative approach as LC-meta and extends it to accommodate the specific characteristics of interpreting studies. Similar to the LC-meta project, the starting point for schema development was a systematic review of existing practice. In the case of UNIC, this review was conducted at a larger empirical scale and across a wider range of modalities. A total of 1,898 publications were screened, resulting in the identification of 125 interpreting corpora described in 382 publications, covering 38 spoken and eight signed languages.

This corpus survey in the form of an ‘overlap and gap’ analysis served two purposes. First, it allowed the identification of metadata elements that recur across interpreting corpora (‘overlaps’) and therefore constitute candidates for core metadata. Second, it made gaps visible between the information in the data and that captured in existing metadata records. Elements that occurred in more than half of the corpora were prioritised as mandatory, while less frequent elements were retained as optional. This distinction mirrors the LC-meta principle of defining a minimal obligatory core that ensures comparability, while preserving flexibility for domain-specific variation.

The UNIC metadata schema demonstrates that conceptual, technical and empirical investigations can be applied to a specific domain while remaining compatible with CLARIN’s overarching metadata ecosystem. As with LC-meta, the UNIC metadata also highlights several common challenges in developing and maintaining such infrastructure, challenges that are likely to recur in many similar community-driven initiatives, which we discuss in section 6.

[15] https://catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu%3Acr1%3Ap_1747312582431&registrySpace=public

5 A Blueprint for Community-Informed Domain-Specific Metadata Development

The development of the Core Metadata Schema for Learner Corpora (LC-meta) and the Unified Interpreting Corpus (UNIC) metadata schema represents systematic efforts within CLARIN-related communities to operationalise FAIR, and, in the case of UNIC, also CARE principles through a community-informed, empirically grounded methodology rather than via top-down standardisation. Taken together, both initiatives suggest a reusable four-step blueprint for other research communities.

First, schema development should begin with a systematic review of existing practice. Metadata variables should be extracted from a wide range of relevant corpora and related publications within the target domain, including corpus documentation, corpus reports and state-of-the-art syntheses of the field (e.g., review articles, methodological overviews or handbook chapters). Based on this review, an initial draft of the metadata schema can be developed.

Second, an iterative and participatory design process should be initiated. Draft versions of the schema should be shared with the research community through workshops, conference presentations, mailing lists, and publicly accessible online documents. Feedback should be systematically collected and regularly integrated into the schema development process, ensuring that concrete community needs drive revisions. This approach will help ensure that successive schema versions reflect actual usage rather than assumptions not grounded in empirical practice.

Third, the schema should explicitly distinguish between obligatory and optional metadata elements, as well as between closed and open vocabularies. These design principles reflect the view that standardisation should establish a minimal common ground necessary for comparability and reuse, while still leaving room for methodological diversity and innovation. Rather than enforcing fixed definitions for potentially contested concepts, the schema should prioritise flexibility and recommend the use of accompanying documentation to clarify and contextualise specific metadata values.

Fourth, the schema should be conceived as a modular and extensible framework. No single schema can capture the full diversity of a domain from the outset. A shared core should therefore be complemented by additional modules developed by sub-communities with specialised expertise. Crucially, the long-term maintenance and evolution of a domain-specific schema should be embedded in appropriate governance structures within the domain's research community and connected to relevant scholarly associations, research centres, and research infrastructures.

Provided that some structural challenges are addressed with CLARIN's support (see Section 6), this combination of empirical grounding, iterative community engagement, and modular extensibility provides a viable blueprint for developing additional domain-specific metadata schemata that are both widely adoptable and responsive to community needs. Together, LC-meta and UNIC illustrate a methodological path for specialised metadata development within CLARIN. LC-meta provides the conceptual and procedural foundation; UNIC builds on this foundation under different empirical, technical and ethical conditions. For CLARIN, this highlights a key role: not to impose uniform metadata semantics, but to support communities in applying, adapting and extending a shared methodological framework. At the same time, developing domain-specific metadata schemata is highly time-consuming and resource-intensive, with limited prospects for external funding, underscoring the critical importance of CLARIN's support in enabling and sustaining these efforts.

In the next section, we propose concrete roles for CLARIN to support communities in developing domain-specific metadata schemata through technical assistance, infrastructure, tooling, governance frameworks that promote long-term sustainability and, last but not least, financial support.

6 How to Support Community-Specific Metadata within a FAIR CLARIN Ecosystem

As mentioned throughout the previous sections, the projects discussed above provide strong examples of obstacles and challenges that are likely to recur in similar community-driven initiatives.

6.1 Fostering Awareness for and Supporting the Creation of Domain-Specific Metadata Schemata

Despite the existence of general metadata schemata such as TEI, Dublin Core, and CMDI, the granular needs of subfields like interpreting studies or learner corpus research require more fine-grained, domain-specific schemata. LC-meta and UNIC underscore this gap. TEI, Dublin Core and CMDI offer broad, flexible structures, but they lack dedicated components for interpreting-specific variables (e.g., transcript-recording alignment, interpreter professional status) or learner-specific information (e.g., age of onset, multilingual background, learning context). Domain-specific schemata like LC-meta and UNIC enable richer, more nuanced metadata tailored to the theoretical and empirical needs of their fields. As such, the schemata complement broader frameworks. They provide the precision and context required to move from general Interoperability to discipline-optimised Reusability and research impact.

At the same time, the two projects also show that ideas and techniques can be reused and adapted across communities, even when their individual requirements differ. This paper examined how FAIR metadata can be realised for specialised corpora whose requirements go beyond generic metadata standards alone. Drawing on the development of LC-meta and the UNIC metadata project, we have argued that domain-specific metadata schemata are not peripheral to FAIR data management, but constitute a necessary condition for achieving Reusability in practice. The central question, therefore, is not whether such schemata should be developed, but how their development could be embedded within a shared research infrastructure such as CLARIN.

Our analysis reinforces a key insight that emerges from both LC-meta and UNIC. Metadata standards are not imposed on a community, but emerge through iterative, collaborative practice. FAIRness, in this sense, is not achieved by enforcing uniformity but by fostering coordination between communities, infrastructures, and services. One of CLARIN's core strengths lies precisely in its capacity to raise awareness and support such coordination by providing shared technical foundations while respecting the epistemic and methodological diversity of its user communities.

6.2 Building and Maintaining Infrastructure

In the case of UNIC, significant funds could be invested in outsourcing the design and implementation of an open-source platform to an external contractor, as funding for the project was provided under the HORIZON-MSCA Action 2022. This approach enabled the fast development of a feature-rich system tailored to the needs of this specific community. At the same time, this has led to a dependence on a codebase whose complexity exceeds what can generally be maintained within an academic research environment. As initial funding dries up, the long-term maintenance of such platforms becomes increasingly challenging. Even when the software is released under an open-source licence, continuity cannot be taken for granted. Bringing new developers into the project requires significant onboarding effort, including familiarisation with the architectural decisions, technology stack, and the domain-specific assumptions embedded in the code. Within a university context, where technical staff turnover is common and long-term software engineering positions are rare, this effort competes with core research and teaching responsibilities. Moreover, the incentive structures within academia are not well-suited for sustained infrastructure maintenance. Contributions to codebases, refactoring, documentation, or technical debt reduction are rarely recognised as scholarly output, despite being essential for the viability of research infrastructures. Consequently, even motivated community members may find it difficult to justify the time investment required to assume responsibility for such a platform. These challenges are not the exception. They are symptomatic of a structural gap between short-term project funding and long-term infrastructural needs.

CLARIN's added value lies in its position as a coordinating infrastructure with a long-term perspective, cross-community reach, and experience in sustaining shared services beyond individual project lifecycles. In this respect, CLARIN could act as an infrastructural stabiliser by providing a minimal, yet reliable framework within which community-specific platform development can be anchored and sustained after the end of project funding. Further support could include long-term hosting arrangements, integration and alignment with persistent identifier mechanisms and support with CLARIN authentication

and authorisation mechanisms.

6.3 Standardisation, Exchange and Usability

Discussions within the CLARIN community make clear that CMDI currently serves as the only commonly agreed meta-standard. At the same time, CMDI was never intended to serve as a single, semantically exhaustive metadata schema. Instead, it provides a flexible framework for defining profiles, understood as community- or domain-specific schema instantiations. This design reflects a structural tension that is intrinsic to CLARIN's mission, namely the need to ensure Interoperability across heterogeneous resources while enabling communities to describe their data with sufficient semantic precision.

A recurring concern among community members is the usability of richly annotated data. Community-specific metadata, because of its complexity and domain orientation, does not easily integrate into generic discovery services such as the VLO without significant information loss. At the same time, reducing such metadata to a minimal common denominator undermines precisely those aspects that make specialised corpora interpretable and reusable within their research contexts. The challenge is therefore not only technical but conceptual: how can metadata representations of different granularities coexist without losing functionality or visibility?

Our metadata model offers one possible response to this challenge. By treating CMDI-based descriptive metadata as the authoritative, infrastructure-facing layer, identified through persistent identifiers and exposed via repositories, CLARIN can provide a stable reference point for all other metadata representations. Richer, community-maintained metadata can then be understood as supersets that extend this authoritative layer, while simplified representations, such as citation metadata, EOSC-compatible HTML summaries, or exports to national discovery services, function as subsets. The persistent identifier serves as the common anchor linking all representations and ensuring their coherence.

This means that while community platforms like UNIC may remain internally technologically diverse, CLARIN could support the definition of stable interfaces for metadata exchange, harvesting, and discovery. By ensuring that key outputs, such as CMDI-compatible metadata, search endpoints, or landing pages, remain Interoperable with CLARIN services, CLARIN can decouple long-term visibility and Reusability from the internal evolution of a platform's codebase. CLARIN B- and C-centres can act as custodians of authoritative CMDI records, ensuring stability and persistence while allowing semantic innovation at higher descriptive layers.

6.4 Financial Support

Developing and maintaining domain-specific metadata standards is a demanding and resource-intensive endeavour that requires sustained expertise at the intersection of research practice and technical infrastructures, as well as careful negotiation between community-specific needs and more general, cross-domain approaches. Complexity is further exacerbated by the fact that many linguistic research communities are inherently multilingual. Their data, research questions and user groups typically span multiple languages, institutions and countries. As a consequence, no single institution, national community, or language group can reasonably be expected to develop and maintain a domain-specific metadata standard on its own. At the same time, no single language or national community is the exclusive beneficiary of such a standard. The benefits are distributed across institutions and countries, while the costs of development, coordination and long-term maintenance are concentrated. Project-based funding schemes tend to support the development of new software or standards, but rarely provide adequate mechanisms for their long-term maintenance and coordination, as illustrated by the experience of projects such as UNIC.

The main tension arises from the structure of existing funding schemes. Local or national funding programmes, for which individual institutions could realistically apply, are typically ill-suited to support the development of domain-specific metadata schemata, and even less so trans-national coordination efforts whose benefits extend beyond a single country. Conversely, European-level funding instruments, while more appropriate in scope, are usually targeted at substantially larger initiatives and infrastructural developments. In this context, CLARIN ERIC could assume an intermediary role by aggregating funding at the infrastructure level and redistributing it in a targeted manner to support coordinated, community-driven metadata work across Knowledge Centres.

Many challenges do not stem from a lack of expertise, but from its fragmentation across countries, institutions, and projects. CLARIN is therefore well-positioned to step in as a broker of shared stewardship by providing targeted, sustained support to Knowledge Centres that coordinate domain-specific, cross-institutional metadata work. By financially enabling these Centres to facilitate community exchange, documentation, and alignment with infrastructural components, CLARIN can help address structural incentive gaps and ensure that domain-specific metadata standards remain both viable and Interoperable over time. In conclusion, supporting FAIR metadata for specialised corpora requires CLARIN to embrace its role as an infrastructure mediator rather than a semantic authority. By enabling layered metadata representations, supporting community-driven schema development, and providing stable points of reference through CMDI and persistent identifiers, CLARIN can help bridge the gap between general Interoperability and domain-optimised Interoperability and Reusability. This approach allows specialised corpora to become truly FAIR, now and in the future.

7 Conclusion

This paper has examined how FAIR metadata can be realised for specialised corpora whose descriptive requirements exceed what can reasonably be expressed through generic metadata standards alone. Drawing on the development of LC-meta and the subsequent extension of its methodology in the UNIC project, we have argued that domain-specific metadata schemata are not peripheral to FAIR data management, but constitute a necessary condition.

The two case studies show that such schemata are most effective when they are developed through community-informed, empirically grounded, and iterative processes, and when their long-term maintenance is embedded in governance structures that connect research communities with infrastructures. At the same time, they also make visible a number of recurring structural challenges, most notably the tension between generic federation and semantic precision, the gap between short-term project funding and long-term infrastructural needs, and the fragmentation of expertise across institutions and countries.

Against this background, CLARIN's role should not be understood as that of an authority imposing uniform metadata standards across all domains. Rather, CLARIN is best positioned to act as an infrastructure mediator that enables the articulation, coordination, and long-term sustainability of community-specific metadata work. By supporting layered metadata representations, stable interfaces, community-driven governance, and targeted coordination through Knowledge Centres, CLARIN can help bridge the gap between general Interoperability and domain-optimised Reusability.

In this sense, the contributions of LC-meta and UNIC reach beyond the two communities examined here. Together, they point towards a viable methodological and infrastructural model for future domain-specific metadata development within CLARIN and illustrate why such work deserves more systematic institutional, technical, and financial support.

References

Bhardwaj, R. K., Nazim, M., & Verma, M. K. (2025). Research data management and FAIR compliance through popular research data repositories: An exploratory study. *Data Technologies and Applications*, *59*(3), 472–492. https://doi.org/10.1108/DTA-12-2022-0477

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Broeder, D., Windhouwer, M., Van Uytvanck, D., Trippel, T., & Goosen, T. (2012). CMDI: A Component Metadata Infrastructure. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini & T. Trippel (Eds.), *Proceedings of the Describing Language Resources with Metadata Workshop at LREC 2012*. European Language Resources Association (ELRA). Retrieved February 1, 2026, from http://lrec.elra.info/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, *19*(1), 1–12. https://doi.org/10.5334/dsj-2020-043

Erjavec, T., & Pančur, A. (2021). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.4133

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Retrieved February 1, 2026, from https://aclanthology.org/L18-1210/

Frey, J.-C., König, A., Stemle, E., Falaise, A., Fišer, D., & Lüngen, H. (2020). The FAIR Index of CMC Corpora. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the prism of digital humanities*. L'Harmattan. https://hdl.handle.net/10863/15984

Frey, J.-C., König, A., Stemle, E. W., & Paquot, M. (2023). A core metadata schema for L2 data. *Book of Abstracts of EUROSLA32*. https://hdl.handle.net/10863/42215

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Gilliland, A. J. (2008). Setting the Stage. In *Introduction to Metadata* (Second Edition, pp. 1–19). Getty Research Institute. Retrieved February 1, 2026, from https://www.getty.edu/publications/resources/virtuallibrary/0892368969.pdf

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. *Cataloging & Classification Quarterly*, *40*(3–4), 17–36. https://doi.org/10.1300/J104v40n03_02

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 1525–1531. Retrieved February 1, 2026, from http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf

König, A., Frey, J.-C., & Stemle, E. W. (2021). Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. *Information*, *12*(5), 199. https://doi.org/10.3390/info12050199

König, A., Frey, J.-C., Stemle, E. W., Glaznieks, A., & Paquot, M. (2022). Towards standardizing LCR metadata. *Book of Abstracts of the 6th Lerner Corpus Research Conference*. https://hdl.handle.net/10863/42994

Lindström Tiedemann, T., Lenardič, J., & Fišer, D. (2018). L2 learner corpus survey: Towards improved verifiability, reproducibility and inspiration in learner corpus research. In I. Skadin & M. Eskevich (Eds.), *CLARIN Annual Conference Proceedings, 2018* (pp. 146–150). CLARIN. Retrieved February 1, 2026, from https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy

Liu, N., & Russo, M. (2025). A value-sensitive metadata schema for interpreting corpora: Implementation on the Unified Interpreting Corpus (UNIC) platform. *Interpreting. International Journal of Research and Practice in Interpreting*, *27*(2), 157–196. https://doi.org/10.1075/intp.00123.liu
OA: https://hdl.handle.net/11585/1013497.

Lušicky, V., & Wissik, T. (2017). Discovering Resources in the VLO: A Pilot Study with Students of Translation Studies: CLARIN Annual Conference 2016 (L. Borin, Ed.). *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, 63–75. Retrieved February 1, 2026, from http://www.ep.liu.se/ecp/136/005/ecp17136005.pdf

Paquot, M. (2024). Learner corpus research: A critical appraisal and roadmap for contributing (more) to SLA research agendas. *Corpus Linguistics and Linguistic Theory*, *20*(3), 567–590. https://doi.org/10.1515/cllt-2024-0014

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, *10*(2), 280–300. https://doi.org/10.1075/ijlcr.24010.paq

Riley, J. (2024). *Seeing Standards: A Visualization of the Metadata Universe*. Platform for Experimental Collaborative Ethnography. Retrieved April 22, 2024, from https://worldpece.org/content/seeing-standards-visualization-metadata-universe-0

Tracy-Ventura, N., Paquot, M., & Myles, F. (2020). The Future of Corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (1st ed.). Routledge. https://doi.org/10.4324/9781351137904

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1029–1034). European Language Resources Association (ELRA). Retrieved April 2, 2026, from https://aclanthology.org/L12-1227/

Velterop, J., & Schultes, E. (2021). An Academic Publishers' GO FAIR Implementation Network (APIN) (A. De Kemp, Ed.). *Information Services & Use*, *40*(4), 333–341. https://doi.org/10.3233/ISU-200102

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018–160018. https://doi.org/10.1038/sdata.2016.18

Wulff, S. (2023). Corpus Research. In A. Chaouch-Orozco, E. Puig-Mayenco, J. Rothman, J. Cabrelli, J. González Alonso & S. M. Pereira Soares (Eds.), *The Cambridge Handbook of Third Language Acquisition* (pp. 683–695). Cambridge University Press. https://doi.org/10.1017/9781108957823.027

Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation

Daniele Melaccio
ILC-CNR, Pisa, Italy
name.surname@cnr.it

Federico Boschetti
ILC-CNR, Pisa, Italy
name.surname@cnr.it

Monica Monachini
ILC-CNR, Pisa, Italy
name.surname@cnr.it

Pietro Sichera
ILIESI-CNR, Rome, Italy
name.surname@cnr.it

Abstract

We present an ontology-based mediation approach for integrating CLARIN language resources into the H2IOSC semantic framework, with the aim of enabling semantic interoperability across the Social Sciences and Humanities. Building on CMDI and CLARIN CCR, our work develops a layered semantic alignment workflow in which CLARIN metadata are progressively reinterpreted through multiple mediation steps. In a first layer, CMDI metadata are aligned with CIDOC CRM and SSHOCro, which are extended through a CLARIN domain ontology that formalises distinctions between datasets, tools, and services as reflected in discovery environments such as the Virtual Language Observatory, Resource Families, and the Language Resource Switchboard. In a second layer, these domain-level representations are projected into H2IOSCro, a Marketplace-oriented ontology that introduces entities, attributes, and typed relations required for operational federation within H2IOSC. The resulting ontology-based representation is finally projected onto the Marketplace operational data model, supporting metadata harvesting, normalisation, validation, and exposure through standard mechanisms such as OAI-PMH. By explicitly connecting semantic modelling, ontology mediation, and platform-level integration, this layered approach enhances discoverability, preserves compatibility with existing CLARIN metadata practices, and contributes to FAIR- and EOSC-aligned interoperability.

1 Introduction

This paper addresses the problem of semantic interoperability between CLARIN and H2IOSC by proposing an ontology-based mediation workflow that connects community-specific metadata practices with a shared semantic framework and an operational Marketplace environment. H2IOSC, the Humanities and Cultural Heritage Open Science Cloud,[1] brings together the Italian nodes of four European research infrastructures — CLARIN-IT, DARIAH-IT, E-RIHS.it, and OPERAS-IT — into a collaborative cluster that provides an open, multidisciplinary environment for advanced research on complex digital data and cultural objects.

In this context, interoperability is a cornerstone: it enables collaboration, data sharing, and resource integration across disciplines and technologies, and aligns H2IOSC with the EOSC FAIR principles. To address this challenge, H2IOSC adopts an ontology-based methodology for integrating heterogeneous domains, in continuity with CLARIN's long-standing commitment to FAIR, sustainable, and reusable metadata practices (de Jong et al., 2018).

Within CLARIN, language-related datasets and tools are described through the Component Metadata Infrastructure (CMDI) (Windhouwer & Goosen, 2022), supported by the CLARIN Concept Registry (CCR; Schuurman and Windhouwer 2015), which provides persistent identifiers and controlled vocabularies (Section 2). CMDI has proven effective in supporting federated discovery and long-term sustainability of language resources. Earlier work such as CMD2RDF (Windhouwer et al., 2017) demonstrated

[1] Official website https://www.h2iosc.cnr.it/

Daniele Melaccio, Federico Boschetti, Monica Monachini, and Pietro Sichera 2025. Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 207–218. DOI: https://doi.org/10.3384/ecp222

the feasibility of serialising CMDI records into RDF, opening the way towards Linked Open Data publication. However, syntactic conversion alone does not ensure semantic interoperability across infrastructures, nor does it provide a stable conceptual basis for cross-domain integration and platform-level reuse.

Several initiatives have explored ontology-based mediation of CMDI metadata. In the context of the PARTHENOS project, CMDI descriptions were mapped to a CIDOC Conceptual Reference Model (CRM)–based conceptual model as part of an effort to harmonise metadata across humanities research infrastructures (Durčo et al., 2018). More recently, the SSHOC project proposed a mapping of CMDI to SSHOC Reference Ontology (SSHOCro) as a proof-of-concept crosswalk for selected metadata standards (Tsouloucha et al., 2021). These efforts demonstrated the conceptual feasibility of aligning CMDI with CRM-based reference ontologies, while also highlighting practical challenges such as the heterogeneity of CMDI profiles and the limited semantic control of certain relation fields. Importantly, these mappings were primarily conceived as exploratory semantic alignment exercises, rather than as components of an operational ingestion pipeline.

Building on this body of work, we present an end-to-end operational workflow that integrates CLARIN metadata into the H2IOSC ecosystem through a layered ontology-based mediation. CMDI metadata is progressively aligned to CIDOC CRM and SSHOCro, extended through CLARIN-oriented domain ontology constructs grounded in established community practices, and further specialised in H2IOSC through a dedicated Marketplace-oriented ontology layer, formalised in H2IOSC Reference Ontology (H2IOSCro). The introduction of this second mediation layer is motivated by the role of the H2IOSC Marketplace as the federated operational environment where resources and services originating from multiple research infrastructures must be ingested, validated, curated, related, and published under a shared catalogue framework. In this setting, semantic interoperability must directly support platform-level requirements such as governance, provenance attribution, controlled categorisation, and structured relations between catalogue entities.

The resulting ontology-based representation is therefore projected onto the Marketplace operational data model, enabling CLARIN resources and services to be exposed and harvested via OAI-PMH[2] and made discoverable and reusable alongside assets from other infrastructures. While reusing the layered design principles introduced in SSHOC, our approach extends them in two key directions. First, ontology extensions are explicitly grounded in CLARIN community catalogues, modelling datasets and services as they are presented through Resource Families and the Language Resource Switchboard. Second, semantic modelling is directly connected to platform integration through H2IOSCro, which provides the conceptual bridge between domain-level descriptions and the operational structures required for Marketplace ingestion and management. In this way, semantic alignment becomes an integral component of an end-to-end interoperability workflow rather than an isolated transformation step.

The contribution of this paper is therefore twofold: (i) we refine and extend existing CMDI-to-SSHOCro mapping efforts by contextualising them within the H2IOSC Common Semantic Framework and introducing CLARIN-oriented domain ontology extensions; (ii) we bridge semantic modelling and platform integration by explicitly linking the ontology layer to an operational data model and standard harvesting mechanisms. This work goes beyond previous mapping efforts by introducing a two-step ontology-based mediation workflow that connects CLARIN metadata to an operational Marketplace environment.

The remainder of the paper is structured as follows. Section 2 introduces CLARIN metadata sources and semantics. Section 3 presents the first mediation layer from CMDI/CCR to CIDOC CRM and SSHOCro, including the SSHOC baseline mapping. Section 4 discusses the second mediation layer towards H2IOSCro. Section 5 describes the projection from the ontology layer to the operational data model and the Marketplace ingestion mechanisms. Section 6 concludes with future directions.

[2]The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a widely adopted protocol for collecting metadata records from distributed repositories.

2 CLARIN Metadata Sources and Semantics

CLARIN is a distributed research infrastructure that provides sustained access to digital language resources and tools for researchers in the humanities and social sciences. Within this ecosystem, interoperability is achieved through a combination of flexible metadata structures, shared semantic references, and federated discovery services. Together, these elements form a coherent metadata framework that supports both local autonomy and cross-centre integration.

At the core of this framework lie the Component Metadata Infrastructure (CMDI) and the CLARIN Concept Registry (CCR), which together provide the structural and semantic foundations for describing CLARIN resources. On top of this metadata layer, a set of access and discovery services — the Virtual Language Observatory (VLO), the Resource Families, and the Language Resource Switchboard — offer complementary views over the same metadata space, supporting discovery, classification, and operational reuse.

2.1 CMDI as the metadata backbone for CLARIN resources

CMDI is a modular, XML-based metadata framework designed to support the description of heterogeneous language resources in a federated infrastructure (Windhouwer & Goosen, 2022). Rather than imposing a single monolithic schema, CMDI allows metadata profiles to be composed from reusable components, which can be combined and extended to address the specific needs of different resource types, communities, and centres.

This design reflects a deliberate balance between standardisation and flexibility. On the one hand, CMDI promotes interoperability by encouraging reuse of components and shared structural patterns. On the other hand, it preserves local expressivity by allowing centres to define profiles tailored to their holdings, whether they describe textual corpora, lexical resources, annotation datasets, or linguistic services. As a result, CMDI has proven scalable and sustainable across the diverse landscape of CLARIN repositories.

From a semantic perspective, CMDI can be understood as a pre-semantic framework. While it supports the use of controlled vocabularies and semantic references, it does not enforce a unified conceptual model across profiles. This choice has been instrumental in fostering adoption, but it also implies that stronger semantic alignment must be achieved downstream, when metadata from different sources are integrated across infrastructures. In this sense, CMDI provides a stable structural substrate upon which ontology-based mediation can later operate.

2.2 The CLARIN Concept Registry as a semantic anchoring layer

The CLARIN Concept Registry (CCR) complements CMDI by providing a shared registry of concepts identified by persistent identifiers (Schuurman & Windhouwer, 2015). CCR concepts are used within CMDI profiles to populate metadata elements with semantically controlled values, mitigating terminological ambiguity across centres and profiles.

Unlike earlier approaches based on data category registries, the CCR adopts a concept-based model, allowing meanings to be defined independently of specific schema elements. Although the CCR is not a formal ontology, it plays a crucial semantic role within the CLARIN ecosystem. CCR concepts act as stable reference points that anchor metadata values to a shared semantic space, enabling consistent interpretation across heterogeneous descriptions.

This semantic anchoring function is particularly relevant in a federated context, where the same conceptual distinction may be realised through different structural choices in CMDI profiles. Here, it is important to distinguish the complementary roles of CMDI and CCR: CMDI primarily provides a flexible structural model for organising metadata descriptions, but remains pre-semantic in the sense that it does not impose shared conceptual commitments across profiles. CCR, instead, operates as a proto-semantic layer, offering a controlled vocabulary of persistent concepts that can be used to stabilise meaning across heterogeneous descriptions. By decoupling terminology from profile structure, CCR enables lightweight semantic convergence within CLARIN, preparing CMDI metadata for subsequent ontology-based mediation and for alignment with more formal semantic frameworks such as CIDOC CRM and SSHOCro.

CMDI organises metadata *syntactically*, CCR stabilises metadata *lexically*, ontologies mediate metadata *conceptually*.

2.3 CLARIN discovery: VLO, Resource Families, Switchboard

CLARIN metadata is exposed and exploited through several discovery and access services that provide different perspectives on the same underlying CMDI descriptions.

The Virtual Language Observatory (VLO)[3] aggregates CMDI records harvested from all CLARIN centres into a unified discovery interface. It supports metadata-driven search and filtering across resource types, languages, and thematic dimensions, serving as the primary entry point for users seeking datasets and tools. The VLO thus operates mainly at the level of record aggregation, emphasising findability and cross-repository discovery.

The Resource Families[4] offer a complementary, conceptually oriented view of CLARIN datasets. Resources are grouped into intuitive categories such as *Reference Corpora*, *Parallel Corpora*, *Treebanks*, or *Lexical Resources*. thus providing a curated semantic classification layer that supports thematic exploration and conceptual orientation, particularly for non-expert users.

On the services side, the Language Resource Switchboard (LRS)[5] focuses on operational reuse by connecting datasets to compatible tools through pragmatic matching mechanisms based on declared formats, languages, and processing capabilities (Zinn, 2018). The Switchboard thus foregrounds functional interoperability and usage scenarios, bridging discovery and analysis.

Taken together, these three access points can be interpreted as complementary semantic views over the same metadata space. The VLO provides a discovery-oriented perspective centred on metadata aggregation; the Resource Families introduce a domain-level conceptual classification of datasets; and the Switchboard reflects an operational perspective focused on tool applicability and reuse. The categories and perspectives embodied in CLARIN's discovery services provide valuable guidance for identifying domain-relevant concepts and relations to be formalised in higher-level semantic frameworks. As shown in the following sections, these implicit semantics play a key role in shaping the ontology-based mediation strategy adopted to integrate CLARIN metadata into SSHOC, first, and then the H2IOSC ecosystem.

3 Layered Semantic Mediation: from CMDI to a CLARIN Domain Ontology

Within CLARIN, early work on exposing metadata in the Semantic Web context focused on the transformation of CMDI records into RDF, most notably through the CMD2RDF approach (Windhouwer et al., 2017). CMD2RDF demonstrated that CMDI metadata can be serialised as RDF triples while preserving the structure and content of the original XML records. This represented a crucial step towards Linked Open Data compatibility and enabled basic integration with RDF-based tools and infrastructures. CMD2RDF operates at a syntactic and structural level. The resulting RDF graphs largely mirror the structure of CMDI profiles, reflecting local modelling choices and profile-specific components. As a consequence, CMD2RDF enables technical interoperability and data exposure.

H2IOSC integrates data, tools, and services from E-RIHS.it, DARIAH-IT, CLARIN-IT, and OPERAS-IT through a Common Semantic Framework (CSF), inspired by the layered interoperability approach developed in the SSHOC project (SSHOC Consortium, 2021a, 2021b, 2021c). The CSF adopts a multi-layer ontology design in which a shared conceptual backbone enables semantic interoperability across infrastructures, while domain-specific extensions preserve the richness and specificity of individual research communities.

Our work builds on the insight provided by CMD2RDF and goes beyond syntactic transformation by introducing an explicit ontology-based mediation layer. Rather than directly translating CMDI structures into RDF, we reinterpret CMDI metadata through a shared semantic framework, allowing resource descriptions to be compared, integrated, and reused across contexts while preserving their original CMDI

[3] Virtual Language Observatory: https://vlo.clarin.eu/
[4] CLARIN Resource Families overview: https://www.clarin.eu/resource-families
[5] CLARIN Language Resource Switchboard: https://switchboard.clarin.eu/

representations. In continuity with SSHOC, this backbone is grounded in CIDOC CRM and its SSH-oriented extension SSHOCro, which provide widely recognised reference semantics for modelling cultural heritage objects, research datasets, tools, and scholarly activities. Rather than replacing existing metadata standards, this layered architecture supports mediation: heterogeneous descriptions remain locally expressed through community practices, but can be aligned within a common semantic space to enable cross-domain discovery and reuse.

Within this framework, CLARIN resources require a dedicated mediation layer that bridges the flexibility of CMDI metadata with the abstraction of top-level reference ontologies. CMDI profiles provide structurally rich but semantically heterogeneous descriptions, while CLARIN discovery services implicitly rely on stable distinctions between datasets, tools, and services. Ontology-based mediation makes these distinctions explicit, providing a conceptual interface between CLARIN-specific practices and the shared SSH interoperability layer offered by CIDOC CRM and SSHOCro.

3.1 CIDOC CRM and SSHOCro as mediation frameworks

CIDOC CRM, standardised as ISO 21127, is a high-level ontology designed to enable semantic interoperability in the cultural heritage and humanities domains. Its event-centric modelling paradigm represents entities, activities, agents, and their relationships in a way that is both expressive and extensible, making it suitable as a cross-domain reference model. SSHOCro builds on CIDOC CRM by introducing extensions tailored to the Social Sciences and Humanities, including constructs for datasets, tools, research activities, and the research data lifecycle (Bekiari et al., 2022). By inheriting the event-centric and bottom-up modelling principles of CIDOC CRM, SSHOCro offers a stable yet flexible semantic backbone that can accommodate diverse disciplinary perspectives. Neither CIDOC CRM nor SSHOCro are intended to replace community-specific metadata standards such as CMDI. Instead, they act as mediation frameworks, providing a shared semantic reference layer to which heterogeneous metadata schemas can be aligned. This makes them particularly well suited for federated environments such as H2IOSC, where interoperability must be achieved without imposing a single upstream metadata schema.

A first attempt to align CMDI metadata with a CIDOC CRM-based ontology was carried out within the SSHOC project, as documented in (Tsouloucha et al., 2021). This work provided a proof-of-concept mapping of selected CMDI profiles to SSHOCro, demonstrating the feasibility of expressing CMDI metadata within a formal semantic framework. A key methodological contribution of this mapping is the explicit distinction between the metadata record and the described scholarly resource, modelled as separate entities connected through an aboutness relation. This distinction is particularly important in federated infrastructures, where metadata records often contain administrative, technical, and provenance information that should not be conflated with the resource itself. The SSHOC mapping exercise also provided a valuable proof of concept, demonstrating the feasibility of expressing CMDI descriptions within a CIDOC CRM–compatible semantic framework, despite the heterogeneity of CMDI profiles and the flexibility of their internal structures. In this sense, the SSHOC first attempt represents a foundational step towards ontology-based mediation for CLARIN metadata, and it directly informs the approach adopted in the present work.

Building on this baseline, we extend the SSHOC mapping perspective by grounding ontology development in CLARIN community practices and by integrating the resulting semantic layer into an end-to-end operational workflow targeting Marketplace ingestion and reuse within H2IOSC.

3.2 Towards a CLARIN domain ontology

Building on these insights, our work introduces an explicit CLARIN domain ontology as an intermediate mediation layer between CMDI metadata and the CIDOC CRM/SSHOCro backbone. The motivation for this layer is twofold. First, CLARIN constitutes a well-defined ecosystem with established conceptual distinctions between datasets, tools, and services that are not explicitly formalised in CMDI. Second, these distinctions are already operationally embedded in CLARIN discovery and access platforms, such as the Resource Families and the Language Resource Switchboard. Rather than mapping CMDI metadata directly to a highly abstract reference ontology, we therefore introduce CLARIN-specific ontology nodes that capture these community practices in a formal and reusable way. This approach preserves the

semantics familiar to CLARIN users while enabling alignment with broader SSH frameworks. These modelling choices are explicitly grounded in CLARIN community practices, rather than being imposed from the reference ontology.

The mapping strategy follows a combined top-down and bottom-up approach. In the top-down phase, high-level CMDI metadata elements — such as resource type, contributors, and funding information — were aligned with CIDOC CRM classes selected for their semantic proximity and cross-domain relevance. The entry point for all resources was established at the level of *E1 CRM Entity* (the most general class in CIDOC CRM), ensuring a coherent semantic foundation. In a second step, SSHOCro was used to refine this alignment by introducing SSH-specific constructs for datasets and tools. To capture the specificity of the CLARIN domain, ontology extensions were then introduced based on established CLARIN catalogues.

Datasets. Under *SHE1_Dataset* (SSHOCro class representing datasets), we introduced top-level subclasses such as *Corpus* and *LexicalResource*, further specialised through subclasses inspired by the CLARIN Resource Families (e.g., *ReferenceCorpus*, *ParallelCorpus*, *Lexicon*). This formalises an existing classification scheme and anchors it within a shared semantic framework.

Tools and services. Under *SHE10_Tool* (SSHOCro class representing tools and services), we defined the superclass *LinguisticTool*, extended with subclasses derived from the Language Resource Switchboard taxonomy. Examples include *CorpusQueryTool*, *PartOfSpeechTagger*, and *NamedEntityRecognitionTool*. This enables tools to be described using the same semantic framework as the datasets they operate on.

All mappings and extensions were formalised in RDF and OWL to ensure compatibility with Semantic Web technologies. Ontology development and validation were carried out using Protégé,[6] with OntoGraf employed for visual inspection of class hierarchies and relationships.

In contrast to the pragmatic matching mechanisms used in the CLARIN Switchboard — where compatibility is determined through file formats or declared input/output types — the ontology-based approach enables explicit semantic assertions linking datasets, tools, activities, and agents. Relationships such as *UDPipe analyses ParlaMint* or *ParlaMint is processed with NameTag* become part of a reusable semantic graph, supporting stable integration and cross-infrastructure interoperability. This ontological ramification provides a structured bridge between the scholarly domain level and the operational requirements of federation. By making CLARIN-specific concepts explicit at the ontology layer, the model prepares the subsequent projection step in which these resources are re-expressed as catalogue-level platform entities, ultimately enabling their integration into the Marketplace-oriented representation introduced through the *H2E1_Item* construct in H2IOSCro.

In the following section, we build on this foundation to describe the transition from the CLARIN domain ontology to the H2IOSC semantic framework, where domain ontologies are integrated under a shared top-level ontology.

4 Layered Semantic Mediation: Extending SSHOCro to H2IOSCro

The first mediation layer ensures that CLARIN datasets and services can be expressed through interoperable and conceptually stable categories, aligned with CIDOC CRM and SSHOCro. However, in H2IOSC interoperability must also support an operational federation scenario: heterogeneous resources originating from multiple RIs must be ingested, validated, curated, related, and published within a shared catalogue. This second set of requirements motivates a further mediation layer, formalised in the *H2IOSC Reference Ontology (H2IOSCro)* as a Marketplace-oriented extension of SSHOCro. This second mediation layer is formally implemented through the H2IOSCro ontology module, which introduces a set of Marketplace-specific classes *(H2E*)* designed by project's partners to represent catalogue entities, governance states, and operational interoperability structures. H2IOSCro is explicitly designed to mirror the conceptual structure of the H2IOSC Marketplace, providing an ontological specification of the same entities, roles, and relations that are later instantiated at the operational level in the Marketplace data model.

[6] Protégé official website: https://protege.stanford.edu/

This second mediation layer is required to bridge the gap between scholarly semantic descriptions and the operational requirements of the federated catalogue.

In H2IOSC, the Marketplace is not only a discovery interface. It is an operational catalogue where resources are managed as platform entities, undergoing governance actions (e.g., validation and publication), retaining provenance to their source infrastructure, and supporting structured links to other catalogue objects. In this context, a purely scholarly description of a dataset or tool is insufficient: the platform needs explicit constructs for (i) catalogue units, (ii) roles and responsibilities, (iii) controlled categorisation, (iv) access and licence information, and (v) typed relations that support navigation and, in perspective, orchestration.

H2IOSCro addresses these needs by introducing platform-oriented modelling constructs that remain compatible with the SSHOC layered approach, while providing explicit semantics for the Marketplace operational lifecycle. This requirement is reflected in the Marketplace specification, where the platform distinguishes between the scholarly resource as a domain entity and its operational representation as a catalogue-managed Item. This distinction anticipates the architecture of the Marketplace data model, in which Items constitute the primary operational units while remaining explicitly grounded in ontology-level constructs defined in H2IOSCro.

4.1 Core H2IOSCro nodes and relations

H2IOSCro adopts an *Item-centric* representation aligned with the H2IOSC Marketplace operational data model. This alignment is intentional: the H2IOSCro Item hierarchy corresponds directly to the structural core of the Marketplace data model discussed in Section 5, ensuring continuity between semantic mediation and operational implementation. The central construct of this layer is the **Item**, formalised as *H2E1_Item*, which acts as the primary catalogue entity managed by the platform. Items provide the operational entry point through which heterogeneous resources and services (datasets, tools, workflows, publications, projects) are exposed, curated, validated, and related within the federated Marketplace environment. In addition to the generic *H2E1_Item* construct, H2IOSCro defines a set of specialised subclasses representing different types of catalogue entities. These include, among others, *H2E6_Service*, *H2E7_Dataset*, and other domain-relevant item types, which provide a more precise semantic characterisation of the resources managed in the Marketplace.

These specialised nodes allow the ontology to distinguish between different categories of Items at the semantic level, while preserving a uniform operational interface in the Marketplace data model. For instance, services are not represented as generic Items but as instances of *H2E6_Service*, enabling more precise modelling of their role, behaviour, and relations within the catalogue. This distinction becomes particularly relevant when projecting CLARIN services into the Marketplace, as illustrated in Section 5.1.

In order to support governance, interoperability, and structured reuse, *H2E1_Item* is connected to a small set of tightly integrated Marketplace-specific node types.

Item governance and lifecycle. Each *H2E1_Item* includes persistent identifiers, multilingual descriptive fields, and governance attributes regulating its lifecycle in the catalogue. In particular, Items support explicit publication and validation states (e.g., `draft`, `published`), enabling the Marketplace to operate as a curated environment rather than a passive harvesting endpoint.

Typed properties and controlled semantics. Descriptive metadata is attached through *H2E2_ItemProperty*, which links an Item to an *H2E3_ItemPropertyCode* and its associated value. This "property code + value" mechanism functions as an extensible application profile: new properties can be introduced through governed codes without restructuring the overall model. Property codes are frequently aligned with established controlled vocabularies adopted in the SSH domain, including `activity` (TADIRAH), `discipline` (OFOS), and `language` (ISO 639-3), as well as operational descriptors such as `keyword`, `resource_category`, `object_format`, and `extent`. This ensures consistent faceted discovery across heterogeneous providers.

Identities, roles, and provenance attribution. Agents and organisations involved in the provision and governance of catalogue entities are represented through *H2E4_Identity*, linked to Items via explicit role assignments modelled as *H2E5_IdentityRole*. Roles capture responsibilities such as `provider`,

`contact`, `author`, `curator`, and `contributor`, ensuring that harvested Items remain traceable to their originating infrastructure and points of accountability.

Typed relations between catalogue entities. Marketplace-level interoperability also requires structured connections between Items. These are formalised through *H2E6_ItemRelation*, which links a source Item to a target Item via a controlled relation vocabulary. Relation types include, among others, `documents / is_documented_by`, `mentions / is_mentioned_in`, `extends / is_extended_by`, and `relates_to / is_related_to`. Such relations provide the minimal operational graph structure required for contextual navigation and reuse in a federated catalogue.

Access, licensing, and operational metadata. H2IOSCro further supports the attachment of access-related entities (e.g., licences, access policies, media objects), ensuring that reuse conditions and governance constraints are represented as first-class operational metadata. Provenance hooks such as `providerId`, harvesting job identifiers, and update timestamps are explicitly modelled to support synchronisation cycles and accountability within ingestion workflows.

Core structural pattern:

H2E1_Item → *H2E2_ItemProperty* → *H2E3_ItemPropertyCode*
H2E1_Item → *H2E4_Identity* (via *H2E5_IdentityRole*)
H2E1_Item → *H2E6_ItemRelation* → *H2E1_Item*

Taken together, these constructs define the platform-oriented semantics introduced by H2IOSCro on top of SSHOCro: a governance-aware catalogue layer that enables semantic mediation to become directly actionable within the Marketplace operational environment.

4.2 H2IOSCro: a Marketplace-oriented ontology

H2IOSCro can be understood as an extension that complements SSHOCro along two axes. First, it introduces an explicit separation between (i) the *domain entity* described at the SSHOCro/CIDOC CRM level (e.g., a dataset as an information object with creators, creation events, formats, etc.) and (ii) its *platform representation* as an Item (governed, validated, published, related, and enriched for faceted discovery). This separation prevents conflation between scholarly semantics and platform lifecycle semantics. Second, it provides a controlled "surface" for catalogue interoperability by stabilising the operational features that the Marketplace needs across providers: property codes aligned to controlled vocabularies, identity roles for governance and provenance, and relation types for catalogue graph navigation. These features are orthogonal to most domain ontologies, yet crucial in a federated Marketplace setting. H2IOSCro does not redefine the scholarly semantics of datasets and tools already captured in SSHOCro, but introduces the governance and catalogue layer required to operationalise them within a federated Marketplace.

CLARIN discovery services such as the Switchboard rely on pragmatic compatibility checks (e.g., input/output formats). In the Marketplace, instead, the catalogue must support stable, typed relations between Items to enable contextual discovery and curated navigation across infrastructures. This shifts interoperability from pragmatic compatibility checks to stable semantic catalogue assertions, supporting contextual navigation and future orchestration scenarios. H2IOSCro supports this by modelling explicit relation objects, enabling patterns such as: (i) a publication `documents` a dataset, (ii) a dataset `relates_to` a tool, (iii) a workflow `mentions` or `relates_to` the services it composes, (iv) a service `extends` another service. These relations provide a minimal operational graph for the federated catalogue and act as anchoring points for future orchestration scenarios.

By introducing Item-centric constructs, controlled properties, identity roles, and typed relations, H2IOSCro prepares the final step of the workflow: the projection from ontology-level constructs to the operational data model adopted by the Marketplace and the corresponding ingestion mechanisms (e.g., harvesting interfaces such as OAI-PMH). In this way, H2IOSCro acts as the semantic bridge between ontology-level mediation and the concrete Marketplace ingestion workflow, where Items are harvested, normalised, validated, and published. H2IOSCro is thus conceived as the ontological specification underlying the Marketplace operational schema, from which a controlled projection is derived to support

harvesting, validation, and publication workflows. The next section describes how this projection is implemented in the operational data model.

4.3 Mapping CLARIN resources and services into H2IOSCro

The transition from the CLARIN domain ontology to H2IOSCro is a projection step: it does not replace CLARIN semantics but wraps them into Marketplace-compatible structures.

Datasets. A dataset instance classified in the CLARIN domain ontology (e.g., *Corpus*, *ParallelCorpus*, *LexicalResource*) is represented in the Marketplace as an Item of the appropriate type. The domain-level classification is retained (as a property and/or as a link to the underlying semantic representation), while operational descriptors are added: controlled categorisation (discipline, activity, resource category), multilingual textual descriptions, access/licence fields, media, governance states, and provenance hooks (provider/job identifiers). For instance, a corpus such as ParlaMint is represented as a Dataset Item enriched with language and discipline codes, access conditions, and provider attribution, while preserving its classification as a CLARIN Corpus at the domain layer.

Tools and services. Similarly, a tool/service instance (e.g., categories inspired by Switchboard such as taggers, parsers, NER tools) is represented as a Marketplace Item that can be discovered and curated consistently across providers. In addition to descriptive metadata, service Items are enriched with operationally relevant properties such as intended audience, standards, formats, and documentation links, supporting reuse-oriented discovery. Similarly, a tool such as UDPipe is represented as a Service Item linked through operational relations to the datasets it can process, enabling reuse-oriented discovery in the Marketplace catalogue.

Publications, projects, and workflows as catalogue objects. H2IOSCro also enables treating publications and projects as first-class catalogue Items, so that documentation and provenance context can be linked to datasets and tools through typed relations. Workflows, in particular, can be represented as catalogue objects connected to services/tools they orchestrate, even if actual execution remains external to the Marketplace.

5 Layered Semantic Mediation: From H2IOSCro to the Operational Data Model

The H2IOSC Marketplace operationalises the federation by providing a shared catalogue where resources and services originating from multiple research infrastructures can be ingested, normalised, curated, and exposed through uniform discovery and governance mechanisms. In this layer, semantic interoperability becomes actionable: ontology-level descriptions, stabilised through the mediation workflow, are projected into a platform-oriented representation that supports validation, publication, provenance attribution, and cross-provider navigation. This projection is grounded in the H2IOSCro ontology introduced in Section 4, and the operational data model implements a subset of its constructs optimised for platform requirements.

The operational data model is centred on the concept of Item, the primary unit of cataloguing and querying. As introduced in Section 4.1, an Item provides a platform-centric representation of a resource or service and aggregates the information required for discovery and reuse, including identifiers, type assignment, multilingual textual descriptions, contextual and access-related information, media, external references, and provenance attributes. This uniform representation allows the Marketplace to manage heterogeneous content without imposing a single upstream schema on providers, but still enabling coherent facets and filters through normalised fields. Item types cover both standard catalogue entities (datasets, tools, publications, projects) and platform-relevant entities such as *service* and *workflow*, which are particularly important to bridge discovery and integration. A distinctive feature of the model is the use of extensible properties governed by codes, implemented through a "property code + value" mechanism. Each `itemPropertyCode` identifies a property and determines its expected data type and, where applicable, the controlled vocabulary to be used. This solution functions as an application profile: it reduces ambiguity, improves cross-provider consistency, and supports evolution of the catalogue without disruptive structural changes.

To ensure effective cross-provider discovery, the Marketplace relies on controlled vocabularies for key

descriptive dimensions, such as disciplinary classification (OFOS), activities (TADIRAH), and language (ISO 639-3). Properties such as `discipline`, `activity`, or `language` only become operationally meaningful if their values are normalised and consistently interpreted across providers. This value-level stabilisation complements ontology-based mediation: controlled vocabularies support comparability of values in the operational catalogue, whereas the ontology layer stabilises semantics at the level of domain concepts and relations.

In a federated catalogue, describing resources alone is insufficient; responsibilities, provision, and contact points must also be represented explicitly. The Marketplace models identities and their roles with respect to Items, supporting structured attribution of `provider`, `contact`, `author`, `curator`, and related responsibilities. By distinguishing identity entities from their role assignments, the model preserves clarity of accountability and supports governance workflows. Provenance information is an integral part of the catalogue record: explicit references to the originating provider and external identifiers are necessary to manage synchronisation, avoid duplication, and preserve traceability, particularly when ingestion is performed through repeated harvesting cycles.

Beyond individual descriptions, the Marketplace supports explicit typed relations between Items, providing a minimal operational graph for contextual discovery and reuse. Relations enable patterns such as linking publications to datasets, tools to training resources, workflows to the services they compose, and projects to their outputs. These operational relations correspond to the *H2E6_ItemRelation* construct defined in H2IOSCro (Section 4.1), ensuring continuity between ontology-level relation semantics and catalogue-level navigation structures.

The federated role of the Marketplace requires reliable mechanisms for collecting and synchronising metadata from external sources. Ingestion is supported through standardised channels, notably OAI-PMH and REST APIs,[7] with workflows that normalise incoming records into the Item-based representation, property codes, and controlled vocabularies. At present, OAI-PMH endpoints are available for CLARIN-IT (developed within H2IOSC) and for OPERAS-IT, and the harvesting chain has been validated through complete ingestion cycles with records imported into the Marketplace staging environment. In practice, the normalisation of CMDI records into the Item-based model requires case-by-case adjustments, as CMDI profiles vary considerably in structure and granularity; certain metadata elements, such as loosely typed relation fields or free-text descriptions, lack sufficient semantic control for fully automated alignment and require manual intervention during the mapping phase. The complementary *providing* functionality, i.e., exposing Marketplace contents to external aggregators, is part of the Marketplace architecture but has not yet been validated end to end; it is nonetheless designed to support future integrations with external federated catalogues, including EOSC-aligned scenarios.

Moving from catalogue discovery to orchestration requires services to be described not only editorially, but also through structured, machine-readable descriptors that expose interfaces, prerequisites, and usage characteristics. In H2IOSC, this role is fulfilled by service manifests, conceived as standardised descriptors that can be harvested and indexed by the Marketplace. Within H2IOSC, orchestration is supported by dedicated components and is not executed by the Marketplace itself. In this context, *workflows* primarily refer to application-level workflows managed through WSO2 Micro Integrator,[8] enabling executable pipelines and integrations between service endpoints. Through the *workflow* Item type, the Marketplace can represent and expose such processes as catalogue objects, while execution and runtime mediation remain the responsibility of the orchestration layer.

In summary, the Marketplace operational data model is the target of an explicit projection from the ontology layer. H2IOSCro provides the Marketplace-oriented ontological specification; the data model then operationalises a controlled subset of these constructs as an application schema optimised for harvesting, validation, indexing, and governance. This projection does not aim at enabling ontology-level reasoning within the Marketplace, but at ensuring semantic continuity between mediation layers and platform implementation.

[7] Representational State Transfer (REST) APIs provide a standard, stateless interface for programmatic access to services and resources over HTTP.

[8] WSO2 Micro Integrator is a lightweight integration runtime used to design and execute application-level integration workflows, supporting mediation, routing, and orchestration of service interactions.

5.1 Illustrative example: representing a CLARIN service in the Marketplace

To illustrate how ontology-level mediation is projected into the operational data model, we consider the example of UDPipe, a widely used linguistic processing service within the CLARIN ecosystem. This example demonstrates how a CLARIN service is represented in the Marketplace while preserving domain-level semantics and enabling platform-level discovery and reuse.

In the Marketplace, UDPipe is represented as an instance of *H2E6_Service*, a specialised subclass of *H2E1_Item* in H2IOSCro. Its classification as a linguistic processing tool is retained through controlled properties aligned with CLARIN and SSH vocabularies. Descriptive metadata (e.g., title, description, supported languages) are combined with operational attributes such as access conditions, documentation links, and governance status.

From a semantic perspective, UDPipe is linked to the CLARIN domain ontology as an instance of *LinguisticTool* (subclass of *SHE10_Tool*), while at the Marketplace level it is enriched with controlled property codes (e.g., activity, language) and connected to other Items through typed relations. For example, UDPipe can be linked to datasets it processes via `relates_to` relations, supporting reuse-oriented discovery.

This example illustrates how the ontology-based mediation approach enables a consistent transition from domain semantics to operational catalogue representation, ensuring both conceptual coherence and practical usability within the federated Marketplace. Table 1 summarises this transformation across semantic mediation layers, from CMDI metadata to Marketplace representation.

Layer	Concept	UDPipe representation
CMDI	Resource type	Declared as a `ToolService` in CMDI XML metadata (`<ResourceType>`)
CLARIN ontology	Domain classification	Interpreted as a `LinguisticTool`, specialised as `PartOfSpeechTagger`
SSHOCro	Semantic alignment	Mapped to `SHE10_Tool` within a shared cross-domain semantic framework
H2IOSCro	Marketplace representation	Recast as `H2E6_Service`, a subclass of `H2E1_Item`, with governance semantics
Marketplace	Operational description	Instantiated as an Item with controlled properties, typed relations (e.g., `relates_to`), and provenance information

Table 1: Step-by-step transformation of a CLARIN service (UDPipe) across semantic mediation layers.

6 Future Directions and Conclusions

Beyond the specific case of CLARIN integration, this work highlights a broader methodological perspective on interoperability in federated research infrastructures. Interoperability is often conceived as a technical challenge of format alignment or protocol compatibility; the experience reported here shows instead that it is fundamentally a matter of semantics. Ontology-based mediation provides a mechanism for making conceptual assumptions explicit and reusable, thereby supporting long-term coherence across infrastructures.

The layered mediation strategy adopted in H2IOSC plays a crucial role in this respect. By clearly separating metadata practices, shared semantic reference models, and platform-oriented representations, this methodology allows individual research infrastructures to preserve their conceptual autonomy while contributing to a common semantic space through mediation layers. From this perspective, CMDI provides a stable pre-semantic structural framework, while the CCR offers semantic anchoring through persistent identifiers and controlled vocabularies. Ontology-based mediation builds on these foundations rather than replacing them, enabling semantic alignment where and when it is required.

The H2IOSC Marketplace further illustrates how semantic modelling gains practical relevance when embedded in operational contexts. As the convergence point of heterogeneous resources and services, the Marketplace is not merely a technical endpoint, but the environment in which ontology-level descriptions are tested against real-world requirements of discovery, governance, and reuse. By projecting semantic representations into an operational data model, the Marketplace makes interoperability actionable,

demonstrating how conceptual alignment can be translated into concrete platform capabilities without disrupting existing infrastructure practices.

Despite these results, several challenges remain. The heterogeneity of CMDI profiles requires partial manual intervention during mapping, and certain metadata elements — such as loosely typed relation fields or free-text descriptions — lack sufficient semantic control for fully automated alignment. Moreover, maintaining consistency between ontology evolution and operational data models requires continuous governance, particularly where semantic ambiguity or underspecified relations occur. These aspects highlight the need for further methodological and technical refinement.

While grounded in the collaboration between CLARIN and H2IOSC, the mediation approach presented here is transferable to other federated settings. Research infrastructures facing similar tensions between local metadata autonomy and cross-domain interoperability may benefit from adopting layered ontology-based strategies that explicitly connect semantic alignment to operational workflows. Such approaches support FAIR- and EOSC-aligned interoperability and provide a sustainable pathway for integrating heterogeneous resources into evolving research ecosystems, while also contributing to semantic governance by stabilising meanings across infrastructures that evolve independently.

Acknowledgments

This work is supported by the H2IOSC Project – Humanities and Cultural Heritage Italian Open Science Cloud, funded by the European Union – NextGenerationEU – NRRP M4C2 – Project code IR0000029.[9]

References

Bekiari, C., Kritsotaki, A., Tsouloucha, E., & Theodoridou, M. (2022). D4.20 SSHOCro (final version) [Publisher: Zenodo]. Retrieved April 11, 2025, from https://zenodo.org/records/6771757

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., & Van Uytvanck, D. (2018). CLARIN: Towards FAIR and responsible data science using language resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Durčo, M., Lorenzini, M., & Sugimoto, G. (2018). Something will be connected – semantic mapping from cmdi to parthenos entities. *Selected Papers from the CLARIN Annual Conference 2017*, *147*, 25–35. https://ep.liu.se/ecp/147/003/ecp17147003.pdf

Schuurman, I., & Windhouwer, M. (2015). CLARIN concept registry: The new semantic registry. *Proceedings of the CLARIN Annual Conference*.

SSHOC Consortium. (2021a, January). *D7.1 system specification of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.1). SSHOC. https://zenodo.org/records/4558302

SSHOC Consortium. (2021b, November). *D7.2 release of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.2). SSHOC. https://zenodo.org/records/5749465

SSHOC Consortium. (2021c, December). *D7.3 final release of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.3). SSHOC. https://zenodo.org/records/5871651

Tsouloucha, E., Kritsotaki, A., Bekiari, C., & Theodoridou, M. (2021, January 31). *D4.19 mapping of two indicative selected standards to the sshocro* (tech. rep. No. D4.19) (SSHOC Project Deliverable). Social Sciences and Humanities Open Cloud (SSHOC). https://doi.org/10.5281/zenodo.4457496

Windhouwer, M., & Goosen, T. (2022, October 10). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *Clarin* (pp. 191–222). De Gruyter. https://doi.org/10.1515/9783110767377-008

Windhouwer, M., Indarto, E., & Broeder, D. (2017). Cmd2rdf: Building a bridge from clarin to linked open data. In J. Odijk & A. Van Hessen (Eds.), *Clarin in the low countries* (pp. 95–103). Ubiquity Press. https://doi.org/10.5334/bbi.8

Zinn, C. (2018). Squib: The language resource switchboard. *Computational Linguistics*, *44*(4), 631–639. https://doi.org/10.1162/coli_a_00329

[9] https://www.h2iosc.cnr.it/

Implementing and Promoting Data Citation for CLARIN Resources at FIN-CLARIN

Mietta Lennes
Department of Digital Humanities
University of Helsinki
Finland
mietta.lennes@helsinki.fi

Ute Dieckmann
Department of Digital Humanities
University of Helsinki
Finland
ute.dieckmann@helsinki.fi

Martin Matthiesen
CSC - IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Tommi Jauhiainen
Department of Digital Humanities
University of Helsinki
Finland
tommi.jauhiainen@helsinki.fi

Jussi Piitulainen
Department of Digital Humanities
University of Helsinki
Finland
jussi.piitulainen@helsinki.fi

Krister Lindén
Department of Digital Humanities
University of Helsinki
Finland
krister.linden@helsinki.fi

Abstract

Citation instructions are provided for all corpora in the Language Bank of Finland to promote good reference practices in the use of research data. By utilizing persistent identifiers (PIDs) and curated metadata, uniform citations can be constructed automatically, even for forthcoming resources. This work outlines how citation components such as authors, publication year, versions, and publisher are currently managed in the Language Bank of Finland, and discusses challenges related to metadata accuracy and interoperability. Data citation practices could be harmonized across CLARIN repositories by supporting data citation more directly via CLARIN compliant metadata.

1 Introduction

Today, the use and reuse of shared data is commonplace in scientific research. In order to ensure that studies can be verified and replicated, researchers should provide persistent references to their data. Data referencing is technically possible, but the practices of citing data are not yet well established in all fields. Moreover, additional effort may still be required from the researcher to gather the necessary details and to construct a data citation in the format requested by the publisher.

In this article, we describe how data set citation practices are encouraged and supported by the Language Bank of Finland (hereafter "the Language Bank")[1]. The Language Bank is a national research infrastructure (RI) service offered by the FIN-CLARIAH RI[2] through the University of Helsinki and CSC - IT Center for Science. It is partly funded by the Finnish Ministry of Education and Culture through various funding instruments. The Ministry and its affiliated organizations are committed to promoting the FAIR principles (data should be Findable, Accessible, Interoperable, and Reusable; see Wilkinson et

[1] https://www.kielipankki.fi/language-bank/

[2] https://www.kielipankki.fi/organization/fin-clariah/

Mietta Lennes, Ute Dieckmann, Martin Matthiesen, Tommi Jauhiainen, Jussi Piitulainen, and Krister Lindén 2025. Implementing and Promoting Data Citation for CLARIN Resources at FIN-CLARIN. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 219–228. DOI: https://doi.org/10.3384/ecp222

al. (2016)). The aforementioned organizations are also strongly committed to develop Fairdata services together and to support the curation and storage of research data.

2 The current situation of data citation

Data repositories, including the Language Bank, should support researchers by systematically offering citation information for the data sets distributed by them, as recommended in the EU data citation guide (Publications Office of the European Union, 2022). Various guidelines have been published over the years, including those by Conzett and De Smedt (2022), the Tromsø recommendations for citation of research data in linguistics (Andreassen, Berez-Kroeker, Collister, Conzett, Cox, De Smedt, McDonnell, & Research Data Alliance Linguistic Data Interest Group, 2019), and the recommendations by CESSDA (Bornatici, Jernung, Alaterä, Tveit Sandberg, Strand, Štebe, & Trtíková, 2025)[3]. The recently published CLARIN Data Citation Guidelines (Matthiesen & Lenardič, 2025) reflect the aforementioned recommendations and are also supported by the Language Bank.

In the past, it has been difficult to cite data properly. Without well-established data citation conventions and persistent identifiers, it can be challenging to determine whether two similar-looking data descriptions appearing in two different contexts would in fact involve the very same data. Without trustworthy repositories and stable download locations, it is difficult to estimate whether a given data set would still remain accessible some years later. As a side-effect, researchers have also been inclined to make copies of all the potentially relevant data they were able to get their hands on.

For a long time, researchers have worked around these limitations by citing the papers describing the data, rather than citing the data sets directly. For derived data sets and data collections, the paper describing the original data set is often cited. However, unlike published papers, digital data sets may change over time for various reasons, and thus it is not possible to rely on research papers alone as the ultimate documents of the research data.

For example, a text corpus created before 1990 would not originally have been encoded in Unicode. However, in the research paper describing the corpus, ISO 8859-1 might have been stated as the character encoding of the texts. Given that the data was still in use a few years later, it would undoubtedly have been converted to Unicode for technical reasons, but the details of the conversion might not have been mentioned in any new scientific paper. Previous research papers can be cited by newer ones, but the original paper would not provide information about the changes that take place at a later point. In case a metadata record is made available, curated, and used systematically as a reference from the beginning, researchers will have consistent access to information on the original data set, including the details of the change in character encoding. In case cross-references are maintained between the metadata records, the researchers will also be able to find any newer, derived, or otherwise related versions of the data (see Matthiesen & Dieckmann, 2019, for details).

Persistent references to data cannot be taken for granted, which can be a replicability issue. For instance, in the data availability survey by Jauhiainen, 2024, the references to data sets were investigated in research articles published between the years 2018 and 2023 on the topic of automatic date detection in texts. The relevant datasets were available and sufficiently described in only three out of 22 articles. Of these three, two data sets were in GitHub repositories without any proper data citation instructions. They were not properly cited in the corresponding articles, either, apart from the links to the repositories. The relevant data set was properly referenced in just one article (by Grabovoy et al., 2021, who had stored their data set in the Mendeley Data service offered by Elsevier[4]).

Habits are difficult to break. As long as supervisors and publishers consider it sufficient to cite scientific articles on research data, proper data citation remains optional. The authors who submit their research to major scientific publications and conferences are usually required to deposit any relevant new data sets to an appropriate repository and to cite the data sets used in their research, but in practice, many researchers do not follow the recommendations.

[3] See also the CESSDA Data Citation Guide, https://datacitation.cessda.eu/

[4] https://doi.org/10.17632/95xt9sngzc.

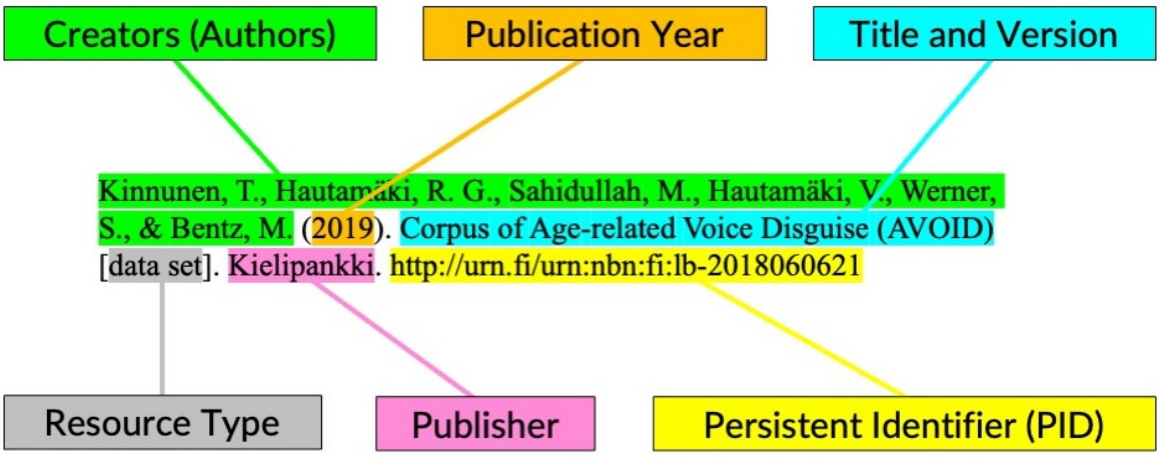

Figure 1: The structure of an example reference to a resource in the Language Bank of Finland.

The adoption of data citation may be slowed down due to the ways of measuring scientific impact. Publishing a data set may not yet count as scientific merit in a similar way to traditional publications.

By providing consistent data citation information, we cannot solve all the above-mentioned issues regarding data citation. Nevertheless, by making citation as easy as possible for researchers, we encourage the adoption of new practices. Let us now describe data citation at the Language Bank of Finland in more detail.

3 Citation Instructions at the Language Bank of Finland

The Language Bank maintains a public website (the Portal)[5] with information about the services of the Language Bank, including lists of the hosted language resources, e.g., corpora and tools. The list of corpora[6] is automatically generated from a separate resource database. The citation instructions for each corpus in the Language Bank are generated from the resource metadata stored in the database.

The instructions appear in the Portal via dynamic links using the resource PID as the citation key. The citation instructions for corpora consist of the following elements: **Creators (Authors)**, **Publication Year**, **Title and Version**, **Resource Type**, **Publisher**, and **PID** (see figure 1).

It is essential to provide a persistent identifier (PID) for each data set (Wittenburg, 2019). In the Language Bank, the PID used in the citation instruction always points to the metadata record of the resource in question. The metadata record provides the necessary details and can be updated when required, while the citation text can remain short and efficient. The citations work in a persistent and uniform way for all resources, regardless of the platform where the resource contents are stored.

In case a new resource is known to have been completed by a researcher and the deposition with the Language Bank is expected to take place in the near future, the Language Bank can provide citation instructions for the forthcoming resource, even before it is officially available.[7] The citation instruction can be reliably generated in the Portal as soon as the necessary details, including the PID, are included in the database.

[5] https://www.kielipankki.fi/language-bank/

[6] https://www.kielipankki.fi/corpora/

[7] "Data should be properly cited, even if not generally available, not fully available, or available only on restrictive terms." (Publications Office of the European Union, 2022)

License information is not included in the resource-specific citation instructions of the Language Bank, since licenses may be updated at a later point and this can be documented in the metadata. Moreover, the date of last access or retrieval has been excluded from the citation instructions. The individual resources that are centrally maintained by the Language Bank are typically not expected to change significantly after their publication date, and if they do change, the changes can be transparently logged in the metadata. Thus, the PID already ensures that a referenced dataset can be reused later, regardless of when it was accessed[8].

Accurate resource metadata are essential for the automatic generation of reference instructions. In this section, we describe why and how certain fields of information are included in the citation instructions of the Language Bank. We also point out some issues that researchers and repositories need to consider.

3.1 Creators (Authors)

According to the current practices in the Language Bank, the ordered list of authors or **creators** is set in the deposition agreement of the resource. When negotiating agreements, the Language Bank encourages depositors to list individual people as creators, as recommended by, e.g., Lenardič and de Maiti Tekavčič (2024), as this is good practice in terms of merit, provenance, and scientific replicability. However, in some cases, it can be appropriate to attribute organizations instead of or in addition to individuals. The depositors make the final decision on authorship.

For each resource in the Language Bank, information about the creators is stored in the database as the first and last names, listed in the order of appearance in the citation. For a more general solution, the list of authors/creators including their contact details, affiliations and the order of citation, should be stated in the metadata record.

Determining reference instructions for older datasets can be challenging, since resource metadata has not always been collected in a consistent way in the past, and the original contributors may no longer be available for consultation. For example, when reviewing the older resource metadata records of the Language Bank, it turned out that some pieces of information had quite frequently been recorded in the wrong places and some details were often missing. For some resources, the publisher had been erroneously mentioned as the creator of the resource, even when the publisher had not really contributed to collecting, organizing, formatting or enriching the data. When trying to track down the parties that should be attributed for specific resources, it is fortunate if the original deposition agreements and/or some relevant email archives are available. For instance the 'IPR holders' or 'Licensors', defined in some metadata records, were not necessarily the same parties that should be cited as the 'creators' or 'authors'. In some cases, where the source version is available under a public license requiring attribution, the attribution information might not be explicitly stated at the source. Thus, a lot of additional effort may be required in case the information about authorship is missing or incomplete and if there is no written agreement about the data.

3.2 Publication Year

For citation instructions, the publication year is interpreted as the year the corpus was published in the Language Bank (even if the collection was previously made available elsewhere), since this is the publication time of the specific resource edition over which the repository has control. The publication year is retrieved from the database and formatted as a four-digit number. For resources that have not yet been published, the text "forthcoming" is shown instead.

3.3 Title and Version

The title of the resource is agreed with the depositor. The same title is stored in the database and in the corresponding metadata record in English and Finnish. The resource title usually includes extensions describing the variant (e.g., Korp or Download[9]) and the version of the resource. Like books, data sets might be published again as new versions or editions. More recent versions can contain more data, or some annotation layers may have been added or updated, e.g. by using an improved annotation method.

[8]For a discussion on assigning new PIDs when datasets change over time, see Matthiesen and Dieckmann (2019).

[9]See https://www.kielipankki.fi/support/corpus-location/

3.4 Resource Type

The title of a corpus or a data set may sometimes be similar to the title of a published paper or research project. To avoid confusions, it is good practice to include information about the type of resource. Square brackets are often used to separate this element in the citation.

The resource type should be expressed in generally comprehensible terms that can be localized in many different languages. As we previously pointed out, excessive or potentially changing details should be avoided in data citations, since further information can be provided via the metadata record. Previously, the resource type *corpus* was used by default in the citation instructions provided by the Language Bank. However, there are many resources that cannot be categorized as 'corpora'. A common vocabulary for all types of resources has not been established. In addition, it is rather typical for language resources to include various combinations of different data types and structures, in which case selecting the resource type would not be straightforward. We therefore decided to use the more generic resource type *data set* (in Finnish, *aineisto*).

3.5 Publisher

In the citation instructions generated by the Language Bank, the publisher of every data set is the Language Bank, identified as "Kielipankki", since the Language Bank is responsible for maintaining the deposited data and for curating the metadata (see Publications Office of the European Union, 2022). Moreover, only Language Bank staff members can make changes to the metadata records.

3.6 Persistent Identifier (PID)

At the Language Bank, the PID of a resource points to the metadata record of the resource in question[10]. Consequently, it is possible to provide unambiguous citation instructions even before the resource has been officially published by the repository. This approach also allows the Language Bank to implement persistent tombstone pages after the data itself is no longer available, according to the FAIR principle A2[11]. In the citation instructions of the Language Bank, the PIDs are also made "actionable", i.e., clickable, as recommended in the data citation guide by the Publications Office of the European Union (2022). The link to the data access location is not included in the citation instructions. The metadata record provides the data access location in CMDI[12] compliant format for humans and machines.

The Language Bank uses URN-based PIDs supported by the National Library of Finland[13]. Handles are also supported as a backup system[14]. In line with Broeder et al. (2009), our PIDs do not contain semantic information. For practical reasons, new PIDs are constructed from dates and running numbers, but even the minting date can not be reliably derived from the identifier.

The PID of the resource points to the metadata record that provides information on how to access and use the resource (including license restrictions, technical documentation or guidelines for the user, information about annotation anomalies, etc.). The metadata can be updated by the Language Bank, e.g., to fix typos or to add new or previously missing information. Minor updates of the content of the resource itself do not result in a new metadata record and a new PID as long as the content is expected to be compatible with the resource version prior to the update. A minor update can be described just by editing the metadata record and the version number after the resource title that appears in the citation. For a more in-depth discussion, see Matthiesen and Dieckmann (2019, section 6).

3.7 Formatting the Output

The citation instructions are offered via the list of corpora in the Language Bank Portal in two languages, English and Finnish. Links to the instructions for individual resources can also be found in the metadata records. The links can be manually constructed if the resource PID is known[15].

[10] For a discussion of the special status of a "citable PID", see Matthiesen and Dieckmann (2019, section 7).

[11] https://www.go-fair.org/fair-principles/a2-metadata-accessible-even-data-no-longer-available/

[12] https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

[13] https://urn.fi/URN:NBN:fi-fe2024051430651

[14] The PIDs can be resolved if one of the systems is unavailable: https://urn.fi/urn:nbn:fi:lb-201710212#Persistent_identifiers

[15] For example, https://www.kielipankki.fi/viittaus/?key=urn:nbn:fi:lb-2024051601&lang=en.

The plain text citation instructions are constructed roughly according to the APA style[16] with one exception: the version information is included in the title. The general structure is in line with other CLARIN centres, e.g., CLARIN DSpace at LINDAT/CLARIAH-CZ[17]. Unlike the Language Bank, CLARIN DSpace does not use the resource type attribute (e.g., "[data set]").

In addition to the citation instructions in plain text, the BibTeX and Zotero versions are also provided. The most commonly used bibliographic styles for BibTeX do not include an entry type for data sets especially. The reference text is generated at the Language Bank with the BibTeX entry type `@misc`, which is also used by CLARIN DSpace and produces reasonable-looking reference items. For Zotero, the format instruction is rendered in BibTeX syntax as `@techreport`, since this produced the best result (first tested in 2015, still holds true in 2025). Finally, an additional link is generated to search for references to the resource from Google Scholar.

3.8 Data Provenance and Relations

Language resources often consist of texts or other materials that have been published elsewhere. Following the recommendations of Andreassen et al. (2019, chapter 1.1), the Language Bank offers citation instructions for entire data sets only. This is where the persistent identifier to the metadata record is useful and efficient. For instance, in case a large corpus of newspaper articles is used as a research data set, a reference can be provided to the entire corpus, without listing the bibliographic details of all the individual texts and their authors. If required, the original sources should be findable via the metadata.

A corpus may consist of several smaller corpora, each of which may have been created by different people. If using an individual text or a specific part of the larger data set, researchers can be instructed to add further references to the parts of the resource. In order to make this possible, sufficient information about the subcorpora should be readily available, e.g., via metadata relations or via external documentation. The necessary details should preferably be provided by the creator of the collection at the time of depositing or publishing the resource.

For instance, the *Uralic UD* resource group[18], available via the Language Bank, includes several versions of a collection of treebanks of Uralic languages from the Universal Dependencies project repositories. The Uralic UD corpus as a whole is provided, maintained and versioned by the Language Bank, but each original treebank has a different list of contributors. Therefore, the users of the resource are instructed to cite the entire corpus (this ensures persistent access), but it is also recommended to include additional references to the individual subcorpora if appropriate (for precise attribution). The details regarding the subcorpora were found in the source repositories on GitHub, and they are provided on the resource group page of the Language Bank, to which a link is available in the metadata records of the individual resource versions.

Resources can thus appear in several versions that may differ not only in the amount of primary data (e.g., original texts or speech recordings that accumulate from one version to the next), but also in the types and amounts of annotation and analysis files. Sometimes, the primary data set has been created by different people than the annotations that were added later. Since the contributions may be independent of each other, it is important not to include the creators of the original data set in the same reference instruction with the creators of the additional annotations without an agreement of all parties.

As an example, *ScotsCorr, the Helsinki Corpus of Scottish Correspondence (1540–1750)*[19] was originally published in the Language Bank in 2017. The suggested citation is as follows:

> Meurman-Solin, A., & Research Unit for the Study of Variation, Contacts and Change in English (VARIENG), University of Helsinki (2017). *Helsinki Corpus of Scottish Correspondence (1540-1750)* [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-201411071

However, new annotations were created later in another, unrelated project and published in the Language Bank as a different data set, the *Parsed Corpus of Scottish Correspondence (pcsc-src)*, included in

[16] https://apastyle.apa.org/style-grammar-guidelines/references/examples/data-set-references
[17] https://doi.org/10.17616/R30G6W
[18] Uralic UD resource group: http://urn.fi/urn:nbn:fi:lb-2022061003
[19] ScotsCorr resource group: http://urn.fi/urn:nbn:fi:lb-202104191

the same resource group. This data set has a citation instruction of its own:

> Gotthard, L. (2025). *The Parsed Corpus of Scottish Correspondence, source* [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2024070601

In addition, the users of the parsed corpus version are instructed to include a separate reference to the original corpus version as appropriate.

As seen from the examples above, the ”web” of the different versions, variants, and combinations of a data set can become quite complex over time. It is likely that not all the relevant changes and possibilities of the intermediate data sets end up being documented in a research paper. However, if it were possible to automatically discover the contributions of the authors and creators of related resources on the basis of the metadata, and if this information could be made available in a uniform and transparent manner, researchers would not be obliged to keep citing both the datasets and the respective papers.

In the Language Bank, the relations between resources are described by using the vocabulary defined by DataCite (see DataCite Metadata Working Group, 2024). However, although the DataCite relations provide a means of expressing provenance information in a machine-readable way, the relations are currently not displayed or utilized by metadata services. Therefore, the landing page where the PID of a resource points may not be able to offer all the information included in the metadata. CLARIN does have a standardized way of accessing the underlying CMDI metadata[20], but this standard is not universal. In order to interpret the relations correctly, the metadata crawler would need to examine the schema.

4 Discussion

The metadata records of the Language Bank already contain almost all the information needed to create the citation instructions. Ideally, the citations could be generated directly from the CMDI metadata and provided to the users via the metadata catalogue. Unfortunately, the order of authors or creators is currently not encodable in the metadata for cases where both authors and institutions should be mentioned. In addition, reference instructions may also need to be generated for forthcoming resources, which are not yet officially displayed via metadata services. Therefore, the information required for the citation instructions is currently retrieved from a separate resource database at the Language Bank, as described above.

There are some risks involved in generating reference instructions automatically from openly distributed metadata. Similarly to all descriptive metadata in CLARIN, the metadata of the resources in the Language Bank of Finland are not licensed, and the records are publicly available via the OAI-PMH[21] interface. Other metadata catalogues, such as the *European Language Grid* (ELG), can harvest and republish the metadata. The metadata may sometimes need to be re-processed to match the needs of the service in question. The external catalogue can then also generate reference instructions in a format of their own. Consequently, the citation instructions may look very different to end users.

Consider this citation of a Language Bank corpus, generated by ELG:

> The Suomi24 Sentences Corpus 2001-2017, Korp version 1.1 (2020, January 01). Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. Source: European Language Grid. https://live.european-language-grid.eu/catalogue/corpus/11759

Then compare the above reference to the recommended citation of the dataset, provided by the Language Bank:

> City Digital Group (2025). The Suomi24 Sentences Corpus 2001-2017, Korp version 1.3 [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2020021803

Ultimately, both references point to the same data. The version numbers differ between the two records, since the metadata provided via ELG was manually imported and it is not synchronized with

[20] see Wittenburg et al. (2024, Chapter 7)

[21] https://www.kielipankki.fi/language-bank/oai-pmh/

the master metadata record that is maintained at the Language Bank. Moreover, during the import, an additional version number was generated by ELG on top of the existing version number, and the link to the catalogue item promoted by ELG has completely replaced the original persistent identifier.

By allowing the metadata to be added into various catalogues, the data sets can gain more visibility. The downside is that the metadata may appear ”out of context”, as shown by the ELG example. To avoid the risk of disseminating contradictory citation instructions for the same data set, new metadata records should always contain a clear link to the original metadata from which they are derived. This metadata should be regularly checked for updates and treated as master metadata. The master metadata should be curated at a single location in agreement with the creators and providers of the original data set. There are also ongoing efforts aiming to promote the propagation of the minimal metadata. For instance, the ”FAIR Signposting” approach to metadata publishing design, endorsed by the EOSC Association, can offer a transparent, compliant, and straightforward mechanism for helping automated agents navigate through metadata spaces to locate the essential FAIR elements: the globally unique identifier (GUID), the data records, and the corresponding metadata records (Wilkinson et al., 2024).

If a data set includes or is a version of other datasets, the authors/creators of the parts or older versions should also be attributed in an appropriate way. All the creators of the subsets should not be listed in the citation instructions of a larger collection, since the list of ”authors” would tend to grow longer and longer, and the individual contributors of the subsets are probably not to be attributed for the entire collection. It is therefore essential that researchers are able to find information about the relations between data sets in the metadata record.

Even a well-formed data citation can only fulfill its purpose if the dataset or at least its metadata remain accessible. It is not enough to provide the data online and mechanically assign a persistent identifier to it, unless the repository is supported and the identifier is curated in the long term. Thus, some infrastructure is required before scientific data citation can fully work.

5 Conclusions and Future Prospects

Citation indices may still reward researchers for their traditional publications rather than their published data sets. However, research data should not be cited by referencing related articles or books alone. A published article cannot be modified after publication and can easily become outdated, whereas the metadata record of the data set can be updated when necessary. Well-curated metadata can maintain and even add to the scientific impact of a data set after the data has been made available. The provenance of the data set can be made more transparent, if the relations of a data set with other resources are described systematically. By ensuring that all the metadata required for citation are available systematically and in an interoperable and user-friendly way, CLARIN can support the adoption of FAIR-compliant data citation practices.

The system for generating citation instructions at the Language Bank of Finland was developed in 2015. Small modifications have been implemented over the years, and the current citation instructions now also comply with the recently published CLARIN Data Citation Guidelines (Matthiesen & Lenardič, 2025). At the Language Bank, the information for all components in the citations is currently retrieved from the internal resource database. This allows the Language Bank to generate citations even for forthcoming resources, since the database includes information about the publication status of each resource. The database also supports the order of author/creator names, unlike the current CMDI metadata profile used by the Language Bank. Ideally, however, the details required for generating the citation instructions should be included in the public metadata records of all data sets in CLARIN.

Metadata records should be widely accessible by default. It is often recommended that metadata be made available in the public domain or under a public license, with practically no limitations to what others can and cannot do with the information. However, if metadata records are copied from one platform to another, converted into a different metadata schema and re-published in isolation from the originals, with no method for retrieving potential updates, the copied metadata can easily become distorted. This is where CLARIN and EOSC could help by developing policies and best practices for displaying and using metadata outside of its original context.

The CMDI components supporting consistent citations should be designed by the CLARIN community and deployed over time by all repositories. This would allow for the development of central citation services, whereas the local consortia could focus on providing high-quality metadata. At the Language Bank, we are working on augmenting our CMDI schema to achieve this goal.

References

Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., De Smedt, K., McDonnell, B., & Research Data Alliance Linguistic Data Interest Group. (2019, December). Tromsø recommendations for citation of research data in linguistics. *Zenodo*. https://doi.org/10.15497/rda00040

Bornatici, C., Jernung, A., Alaterä, T. J., Tveit Sandberg, L., Strand, K., Štebe, J., & Trtíková, I. (2025, March). CESSDA recommendations on data citation: Practical recommendations for key stakeholders. *Zenodo*. https://doi.org/10.5281/zenodo.15043854

Broeder, D., Dreyer, M., Kemps-Snijders, M., Witt, A., Kupietz, M., & Wittenburg, P. (Eds.). (2009). *Persistent and unique identifiers*. https://office.clarin.eu/pp/D2R-2b.pdf

Conzett, P., & De Smedt, K. (2022). Guidance for citing linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 143–155). The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/12200.003.0015

DataCite Metadata Working Group. (2024). *DataCite metadata schema documentation for the publication and citation of research data and other research outputs* (Version 4.6). DataCite e.V. https://doi.org/10.14454/mzv1-5b55

Grabovoy, A., Bakhteev, O., & Chekhovich, Y. (2021). The automatic approach for scientific papers dating. *2021 Ivannikov Ispras Open Conf. (ISPRAS)*, 107–113. https://doi.org/10.1109/ISPRAS53967.2021.00020

Jauhiainen, T. (2024). Data availability and evaluation reproducibility for automatic date detection in texts, a survey. *Digital Humanities in the Nordic and Baltic Countries Publications*, *6*(1). https://doi.org/10.5617/dhnbpub.11511

Lenardič, J., & de Maiti Tekavčič, K. P. (2024). The citation of language resource technologies in CLARIN. In V. Vandeghinste & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference, 15 – 17 October 2024, Barcelona, Spain* (pp. 46–50). CLARIN ERIC. https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Matthiesen, M., & Dieckmann, U. (2019). A PID is a promise — Versioning with persistent identifiers. In I. Skadina & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018* (pp. 103–112, Vol. 159). CLARIN ERIC. https://doi.org/10.3384/ecp159

Matthiesen, M., & Lenardič, J. (2025). CLARIN Data Citation Guidelines. https://doi.org/10.34733/DOC-189

Publications Office of the European Union. (2022). *Data citation – A guide to best practice*. https://data.europa.eu/doi/10.2830/59387

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Grootveld, M., Dennis, R., Hecker, D., Huber, R., Soiland-Reyes, S., Van de Sompel, H., Czerniak, A., Thurston, M., Lister, A., & Gaignard, A. (2024). Report on FAIR Signposting and its uptake by the community. *Zenodo*. https://doi.org/10.5281/zenodo.10490289

Wittenburg, P. (2019). From Persistent Identifiers to Digital Objects to Make Data Science More Efficient. *Data Intelligence*, *1*(1), 6–21. https://doi.org/10.1162/dint_a_00004

Wittenburg, P., Van Uytvanck, D., Zastrow, T., Straňák, P., Broeder, D., Schiel, F., Boehlke, V., Reichel, U., & Offersgaard, L. (2024). *CLARIN B Centre checklist* (tech. rep. No. Version 7.4.1). CLARIN.

Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring Communities and Orchestrating Digital Services

Pietro Sichera ILIESI, CNR, Italy pietro.sichera@cnr.it
Monica Monachini ILC, CNR, Italy monica.monachini@cnr.it
Valeria Quochi ILC, CNR, Italy valeria.quochi@cnr.it
Nicola Giampietro ILIESI, CNR, Italy nicola.giampietro@cnr.it
Vittoria Fabiani ILIESI, CNR, Italy vittoria.fabiani@cnr.it
Roberta Bianca Luzietti ILC, CNR, Italy robertabianca.luzietti@cnr.it
Roberta Ottaviani ILC, CNR, Italy roberta.ottaviani@cnr.it
Daniele Melaccio ILC, CNR, Italy daniele.melaccio@cnr.it
Laura Baggiani ILIESI, CNR, Italy laura.baggiani@cnr.it

Abstract

This paper presents two components developed within the Humanities and Cultural Heritage Italian Open Science Cloud, H2IOSC, the Observatory and the Marketplace. Together, they support the SSH federated national cluster of Research Infrastructures by combining evidence-based community monitoring with an environment for discovering and combining digital resources and services. The Observatory adopts a mixed methods approach, integrating quantitative and qualitative instruments to document researchers' practices, needs, and expectations across partially overlapping domains. The Marketplace provides a catalogue of services and a workflow environment based on controlled ingestion, semantic normalisation, and interoperability with external sources. We also describe how the two components interact, linking community analysis with service provision. This interaction shows how CLARIN-IT and OPERAS-IT work together within H2IOSC to monitor uptake and inform service development.

1 Introduction

The Open Science paradigm has created new challenges and opportunities for the Social Sciences and Humanities (SSH) and Cultural Heritage (CH) disciplines. While the availability of digital tools and datasets is constantly increasing, fragmentation still characterises the research landscape for many domains. Researchers often struggle to navigate various independent repositories, incompatible data formats, and isolated service platforms, which hinders the full realization of the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016).

In this context, the Italian nodes of four major European Research Infrastructures (namely, CLARIN-IT, DARIAH-IT, E-RIHS.it, and OPERAS-IT) jointly contribute to the creation of a federated cluster of Research Infrastructures (RIs), aiming to overcome these barriers by fostering an open, federated digital ecosystem for SSH research communities in Italy, while maintaining solid connections with corresponding European counterparts within the European Open Science Cloud (EOSC): the Humanities and Cultural Heritage Open Science Cloud (H2IOSC hereafter, Degl'Innocenti et al., 2023).

One output of this initiative is the development of two key components, the H2IOSC Marketplace and its Observatory. The Marketplace serves as the main access point for the cluster. Unlike traditional static catalogues, it acts as an active orchestration layer allowing researchers to discover services from different providers and combine them into coherent workflows. The Observatory complements this function by monitoring community needs, service uptake, and user experiences through survey data, focus groups, and usage indicators. This establishes a feedback loop: usage data from the Marketplace informs the Observatory, which in turn provides evidence that can inform the development of the service portfolio offered by the infrastructures. By integrating these components, the project supports more coordinated access to research tools and services across communities that have often operated separately. This paper describes the architectural design and the functional implementation of the Marketplace and the Observatory, illustrating how they support a dynamic and collaborative SSH research environment.

The remainder of this paper is organised as follows: Section 2 contextualizes the project within the European framework of Research Infrastructure clustering and introduces the H2IOSC initiative. Section 3 describes the two main instruments devised for community data collection: the survey, implementing a quantitative approach, and the focus groups, embodying a qualitative approach. Section 4 describes the

Pietro Sichera, Monica Monachini, Valeria Quochi, Nicola Giampietro, Vittoria Fabiani, Roberta Bianca Luzietti, Roberta Ottaviani, Daniele Melaccio, and Laura Baggiani 2025. Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring Communities and Orchestrating Digital Services. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 229–241. DOI: https://doi.org/10.3384/ecp222

design and technical implementation of the H2IOSC Observatory. Section 5 presents the architecture of the Marketplace, highlighting its orchestration capabilities. Section 6 demonstrates the strategic integration between the Observatory and the Marketplace, illustrating the feedback loop behind the ecosystem. Finally, Section 7 draws conclusions and outlines future directions for the Marketplace and Observatory.

2 Context and Related Works: From Isolation to Federation

For at least the past two decades, the European Commission has heavily invested in Research Infrastructures (RIs) for the Humanities, Social Sciences, and Cultural Heritage. These investments have been driven by policy priorities increasingly aligned with Open Science and the FAIR principles, aiming to transform fragmented resources into a coherent scholarly ecosystem (European Commission: Directorate General for Research and Innovation, 2025). In the past, digital resources in the SSH domain were developed in isolation, leading to data silos. To address this, last-decade(s) EU strategies encouraged the construction and spread of distributed research infrastructures as a means for fostering both the data sharing and adherence to open science and FAIR principles also in the SSH disciplines, that were lagging behind life- and hard-sciences.

This resulted in a proliferation of infrastructural initiatives both at EU and at national levels, that now run the risk of limiting sharing and hampering interoperability. In line with this trajectory, recent EU strategies encourage the aggregation of RIs into "clusters", aiming to optimize resources, avoid duplication of efforts, and foster even more interdisciplinary collaboration. A notable and successful example of this is the European Open Science Cloud (EOSC), which promotes the reuse and sharing not only of data but also of tools, methodologies, protocols, and workflows (Almeida et al., 2017). Following this supranational direction, several European countries have begun to operationalise this clustering model at the national level. Notable examples of integrated national nodes include, among others: France, with the *HumaNum* infrastructure[1] which aggregates services for digital humanities; Germany: with the *Text+* consortium, part of the NFDI initiative; Switzerland with *SSHOC-CH*[2], which brings together the Swiss national nodes of CLARIN, DARIAH and OPERAS.

The Italian strategic response to this trend is the H2IOSC initiative[3], which aims to answer the need for a coordinated federation of infrastructures (Degl'Innocenti et al., 2023). It establishes a cross-disciplinary infrastructure cluster designed to support end-user support, data stewardship, resource curation, and training activities for the national scholarly community. Within this federated model, CLARIN-IT plays a pivotal role focusing on linguistic data management, standardization, and service integration. Its contribution ensures that the Italian ecosystem remains fully aligned with the European CLARIN standards and related EOSC initiatives. By integrating CLARIN-IT's specific capabilities with the broader H2IOSC architecture, the project aims to implement a "system of systems" where specialized vertical nodes contribute to a horizontal, interoperable marketplace.

Similarly, OPERAS-IT contributes the perspective and practices of open scholarly communication, covering areas such as peer review, open access publishing, and dissemination of research outputs. Within H2IOSC, OPERAS-IT is responsible for the development of the Marketplace and has carried out pilot implementations, while also coordinating the qualitative community monitoring activities described in Section 3.2. The Observatory itself is coordinated by CLARIN-IT under WP2, ensuring that the analytical framework remains aligned with CLARIN's metadata standards and community practices.

This integration is not unidirectional. While H2IOSC benefits from CLARIN ERIC's established standards and discovery services, it also contributes back to the broader ecosystem, for instance by cataloguing internationally maintained CLARIN services alongside nationally developed tools in a shared Marketplace, and by formalising implicit semantic distinctions from CLARIN's metadata practices into a reusable domain ontology (Melaccio et al., 2025). The relationship with CLARIN ERIC is further discussed in Section 7.

[1] https://www.huma-num.fr/
[2] https://sshoc.ch/
[3] https://www.h2iosc.cnr.it/

3 Mixed-Methods Approach to Community Monitoring

This section outlines the mixed-methods strategy adopted to investigate the needs, practices, and expectations of the communities served by the four H2IOSC Research Infrastructures. It combines a quantitative survey, designed to collect comparable evidence at federation level, with a qualitative focus-group campaign aimed at validating and enriching survey findings and capturing emergent themes. Together, these instruments provide the empirical basis for the Observatory's monitoring activities and inform the evolution of the service portfolio and workflows exposed through the H2IOSC Marketplace.

3.1 The quantitative approach

The surveying activity was jointly designed by the four H2IOSC Research Infrastructures (CLARIN-IT, DARIAH-IT, E-RIHS.it, and OPERAS-IT) to collect comparable evidence on communities, practices, and needs across partially overlapping domains. A single questionnaire was intentionally adopted to (i) reduce respondent burden and community fatigue, and (ii) enable an integrated analysis of requirements and expectations at federation level.

The questionnaire was collaboratively drafted to address both shared and RI-specific information needs, covering: respondent profiles, awareness and use of RIs, training needs, research outputs, and the description of digital resources, tools, and services used or produced by participants. Before wide release, the instrument underwent a two-step validation process: an initial pilot with a small group of researchers and professors representative of the target domains, followed by a second review with members of key disciplinary associations. Feedback from these phases informed a revised version of the questionnaire, structured into two parts (profile/needs and resources/services), which was then disseminated broadly through mailing lists, social media, RI and project websites, conference presentations, and outreach events (Luzietti et al., 2024, 2025; Monachini, Quochi, Luzietti, Moscati, et al., 2025; Monachini, Quochi, Luzietti, Spadi, et al., 2025).

To encourage participation while respecting privacy, personal data collection was minimized and limited to a contact email address used exclusively for follow-up communication and engagement activities.

3.2 The qualitative approach

The questionnaire distribution phase was followed by a qualitative mapping exercise to:

- validate the data collected through the questionnaire;
- identify new concepts not included in the questionnaire;
- foster a sense of belonging.

Four focus groups were conducted, coordinated by the OPERAS-IT node. Twenty-four researchers from the four communities were involved, including five archaeologists, four philologists, eight linguists, and seven philosophers. The meetings were video-recorded and automatically transcribed using Microsoft Teams. The researchers corrected the transcripts and processed them using the MAXQDA26 tool to produce a textual corpus of 59,724 tokens. The corpus then underwent a preliminary analysis to assess lexical richness and vocabulary diversity in the language of the researchers' involved. A sample of unique words was selected from the corpus (The sample was stratified and weighted according to the following variables: age, biological characteristics, the scientific fields of the four IRs). There are 930 word types in the corpus with a Type-Token Ratio (TTR) of 0.0156[4].

Knowledge of and use of research infrastructures (RIs) were proposed as discussion prompts in the focus groups. Most participants stated that they were familiar with RIs and used them frequently. To specifically examine the researchers' opinions, the corpus was analyzed a second time. Stop words, filler words, and idiomatic repetitions were removed. One-on-one exchanges between researchers and

[4]Type-Token Ratio (TTR) is a well-known quantitative measure used in linguistic disciplines to assess lexical richness. It is calculated by dividing the number of unique words (or word"Types") by the total number of their occurrences (or "Tokens") and multiplying by 100, Types /Tokens ×100. In MAXQDA this measure excludes stop words (common and highly frequent words such as *the*, *of* and *that*) to avoid noise during topical analysis of the data.

participants that were unrelated to the research topic were also eliminated. As a result, the reference corpus was reduced to 11,177 tokens, with a TTR of 0.0042. The word cloud in Fig. 1 was constructed from this reference corpus.

Figure 1: Word Cloud referring to the four focus groups

This image illustrates how researchers at the four RIs can exploit the opportunity presented by tools such as the H2IOSC infrastructure.

The tokens in the word cloud, whose absolute and relative frequencies are partially shown in Fig. 2, were converted into codes using the MAXQDA26 software.

Words	Frequencies	%
infrastructure	234	2,09
research	220	1,97
tools	121	1,08
fair	60	0,54
digital	57	0,51
open access	41	0,37
linguistics	37	0,33
publication	18	0,16
computational	17	0,15
Clarin	**13**	**0,12**
interoperability	8	0,07

Figure 2: Table of absolute frequencies and percentages of the words included in the analysis

The resulting codes were visualized in a graph showing the relationships between them, creating a “code map.” (Fig. 3, Area 1).

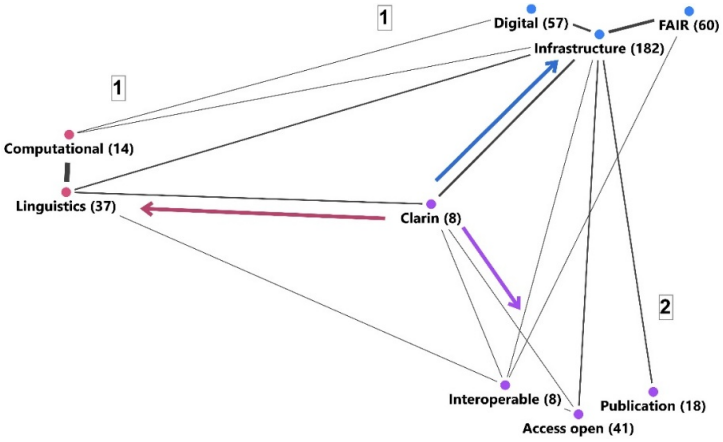

Figure 3: Concept map of stimuli on knowledge, use and requests for the CLARIN infrastructure

Using the "code map" tool, the relationships between all the codes and the "CLARIN" code were analyzed (Fig. 3). Regarding the linguistic domain, all eight researchers reported being aware of research infrastructures, and among these, CLARIN is the most widely used. Researchers associated the name CLARIN with the linguistic, computational, digital, infrastructure, and FAIR principles codes.

These code co-occurrences were then examined through MAXQDA's concept mapping functionality, which visualises proximity and frequency-based relationships between codes. The resulting map (Fig. 3, Area 2) revealed three convergent thematic clusters: a shared emphasis on resource availability and interoperability across RIs, a concern for the visibility and discoverability of research outputs, and a strong orientation towards Open Science compliance. These clusters were further validated by cross-referencing them with responses from the quantitative survey, confirming their representativeness across the broader community. The three expectations reported below correspond directly to these thematic clusters.

In addition, the expectations identified regarding the confederate infrastructure (Fig. 3, area 2) are:

- **Expansion of Digital Heritage:** H2IOSC should increase the availability of corpora and structured educational materials as resources that comply with FAIR principles. Linguists expect to have the resources of the four member research institutions immediately available in a single Marketplace.
- **Dissemination and Impact:** Leveraging the centralized infrastructure to enhance the visibility of linguistic research outputs, facilitating their discovery on a multidisciplinary scale. Users hope that H2IOSC will act as a catalyst for the outputs (publications) currently hosted in CLARIN.
- **Progress towards Open Science:** Integration into the H2IOSC Marketplace is seen as a critical opportunity to increase the CLARIN node's compliance with Open Science protocols.

The analysis of lexical relationships places CLARIN at the center (Fig. 3), highlighting that it is a digital linguistics infrastructure already aligned with open science and open access dissemination. The inclusion of linguists' needs in the H2IOSC Marketplace contributes to the development of a more inclusive research environment and may help CLARIN, as well as the other RIs, address interdisciplinary requirements more effectively. Our results suggest that infrastructure federation is not only a technical process, but also has important organizational and human dimensions. It can be informed by community feedback and, in turn, can inform methodological developments in related disciplines. In particular, the Observatory is the dedicated instrument through which this evidence-based community feedback, collected periodically through qualitative and quantitative methods, is gathered and returned to the communities.

4 The H2IOSC Observatory: Monitoring the Evolving Landscape of Digital Services

The H2IOSC Observatory is designed as a long-term instrument for monitoring the development and socio-economic impact of the four Research Infrastructures (RIs) of the H2IOSC federation. It provides

a continuous framework for analyzing the evolving landscape of digital resources, tools, and services in the domains of Linguistics, Digital Humanities, and Cultural Heritage, with particular attention to how these infrastructures evolve over time in relation to both community practices and infrastructural developments.

Originally designed as an internal instrument for monitoring communities and supporting evidence-based adjustments of services and priorities, the Observatory was redefined as a stable, publicly accessible online service. This evolution reflects its broader function as a shared environment for documenting researcher behaviors, practices, and needs across the H2IOSC communities. By displaying high-quality empirical data on researcher behaviors, practices, and needs, the Observatory establishes a data-driven feedback loop between community activities and the digital service environment of the H2IOSC ecosystem. The Observatory website is not intended as a space for active data collection, but rather as a curated and publicly accessible environment for the structured dissemination and interpretation of evidence produced within H2IOSC. This design choice is reflected in the organization of dedicated sections for different methodological instruments, such as surveys and focus groups, which are exposed through clearly separated thematic tabs (see Fig. 4). This structure supports transparency, reuse, and long-term comparability of results, while also facilitating navigation across heterogeneous forms of evidence.

Figure 4: Thematic sections of the H2IOSC Observatory dedicated to survey and focus-group evidence, exemplifying the platform's structured organisation of heterogeneous monitoring data.

4.1 Community-oriented monitoring

The Observatory operationalizes the mixed-methods approach described in Section 3, integrating quantitative and qualitative instruments such as surveys, focus groups, interviews, and landscape mapping activities (Tashakkori & Creswell, 2007). These instruments support the systematic investigation of the behaviors, needs, and practices of current and prospective RI users, providing an empirical basis for the analysis of both declared expectations and emerging research dynamics.

Alongside its analytical function, the Observatory also supports community engagement and participatory monitoring. By involving researchers, institutional stakeholders, and prospective users in the assessment and discussion of services, priorities, and gaps, it supports the co-definition of offerings, and

the development of infrastructures that respond to documented user needs and practices. In this sense, the Observatory bridges community-oriented analysis and infrastructural decision-making, helping to keep service development grounded in empirical evidence (Luzietti et al., 2024).

4.2 Platform implementation and content organisation

From an implementation perspective, the Observatory has been developed as a modular web platform, fully aligned with the approved UX/UI design and with the shared visual and functional ecosystem of H2IOSC infrastructures. The platform adopts a mobile-first approach and devotes specific attention to accessibility and usability requirements, in line with WCAG 2.1 principles[5], ensuring inclusive access and compliance with public-sector standards.

The organisation of contents reflects the core instruments of the Observatory and translates methodological choices into navigable sections. These include survey results and analytical summaries, qualitative outcomes of focus groups and interviews, and landscape mapping activities outlining available tools, datasets, and services. In addition, monitored projects are integrated through the Digital Platform for Cultural Heritage Research (DHeLO) platform via dedicated APIs, enabling the Observatory to connect empirical monitoring with a broader representation of infrastructural initiatives across the federation.

Selected usage indicators originating from the Marketplace are incorporated into the Observatory through dedicated APIs, complementing survey-based evidence with behavioural traces from real interactions with the service catalogue (see also Section 6).

4.3 Role within the H2IOSC ecosystem

The long-term sustainability of the Observatory is ensured through both organisational and technical measures. The Observatory Scientific Committee is responsible for guaranteeing methodological rigour, analytical depth, and the quality of published contributions, acting as a supervisory and theoretical reference point in alignment with Open Science principles (Rossi et al., 2022). At the same time, the platform includes a dedicated back-office environment based on a modular Content Management System and a Digital Asset Management component, enabling structured editorial updates and progressive enrichment of contents without continuous technical intervention.

Within the broader H2IOSC ecosystem, the Observatory operates as both a reflective research environment and a strategic infrastructural component. By consolidating heterogeneous evidence into a coherent monitoring framework, it supports adaptive governance of the federation and prepares the ground for the service orchestration mechanisms implemented in the H2IOSC Marketplace. In this perspective, the Observatory provides the interpretative and analytical layer of the ecosystem, while the Marketplace constitutes its operational interface, together enabling an evidence-based and continuously evolving research infrastructure.

5 The H2IOSC Marketplace as an Operational Infrastructure and Source of Evidence

The H2IOSC Marketplace is conceived as a platform for the consultation and management of resources and services exposed by the Italian nodes of the federated infrastructures, clearly separating functionalities aimed at end users from those devoted to governance and curation. It combines controlled metadata ingestion, semantic normalisation, indexing, and presentation, and is designed for interoperability with external sources and aggregators.

5.1 Three-layer architecture

From an architectural standpoint, the Marketplace adopts a three-component structure: Back-End, Back-Office, and Front-End. The Back-End constitutes the application core and concentrates the exposure of APIs; the Back-Office is the operational environment for authorized users (administrators, curators, editors), responsible for curatorial management and configuration; the Front-End is the public interface for consultation, search, and catalog navigation. This separation is not an implementation detail, but an infrastructural constraint: it allows governance functions to be isolated from user-facing access paths and,

[5] https://www.w3.org/TR/WCAG21/

above all, enables programmatic access to Marketplace data through APIs without depending on the UI. The technical design includes infrastructural components dedicated to persistence and search, employing technologies such as SOLR, Redis, and MongoDB, as well as integration components to interact with external systems.

5.2 Identity management

Identity and authentication management is based on Keycloak. This aspect is relevant for the relationship with the Observatory for two reasons: on the one hand, it enables the separation of public access for consultation from the use of services that require credentials; on the other hand, it creates the technical conditions to distinguish types of interaction and use (anonymous vs. authenticated) in compliance with access policies.

5.3 A data model as the basis for observability

The descriptive uniformity of the catalogue is ensured by a data model organised around the *Item* entity, intended as the main unit of cataloguing for heterogeneous types (for example, tools, datasets, publications, projects, and services). Items are described through a structured set of fields, properties, and relations; the mandatory status and formats of fields are defined to ensure consistency and interoperability. This choice has a direct consequence for the relationship with the Observatory: it makes both the composition of the catalogue and its temporal evolution observable and comparable in a systematic way, because variability in descriptions is constrained by schemas and controlled vocabularies. In particular, lexical and semantic normalisation through managed dictionaries and vocabularies enables reliable metrics (e.g., by Item type, disciplinary area, or access mode) that the Observatory can reuse directly.

5.4 Feeding the Marketplace

The feeding of the Marketplace is designed to combine automated import and controlled management, with the aim of maintaining alignment and traceability with respect to the original sources. For the acquisition and integration of metadata and Items, mechanisms based on the OAI-PMH protocol and on REST APIs are envisaged, with a dedicated endpoint `/api/items/harvest` and `Bearer`-type authentication, and with a data schema aligned with the OAI-PMH model. Beyond enabling ingestion, this approach defines a technical contract that makes it possible to distinguish what is curated locally from what is synchronized with external providers.

On the Back-Office side, harvesting mechanisms are operationalized through a harvesting console that allows users to create, modify, and schedule jobs, monitor their executions, and download outcome reports. The orientation is explicitly curatorial: the Marketplace does not merely "collect", but provides tools to govern the ingestion process, including filters, sorting, and monitoring of the status of jobs and processed items.

In parallel with harvesting, the design envisages providing APIs to export Items towards other marketplaces or aggregators, through token-authorized access and an OpenAPI description. Providing is therefore part of the Marketplace's federated positioning: not only ingestion, but also the ability to "expose" contents in exchange formats and to support scenarios in which third-party systems select and reuse the catalog contents.

5.5 Descriptive workflows and executable workflows

A qualifying element of the H2IOSC Marketplace is workflow management as a orchestration of services. The implementation includes descriptive workflows, which represent and document how different services can be combined, and executable workflows (Sichera et al., 2024), in which the orchestration is designed and guided at run-time, subject to compatibility constraints between output/input pairs and adherence to specifications such as OpenAPI 3. In this setting, the Marketplace does not merely describe resources, but encodes operational relationships among services that may also come from different sources.

In the technical design, the construction and management of executable workflows and integration is associated with the use of WSO2 API Manager (for API management and exposure) and WSO2 Micro

Integrator (for orchestration), coupled with the management of manifest files for service description.

At the time of writing, the harvesting pipeline for CLARIN-IT services has been fully developed and validated. The Marketplace catalog already includes, or will include at launch, a range of services provided by CLARIN-IT within H2IOSC — such as an eScriptorium instance for handwritten text recognition, a deposit service, and a training platform — alongside internationally maintained CLARIN services such as UDPipe and resources such as ParlaMint. This combination of national and international assets illustrates the federated nature of the catalog and its alignment with the broader CLARIN ecosystem. An example of cross-infrastructure orchestration is provided in Section 6.4.

6 The Interaction between Marketplace and Observatory as an Infrastructural Mechanism

The relationship between the Marketplace and the Observatory, in the H2IOSC design, is explicitly conceived as a continuous synergy between two complementary nodes of the same federated research infrastructure. The Marketplace aggregates and exposes digital assets and services from the participating infrastructures, while the Observatory analyses the use and evolution of such resources over time, with the aim of supporting adaptive management of the ecosystem and aligning the service offer with emerging needs of the community. This interaction is not incidental, but intentionally designed as an infrastructural mechanism in which observation, interpretation, and service provisioning are functionally decoupled yet operationally interconnected.

6.1 *Ex ante* and ongoing evidence of use

The Observatory inherits and systematises evidence built with replicable mixed-methods instruments (questionnaires, interviews, focus groups, and semi-automated collection protocols), producing a knowledge base on the needs and practices of the reference communities. The opening and use of the Marketplace add a further dimension: the possibility of correlating what users declare or discuss with what they actually explore and use in the catalogue, making expectations and observed behaviour comparable. In this perspective, the Marketplace becomes a generator of traces and metrics that transform observation from a periodic snapshot into a continuous process. This ongoing monitoring is supported by usage statistics and interaction indicators, which provide continuous evidence of service uptake and navigation dynamics within the platform (see Fig. 5). For instance, the Observatory dashboard currently displays the distribution of catalogue Items by type, the temporal evolution of published resources, registered user counts over time, and saved search activity, all drawn from the Marketplace via dedicated APIs. These indicators, though preliminary, already enable the identification of trends in community engagement and catalogue growth. In this sense, the Marketplace enables a form of infrastructural observability, whereby user interactions, navigation paths, and service combinations can be analysed as structured signals rather than isolated events.

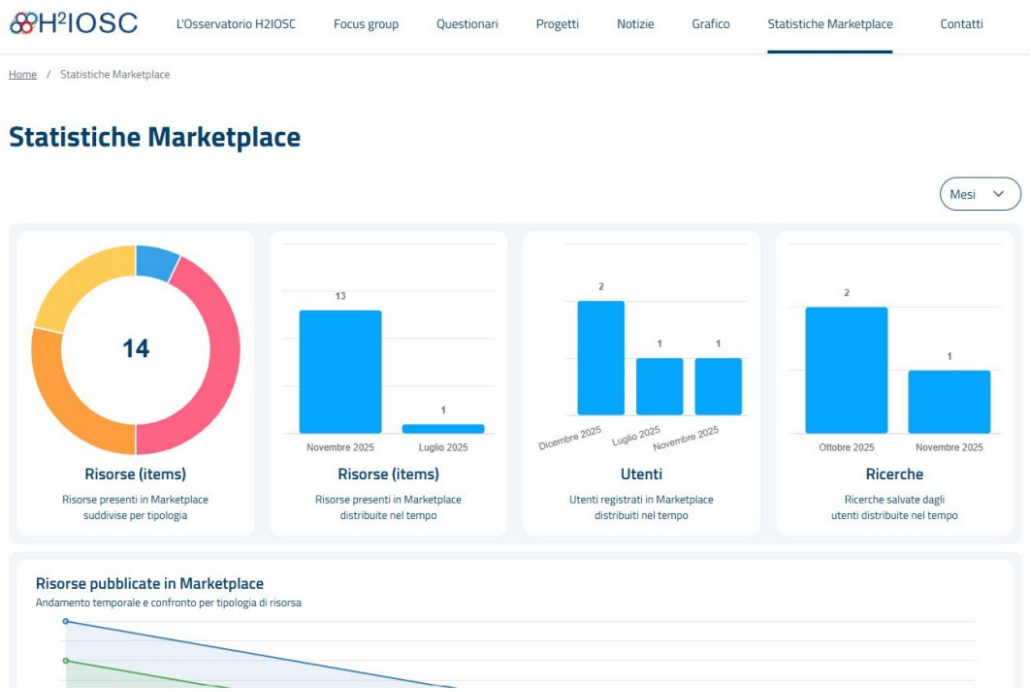

Figure 5: Usage statistics from the H2IOSC Marketplace as presented in the Observatory dashboard for monitoring catalogue growth and user engagement.

6.2 Service provision and analysis

The Observatory can be considered a modular web application developed "in synergy" with the Marketplace, with a bidirectional flow of information: the Marketplace aggregates and exposes resources, the Observatory analyses their use and evolution and returns insights that guide subsequent developments.

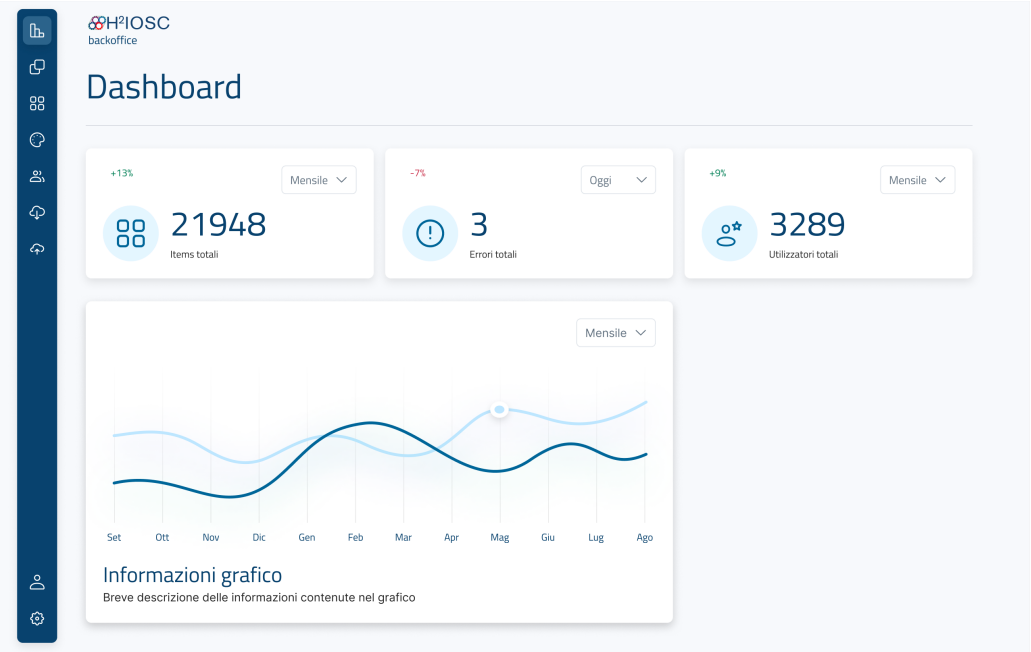

Figure 6: Analytical dashboard based on Marketplace data, used to observe user behaviour, resource uptake, and patterns of interaction across the platform.

The integration of the two platforms aims to enable monitoring of resource ingestion, user behavior, and service uptake. This defines a minimum set of requirements for interoperability between the two

platforms: availability of structured data on the catalog (composition and changes), availability of usage metrics (with adequate granularity and access rules), and semantic consistency to enable meaningful aggregations. The "Item-centric" model and the controlled vocabularies of the Marketplace thus become a technical prerequisite for observational analysis, because they stabilize the dimensions along which to measure and compare. This separation of concerns is technically supported by stable data models, controlled vocabularies, and API-based exchanges, which allow analytical components to evolve independently from service provisioning layers.

6.3 Catalogue and process data

Observational interaction concerns not only which Items exist in the catalog, but also how they reach it and how they are maintained. Harvesting and providing mechanisms, as well as the management of ingestion jobs and operational reports, produce process data that can be interpreted as indicators of the maturity and reliability of the catalog feeding chains. In particular, the ability to distinguish between functionalities implemented and verifiable in staging and functionalities implemented but not verifiable end-to-end due to external prerequisites helps to avoid improper readings of the metrics, and suggests that the Observatory should incorporate metadata about the "state" of the platform and of the flows, in addition to usage metrics. Such information is particularly relevant for governance purposes, as it enables infrastructure managers to distinguish between structural limitations, integration bottlenecks, and actual patterns of community uptake. As a result, observability is not retrofitted onto the system, but emerges from design choices that constrain variability and make catalog evolution analytically tractable.

6.4 From observational feedback to action

The operational value of the Observatory–Marketplace cycle lies in transforming insights into verifiable actions. A first trajectory concerns the evolution of the catalog: evidence on emerging needs and domains can guide scouting and integration activities through harvesting from external sources and through curatorial processes in the Back-Office. The Marketplace design explicitly requires full interoperability with European catalogs (including EOSC and SSHOC), which, within the observational cycle, can also be read as a strategy to reduce the time between identifying a need and making relevant resources available, by exploiting federation and exchange channels.

A concrete example of how community evidence informs service design is provided by the workflow orchestration capabilities of the Marketplace. Focus group findings (Section 3.2) highlighted that researchers across disciplines regard interoperability and the ability to combine services from different infrastructures as a key expectation. This evidence directly informed the development of executable cross-infrastructure workflows. For instance, a workflow has been designed that takes a IIIF manifest served by E-RIHS.it as input, sends it to the CLARIN-IT eScriptorium instance for handwritten text recognition, and automatically imports the resulting transcription into a TEI Publisher instance for publication and annotation. This pipeline illustrates how the Marketplace translates declared community needs into operational service compositions that span multiple federated infrastructures.

A second trajectory concerns the quality of the presented information: the Back-Office includes a Content Management System that enables the management of editorial pages and the creation of content (simple pages, collections, blog posts), complementing the technical description of Items with layers of contextualisation and usage-oriented guidance. In a data-driven cycle, the Observatory can help identify areas of the website and catalogue that require greater clarity or informational completeness, intervening not only in the addition of new Items but also in their presentation and in the construction of editorial paths consistent with observed usage practices. In this way, observational evidence is translated into concrete curatorial, interoperability, and editorial decisions, closing the loop between monitoring, interpretation, and infrastructural action. Taken together, the Observatory–Marketplace interaction exemplifies how analytical monitoring and service orchestration can be combined into a single adaptive mechanism within federated research infrastructures.

7 Conclusions and Future Work

In the near future, the Observatory will be further developed by cyclically running the defined monitoring strategies as well as extending current data collection methods with additional monitoring strategies and impact assessment procedures. The aim is to improve the identification of changes in community needs and service usage patterns, in order to inform decisions about the extension of available tools and possible technical developments.

In this context, requests for additional data from the Marketplace, such as performance indicators or engagement metrics, may offer useful insights for refining both the Observatory's analytical methods and its overall development.

At the same time, the Marketplace may expand its functionality and interoperability on the basis of feedback emerging from the Observatory's activities. This may include the introduction of new modules for orchestrating complex workflows and procedures for integrating third-party services and external infrastructures. The effectiveness of these enhancements will depend on continued coordination between the two components: the Observatory will provide ongoing monitoring and identify areas for improvement; the Marketplace will supply the data and resources required to support incremental, modular updates and alignment with user needs.

In this context, the partnership between infrastructures will continue to develop through iterative exchanges between the Observatory and the Marketplace. This interaction supports an integrated Open Science ecosystem, where coordination between research infrastructures contributes to the development of tools, broadens user engagement and increases the use of available services, thereby improving the uptake of open science digital approaches in humanities, social and cultural heritage sciences.

In addition, the work carried out within H2IOSC has also contributed back to the CLARIN ecosystem. The ontology-based mediation approach developed for aligning CLARIN-IT metadata with the H2IOSC Marketplace enables interoperability between heterogeneous schemas while preserving semantic consistency. This approach represents a concrete methodological contribution that can be reused by other CLARIN national nodes facing similar integration challenges, and may support the broader harmonisation of services and metadata within CLARIN ERIC and EOSC-related environments (Melaccio et al., 2025).

Finally, H2IOSC is conceived in continuity with the EOSC vision, as a federation of Italian nodes of European ESFRI RIs already involved in EOSC. CNR is a member of the EOSC Association, and several H2IOSC team members have participated in EOSC-related and SSHOC initiatives. H2IOSC draws on these experiences. In particular, the Marketplace and the Observatory are informed by the EOSC and SSH Open Marketplace, while adapting their functions to national and domain-specific requirements. The Observatory is intended to complement the EOSC Observatory by providing more detailed, community-driven information on research practices and service usage at the community level. These developments point to a layered model of integration, in which national infrastructures contribute specialized components that can be connected and may enrich their European counterparts. In this perspective, CLARIN-IT within H2IOSC remains aligned with the mission of CLARIN ERIC, while OPERAS-IT continues its participation and strengthens its role within OPERAS, ensuring that both infrastructures contribute to and benefit from broader open science initiatives at national and European levels.

Acknowledgments

This work is supported by the H2IOSC Project - Humanities and Cultural Heritage Italian Open Science Cloud funded by the European Union – NextGenerationEU – NRRP M4C2 - Project code IR0000029 - https://www.h2iosc.cnr.it/

References

Almeida, A. V. d., Borges, M. M., & Roque, L. (2017). The european open science cloud: A new challenge for europe. *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*. https://doi.org/10.1145/3144826.3145382

Degl'Innocenti, E., Monachini, M., Bucciero, A., Pasini, E., Fanini, B., & Frontini, F. (2023). H2iosc: Humanities and heritage open science cloud. *La Memoria Digitale: Forme Del Testo e Organizzazione Della Conoscenza. Atti Del XII Convegno Annuale AIUCD*, 63–64.

European Commission: Directorate General for Research and Innovation. (2025). The european strategy on research and technology infrastructures [Accessed: 2025-10-22].

Luzietti, R. B., Spadi, A., Giampietro, N., Giacomo, M., Caravale, A., D'Eredità, A., Caradonna, M., Moscati, P., Quochi, V., Monachini, M., & Degl'Innocenti, E. (2024). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an open science cloud. In A. Di Silvestro & D. Spampinato (Eds.), *Me.te. digitali. mediterraneo in rete tra testi e contesti, proceedings del xiii convegno annuale aiucd*. AIUCD Associazione per l'Informatica Umanistica e la Cultura Digitale.

Luzietti, R. B., Spadi, A., Giampietro, N., Mancuso, G., Caravale, A., D'Eredità, A., Caradonna, M., Moscati, P., Quochi, V., Monachini, M., & et al. (2025). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an open science cloud. *Umanistica Digitale*, *9*(20), 419–439. https://doi.org/10.6092/issn.2532-8816/21226

Melaccio, D., Boschetti, F., & Monachini, M. (2025). Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation. *Proceedings of the CLARIN Annual Conference 2025*. https://doi.org/10.5281/zenodo.17357825

Monachini, M., Quochi, V., Luzietti, R. B., Moscati, P., Caravale, A., Chirivì, A., Degl'Innocenti, E., Spadi, A., & Caradonna, M. (2025, January). *D2.1 first report on the H2IOSC landscapes* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.14680021

Monachini, M., Quochi, V., Luzietti, R. B., Spadi, A., Mancuso, G., D'Eredità, A., Caravale, A., Giampietro, N., Caradonna, M., & Melaccio, D. (2025, October). *D2.2 updated report on the H2IOSC landscapes* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.17737101

Rossi, G., Caso, R., Castelli, D., & Giglia, E. (2022). National plan for open science.

Sichera, P., Marras, C., & Pasini, E. (2024, November). Orchestrazione api per workflow applicativi nell'ambito delle digital humanities. https://doi.org/10.5281/zenodo.14187534

Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. *Journal of mixed methods research*, *1*(1), 3–7.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

UPSKILLS Two Years on: Teaching about Language Resources and FAIR Data Principles with CLARIN

Iulianna van der Lek
CLARIN ERIC
Utrecht, The Netherlands
i.vanderlek@uu.nl

Maja Miličević Petrović
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
maja.milicevic2@unibo.it

Silvia Bernardini
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
silvia.bernardini@unibo.it

Adriano Ferraresi
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
adriano.ferraresi@unibo.it

Olga Arsić
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
olga.arsic2@unibo.it

Abstract

The ongoing technological developments have a major impact on language-related research and professional activities. This highlights an increasing demand for new skills and profiles, as well as a greater focus on making language data and resources open and FAIR. In this paper, we present long-term efforts that address these issues, starting from the Erasmus+ project UPSKILLS (2020–2023). We specifically focus on follow-up activities at the University of Bologna in collaboration with CLARIN ERIC, proposing a conceptual framework that combines progressive competence development across all study levels with infrastructure-based pedagogy to support discipline-specific Open Science education. Using CLARIN as a learning environment, lecturers can modernize curricula, improve students' FAIR data literacy, and improve graduate employability.

1 Introduction

The rapid advances in language technologies require adaptation not only on the part of individual researchers and language professionals, but also through initiatives of higher education institutions offering language- and linguistics-related degree courses. One of the recent projects with related objectives was the Erasmus+ strategic partnerships UPSKILLS: *UPgrading the SKIlls of Linguistics and Language Students*.[1] This three-year project (2020–2023) sought to identify and address the gaps and mismatches in the skills of linguistics and language students with respect to the demands of the current job market and the broader research landscape. The project's results received the Erasmus+ good practice label[2] and were ranked as the second runner-up in the 2023 competition for the best Digital Humanities training materials.[3] Two years later, we revisit the results by focusing specifically on the impact the project has had on one of the consortium partners, the Department of Interpreting and Translation (DIT) of the University of Bologna (UNIBO). We expand on the original UPSKILLS project results by exploring in particular how the CLARIN research infrastructure can be integrated into teaching practices to raise awareness among students about the basic principles of FAIR research data management, as well as the vast body of knowledge, expertise, data and tools available in open access.

[1] https://upskillsproject.eu

[2] https://erasmus-plus.ec.europa.eu/projects/search/details/2020-1-MT01-KA203-074246

[3] http://dhawards.org/dhawards2023/results/

Iulianna van der Lek, Maja Miličević Petrović, Silvia Bernardini, Adriano Ferraresi, and Olga Arsić 2025. UPSKILLS Two Years on: Teaching about Language Resources and FAIR Data Principles with CLARIN. *Selected papers from the CLARIN Annual Conference 2025. Ed. by Cristina Grisot and Thalassia Kontino.* Linköping Electronic Conference Proceedings 222, pp. 242–254. DOI: https://doi.org/10.3384/ecp222

The paper is structured as follows. Section 2 presents the UPSKILLS project, outlining its main results, the spin-off materials hosted at UNIBO, as well as the changes that DIT implemented in its study programmes. Section 3 describes the CLARIN-UNIBO collaboration and the topics covered in the CLARIN guest lectures at DIT. Section 4 presents the conceptual framework and a matrix of proposed learning outcomes and potential activities using CLARIN's research infrastructures, which could be embedded in any language-related research programme. Section 5 concludes with reflections on how to ensure the long-term impact of the project and a call to the CLARIN community to help improve the proposed framework.

2 From UPSKILLS to UpskilLing@UNIBO

2.1 A summary of UPSKILLS outputs

In the UPSKILLS project, researchers from eight partner institutions across Europe worked toward developing a new curriculum component and a new teaching approach targeting the skills needed for industry and academic research among (primarily) undergraduate students of language-related subjects. The deliverables comprise: (1) a needs analysis, based on several components and summarized through a new target profile of *language data and project specialist*, skilled in the disciplinary domain, but also in technology, research and data and project management (Miličević Petrović et al., 2021), (2) a series of guidelines and best practices on research-based teaching and the use of research infrastructures (Simonović et al., 2023; van der Lek et al., 2023), (3) a set of Moodle-based materials for online and blended teaching and learning in the areas identified as in need of improvement (Gledić et al., 2023),[4] and (4) educational games developed for, or adapted within the project (Arsić et al., 2023).[5]

UNIBO and CLARIN collaborated on the development of guidelines and teaching materials, which focused on language resources, namely the *Processing Text and Corpora* course (Bernardini et al., 2023)[6], the course *Introduction to language data: Standards and repositories* (van der Lek & Fišer, 2023)[7] and the *Guidelines for Integrating Research Infrastructures into Teaching* (van der Lek et al., 2023). While parts of the materials produced in the project, included those by UNIBO and CLARIN, were piloted at its end, during the UPSKILLS Summer School in August 2023,[8] more extensive inclusion of the content into UNIBO curricula took place in the years that followed.

2.2 Reforming the UNIBO curriculum

Towards the end of the UPSKILLS project and upon its conclusion, the UNIBO team developed additional materials, partly adapted to meet the needs of the local target community, students of intercultural communication, translation and interpreting in Italy. The needs analysis was complemented by a survey that gives students an initial indication of the UPSKILLS subprofile they might wish to pursue, i.e. language data analyst, language data scientist, language data manager, language project manager, depending on their inclination towards data analysis versus organizational tasks. Furthermore, the original UPSKILLS corpus of job advertisements, a central component in the needs analysis, was accompanied by an updated 2023 edition (with a 2026 edition in preparation), providing more up-to-date information, especially on AI-related job positions.[9] The resources converged in a website that helps students discover a suitable profile and leads them to relevant learning materials.[10]

At the same time, DIT started partially reforming its degree programmes in line with the UPSKILLS findings. At the MA level, the degree course in Specialized Translation gained a new curriculum in Trans-

[4] All courses can be viewed directly on Moodle (upon free registration) at https://upskills.fil.bg.ac.rs, and they can be downloaded from the CLARIN.SI repository at http://hdl.handle.net/11356/1865. Descriptions of individual courses, with access and download links, can be found at https://upskillsproject.eu/deliverables/io3/upskills_learning_materials/

[5] https://upskillsproject.eu/deliverables/io4/

[6] https://upskillsproject.eu/project/text_processing/

[7] https://upskillsproject.eu/deliverables/io2/research-infrastructures/, available in HTML format at https://www.clarin.eu/content/introduction-language-data-standards-and-repositories-0; see also the Italian translation and adaptation by van der Lek et al., 2024 at https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it.

[8] https://upskillsproject.eu/events/summer-school/

[9] https://bellatrix.sslmit.unibo.it/noske/upskills/#open

[10] https://upskilling.dipintra.it

lation and Technology, heavily focused on research skills, and incorporating courses such as *Profession-Based Research*, *Machine Translation*, *Natural Language Processing*, and *Language Data Analysis*. Later on, a module called *Language, Technology, Research* was introduced as an option in the BA course in *Languages and Technologies for Intercultural Communication*, with courses in *Computational Thinking* (I year), *Text Processing* (II year) and *Researching Language* (III year). Finally, the principles of research-based teaching have been implemented, most notably in the MA course in *Corpus Linguistics*, jointly with an evaluation approach based on peer assessment.

In addition, a MOOC titled *Processing Texts and Corpora: An Introduction* (Bernardini et al., 2025) has recently been created and made available on the UNIBO Open Knowledge platform,[11] as part of the project *Edvance – Digital Education Hub per la Cultura Digitale Avanzata*.[12] This course revisits a number of topics addressed in its UPSKILLS predecessor *Processing Texts and Corpora*, elaborating on CLARIN-related topics. Its unit "Corpus sourcing" focuses on CLARIN ERIC: it explains how to locate corpora via CLARIN's Virtual Language Observatory (VLO)[13] and how to select tools via the CLARIN Switchboard,[14] and helps learners assess suitability for reuse based on metadata, licensing, and documentation.[15] The original UPSKILLS outputs and later additions are both shown in Figure 1.

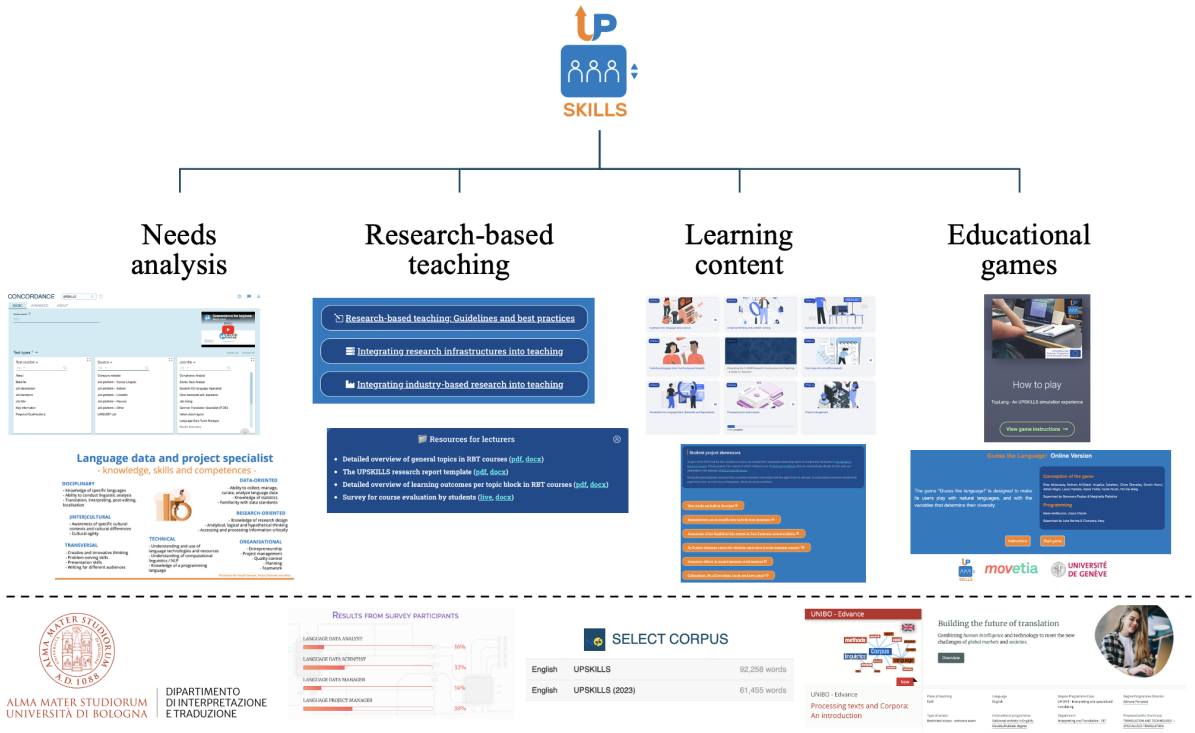

Figure 1: The UPSKILLS and additional UNIBO outputs.

3 Embedding FAIR competences through CLARIN

3.1 Focus on FAIRness

Open Science, defined as "an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone" (UNESCO, 2021, p. 7), includes as one of its cornerstones the **FAIR data principles** (*Findability, Accessibility, Interoperability, and Reusability*). As stated by Wilkinson et al. (2016), these foundational principles apply to data in the conventional sense, as well as the tools and workflows that led to that data.

[11] https://book.unibo.it/
[12] https://www.edvance.it/
[13] https://vlo.clarin.eu/?0
[14] https://switchboard.clarin.eu/
[15] For another Italian experience, see Pedonese et al. (2025) on the use of UPSKILLS materials at the University of Ferrara.

Although Open Science and FAIR data principles are increasingly embedded in research policy and funding requirements, their integration into university curricula remains largely fragmented. Several initiatives[16] acknowledged a lack of awareness among lecturers, students and researchers of research data management (RDM) practices, including knowledge of principles to make digital resources FAIR. Consequently, the *FAIR Competence Framework for Higher Education* (Demchenko et al., 2021) proposed a set of core competencies for FAIR data education that universities can use to integrate relevant content into their curricula. Students, teachers, and researchers from all disciplines are encouraged to acquire fundamental skills for Open Science, including the ability to effectively interact with **federated research infrastructures** and Open Science tools for collaborative research. Furthermore, research on RDM education consistently highlights the importance of embedding data skills within disciplinary teaching rather than offering generic, de-contextualized training (Cox et al., 2017). Effective approaches emphasize alignment with research data lifecycle, use of authentic data and tools, and progression from awareness to independent practice.

Since Open Science originated largely in the natural sciences, its application to the humanities has required adaptation to disciplinary epistemologies, data types and ethical concerns (Borgman, 2015). Data in the humanities are often heterogeneous, multilingual, and context-dependent, raising challenges for standardization and reuse (Edmond, 2020). In language-related disciplines specifically (Translation Studies, Corpus Linguistics, Language Technology, and others), researchers and students routinely work with complex language resources such as corpora, parallel texts, linguistically or otherwise annotated data, or terminology databases, where FAIR principles have proven highly relevant, and metadata quality, documentation, and interoperability directly affect reuse and reproducibility (Krauwer & Hinrichs, 2014). Yet, explicit training in research data management, licensing, metadata and data publication is often limited in this area.

Research infrastructures such as CLARIN are commonly framed as enablers of advanced research, but work in digital humanities and NLP pedagogy argues that infrastructures can also function as learning environments (De Smedt et al., 2024). Exposing learners to real-world data, standards, repositories, and infrastructures, supports experiential learning and fosters data literacy grounded in disciplinary practice. This paper proposes a progressive, infrastructure-based teaching framework for Open Science, FAIR data and research data management in language-related disciplines based on the materials CLARIN produced in the UPSKILLS project, namely the course *Introduction to Language Data: Standards and Repositories* (van der Lek & Fišer, 2023) and *Best Practice Guidelines for Integrating Research Infrastructures into Teaching* (van der Lek et al., 2023). The materials demonstrate how instructors can use the CLARIN open data repositories, language resources, and integrated NLP tools to promote the added value of Open Science and guide students in applying FAIR data principles in language research.[17]

3.2 Lectures on CLARIN resources at UNIBO

Throughout the academic year 2024/2025, CLARIN-related topics were introduced in the DIT curriculum both as in-class guest lectures and embedded in an e-learning course. In March 2025, a total of three 90-minute guest lectures were delivered in two courses within DIT's degree programmes at UNIBO, with the objective of familiarizing the students with the CLARIN research infrastructure and raising awareness about basic concepts of Open Science, FAIR data principles and research data management. After the guest lectures, the materials were adapted for the MOOC *Processing Texts and Corpora: An Introduction* (Bernardini et al., 2025). The content of all courses was based on the materials CLARIN developed in UPSKILLS (van der Lek & Fišer, 2023).

The undergraduate course in *Language, Technology, Research II: Text Processing* aimed to introduce BA-level students to basic text processing methods and teach them how to design and build compara-

[16] In addition to the UPSKILLS projects, especially the EOSC Working Group on Skills & Training (https://eosc.eu/opportunity-area-exp/oa5-skills-training-rewards-recognition-upscaling/) and their publication *Digital Skills for FAIR and Open Science* (https://data.europa.eu/doi/10.2777/59065).

[17] These materials were built on the handbook *How to be FAIR with your data* (Engelhardt et al., 2022), comprising ready-made lesson plans and guidelines for teachers, and *The Open Handbook of Linguistics Data Management* (Berez-Kroeker et al., 2022), a useful toolkit for linguists.

ble and parallel corpora. After being introduced to character encodings, file formats, and principles of corpus design and construction, the students had to build corpora of various kinds: scientific articles, tasting notes on beverages other than wine, and manuals of household appliances. Before the students started with data collection, they were introduced to data discovery through two sessions organized by a CLARIN guest lecturer who showed how to locate already digitized text collections available in public repositories, at the same time gradually introducing the students to basic concepts in linguistic data management and FAIR principles. The sessions were attended by 10 students with different language combinations, and three lecturers with a background in translation and interpreting studies. The topics described below were covered in two sessions organized at the undergraduate level, and one at MA level.

3.2.1 A gentle introduction to project and research data management

The opening session introduced shared best practices in project and research data management, with a particular focus on teamwork for collecting and building resources such as corpora. As students were working in pairs to collect texts from the web to build their corpora, some user-friendly cloud-based tools were shown for collaboration and task management, e.g. Trello.[18] Additionally, the concept of data management plan (DMP)[19] was introduced to show how it can be used to plan for each step in the project: data collection, documentation and metadata, ethics and legal compliance, storage and backup, data sharing and how to split responsibility and resources during the project.

This approach was adopted because, while students typically gain practical project management experience only during internships, university lectures can also introduce good project and (research) data management practices as early as the undergraduate level. This early academic exposure, which often includes joint work when collecting and building resources such as corpora, allows students to refine essential soft skills in planning and managing complex data workflows throughout their degree, workflows essential for the professional profile of the "language data and project specialist".

3.2.2 Locating linguistic data in public repositories

UNIBO students also learned how to find research data, starting from their institutional repository, and using registries such as re3data (REgistry of REsearch Data REpositories) to locate domain-specific repositories used in linguistics, such as CLARIN, TROLLing (Tromsø Repository of Language and Linguistics), and OLAC (Open Open Language Archives Community).[20]

Specifically, the CLARIN representative demonstrated how to discover language data in CLARIN's Virtual Language Observatory and Resource Families. By browsing the VLO, a search portal containing over a million records, students learned that CLARIN helps researchers make their published language resources FAIR, e.g. the resources are described with consistent metadata (through CMDI – Component Metadata Infrastructure) and are assigned a persistent identifier (a Handle) ensuring that the resource can be reliably retrieved and cited in their work even if the physical location of the data changes. After locating a suitable digital text collection they could potentially download and use to build their own corpus, the students first needed to review the metadata to identify whether the resource allowed public, academic or restricted access. In the case of restricted access, students got to test the CLARIN federated login system, i.e. the single-sign on,[21] which allowed them to log in to the repository hosting the respective resource with their institutional credentials.

3.2.3 FAIR assessment of corpora

The students learned about the FAIR data principles through a hands-on activity that involved evaluating the available corpora from a FAIR perspective. They were split into groups and asked to pick a corpus from a proposed list and assess its FAIRness based on a checklist inspired by Frey et al. (2019). The students went through each stage of FAIR, checking whether the corpus and its metadata were easily accessible for use, whether the corpus was assigned a unique and persistent identifier (PID), whether

[18] https://trello.com

[19] https://www.openaire.eu/what-is-a-data-management-plan

[20] https://www.re3data.org, https://site.uit.no/trolling/, http://www.language-archives.org/

[21] https://www.clarin.eu/content/identity-provider-technical-details

it was developed using common standards and formats, was properly documented, and was assigned a clear and accessible license usage (e.g., a Creative Commons License).

3.2.4 Corpora depositing and sharing

To make students aware of the depositing workflow, a brief demonstration of the Italian ILC4CLARIN repository[22] was given. The repository is hosted at the Institute for Computational Linguistics of the National Research Council in Pisa. Scholars in Italy can use the repository to search for linguistic data and tools, as well as deposit their data to safely store them, ensuring that it is also findable and properly credited if reused by other scholars in research or teaching. To be able to deposit data, the data owner is prompted to log in using their institutional credentials. Once logged in, the depositing workflow is relatively straightforward. For example, corpus data needs to be provided with metadata in standard formats recommended by the CLARIN community[23]. The depositor needs to fill in all the required metadata, after which the submission is curated by the repository manager. Once published, the corpus metadata is freely accessible not only via the Italian repository but also via the Virtual Language Observatory.

Although a formal evaluation of the activities was not conducted, the interactions during the sessions showed that the students appreciated being introduced to general and linguistic-specific repositories, which they would otherwise have not been able to discover through simple Google searches.

3.2.5 Emerging career paths

In the master's course in *Profession-Based Research*, part of DIT's mentioned Translation and Technology curriculum, a CLARIN guest lecture was delivered to ten students and two lecturers. This session introduced students with backgrounds in modern languages and translation to the CLARIN research infrastructure, highlighting potential career paths, such as data stewards, communication specialists, and training officers supporting researchers doing language research.

It was highlighted that graduates of language and linguistic programmes with strong language data management skills are well suited to take on data stewardship roles across all sectors where their disciplinary expertise is required to enhance the findability and reusability of language data through high-quality metadata, and manage the legal and ethical constraints governing access and reuse.

A summary of activities promoted by CLARIN and UNIBO is provided in Figure 2; the figure also shows some of the key elements of the next section, which introduces recommendations that extend beyond any single university or degree course.

4 Learning Objectives and Curriculum Design

4.1 Experience-based recommendations on how to teach about language resources

The CLARIN teaching experience at UNIBO confirmed the need for a more systematic approach for integrating CLARIN into university curricula. The authors of the present paper propose a conceptual framework for doing so based on three interrelated pedagogical principles:

1. **Situated Open Science learning**, which implies that Open Science and FAIR data principles are embedded in concrete research and data practices as early as possible. Students at all levels should work with real corpora and public data repositories, making openness and reusability tangible.
2. **Progressive competence development across educational levels**, which aligns with competence-based approaches to Open Science education (European Commission, 2017; Whyte et al., 2024); specific progressive learning objectives are summarized in Table 1 below, based on a learning trajectory that gradually moves from recognition to original contribution.
3. **Infrastructure-based pedagogy**, which supports end-to-end learning across the research data lifecycle, based on CLARIN's ecosystem of services for language data discovery, standards, tools and repositories.

[22] https://ilc4clarin.ilc.cnr.it/en/

[23] https://standards.clarin.eu/sis/views/view-centre.xq?id=ILC4CLARIN

Figure 2: Joint CLARIN-UNIBO activities.

In other words, our approach puts students at the centre, with hands-on activities based on field-specific data and tools being given more importance than frontal teaching and generic examples, and with FAIR-related topics being presented jointly with other topics related to text processing.

Using also the learning objectives designed for the UPSKILLS course *Introduction to Language Data: Standards and Repositories* as a topical basis for the conceptual framework, Table 1 lists the main topics related to infrastructure and shows the progressive development of competence using level-specific keywords. A more elaborate version of the table is available online as a Google sheet,[24] i.e. in the form of a matrix, and it has been discussed with fellow researchers and teachers during the CLARIN Annual Conference in Vienna, 2025. The subsections below elaborate on these learning objectives and translate them into more specific recommendations that can be implemented in curriculum design.

Topic	**Learning objectives**		
	BA	MA	PhD
Open Science and FAIR Data Principles	Recognize	Use	Apply
Research Infrastructures and Data Repositories			
Finding and (Re)Using Language Resources	Define	Assess	Solve
Analyzing Language Resources			
Building and Managing Language Resources	Explain	Analyze	Contribute
Sharing and Archiving Language Resources			

Table 1: A summary of language resource topics and learning objectives.

4.2 Bachelor's level: Awareness, discovery and responsible use

At the undergraduate level, teaching should raise awareness and build foundational literary in Open Science, FAIR data, and language resources, with a strong emphasis on discoverability, reuse and basic standards rather than data creation.

[24] https://tinyurl.com/yyhckrtj

Recommended approach for teachers

- Introduce Open Science concepts early, focusing on Open Access, Open Data, FAIR and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) principles, and why they matter in language-related research.
- Use CLARIN as an entry point to research infrastructures, helping students understand what a research infrastructure is and how it supports research communities.
- Emphasize finding and reusing existing language resources rather than producing new ones.

Key topics and learning outcomes

Students should be able to:

- Explain Open Science, Open Access, Open Data and FAIR and CARE principles.
- Understand what research data and language resources are, and distinguish between major resource types (written corpora, speech corpora, multimodal corpora, etc.).
- Discover language resources using CLARIN tools, such as the Virtual Language Observatory and Content Search.
- Recognize and compare licence conditions and basic legal/ethical constraints for reuse.
- Gain introductory exposure to annotation methods, corpus analysis tools, and common standards from written and speech corpora.
- Understand the benefits and challenges of sharing and archiving data, and the role of CLARIN centres and repositories.

Activities

- Hands-on discovery exercises in the VLO, Content Search and Resource Families of corpora and short reflective assignments on FAIRness and licensing

These suggestions are in line with the *Minimum Viable Skillsets* framework developed in the Skills4EOSC project[25] in the domain of Open Science and FAIR data principles and practices (White and Green, 2022, see also Whyte et al., 2023), according to which students at this level need to have foundational digital research skills, understand why it is important for society that research data is open and FAIR, and have good organization and documentation skills. The notion of a research infrastructure can also be introduced at this level and the CLARIN infrastructure can be used to show examples of published digitized text collections that are available and can be explored with free tools and services. By browsing the resources in the VLO, students can learn that published data and tools are described with consistent metadata, carry licences with different possibilities of reuse (public, academic, restricted), and can be located and cited using a persistent identifier (e.g., Handle or DOI). Teachers can also use the VLO to show students how to search for text collections in plain text format, download them and upload them to their preferred corpus-building tool. For instance, the VLO provides direct access to Voyant Tools[26] via the CLARIN Switchboard. As a free, open-source, and web-based platform, Voyant Tools can be introduced at the undergraduate level to gently introduce students with a humanities background to basic quantitative text analysis and visualization. For more advanced text analysis, the CLARIN resource families of corpora[27] can be used to show students how they can explore different types of corpora through the integrated concordancers (NoSketch Engine, Korp, or KonText). Finally, students can be introduced to the benefits and challenges of sharing and archiving language resources, as well as to the basic taxonomy of repositories (generic, discipline-specific, institutional) and to the idea that "responsible openness" may require additional safeguards when working with sensitive data, in line with CARE principles.

[25] https://www.skills4eosc.eu/
[26] https://voyant-tools.org/
[27] https://www.clarin.eu/resource-families

4.3 Master's Level: Applying FAIR principles and managing data

At the MA level, teaching should shift towards active data management, evaluation, and informed use, with students learning to apply FAIR principles in practice and work more independently using infrastructure services and available tools for language data processing and analysis.

Recommended approach for teachers

- Frame CLARIN as a practical research environment, not just a catalogue.
- Integrate FAIR assessment, standards and workflows into research methods.
- Encourage students to critically evaluate data quality, repositories and tools.
- Direct students to CLARIN Knowledge Centres (K-centres) for specialized expertise on language-specific corpora, methods and tools.

Key topics and learning outcomes

Students should be able to:

- Identify data sources on the web relevant for collecting open data for research projects in language/linguistics (e.g. repositories of research data, or data from governments and industry).
- Assess how FAIR existing language resources are, including repository practices.
- Distinguish between research data, metadata and documentation, and identify relevant data sources.
- Evaluate the quality and suitability of open data for specific research questions.
- Identify legal and ethical issues in language data collection and reuse.
- Understand the research data lifecycle, including planning, documentation, storage and sharing.
- Apply recommended community standards, file formats, and best practices for building and managing language resources.
- Use tools for language resource management, annotation, querying and basic programming-based analysis.
- Understand data publication, citation, and reuse, and how CLARIN repositories implement FAIR principles.

Activities

- Start by designing simple assignments that follow a full research cycle: discover data in the VLO, analyse data with Switchboard tools, export and interpret results, document and share outputs via a general-purpose repository, such as Zenodo. Such assignments will teach students to think in terms of "workflows".
- Embedd CLARIN resources into project-based assignments, e.g. write a simple data management plan before the start of the project, prepare a small dataset for potential deposit in a linguistic repository hosted by a CLARIN national consortium.

In line with the *Minimum Viable Skillsets*, at the MA level, students are expected to be able to integrate Open Science skills in their project workflows and assess FAIR concepts in their respective field of study, understand the data life cycle when conducting research. A useful starting point is to clarify foundational terminology by distinguishing *research data*, *datasets* and *metadata*, and by discussing how metadata quality affects findability and reuse. The research infrastructure can be used to encourage students to explore available resources and services, starting from a research question, and to apply the FAIR guiding principles to assess whether the published resources are suitable for reuse. For example, the CLARIN resource families of corpora can be used for interdisciplinary and comparative research; the Switchboard can be used to incorporate natural language processing (NLP) tools in the classroom and teach students basic NLP methods, such as text segmentation, tokenization, part-of-speech tagging,

syntactic parsing, and named entity recognition. In parallel, students can practice locating and accessing resources via persistent identifiers assigned by repositories, and become familiar with common standards and file formats used in CLARIN-oriented workflows. To strengthen good scholarly practice, they can be taught established guidelines for citing research data, as well as practical routines for sharing, reusing and citing language resources. After getting acquainted with basic NLP methods, students can be taught more advanced digital skills, such as programming in Python and R.

4.4 Doctoral level: Advanced research data management, publishing and stewardship

At PhD level, teaching should support researchers as data producers and stewards, focusing on more advanced research data management, legal/ethical responsibility, and data publication within their institutional repository or a domain-specific repository, e.g. a CLARIN national centre.

Recommended approach

- Treat Open Science and FAIR data as integral to doctoral research quality.
- Align training with the RDM requirements of funders and institutions.
- Use CLARIN to demonstrate end-to-end data stewardship, from planning to publication.

Key topics and learning outcomes

Doctoral candidates should be able to:

- Design and implement a research data management plan (DMP) for language-centric research.
- Identify high-level workflows in the creation and processing of language resources.
- Critically evaluate NLP and computational tools in terms of reproducibility and reusability.
- Address complex legal and ethical issues, including handling sensitive personal data and obtaining consent.
- Prepare language resources for deposit in a FAIR-compliant repository.
- Deposit, publish and cite datasets using CLARIN repositories and persistent identifiers.
- Create and share virtual collections to support transparency and reproducibility.

Activities

- Integrate CLARIN training into doctoral seminars and supervision, linking dataset publication directly to dissertation outputs and Open Science practices.

Students at this level are usually required to follow general Open Science and research data management courses, and are expected to be able to apply OS practices and FAIR data principles in their research (e.g., research data storage and archiving, handling legal and ethical issues surrounding the research data, selecting appropriate licences to protect potentially sensitive data). The infrastructure can be used to teach students how to find a suitable CLARIN repository to deposit language data and resources they may have created in their PhD project (e.g., corpora, audio and video recordings, annotations, grammars), and make them available to the community in a reliable manner, following established good practices in research ethics. If resources contain sensitive data, they can be protected by an institutional login, data encryption, and specific terms and conditions for data protection. Finally, doctoral students can take responsibility for preparing resources for depositing and archiving (packaging, metadata completeness, documentation, format choices), addressing legal and ethical issues in data sharing (e.g., personal data and consent, access restrictions), and depositing and publishing resources in a repository under conditions that balance openness with the protection of sensitive data.

5 Concluding remarks

Drawing on the UPSKILLS project and subsequent pedagogical activities at the University of Bologna, this paper proposed an integrated approach that combines progressive competence development with infrastructure-based pedagogy to foster sustainable and discipline-specific Open Science education. The conceptual framework described in Table 1 can help curriculum designers transition from abstract policy-driven instruction to situated learning using authentic language data from the CLARIN repositories and professional workflows. This framework refines the learning objectives of the *Introduction to Language Data: Standards and Repositories* course, employing a graduated matrix that moves from foundational recognition of FAIR principles at the Bachelor's level to integrated assessment at the Master's level and full professional application and data stewardship at the Doctoral level. Such a model directly addresses the data-related skills gap, supporting a smoother transition from student to researcher, and at the same time preparing students for emerging professional profiles, such as the *language data and project specialist* and *data steward*.

Although the proposed framework represents a significant step towards infrastructure-based pedagogy, it is not without limitations. A primary objective for our future research is to conduct a robust empirical evaluation of the framework's effectiveness in diverse classroom settings to validate its impact on student learning. To address this, we intend to refine our proposal through collaborative engagement with the broader academic community, inviting lecturers and trainers within the CLARIN network to contribute their expertise and feedback on the matrix of learning objectives. Another path worth exploring is the systematic deployment of the interactive research tracking tool, developed by consortium partners at the University of Zurich, which could enable lecturers to monitor student progress and provide targeted feedback during complex data-driven projects at all study levels. By integrating these practical tracking mechanisms with the resources available in the CLARIN Learning Hub, we aim to ensure the framework remains aligned with the evolving needs of the *language data and project specialist* profile and the broader European Open Science landscape.

References

Arsić, O., Assimakopoulos, S., Bernardini, S., Ferraresi, A., & Vella, M. (2023). *Enhancing learning with branching scenario simulations: Guidelines to using H5P content blocks* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.8353727

Berez-Kroeker, A., McDonnell, B., Koller, E., & Collister, L. (Eds.). (2022). *The Open Handbook of Linguistic Data Management*. The MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001

Bernardini, S., Ferraresi, A., & Tedesco, N. (2023). Processing texts and corpora. https://upskillsproject.eu/project/text_processing/

Bernardini, S., Polizzi, D., & Tedesco, N. (2025). Processing texts and corpora: An introduction [UniBOOK – UniBO Open Knowledge platform]. https://apps-book.unibo.it/learning/course/course-v1:UNIBO+DEH07+aa_24-25/home

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=3339930

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, *68*(9), 2182–2200. https://doi.org/10.1002/asi.23781

De Smedt, K., van der Lek, I., van den Heuvel, H., Balvet, A., Janssen, M., Cinková, S., Sanz, A., Assimakopoulos, S., & ten Bosch, L. (2024). CLARIN in training and education. In K. Lindén, T. Kontino, & J. Niemi (Eds.), *Selected papers from the clarin annual conference 2023*. https://doi.org/10.3384/ecp210011

Demchenko, Y., Stoy, L., Engelhardt, C., & Gaillard, V. (2021). *D7.3 FAIR competence framework for higher education (Data stewardship professional competence framework)* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.5361917

Edmond, J. (Ed.). (2020). *Digital Technology and the Practices of Humanities Research*. Open Book Publishers. https://books.openedition.org/obp/11899

Engelhardt, C., Barthauer, R., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Hansen, K. K., Kuusniemi, M. E., … Zhou, B. (2022). *How to be FAIR with your data: A teaching and training handbook for higher education institutions*. Universitätsverlag Göttingen. https://doi.org/10.17875/gup2022-1915

European Commission. (2017). *Providing researchers with the skills and competencies they need to practise Open Science*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/121253

Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC corpora? *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)*, 25–30. https://cmccorpora19.sciencesconf.org/resource/page/id/15

Gledić, J., Đukanović, M., Budimirović, J., Soldatić, N., Miličević Petrović, M., Bernardini, S., Ferraresi, A., Tedesco, N., van der Lek, I., Fišer, D., Puskas, G., Pallottino, M., Berthouzoz, M., Kraš, T., Podboj, M., Simonović, M., Samardžić, T., van der Plas, L., Tanti, M., … Vella, M. (2023). Upskills teaching and learning content [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1865

Krauwer, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: Resources and tools for e-humanities scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1525–1531. https://dspace.library.uu.nl/handle/1874/307981

Miličević Petrović, M., Bernardini, S., Ferraresi, A., Aragrande, G., & Barrón-Cedeño, A. (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.5030929

Pedonese, G., Frontini, F., del Fante, D., & Federici, E. (2025). Adapting UPSKILLS learning modules to the university curricula: Best practices and lessons learnt from the H2IOSC training experience at the University of Ferrara. *Selected papers from the CLARIN Annual Conference 2024*, 37–47. https://doi.org/10.3384/ecp216.04

Simonović, M., Arsenijević, B., van der Lek, I., Assimakopoulos, S., ten Bosch, L., Fišer, D., Kraš, T., Marty, P., Miličević Petrović, M., Milosavljević, S., Tanti, M., van der Plas, L., Pallottino, M., Puskas, G., & Samadžić, T. (2023). *Research-based teaching: Guidelines and best practices* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.8176220

UNESCO. (2021). UNESCO Recommendation on Open Science. https://doi.org/10.54677/MNMH8546

van der Lek, I., Fišer, D., Frontini, F., & Pedonese, G. (2024). Introduzione ai dati linguistici: Standard e archivi digitali [Zenodo]. https://doi.org/10.5281/zenodo.13911935

van der Lek, I., & Fišer, D. (2023). Introduction to language data: Standards and repositories. https://upskillsproject.eu/project/standards_repositories/

van der Lek, I., Fišer, D., Samardžić, T., Simonović, M., Assimakopoulos, S., Bernardini, S., Miličević Petrović, M., & Puskas, G. (2023). *Integrating research infrastructures into teaching: Recommendations and best practices (version 2)* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.8114407

White, A., & Green, D. (2022). Minimum Viable Skillsets: A briefing for competence centres. https://doi.org/10.5281/zenodo.14865952

Whyte, A., Green, D., Avanço, K., Di Giorgio, S., Gingold, A., Horton, L., Koteska, B., Kyprianou, K., Prnjat, O., Rauste, P., Schirru, L., Sowinski, C., Torres Ramos, G., van Leersum, N., & Lazzeri, E. (2024, June). Minimum Viable Skillsets - Catalogue of Open Science Career Profiles. https://doi.org/10.5281/ZENODO.11469300

Whyte, A., Green, D., Avanço, K., Di Giorgio, S., Gingold, A., Horton, L., Koteska, B., Kyprianou, K., Prnjat, O., Rauste, P., Schirru, L., Sowinski, C., Torres Ramos, G., van Leersum, N., Sharma, C., Méndez, E., & Lazzeri, E. (2023). *D2.1 catalogue of open science career profiles - minimum viable skillsets* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.8101903

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18