CLARIN

Selected papers from the
CLARIN Annual Conference 2024
Barcelona, Spain

Selected papers from the

CLARIN Annual Conference 2024

Barcelona, Spain, 15–17 October 2024

edited by Vincent Vandeghinste and Thalassia Kontino

CLARIN

Common Language Resources and Technology Infrastructure

Front Cover Illustration:
Picture Composition by CLARIN ERIC

Linköping Electronic Conference Proceedings 216
eISSN 1650-3740 (digital) • ISSN 1650-3686 (print) 2024
ISBN 978-91-8118-188-3 (PDF)

Introduction

Vincent Vandeghinste
Program Committee Chair
Dutch Language Institute, Netherlands,
KU Leuven, Belgium
vincent.vandeghinste@ivdnt.org

Darja Fišer
Executive Director of CLARIN ERIC
Institute of Contemporary History
Slovenia
darja.fiser@inz.si

CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform that is accessible to everyone interested in language. CLARIN offers access to language resources, technology, and knowledge, and enables cross-country collaboration among academia, industry, policy-makers, cultural institutions, and the general public. Researchers, students, and citizens are offered access to digital language resources and technology services to deploy, connect, analyse and sustain such resources. In line with the Open Science agenda, CLARIN enables scholars from the Social Sciences and Humanities (SSH) and beyond to engage in and contribute to cutting-edge, data-driven research based on language data in a range of formats and modalities. The CLARIN infrastructure is run by CLARIN ERIC[1], a consortium of participating countries and institutes that was established in 2012 and has grown considerably in size since. Currently there are 27 member countries and around 240 associated research institutions, which are all encouraged and supported to be represented at the annual conference.

The CLARIN Annual Conference is the main annual event for those working on the construction and operation of CLARIN across Europe, as well as for representatives of the communities of use in the humanities and social sciences. The event is central for the CLARIN community and is one of the crucial instruments for CLARIN to function as a knowledge hub. At the conference, consortia from all participating countries and the various communities of users meet in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. Moreover, CLARIN 2024 was also intended for the wider humanities and social sciences communities in order to exchange ideas and experiences within the CLARIN infrastructure. This includes the design, construction and operation of the CLARIN infrastructure, the data, tools and services that it contains or for which there is a need, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure.

In January 2024, a call was issued for which 61 abstracts were submitted. All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 61 submitted abstracts, 43 submissions were accepted for presentation at the conference (acceptance rate 0.70). The Proceedings with extended abstracts are available at https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf. The CLARIN Annual Conference 2024 took place in Barcelona, Spain, from 15 to 17 October 2024 as a hybrid event.

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 2024*. All full-length paper submissions underwent the review cycle again, and 13 papers were accepted:

[1] http://www.clarin.eu

Vincent Vandeghinste and Darja Fišer 2024. Introduction. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. i–iv. DOI: https://doi.org/10.3384/ecp216

- The paper by Wisik *An Infrastructural Approach to Terminology Work: The Case of Research Infrastructures* explores the role of research infrastructures with regard to terminology work in institutional settings.

- The paper by Steingrímsson and Sigurðsson *Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset* describes a dataset of multiword expressions and evaulates its use in the context of machine translation.

- The paper by Ahltorp and Skeppstedt *Word Rain as a Service: Making semantically structured word clouds available to everyone* describes the Word Rain text visualisation technique and how it is made available as a service and as open source code.

- The paper by Pedonese, Frontini, Del Fante and Federici *Adapting UPSKILLS Learning Modules to the University Curricula: Best Practices and Lessons Learnt from the H2IOSC Training Experience at the University of Ferrara* details the steps taken to adapt and integrate the training materials developed by CLARIN ERIC in two bachelor's degree courses and one master's degree course at the University of Ferrara.

- The paper by Körner, Eckart, Helfer and Kretschmer *An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences* discusses the development of the Federated Content Search in recent years, with newly formulated application scenarios, newly opened up user groups and newly developed tools and user interfaces.

- The paper by Zinn and Trippel *On the Successful Migration of Languages Resources from one Repository to Another* describes the actual challenges in the migration process of language resources, the deviations from the original scenario and the compromises needed, and finally, how to succeed to get all data transferred in a safe and information-preserving manner.

- The paper by Erjavec, Ljubešić, Meden, Kuzman, Laskowski, Javoršek, Krek, Tomazin, Dobrovoljc, Arhar Holdt, Jakob Lenardič and Darja Fišer *CLARIN.SI, the Slovenian node of CLARIN: ten years on* presents the organisation and services offered by the Slovenian research infrastructure for language resources and technologies CLARIN.SI.

- The paper by Navarretta, Haltrup Hansen and Jongejan *Policy Domains and the Speakers' Gender in ParlaMint-DK* describes the ParlaMint-DK 4.1 corpus, which consists of the Danish parliament speeches from 2014 to 2022 annotated with 20 general policy domains mapped to the codebook of the Comparative Agendas Project.

- The paper by Illig, Diewald, Kamocki and Kupietz *Managing Access to Language Resources in a Corpus Analysis Platform* presents an approach to maximize user access to corpus data while protecting rights holders' legitimate interests. Query rewriting techniques and authorization procedures allow for modeling license terms in detail, enabling broader applications, offering an alternative to methods that only model a greatest common denominator of licenses, thereby limiting the possibilities for using the data.

- The paper by Daza and Fokkens *Choosing the Right Tool for You: Informed Evaluation of Text Analysis Tools* exemplifies a methodology that stimulates critical evaluation and detailed analysis of automatic outputs of NLP tools. It particularly analyzes the case of choosing the best Named Entity Recognition tool for a corpus of Dutch biographies.

- The paper by Ecker, Fischer, Schwarz, Trippel, Werthmann and Wilm *Unlocking the Corpus: Enriching Metadata with State-of-the-Art NLP Methodology and Linked Data* describes an approach to add semantic metadata on the text level to facilitate the search over data, enriching metadata with three methods: named entity recognition, keyword extraction, and topic modeling.

- The paper by van Gompel and Windhouwer *FAIR Tool Discovery: an automated software metadata harvesting pipeline for CLARIAH* describes a pipeline that harvests software metadata from the source, detects existing hetero geneous metadata formats already in use by software developers, and converts them to a single uniform representation.
- The paper by Alkorta, Farwell, Fernandez de Landa, Altuna, Estarrona, Iruskieta, Arregi, Goenaga, Arriola *CLARIAH-EUS: A Strategic Network Helping Basque Country Researchers to Participate in European Research Infrastructures* outlines the rationale behind establishing CLARIAH-EUS, a node within CLARIAH-ES, focused on Basque or Basque culture-related research in the humanities, arts, and social sciences.

We thank all PC members and reviewers for their evaluation efforts, Thalassia Kontino from the CLARIN Office for support, and our colleagues at Linköping University Electronic Press for smooth publication. The 2024 programme subcommittee included Vincent Vandeghinste (chair), Stelios Piperidis, Cristina Grisot, and Krister Lindén.

Programme Committee Members

- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven, Belgium (chair)
- Starkaður Barkarson, Árni Magnússon Institute, Iceland
- Lars Borin, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Cristina Grisot,University of Zurich and at the Swiss National Center for Data & Services for the Humanities DaSCH, Switzerland
- Eva Hajičová, Charles University, Czechia
- Krister Lindén, University of Helsinki, Finland
- Monica Monachini, Institute of Computational Linguistics, Italy
- Costanza Navarretta, University of Copenhagen, Denmark
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, ILSP, Greece
- German Rigau, Basque Center for Language Technology, Spain
- Gijsbert Rutten, Leiden University, The Netherlands
- Kiril Simov, Bulgarian Academy of Sciences
- Inguna Skadiņa, University of Latvia
- Gunn Inger Lyse Samdal, University Library, University of Bergen, Norway

- Sara Stymne, Uppsala University, Sweden
- Marko Tadić, University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Hungarian Academy of Sciences
- Joshua Wilbur, University of Tartu, Estonia
- Tanja Wissik, Austrian Academy of Sciences
- Andreas Witt, University of Mannheim, Germany
- Friedel Wolff, North-West University, South Africa
- Martin Wynne, University of Oxford, UK

Contents

An Infrastructural Approach to Terminology Work: The Case of Research Infrastructures

Tanja Wissik
Austrian Centre for Digital Humanities and Cultural Heritage
Austrian Academy of Sciences
Vienna, Austria
tanja.wissik@oeaw.ac.at

Abstract

This study explores the role of research infrastructures, in particular the role of CLARIN and DARIAH, with regard to terminology work in institutional settings (academic and non-academic) by analyzing a body of qualitative interview data, collected in 2023 across Europe. The contribution also discusses how research infrastructures (RIs) could reach out to new non-academic communities e.g., in the public sector and to reevaluation of existing terminology infrastructure models and include research infrastructures.

1 Introduction

It is not new to use the term *infrastructure* for the organisation of terminological collaboration and terminological activities (Pilke et al. 2021, 101). Already Galinski (1998) described an infrastructural approach to terminology, dividing terminological infrastructures into horizontal and vertical infrastructures. The horizontal infrastructure includes five elements: "terminology (planning) policy, terminology creation centres, terminology information and documentation centres, terminology associations and corporate cooperation groups led by the private sector" (Galinski, 1998). The vertical infrastructure concerned the different ways of carrying out terminological activities within different domains (Galinski, 1998). An adapted version of this infrastructural model (Galinski & Giraldo, 2023) is used in a cooperation project between Austria and Mongolia[1]. This model includes five horizontal layers "(1) Individual users and data creators (as well as their groups and networks), (2) Intermediaries and service providers to individual users/groups, (3) Organizations/networks of data creators and curators, (4) Terminology infrastructure coordination authority/ies, and (5) Policies and high-level authorities. At all levels six vertical functions and activities occur: (A) Organizational aspects, (B) Services, (C) Education and R &D, (D) ICT systems and tools, (E) Support and promotion, and (F) Legal & technical regulations." (Galinski & Lušicky, 2024). In these existing infrastructural models for terminology, research infrastructures are not explicitly included, however there are several connecting points between terminology work and research infrastructures (e.g., Andersen & Gammeltoft, 2022; Wissik & Declerck, 2020; Wissik, 2022). Stakeholders in terminology work can be on the one hand data providers and on the other hand they can be users of data, tools and services provided by RIs and benefit from the knowledge sharing infrastructure to exchange knowledge and promote collaboration.

However, there is little insight into the role and use of such research infrastructures, in particular CLARIN and DARIAH, within the community of stakeholders involved in terminology work, especially in institutional settings, besides some case studies (e.g., Andersen & Gammeltoft, 2022). So, this paper

[1] The name of the project is "Terminology planning strategy and terminology infrastructure for Mongolia to support scientific and educational development and innovation".

Tanja Wissik 2024. An Infrastructural Approach to Terminology Work: The Case of Research Infrastructures. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 1–12. DOI: https://doi.org/10.3384/ecp216

wants to close this gap and explores the role of research infrastructures with regard to terminology work in institutional settings (academic and non-academic) based on qualitative interview data. The paper is structured as follows: after an introduction, the research method is described, and the results are discussed. The contribution focuses on resources (e.g., corpora) and repositories as possible links between research infrastructure and the community of stakeholders involved in terminology work, mentioning also other possible areas of cooperation such as training materials and tools.

2 Background

2.1 Research Infrastructures in the Humanities

Among the first Research Infrastructures (RIs) to support research in the Arts and Humanities were libraries, museums and archives (Moulin et al., 2011). In today's digital age a number of RIs have emerged at European and national level to support digital research. These RIs offer technical infrastructures in a more stable and sustainable way than research projects that run only short period. As technical infrastructure they provide resources, tools and services to the scientific community in order to support top-level research activities. Furthermore, RIs provide a social infrastructure for collaboration and knowledge exchange and they act as promoter for the use of common methods and standards. RIs also play a crucial role in training and educating future generations of researches and research engineers (Wissik & Declerck, 2020).

In the Social Sciences and Humanities (SSH) the European Strategy Forum on Research Infrastructures (ESFRI) recognizes five large European research infrastructures and six research infrastructure projects (ESFRI 2021). RIs can be generic or domain specific. As generic research infrastructures, for this paper, we understand research infrastructures that can be used by researchers from a variety of research fields within the SSH. In the Humanities, e.g., CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities) count as generic research infrastructures (Doel & Maes, 2012) and for example EHRI (European Holocaust Research Infrastructure) can be seen as a domain specific research infrastructure (Wissik & Declerck, 2020).

In the following we will describe the two generic RIs CLARIN and DARIAH that are most relevant for the field of terminology research and practice.

CLARIN stands for Common Language Resources and Technology Infrastructure) and it was established as an ERIC in 2012 with "the mission to create and maintain a digital infrastructure to support the sharing, use and sustainability of language data (in written, spoken, or multimodal form) available through repositories from all over Europe, in support of research in the humanities and social sciences and beyond" (de Jong et al, 2022). Currently, CLARIN currently has 26 member countries that operate a distributed network of data, service and knowledge centres. Through its network, CLARIN provides access to digital language resources, services and expertise to scholars, researchers, and students from different disciplines in the SSH.

The other relevant generic infrastructure in the humanities is DARIAH, which stands for Digital Research Infrastructure for the Arts and Humanities, and was established in 2014 as an ERIC with the mission "to empower research communities with digital methods to create, connect and share knowledge about culture and society." (DARIAH, 2025) Currently DARIAH has 22 member countries and 17 cooperating partners in non-member countries. DARIAH represents a network of "people, expertise, information, knowledge, content, methods, tools and technologies." (DARIAH, 2025a).

In this section, the two generic RIs CLARIN and DARIAH were introduced. In the next section we will discuss some offerings from those RIs, that could be of interest for stakeholders working in the area of terminology.

2.2 Research Infrastructures and Terminology

As already mentioned before, there are several connecting points between terminology work and research infrastructures (e.g., Andersen & Gammeltoft, 2022; Wissik & Declerck, 2020; Wissik, 2022). In the following we will look into some of these connecting points by discussing services offered by RIs.

Stakeholders in the area of terminology on the one hand produce language data as a result of the terminology workflow. The data can be in the form of terminology databases, glossaries or terminological dictionaries. Furthermore, they also might create specialized corpora, focusing on a specific domain or subdomain. On the other hand, they often consult already existing language resources, such as terminology databases, glossaries, dictionaries or corpora in the process as well. Therefore, we will look into ways of accessing existing language resources as well as services for depositing language resources that are provided by the before mentioned RIs.

In CLARIN, there are several ways of accessing the language resources and tools that are available within the infrastructure: One way of accessing is via the Virtual Language Observatory (VLO), a search catalogue to find language resources via different facets (Van Uytvanck et al., 2012; Goosen & Eckart, 2014); Another way of accessing is via the so-called CLARIN Resource Families, a manually curated overview of language resources and tools organized by type (e.g., glossaries, corpus query tools) (Fišer et al., 2018). Another option of accessing is directly via the data repositories hosted by the CLARIN B-Centres network.

Besides exploring and searching for available resources via their metadata, it is also possible to search for patterns within the language resources directly online. One option is the Federated Content Search (FCS)[2], which allows the search in different language resource, hosted at different centres, at the same time. The FCS was originally created for searching in full-texts with optional annotation layers. Recently a new LexFCS extension (Eckart et al., 2023) was designed to make also lexical resources such as dictionaries, word lists or semantic wordnets, searchable via the FCS. A first implementation of this extension is included in the Text+ FCS Aggregator (Körner et al., 2024).

Furthermore, it is also possible to search in individual corpora via provided concordance tools by dedicated CLARIN B-Centres (e.g., CLARIN.si or LINDAT/CLARIAH-CZ).

Another service, that is provided by the CLARIN infrastructure is the depositing service for language resources. This service is offered by CLARIN B-Centres, who maintain certified data[3] repositories in order to archive and disseminate language resources. Depositing language resources is one essential step in making the language resources findable, accessible, interoperable and reusable.

Since DARIAH is serving a broader community, there are very diverse tools and services listed in the DARIAH Tools and Services Catalogue. In particular, the Vocabs services, hosted and maintained by the ACDH-CH might be of interest for the terminology community. The Vocab services contain on one hand a controlled vocabulary repository and on the other hand an editor in order to collaboratively create and maintain controlled vocabularies.

Furthermore, the SSH Open Market place (König et al. 2024), a discovery portal maintained by the three RIs DARIAH, CLARIN and CESSDA, can be of interest for the terminology community, on one hand to search for tools and services but also to publish and share their own tools and services with the broader SSH community. When searching for the keyword "terminology" in the SSH Open Market place, 36

[2] https://www.clarin.eu/content/content-search

[3] A Clarin B-Centre needs to have the Core Trust Seal and must fulfill the requirements set by the CLARIN Technical Centre Assessment Committee. More information available here https://www.clarin.eu/content/clarin-b-centre-assessment

resources were listed (21 data sets, 10 tools and services, 4 publications and 1 training material. However, no workflow was listed[4]).

Besides a technical infrastructure, RIs also provide expertise and a space for knowledge exchange. CLARIN maintains a knowledge infrastructure, comprising a network of CLARIN Knowledge Centres, so-called CLARIN K-Centres, to make expertise and knowledge available in a structured way. When searching with the keyword "terminology" through the CLARIN K-Centre inventory, there are 7 CLARIN K-Centres listed with expertise in the area of terminology: CLARIN-ELEXIS Knowledge Centre for Lexicography, Czech CLARIN Knowledge Centre for Corpus Linguistics, CLARIN Knowledge Centre for Dutch, CLARIN K-Centre for Natural Language Processing in Greece, CLARIN Knowledge Centre for Polish Language Technology, CLARIN Knowledge Centre for The Languages of Sweden and CLARIN K-Centre for Terminology Resources and Translation Corpora.

In the case of DARIAH, a lot of expertise can be found in the DARIAH Working Groups. The goal of the DARIAH Working groups is "to consolidate infrastructure and scholarship in certain areas of research and to create or reinforce the network of expertise inside of DARIAH" (Scharnhorst et al. 2019, 9). At the moment there is no WG specifically targeting the terminology community, but there are several working groups where there are connecting points: for example, with the WG "Lexical Resources"[5] because they are dealing with all kinds of digital resources not only dictionaries, also lexicons, thesauri, word lists etc. and they are not only working with semasiological methods and standards but also with onomasiological methods and standards. Another connection point can be seen with the WG "Multilingual DH", that want to enhance digitally-enabled research in under-resourced languages by adapting or developing tools for those languages. Furthermore, the Research Data Management WG might be of interest for the terminology community, regarding best practices in data management, the use of data repositories, FAIR data principles etc.

2.3 Related Research

There was some qualitative research done on interinstitutional cooperation in terminology work by Chiocchetti & Ralli (2013). They conducted expert interviews with 17 terminologists in institutional settings. However, when the interviews were conducted between 2011 and 2012, research infrastructures in the Humanities were fairly new, so the topic of research infrastructure was not discussed in their contribution. Budin (2015) provided a theoretical discussion on the role of research infrastructure, particularly CLARIN, in the context of computational terminology, without evidence from empirical data. For the ELRC White Paper (ELRC, 2022) on "AI for a multilingual Europe – Why Language Data Matters" a survey was conducted to get insights into the current use and importance of language technologies and into common European practices with respect to translation, data management and sharing in public administrations and SMEs. 73 people responded to the survey: their answers were the basis for the ELRC White Paper. The national CLARIN consortia were mentioned several times in the ELRC White Paper in the country report section.

Furthermore, there are recent initiatives, besides the already known catalogues and registry (e.g., Virtual Observatory, ELRC Share, the ELRA Universal Catalogue), to collect terminology data as shown by the TeresIA project (Maroto, 2024). In this project a survey for terminology providers was set up to collect information on the data and metadata in order to integrate these terminology resources into a newly created meta-search portal[6].

In this section we have discussed those parts of the technical and social infrastructure of RIs, that might be relevant for the stakeholders in terminology work and we have reviewed some related research.

[4] Here the search performed on January 29, 2025 can be found https://marketplace.sshopencloud.eu/search?q=Terminology

[5] https://www.dariah.eu/activities/working-groups/lexical-resources/

[6] https://proyectoteresia.org/colaborar

However, until now there was little insights, which role RIs, especially CLARIN and DARIAH, have within the community of practice.

3 Method

The present contribution is part of a larger study, carried out in 2023, exploring the role and impact of new technologies and new paradigms, such as open data, on terminology work performed in institutional settings and how workflows, tasks and roles are influenced consequently.

To gain insight in this area, 15 semi-structured expert interviews (Meuser & Nagel, 1991: 443) were conducted with individuals involved in terminology workflows in different institutional settings in different roles to better understand terminology workflows in the digital age (Wissik, 2024).

Experts that were eligible for this study had to work in an institution[7] that (1) performs practical terminology work and (2) maintains a publicly accessible terminology database or terminological dictionaries. Moreover, the interviewees had to be regularly involved in practical terminology workflows. An initial selection of candidates was identified through relevant terminology networks. Additionally, relevant international organizations were added to the list. From this initial selection of candidates, 15 experts consented to participate in the study (Wissik, 2025). One limitation of the study is the small sample size. However, the final sample covered academic and non-academic institutions: 4 regional/state administrations in Europe, 2 European institutions, 2 international organizations, and 7 academic/research institutions in Europe (universities or academy-based terminology institutes and language centres). There were both, institutions from CLARIN and/or DARIAH member countries and non-member countries (see Table 1). Moreover, the sample covered institutions with different sizes of publicly accessible terminology database, ranging from large databases with over 400,000 entries to smaller ones with fewer than 10,000 entries and all major types of terminology work: systematic, ad hoc, translation-oriented/text-based, preparatory for standardization (e.g., terminology planning), proactive, and a posteriori terminology work. Furthermore, different roles within the terminology workflow were represented including directors of terminology units, terminologists, terminology managers/coordinators, technology managers, developers/IT experts, and members of standardizing committees (Wissik, 2025).[8] So, even though given the small sample size, it reflects most of the scenarios of institutional terminology work. And experiments with qualitative interview data have shown, that more interview does not necessarily mean more information and that already with 12 interviews data saturation was reached (Guest et al. 2006)

The questions for the expert interviews were based on a previous study on terminology workflows by Chiocchetti and Ralli (2012, Attachment A) and were modified for this study (see question list on Zenodo).[9]

Table 1. Interview participant profiles (adapted from Wissik, 2025)

[7] Experts from commercial settings or freelancer were not included in this study.

[8] For more details on the interview participants see Wissik (forthcoming).

[9] List of questions is available on Zenodo with the following link https://doi.org/10.5281/zenodo.11144968

Interview Number	Role	Type of Institution	CLARIN Member	DARIAH Member	Less-resourced language
INT 1	Developer / IT expert	Research / Academic	no	yes	yes
INT 2	Technology manager	Research / Academic	no	yes	yes
INT 3	Head Terminology Unit/ Terminologist	Research / Academic	yes	yes	yes
INT 4	Member of Terminology Committee	Research / Academic	no	yes	yes
INT 5	Head of Terminology and Legal Translation Unit, Deputy Director for Development	Administration (national level)	yes	cooperating partner	yes
INT 6	Terminologist	Administration (regional level)	yes	yes	yes
INT 7	Terminologist	Research / Academic	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	Yes
INT 8	Head Unit Project Management / Terminologist	Administration (regional level)	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	yes
INT 9	Terminologist	Research / Academic	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	yes
INT 10	Terminology Coordinator / Terminology Manager	Intergovernmental Organisation	no	no	yes
INT 11	Head Unit Terminology / Terminologist	Intergovernmental Organisation	no	no	no
INT 12	Technology Manager	Intergovernmental Organisation	no	no	no
INT 13	Head Unit Terminology / Terminologist	Intergovernmental Organisation	no	no	yes
INT 14	Terminology Coordinator / Terminology Manager	Research / Academic	yes	Cooperating Partner	yes
INT 15	Head Unit Terminology / Terminologist	Administration (regional level)	Observer	yes	no

Most interviews were conducted in English, two interviews were conducted in German. The transcribed and anonymized interviews were analyzed by using a thematic qualitative text analysis (Kuckartz, 2014). The data was encoded with CATMA (Gius et al., 2023), an open-source annotation tool that allows to create your own categories to annotate the data (Wissik, 2025).

This study explored the role of RIs in particular the role of CLARIN and DARIAH with regard to terminology work in institutional settings. The relevant questions asked in this context where (1) which material/resources do the participants use for their terminology work and (2) if the participants use or create corpora when doing terminology work and if they publish the corpora, they have created. Furthermore, (3) if they deposit their terminological data in data repositories and (4) if they collaborate with Research Infrastructures, in particular CLARIN and/or DARIAH.

4 Results

When analysing the interview data, several aspects regarding the actual role of RIs in terminology work and potential connection points emerged. In this contribution we will focus on the following aspects: use of corpora and other language resources when compiling terminological resources, use of data repositories for the created language resources and collaboration or engagement with RIs. All of these aspects will be described and discussed below and illustrated with examples from the interview data.

4.1 Corpora and other language resources

When compiling terminological resources, documentation is collected to extract the terms to be included into the final resource. Therefore, using digital corpora in terminology work is not new (Bowker, 1996; Pearson, 1998). When doing ad hoc terminology work, for example, answering requests from query services (Žagar Karer & Fajfar, 2023), terminologists usually resort to already existing corpora:

> "We use already compiled corpora, for example general language corpora, to check how a specific term is used or some domain corpora just to check whether they [terms] are used and frequencies, when we deal with several term candidates, which one is more represented which one is less, that kind of things." (INT 7)

Besides corpora also other language resources are used for researching and verifying terms such as already existing dictionaries:

> "[...] we search through dictionaries, [...] depends on the type of the problem and in specialized texts and in corpora. We have a lot of corpora in [Name of Country], we have corpora of academic texts, and general corpora and all sorts of specialized corpora, so we can check in different kind of corpora to see the situation in language." (INT 3)

When doing systematic terminology work for a specific domain in order to create for example, a specialized dictionary or to enrich a terminology database with a new domain, terminologists also create specialized corpora from scratch: "In the beginning when we started compiling […] dictionaries we always prepare a specialized corpus of the texts that the experts give them [the texts] to us and then we the terminologist, prepare wordlist" (INT 3).

Another resource that was mentioned, especially in the translation context, were parallel corpora: "[a]s part of our work in terminology we have built up parallel corpora" (INT2). A special type of parallel corpora, translation memories were also mentioned when asked for the use of corpora: "No [we don't use corpora for systematic terminology work], we use our databases primarily and translation memories for that but corpora we use when answering these consultations were there more language problems, morphology problems and those cases." (INT5)

Other interview partners have stated, that they tried to create their own corpora, but they encountered various obstacles, such as little data for specific less-resourced language or missing specific IT support, because the institution has only a generic IT department and not a specific IT department that supports only language technologies (INT 7):

> "We made some experiments in creating our own corpora, but we did not get good results. Because we have little data on [Name of less-resourced language]. Corpus was compiled [in] [Name of well-resourced language]. You really need a large spectrum to cover all the areas to have a balanced corpus and we never ended with having balanced corpora. There were always some areas that were more present. Not all the subject fields we wanted were represented. And because we did not have an IT department, that only work for us, for our need, it did not work for us." (INT 7)

The snippets from the interview show, that in institutional terminology work on the one hand already existing corpora are consulted, as well as other already existing language resources such as dictionaries and on the other hand specialized corpora are created from scratch. When asked about publishing those corpora, some interview partners mentioned, that they publish their corpora on their own website, depending on legal constrains:

> “We do publish a lot of our corpora. We have a national corpus project as well […] and on our own website we have the contemporary corpus of [Name of less-resourced language] which is about 100 million worth of contemporary [Name of less-resourced language] published over the last decade or so and we have a parallel corpus of legislative materials from EU legislation and national legislation and they are both freely available on our website. Some of the material is to download, some of them are there only to search, again depending on the copyright restrictions.” (INT 4).

However, most interview participants create the corpora for internal use only. They do not publish them or deposit them in a repository: “We usually keep it as a working material. It's more like really like a stage in preparing a dictionary, it is not annotated with POS [part of speech]. I would take us too much time for this.” (INT3). Sometimes they are also shared through a corpus management platform but they are not published:

> “[W]e store them [corpora] in SketchEngine, it's collaborative so you can share the corpus with other people in the organization or outside the organization so that's very useful and we typically leave it there. I mean we don't export them we don't. Sometimes we will use them but we don't publish them or we don't, you know, otherwise store them except for sketch engine where we have a license and some storage.” (INT 13).

4.2 Data repositories

Data repositories are a way to archive language resources for the long-term and make the language resources (e.g., corpora, terminological resources) available to the community in a reliable way. Usually, data repositories assign persistent identifiers and therefore the data can be cited easily. Through the available metadata it is also possible to search for the language resources efficiently.

Regarding the use of repositories, Wissik (2024) analysed the use of data repositories for terminological data in general in the context of sustainability and the findings showed, that it was not a very common practice among the interviewees to store terminological data in a data repository. However, most of them had multiple other access points to their data and alternative data backup strategies.

For this contribution we only analysed the use of data repositories that are related to RIs. Only one interview participant reported, that their terminological data is stored in a national CLARIN repository (INT 3) and one interview participant reported, that they had recently talked with a national CLARIN representative also about the possibility of archiving the data in a CLARIN repository in the future (INT 14). For publishing their corpora, none of the interview participants used data repositories (see also section 4.1).

4.3 Collaboration or engagement with Research Infrastructures such as CLARIN or DARIAH

A part of the interview was also dedicated to participation in networks and collaborations with other institutions with special focus on CLARIN and/or DARIAH. When looking at the interview profiles in Table 1, it shows, that out of the 15 interview partners, 11 interview partners were from a country that are part of CLARIN and/or DARIAH or at least cooperating partners.

All the interviewees mentioned that they are active in different networks and associations regarding terminology, or specific languages etc., as institutions or as individuals. Several of the non-academic institutions mentioned collaborations with universities. Regarding the explicit collaboration with CLARIN and DARIAH, most academic institutions were aware of both RIs, and some had direct links. Besides using for example, the repositories in the CLARIN infrastructure, also activities in committees were mentioned: “I think one of my colleagues is member in CLARIN, she is active member in some committee” (INT 3). However, most of the units responsible for terminology had no direct links: “I think [Name of University] possibly has some DARIAH links, but not our unit.” (INT 4). Furthermore, some interview partners mentioned, that they are planning to collaborate in the future: “And regarding CLARIN and DARIAH we are not collaborating with this research infrastructures at the moment but we are considering such collaboration in the future.” (INT 9). Several interview participants, especially those from non-academic institutions, were not familiar with CLARIN and DARIAH (e.g., INT 5, INT 6, INT 8).

5 Discussion

Regarding the use of corpora, most of the interviewees reported, that they were using already existing corpora, especially in the context of ad hoc terminology, but also other resources such as dictionaries. In these cases, also corpora from CLARIN national consortia were mentioned and used. However, not all interviewees were aware of the available variety of resources through CLARIN and DARIAH. The possible ways of accessing these resources, via the VLO via the CLARIN Resource Families and via the SSH Open Market Place could be promoted further to this target group. Especially Corpora of Academic Texts, Legal Corpora and Parallel Corpora, Dictionaries and Glossaries might be of particular interest for the terminology community. As described before, in the interview data, translation memories were mentioned as a resource, and not so often parallel corpora. In the CLARIN Resource Family for Parallel Corpora, also Translation Memories are listed, but it is not mentioned in the title. If it is not too long, also Translation Memories could be integrated into the title, to show that Translation Memories are included and that this CLARIN Resource Family is among the hit list, when searching for Translation Memories. Besides the resources themselves, also the possibilities of the Federated Content Search and the use of online concordance tools could be beneficial for terminology community, especially when checking for example the use and frequencies of several term candidates.

Furthermore, consideration should be given to adapt the LexFCS extension to terminology resources or to design and implement a new extension for the FCS in order to search through distributed (multilingual) terminology data, which is already available in the CLARIN infrastructure.

Regarding the creation of their own corpora, several participants mentioned, that for systematic terminology work, they are compiling their own corpora. When ask about publishing, only few participants reported, that they are publishing their own corpora, mainly on their own websites. None of the participants published their corpora in a data repository. Most of the interview partners do not publish their corpora, because they only see them as internal working material towards the final product, the terminological resource. Furthermore, often these corpora are not annotated with part of speech, because it would be too time consuming to add them, and therefore the corpora are also considered by the creators less valuable for others and are therefore not shared.

It can be seen, that most institutions, creating terminological resources have also other valuable language resources, that are not yet shared with a wider community but could be of interest for the CLARIN and also DARIAH community. So, RIs could promote in this community the value of sharing and reusing language resources, and that language resources do not always need to be annotate with for example Part of Speech Tagging, in order to be valuable for certain user groups.

During the interviews also challenges when creating their own corpora were mentioned. For example, the lack of not enough IT support. In these cases, CLARIN could provide training and training materials to terminologists, how to create and maintain corpora, so that terminologists are capable to do it on their own without or with minimal support from IT units.

Another topic that was discussed in the interviews were data repositories. The use of depositing services offered by RIs for publishing terminological data, i.e., to publish a terminological dictionary or the data export of a terminology database in a data repository, was not a very common practice at the time of the interviews. In fact, only one participant mentioned, that they deposit their data, where copyright allows it, in a national CLARIN repository. This is in line with the findings in Wissik (2024: 110) that the use of data repositories, in general, is not a widespread practice in this community. So, in this respect, awareness raising regarding the use of data repositories, and the benefits of it, e.g., sustainability of the data, adhering to the FAIR data principles, additional dissemination channel for the data in this community would be needed, in both academic and non-academic settings. This could be done together by the RIs through dedicated CLARIN K-Centres, and through dedicated DARIAH WGs such as DARIAH WG on "Lexical Resources" and "Research Data Management".

A part of the interview was also dedicated to networks and collaborations with special focus on CLARIN and/or DARIAH. All of the interview participants were very active in relevant networks and associations regarding terminology or specific languages, regardless if they were academic or non-academic institutions. Most non-academic institutions also mentioned, that they have collaborations with academic institutions such as universities. However, most participants from non-academic institutions did not have collaborations with RIs and some of them were not familiar with CLARIN and DARIAH, even though their institutions were located in a member country or the country had at least a K-Centre (in case of CLARIN). These results are not so surprising, as the priority within the RIs so far was to reach out to academic users and to broaden the academic user base. However, recently RIs started to engage with non-academic communities as well, such as the public sector (e.g., Lyding et al., 2022). In the case of CLARIN, with the help of dedicated K-Centres, training materials could be created, that specifically target the terminology community. Furthermore, an ambassador from the terminology community within the public sector could be used to get engage with others in the public sector. By recruiting ambassadors also from the public sector, the already existing and successful CLARIN ambassadors programme could be used to reach out to the public sector. Furthermore, CLARIN and DARIAH could engage with this community via terminology or language associations, where they are members. Furthermore, terminology communities that already benefit from RIs like in Norway (Andersen & Gammeltoft, 2022) could be used as success stories to highlight the benefits of RIs in terminology work. Another finding of the analysis was, that interviewees that were aware of CLARIN and/or DARIAH, reported, that the specific academic unit involved in terminology work was not having links with these RIs. It is clear, that institutions such as universities are complex but it could be worthwhile to investigate, how to best target relevant users within an institution that is already part of an RI consortium.

More theoretically, it would be beneficial to integrate research infrastructures such as European Research Infrastructure Consortia (ERICs) and also similar initiatives such as the European Digital Infrastructure Consortia (EDIC)[10] into the terminology infrastructure models proposed by Galinski (1998) or Galinski & Giraldo (2023).

6 Concluding remarks

This contribution has discussed the potential role of Research Infrastructures such as CLARIN and DAIRAH in practical terminology work and explored the actual role of those RIs in practical terminology work in institutional settings by analysing 15 recorded expert interviews with a qualitative approach. To sum up, the case study has shown, that the role of RIs in terminology work in institutional settings has potential but is still expandable. Due to the small size of the sample, it is difficult to genialize the results. However, since the 15 expert interviews covered most current terminology approaches and most common scenarios where terminology work is done in institutional settings, the study contributes to our understanding of the current relation between Research Infrastructures and the terminology community, especially in institutional settings.

Furthermore, the contribution has discussed possible measures how CLARIN and DARIAH could engage more with the terminology community, for example by involving more dedicated CLARIN K-Centres and DARIAH WGs in awareness raising and training measures. The contribution also discussed measures on how to specifically target potential new user groups in the public sector. One suggestion was to expand the CLARIN ambassador programme by recruiting also ambassadors from the public sector. So, for examples, stakeholders in terminology work for example in public administration could act as ambassadors to engage with non-academic communities who could benefit from the data, services

[10] One example of an EDIC is the Alliance for Language Technologies EDIC (ALT-EDIC). More information can be found here https://alt-edic.eu/about-us/.

and knowledge provided by RIs. In this contribution, we have only discussed language resources and depositing services and the sharing of expertise as possible connecting points between the terminology community and RIs. However, also other connecting points could be discussed such as digital tools. Several tools to manage, edit and visualize data play an important role in terminology work (e.g., terminology management systems, corpus management tools, term extraction tools) which could be also a possible area of interaction with research infrastructures. Another area of possible interaction could be the use of AI and LLMs in terminology work especially for less-resourced languages: There the CLARIN K-Centre for Terminology Resources and Translation Corpora and the CLARIN K-Centre for LLMs4SSH could collaborate.

Moreover, a reevaluation of existing terminology infrastructure models and integration of research infrastructures and similar constructs into the terminology infrastructure models is recommended.

Acknowledgement

This contribution was supported in part by the BMBWF's funding programme Cooperation Development Research by the OeAD (Project No: KOEF 09/2019) and CLARIAH-AT.

References

Andersen, G, & Gammeltoft, P. (2022). The Role of CLARIN in Advancing Terminology: The Case of Termportalen – the National Terminology Portal for Norway. In Fišer, D. and Witt, A. (eds.) CLARIN: The Infrastructure for Language Resources, Berlin, Boston: De Gruyter, 249–274.

CLARIN: The Infrastructure for Language Resources, ed. by Darja Fišer and Andreas Witt, 249–274. Berlin, Boston: De Gruyter.

Budin, G. (2015). Digital Humanities, Language Industry, and Multilingualism: Global Networking and Innovation in Collaborative Methods." In: CIUTI-Forum 2014: Pooling Academic Excellence with Entrepreneurship for New Partnerships. 423 – 448, Lausanne.

Bowker, L. (1996). A Corpus-Based Approach to Terminography. *Terminology, 3*(2), 27–52.

Chiocchetti, E. & Ralli, N. (2013). Let's do it together: Instances of cooperation in terminology work: Roles, tools, needs and difficulties. In Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria, 515 – 524.

DARIAH (2025). Mission & Vision. In DARIAH Website. https://www.dariah.eu/about/mission-vision/

DARIAH (2025a). DARIAH in a Nutshell. In: DARIAH Website. https://www.dariah.eu/about/dariah-in-nutshell/

De Jong, F. Van Uytvanck, D., Frontini, F., Van den Bosch, A., Fišer, D. and Witt, A. "Language Matters". In Fišer, D. and Witt, A. (eds.) CLARIN: The Infrastructure for Language Resources, Berlin, Boston: De Gruyter, 31–58.

Doel, W. van den and Maes, K. 2012. Social Sciences and Humanities: Essential Fields for European Research and in Horizon 2020. League of European Research. Accessed January, 17 2025 https://www.leru.org/ files/Social-Sciences-and-Humanities-Essential-Fields-for-EuropeanResearch-and-Horizon-2020-Full-paper.pdf

ELRC (2022). AI for a Multilingual Europe. Why Language Matters. ELRC White Paper. Saarbrücken: Research Center for Artificial Intelligence (DFKI).

Fišer, D., Lenardič, J., Erjavec, T. (2018), CLARIN's Key Resource Families. In N. Calzolari (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan.

Galinski, C. (1998). Terminology infrastructures and the terminology market in Europe. *TRANS Internet-Zeitschrift für Kulturwissenschaften, 0*. https://www.inst.at/trans/0Nr/galinski.htm

Galinski, C. & Giraldo, S. (2023). Mongolian Infrastructure Layers. Presentation at the Project Meeting (Project Nr. KOEF 09/20), October 2023, Vienna.

Galinski, C. & Lusicky, V. (2024). Empowering Knowledge Societies Using Terminological Approaches – How Top-Down Policies Can Meet Bottom-Up Application. Book of Abstracts of the International Conference “Terminology in Scientific and Technological Development (TSTD) – 2024, September 13, 2024, Ulaanbaatar, Mongolia, 7–8.

Goosen, T., & Eckart, T. (2014). Virtual Language Observatory 3.0: What’s New? Proceedings of the CLARIN Annual Conference 2014, Soesterberg, Netherlands.

Gius, E. et al. (2023). *CATMA 7 (Version 7.0)*. Zenodo. DOI: 10.5281/zenodo.1470118.

Guest, G., Bunce, A. & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability”. *Field Methods* 18 (1), 59–82.

König, A., Barbot, L., Grisot, C., Michael Kurzmeier, Gray, E. J. (2024). SSH Open Market Place and CLARIN. In Lindén, K., Kontino, T. Niem, J. (eds.). Selected Papers from the CLARIN Annual Conference 2023.

Körner, E., Eckart, T., Helfer, F. & Uwe Kretschmer, U. (2024). Federated Content Search: Advancing the Common Search Infrastructure. In Kontion, T. & Vandeghinste, V. (eds). CLARIN Annual Conference Proceedings 2024, 15 – 17 October 2024 Barcelona, Spain, 41–45.

Kuchartz, U. (2014). *Qualitative Text Analysis: A Guide to Methods, Practice and Using Software*. London: Sage.

Lyding, V., Stemle, E., König, A. (2022). Collaborating on Language Resource Infrastructures with Non-ResearchPartners: Practicalities and Challenges. In Monica Monachini and Maria Eskevich (eds.). Selected Papers from the CLARIN Annual Conference 2021, 88–100.

Maroto, N. (2024). TeresIA. Spanish Access Portal to Terminologies and Artificial Intelligence Services. In Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024), Granada, Spain, June 27-28, 2024.

Meuser, M., & Ulrike Nagel, U. (1991). “Experteninterviews – vielfach erprobt, wenig bedacht.” In Gerz, D. & Karaimer, K. (eds.) Qualitative-empirische Sozialforschung. Konzepte, Methoden, 441–471. Opladen: Westdeutscher Verlag.

Moulin, C., Nyhan, J., Ciula, A., Kelleher, M., Mittler, E., Tadić, M., Ågren, M., Bozzi, A., Kuutma, K., (2011) Research Infrastructures in the Digital Humanities. ESF Science Policy Briefing 42. Strasbourg: European Science Foundation (ESF).

Pearson, J. (1998). *Terms in Context*. Amsterdam: Benjamins.

Pilke, N., Nissilä, N. & Landqvist H. (2021). Organising terminology work in Sweden from the 1940s onwards. Participatory expert roles in networks. *Terminology Terminology, 27(1),* 80–109.

Scharnhorst, A. Morselli, F. Admiral, F. (2019). DARIAH Working Groups Policy Statement 2019, v05. DARIAH ERIC. https://www.dariah.eu/wp-content/uploads/2019/09/DARIAH-Working-Groups-Policy-Statement_v5.pdf

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In N. Calzolari (Ed.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012, 1029–1034. European Language Resources Association (ELRA).

Wissik, T. (2025). Impact of automatic term extraction on terminology work: a qualitative interview study in institutional settings. *Terminology 31(1)*, 110–135.

Wissik, T. (2024). Dimensions of sustainability in terminology practices in institutional settings. *Terminology Science & Research*, 27, 93–116.

Wissik, T. (2022). Research Infrastructures and Lexicography: An European Perspective. *Mongolian Terminology Studies,* 133–43. Ulaanbaatar: Mongolian Academy of Sciences.

Wissik, T. & T. Declerck, T. (2020). Using an Infrastructure for Lexicography in the Field of Terminology. *Terminologie & Ontologie: Théories et Applications Actes de la conférence TOTh 2019,* 365–379. Chambéry: Presses, 365–379.

Žagar Karer, M., & T. Fajfar, T.(2023). Terminological problems of terminology users: Analysis of questions in terminological counselling service on the Terminologišče website. Terminology 29(2): 78–102.

Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset

Steinþór Steingrímsson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
steinthor.steingrimsson@arnastofnun.is
einar.freyr.sigurdsson@arnastofnun.is

Abstract

Multiword expressions (MWEs) are generally problematic for machine-translation systems. In this paper, we (i) describe a set, available on CLARIN-IS, of appr. 1,000 idiomatic MWEs which have been translated into English; (ii) use the set as a template for a hidden evaluation set, to be used in a new leaderboard for Icelandic language technology, and (iii) evaluate – using both automatic and manual approaches – four MT systems' abilities to translate MWEs from Icelandic to English using both datasets. We find that traditional transformer-based MT systems evaluated commonly fail when translating idiomatic expressions, while LLMs do much better.

1 Introduction

Multiword expressions (MWEs) are a frequent phenomenon in natural language and speech. Proper handling of MWEs is important for various natural language processing (NLP) tasks, such as machine translation (MT), bilingual lexicon induction and information extraction. It is difficult to provide clear boundaries for what constitutes a MWE and what does not. The term can be used to describe fixed or semi-fixed phrases, compounds, idioms, phrasal verbs or collocations – in general, any sequence of words that acts as a single unit on some level (Calzolari et al., 2002).

In this paper, we introduce a set of approximately 1,000 Icelandic MWEs,[1] along with their translations into English as well as structured information about their usage. We consider all the MWEs in our dataset to be *idiomatic expressions*, i.e. idioms with an intended meaning that diverges from the literal meaning of the words constituting the expression, and therefore, they usually cannot be translated word for word.

Machine-translation systems generally do not handle MWEs well, and even though they are an important part of generating fluent translations, they can be a blind spot for traditional automatic evaluation approaches, such as BLEU (Papineni et al., 2002) or chrF++ (Popović, 2017). This applies especially in cases where there is more than one "right" answer, as the traditional lexical metrics cannot identify what goes wrong in a translation. The Icelandic MWE dataset was compiled for use with MT, and can be used either to augment training sets with sentence pairs containing common idiomatic expressions, or for evaluating the capabilities of MT systems to translate such expressions. We show how the dataset can be used to evaluate the capabilities of four MT systems to translate MWEs, by evaluating the systems in three different ways: using traditional automatic approaches, using automatic evaluation of MWE translations, and by manually evaluating the output.

For an accurate evaluation of an MT system capabilities, the evaluation data should not be used for training the system. Large language models (LLMs), which demand enormous amounts of data, are primarily trained on web-crawled data. When a dataset is made available online for a prolonged period of time, it becomes increasingly likely that it has been gobbled by spiders crawling the web for LLM training data. Therefore, it can be said that openly available evaluation sets start to rot, meaning that they become increasingly untrustworthy, as soon as they are put online. To still be able to gauge the capabilities of MT systems in translating MWEs from Icelandic into English, we tackle this problem by using the publicly available dataset from the CLARIN-IS repository as a template for another dataset, Hidden Idiomatic Expressions for Machine Translation Evaluation (HIDEMATE). In order to hinder data

[1] http://hdl.handle.net/20.500.12537/275 (Halldórsson et al., 2022).

Steinþór Steingrímsson and Einar Freyr Sigurðsson 2024. Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 13–24. DOI: https://doi.org/10.3384/ecp216

leakage, HIDEMATE will not be made publicly available and will only be used as a part of a machine translation (MT) evaluation suite to be deployed on a leaderboard for Icelandic language technology, set to open in the fall of 2025. The dataset will be controlled by the Árni Magnússon Institute for Icelandic Studies (AMI), stored on internal servers and evaluations using the datasets will only be carried out by AMI staff.

In the paper we will describe how we evaluate MT systems using these datasets as well as carrying out initial evaluations. Section 2 discusses related work. In Section 3 we briefly describe the MWE datasets whereas Section 4 discusses MT evaluation. We demonstrate our results in Section 5, discuss future work in Section 6 and conclude in Section 7. In Section 8, we discuss potential limitations of this work.

2 Related Work

In recent years, idiomatic expressions (and MWEs in general) have been the focus of much work in natural language processing (NLP), not least in MT research. Stap et al. (2024) introduce a dataset containing one thousand idioms in context with translations for three translation directions, English→German, German→English and Russian→English, and use it to evaluate LLMs fine-tuned using their fine-tuning approach. Tang (2022) present a parallel English dataset of Chinese idioms, and Ármannsson et al. (2024) introduce a dataset for evaluating English→Icelandic idiomatic expressions and an approach for how to go about the evaluation.

In Ármannsson et al.'s (2024) submission to the WMT24 test suite subtask, they focus on evaluating the capabilities of participating models in translating idiomatic expressions from English to Icelandic. This is the first published attempt to systematically compare translational capabilities of automatic systems working with Icelandic in terms of MWEs. They use two lists of words for each idiom, one with literal translations, which are a negative match when translating idiomatic expressions, and one with a positive match, which contains the words of the most likely idiomatic expression. In our work we use this latter type of list, but not the former. The idiom *taka <einhvern> til bæna*, lit. 'take <someone> to prayers', can be translated at least as 'read <someone> the riot act' and 'take <someone> to task'. For a successful translation, according to our automatic evaluation, we require the translation to either include all of the words *read*, *riot*, *act* or all of the words *take*, *to*, *task*.

3 Collecting the Multiword Expressions

The set of multiword expressions, distributed on the Icelandic CLARIN repository,[2] contains approximately 1,000 Icelandic idioms processed from the ISLEX dictionary (Úlfarsdóttir, 2014). They are listed with their English idiomatic equivalent and literal meaning in both languages, as well as example sentences and keywords. The idioms are, in most cases, syntactically mobile, which is why case information is included.

The idioms were processed from a list of 4,000 MWEs in the ISLEX database. The idioms are ordered alphabetically according to the first keyword of each idiom and each line contains the following categories: 1) Icelandic idiom; 2) English equivalent; 3) Meaning of the Icelandic idiom; 4) Meaning of the English idiom; 5) An example sentence with the Icelandic idiom; 6) An example sentence with the English idiom; 7) An example sentence with the meaning of the Icelandic idiom; 8) An example sentence with the meaning of the English idiom; 9) Keywords in the Icelandic idiom, lemmatized (in some cases in the plural). Table 1 exhibits an example from the dataset where all nine columns are shown. The first four categories contain type information, cf. for example, the idiom *rétta <einhverjum> hjálparhönd* 'lend <someone> a helping hand', which is listed as follows: <NP1-nom> rétta <NP2-dat> hjálparhönd; <NP1> lend <NP2> a helping hand; <NP1-nom> hjálpa <NP2-dat>; <NP1> help <NP2>.

Where more than one English equivalent, translation or sense are possible, alternatives are separated with a pipe symbol. Keep in mind that there is not always a 1-1 relation between the example sentences. For example, the Icelandic idiom *Það er fokið í flest skjól* is translated as 'We're at the end of our tether', where the Icelandic expletive *það* 'it, there' makes way for a personal pronoun in English. The use of other symbols in the file is as follows: alternatives within the same segment are separated with a slash (/),

[2] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/275

Source idiom <NP1-nom> rétta <NP2-dat> hjálparhönd	**Target idiom** <NP1> lend <NP2> a helping hand
Source sense <NP1-nom> hjálpa <NP2-dat>	**Target sense** <NP1> help <NP2>
Source idiomatic example Sigurður rétti Guðmundi hjálparhönd	**Target idiomatic example** Sigurður lent Guðmundur a helping hand
Source sense example Sigurður hjálpaði Guðmundi	**Target sense example** Sigurður helped Guðmundur
Keywords hjálparhönd	

Table 1: An example of an idiom in our dataset.

as in, e.g., the idiom *vera klár/tilbúinn í slaginn* ('be ready to rumble'), and optional parts of idioms are in parentheses, e.g. the idiom *bretta upp ermar(nar)* ('roll up one's sleeves') or *vera sjálfs sín(s) herra* ('be one's own boss').

There are a few examples of duplicate lines in the file with respect to the source idiom, but only in cases where the respective meaning can be considered twofold, as for example in the idiom *ganga ekki heill til skógar*, which can either refer to physical or mental health, i.e. 'be under the weather' (physical) or 'not be playing with a full deck' (mental).

Users of the dataset will note that the Icelandic male names *Sigurður* and *Guðmundur* are used as actors in the example sentences. This is for the sole reason that they have different inflectional forms for each case (nom. *Sigurður/Guðmundur*, acc. *Sigurð/Guðmund*, dat. *Sigurði/Guðmundi*, gen. *Sigurðar/Guðmundar*).

For the MT evaluation, we process the data in a slightly different way than in the distribution file. We number each segment, and while we only use the example phrases and their translations, where there are alternatives within the segments we generate all possible pairs. The generated pairs then get the segment number and the evaluation results are weighted so that all segments in the dataset have the same weight in the final score. Furthermore, we add a list of words that should be included in the MT translation of the idiom, and that list is used for the automatic evaluation of idiom translation. The processed data, along with all scripts, are made available on GitHub.[3]

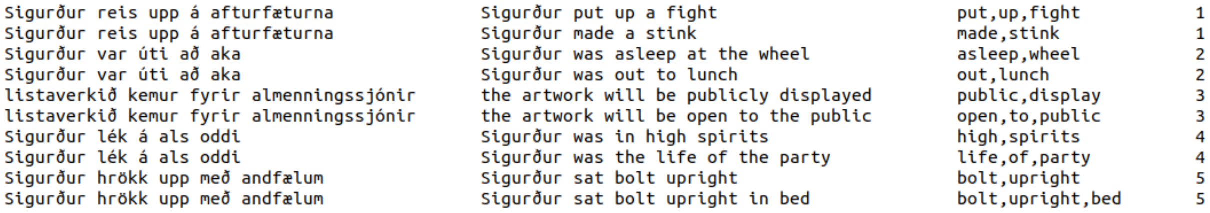
HIDEMATE (Hidden IDiomatic Expressions for MAchine Translation Evaluation) is processed in the same way and only contains data fields necessary for evaluation. Figure 1 shows examples of how the dataset from the CLARIN-IS repository is prepared for evaluation. The first column contains the source sentence in Icelandic. The second column contains a reference translation in English. Note that for each source sentence there can be one or more reference translations. The keywords that should be included in an MT translation of the idiom are in column 3. Column 4 has an id for the source sentence, as sometimes the same source sentence has multiple translations.

4 Evaluating Machine-Translation Systems

When choosing which MT system to use for a given task, the ability to translate MWEs can be a deciding factor. When we evaluate MT systems' ability to translate idioms from one language to another, we may want to punish incorrect literal translations. An example would be if an MT system would translate *kick the bucket*, meaning 'die', from English to, say, Icelandic and we would find both the word *sparka* 'kick' and *fata* 'bucket' as *sparka í fötuna* can only mean that someone strikes a bucket with their feet.

It is therefore important to be able to test these capabilities. To this end, we run three evaluation

[3] https://github.com/stofnun-arna-magnussonar/IdiomaticExpressions



```
Sigurður reis upp á afturfæturna           Sigurður put up a fight                    put,up,fight          1
Sigurður reis upp á afturfæturna           Sigurður made a stink                      made,stink            1
Sigurður var úti að aka                    Sigurður was asleep at the wheel           asleep,wheel          2
Sigurður var úti að aka                    Sigurður was out to lunch                  out,lunch             2
listaverkið kemur fyrir almenningssjónir   the artwork will be publicly displayed     public,display        3
listaverkið kemur fyrir almenningssjónir   the artwork will be open to the public     open,to,public        3
Sigurður lék á als oddi                    Sigurður was in high spirits               high,spirits          4
Sigurður lék á als oddi                    Sigurður was the life of the party         life,of,party         4
Sigurður hrökk upp með andfælum            Sigurður sat bolt upright                  bolt,upright          5
Sigurður hrökk upp með andfælum            Sigurður sat bolt upright in bed           bolt,upright,bed      5
```

Figure 1: Examples of sentences prepared for evaluation.

experiments. First, we simply evaluate the MT output using traditional automatic approaches. We apply the common evaluation metrics BLEU, chrF++ and COMET. Second, we devise a simple automatic approach that classifies translations in two groups: translations likely to have correctly handled the MWE and translations that failed to do so. Third, we manually evaluate all translations to be able to confirm or reject the adequacy of the automatic approach.

4.1 MT Systems

We compare four systems capable of Icelandic to English machine translation: two LLMs and two dedicated translation systems. All these systems are publicly available and as of early 2025 are all commonly used by the Icelandic-speaking population. These systems are OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, Google Translate and the translation system on M.is.

4.1.1 GPT-4o

ChatGPT caught the world by storm in late 2022. The chatbot was built on a large language model, GPT-3, taking advantage of reinforcement learning with human feedback (RLHF) to create impressive conversational abilities. Since then multiple improvements have been made as well as different versions of the underlying language model. Among the many uses of the system is automatic translation, as the underlying LLMs are trained on data in multiple languages and have cross-lingual capabilities. In our experiment we use GPT-4o, which the OpenAI website states is their "versatile, high-intelligence flagship model". [4] We employ the default GPT-4o version at the time of our experiment, `gpt-4o-2024-08-06`.

We accessed the model through API to carry out a three-shot translation. The examples given to the system were the ones used to collect translations from LLMs for the WMT 24 general translation task (Kocmi, Avramidis, Bawden, Bojar, Dvorkovich, Federmann, Fishel, Freitag, Gowda, Grundkiewicz, Haddow, Karpinska, Koehn, Marie, Murray, et al., 2024).[5] We set the temperature to a rather low number, 0.2, for more deterministic output. The request template is shown in Figure 2.

4.1.2 Claude 3.5 Sonnet

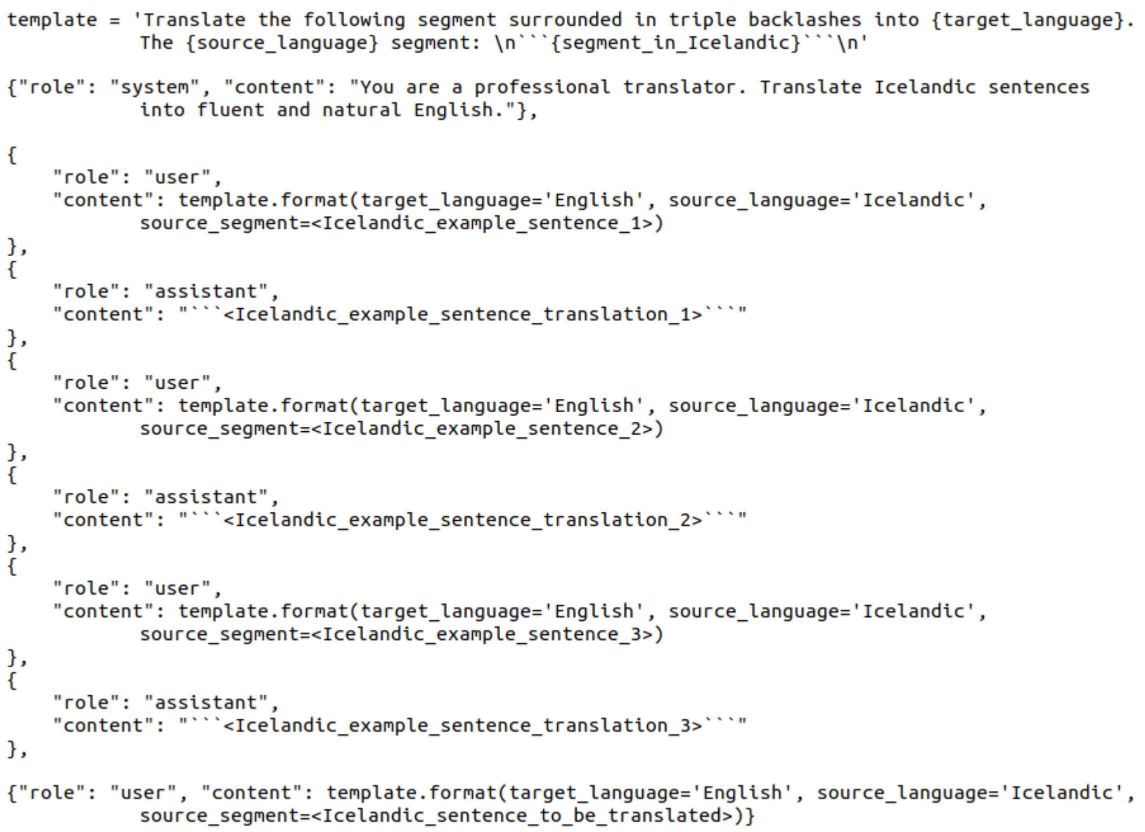
As well as being widely used, Anthropic's Claude 3.5 was the highest scoring openly available system in the WMT 24 General Translation Task (Kocmi, Avramidis, Bawden, Bojar, Dvorkovich, Federmann, Fishel, Freitag, Gowda, Grundkiewicz, Haddow, Karpinska, Koehn, Marie, Monz, et al., 2024) for the English→Icelandic translation direction. It is thus an obvious choice for our MWE translation evaluation for the opposite direction, Icelandic→English.

We employ Claude 3.5 Sonnet, which is the "most intelligent model" according to Anthropic's website.[6] For our experiments we select the default Claude 3.5 Sonnet version at the time of our experiment, `claude-3-5-sonnet-20241022`. As with the OpenAI model we set the temperature to 0.2, and format the request template in the same way as before, shown in Figure 2.

[4] https://platform.openai.com/docs/models#gpt-4o, accessed on February 6th, 2025.

[5] Made available here: https://github.com/wmt-conference/wmt-collect-translations

[6] https://docs.anthropic.com/en/docs/about-claude/models, accessed on February 6th, 2025.



````
template = 'Translate the following segment surrounded in triple backlashes into {target_language}.
           The {source_language} segment: \n```{segment_in_Icelandic}```\n'

{"role": "system", "content": "You are a professional translator. Translate Icelandic sentences
           into fluent and natural English."},

{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_1>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_1>```"
},
{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_2>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_2>```"
},
{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_3>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_3>```"
},

{"role": "user", "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_sentence_to_be_translated>)}
````

Figure 2: The template for the requests to gpt-4o and Claude 3.5 Sonnet APIs.

4.1.3 Google Translate

For many people, Google Translate has been the go-to MT system for Icelandic since the language was added in 2009. Google Translate is fast and accessible and widely used. It has commonly been used by the Icelandic MT research community to compare their models to, ever since the first paper on Icelandic MT using statistical and neural methods was published by Jónsson et al. (2020). For our experiment we access Google Translate through an API.[7]

4.1.4 m.is

m.is[8] is a dictionary and language technology portal, which opened in mid-2024 and is hosted by AMI. It is aimed at students of Icelandic, native speakers of Icelandic, as well as second language learners. The website gives access to a contemporary dictionary of Icelandic, a database of inflections, Icelandic→English and Icelandic→Polish bilingual dictionaries as well as an MT system translating both ways between Icelandic and English.

The MT system on m.is is based on the submission to the WMT24 general news translation task by Jasonarson et al. (2024). While the WMT task was only to translate English→Icelandic, a model was also trained in the other direction, Icelandic→English, to generate backtranslations for training the final system.

The submission was based on four transformer (Vaswani et al., 2017) models of various sizes, with each of them generating multiple candidates. The final translation is selected using COMET (Rei et al., 2020). For the backtranslations from Icelandic to English only one model was trained, a Transformer$_{\text{BIG}}$ model of approximately 200M parameters. This model was then deployed to carry out

[7] Google Translate was used to translate the sentences on February 6, 2025.

[8] https://m.is

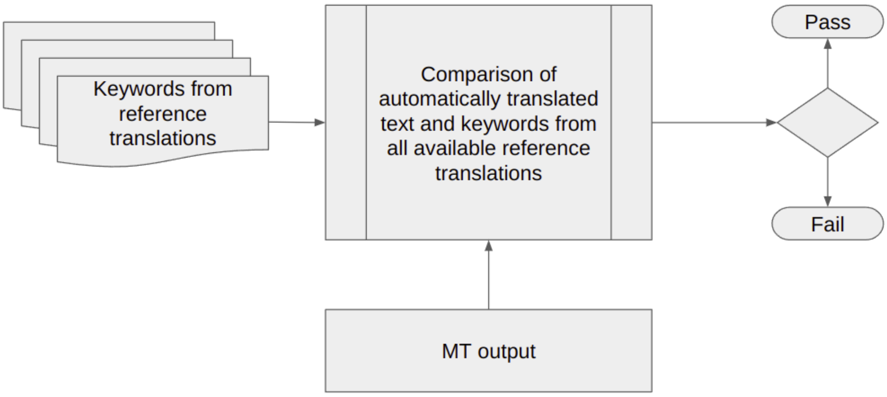

Keywords from reference translations
Comparison of automatically translated text and keywords from all available reference translations
Pass
Fail
MT output

Figure 3: The AutoIE process.

Icelandic→English translations on m.is, by generating five candidates, and having COMET selecting the best one for output.[9]

4.2 Automatic Evaluation Approaches

In order to make a general comparison of the MT systems used, we calculated BLEU, chrF++ and COMET scores for translations of all sentences in the two datasets. The scores for BLEU[10] and chrF++[11] were calculated using Sacrebleu (Post, 2018). Sacrebleu signatures are given in footnotes and results reported in Table 2 for the dataset from the CLARIN-IS repository and in Table 4 for HIDEMATE.

4.2.1 Automatic Idiom Evaluation

For our experiments, we devised a simple automatic approach, designated AutoIE (Automatic Idiom Evaluation), to gauge how well the MT systems managed to process the idiomatic expressions. Each machine-translated output is assigned a pass or a fail. The translation gets a pass if it contains all content words of the translation in the dataset. For example, for the sentence *Sigurður fékk sér kríu*, translated in the dataset as 'Sigurður took a nap', the MT translation has to contain the words 'took' and 'nap' to receive a pass. If it does not contain both words, it is assigned a fail. Figure 3 shows how the AutoIE algorithm is passed a translation candidate from an MT system and provided with one or more reference translations and keywords for each one. If all the keywords for any of the reference translations is found in the MT translation candidate, the system assigns a pass. Otherwise it assigns a fail.

Note that each source sentence can have multiple different acceptable translations, but they need to be defined in the reference set. The reference file for the dataset from the CLARIN-IS repository has been made available in the project's GitHub-repository.[12]

4.2.2 BLEU

The BLEU score, introduced in 2002 (Papineni et al., 2002), was the de facto standard for MT Evaluation for twenty years and only in the early 2020s have other metrics come to topple its dominance, with 2024 being the first year that BLEU was not used as one of the evaluation metrics for the WMT general translation task. BLEU compares an automatically translated text to one or more human-written reference translations. It does so by checking how many n-grams in the candidate translation appear in the

[9] The m.is translation engine was used to translate the sentences on February 6, 2025.

[10] BLEU|nrefs:3|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[11] chrF2|nrefs:3|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

[12] https://github.com/stofnun-arna-magnussonar/IdiomaticExpressions

reference translation, with the default being all n-grams up to 4-grams. It also applies a brevity penalty for translations that are much shorter than the reference.

While BLEU is easy to use and can give a good idea of the likeliness of a candidate translation and the reference, over the years it has been criticized for only measuring surface features while not considering meaning or context. It also does not evaluate syntax, grammar or sentence fluency and cannot evaluate long-range relationships between words. It depends on reference translations, and a single reference may not cover all possible valid translations as BLEU does not account for synonyms or paraphrases.

These drawbacks often make it hard to use BLEU to evaluate translations rich in idiomatic expressions as they are likely to have multiple valid translations, which can be idiomatic or literal in some cases, but not in others. In our experiment we will calculate BLEU scores for the translations and compare it to the AutoIE score as well as manual evaluation.

4.2.3 chrF++

chrF++ is another surface-level evaluation metric for MT. It measures precision and recall of character n-grams, which makes it more suitable when translating into morphologically rich languages, such as Icelandic. chrF++ also considers word n-grams, which improves its correlation with human judgment, but it is not as strict as BLEU in penalizing word reordering. Like BLEU it has multiple drawbacks related to it not considering meaning or context or properties appertaining to that.

4.2.4 COMET

Unlike the traditional n-gram based metrics, COMET uses neural models to compare an automatically generated translation to the source sentence and a reference translation. These inputs are processed by multilingual transformer models to generate semantic representation. These models have been trained on human judgments of MT quality, such as the publicly available human evaluations from the WMT evaluation campaigns pertaining to the general translation tasks which take place each year.

COMET has come to be quite popular in recent years, especially when evaluating language pairs for which there are substantial resources of training data. Multiple different COMET models have been made available. For evaluating our translation experiments we use wmt22-comet-da.[13]

4.3 Manual Evaluation

To assess whether our automatic approach is useful a sample of translations from each set, 220 sentences from the dataset from the CLARIN-IS repository and 220 sentences from HIDEMATE, for each of the four MT systems, were evaluated by a linguist whose task was only to look at the MWE and assess whether it was translated correctly. The evaluator would select one of three options: *Correct translation*, *Incorrect translation* and *Unusual translation but can be understood*. It can be difficult to determine whether a translation is correct and this last category, *Unusual translation but can be understood*, was sometimes used if the evaluator was uncertain, for example, if he found a translation to be fairly good but he did not feel it captured the meaning perfectly. Also in this category were translations where the meaning was captured but the wording was not entirely what one would expect. The linguist was only to look at the MWE and disregard all other possible errors in the translation. The results for the manual evaluation for the CLARIN-available dataset are given in Table 3 and the manual evaluation for HIDEMATE in Table 5.

Upon inspecting the outputs, we find that idioms that can be translated word by word from Icelandic into English, such as *Sigurður var úlfur í sauðargæru* (‘Sigurður was a wolf in sheep’s clothing’) and *Sigurður bjargaði andlitinu* (‘Sigurður saved face’; lit. ‘Sigurður saved **the** face’), are most likely to be translated correctly. Idioms that require translating into an idiom that has the same meaning but uses a different metaphor are less likely to be translated correctly. Example of that could be *Sigurður er eldri en tvævetur*, literally ‘Sigurður is older than two winters old’, which would normally be translated into ‘Sigurður was not born yesterday’, or an idiom containing words where the most common sense is not the one carried in the idiom, such as *Sigurður rak lestina* (‘Sigurður trailed behind’) which contains the word *lest*, perhaps most commonly meaning a locomotive train and translated as ‘train’.

[13] https://huggingface.co/Unbabel/wmt22-comet-da

MT System	BLEU	chrF++	COMET	AutoIE (%)
gpt4-o1	29.0	57.7	0.6559	25.3
Claude 3.5 Sonnet	29.9	53.0	0.6572	26.3
Google Translate	19.4	51.6	0.6166	15.8
m.is	23.2	52.0	0.6434	14.8

Table 2: Automatic evaluation of the four MT systems on the CLARIN-available dataset.

MT System	Correct (%)	Understandable (%)	Incorrect (%)
gpt4-o1	68.5	9.6	21.9
Claude 3.5 Sonnet	69.0	8.7	22.3
Google Translate	35.6	7.8	56.6
m.is	38.4	9.6	52.0

Table 3: Manual evaluation of the four MT systems on the CLARIN-available dataset.

MT System	BLEU	chrF++	COMET	AutoIE (%)
gpt4-o1	34.9	61.3	0.6867	35.5
Claude 3.5 Sonnet	36.8	57.9	0.6861	34.1
Google Translate	25.1	54.8	0.6405	21.8
m.is	26.3	52.9	0.6494	20.0

Table 4: Automatic evaluation of the four MT systems when translating HIDEMATE.

MT System	Correct (%)	Understandable (%)	Incorrect (%)
gpt4-o1	81.1	6.0	12.9
Claude 3.5 Sonnet	80.7	7.8	11.5
Google Translate	44.0	6.5	49.5
m.is	44.2	10.2	45.6

Table 5: Manual evaluation of the four MT systems on the HIDEMATE dataset.

5 Results

The LLMs scored highest on all the metrics, usually by a large margin. In the automatic evaluation, they translated over 25% of the idiomatic expressions in the CLARIN-available dataset correctly (see AutoIE in Table 2) and more than a third in the HIDEMATE dataset (see AutoIE in Table 4). We can see that there is a slight variation in the order of the systems by which metric is used. Claude always obtains the highest BLEU scores, while GPT4-o1 obtains the highest chrF++ scores. Similar variations can be seen when we compare Google Translate and m.is, with m.is scoring higher on BLEU and COMET, but Google Translate obtaining higher chrF++ scores on the HIDEMATE dataset and a very close chrF++ score on the CLARIN-available dataset. Google Translate also scores higher on the AutoIE metric.

This shows that the choice of evaluation metric is important when it comes to selecting an MT system for a given task. While all the metrics give the general picture, there is a considerable variation in how close the scores are between two system. We also see that when the traditional metrics show a difference between two systems, like for m.is and Google Translate in our experiment, this does not necessarily indicate that the “better” system is more adept in translating something like idiomatic expressions.

The results in the manual evaluation demonstrate that the LLMs get a lot higher score than Google Translate and m.is, which are both based on encoder-decoder models. gpt4-o1 is not far behind Claude3.5 Sonnet in the CLARIN-available dataset but scores slightly higher in the HIDEMATE dataset. In both cases though, the difference is negligible. While Google Translate and m.is translate considerably fewer idioms correctly than the LLMs, there is little difference between the two. m.is obtains a slightly higher number of correct translations for both datasets, but the difference between the two is more pronounced in the number of incorrect translations, where Google Translate does rather worse than m.is.

When we compare the manual evaluation to the results of AutoIE, we find that the automatic metric only accepts approximately half of what the human evaluator accepts. However, even though it does not give more accuracy than that, its failings seem to apply to all datasets and all MT systems, and gives very similar results, although Google Translate scores slightly higher than m.is on AutoIE, while m.is scores slightly higher on the manual evaluation. Overall the results indicate that the AutoIE metric gives useful information on the capabilities of translation systems when translating idiomatic expressions.

Our experiments also show that LLMs are considerably better than smaller transformer models in translating idiomatic expressions. The LLMs do produce less literal outputs, compared to the transformer models and this seems to be particularly true when dealing with idiomatic expressions. Our results are in line with previous research, for example Ármannsson et al. (2024), which found that while the best transformer-based models could be equally good or even better than LLMs in cases where a literal translation was called for, they were no match when it came to idiomatic expressions.

Finally, even though we do not intend to publish the HIDEMATE evaluation set or making it accessible in any way, it is important to be able to describe its content and give examples of the kind of data that is included in it. We modeled HIDEMATE on a dataset of idiomatic expressions that has been made available on CLARIN and ideally the two datasets should give very similar scores when evaluating the same system. While both datasets give results close enough to give the same general idea, there is quite some difference between the scores.

6 Future Work

In some cases, each Icelandic example is given only one translation in our datasets, although more translations may be valid. Adding additional valid translations for each example would be useful in order for automatic evaluations to align more closely to manual evaluation scores when the datasets are used with the AutoIE approach to evaluate the capabilities of MT systems to translate idiomatic expressions. By comparing the translations deemed correct in the human evaluation to the translations given in the datasets, we can add more valid translations. We intend to do so for future versions of the datasets in order to make them even more viable for automatic evaluation.

We have not divided our datasets into different types of MWEs, depending on how transparent or opaque they are. However, that might be helpful for determining what kind of idioms the MT systems are best at translating. For some opaque idioms, all the MT systems fail. An example is *Sigurður dró ýsur* which means that Sigurður nodded off or dozed but the MT systems translate it literally, as ‘Sigurður pulled/caught haddocks’. By categorizing the idioms with respect to transparency, we can make sure that the different datasets we are evaluating are as similar to each other as possible.

Our results show that there is great enough difference in two datasets to result in quite different scores. We want to investigate further why that is, and in a later version try to take care to remedy this difference.

Finally, we intend to use the openly available dataset introduced here as a supplemental data for MT training, and investigate if that will increase the capabilities of an MT system to translate idioms, as measured by our hidden dataset, HIDEMATE.

7 Conclusions

The evaluation results, and the result analysis, indicate that the traditional MT systems commonly fail when translating idiomatic expressions, while LLMs usually do not. Specialized evaluation sets, such as the one introduced in this paper, can be used to gauge the capabilities of different systems. The simple automatic approach introduced here provides results in line with a thorough manual evaluation, indicating that it may be sufficient to help in the selection of the best system in this regard, when needed.

8 Limitations

There are various potential limitations to our work and here we mention a few of them.

While we want to prevent data leakage by building a hidden evaluation set, HIDEMATE, we process the idioms from ISLEX which is available online. It could be the case that the ISLEX data is among the web-crawled texts in the LLMs’ training data. While that data is not formatted to be made available in sentence pairs, as an evaluation set is, the proximity of idioms to their explanations and translations in the online dictionary could help LLMs learn these exact idioms when trained on the web-scraped dictionary data.

It is not always clear cut what constitutes an MWE which is not an idiom and what is clearly an idiom. There may be a gray area there. When compiling our datasets we select the MWEs which we consider idiomatic, in some cases others might disagree.

In our manual evaluation, all four translations for a given sentence appear in the same order. This could lead to a bias if the worst model or models do something unusual while the best models do not, as the evaluator may lean towards what he can expect beforehand.

AutoIE is a word inclusion test, testing for a non-literal translation of idioms. It is potentially prone to recognising presence of the proper translation while, in fact, the words in the target language may not form a proper expression. While it does not seem to our evaluator to be a common phenomenon, estimating how common it is could give us more insight into the feasibility of our automatic approach.

Finally, while the evaluator is a native speaker of Icelandic and speaks English fluently, he is not a native speaker of English. That is a clear limitation when judging whether translations into English are correct or not. However, as he was the only evaluator, there should be internal consistency in his judgments.

Acknowledgments

We thank three anonymous reviewers for helpful comments which improved the paper. The making of the CLARIN-available dataset (Idiomatic Expressions (Icelandic and English) 22.09) was funded by the Ministry of Education, Science and Culture in Iceland as part of the Language Technology Programme for Icelandic 2019–2023. We thank our collaborators who worked on the dataset: Björn Halldórsson, Árni Davíð Magnússon and Finnur Ágúst Ingimundarson.

References

Ármannsson, B., Hafsteinsson, H., Jasonarson, A., & Steingrímsson, S. (2024). Killing Two Flies with One Stone: An Attempt to Break LLMs Using English-Icelandic Idioms and Proper Names. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 451–458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.31

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 1934–1940). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf

Halldórsson, B., Magnússon, Á. D., Ingimundarson, F. Á., Sigurðsson, E. F., Steingrímsson, S., Jónsdóttir, H., & Úlfarsdóttir, Þ. (2022). Idiomatic Expressions (Icelandic and English) 22.09 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/275

Jasonarson, A., Hafsteinsson, H., Ármannsson, B., & Steingrímsson, S. (2024). Cogs in a Machine, Doing What They're Meant to Do – the AMI Submission to the WMT24 General Translation Task. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 253–262). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.18

Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., & Loftsson, H. (2020). Experimenting with Different Machine Translation Models in Medium-Resource Settings. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Text, Speech, and Dialogue* (pp. 95–103). Springer. https://doi.org/10.1007/978-3-030-58323-1_10

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Murray, K., Nagata, M., Popel, M., Popovic, M., Shmatova, M., … Zouhar, V. (2024). Preliminary WMT24 Ranking of General MT Systems and LLMs. *arXiv preprint*. https://doi.org/10.48550/arXiv.2407.19884

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://www.aclweb.org/anthology/P02-1040

Popović, M. (2017). chrF++: words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 612–618). Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-4770

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://aclanthology.org/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP) (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Stap, D., Hasler, E., Byrne, B., Monz, C., & Tran, K. (2024). The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. 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. 6189–6206). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.336

Tang, K. (2022). PETCI: A Parallel English Translation Dataset of Chinese Idioms. *arXiv preprint*. https://arxiv.org/abs/2202.09509

Úlfarsdóttir, Þ. (2014). ISLEX – a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2820–2825). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/672_Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5999–6009). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Word Rain as a Service:
Making semantically structured word clouds available to everyone

Magnus Ahltorp
Nakajima Koen Research Institute
Stockholm, Sweden
magnus@nakajimakoen.org

Maria Skeppstedt
Centre for Digital Humanities
and Social Sciences Uppsala
Department of ALM
Uppsala University, Sweden
maria.skeppstedt@abm.uu.se

Abstract

The Word Rain text visualisation technique is a novel approach to the classic word cloud that uses word embeddings to make the visualisation useful for exploring the word content of a text or corpus. Downloading and running code for generating word rain visualisations can, however, be prohibitively difficult or cumbersome for non-technical users and for casual evaluation. These use cases would consequently benefit greatly from a streamlined interface. We have therefore collected everything needed for generating word rain visualisations in a web-based service, and made it available as a SWELANG CLARIN K-centre resource. The web service, as well as the code for generating word rains, is made available as open source. The web service is deployed at: https://wordrain.isof.se.

1 Introduction

Word clouds are widely used for informally illustrating the most prominent words in a text or corpus. Their popularity likely stems from their usefulness as a graphic design element, the simple, self-explanatory nature of the visualisation, as well as from the prevalence of easy-to-use tools for generating word clouds. The visualisation consists of a graph, where the words are displayed in a font size that reflects the prominence of the words. Typically, the more frequent a word in the text visualised, the larger the font size used for displaying the word. The words are often positioned in a random or alphabetical order, in a fashion that aims to maximize the visual design of the word cloud.

Traditional word clouds are, however, also criticised because of the lack of semantic relevance in the placement of words (Barth et al., 2014). The reader can therefore be misled into thinking there is a relationship between nearby words where there is none. The absence of a semantically meaningful word positioning also makes the word cloud visualisation unsuitable as a tool for practically exploring and analysing texts (Cao & Cui, 2016; Hicke et al., 2022). Since the traditional word cloud does not provide any guidance to where in the graph to zoom in to read words displayed in a small font size, the inclusion of less prominent words in the graph has no practical text exploration purpose – only a cosmetic function. In addition, typical text analysis tasks, e.g., to manually create semantic categories among the most prominent words in a text, or to compare the content of pairs of texts, are also very difficult to carry out when the words are positioned in a non-semantically relevant order.

Despite the critique directed towards the traditional word clouds as a tool for practically exploring and analysing texts, the word clouds are often used for these purposes (Hicke et al., 2022). We hypothesise that the aforementioned prevalence of easy-to-use tools for generating traditional word clouds – and the absence of similar tools for generating more practically useful visualisations – is an important reason for word clouds still being used for these tasks. The project described here therefore aims to direct researchers – as well as people in general – to the novel and more practically useful Word Rain text visualisation technique. By streamlining the process of generating word rains to the point where it is as easy to generate as a word cloud is today, our goal is to simplify the creation of text visualisations that can be used practically for exploring and analysing texts.

Magnus Ahltorp and Maria Skeppstedt 2024. Word Rain as a Service: Making semantically structured word clouds available to everyone. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 25–36. DOI: https://doi.org/10.3384/ecp216

We will here start by providing a background which describes the Word Rain text visualisation technique. Then, we will concretise the theoretical description by discussing example word rains generated using the Word Rain web service. Thereafter, we will provide a practical description for how to use the Word Rain web service to generate word rain(s) from one or several texts. We will then conclude with a summary of the work conducted and by outlining possible future directions for the Word Rain web service.

2 The Word Rain text visualisation technique

There are many tools for visualising prominent words in a text that build on animation and/or more or less advanced user interaction (Liu et al., 2015; Wang et al., 2018; Xie et al., 2024). While such functionality increases the possibilities to carry out different types of text exploration and text analysis tasks, it does not – as the word cloud – provide a simple, self-explanatory static graph that, for example, can be included in an article or printed on a poster. That is, these tools are not suitable replacements to the traditional word cloud.

Other approaches that extend the traditional word cloud by generating a static graph, and which can be used in the same contexts as the word cloud, typically focus on solving only one of the problems associated with the traditional version of the visualisation technique. There are, for instance, approaches that make it easier to compare pairs of texts, either by generating graphs that simultaneously visualise several texts or by generating series of graphs where a word is always given the same position in the graph (Burch et al., 2014; Cui et al., 2010; Diakopoulos et al., 2015; Herold et al., 2019; Lee et al., 2010). There are also other approaches where a visualisation similar to the traditional word clouds is generated, except that semantically similar words are positioned close to each other (often in combination with a semantically motivated colour coding), and that aim to solve the problem of word clouds not supporting a manual categorisation of prominent words (Barth et al., 2014; Wu et al., 2011; Xu et al., 2016).

With the Word Rain technique (Skeppstedt et al., 2024), in contrast, we propose a solution to all of the above-mentioned problems associated with traditional word clouds – a solution that still produces a static, self-explanatory graph as its output. The most important difference from the traditional word cloud is that the Word Rain technique uses the position on the x-axis, as well as the position on the y-axis for conveying information. The semantic relevance of the word positioning is achieved by letting a word's position on the x-axis represent the semantics of the word. The position is automatically determined by using multi-dimensional word embedding vectors that represent the words and reducing the vectors to one dimension. The position of a word on the y-axis is, instead, primarily based on word prominence. To avoid words overlapping when they have a similar position on the x-axis, less prominent words have to yield to more prominent ones. That is, the less prominent word "rains down" in the graph, until it reaches a lower y-position where its extension no longer overlaps with the extension of a more prominent word.

With this technique, words with a similar meaning will be positioned close to each other on the x-axis, creating (partly) vertical clusters of semantically similar words, with the more prominent ones typically at the top of the cluster. Thereby, the horizontal position of the words, and the clusters of similar words, can assist the user when manually creating categories of prominent words in a text. The semantic word positioning also has the effect that the more prominent words – which are displayed in a large font size – can guide the user to semantically interesting areas in the graph, where the user can zoom in and read semantically similar words displayed in a small font size. Finally, several word rains can be generated with the same embedding projection on the x-axis, making it possible to compare corpora. This comparison can be conducted on the level of each individual word, but semantic clusters of words can also be compared.

The word embedding model used for creating the semantic word positioning could either consist of a pre-trained word embedding model, or the user could train their own model, e.g., on the texts that are to be visualised. The multidimensional word embedding vectors are projected down to the one-dimensional x-axis using t-SNE dimensionality reduction.

There are two configurations for determining word prominence, one is *term frequency*, i.e. the frequency of the words in the text visualised, and the other is *term frequency–inverse document frequency*, i.e., a measure that down-weights words that occur in many texts. The word prominence is not only indicated by a word's position on the y-axis, but – similar to the traditional word cloud – its prominence is also used to determine the font size used for displaying the word. In addition, there is a row of bars associated with the words in the graph, where the height of a bar is proportional to the prominence of its associated word. The bars do not only function as an additional indication of word prominence, but also emphasise which semantic regions on the x-axis that are heavily populated in the text.

We have previously showcased the Word Rain visualisation technique on text comparison and dictionary development tasks, and we have also performed a user study (Skeppstedt et al., 2024). In addition, we have used the Word Rain visualisation technique in a digital history project to explore longitudinal changes in a temporal corpus (Skeppstedt & Aangenendt, 2024; Skeppstedt et al., 2025).

3 Examples of graphs generated with the Word Rain web service

In Figure 1, we have used the Word Rain web service to generate three different word rains with the top 300 words/bigrams, from three different corpora: the upper from the EuroParl-UdS corpus (Karakanta et al., 2018), spanning the years 1999-2017, the middle (House of Commons) and lower (House of Lords) from the British part of the ParlaMint corpus (Erjavec et al., 2023), only for the year 2017.

The Word Rain service removes stop words for the language of the texts, which is specified by the user before the word rains are generated. For the English texts used for the example, the stop word list employed by the web service used is the "English" list from NLTK (Bird, 2002). The English word embedding model has been retrieved from the NLPL word embeddings repository[1] (number 40, English CoNLL17 corpus, Word2Vec Continuous Skipgram).

The three graphs have been generated as one series of graphs, sharing the same word embedding projection on the x-axis. Word position on the x-axis can thereby be used for comparing the three word rains. As prominence measure term frequency is used.

We can, for instance, clearly see that *government* is prominent in both the House of Commons and House of Lords corpora. Since the word rains share the the same x-axis projection, we can also look in the European Parliament word rain at the same x coordinate, and if we look closely enough, we will find the same word there, but considerably smaller.

On the other hand, we can look at a very prominent word in the House of Lords corpus: *noble*. If we look in the European Parliament and House of Commons word rains, we will not find this word among the top 300 visualised. We will, however, find words that are used similarly nearby, such as *hon* (short for Honourable, title of member of parliament) in the House of Commons, and *mr* in the European Parliament. The Word Rain technique, thus, not only supports us in comparing graphs on an individual word level, as the for the word *government*. We can also make a comparison on a semantic category level: All three texts contain words denoting titles, but which titles are used differ between the three corpora. We can also detect clusters of words that only exist among the top 300 most prominent words in one of the corpora. For instance, the House of Lords corpus has a cluster of words denoting education (in violet, e.g., including the words *student/education/universities/academic*), which is only represented by the word *education* in the other British corpus, and not at all represented in the European Parliament corpus. The European Parliament data instead has a cluster of prominent words referring to EU and Europe (also in violet, e.g., including *European/EU/commission/council*). This word cluster also occurs in the graphs for the other two corpora, but it is not at all prominent in the other two graphs.

The discussion of differences and similarities between the graphs also exemplifies how the word positioning supports a manual, semantic categorisation of prominent words. We have, for instance, mentioned that prominent words belong to categories such as *titles*, *education* and words related to *EU/Europe*, i.e., categories that are easy to detect in the graphs thanks to the semantically motivated word positioning. We have also seen how the semantic positioning guides us to zooming in into interesting areas in the graph. For instance, the word *students*, in the House of Lords corpus and a corresponding cluster of many bars in the semantic area close to this word, made us zoom into this area, where we found more words on the topic of education. Thereby, in contrast to a traditional word cloud, there is a point of also including

[1] https://vectors.nlpl.eu/repository/

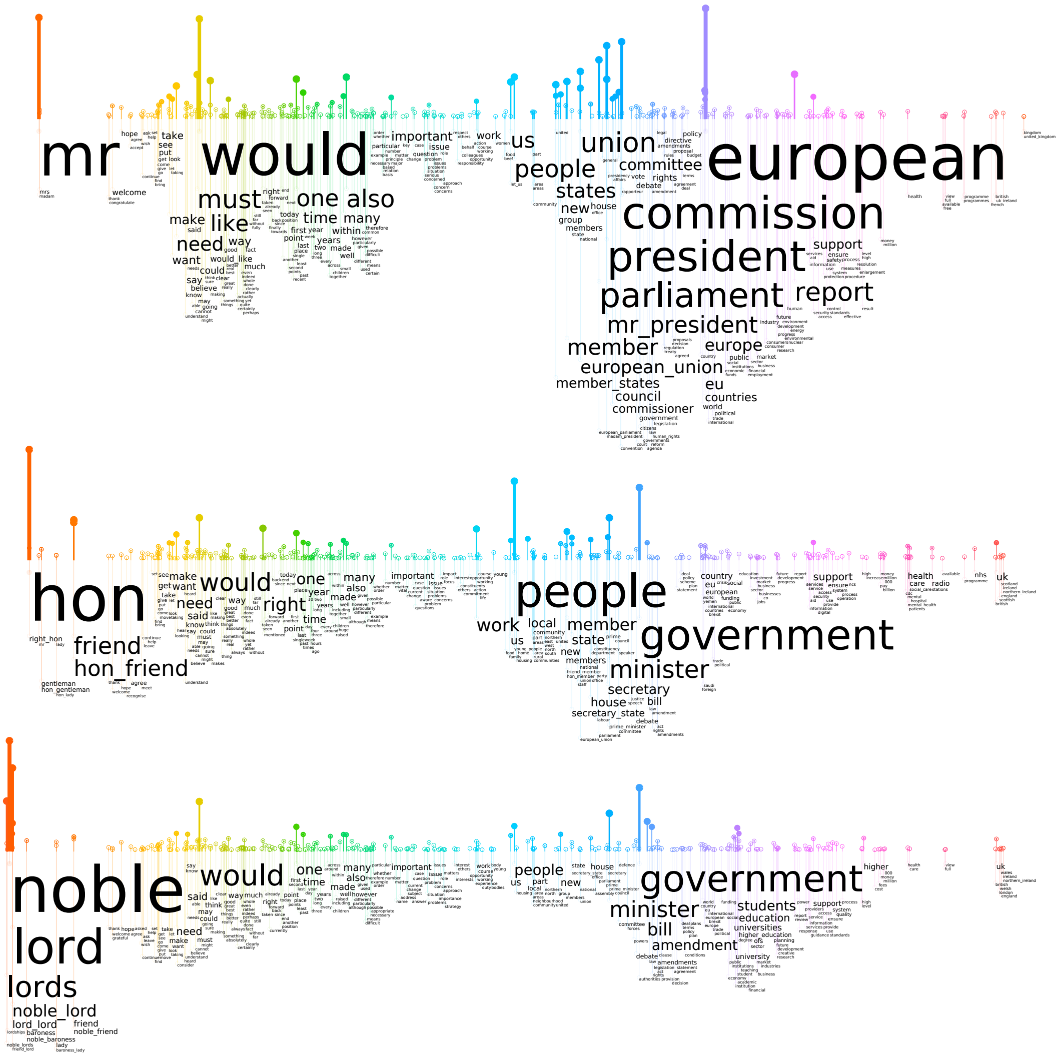


Figure 1: Example Word rains comparing corpora by using the same x-axis. The texts are from the European Parliament (EU), the House of Commons (UK) and the House of Lords (UK). The word rains include the top 300 most prominent words.

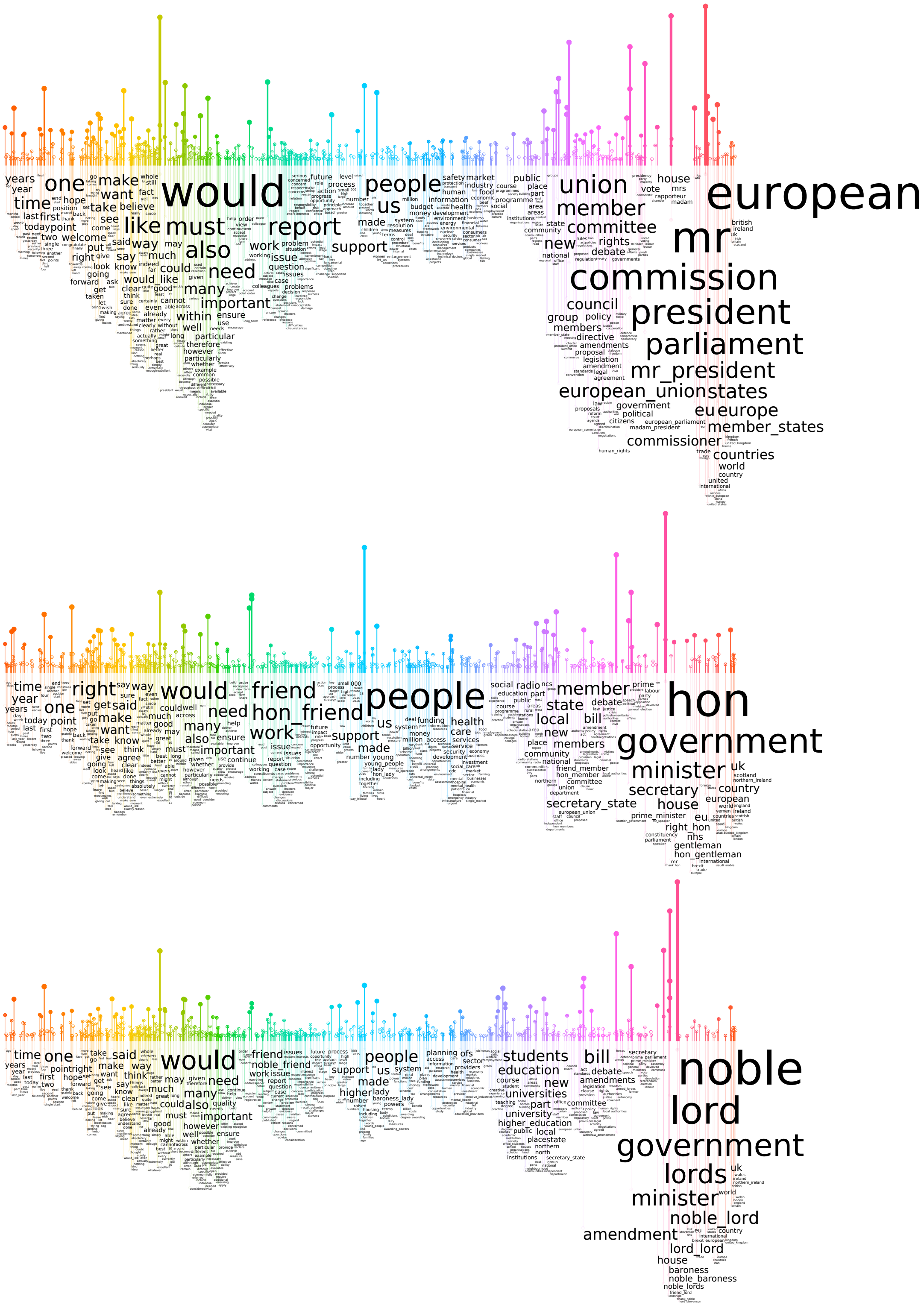
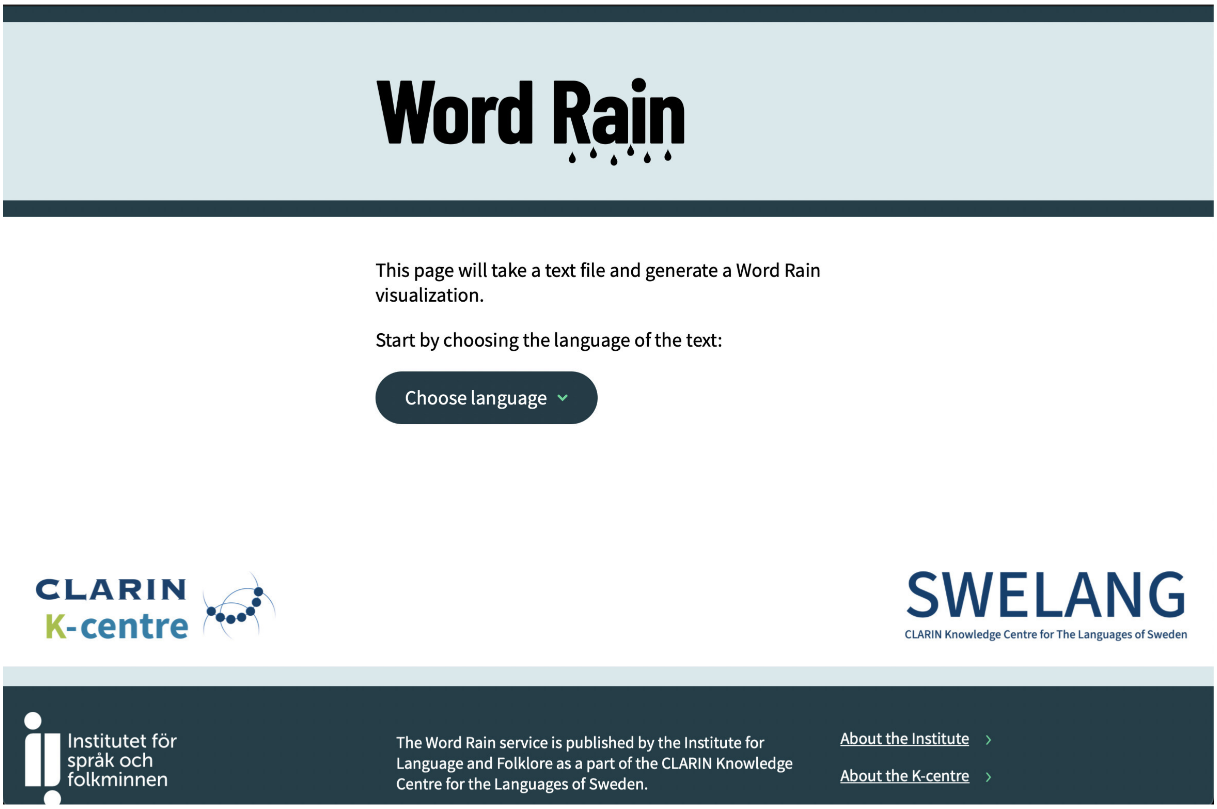


Figure 2: Example Word rains comparing corpora by using the same x-axis. The texts are from the European Parliament (EU), the House of Commons (UK) and the House of Lords (UK). The word rains include the top 600 most prominent words.



Word Rain

This page will take a text file and generate a Word Rain visualization.

Start by choosing the language of the text:

Choose language
CLARIN K-centre
SWELANG
CLARIN Knowledge Centre for The Languages of Sweden
Institutet för språk och folkminnen

The Word Rain service is published by the Institute for Language and Folklore as a part of the CLARIN Knowledge Centre for the Languages of Sweden.

About the Institute
About the K-centre

Figure 3: The Word Rain service web site at https://wordrain.isof.se/.

words displayed with a small font, as they can be reached through their more prominent semantic neighbours. Figure 1 also exemplifies how the coloured bars above the words give a sense of where there are prominent words, without the bias of word length.

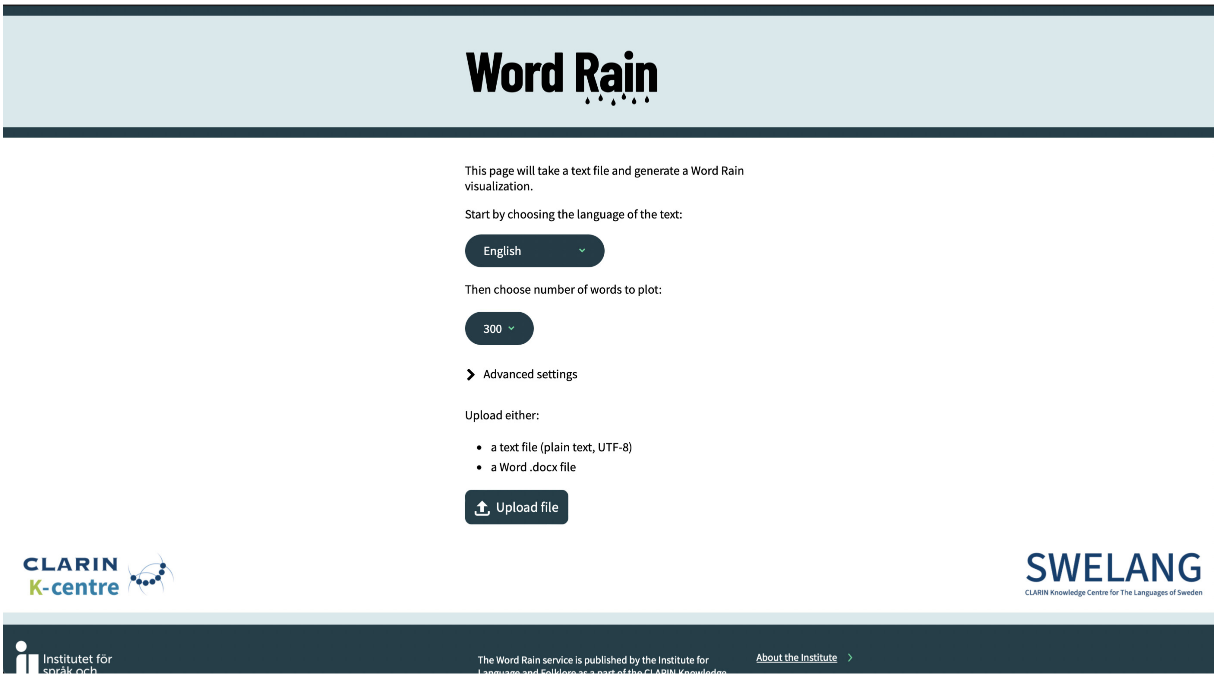
Figure 2 shows a word rain visualisation of the same three corpora, but the 600 most prominent words are instead included in the graph. Note that this graph, thereby, shows another series of word rains, with a different semantic projection on the x-axis than what was used for the graphs in Figure 1. The same type of analysis, as exemplified for the graphs showing the top 300 words, can be conducted on the word rain series containing the top 600 words.

Together, these properties of Word Rain mean that it can be meaningful to dig deeper, and use the graph for actual exploration and analysis, as opposed to traditional word clouds, where the information content in practice is not greater than a simple sorted list of words.

It should, however, be noted that the main goal of the Word Rain visualisation technique is still the same as that of a sorted word frequency list or of a word cloud, i.e., to provide an overview of the text content by extracting and displaying its most prominent words. The goal of the Word Rain visualisation technique is, thus, *not* to represent the original word embeddings in as much detail as possible, but to use the word embeddings to order the prominent words in a meaningful way. Thereby, this novel visualisation technique is able to better support the exploration and analysis of the most prominent words in a text.

4 The Word Rain web service

Word clouds can currently be generated by a large number of tools, from numerous online websites to software libraries that can be easily integrated into custom solutions. To generate word rains, on the other hand, the user has previously been required to download the Python code, install relevant packages, find (or train) a suitable language model, and finally write a small program using the Word Rain library.



Word Rain

This page will take a text file and generate a Word Rain visualization.

Start by choosing the language of the text:

Then choose number of words to plot:

Upload either:

- a text file (plain text, UTF-8)
- a Word .docx file

The Word Rain service is published by the Institute for

Figure 4: After the user has chosen a language, the basic configuration options for the text processing are shown.

This means that, even though we believe the Word Rain technique is a more suitable visualisation for many applications, most people would not have the time or technical know-how to generate a word rain from their data.

We have therefore collected everything needed for generating word rains in a web-based service, where the user can just upload one or more text documents and choose parameters for the visualisation. The service is both available as a web site at https://wordrain.isof.se/ and as open source code that can be easily deployed in standard web server environments.[2] At the moment, the instance deployed at wordrain.isof.se offers language models for Swedish, English, Finnish and Yiddish.

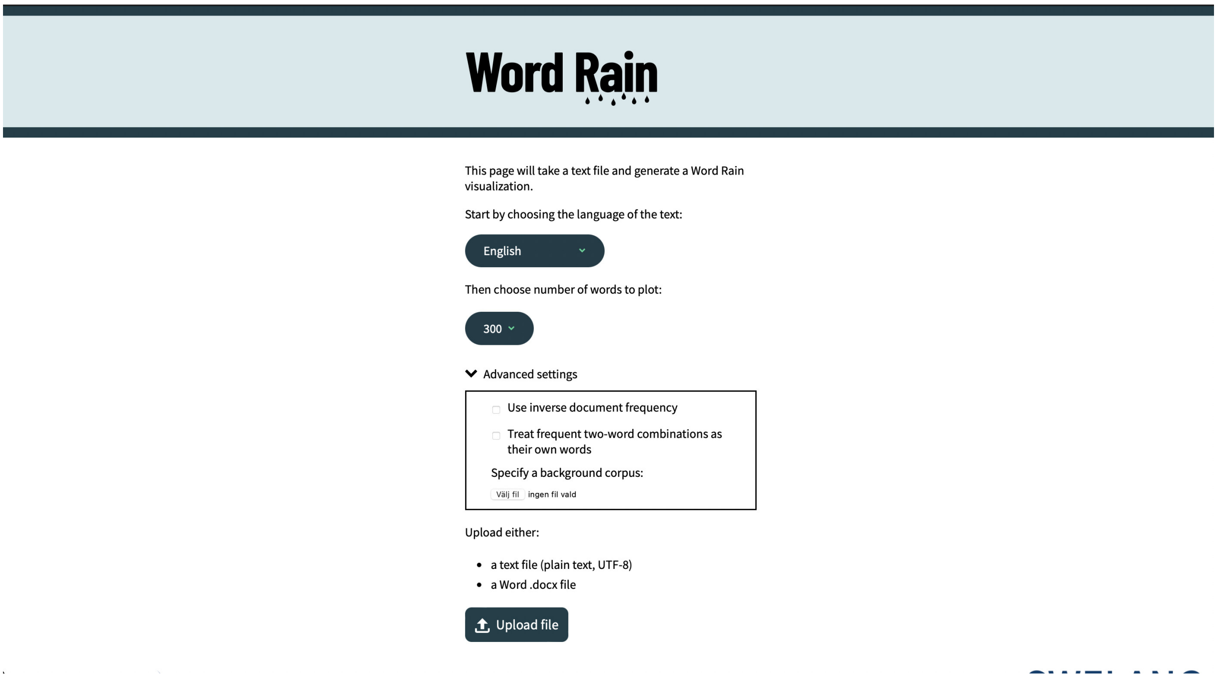
In the current version, the published code only supports plain UTF-8 text files. In order to support different file types without changing the code, the service provides a plug-in architecture where additional Python modules can be specified. Each plug-in has the opportunity to recognize and extract text from a document file format. The above-mentioned web site uses this plug-in architecture to support .docx (OOXML) files.

The web server code uses the main Word Rain Python library.[3] We made a few adaptions and additions to this library when developing the Word Rain web service. One of these regards how the word embedding model files are read. In the example cases, which we provide in the Word Rain code repository, the word embedding model files are read using the Gensim library (Rehurek & Sojka, 2011). However, the functionality provided by Gensim reads the whole model into memory. Since we aim to support several languages at once in the web server, this can quickly use a lot of memory. Word Rain only reads vectors for the words that are actually plotted, meaning the performance of each access is not critical. We therefore developed our own model reading library that initially only reads an index of the words into memory, and then reads each vector from the model file as needed.

Apart from the selection of a language model, the parameters for configuring the word rains can be grouped into two categories: parameters controlling the *text processing*, and parameters controlling the

[2] The source code for the web service is available at https://github.com/sprakradet/wordrain-service

[3] Available at https://github.com/CDHUppsala/word-rain



Word Rain

This page will take a text file and generate a Word Rain visualization.

Start by choosing the language of the text:

English

Then choose number of words to plot:

300

Advanced settings

□ Use inverse document frequency

□ Treat frequent two-word combinations as their own words

Specify a background corpus:

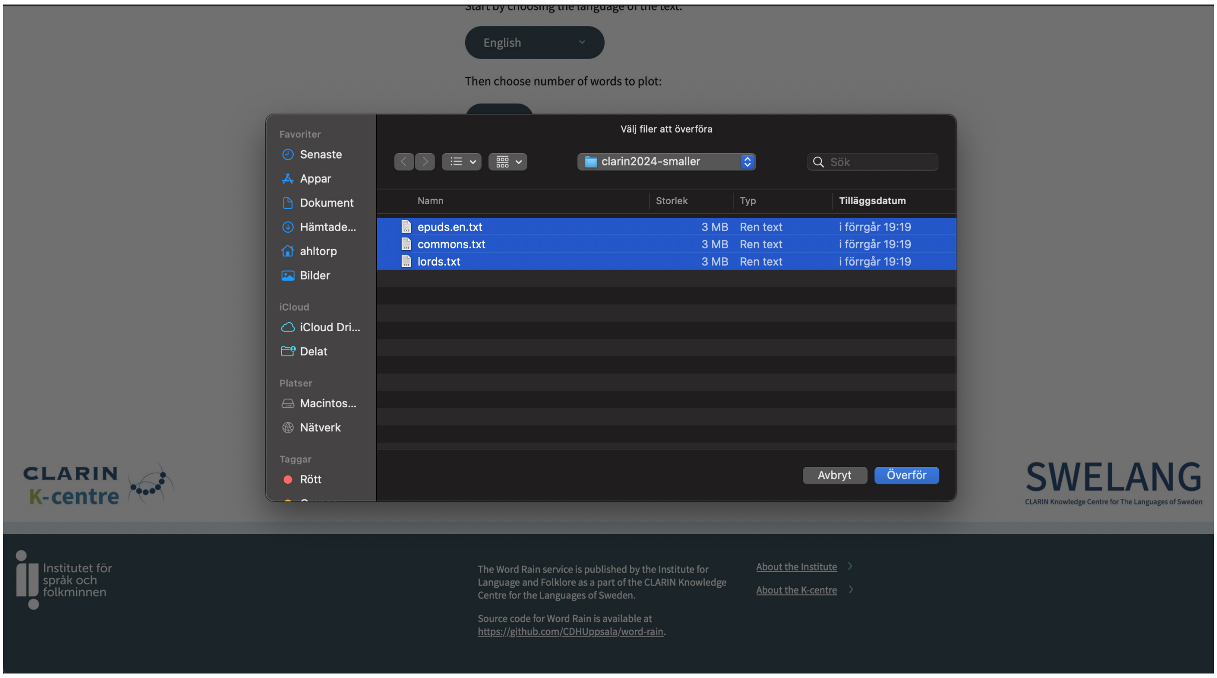
Välj fil ingen fil vald

Upload either:

- a text file (plain text, UTF-8)
- a Word .docx file

Upload file

Figure 5: More advanced settings for the text processing.



Start by choosing the language of the text:

English

Then choose number of words to plot:

Välj filer att överföra

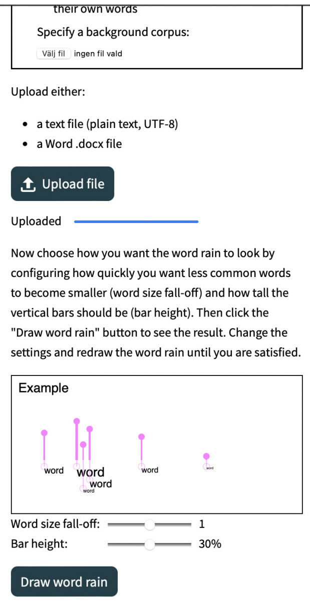
clarin2024-smaller

Namn	Storlek	Typ	Tilläggsdatum
epuds.en.txt	3 MB	Ren text	i förrgår 19:19
commons.txt	3 MB	Ren text	i förrgår 19:19
lords.txt	3 MB	Ren text	i förrgår 19:19

The Word Rain service is published by the Institute for Language and Folklore as a part of the CLARIN Knowledge Centre for the Languages of Sweden.

Source code for Word Rain is available at https://github.com/CDHUppsala/word-rain.

Figure 6: Choosing the texts to visualise.



their own words

Specify a background corpus:

Upload either:

- a text file (plain text, UTF-8)
- a Word .docx file

Now choose how you want the word rain to look by configuring how quickly you want less common words to become smaller (word size fall-off) and how tall the vertical bars should be (bar height). Then click the "Draw word rain" button to see the result. Change the settings and redraw the word rain until you are satisfied.

Word size fall-off: 1

Bar height: 30%

Figure 7: Choosing configuration for the graphical presentation.

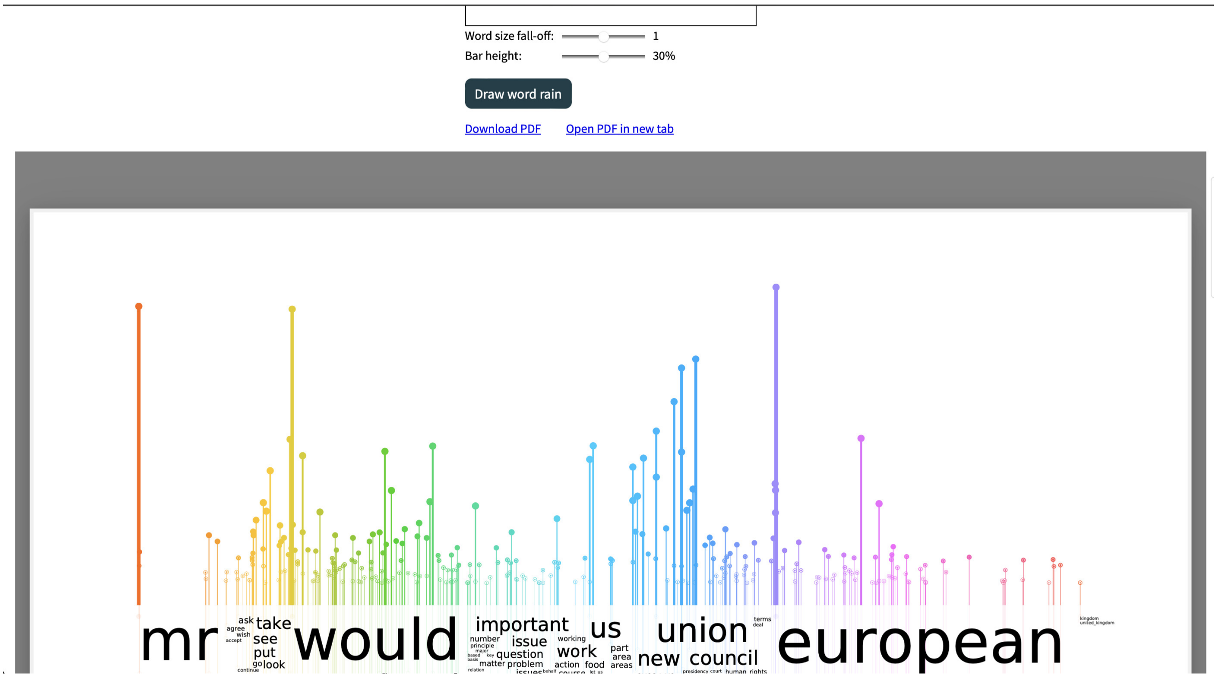
graphical presentation. Examples of text processing parameters are whether to use inverse document frequency and/or a background corpus for prominence calculations, the maximum size of n-grams that should be treated as one term, and the desired number of words to display. Graphical presentation parameters include how much space to dedicate to the vertical bars, and how sharp the drop-off in font size should be. In the following paragraphs, we will practically show how to configure the different parameters.

Figure 3 shows the information presented when the web page of the service is first loaded. When the user has chosen one of the languages offered by the deployed instance of the web service, the one basic parameter controlling the text processing becomes visible. As shown in Figure 4, this parameter consists of the number of words to include in the plot. The default value is the top 300 most prominent words, but the user can also choose to include the top 600 most prominent ones.

It is also possible to select more advanced settings for the text processing carried out. Figure 5 shows the three more advanced user settings currently available: i) To use the term frequency–inverse document frequency as the prominence measure instead of the standard term frequency measure, ii) to include prominent bigrams in the graph, and iii) to use a background corpus for the inverse document frequency calculations.

After the configuration for the text processing has been carried out, the text(s) to visualise can be uploaded. This is done by pressing the “Upload file” button, which results in a dialogue box for choosing files. This is illustrated in Figure 6, and as shown in the figure, it is possible to select multiple files. When the files have been uploaded, a progress bar is displayed to the user while the top *n* most prominent words are extracted from each of the uploaded texts, and a t-SNE projection is calculated from the matrix of word embeddings representing the top *n* words in all texts.

After the text processing has been carried out, the user can select configuration parameters for the graphical presentation. There are two parameters to set: i) “Word size fall-off”, and ii) “bar height”. The “word size fall-off” rate governs how quickly the font size of the less prominent words are to decrease in the graph. “Bar height” decides the height of the vertical bars. The value for these configuration options are set by using sliders, as shown in Figure 7.


Word size fall-off: 1

Bar height: 30%

Figure 8: The word rains have been generated.

By pressing the “Draw word rain” button, the user can generate word rains with the configuration options chosen. The user is then again presented with a progress bar, while the graphs are being generated. When the graphs have been produced, they are then shown to the user, as in Figure 8. The user can then either change the configuration for the graphical presentation and generate new word rains, or – if satisfied with the design of the word rains – download the PDF-file generated.

5 Conclusions and future work

Providing an easy-to-use tool for generating word rains is crucial to widespread adoption. Many potential users do not have the technical skills required to run the Word Rain Python code library. Others do not have the time to use a code library for evaluating a new visualisation technique, even if they would eventually want to set it up themselves, should they decide to use it in their workflow. The tool is therefore a welcome addition to the SWELANG CLARIN K-centre repertoire, especially for the K-centre’s focus: the languages of Sweden. Offering it for English as well from the start makes it useful for prototyping Word Rain use from the whole CLARIN user community.

Future work includes the addition of more configuration parameters. We do not aim to include all of the configuration options offered by the Python code library into the web service, but we believe there are a number of additional configuration options that might be useful to add to the service, and which would not make the service too complex to use. We have, for instance, received requests to add the possibility of uploading a user-defined stop word list to the web service. We also plan to add support for more languages, starting with the Swedish national minority language Meänkieli.

There are also a number of technical improvements that might be relevant to implement. For instance, processing large corpora uses large amounts of memory for the calculation of word frequency. This could probably be optimised, by for example running a pre-processing pass eliminating all words under a certain frequency.

Acknowledgments

The development of the Word Rain web service is funded by the Swedish Research Council: Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

The development of the Word Rain Python library code is funded by the Swedish Research Council: Swe-CLARIN/The National Language Bank of Sweden (2017-00626), Huminfra (2021-00176) and InfraVis (2021-00181).

References

Barth, L., Kobourov, S. G., & Pupyrev, S. (2014). Experimental comparison of semantic word clouds. In J. Gudmundsson & J. Katajainen (Eds.), *Experimental algorithms* (pp. 247–258). Springer International Publishing. https://doi.org/10.1007/978-3-319-07959-2_21

Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Burch, M., Lohmann, S., Beck, F., Rodriguez, N., Di Silvestro, L., & Weiskopf, D. (2014). RadCloud: Visualizing multiple texts with merged word clouds. *Proceedings of the International Conference on Information Visualisation*, 108–113. https://doi.org/10.1109/IV.2014.72

Cao, N., & Cui, W. (2016). *Introduction to text visualization* (Vol. 1). Atlantis Press. https://doi.org/10.2991/978-94-6239-186-4

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M., & Qu, H. (2010). Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, *30*(6), 42–53. https://doi.org/10.1109/MCG.2010.102

Diakopoulos, N., Elgesem, D., Salway, A., Zhang, A., & Hofland, K. (2015). Compare Clouds: Visualizing text corpora to compare media frames. *Proceedings of the 2015 IUI Workshop on Visual Text Analytics*.

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. (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1859

Herold, E., Pöckelmann, M., Berg, C., Ritter, J., & Hall, M. M. (2019). Stable word-clouds for visualising text-changes over time. *Digital Libraries for Open Knowledge*, 224–237. https://doi.org/10.1007/978-3-030-30760-8_20

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

Karakanta, A., Vela, M., & Teich, E. (2018). EuroParl-UdS: Preserving and extending metadata in parliamentary debates. *Proceedings of the LREC 2018*.

Lee, B., Riche, N. H., Karlson, A. K., & Carpendale, S. (2010). SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1182–1189. https://doi.org/10.1109/TVCG.2010.194

Liu, X., Shen, H.-W., & Hu, Y. (2015). Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization*, *14*(2), 168–180. https://doi.org/10.1177/1473871614534095

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Skeppstedt, M., & Aangenendt, G. (2024). Using the Word Rain technique to visualize longitudinal changes in periodicals from the Swedish Diabetes Association. *Proceedings of the Workshop on Visualization for Natural Language Processing*. https://doi.org/10.2312/vis4nlp.20241132

Skeppstedt, M., Ahltorp, M., Aangenendt, G., & Söderfeldt, Y. (2025). Further developing the Word Rain text visualisation technique in a digital history project. *Digital Humanities in the Nordic and Baltic Countries Publications*, *7*(2). https://doi.org/10.5617/dhnbpub.12292

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*. https://doi.org/10.1177/14738716241236188

Wang, Y., Chu, X., Bao, C., Zhu, L., Deussen, O., Chen, B., & Sedlmair, M. (2018). EdWordle: Consistency-preserving word cloud editing. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 647–656. https://doi.org/10.1109/TVCG.2017.2745859

Wu, Y., Provan, T., Wei, F., Liu, S., & Ma, K.-L. (2011). Semantic-preserving word clouds by seam carving. *Computer Graphics Forum*, *30*(3), 741–750. https://doi.org/0.1111/j.1467-8659.2011.01923.x

Xie, L., Shu, X., Su, J. C., Wang, Y., Chen, S., & Qu, H. (2024). Creating emordle: Animating word cloud for emotion expression. *IEEE Transactions on Visualization and Computer Graphics*, *30*(8), 5198–5211. https://doi.org/10.1109/TVCG.2023.3286392

Xu, J., Tao, Y., & Lin, H. (2016). Semantic word cloud generation based on word embeddings. *Proceedings of the IEEE Pacific Visualization Symposium*, 239–243. https://doi.org/10.1109/PACIFICVIS.2016.7465278

Adapting UPSKILLS Learning Modules to the University Curricula: Best Practices and Lessons Learnt from the H2IOSC Training Experience at the University of Ferrara

Giulia Pedonese
CNR Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy
giulia.pedonese@cnr.it

Francesca Frontini
CNR Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy
francesca.frontini@cnr.it

Dario Del Fante
Department of Humanities University of Ferrara, Ferrara, Italy
dario.delfante@unife.it

Eleonora Federici
Department of Humanities University of Ferrara, Ferrara, Italy
eleonora.federici@unife.it

Abstract

This paper details the steps taken to adapt and integrate the training materials developed by CLARIN ERIC in two bachelor's degree courses and one master's degree course at the University of Ferrara. The workflow applies the shared methodology developed within the Humanities and Heritage Italian Open Science Cloud project. It modifies the training materials of the UPSKILLS course "Introduction to Language Data: Standards and Repositories" according to the needs of three target courses focusing on English to Italian translation: English Language Course for Tourism, English Language for Translation and English Language and Linguistics for Humanities, Arts and Archaeology. The result of this pilot is a documented example of how CLARIN services can be integrated into university teaching, including initial teacher training, and providing an opportunity to discuss the topic and a use case for trainers who intend to include CLARIN in their courses.

1. Introduction

The Humanities and Cultural Heritage Italian Open Science Cloud project (H2IOSC) (Degl'Innocenti et al. 2023) aims to create a federated cluster of the services and resources developed by the national nodes of four research infrastructures for Open Science that are part of the European Strategy Forum on Research Infrastructure (ESFRI) roadmap in the area of social and cultural innovation.[1] One is CLARIN-IT, the Italian consortium of the Common Language Resource and Technology Infrastructure.[2] In line with other projects of national and international scope, H2IOSC devotes an entire work package to training and education: the Work Package 8 (WP8), whose aim is to define a comprehensive shared strategy for training at the level of single infrastructures and for the whole cluster based on the needs of

[1] 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.

[2] The participating research infrastructures are: Digital Research Infrastructures for the Arts and Humanities (DARIAH); European Research Infrastructure for Heritage Science (E-RIHS), Common Language Resource and Technology Infrastructure (CLARIN) and Open Scholarly Communication in the European Research Area for Social Sciences and Humanities (OPERAS). See the official website: https://www.h2iosc.cnr.it/home/

Giulia Pedonese, Francesca Frontini, Dario Del Fante and Eleonora Federici 2024. Adapting UPSKILLS Learning Modules to 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. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 37–47. DOI: https://doi.org/10.3384/ecp216

the community as identified with the landscaping activity of Work Package 2. The repurposing of reusable teaching materials planned within H2IOSC WP8 is in line with the teaching strategy of other European projects: first, with the UPSKILLS[3] project, which aims to integrate university curricula of language subjects with skills required by the labor market, and for which CLARIN developed the course "Introduction to Language Data: Standards and Repositories" (Section 2). Secondly, repurposing teaching materials for different audiences aligns with the aims of the Skills for the European Open Science Commons project (Skills4EOSC)[4] to expand European researchers' skills by creating European competence centres, with which H2IOSC training has aligned (Section 2). Therefore, this initiative appears to be embedded in a broader context, which aims not only to offer educational materials suited to the professional needs of a learner audience but also to create a network of trainers who can exchange skills and materials.

This paper aims to show the application of the training strategy detailed in WP8 to the modules designed by CLARIN-IT within activity 8.2: "Teach CLARIN, teach with CLARIN: training, communication and impact", adapting them to the training needs of the University of Ferrara. This experience resulted in a case study to show the potential of such a strategy when integrating the CLARIN research infrastructure into teaching. First, there will be an overview of the H2IOSC training methodology and its applications to CLARIN courses (Section 2) and a background on the training needs of the selected curricula at the University of Ferrara (Section 3). Then, Section 4 and its subsections will detail the steps of this pilot. Finally, Sections 5 and 6 will highlight the steps taken in 2025 and the lessons learnt from this experience.

2. H2IOSC Training Methodology Applied to CLARIN Courses

The H2IOSC project's training strategy is aimed primarily at the Italian Social Sciences and Humanities communities and offers tailored training on the resources and tools available in the participating infrastructures. An essential pillar of this strategy is the train-the-trainers perspective, which aims to help teachers of university and professional courses enhance their skills and those of their students. To this end, in WP8, we developed a shared methodology for designing training materials as digital objects compliant with the FAIR principles (Wilkinson et al., 2016). We implemented two training platforms to facilitate their reuse: a Learning Management System for students to access course materials in a modular and intuitive way, and a repository for teachers to browse and download learning resources from. As the development of those platforms was still ongoing while the lessons in Ferrara took place, we will focus on the methodology we applied (Pedonese et al. 2024) to make the course materials accessible and reusable on different devices. However, the integration of these technologies is planned in the continuation of the pilot.

While developing a methodology to align all four research infrastructures on the adaptation and creation of reusable training materials, we identified the *FAIR-by-Design Methodology for Learning Materials* as the standard established by the Skills4EOSC project, funded by the European Union under Horizon Europe and aimed to enhance Open Science skills in Europe by promoting FAIR practices. This framework ensures training materials align with the FAIR principles (Findable, Accessible, Interoperable, Reusable) to enhance their reusability in the scientific community. It adopts a backwards instructional design process, expanded with six steps: planning, development, quality check, publication, and dissemination, emphasising metadata, interoperability, and granularity and includes practical tools like checklists and templates to support implementation (Filiposka et al. 2023). The primary references we have drawn from the methodology, adapting it to the needs of our disciplinary area, have been the adoption of a standard metadata schema and the definition of the learning object as the smallest unit to which it applies.

[3] UPSKILLS stands for "UPgrading the SKIlls of Linguistics and Language Students": it was an Erasmus+ strategic partnership for higher education running from September 2020 to August 2023. See the official website: https://upskillsproject.eu/

[4] It ran from 2022 to 2025, with the goal of building a comprehensive training ecosystem. The project brought together international partners to advance open, interoperable scientific practices. See the official website: https://www.skills4eosc.eu/

To effectively describe and share H2IOSC learning materials, we adopted the metadata set developed by the Research Data Alliance (RDA), the Minimal Metadata Schema for Learning Resources (Hoebelheinrich et al. 2022). As shown in Fig.1, this dataset includes 14 fields divided between descriptive information (title, author, etc.), access information (licence, URL, etc.) and didactic information (type of resource, target group, initial level, learning outcomes).

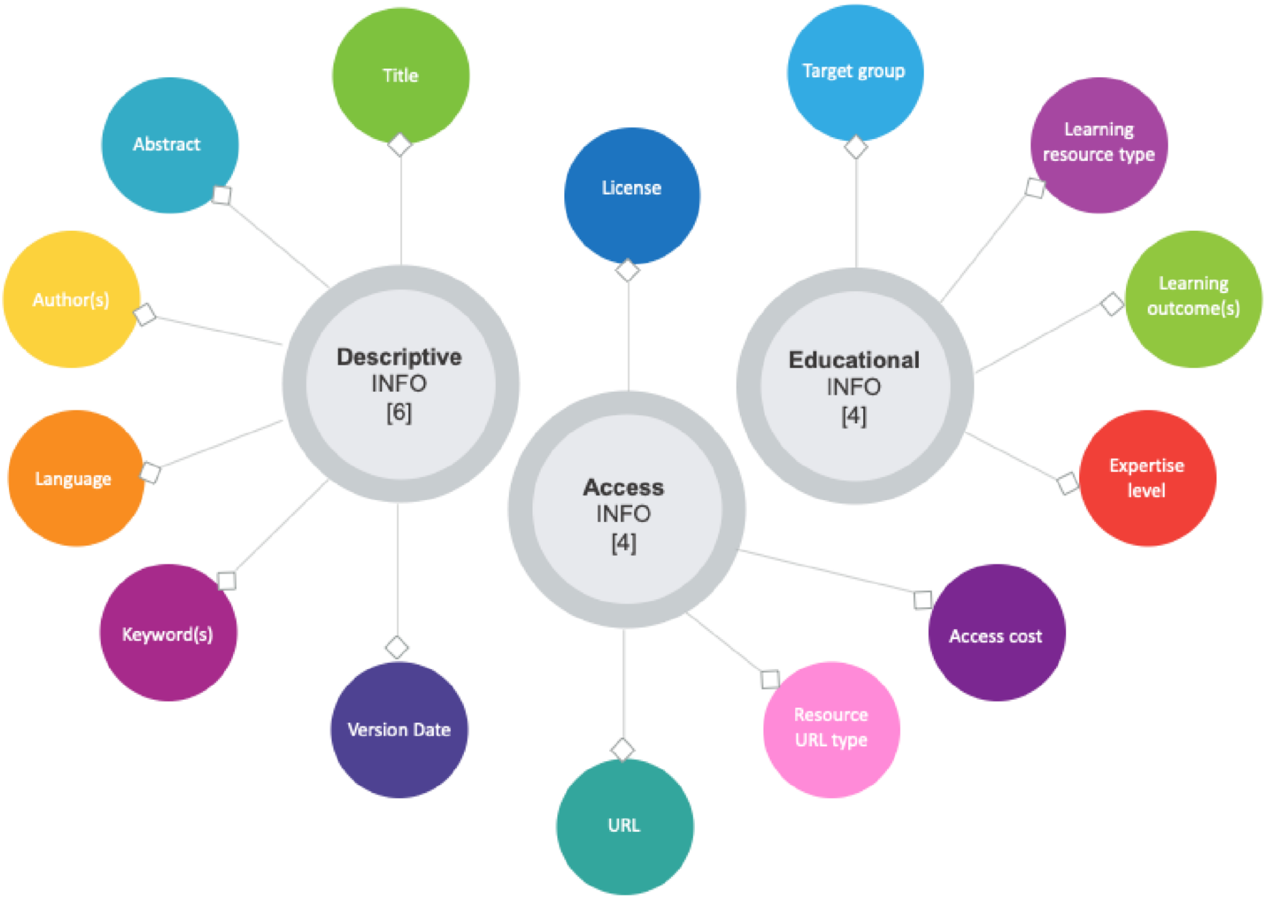
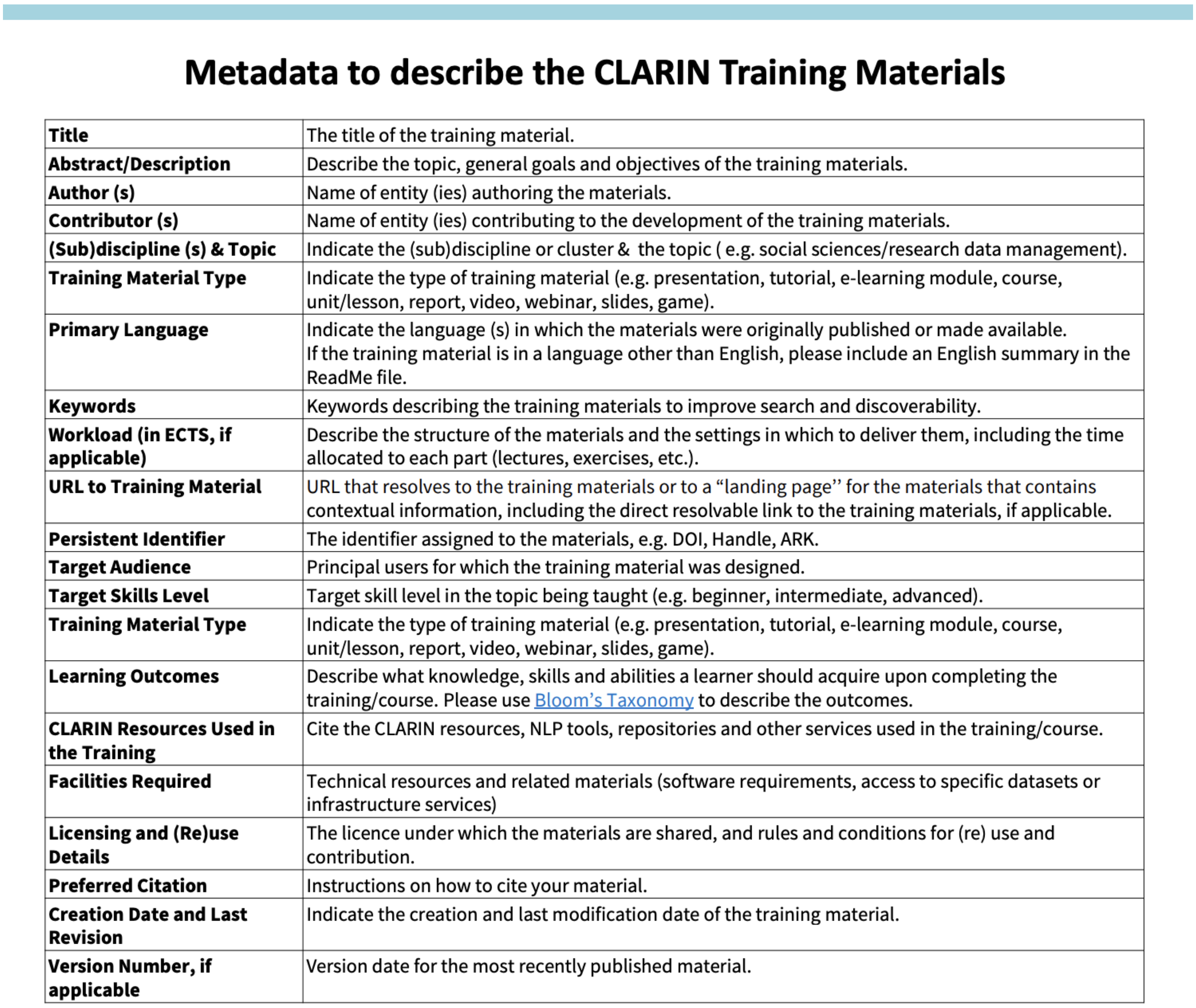


Fig. 1. RDA Minimal Metadata Set with Definitions (Hoebelheinrich et al. 2022).

This highly flexible model has already been adapted to the needs of CLARIN (van der Lek et al. 2023 c), which has invested in training by promoting standard practices and creating a community of trainers through the CLARIN Trainers’ Network. Fig. 2 provides an example of the CLARIN adaptation. In H2IOSC, we have adopted the modified model by including specific fields such as *contributors*, *workload in ECTS*, *PID*, *version date,* and *standard citation*.



Metadata to describe the CLARIN Training Materials

Title	The title of the training material.
Abstract/Description	Describe the topic, general goals and objectives of the training materials.
Author (s)	Name of entity (ies) authoring the materials.
Contributor (s)	Name of entity (ies) contributing to the development of the training materials.
(Sub)discipline (s) & Topic	Indicate the (sub)discipline or cluster & the topic (e.g. social sciences/research data management).
Training Material Type	Indicate the type of training material (e.g. presentation, tutorial, e-learning module, course, unit/lesson, report, video, webinar, slides, game).
Primary Language	Indicate the language (s) in which the materials were originally published or made available. If the training material is in a language other than English, please include an English summary in the ReadMe file.
Keywords	Keywords describing the training materials to improve search and discoverability.
Workload (in ECTS, if applicable)	Describe the structure of the materials and the settings in which to deliver them, including the time allocated to each part (lectures, exercises, etc.).
URL to Training Material	URL that resolves to the training materials or to a "landing page" for the materials that contains contextual information, including the direct resolvable link to the training materials, if applicable.
Persistent Identifier	The identifier assigned to the materials, e.g. DOI, Handle, ARK.
Target Audience	Principal users for which the training material was designed.
Target Skills Level	Target skill level in the topic being taught (e.g. beginner, intermediate, advanced).
Training Material Type	Indicate the type of training material (e.g. presentation, tutorial, e-learning module, course, unit/lesson, report, video, webinar, slides, game).
Learning Outcomes	Describe what knowledge, skills and abilities a learner should acquire upon completing the training/course. Please use Bloom's Taxonomy to describe the outcomes.
CLARIN Resources Used in the Training	Cite the CLARIN resources, NLP tools, repositories and other services used in the training/course.
Facilities Required	Technical resources and related materials (software requirements, access to specific datasets or infrastructure services)
Licensing and (Re)use Details	The licence under which the materials are shared, and rules and conditions for (re) use and contribution.
Preferred Citation	Instructions on how to cite your material.
Creation Date and Last Revision	Indicate the creation and last modification date of the training material.
Version Number, if applicable	Version date for the most recently published material.

Fig. 2. Metadata to Describe the CLARIN Training Materials (van der Lek et al. 2023 c).

According to the best practices recommended in Skills4EOSC, we considered the single lesson as a learning object, the minimum unit to apply the FAIR principles and be described with the RDA metadata schema. The learning object, as defined in the methodology, is "any digital resource that supports learning developed around a single learning objective defined as a package of a lesson, activity and assessment with a concrete learning outcome" (Filiposka et al. 2023:8).[5] While this certainly facilitates the reusability of training materials for future reuse, during the course delivery phase to students, it was applied to teaching by dividing the module into several lectures of no more than 25 minutes alternating with multiple-choice quizzes using interactive software such as *Kahoot!* and *Mentimeter*, implementing gamification strategies. Furthermore, we adopted open Creative Commons licenses and standard citations while sharing the materials with students and improved the accessibility of teaching materials by converting source formats into even more open and editable formats, possibly open source.[6]

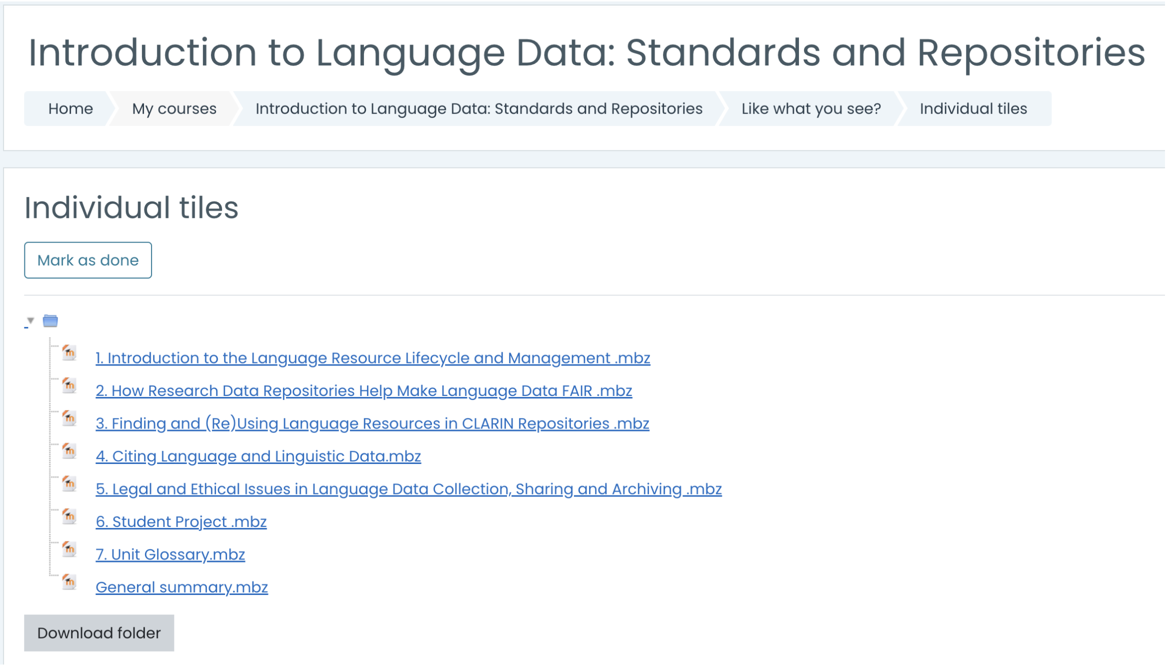
In the first half of the project, the CLARIN-IT consortium, represented by its founder member and host of the ILC4CLARIN service provider centre at the CNR Institute of Computational Linguistics,[7] worked on applying the Skills4EOSC methodology to the UPSKILLS course "Introduction to Language Data: Standards and Repositories" (van der Lek et al. 2023 a) which was already reusable in compliance

[5] For a more extensive definition, see also pp. 23-24 of the deliverable.

[6] More detais on the H2IOSC methodology for training materials in Pedonese et al., 2024 now submitted for reviewing in Umanistica Digitale.

[7] https://ilc4clarin.ilc.cnr.it/

with the FAIR principles and accessible on the project's Moodle platform.[8] As shown in Fig.3, the user could download each module of this course by selecting it from the course structure.



Introduction to Language Data: Standards and Repositories

Individual tiles

1. Introduction to the Language Resource Lifecycle and Management .mbz
2. How Research Data Repositories Help Make Language Data FAIR .mbz
3. Finding and (Re)Using Language Resources in CLARIN Repositories .mbz
4. Citing Language and Linguistic Data.mbz
5. Legal and Ethical Issues in Language Data Collection, Sharing and Archiving .mbz
6. Student Project .mbz
7. Unit Glossary.mbz
General summary.mbz

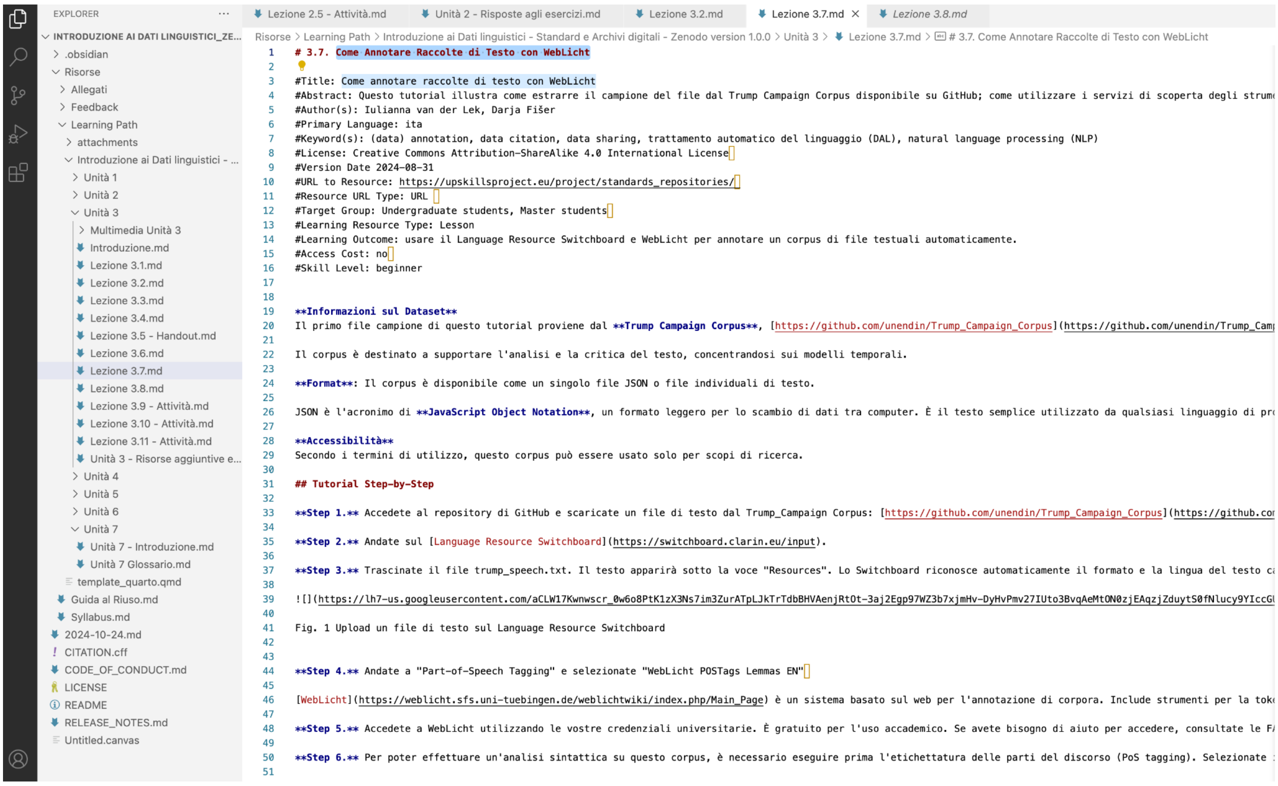
Fig. 3. Original Training Materials on UPSKILLS Moodle Platform (van der Lek et al., 2023 a).

The course is designed to provide instructors and students in BA/MA linguistic programs with educational resources and practical activities on using research data repositories within the lifecycle of linguistic data, aligning with the FAIR and Open Science principles. Its learning objectives are using certified repositories to discover, share, publish, and store linguistic resources and datasets and applying integrated services and tools from repositories to process, annotate, and analyse various types of corpora according to community standards.

To maximise its overall reusability and adapt the course materials ourselves, we first asked the creators to access the editable version of the learning block, since the only available version was in Moodle predefinite format (.mbz file), which we could download from the original project's Moodle and was only editable via the same platform. We then translated the entire course into Italian and published it as a structured aggregate of Markdown[9] files following the Skills4EOSC methodology (van der Lek et al. 2024). Fig. 4 shows an example of training material in a Markdown file: this highly flexible and compatible annotation allows users to easily integrate these training materials into their editing tools.

[8] https://upskillsproject.eu/. The course is available on the UPSKILLS Moodle platform: https://upskillsproject.eu/project/standards_repositories/. For CLARIN's activities in UPSKILLS and the training materials produced see also https://www.clarin.eu/content/upskills-learning-and-teaching-materials (Gledić et al. 2023).

[9] https://www.markdownguide.org/



```
# 3.7. Come Annotare Raccolte di Testo con WebLicht

#Title: Come annotare raccolte di testo con WebLicht
#Abstract: Questo tutorial illustra come estrarre il campione del file dal Trump Campaign Corpus disponibile su GitHub; come utilizzare i servizi di scoperta degli strume
#Author(s): Iulianna van der Lek, Darja Fišer
#Primary Language: ita
#Keyword(s): (data) annotation, data citation, data sharing, trattamento automatico del linguaggio (DAL), natural language processing (NLP)
#License: Creative Commons Attribution-ShareAlike 4.0 International License
#Version Date 2024-08-31
#URL to Resource: https://upskillsproject.eu/project/standards_repositories/
#Resource URL Type: URL
#Target Group: Undergraduate students, Master students
#Learning Resource Type: Lesson
#Learning Outcome: usare il Language Resource Switchboard e WebLicht per annotare un corpus di file testuali automaticamente.
#Access Cost: no
#Skill Level: beginner


**Informazioni sul Dataset**
Il primo file campione di questo tutorial proviene dal **Trump Campaign Corpus**, [https://github.com/unendin/Trump_Campaign_Corpus](https://github.com/unendin/Trump_Camp

Il corpus è destinato a supportare l'analisi e la critica del testo, concentrandosi sui modelli temporali.

**Format**: Il corpus è disponibile come un singolo file JSON o file individuali di testo.

JSON è l'acronimo di **JavaScript Object Notation**, un formato leggero per lo scambio di dati tra computer. È il testo semplice utilizzato da qualsiasi linguaggio di pro

**Accessibilità**
Secondo i termini di utilizzo, questo corpus può essere usato solo per scopi di ricerca.

## Tutorial Step-by-Step

**Step 1.** Accedete al repository di GitHub e scaricate un file di testo dal Trump_Campaign Corpus: [https://github.com/unendin/Trump_Campaign_Corpus](https://github.com

**Step 2.** Andate sul [Language Resource Switchboard](https://switchboard.clarin.eu/input).

**Step 3.** Trascinate il file trump_speech.txt. Il testo apparirà sotto la voce "Resources". Lo Switchboard riconosce automaticamente il formato e la lingua del testo ca

![](https://lh7-us.googleusercontent.com/aCLW17Kwnwscr_0w6o8PtK1zX3Ns7im3ZurATpLJkTrTdbBHVAenjRtOt-3aj2Egp97WZ3b7xjmHv-DyHvPmv27IUto3BvqAeMtON0zjEAqzjZduytS0fNlucy9YIccGl

Fig. 1 Upload un file di testo sul Language Resource Switchboard


**Step 4.** Andate a "Part-of-Speech Tagging" e selezionate "WebLicht POSTags Lemmas EN"

[WebLicht](https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page) è un sistema basato sul web per l'annotazione di corpora. Include strumenti per la toke

**Step 5.** Accedete a WebLicht utilizzando le vostre credenziali universitarie. È gratuito per l'uso accademico. Se avete bisogno di aiuto per accedere, consultate le FA

**Step 6.** Per poter effettuare un'analisi sintattica su questo corpus, è necessario eseguire prima l'etichettatura delle parti del discorso (PoS tagging). Selezionate :

```

Fig. 4. Training Materials' Adapted Structure in Markdown Editor (van der Lek et al., 2024).

While the course was being translated, we selected the parts that could be reused for teaching at the University of Ferrara: we adapted Lessons 1.1 to 1.4 on learning resources, disciplinary repositories and lessons 3.1 to 3.3. on CLARIN core services that implement the FAIR principles, such as the Virtual Language Observatory and the Language Resource Switchboard, and an introduction to the fundamentals of Natural Language Processing. Those were the main points on which we wanted to raise the interest of the students of the selected courses in Ferrara (Section 3), as language studies, both in academia and business, can no longer do without proper data management.

3. Teaching Digital Humanities at the University of Ferrara

The University of Ferrara is a dynamic environment where digital humanities have the potential to fruitfully interact with linguistic research, especially with CLARIN's support. The University Language Centre of the University of Ferrara has been a CLARIN-IT member since 2023, providing metadata related to language resources for the ILC4CLARIN repository centre. Thus, integrating CLARIN services into teaching is a way to educate future researchers and professionals in an increasingly digitised context, providing them with tools and resources to support them in their careers. This is why a collaboration was established between the CNR-ILC personnel dedicated to training in the H2IOSC project and the CLARIN-IT referent for the University Language Centre, who also happen to be lecturers of linguistic courses at BA, MA and PhD level.

More specifically, the collaboration aimed to acquaint the students with the fundamentals of digital humanities using CLARIN tools and services. To this end, we selected three courses Prof. Federici and Dr. Del Fante taught. We included two lessons at the end of each course, providing an introduction to the use of language resources and language technologies, how to access CLARIN tools, how to use a repository such as ILC4CLARIN, and why practices for accessing, publishing and reusing data on the web like the FAIR principles can have a direct impact on their student careers. So, we chose three English Language and Linguistics courses belonging to three different degree programs respectively: 1) Bachelor's degree in Humanities, Arts, and Archaeology; 2) Master's degree in Foreign Languages and Literature; 3) Bachelor's degree program in Manager of Cultural Itineraries.

The English Language for Humanities, Arts, and Archaeology course (1) aims to provide students with critical tools to understand English information. The course focuses on three main topics: English critically approached as a global language, the notion of discourse, and communicative strategies for representing people and identities in political communication and social media contexts. The English Language course for Modern Languages and Literature (2) consists of a theoretical part, which offers a panorama of the main aspects of Cultural Translation, and a practice section, which will translate literary texts of various periods and contexts from English to Italian. Finally, the English Language for Tourism course (3) aims to equip students with linguistic and metalinguistic tools essential for developing an awareness of the English tourist language. This goal is achieved through linguistic and cultural analysis of various contemporary tourism text types. These were selected as pedagogical spaces to introduce students to CLARIN technologies, and the UPSKILLS course was adapted to fit each skill level and focus. The adaptation process was relatively easy, as the course materials were already structured in a highly flexible way, allowing us to combine multiple topics and adjust the content as needed.

As this was the first experiment of this kind, we tried to include different disciplinary areas encompassing a broad selection of professional profiles, to test the flexibility of the teaching materials that were made available. In parallel, we devised a method to gather as much feedback as possible to improve the teaching-learning experience for future editions.

4. CLARIN-IT Training at the University of Ferrara

The training initiative stemmed from preexisting contacts between the CNR-ILC and the University of Ferrara and followed the train-the-trainer approach. The first meeting was held online to align Dr. Dario Del Fante, teacher and researcher at the University of Ferrara, with the latest available resources, define the lesson plan, and schedule an assessment strategy. Then, we held the lessons in person at the University of Ferrara between April and May 2024, finally assessed the students, and gathered feedback via a questionnaire. The following subsections will detail the phases of the pilot as shown in Fig. 5.

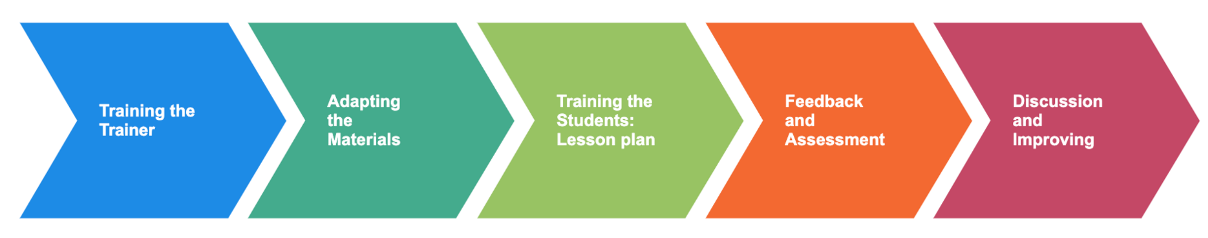


Fig. 5. Phases of the Pilot.

4.1 Training the Trainer

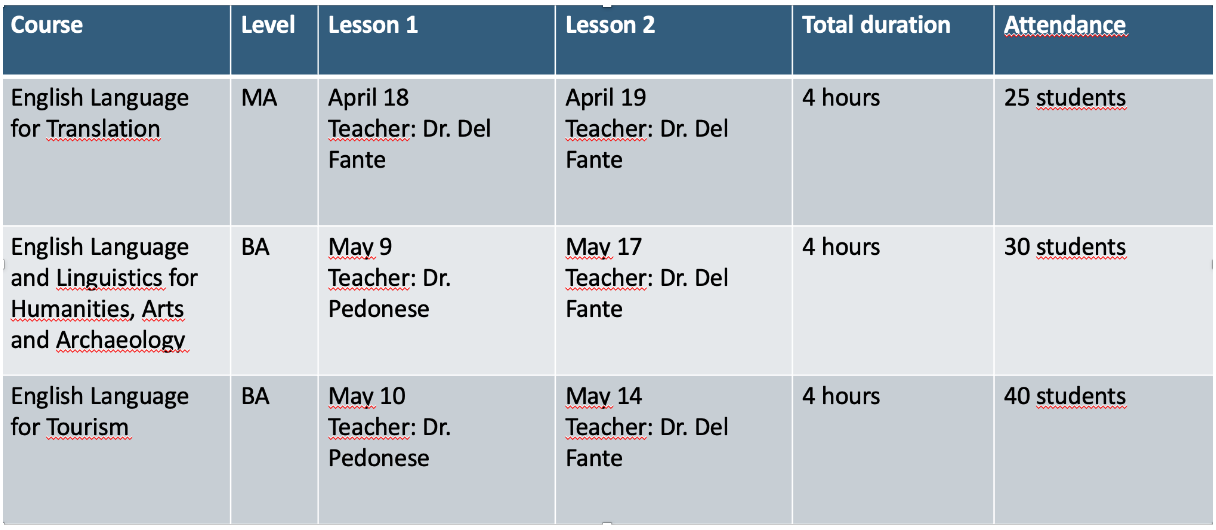
The first meeting was held online and consisted of the presentation of the primary teaching resources currently available for the Italian SSH community. The materials from the adapted UPSKILLS course and the facilitation guide developed within the H2IOSC project on how to reuse them were presented. A precious resource in this phase was the UPSKILLS "Guidelines on integrating Research Infrastructures into Teaching: Recommendations and Best Practices" (van der Lek et al. 2023 b), an up-to-date tool to enhance research-based teaching, presenting the case study of CLARIN. Since Dr. Del Fante was already familiar with CLARIN core services, the training session focused mainly on their latest functionalities and the discovery process of language resources that could be relevant to students.

Upon requests for specific translation technologies to teach students, further inquiries were made by the CLARIN-IT trainer, which resulted in a fruitful discussion with the CLARIN ERIC Training Officer, Iulianna van der Lek. This led to the definition of a CLARIN Café on Translation Technologies and Workflows for SSH Research,[10] held on June 14, 2024. The seminar was aimed at providing instructors, researchers, and CLARIN infrastructure staff with the tools to introduce these technologies in their teaching and research activities.[11] As the initiative originated directly from the training needs of CLARIN consortium members and was then extended to the infrastructure's broader community via the CLARIN Trainers' Network, it was one of the many results of the collaboration with the University of Ferrara.

[10] https://www.clarin.eu/event/2024/clarin-cafe-translation-technologies-workflows-ssh-researchers
[11] https://www.h2iosc.cnr.it/h2iosc-training-clarin-cafe/

4.2 Training the Students

For teaching the students of all three courses, we developed three modules of two lessons each to be held in person in Ferrara: 1) English Language Course for Tourism: two lessons of two hours each for 40 students; 2) English Language for Translation: two lessons of two hours each for 25 students; 3) English Language and Linguistics for Humanities, Arts and Archaeology: two lessons of two hours each for 30 students. For clarity, the lesson plan is shown in Fig. 6.



Course	Level	Lesson 1	Lesson 2	Total duration	Attendance
English Language for Translation	MA	April 18 Teacher: Dr. Del Fante	April 19 Teacher: Dr. Del Fante	4 hours	25 students
English Language and Linguistics for Humanities, Arts and Archaeology	BA	May 9 Teacher: Dr. Pedonese	May 17 Teacher: Dr. Del Fante	4 hours	30 students
English Language for Tourism	BA	May 10 Teacher: Dr. Pedonese	May 14 Teacher: Dr. Del Fante	4 hours	40 students

Fig. 6. Lesson Plan.

For the first lesson of each module, we adopted the same format, tailoring the examples and case studies to the disciplinary interests of each course. This introductory lesson offered a detailed overview of the CLARIN research infrastructure, emphasising the support provided by CLARIN Knowledge Infrastructure and giving information on funding opportunities. Additionally, we introduced the use of CLARIN Language Resource Families and the main functions of the metacatalogue Virtual Language Observatory.

We then dedicated the second lesson of each module to the practical application of CLARIN core services, and, in this case, we tailored the content to the specific needs of each course. Each resource and tool we chose to show the students was accessed through the Virtual Language Observatory to familiarise the class with CLARIN services, and we encourage students of all three courses to include those digital tools in their research projects. In the first course, we implemented the recently released ParlaMint 4.0[12] to analyse cross-linguistic political communication in Italy and the UK: we demonstrated how to use metadata to analyse textual data differently and provide a more comprehensive analysis of the texts. In the second course, we focused on using the parallel corpora available in the CLARIN infrastructure to assist translators. Specifically, we utilised the Intercorp parallel corpus (Čermák & Rosen 2012) to provide exercises addressing translation issues. We also utilised *Treq - the translation equivalent database*[13] - to compare the best translation equivalents and synonyms available in the corpus. For the third course, we demonstrated the utility of Voyant Tools in analysing the most recurrent features of promotional language used on different international tourist websites.

[12] ParlaMint is a CLARIN flagship project aimed at facilitating cross-linguistic and cross cultural analysis of political discourse by developing a multilingual and comparable corpus of parliamentary debates from 29 European countries convering the period from 2015 to 2022. The corpus contains over 1 billion words and features rich metadata on approximately 24,000 speakers. The data is uniformly encoded and linguistically annotated up to the level of Universal Dependencies syntax and named entities: https://www.clarin.eu/parlamint ParlaMint 4.0, released in October 2023, expanded the coverage, enhanced metadata and provided new data quality improvements.https://lindat.mff.cuni.cz/services/teitok/parlamint-40/index.php?action=subselect&id=GB

[13] https://treq.korpus.cz/index.php

5. Feedback and Assessment

Since this is a developing project, we submitted a feedback form to the students along with the final quiz to assess their knowledge of the course's core concepts in the short term and the quality of the course itself. The feedback and assessment form was the same for all three classes, and 20 participants completed it. In the quiz, students were asked questions such as defining Open Science, choosing the correct sequence of FAIR principles and completing a Data Management Plan definition. An average of 80% of the students answered these questions correctly, showing they had learnt about those topics and could comprehend and use CLARIN core services. As for the quality feedback, we asked the students to rate their satisfaction, effort, and relevance of the course activities on a scale of 1 to 5.

An important aspect to note was student satisfaction, which scored 3.67 out of 5. This figure, cross-referenced with the good performance of the participating students in the assessment exercises, shows that the course was well-tuned to the students' abilities. On the other hand, the students' perception of the usefulness of the topics covered was a critical aspect, which scored 3.42 out of 5. Even though lecturers always emphasised the possibilities of the CLARIN tools demonstrated in lectures, including this information within the final part of the courses may have been ungraded. They may have led students to think it was a secondary part.

For this reason, we have decided to reshape the second edition of the course in the 2024-2025 academic year by devoting more time to these topics from the first semester of classes and selecting a smaller cohort of students upstream, focused exclusively on language studies. This allowed us to take a more integrative approach to the students' university curriculum, which could accompany them in developing skills geared towards professional figures leaving university and eventually employed in the publishing and cultural industries.

6. Further Developments

Starting from the lessons learnt in this first experiment, for the 2024-2025 edition, which is still ongoing, we restricted the selection to two courses explicitly focusing on linguistics and translation technologies: English Language (MA) and English Language III (BA). We have also strengthened the assessment plan by including the final quiz and feedback form as necessary elements of the course evaluation to encourage all students' participation. In addition, we decided to add a preliminary questionnaire to test students' prior knowledge of the topics covered in the lectures, which in this case focused on machine and computer-aided translation tools. Surprisingly, all the students already use machine translation tools such as DeepL and Google Translate in their everyday lives, but no students can write down a definition of these technologies. So, we have tried to bridge the gap between their knowledge and their actual use of these tools, delving into the workings of the different types of machine translation and highlighting the possibilities of integrating them into computer-assisted tools such as MateCat.[14]

In addition, this time, we adapted and reused the materials of a new H2IOSC course along with the previous year's UPSKILLS materials. In a specific lesson restricted to MA students, we illustrated the fundamentals of the Linked Data paradigm, repurposing a part of the course Linguistic Linked Open Data for Humanists, which was initially created for international MA and PhD students participating in the Lisbon Summer School in Linguistics (Khan et al. 2024). As this new version of the pilot at the University of Ferrara is still in progress, we only have partial results from the first three lessons held in November 2024. However, we already saw a significant shift in the students' participation and engagement thanks to the more structured assessment plan and the interactive teaching strategies we implemented, primarily through gamification software like *Kahoot!* and Mentimeter.

Since the collaboration between CNR-ILC and the University of Ferrara remains active, we continued the lectures in the second semester. We further adapted the approach to another English language course for students in the master's degree program in Education, Communication and Digital Citizenship. In the lecture on April 9, data management aspects were emphasised, and a Data Management Plan

[14] https://www.matecat.com/

template was proposed along with the presentation of the data steward, who could be one of the professional outlets envisaged by the course of study.

Beyond the individual adaptations, the relevant aspect is the scalability of the approach, which includes an instructional design methodology that ensures easy reuse and updating of the material. The starter course, “Introduction to Language Data: Standards and Repositories,” was very intuitive to repurpose because it was structured in highly editable modules. Subsequent repurposing in H2IOSC allowed it to be placed within an integrated teaching infrastructure, and it ensured that a wider cohort of students with diverse backgrounds and different professional profiles could access the material.

7. Results

This series of training events at the University of Ferrara allowed us to collect insight into using CLARIN language resources and services in academic teaching activities and how to develop new modules reusing existing courses, namely the UPSKILLS training materials. Thanks to this experience, we could better understand the steps needed to train a trainer who is already familiar with CLARIN resources but has new and challenging requests to which CLARIN-IT needs to respond with the help of the broader international consortium. Another lesson learnt is the necessity of adapting methods, standards, tools and resources to make them relevant to the Italian community, which is only possible thanks to disseminating success stories on both national and international levels.

References

Čermák, F. – Rosen, A. 2012. The case of InterCorp is a multilingual parallel corpus. In *International Journal of Corpus Linguistics*, 17(3), 411–427.

Degl’Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fanini, and Francesca Frontini. 2023. ‘H2IOSC: Humanities and Heritage Open Science Cloud’. In *La Memoria Digitale: Forme Del Testo e Organizzazione Della Conoscenza. Atti Del XII Convegno Annuale AIUCD*, edited by Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, and Francesco Stella, 63–64. https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf.

Filiposka, Sonja. 2023. ‘D2.2 Methodology for FAIR-by-Design Training Materials’, August. https://doi.org/10.5281/ZENODO.8305540.

Frontini, Francesca, and Monica Monachini. 2023. ‘Infrastrutture Digitali per Le Scienze Umane e Sociali’. In *Digital Humanities. Metodi, Strumenti, Saperi*, 197–213. Roma: Carocci. https://www.carocci.it/prodotto/digital-humanities.

Garcia, Leyla, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, et al. 2020. ‘Ten Simple Rules for Making Training Materials FAIR’. *PLOS Computational Biology* 16 (5): e1007854. https://doi.org/10.1371/journal.pcbi.1007854.

Gledić, Jelena, Maja Đukanović, Jelena Budimirović, Nađa Soldatić, Maja Miličević Petrović, Silvia Bernardini, Adriano Ferraresi, et al. 2023. ‘UPSKILLS Teaching and Learning Content’. *Https://Upskillsproject.Eu/*, August. https://www.clarin.si/repository/xmlui/handle/11356/1865.

Hoebelheinrich, Nancy J, Katarzyna Biernacka, Michelle Brazas, Leyla Jael Castro, Nicola Fiore, Margareta Hellström, Emma Lazzeri, et al. «Recommendations for a Minimal Metadata Set to Aid Harmonised Discovery of Learning Resources», 9 June 2022. https://doi.org/10.15497/RDA00073.

‘Integrating Research Infrastructures into Teaching: Recommendations and Best Practices’. n.d. Accessed 6 February 2025. https://doi.org/10.5281/zenodo.8114407.

Khan, A. F., Pedonese, G., Mallia, M., Frontini, F., Quochi, V., & Squadrito, E. (5 July 2024). Linguistic Linked Open Data for Humanists. Lisbon Summer School in Linguistics, Lisbon. Zenodo. https://doi.org/10.5281/zenodo.13897931.

Pedonese Giulia, Frontini Francesca, Ottaviani Roberta, Boschetti Federico, Spadi Alessia, Francalanci Lucia, Scognamiglio Alessia, Restaneo Pietro, Chaban Antonina, Striova Jana, Benassi Laura ‘Materiali didattici come oggetti digitali FAIR: una metodologia condivisa per la formazione in H2IOSC’ (AIUCD 2024), *in* a cura di: Di Silvestro, Antonio; Spampinato, Daria (2024) Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale

AIUCD2024. Catania: AIUCD, p. 595. ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927. In: Quaderni di Umanistica Digitale.

'UNESCO Recommendation on Open Science - UNESCO Digital Library'. n.d. Accessed 1 December 2023. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en.

van der Lek, Iulianna, Darja Fišer, Francesca Frontini. «Introduction to Language Data: Standards and Repositories». In UPSKILLS Learning Content, 2023. (van der Lek et al., 2023 a) https://upskillsproject.eu/project/standards_repositories/.

van der Lek, Iulianna, Darja Fišer, Francesca Frontini, e Giulia Pedonese. «Introduzione ai Dati Linguistici: Standard e Archivi Digitali». 31 August 2024. https://doi.org/10.5281/zenodo.13911935.

van der Lek, Iulianna, Darja Fišer, Tanja Samardzic, Marko Simonovic, Stavros Assimakopoulos, Silvia Bernardini, Maja Milicevic Petrovic, e Genoveva Puskas. «Integrating Research Infrastructures into Teaching: Recommendations and Best Practices». Zenodo, 31 agosto 2023. (van der Lek et al., 2023 b) https://doi.org/10.5281/zenodo.8114407.

van der Lek, Iulianna, Francesca Frontini, Darja Fišer, Alexander König 2023. «Making the CLARIN Training Materials FAIR-By-Design», CLARIN Annual Conference 2023, Leuven 16 October - 18 October 2023. Deposited 20 January 2025. (van der Lek et al., 2023 c) https://doi.org/10.5281/zenodo.14699078.

Wilkinson, Mark, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences

Erik Körner, Thomas Eckart, Felix Helfer, Uwe Kretschmer
Saxon Academy of Sciences and Humanities in Leipzig
Leipzig, Germany
{koerner,eckart,helfer,kretschmer}@saw-leipzig.de

Abstract

The general idea and implementation of a federated search infrastructure component that allows querying both full-text resources and their linguistic annotations is a prominent part of the CLARIN project and is closely interconnected with the other components of its decentralised European-scale research data infrastructure. Since its beginnings, the Federated Content Search (FCS) has been continuously improved and by now fulfils its original goals that were formulated more than 12 years ago. During the last years, development of the FCS has accelerated massively with newly formulated application scenarios, newly opened up user groups and newly developed tools and user interfaces. This paper gives a summary of the developments of recent years and the topics that are currently being worked on. In addition to the further development of existing modules, this includes in particular the consideration and implementation of new requirements reflecting a rapidly evolving research infrastructure landscape.

1 Introduction to the Federated Content Search

The Federated Content Search (FCS) is a system for facilitating the discovery and retrieval of linguistic resources distributed across various data providers. The key idea is to enable users to search through different resource types (including text corpora and dictionaries) stored in multiple repositories in parallel through easy-to-use interfaces. Instead of requiring users to search each individual resource separately, the FCS aggregates results from these resources and provides a search capability that allows users to access and explore a wide range of linguistic content seamlessly. The FCS was developed in the context of the European CLARIN project. Today, it offers access to more than 500 resources[1] provided by scientific institutions from twelve European countries.

The general idea of the Federated Content Search was outlined in 2012 (Stehouwer et al., 2012) and mentioned a number of topics as central characteristics. They include searching in research data content – as opposed to searching metadata records – , support of distributed resources, use of a standard protocol (SRU/CQL, OASIS, 2013), and consideration of possible future extensions. The FCS as it is today represents an easy-to-use solution to execute complex linguistic queries on large distributed data sets. It does not replace the powerful corpus query engines already in use, but provides a lightweight interface to run parallel queries on them. The connection between the FCS and the respective local search engine is made via so-called “FCS endpoints” which map FCS queries to the locally supported query language and convert local data formats into FCS-compatible formats.

In the aforementioned publication, various topics were described as possible future priorities, including the provision of bindings for popular data indexing systems, integration with other CLARIN services, support for access-restricted resources via suitable control mechanisms, and the use of ISOcat (Kemps-Snijders et al., 2009) as a semantic integration layer between the common infrastructure and specific endpoint implementations. Version 1.0 of the FCS specification (Schonefeld et al., 2014) two years after

[1] FCS Aggregator, https://contentsearch.clarin.eu/, as of April 2025

Erik Körner, Thomas Eckart, Felix Helfer and Uwe Kretschmer 2024. An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 48–59. DOI: https://doi.org/10.3384/ecp216

the original paper incorporated large parts of these ideas and was superseded by version 2.0 (van Uytvanck et al., 2017) in 2017, which introduced a new query language *FCS-QL* and a new data view to enhance access to annotated text corpora.

The vast majority of the initially planned ideas have now been implemented and, in most cases, continuously improved and expanded in several iterations. Of the original ideas, only two topics are currently still unfinished, which have significantly different probabilities for implementation: the support of access mechanisms for querying protected resources – which is currently in progress – and the use of ISOcat as an integration layer which is no longer part of the development goals[2].

The FCS continues to be a central component of the CLARIN infrastructure and is a prominent part of its current work plan, but has also established itself in institutions and projects beyond this. A prominent example is the establishment of the FCS as the central search component in the German infrastructure consortium *Text+*[3], which has already significantly increased the number of available resources. In the same context, new user interfaces are currently being developed to improve usability and user-friendliness and new usage scenarios are increasingly being supported that benefit from the flexible extensibility of the FCS standard.

2 New Requirements in a Changing Resources Landscape

The original concept of FCS focussed on resources that are mostly available as tokenized continuous text (such as plain text documents, text corpora, or text collections) with optional token-based linguistic annotations that are freely accessible in a decentralized infrastructure.

These key aspects are still valid for most of the resources provided. However, it is becoming increasingly clear that this does not adequately reflect the diversity of linguistic resources or resource types and that the all too common scenario of trying to access protected resources (due to publisher restrictions, copyright or personal rights, etc.) limits the availability of text resources that are highly relevant for research. Research projects are often oriented towards and organised based on specific resource types, with sometimes significantly varying technical requirements and traditions. In the German *Text+* project, for example, this is illustrated by its organisation into three data domains “Collections”, “Lexical Resources” and “Editions”. During years of project work, it became clear that the FCS in its current form not always sufficiently covers their requirements. This becomes especially obvious for lexical resources (such as dictionaries, word lists, lexical-semantic networks), which differ greatly from full-text-oriented resources due to their sometimes complex internal structure and which also require a different set of queryable types of information.

The increasing importance of central knowledge bases – such as Wikidata (Vrandečić & Krötzsch, 2014), VIAF[4], or the Integrated Authority File GND (German National Library (DNB), 2024) – for the discovery of and linkage between digital resources in the humanities research cannot be sufficiently supported within the existing FCS implementation. This support of authority files or referenceable knowledge bases with similar functionality is addressed within the *EntityFCS* specification. The first corpora and dictionaries that make use of this functionality are already available.

In comparison, the support of access-restricted resources via an Authentication & Authorization Infrastructure (AAI) is a long-standing requirement. This includes the option for users to log in via their known identity provider if required and to forward attributes to FCS endpoints that can allow access to resources that are only available to restricted user groups or even individual users. Seamless integration into existing user interfaces is a key aspect in ensuring a high level of usability here. However, developments in recent years have also led to adapted requirements in this area. Especially “reference-only results” – indicating only the presence of a result at a specified location – or support of *derived text formats* (Schöch et al., 2020) are being discussed for cases where the actual content cannot be provided due to contractual or legal reasons.

[2] Due to its discontinuation in 2014.

[3] Text+, https://www.text-plus.org/en

[4] Virtual International Authority File, https://viaf.org/

Another problem that illustrates the diversity of language resources in the context of incomplete standardisation is the inadequate representation of character sets, which often occur in the field of historical languages and which are not always adequately supported by standards such as Unicode[5] or by established and freely available fonts. Especially for a federated search platform, resulting display artifacts pose a major problem for an appropriate representation of results. This problem is addressed by an optional extension of the FCS specification that allows endpoints to provide font information for resources, which can be presented to the user in the search interfaces.

This paper is structured along these new requirements, starting with a brief introduction as to why the FCS is a solid foundation for these types of customisations: the flexible structure of its specification that allows backward-compatible extensions (section 3). In the following sections, these extensions are described in more detail starting with the support of new resource types in the *LexFCS* (section 4) and new kinds of requests in the *EntityFCS* (section 5). The support of queries on access-restricted resources by integrating the FCS into an Authentication and Authorization Infrastructure (AAI) and of appropriate UI representation of resources via external font support is explained in section 6 and section 7. In section 8 an overview is given of different means to improve usability by a revised development and documentation process which significantly facilitates participation in the development process and the provision of own resources and their use by end users. The publication concludes with suggestions for improvement that are currently being evaluated (section 9) or are part of the short- and medium-term time plan.

3 Flexibility through the FCS Extension System

Extensibility and backwards compatibility are key factors and important design principles of the FCS. The SRU/CQL protocol (OASIS, 2013) contains a number of supporting mechanisms that were the reason for its choice as the technical basis of the platform. This allows the FCS to adapt to evolving user requirements while maintaining compatibility for endpoints and client libraries.

This flexibility includes extended API calls, support for new data formats (so-called "Data Views") and new query languages or optional extensions to standardized schemata. All work presented in the following sections ultimately uses these mechanisms. Extended endpoint implementations are – in the vast majority of cases – still accessible to existing client libraries and fit easily into the existing infrastructure.

Figure 1 summarises the overall FCS architecture, highlighting current extensions of its ecosystem.

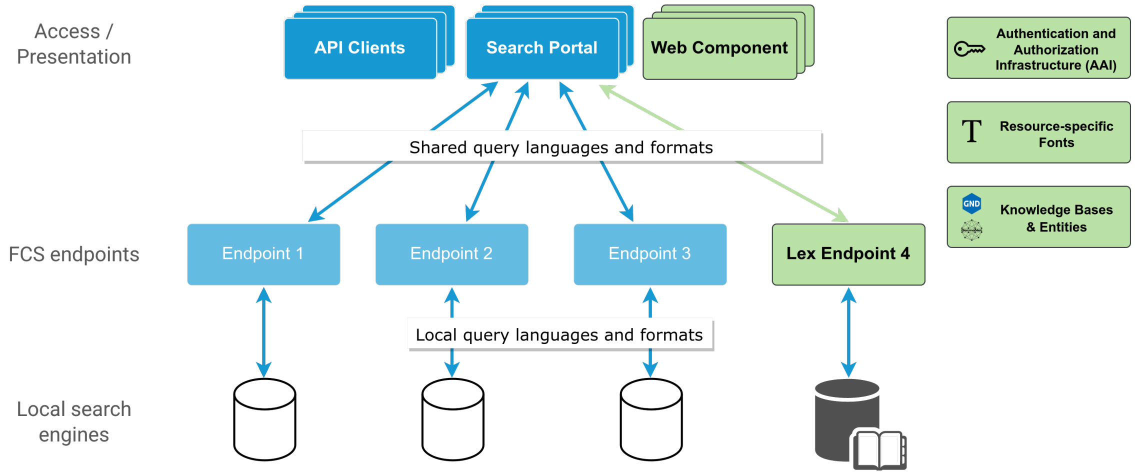


Figure 1: General architecture of the FCS, highlighting currently developed extensions in green

4 LexFCS – Lexical Resources in the FCS

The FCS was initially developed to offer a federated content search on simple full texts but later extended for more complex linguistic search patterns on tokenized texts with optional annotation layers. Common to these is that searches are based on "flat" text sequences, making it difficult to map more

[5] Unicode – The World Standard for Text and Emoji, https://unicode.org

complex structures into this format. However, this excludes resource types such as lexical resources, including dictionaries, word lists, or semantic wordnets, which typically have a more complex structure, like graphs, feature structures, or simple property value structures.

A first working proposal had been prepared by a dedicated working group in Text+ for a new *LexFCS* extension (Eckart et al., 2023). First prototypes of FCS endpoints and result presentation in FCS clients were developed based on an early and temporary specification. Collected feedback and input from discussions at conferences and workshops were incorporated into the current, revised version of the LexFCS specification (Körner et al., 2024) that proposes the following two major additions, which can only be presented here briefly.

(1) A key-value based data model which structures lexical information in *Entries* that contain typed *Field* elements each having *Values* and supplementary attributes, and which is serialized into a new **Lex Data View**. Each *Entry* represents an individual, self-contained dictionary entry, containing information about a lemma without making assumptions about its type. Fields group properties of a lemma by their information type such as basic information (e.g., word form, spelling variants, identifiers, transcriptions), relations to other entries or external resources (references, citations), morphosyntactic (part-of-speech, segmentation), semantic (like sentiment, synonyms, hyponyms), or frequency-related data as well as more prosaic information (definition, etymology). The actual content is given in *Value* elements and may contain additional meta information about content language, internal or external references, and more.

The internal reference and grouping mechanism is achieved via IDs which allow custom highlighting in the user interface of related elements but can also be used to structure information hierarchically based on given definitions and etymologies.

Regardless of this, the representation of a lexical record in the style of a printed dictionary with additional annotations is now supported with the **LexHITS Data View**. It is a backward-compatible upgrade of the HITS Data View that allows inline annotation of lexical information in full text.

(2) Extending the *Contextual Query Language* (CQL[6]) standardized by the US Library of Congress and the standards organisation OASIS, we propose **LexCQL** as the query language dedicated to querying lexical entries. It defines searchable indexes based on the *Fields* of the aforementioned data model with relations and relation modifiers to allow both simple and complex queries.

LexCQL supports the default search relation = for relaxed and the additional == for strict equality matching. Relation modifiers for case sensitivity, (umlaut) normalization, matching, etc. can be used to refine queries. In addition, the new relation `is` can be used to search based on concept or reference URIs instead of plain value strings, e.g. requesting a noun via its definition according to the Universal Dependencies project (`https://universaldependencies.org/u/pos/NOUN`) instead of highly ambiguous search terms (like “`N`”). Boolean operators `AND`, `OR`, `NOT` and parentheses allow to combine and build complex queries.

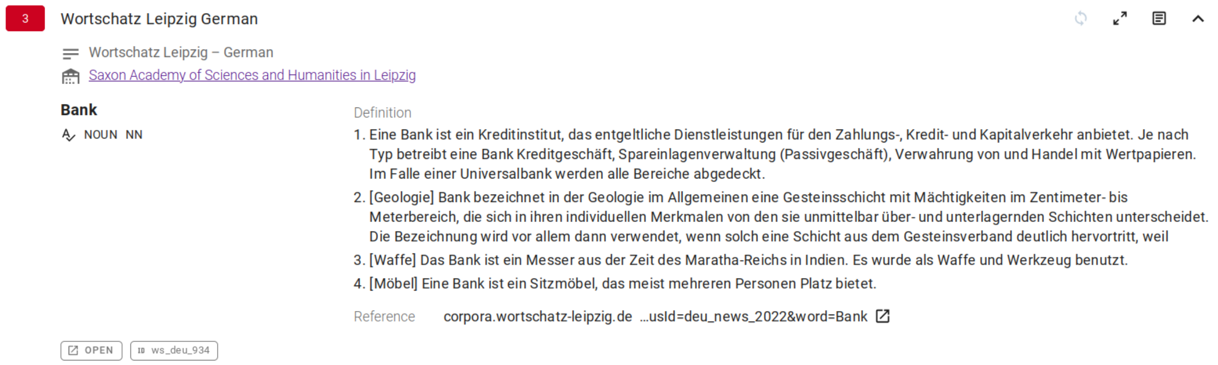
Based on the *Values*’ language information and a language index for the overall *Entry*, queries in multilingual dictionaries are also supported.

Current work has been focused on the standardisation of terminology and to provide a data model and query language to represent and search through potentially complex lexical records while maintaining a high level of flexibility to allow compatibility with many types of lexical resources. The Text+ FCS Aggregator[7] already includes a first implementation of this extension with a few endpoints providing dictionary data. Another way of representing LexFCS results has been implemented in an alternative user interface (cf. Figure 2, more about this search integration in section 8).

Ongoing work will address further technical details to provide improved user interfaces and to offer guidelines for new data providers about available features and on how to integrate their own custom formats into the LexFCS.

[6] Please note that in this paper the term “CQL” is always referring to the *Contextual Query Language*.

[7] LexFCS search in the Text+ FCS Aggregator, https://fcs.text-plus.org/?queryType=lex



Wortschatz Leipzig German

Wortschatz Leipzig – German

Saxon Academy of Sciences and Humanities in Leipzig

Bank

NOUN NN

Definition

1. Eine Bank ist ein Kreditinstitut, das entgeltliche Dienstleistungen für den Zahlungs-, Kredit- und Kapitalverkehr anbietet. Je nach Typ betreibt eine Bank Kreditgeschäft, Spareinlagenverwaltung (Passivgeschäft), Verwahrung von und Handel mit Wertpapieren. Im Falle einer Universalbank werden alle Bereiche abgedeckt.
2. [Geologie] Bank bezeichnet in der Geologie im Allgemeinen eine Gesteinsschicht mit Mächtigkeiten im Zentimeter- bis Meterbereich, die sich in ihren individuellen Merkmalen von den sie unmittelbar über- und unterlagernden Schichten unterscheidet. Die Bezeichnung wird vor allem dann verwendet, wenn solch eine Schicht aus dem Gesteinsverband deutlich hervortritt, weil
3. [Waffe] Das Bank ist ein Messer aus der Zeit des Maratha-Reichs in Indien. Es wurde als Waffe und Werkzeug benutzt.
4. [Möbel] Eine Bank ist ein Sitzmöbel, das meist mehreren Personen Platz bietet.

Reference corpora.wortschatz-leipzig.de ...usId=deu_news_2022&word=Bank

Figure 2: Representing the micro structure of a dictionary entry using the new Lex Data View in a user interface prototype,

5 EntityFCS – Entity-oriented Search

Authority files are important tools to ensure consistency and accuracy in the identification of entities such as persons, locations, organizations, events, and more. Resources that are annotated accordingly allow access to these entities despite language boundaries, potential variations in spelling or transliteration and – even more important – in case of lexical ambiguity the specific intended meaning can be described and referenced. Authority files have therefore the potential to improve data discoverability and interoperability in a distributed and diverse research environment significantly.

In the field of Humanities and Social Sciences, various knowledge bases are increasingly used to enhance existing resources or are considered from the start for born-digital resources. In the context of this section, the term "authority file" is meant in a deliberately broad sense, including all systems that define specific meanings for entities in a field of knowledge and provide persistent means for their identification and global access. As a consequence, not only traditional authority files like the *Virtual International Authority File* VIAF or the *Integrated Authority File* GND (German National Library (DNB), 2024) are considered, but collaborative knowledge bases such as Wikidata (Vrandečić & Krötzsch, 2014) or GeoNames[8] and sense-based lexical databases like GermaNet (Hamp & Feldweg, 1997) and Princeton WordNet (Fellbaum, 1998). The growing importance of Knowledge Graphs for structuring diverse research landscapes and providing access to resources is only conceivable taking such semantic "anchors" into account.

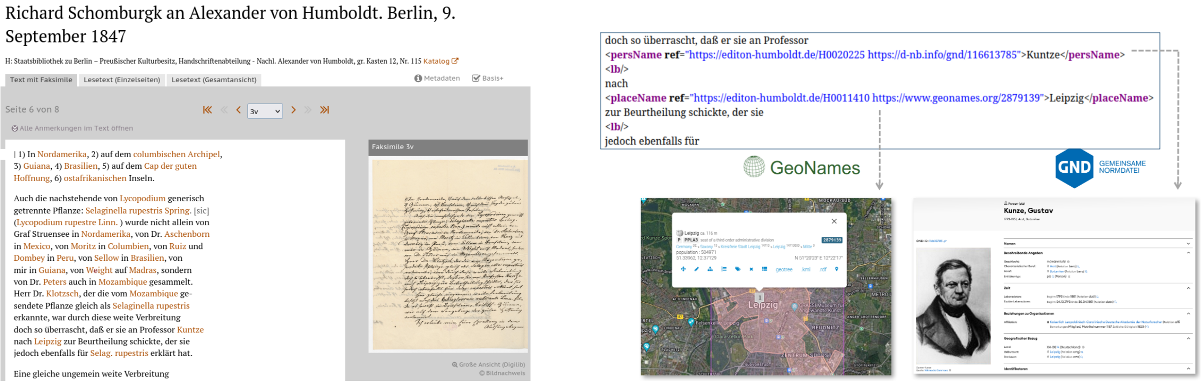
Figure 3 demonstrates the use of two of these knowledge bases in letters to Alexander von Humboldt at the *edition humboldt digital*[9].

The task of resolving a mention in a text to an appropriate entry in an authority file is called *entity linking* (Hachey et al., 2013). While entity linking can greatly enrich textual resources, the process is often laborious and difficult, especially for linking to a large authority file, where a mention could refer to multiple different entries (e.g. different people with the same name). In addition, previous automated approaches often underperform, especially for non-English languages (see benchmark by Schwarz and Barth, 2024). Hence, data with annotations of this kind are still quite rare. However, with the continued development of new approaches to entity linking, particularly those utilizing the advancement of large language models, such as Qi et al., 2024 and knowledge graphs, such as Ayoola et al., 2022, it is to be expected that data with mentions linked to an authority file entry will become more common over time. Search applications that support these types of annotations can then further help with the accessibility and usability of the referenced data.

The *entity-oriented search* paradigm (Balog, 2018) focuses on retrieval mechanisms on the basis of

[8] GeoNames, https://www.geonames.org

[9] edition humboldt digital, https://edition-humboldt.de



Richard Schomburgk an Alexander von Humboldt. Berlin, 9. September 1847

H: Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Handschriftenabteilung - Nachl. Alexander von Humboldt, gr. Kasten 12, Nr. 115 Katalog

Seite 6 von 8

| 1) In Nordamerika, 2) auf dem columbischen Archipel, 3) Guiana, 4) Brasilien, 5) auf dem Cap der guten Hoffnung, 6) ostafrikanischen Inseln.

Auch die nachstehende von Lycopodium generisch getrennte Pflanze: Selaginella rupestris Spring. [sic] (Lycopodium rupestre Linn.) wurde nicht allein von Graf Struensee in Nordamerika, von Dr. Aschenborn in Mexico, von Moritz in Columbien, von Ruiz und Dombey in Peru, von Sellow in Brasilien, von mir in Guiana, von Weight auf Madras, sondern von Dr. Peters auch in Mozambique gesammelt. Herr Dr. Klotzsch, der die vom Mozambique gesendete Pflanze gleich als Selaginella rupestris erkannte, war durch diese weite Verbreitung doch so überrascht, daß er sie an Professor Kuntze nach Leipzig zur Beurtheilung schickte, der sie jedoch ebenfalls für Selag. rupestris erklärt hat.

Eine gleiche ungemein weite Verbreitung

(a) Online presentation,

```
doch so überrascht, daß er sie an Professor
<persName ref="https://editon-humboldt.de/H0020225 https://d-nb.info/gnd/116613785">Kuntze</persName>
<lb/>
nach
<placeName ref="https://editon-humboldt.de/H0011410 https://www.geonames.org/2879139">Leipzig</placeName>
zur Beurtheilung schickte, der sie
<lb/>
jedoch ebenfalls für
```

(b) TEI encoding of authority file references

Figure 3: Referencing entities in a letter to Alexander von Humboldt using GeoNames and the GND (Project *edition humboldt digital*)

linked entities. However, the FCS in its original specification does not offer any support for this kind of queries. The focus on linguistic annotations with a limited set of annotation layers designed for this dedicated purpose (such as lemma, orthographic transcription and similar) currently supports queries best through the use of the *pos* layer, with which – using suitable vocabulary – at least the information on entity types can be accessed.

The *EntityFCS* extension of the FCS supports search on entity- and sense-annotated full texts as well as lexical resources using global references. As all supported FCS query types have different specifics, there are slightly different characteristics in the respective specification customisation. They both entail modifications to the query language and support of the entity annotation layer in the respective data view:

- **Advanced Search** queries allow searching on the additional layer type "entity" as an optional extension to the existing annotation layers. This is represented as a new layer both in the query language *FCS-QL* and in the serialization in the *Advanced Data View*.
- **Lexical Search** queries allow access to entity references via the field "senseRef" that is supported both in the query language *LexCQL* and in the *Lex Data View*.

Figure 4 shows an example result set in a new web interface (for more details, see section 8), where all textual references to the German city Leipzig are queried based on its GND identifier in a corpus of the project "Letters and Records regarding the Church Politics of Frederick the Wise and John the Constant from 1513 to 1532"[10]. The short description and illustrative image in the blue tooltip are dynamically loaded from the *lobid* API[11].

The described functionality is already part of the current LexFCS specification (Körner et al., 2024). Its support in the multi-layer advanced search is still in the preparatory stage.

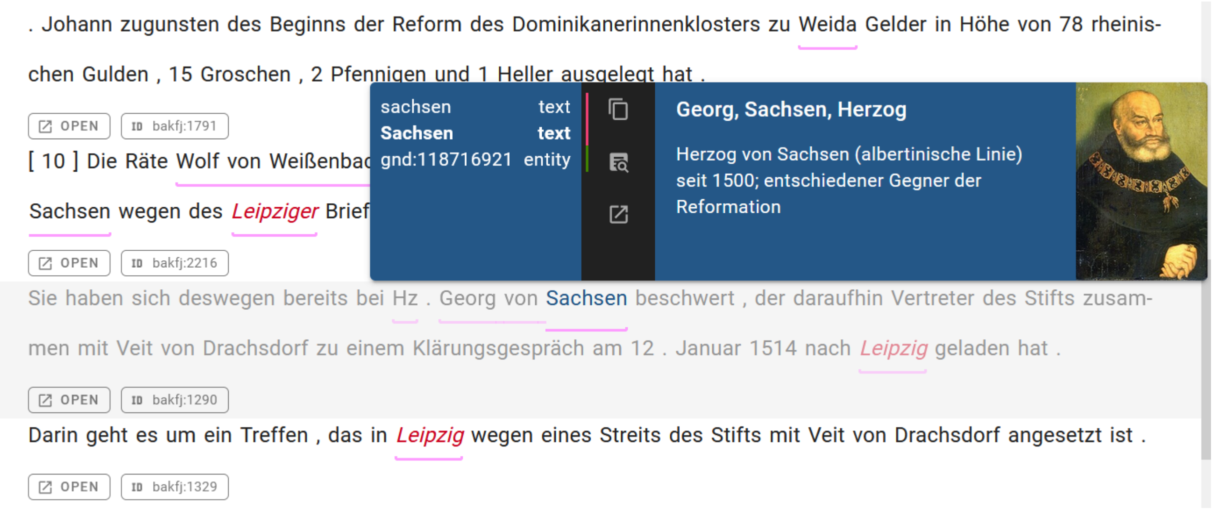
6 AAI – Querying Access-restricted Resources

Data protection and copyright laws, licensing, and other reasons often impede general public access to linguistic resources. In order to make restricted content available via the FCS anyway, with interested parties having a better overview of potentially relevant data and data providers having the option to offer at least some kind of access, integration of a decentralized *Authentication and Authorization Infrastructure* (AAI[12]) is required. This enables data providers to check search requests to their endpoints if they

[10] Project homepage, https://bakfj.saw-leipzig.de/

[11] Linked Open Data (LOD) services for libraries, https://lobid.org/index-en

[12] A brief overview on "What is AAI?", https://aai.mpg.de/aai.shtml



. Johann zugunsten des Beginns der Reform des Dominikanerinnenklosters zu Weida Gelder in Höhe von 78 rheinischen Gulden , 15 Groschen , 2 Pfennigen und 1 Heller ausgelegt hat .

OPEN ID bakfj:1791

[10] Die Räte Wolf von Weißenbac

Sachsen wegen des *Leipziger* Brief

sachsen	text
Sachsen	**text**
gnd:118716921	entity

Georg, Sachsen, Herzog

Herzog von Sachsen (albertinische Linie) seit 1500; entschiedener Gegner der Reformation

OPEN ID bakfj:2216

Sie haben sich deswegen bereits bei Hz . Georg von Sachsen beschwert , der daraufhin Vertreter des Stifts zusammen mit Veit von Drachsdorf zu einem Klärungsgespräch am 12 . Januar 1514 nach *Leipzig* geladen hat .

OPEN ID bakfj:1290

Darin geht es um ein Treffen , das in *Leipzig* wegen eines Streits des Stifts mit Veit von Drachsdorf angesetzt ist .

OPEN ID bakfj:1329

Figure 4: Entity-oriented search for the GND entity “4035206-7” (city of Leipzig),

meet their specific identification or authorization requirements. *Shibboleth*[13] with its support for Single Sign-on (SSO) authentication in particular, has established itself for these purposes and is widely used, which makes it an ideal candidate for integration into the FCS. This implies to a greater extent that users have an academic background, but this is also reflected in access requirements for restricted resources, which is often *academic users only*. However, the distributed nature and different requirements for data protection and applications lead to very diversely configured identity providers, so available attributes to describe user identities are limited to email or anonymous identifiers[14] which only allows for relatively shallow verification.

The AAI specification extension proposes the integration of FCS clients (like the FCS Aggregator) as Service Providers where users can be authenticated if required. This authentication status will be forwarded securely and verifiably to FCS endpoints that require this information for access to their resources. The endpoint decides if the provided information meets the resource’s requirements, e.g. by checking for supported affiliations or user IDs, and return the results in case of success. In cases where this approach is not sufficient, dedicated diagnostics in the FCS protocol allow to state that results are present but can not be provided with an included reference to where users can request access. This “last resort” approach can at least provide some overview and an indication of potentially relevant data records at an institution for a user request.

The CLARIN FCS task force is currently reviewing a draft of the specification[15] and initial prototypes (including software libraries and deployment) are being evaluated to check if all requirements are met.

7 Supporting Resource-specific Fonts

The Unicode[5] standard comprises a wide variety of characters and symbols and enables the encoding of most texts. Nevertheless, as a global standard there are limitations on what is, can and will be included in the standard. In particular, historical and newer, non-standard applications make use of *Private Use Areas* in UTF-8 to map their own symbols into unused areas. This leads to issues for the general public when texts require custom fonts for correct rendering (e.g. font issues for lexical resources in Figure 5).

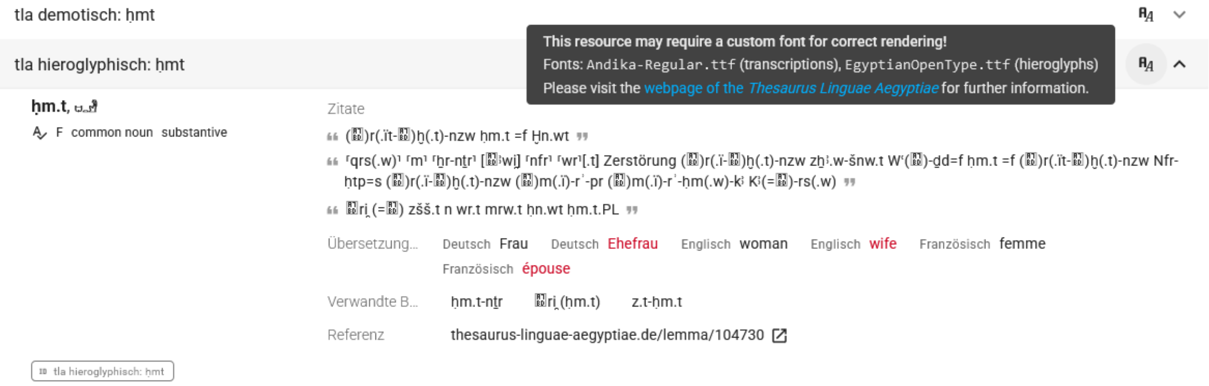
In the humanities, the necessity for custom fonts is particularly pronounced due to the diverse range of

[13] Identity management software, Shibboleth Consortium, https://www.shibboleth.net/

[14] The SAML attributes *eduPersonPrincipalName*, *eduPersonTargetedID*, or *mail*, were determined to be the minimal set that most Identity Providers can offer and whichever information is available will be used, see https://tools.aai.dfn.de/entities/ for an overview in the DFN-AAI network.

[15] FCS AAI specification (draft), https://clarin-eric.github.io/fcs-misc/fcs-aai-specs/fcs-aai.html

scripts, symbols, and notational conventions used across different disciplines. Standardization processes, such as in the case of Unicode, often take considerable time, sometimes spanning several years. This delay poses challenges for projects that require immediate solutions for encoding and rendering specialized texts. Even when an encoding standard is eventually extended, integrating these updates into widely available fonts is not instantaneous, as font developers need additional time to implement and distribute the changes. Consequently, researchers and institutions must rely on custom fonts as interim solutions to ensure accessibility and readability of critical textual resources.



tla demotisch: ḥmt

tla hieroglyphisch: ḥmt

This resource may require a custom font for correct rendering!
Fonts: `Andika-Regular.ttf` (transcriptions), `EgyptianOpenType.ttf` (hieroglyphs)
Please visit the webpage of the *Thesaurus Linguae Aegyptiae* for further information.

ḥm.t, ꜥ

F common noun substantive

Zitate

“ (□)r(.ı͗t-□)ẖ(.t)-nzw ḥm.t =f Ḥn.wt ”

“ ⸢qrs(.w)⸣ ⸢m⸣ ⸢ẖr-nṯr⸣ [□ꜣwi͗] ⸢nfr⸣ ⸢wr⸣[.t] Zerstörung (□)r(.ı͗-□)ẖ(.t)-nzw zẖꜣ.w-šnw.t Wꜥ(□)-ḏd=f ḥm.t =f (□)r(.ı͗t-□)ẖ(.t)-nzw Nfr-ḥtp=s (□)r(.ı͗-□)ẖ(.t)-nzw (□)m(.ı͗)-rꜣ-pr (□)m(.ı͗)-rꜣ-ḥm(.w)-kꜣ Kꜣ(=□)-rs(.w) ”

“ □ri̯ (=□) zšš.t n wr.t mrw.t ḥn.wt ḥm.t.PL ”

Übersetzung...	Deutsch	Frau	Deutsch	Ehefrau	Englisch	woman	Englisch	wife	Französisch	femme
	Französisch	épouse								

Verwandte B... ḥm.t-nṯr □ri̯ (ḥm.t) z.t-ḥm.t

Referenz thesaurus-linguae-aegyptiae.de/lemma/104730

tla hieroglyphisch: ḥmt

Figure 5: Font issues in *Thesaurus Linguae Aegyptiae* with information about custom fonts

A few interesting use-cases have already been identified where data providers plan to apply this new font extension: the *Thesaurus Linguae Aegyptiae* (BBAW, 2025), which requires fonts for both Egyptian texts and transcriptions, the *Hamburg Sign Language Notation System* (HamNoSys, Hanke, 2021), the *KompLett* font used for historical German texts, and *Landsmålsalfabetet* (Swedish Dialect Alphabet, Lundell, 1928) used for documenting dialects.

The proposal aims to provide a simple augmentation of the Endpoint Description to specify which resources may require certain fonts. Integrations in FCS clients can then either automatically apply the font in their front-end for results or offer a hint to users that further actions (e.g., downloading and installing a specific font) are required to enable correct rendering. Furthermore, the integration of information about required fonts and their sources not only enhances the representation and communication within FCS clients but also provides direct benefits for researchers who download FCS search results for further local analysis. Ensuring that font requirements are clearly documented allows scholars to work with the data in their own analytical environments without encountering unexpected rendering issues.

To ensure seamless integration of custom fonts within the FCS framework, a standardized approach to font metadata must be defined. This includes specifying font requirements within the metadata of the respective resources, enabling clients to recognize and apply the appropriate fonts dynamically. Additionally, font distribution mechanisms should be considered to facilitate user access, such as linking to web-hosted font repositories or embedding font files where permissible. Moreover, it is essential to address potential challenges related to the licensing and copyright constraints of custom fonts. Many specialized fonts are either proprietary or have restrictive licensing conditions that may prevent straightforward distribution. Therefore, FCS clients and providers must implement mechanisms to inform users about licensing terms and potential restrictions while offering alternatives, such as open-source fonts, whenever possible.

Currently, active work is underway to develop and refine this extension. Initial prototypes have already been built, demonstrating the feasibility of integrating font metadata into the FCS framework. Furthermore, discussions and exchanges with the community are ongoing to gather feedback and improve the approach based on real-world use cases and requirements. This will contribute to a more user-friendly and inclusive experience for diverse linguistic and historical text resources.

8 Improving Usability for all User Groups

An open, federated system like the FCS depends on a high degree of usability for all types of users, including developers, resource providers or end-users. In recent years, the FCS ecosystem has changed significantly to accommodate this goal. This includes a shift to popular working environments (GitHub/GitLab[16]) for supporting modern, open development processes, an expanded documentation based on easily editable formats (e.g. AsciiDoc[17]), provision of an increasing number of software libraries for a wide variety of use cases and the development of improved user interfaces.

To make the code base and documentation more accessible, it was moved to GitHub.com. The working format for the FCS specification and other documents[18] was changed to AsciiDoc, to support improved editing, cross-linking and various output formats. Additional material like FCS Endpoint Development tutorials and extensive presentation slides[19] – focusing on different user groups – were created to ease development efforts.

Besides the documentation itself, various communication channels and activities are offered to support end-users, endpoint developers and operators. This includes the CLARIN Forum[20] to transparently publish news as well as allowing user interaction and feedback, workshops for FCS endpoint development (e.g., organised by the Saxon Academy of Sciences and Humanities in Leipzig (SAW) for Text+[21]) and support via help desks. More hackathons for developers and workshops for end-users focusing on specific usage scenarios are currently planned. All information material created in the process will be made available on the respective information channels as well.

On the software side, reference libraries that were initially developed in `Java` have now also been translated to `Python` to accommodate a greater variety of technology stacks. These libraries are continually being improved and extended. Numerous additional endpoint implementations in other languages exist as well – the open-source ones can be found in the aforementioned *Awesome FCS* list. To further assist FCS developers, various support tools are provided, many of them *dockerized* for an easier setup. A particularly important application is the **FCS Endpoint Validator**[22] which provides immediate feedback about the conformance of an endpoint to the FCS specification and which has been completely rewritten and extended to cover more test cases. It now offers additional features for configuration and reporting in response to evolving user requirements.

To facilitate end-user engagement, the FCS's central search application (the *FCS Aggregator*) has been revised several times since its initial release and has been continuously developed further. The latest major changes include an improved RESTful API[23] which can also be accessed from external applications. There is also a new Web component based on this API using the `Vue.js` framework. This component allows new forms of access to FCS endpoints and the presentation of results in an alternative web interface. In particular, this application supports a first integration of the EntityFCS based on the GND (cf. Figure 4), the improved presentation of lexical resources (cf. Figure 2), and also the option of integrating this new search interface into your own application with little effort. This is already implemented as part of the FCS integration in the central Text+ portal[24] (cf. Figure 6) and in the first repository websites that want to support search in their local resources[25] with a user-friendly interface.

9 Conclusion and Further Work

The Federated Content Search has proven and established itself as a platform that can react flexibly to changing or new requirements thanks to its open architecture and can therefore be easily integrated into

[16] For an overview of related repositories, see https://github.com/clarin-eric/awesome-fcs

[17] A plain text markup language for writing technical content, https://asciidoc.org/

[18] Overview page of compiled documents: https://clarin-eric.github.io/fcs-misc/

[19] Presentations slides for FCS endpoint development, https://clarin.eu/fcsdevguide

[20] All topics tagged with *fcs* in the CLARIN Forum, https://forum.clarin.eu/tag/fcs

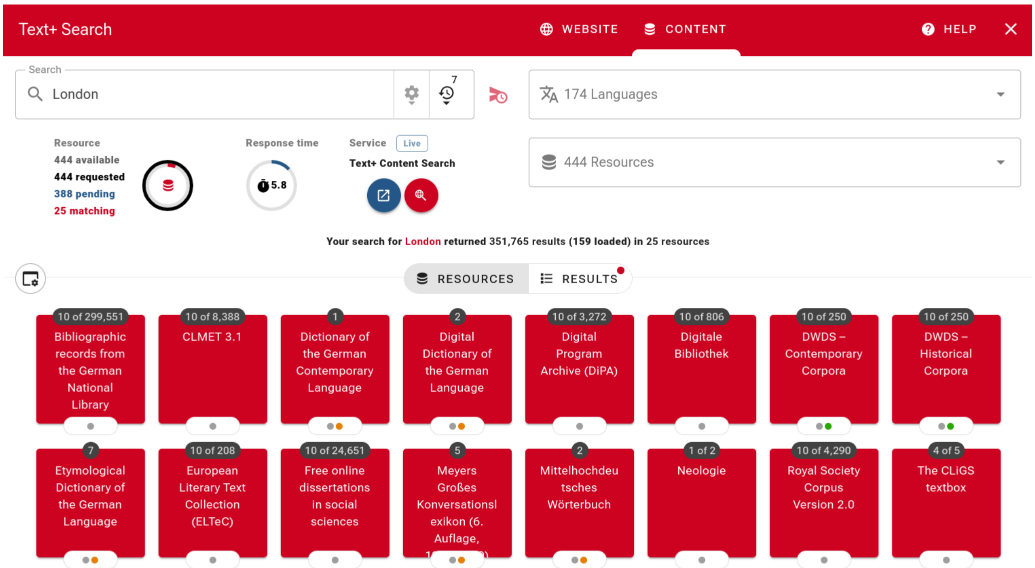
[21] Blog post about the FCS Endpoint Development Hackathon, https://textplus.hypotheses.org/9750 (in German)

[22] Official FCS Endpoint Validator for FCS / SRU protocol conformity and feature checks, https://www.clarin.eu/fcsvalidator

[23] OpenAPI-compliant description of the FCS Aggregator REST API, https://contentsearch.clarin.eu/openapi.json

[24] FCS Web Component integration in Text+ webpage, https://text-plus.org/en#action-open-search?tab=content

[25] FCS integration at the SAW Leipzig repository with a filtered resources list, https://repo.data.saw-leipzig.de/en#open-fcs



Search: London

7

174 Languages

444 Resources

Resource
444 available
444 requested
388 pending
25 matching

Response time 5.8

Service Live
Text+ Content Search

Your search for London returned 351,765 results (159 loaded) in 25 resources

10 of 299,551 Bibliographic records from the German National Library

10 of 8,388 CLMET 3.1

1 Dictionary of the German Contemporary Language

2 Digital Dictionary of the German Language

10 of 3,272 Digital Program Archive (DiPA)

10 of 806 Digitale Bibliothek

10 of 250 DWDS – Contemporary Corpora

10 of 250 DWDS – Historical Corpora

7 Etymological Dictionary of the German Language

10 of 208 European Literary Text Collection (ELTeC)

10 of 24,651 Free online dissertations in social sciences

5 Meyers Großes Konversationslexikon (6. Auflage, 1

2 Mittelhochdeutsches Wörterbuch

1 of 2 Neologie

10 of 4,290 Royal Society Corpus Version 2.0

4 of 5 The CLiGS textbox

Figure 6: Result overview for a FCS query on the webpage of the Text+ project, 🔗

new research contexts. In recent years, specific application-driven usage scenarios have led to various lightweight extensions and customisations, many of which are already in active use. The work presented not only illustrates the dynamics of the development to date, but also forms the basis for further short and medium-term planning. The focus for the coming years will be on further lowering the entry barrier for users and developers and adapting current developments in the area of large language models, such as the use of the FCS in the context of Retrieval Augmented Generation (RAG).

More FCS extensions are currently being worked on, but are still in the user requirements analysis phase or consist of only initial prototypes for evaluation and testing purposes. The potential support for queries on syntactic structures (e.g., in dependency treebanks) is a fairly recent development. A proposal for such a new type of query language might lead to a new *SyntacticFCS* extensions. Another aspect under investigation is an improved description of results through structured metadata. The FCS specification is currently very limited on how much additional metadata can be provided and used by software clients and end users. Data providers with resources that, for example, aggregate texts from various sources such as newspaper collections, require ways to enrich individual results with more descriptive metadata, including publication and release date, title, or author information.

In any case, for each extension, a decision must be made as to whether the specific need justifies the corresponding work, for example whether a sufficient number of resource providers can or want to support the respective type of information or the respective use case. To ensure the functionality and interoperability of the entire FCS infrastructure, issues such as backward compatibility and mechanisms for explicitly opting out of individual functionality play an important role here.

Acknowledgments

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e. V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460033370. The authors would like to thank for the funding and support. Furthermore, the authors would like to thank all members of the Text+ data domain *Lexical resources* and the *CLARIN FCS taskforce* for their continuous work.

References

Ayoola, T., Fisher, J., & Pierleoni, A. (2022, July). Improving entity disambiguation by reasoning over a knowledge base. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2899–2912). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.210

Balog, K. (2018). *Entity-Oriented Search* (Vol. 39). Springer Cham. https://doi.org/10.1007/978-3-319-93935-3

BBAW. (2025). Thesaurus Linguae Aegyptiae [Accessed: 2025-04-09]. https://thesaurus-linguae-aegyptiae.de

Eckart, T., Herold, A., Körner, E., & Wiegand, F. (2023). A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, 280–292. https://elex.link/elex2023/wp-content/uploads/elex2023_proceedings.pdf

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books. https://mitpress.mit.edu/9780262561167/

German National Library (DNB). (2024). The Integrated Authority File (GND) [Accessed: 2024-06-24]. https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with wikipedia [Artificial Intelligence, Wikipedia and Semi-Structured Resources]. *Artificial Intelligence*, *194*, 130–150. https://doi.org/https://doi.org/10.1016/j.artint.2012.04.005

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. https://aclanthology.org/W97-0802

Hanke, T. (2021). Hamnosys. https://doi.org/10.25592/uhhfdm.9725

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. (2009). ISOcat: Remodelling metadata for language resources. *IJMSO*, *4*, 261–276. https://doi.org/10.1504/IJMSO.2009.029230

Körner, E., Eckart, T., Kretschmer, U., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T., Rau, F., Klee, A., Werning, D., Blöse, D., & Zinn, C. (2024). *Federated Content Search for Lexical Resources (LexFCS): Specification*. https://doi.org/10.5281/zenodo.7849753

Lundell, J. A. (1928). The swedish dialect alphabet. *Studia Neophilologica*, *1*(1), 1–17. https://doi.org/10.1080/00393272808586721

OASIS. (2013). *searchRetrieve v1.0*. Organization for the Advancement of Structured Information Standards. http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html

Qi, L., Yongyi, H., Defu, L., Zhi, Z., Tong, X., Che, L., & Enhong, C. (2024). Unimel: A unified framework for multimodal entity linking with large language models. https://arxiv.org/abs/2407.16160

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., & Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*, *5*. https://doi.org/10.17175/2020_006

Schonefeld, O., Eckart, T., Kisler, T., Draxler, C., Zimmer, K., Ďurčo, M., Panchenko, Y., Hedeland, H., Blessing, A., & Shkaravska, O. (2014). *CLARIN Federated Content Search (CLARIN-FCS) – Core Specification*. https://www.clarin.eu/content/federated-content-search-core-specification

Schwarz, P., & Barth, F. (2024, March). Classification and linking of named entites. https://doi.org/10.5281/zenodo.10893761 Workshop contribution in Pollin, Christopher, et al. ”Workshop generative KI, LLMs und GPT bei digitalen Editionen”.

Stehouwer, H., Durco, M., Auer, E., & Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, 3255–3259. http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf van Uytvanck, D., Olsson, L.-J., Schonefeld, O., Eckart, T., Körner, E., Kisler, T., Fischer, P. M., & Illig, E. M. (2017). *CLARIN Federated Content Search (CLARIN-FCS) – Core 2.0*. https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

On the Successful Migration of Languages Resources from one Repository to Another

Claus Zinn and Thorsten Trippel
Department of Linguistics
University of Tübingen
Keplerstr. 2, 72074 Tübingen, Germany
claus.zinn@uni-tuebingen.de
thorsten.trippel@uni-tuebingen.de

Abstract

More than five years ago, we crafted a detailed scenario for migrating our research data from our locally-maintained, departmental repository to an external, institutional repository for which we had only little control over. Now, with the rising cost of updating and maintaining our repository software to the latest version, personnel fluctuation, and the opportunity to use data services of a newly founded Digital Humanities Center, we decided to put into practise the scenario step by step. This paper describes the actual challenges we encountered in the migration process, the deviations from the original scenario and the compromises we needed to make, and finally, how we succeeded to get all data transferred in a safe and information-preserving manner.

1 Introduction

The maintenance of a research data repository comes with substantial costs. While a large part of the effort is devoted to data curation, metadata annotation and ingestion as well as to the communication with data depositors and consumers, there is a significant workload involved in keeping the repository software up-to-date. Security updates are a major concern; at short notice, they must be deployed in a running repository system to make it less vulnerable to external threats. From time to time, old versions of repository systems are deprecated and stop benefiting from security patches. In this case, one has to perform a major upgrade to a newer version of the software. Costs related to software maintenance are rising, and some institutions may consider migrating their research data to an external, infrastructural organisation that is experienced with research data management (RDM), already hosts research data from a variety of other disciplines, and has trained staff. Scaling up certainly helps to keep expenditures in check. Ultimately, the turnaround of key personnel triggered our migration process. The process was aided by a migration workflow that was started more than five years ago (Trippel & Zinn, 2018, 2021). We have now executed the workflow, and report on the actual challenges and difficulties we encountered, some of which were anticipated, others newly arose unexpectedly from new, unforeseen technical requirements.

2 Background

In the absence of proper research data management services at a university-wide level, in 2010, our department kickstarted its own repository system. The "Tübingen Archive for LAnguage Resources" (TALAR) was targeted at researchers of two Collaborative Research Centres[1] (CRC-441, CRC-833) to provide them with a centralised storage solution for all data they created. Also, the data created by CRC-external activities of our institution got a new, central archiving home. We started with a system based upon Fedora Commons[2] (version 3), which we extended with a number of essentials such as a shell-based environment to support data ingestion and rights management, and an OAI-PMH port[3] to make

[1] German:"Sonderforschungsbereich", abbreviated as "SFB"

[2] https://fedora.lyrasis.org

[3] See https://www.openarchives.org/pmh/

Claus Zinn and Thorsten Trippel 2024. On the Successful Migration of Research Data. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 60–70. DOI: https://doi.org/10.3384/ecp216

available metadata to external parties. Due to security reasons, we later updated the system to version 4 of Fedora. Security patches available for this version where applied whenever possible.

Having control over your own repository comes with a significant amount of responsibility, but it also opens up a design space around many aspects of research data management, *e.g.*, how to best describe research data with metadata; and what research data should be accepted for ingestion (only internal data stemming from your own institution, or data from other institutions, or only data passing some quality threshold)? In the past decade, we have fitted the Fedora repository system with a number of bells and whistles, namely, a web-based, graphical user interface to serve as a front-end of the repository for both users and archivists (Dima, Henrich, et al., 2012), the Bagman software for researchers to help them describing and transferring their data to the archive (Zinn, 2022), and the ProFormA editor to help them annotating their research data with metadata (Dima, Hoppermann, Hinrichs, Trippel, & Zinn, 2012). We have also defined, and redefined, a good number of CMDI profiles to have adequate means to describe different types of resources with rich metadata, and we have also implemented crosswalks between CMDI to Dublin Core and MARC-21 (Zinn, Trippel, Kaminski, & Dima, 2016).

The repository software used to run on a designated, virtual, Unix-based machine at the university's computing centre. It benefited from regular back-ups, and all its content was regularly mirrored onto a system running in a different physical location.

Our repository has been awarded with the *Data Seal of Approval*[4] in 2013 and 2015 and with the Core Trust Seal[5] afterwards (last renewed in 2023). From its initial days, our repository has also been a certified CLARIN-B centre.[6] The TALAR repository took part in the CLARIN harvesting infrastructure.[7] The CLARIN OAI-PMH harvest manager contacted TALAR at regular intervals to download its 670+ dataset descriptions. The harvesting result showed up in the CLARIN Virtual Language Observatory (VLO).[8] VLO users could browse all metadata in the VLO and could click on the handles that pointed to the dataset's landing page or to individual resources of the dataset.[9]

More than five years ago, within the NaLiDa project[10], a detailed migration workflow was crafted to migrate all research data to another repository, maintained by the university library of the University of Tübingen (Trippel & Zinn, 2018, 2021). Due to changes in personnel, and the arising opportunity to store research data in the newly-founded Digital Humanities Center, we are now putting the migration workflow into practice.

By and large, the main issues of the migration workflow outlined at the time are still the main issues to be tackled: user authentication and authorization, metadata harmonisation, and persistent identifier management. At the time, the migration would have benefited from a common technological base as both the source and the target repository system were based on Fedora. In the meanwhile, however, the university library has expanded its eScience services into a newly founded *Digital Humanities Center (DHC)*.[11] In due course, the DHC staff also deployed new repository software, which was now based on InvenioRDM.[12] This has complicated the migration process along all dimensions.

Our migration process was driven by the technical state and content of TALAR (henceforth, *source repository*), and the technical and organisational requirements of the target repositories.[13] We now give a more detailed description of the source repository.

[4] http://www.datasealofapproval.org

[5] https://www.coretrustseal.org

[6] https://www.clarin.eu/content/certified-b-centres

[7] https://centres.clarin.eu/oai_pmh

[8] At https://vlo.clarin.eu, select collection "Tübingen Archive of Language Resources (TALAR)".

[9] We also used to provide an HTML-based representation of all metadata with a sitemap through our institutional web-server, supporting researchers to discover all data more easily.

[10] http://www.sfs.uni-tuebingen.de/nalida/en/

[11] See https://uni-tuebingen.de/forschung/forschungsinfrastruktur/digital-humanities-center/.

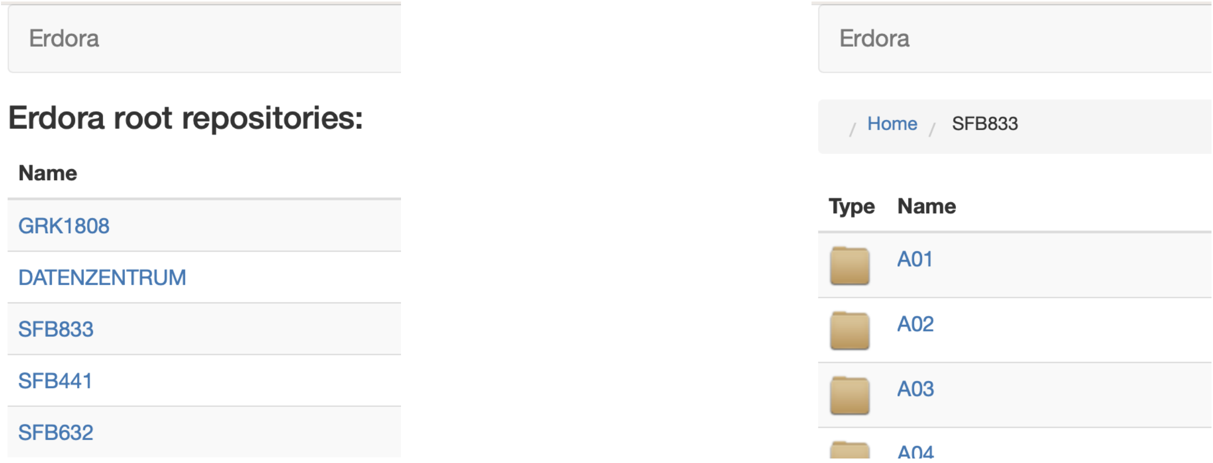
[12] https://inveniosoftware.org/products/rdm/

[13] As will be explained further below, we decided to use the DHC repository for genuine research data and to use the Zenodo repository for all other data.

3 The TALAR Repository

The TALAR repository was operational from 2010. At the time of migration, it contained 673 datasets, totalling hundreds of gigabytes of data. Each dataset came with CMDI-based metadata, for which we have developed a number of CMDI profiles to accurately describe the various types of research data we host: text corpora, speech corpora, lexical resources, tools, web services, experimental data, and teaching material.

Each dataset was addressed by a persistent identifier using the handle system[14]. Part identifiers were used to address individual files inside a dataset. Handles without part identifiers pointed to the HTML-based *landing page* of a dataset, which the TALAR website automatically rendered by using a CMDI-2-HTML based transformation of the dataset's CMDI file. [15].



Erdora

Erdora root repositories:

Name
GRK1808
DATENZENTRUM
SFB833
SFB441
SFB632

(a) Root Structure

Erdora

/ Home / SFB833

Type	Name
📁	A01
📁	A02
📁	A03
📁	A04

(b) WP Structure

Figure 1: The ERDORA GUI

Fig. 1a shows the top-level organisation of all research data in TALAR. In fact, most of our research data originated from two collaborative research centres, SFB441 and SFB833. But the repository also hosted resources that we have developed in our own department such as the lexical-semantic wordnet GermaNet (Hamp & Feldweg, 1997), and treebanks such as TüBa-D/Z[16], and other annotated text corpora such as the Index Thomisticus[17]. Those resources were hosted in the Fedora node "DATENZENTRUM". Resources from a *Graduiertenkolleg* and from an institution-external CRC were stored in GRK1808 and SFB632, respectively.

In TALAR, all research data were hierarchically structured. Fig. 1b shows a fragment of the SFB833 tree, mirroring the working packages of this collaborative research centre. Usually, a work package node contained multiple datasets, and often, each dataset itself exhibited a complex directory structure. Rarely was such data compressed in archive files, say in ZIP format. In fact, the CRC-833 research data had quite a few datasets that consisted of hundreds of individual files, each of which was addressable by a handle-based part identifier. Moreover, the underlying Fedora Commons software made is possible to assign access permissions at individual directory and file levels. In TALAR, those *Access Control Lists* were used to give authenticated individual users (usually, members of the two CRCs) read access to (non-public) datasets.

Note that the CMDI-based description of a dataset listed all its resources and defined their handle-based address space, see Fig. 2a. Handles with part identifiers allowed users to directly download in-

[14] https://handle.net

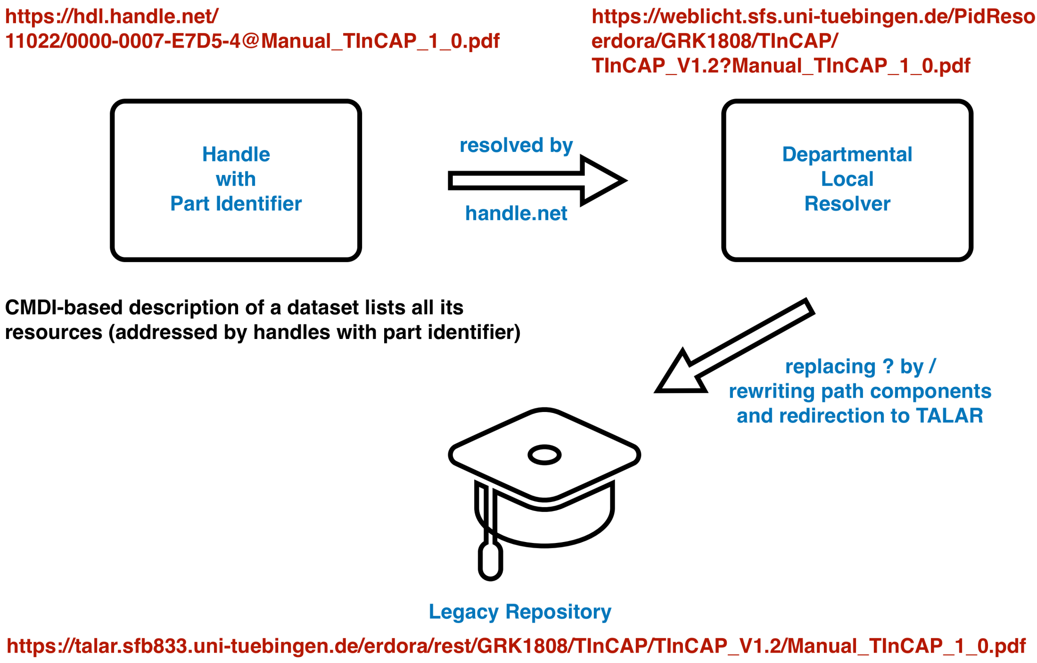
[15] For this purpose, a processing instruction in the XML file invoked an XSL transformation.

[16] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/tueba-dz/

[17] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/index-thomisticus-baumbank/

```
<cmd:ResourceProxyList>
  <cmd:ResourceProxy id="TInCAPexportv12xml">
    <cmd:ResourceType mimetype="text/html">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4@TInCAP_export_v1.2.xml</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="ManualTInCAP10pdf">
    <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4@Manual_TInCAP_1_0.pdf</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="landingpage-11022-0000-0007-E7D5-4">
    <cmd:ResourceType mimetype="text/html">LandingPage</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4</cmd:ResourceRef>
  </cmd:ResourceProxy>
```

(a) Handles with part identifiers in the CMDI file


https://hdl.handle.net/
11022/0000-0007-E7D5-4@Manual_TInCAP_1_0.pdf
https://weblicht.sfs.uni-tuebingen.de/PidResolver/
erdora/GRK1808/TInCAP/
TInCAP_V1.2?Manual_TInCAP_1_0.pdf
Handle
with
Part Identifier
resolved by
handle.net
Departmental
Local
Resolver
CMDI-based description of a dataset lists all its
resources (addressed by handles with part identifier)
replacing ? by /
rewriting path components
and redirection to TALAR
Legacy Repository
https://talar.sfb833.uni-tuebingen.de/erdora/rest/GRK1808/TInCAP/TInCAP_V1.2/Manual_TInCAP_1_0.pdf

(b) Handles and PID Resolving for TALAR-based Resources

Figure 2: Handles and how they were resolved

dividual files of a research dataset. Fig. 2b depicts the process that took place when users clicked on such persistent identifiers. Note that the rewriting behaviour of part identifiers has been configured per handle prefix, where `@partIdentifier` was rewritten to `?partIdentifier`. Also note that the main part of the handle did not point directly to the source repository but to a locally maintained local resolver, which rewrote the URL fragment weblicht.sfs.uni-tuebingen.de/PidResolver/erdora/ to talar.sfb833.uni-tuebingen.de/erdora/rest/ and replaced the question-mark announcing the part identifier with a forward slash, see the URL pointing into the source repository at the bottom of the figure.

Data Curation In normal operation, once a dataset was ingested into a repository, there was little, if any, work, associated with it. Occasionally, researchers contacted us to update the metadata with new contact information, and when a new version of a dataset was released, a new dataset and its updated metadata was ingested into the repository (equipped with a new handle and a number of part identifiers). Licences were rarely changed.

In the past decade, several archive manager took care of research data management, and consequently, their are subtle differences in archiving standards.[18] During the times, we have improved the CMDI

[18] We found some of the research data poorly described; we occasionally encountered CMDI-based metadata that, for instance, used obscure titles, failed to specify resource creators, or had ambiguous licence statements.

profiles in use, clearly separating administrative metadata (such as project-related information) from resource-type specific information (such as for the description of lexical resources or text corpora). In the last years, Bagman, a web-based software that supports researchers and archive managers alike to ingest research data into the repository, was developed and deployed. With Bagman, users pack their data using the BagIt format (Kunze, Littman, Madden, Scancella, & Adams, 2018) and fill out forms to describe it with metadata. All this information is then used to automatically instantiate the appropriate CMDI profile. The automatic generation of CMDI files improved the quality of the metadata considerably and proved superior to the manual creation of such data where XML editors such as ProForma or simple text editors were used.

The migration of the entire repository created a situation were we stepped back and took a bird's-eye view, seeing all the datasets accumulated so-far for the first time in their entirety. This perspective revealed a number of organisational issues that the migration sought to address:

- a significant part of the the CRC-833 collection have datasets that consist of only a single text file (in all cases some scientific text in PDF format). Those 265 PDF articles played the role of a reference collection to support scientific research across the many members of the CRC-833. While some of those papers have been published in journals, others were manuscripts submitted for review or were in press. For these reasons, all datasets were assigned a non-public licence.

- we found thirteen datasets that should have received their own root node. Each dataset contains teaching material (presentation slides, exercises, data files) targeted at graduate and PhD students. These datasets were assigned an open CC-BY licence.

- the repository also contained 54 service descriptions for the WebLicht workflow engine (Hinrichs, Zastrow, & Hinrichs, 2010). WebLicht contacts a harvester[19] whose task is to source available service descriptions from our repository and others parties via their respective OAI-PMH endpoints.[20] It is paramount that those CMDI-based WebLicht service descriptions in TALAR must continue to be available. Note, however, that these services were ingested into TALAR not for archival but for pragmatic reasons, namely, for the sole purpose to be discoverable via an OAI-PMH port.

- the repository also contained 28 publications from the TLT-13 conference "Treebanks & Linguistic Theories", which took place in Tübingen in 2013. Their ingestion into TALAR made them publicly accessible via a persistent identifier. As no research data was attached to any of the papers, a simple preprint server would have been a better match than TALAR.

- there were two sparsely populated root nodes; the subdirectory SFB632 (CRC-632) had only two datasets, and the directory GRK 18080 ("Graduiertenkolleg 1808") had only 5 datasets.

The bird's-eye view revealed that our datasets consisted of four different types: genuine research data, scientific publications, teaching material, and software configuration data. We decided to only move genuine research data to FDAT, the repository of the Digital Humanities Center.[21] All other data must be moved to different homes.

We started the migration process with the following goals in mind: (i) preserve the hierarchical structure of research data; (ii) migrate all research data (independent of its age and quality); (iii) ensure that there is no information loss in terms of metadata; and (iv) strive for minimal service disruptions.

4 Migration

Originally, we thought that the repository of the Digital Humanities Center of the University of Tübingen (FDAT) should receive all of our TALAR data. However, we decided that our CRC-833 collection of

[19] http://weblicht.sfs.uni-tuebingen.de/apps/harvester/resources/services

[20] At last count, 572 services are harvested in total, from many dozen repositories. The URL request for our repository (TALAR) is https://talar.sfb833.uni-tuebingen.de/erdora/rest/oai?verb=ListRecords&metadataPrefix=cmdi&set=WebLichtWebServices.

[21] FDAT is an acronym built from the German word 'Forschungsdaten'.

research articles, despite their non-public licence, as well as the collection of teaching material and the collection of TLT-13 papers should be migrated to elsewhere and that Zenodo[22] was found a better fit. For the genuine research data, we created designated *communities* for data from the two main CRCs in FDAT, and for our own research data (and the one from SFB632 and GRK1808) we defined the TALAR community there. Since neither FDAT nor Zenodo offers an OAI-PMH port, none of the CMDI files carrying WebLicht services descriptions could be migrated. Instead, we extended the source code of WebLicht, which is software we developed in house. The workflow engine has been enabled to also "harvest" service descriptions from subdirectories attached to WebLicht's source code distribution; here, a new subdirectory was created to host the 54 service descriptions.

Table 1 depicts the new organisation.

Type of resource	**Target**	URL of the Community	Number
PDF Articles (CRC-833)	Zenodo	https://zenodo.org/communities/sfb-833-literature	265
Teaching Material	Zenodo	https://zenodo.org/communities/talar-teaching-material	13
TLT	Zenodo	https://zenodo.org/communities/tlt13	28
Research data (CRC-441)	FDAT	https://fdat.uni-tuebingen.de/communities/crc441	28
Research data (CRC-833)	FDAT	https://fdat.uni-tuebingen.de/communities/crc833	227
All other research data	FDAT	https://fdat.uni-tuebingen.de/communities/talar	13
WebLicht Service data	–	`transferred to WebLicht Github Repository`	54

Table 1: Overview of Target Repositories

4.1 Migration to Zenodo

Like FDAT, Zenodo is a repository system that is based upon InvenioRDM, and hence, it also allows the organisation of research data into communities. For our purposes, three communities were created, one to hold the literature from the CRC-833, a second one to take on TALAR's teaching material, and a third one to hold the TLT papers.

The ingestion of literature data was rather straightforward. With each PDF file being complemented with a CMDI-based metadata description, we wrote an XSLT stylesheet to extract the relevant metadata into the required DataCite[23] fields, namely, author, title, publication date, and description. It showed that the CMDI files did not have more information that needed to be preserved, and therefore, no information loss incurred. Consequently, we did not ingest any CMDI files to the Zenodo CRC-833 community. Also, due to copyright issues, it was required that all research data in the Zenodo community "sfb-833-literature" was restricted. Interested parties can contact the community curator for which we have created a new special-purpose email account.[24]

The teaching material had a complex nature. They usually consisted of many files, some of which were hierarchically structured, and used a variety of different data formats. We found their CMDI-based description rather shallow, not making use of the potential that the CMDI profile "CourseProfile" offered. As a result, we also omitted the ingestion of CMDI profiles to Zenodo. The teaching material of each dataset was archived in ZIP format to preserve their hierarchical structure, and subsequently ingested into the Zenodo community "talar-teaching-material" with a CC-BY licence.

The TLT data consisted of 28 PDF files with corresponding CMDI-based annotations that carried little information other than DataCite fields. Without loss of information, only the TLT papers were ingested into the TLT community, but not their CMDI metadata.

In sum, 306 datasets (research articles, teaching material and TLT-13 papers) left the realm of Tübingen University and found their new home in the Zenodo repository. All SFB-833 literature data was automatically ingested into Zenodo using a Python-based script that makes use of the Zenodo developer

[22] https://zenodo.org

[23] https://datacite.org

[24] The email account data-steward@semsprach.uni-tuebingen.de also answers requests from the other Zenodo community and the newly created FDAT communities.

API.[25]. The teaching material as well as the TLT-13 data was manually ingested into Zenodo.

4.2 Migration to FDAT, the Institutional Research Data Repository

There were less than 300 datasets still to be taken care of. In terms of content and size, they constituted the "real" research data. To mirror the high-level structure of the TALAR source repository, we have created three communities on the new institutional FDAT repository, see Tab. 1: (i) the CRC-441 community to host all data stemming from the Collaborative Research Centre "Linguistic Data Structures", (ii) the CRC-833 community to host all data originating from the Collaborative Research Centre "Emergence of Meaning", and (iii) the "Tübingen Archive for LAnguage Resources" (TALAR) community.

The target repository defined a number of hard constraints that we needed to deal with:

- CMDI-based metadata is not an accepted metadata standard for the description of research data; all research data must be described using DataCite;
- the size of each dataset is limited to 100 GB and cannot contain more than 100 individual files;
- FDAT only supports lists of resources, not hierarchically structured ones;
- access to a dataset is either restricted for all users or available to all. No individual rights can be associated with the files of a dataset; and
- all datasets must be addressed with DOI-based persistent identifiers; existing handle-based persistent identifiers cannot be reused.

Apart from these constraints, FDAT offered two benefits: it gives support for the versioning of datasets as such dependencies between datasets can be spelled out explicitly; and it supports the creation of communities so that research data that stemmed from different players could be easily grouped together.

While there was little data curation necessary for transferring literature data or course material from the source repository to the three Zenodo communities, the situation was different for the remaining, genuine research data.

Moreover, many of our research datasets have a deep directory structure, which is not supported by FDAT. In these cases, the hierarchy was "flattened", usually by replacing them with ZIP archives so that their unarchiving reestablishes the hierarchy. Moreover, some datasets had information duplicated. Sometimes, the files of a dataset were complemented with a ZIP archive that also held all files. Such redundancy was removed, consistently in preference for the ZIP archive.

Also, some datasets failed to attribute the agency that funded the project producing the research data. In these cases, the funder for the two CRCs was manually added to the DataCite metadata. Last, while some CMDI-based metadata used authority file information from GND (https://gnd.network) and VIAF (https://viaf.org) to uniquely identify persons associated with the research data, no such information was given for the organisations the persons work for. In DataCite, we have complemented information about organisations with their ROR identifier.[26]

Given the complexity of the research data, the migration to FDAT required us to review each dataset individually. During the review, we observed other, mostly minor, oversights or flaws in the metadata, which we corrected in due course. We also contacted some of the researchers that produced the research data to review our migration work. This made us realize that an acceptable translation from CMDI to DataCite is far from trivial, and must take into account a few subtle but important aspects. In the CMDI metadata, for instance, we find the CMDI component `Project`[27] , which contains information about the project in which a resource was created. In the first iteration, we falsely associated the person mentioned in the Project component as the *creator* of the resource (with role "Project leader"), where in fact, the

[25] https://developers.zenodo.org

[26] The German Research Foundation has the ROR identifier https://ror.org/018mejw64.

[27] See the CLARIN component registry at https://catalog.clarin.eu/ds/ComponentRegistry; select group name "NaLiDa" and search for "Project".

DataCite *contributor* with that role should have been used. As a result, the dataset's citation as generated by the FDAT GUI was misleading as it omitted the actual creators of the resource.

To avoid any loss of information given in the CMDI metadata description of a dataset, we made the CMDI file an integral part of the dataset itself. While the DataCite metadata can be altered after the publication of a dataset, this is not the case for the research data itself. Consequently, it was crucial to ensure that all CMDI-based metadata was in its final, publishable state. As a result, each CMDI file was diligently reviewed by the communities' curator before the entire dataset it describes was published.

4.2.1 User authentication and Authorization

While the source repository had an expressive rights management system in place, the target repository had no capabilities for user authentication and authorization to cater for such personalised access. Data still under publication embargo will continue to be inaccessible to *all* users in the target repository; and this includes the data creators. For most datasets, the embargo date has been set to September 30, 2026. Interested parties must contact the data steward of the FDAT communities (CRC 833, CRC 441, & TALAR), or the contact persons specified in the DataCite metadata. Once the embargo data passes, all research data in FDAT will become publicly available under a CC-BY licence.

4.2.2 Persistent Identification

The FDAT repository requires the use of DOI-based persistent identifiers.[28] In principle, the FDAT repository allows the addressing of individual files of a dataset. For this purpose, however, no DOI can be used, only the resolved URL. [29] However, the FDAT administrator cannot guarantee that the resolved URL, or its current path, will continue to be serviced in the future. If we were to continue supporting part identifiers, then we must continue to support our local resolver, see Fig 2b and maintain its mapping table to do the rewriting, and change the rewriting whenever the target URL of FDAT changes.

To minimize our commitments, and to keep things simple, we stopped our support for part identifiers; access to all research data is via their new FDAT-based landing page. All handles that we registered at `handle.net` now directly point to their respective DOI handles of FDAT, hence bypassing any local URL resolver, which we have therefore retired. Handles, with or without part identifiers used in CMDI files before (see Fig. 2a) have been replaced by their respective DOIs (see Fig. 3a). Users trying to invoke legacy handles with part identifiers, say because they have bookmarked them, will find themselves redirected to the landing page of the dataset in FDAT.

Fig. 3b illustrates the resolving of a handle-based persistent identifier with a part identifier, which gets redirected to doi.org. Note that the part identifier initiated with the "at"-sign "survives" the first redirection; the part identifier now appears as URL argument of the DOI-based address. However, when doi.org resolves the DOI-based URL, it ignores all its arguments so that the final URL is free from any path information.

4.2.3 Metadata Provision

The migration required us to convert CMDI-based metadata to DataCite. To minimize information loss, some resource-specific metadata (such as information about, say an experimental design, or the number of entries in a lexical resource) was written into one of Datacite's description fields. Moreover, the original CMDI-based description became part of the research data stream so that users can consult the metadata for information that cannot be (or has not been) mapped to DataCite.

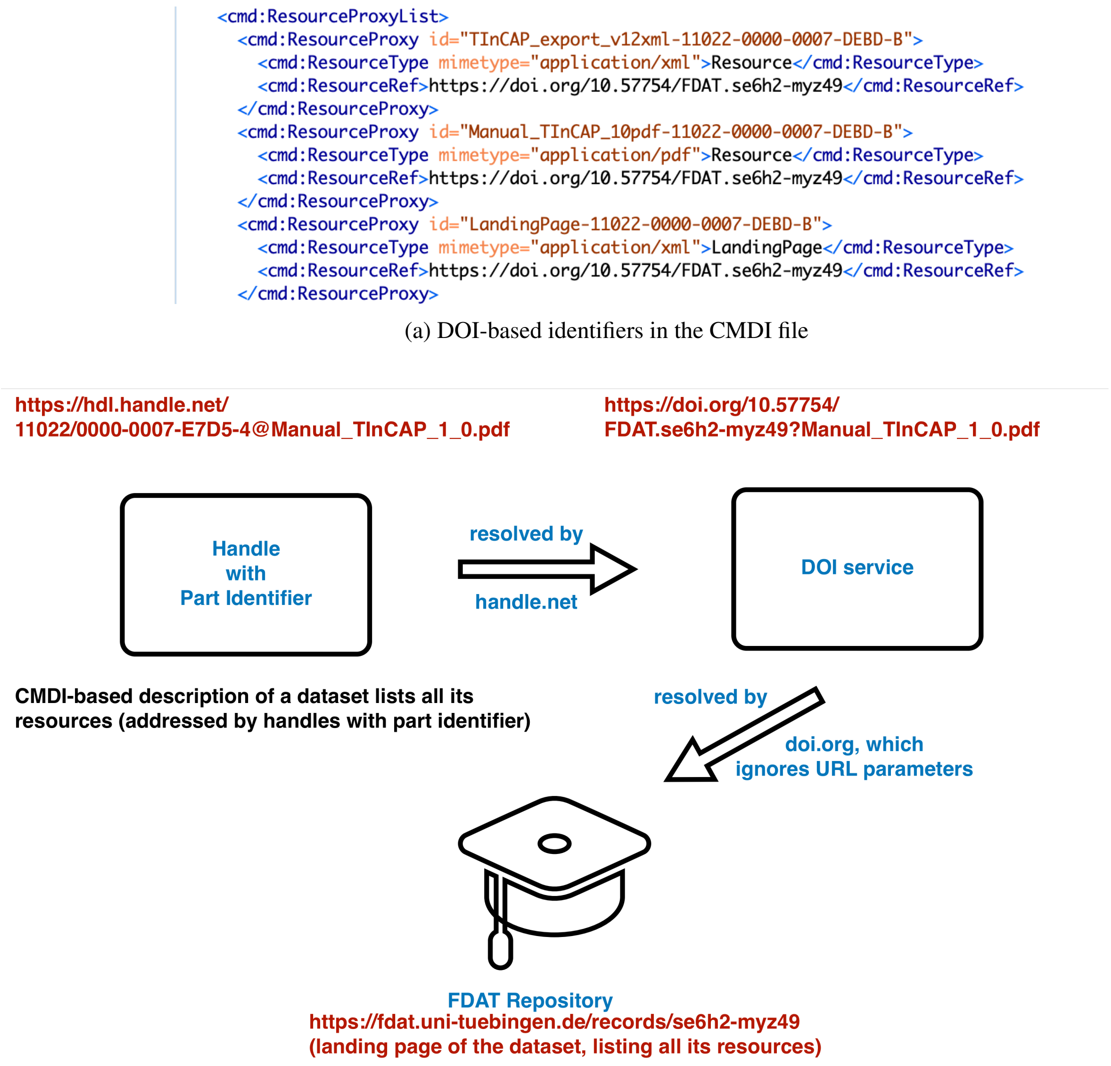
The source repository offered OAI-PMH harvesting for metadata in DC, MARC 21, and CMDI. Zenodo and FDAT, however, only support metadata harvesting for DataCite. With the source repository being shut down, we had to search for an alternative solution to OAI-PMH servicing that provides CMDI-based metadata to the Virtual Language Observatory (VLO) and other interesting parties.

For this purpose, we have implemented our own OAI-PMH endpoint in the Prolog programming language.[30] The endpoint's content is sourced from a directory that contains the CMDI files of all data

[28] The DOI https://doi.org/10.57754/FDAT.c0cvj-4vk83, for instance, resolves to https://fdat.uni-tuebingen.de/records/c0cvj-4vk83.

[29] For example, https://fdat.uni-tuebingen.de/records/c0cvj-4vk83/files/de-nn-com-sem_8005_compounds.txt

[30] http://textplus.sfs.uni-tuebingen.de:8088/api/oai?verb=Identify



```
<cmd:ResourceProxyList>
  <cmd:ResourceProxy id="TInCAP_export_v12xml-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/xml">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="Manual_TInCAP_10pdf-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="LandingPage-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/xml">LandingPage</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
```

(a) DOI-based identifiers in the CMDI file

https://hdl.handle.net/
11022/0000-0007-E7D5-4@Manual_TInCAP_1_0.pdf
https://doi.org/10.57754/
FDAT.se6h2-myz49?Manual_TInCAP_1_0.pdf
Handle
with
Part Identifier
resolved by
handle.net
DOI service
CMDI-based description of a dataset lists all its
resources (addressed by handles with part identifier)
resolved by
doi.org, which
ignores URL parameters
FDAT Repository
https://fdat.uni-tuebingen.de/records/se6h2-myz49
(landing page of the dataset, listing all its resources)

(b) Handles and PID Resolving for TALAR-based Resources

Figure 3: Handles and PID Resolving for TALAR Resources (after migration)

migrated to FDAT. The name of each file follows a naming scheme that encodes two pieces of metadata, namely, the FDAT community, and the original publication date of the research data as well as a running number (*e.g.*, `sfb833_2020-10-26_0004.xml`). Each CMDI-based file is mirrored by a metadata description in Dublin Core, which has been automatically generated using an XSL-based crosswalk (Zinn et al., 2016). [31] – At September 30, 2026 latest, our simple OAI-PMH service will be transferred to another (yet to be identified) organization to ensure that all research data remains visible to the VLO for the foreseeable future.

Tombstoned data In contrast to our initial aim to migrate all research data independent of its age and quality, we encountered ten datasets where we made an exception. In the source repository, we found occasional datasets that were either incomplete, clearly did not merit long-time archiving, or constituted merely test data. Such data were tombstoned. For this purpose, we created a ’Metadata only’ entry in FDAT, see https://doi.org/10.57754/FDAT.kvd3z-7w002. It lists all the tombstoned datasets via their respective handles, which in turn now all point to the aforementioned DOI.

[31] Recall that CMDI descriptions for all data migrated to Zenodo have been removed; the TALAR collection in the VLO will hence shrink to around 300 datasets of genuine research data.

Transitional period We have now migrated all research data to the new repositories. All handles have been redirected to DOI-based identifiers pointing to FDAT or Zenodo. The new OAI-PMH endpoint has replaced our legacy endpoint so that CMDI-based descriptions in the VLO now only have DOI-based identifiers (saving one level of redirection). Users with bookmarked handle-based identifiers, with or without part identifiers, will always be directed to a research dataset's landing page. Those interested in a particular part of the dataset may now be required to download a ZIP-based archive of the entire dataset to then extract the specific item of their interest.

For new research data, our institution will continue to provide help to researchers who would like to archive their research data in a trusted, sustainable environment. For the data we accept, end-users can expect to receive the same quality of service as before; annotation of research data with CMDI profiles will continue. The CMDI metadata will be part of the dataset, and distributed via our OAI-PMH service. Upon the closure of the institute, all users will be directed to another CLARIN-B centre or asked to contact the FDAT archive manager. At this point, FDAT users will not be required to provide CMDI-based metadata annotations.

5 Conclusion

The migration of research data from one repository system to another is no easy matter and is bound to create issues that cannot always be resolved without making compromises. The actual effort for migrating all data involved numerous internal discussions, coordination with the FDAT manager, the authoring of XSL-based stylesheets for ingest mechanisation, the adaptation of the WebLicht software to fetch their CMDI-based service descriptions in a modified manner, the editing of CMDI files, and the reconfiguration of handles to point to new target DOIs. Our discussion shows that the migration of research data needs careful planning and execution, and that any migration efforts must be started well in advance.

The migration of research data freed us from maintaining a good number of software packages. ProForma, the Erdora shell and its GUI, the OAI-PMH plugin, the local resolver, and the entire Fedora repository has been retired. The only software that is run for a longer period of time is the OAI-PMH service to make available all CMDI files to the CLARIN VLO, ensuring that the visibility of our language resources stays high.

Our department has offered a repository system since 2010; it started at a time where research data management was underdeveloped in the infrastructural institutions of our university. At the time, we had little alternatives to perform research data management other than doing it ourselves. Supported by national and international funding from CLARIN, we developed an entire eco-system around RDM. This time is now coming to an end. The old repository, which served the CLARIN community well, has been dismantled. What remains are our Github repositories that continue to hold our source code for the software we built. In particular, we hope that the Bagman software attracts some interest from other parties as it provided tremendous support to our archiving workflow.

For us, our work in research data management does not stop. We will continue to advise researchers in these matters. We will continue to take on new research data, help getting it properly annotated with CMDI metadata, ingest it into FDAT, and making available its metadata via OAI-PMH to the CLARIN Virtual Language Observatory and now also to the Text+ registry.[32] Relieved from the burden of running the daily business of keeping a repository running and up-to-date, we can focus on other important aspects of research data management.

6 Acknowledgements

In the last decade, the TALAR repository profited from a number of research grants, in particular, the German Research Foundation within the NaLiDa project which ran von 2009 to 2016[33]. We also profited from funding within the national CLARIN-DE project and the pan-European CLARIN project.

[32] See https://registry.text-plus.org.

[33] https://gepris.dfg.de/gepris/projekt/88614379

The migration work described in this paper has been supported by the Text+ project[34], funded by the German National Science Foundation (DFG), project reference 460033370.

We thank the repository manager of FDAT, Dr. Steve Kaminski, for his invaluable (and continuing) support during the migration of our data.

References

Dima, E., Henrich, V., Hinrichs, E., Hinrichs, M., Hoppermann, C., Trippel, T., Zastrow, T., & Zinn, C. (2012). A repository for the sustainable management of research data. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)* (pp. 3586–3592). ELRA.

Dima, E., Hoppermann, C., Hinrichs, E., Trippel, T., & Zinn, C. (2012). A metadata editor to support the description of linguistic resources. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)* (pp. 1061–1066). ELRA.

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop automatic information extraction and building of lexical semantic resources for nlp applications*. Madrid, Spain.

Hinrichs, M., Zastrow, T., & Hinrichs, E. (2010). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In *Proceedings of the seventh conference on international language resources and evaluation (LREC)*, ELRA.

Kunze, J., Littman, J., Madden, E., Scancella, J., & Adams, C. (2018). *The bagit file packaging format (v1.0)*. RFC 8493, DOI 10.17487/RFC8493. See https://www.rfc-editor.org/info/rfc8493.

Trippel, T., & Zinn, C. (2018). Lessons learned: On the challenges of migrating a research data repository from a research institution to a university library. In *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018*, ELRA.

Trippel, T., & Zinn, C. (2021). Lessons Learned: On the Challenges of Migrating a Research Data Repository from a Research Institution to a University Library. *Language Resources and Evaluation*, *55*, 191–207. Springer.

Zinn, C. (2022). Bagman – a tool that supports researchers archiving their data. *Linköping Electronic Conference Proceedings*, *189*, 181–189. Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.

Zinn, C., Trippel, T., Kaminski, S., & Dima, E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In *Proceedings of the tenth international conference on language resources and evaluation (LREC)*, ELRA.

[34] See https://www.text-plus.org/en/home/.

CLARIN.SI, the Slovenian node of CLARIN: ten years on

Tomaž Erjavec
Jožef Stefan Institute
ZRC SAZU
tomaz.erjavec@ijs.si

Nikola Ljubešić
Jožef Stefan Institute
University of Ljubljana
nikola.ljubesic@ijs.si

Katja Meden
Jožef Stefan Institute
Inst. of Contemporary History
katja.meden@ijs.si

Taja Kuzman
Jožef Stefan Institute
taja.kuzman@ijs.si

Cyprian Laskowski
University of Ljubljana
cyp@cjvt.si

Jan Jona Javoršek
Jožef Stefan Institute
jona.javorsek@ijs.si

Simon Krek
University of Ljubljana
Jožef Stefan Institute
simon.krek@ijs.si

Mateja Jemec Tomazin
ZRC SAZU
mateja.jemec-tomazin@zrc-sazu.si

Kaja Dobrovoljc
University of Ljubljana
Jožef Stefan Institute
kaja.dobrovoljc@ijs.si

Špela Arhar Holdt
University of Ljubljana
spela.arharholdt@ff.uni-lj.si

Jakob Lenardič
Inst. of Contemporary History
CLARIN ERIC
jakob.lenardic@inz.si

Darja Fišer
Inst. of Contemporary History
CLARIN ERIC
darja.fiser@inz.si

Abstract

The paper presents the organisation and services offered by the Slovenian research infrastructure for language resources and technologies CLARIN.SI. We introduce the governance, organisational structure, and technical components of the infrastructure, followed by a description of its web applications with a focus on the repository and concordancers. Next, we provide an overview of support activities offered by CLARIN.SI, which includes services of the CLASSLA knowledge centre for processing South Slavic languages, financial support for projects, and facilitation of conferences and workshops. We also present the involvement of CLARIN.SI in national and European projects, with its sister national infrastructure nodes of DARIAH and CESSDA, and in the work of CLARIN ERIC.

1 Introduction

The CLARIN.SI consortium was established in 2014, and CLARIN.SI became a member of CLARIN ERIC in 2015. To date, the only publication in English[1] that comprehensively presented CLARIN.SI was published shortly after its establishment (Erjavec et al., 2014), where we described the initial steps of the research infrastructure (RI) and plans for further work. This paper summarises the state of the infrastructure ten years later. In Section 2, we present the organisational structure and management of the RI. In Section 3, we describe the CLARIN.SI repository of language resources and tools. Section 4 presents online services, with a focus on online corpus concordancers, and Section 5 focuses on support activities (knowledge centres, project support, organization of events and other dissemination). In Section 6, we describe the involvement of CLARIN.SI in national and European projects, while Section 7 provides conclusions and plans for further work.

2 Management of CLARIN.SI

The CLARIN.SI infrastructure is based at the Jožef Stefan Institute (JSI), Slovenia's largest research institute. Most of the computer equipment is located at JSI, and the institute also ensures the security, maintenance, and continuous operation of the infrastructure's online services. Three organisational units of the JSI, namely the Department of Knowledge Technologies, the Artificial Intelligence Laboratory and the Centre for Network Infrastructure, are involved in the management and technical maintenance of the infrastructure.

[1] However, see Erjavec et al. (2022a) for a more recent presentation in Slovenian.

Tomaž Erjavec, Nikola Ljubešić, Katja Meden, Taja Kuzman, Cyprian Laskowski, Jan Jona Javoršek, Simon Krek, Mateja Jemec Tomazin, Kaja Dobrovoljc, Špela Arhar Holdt, Jakob Lenardič and Darja Fišer 2024. CLARIN.SI, the Slovenian node of CLARIN: ten years on. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 71–85. DOI: https://doi.org/10.3384/ecp216

2.1 CLARIN.SI Consortium

CLARIN.SI is organised as a consortium, currently comprising 12 partner institutions. The consortium brings together all the key institutions involved in the development or use of language resources and technologies in Slovenia, including research institutes, universities, companies, and an association:

- Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU), in particular its Fran Ramovš Institute for the Slovenian Language, which collects and analyses linguistic materials and produces fundamental works in Slovenian linguistics, especially dictionaries.
- Jožef Stefan Institute (JSI), from where the RI is coordinated, and its repository and most of its services are maintained. The departments involved in the IR also have a strong track record in the development of language resources and tools.
- Institute of Contemporary History (INZ), which coordinates DARIAH-SI and leads the only long-term research programme in Slovenia focused on digital humanities.
- Science and Research Center Koper (located close to the borders with Italy and Croatia), in particular its Institute for Linguistic Studies, which focuses on languages (and cultures) in contact.
- University of Ljubljana, in particular its Centre for Language Resources and Technologies (Center za jezikovne vire in tehnologije, CJVT), which coordinates research in corpus linguistics and language technologies, while developing and maintaining fundamental digital language resources and tools for contemporary Slovenian. The centre is a part of the Network of research and infrastructural centres at the University of Ljubljana, and its activities are carried out at six different faculties, e.g., the Faculty of Computer and Information Science and the Faculty of Arts facilitating interdisciplinary development of written language and speech technologies, and the Faculty of Social Sciences, the Slovenian node of the sister RI for social sciences CESSDA.
- University of Maribor, in particular its Faculty of Electrical Engineering and Computer Science, where language technology and especially speech technology research are performed.
- University of Nova Gorica, in particular its Center for Cognitive Science of Language, which specialises in formal theoretical and experimental linguistics.
- University of Primorska, in particular its Faculty for Mathematics, Natural Sciences and Information Technologies that is active in the fields of machine translation and knowledge discovery.
- National and University Library, which collects, documents, preserves and archives the written cultural and scientific heritage of the Slovenian nation. The Library has joined the consortium only in 2024 and is, of course, an extremely welcome partner in the view of future cooperation (e.g., in transforming parts of its large digital library dLib into language corpora).
- Slovenian Society for Language Technologies (SDJT, est. 1998) promotes the development of language technologies for Slovenian with a focus on open-source solutions. It was for many years the only organiser of the biennial conference series “Language Technologies and Digital Humanities”, and remains the main one.
- Alpineon specialises in developing state-of-the-art computer vision and products involving speech and translation.
- Amebis develops products in the fields of language technology and electronic publishing. It develops text and speech corpora, plug-in language processing modules for Slovenian, and machine translation and speech synthesis systems.

Decisions on the management of the RI are made or confirmed by the CLARIN.SI Management Board, where each partner has one representative and an unlimited number of deputies. Communication occurs through the Board's mailing list, currently comprising 40 subscribers, and during annual meetings where the RI's activities over the past year are reviewed and plans for the following year are formulated.

The operation of the CLARIN research infrastructure in Slovenia is thus based on the needs and consensus of the key stakeholders in digital linguistics and language technologies, as well as digital humanities and social sciences.

2.2 Technical Infrastructure

CLARIN.SI maintains a bilingual (Slovenian, English) website that presents the RI and its services. The website also provides contact information for assistance or advice for users, and password-protected internal pages accessible to members of the Board of Directors, containing founding documents, minutes of meetings, and relevant CLARIN ERIC meeting minutes. In 2024, the CLARIN.SI website received approximately 30,000 visits from users worldwide. As presented in Figure 1, website visits have been consistently increasing over time.

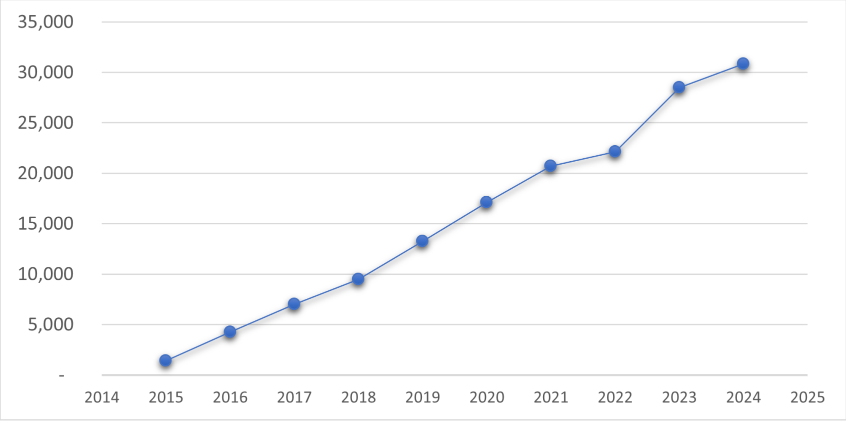


Figure 1: Number of visits to the CLARIN.SI website over time.

Technical documentation is maintained on an internal WordPress installation, where set-up and maintenance procedures for all CLARIN.SI online services are documented. A Redmine installation at the University of Ljubljana is used for managing requests to address identified issues and for planning upgrades.

All changes to critical online services are first tested on the dedicated development servers, where the functioning of the software, documentation and language resources are assessed before deployment on the production servers. The operation of the online services is monitored locally using NAGIOS, while the functioning of the repository is independently verified by the CLARIN ERIC Icinga. In case of errors, service administrators are notified immediately, and can promptly rectify the issue.

3 The CLARIN.SI Repository

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, which was developed within LINDAT/CLARIAH-CZ at the Institute for Formal and Applied Linguistics at Charles University in Prague. In addition to Slovenia and the Czech Republic, the platform is also used by a number of other national CLARIN repositories, which together represent a third of all regular members of CLARIN ERIC.

3.1 Quality Assurance

Besides ADP, the CLARIN.SI repository is the only one in Slovenia accredited with the Core Trust Seal certificate, which certifies it as a trustworthy data repository. In accordance with the CLARIN ERIC strategy, the repository implements FAIR principles. The European agenda for open science and the principles of FAIR CLARIN (Jong et al., 2018) are followed using the following instruments:

- AAI academic authentication, which operates using the SSO ("Single sign-on") system. This separates identity providers (e.g., the Slovenian academic network Arnes, the universities and other academic institutions) and service providers (e.g., the repository), allowing users to log into the repository without creating an account on CLARIN.SI. Instead, they use their EduGain username and password with the chosen identity provider.
- Permanent identifiers of entries via the Handle system, which enables the assignment of a permanent URL address to each repository entry.
- Involvement in international online metadata aggregators, such as OpenAIRE, Re3data, and, since 2022, the European Language Grid. CLARIN (and hence CLARIN.SI) was one of the first RIs to be included in the European Open Science Cloud (EOSC). Within CLARIN(.SI), CMDI (Component MetaData Infrastructure) recommendations are followed for metadata records, with export and metadata harvesting supported in the Dublin Core standard.
- A rich selection of licenses, ranging from open ones such as Creative Commons to more restrictive ones that require prior registration and a digital signature of the resource usage agreement.
- Explicit terms of use, which define the rights and obligations of both repository managers and users.
- Instructions for depositing entries, which describe the resource submission process with special emphasis on the required metadata and its format, ensuring uniform and complete metadata records.
- Instructions for encoding deposited data, which specify acceptable record formats and data marking methods, and also include general instructions for the preparation of high-quality and harmonized data. Unlike most other CLARIN repositories (Lenardič & Fišer, 2022b), which typically only list the acceptable formats, CLARIN.SI also offers broader guidance, which is particularly helpful for humanities researchers who may lack advanced computer and data management skills.
- A FAQ about various aspects of the repository and depositing data.

In addition to its main purpose of describing language resources, the CLARIN.SI repository differs from general self-archiving repositories such as Zenodo by ensuring high quality of the deposited language resources and their metadata, as each entry undergoes a careful review by one of the repository editors before publication to ensure it meets the CLARIN.SI criteria. If the criteria are not met, the editor rejects the entry with an explanation of the errors, and, in prearranged cases, assists in correcting the resource.

3.2 Usage and Impact

The repository, at the time of writing, contains 651 entries[2] produced by more than 1,000 authors from over 100 institutions and totalling about 5TB of data. About half of the entries are for or include Slovenian, and about 220 entries include other South Slavic languages (cf. Section 5.1). The top contributors are Jožef Stefan Institute (248 entries), University of Ljubljana (174) and ZRC SAZU (74).

It should be noted that, as a rule, the repository accepts only entries that include data, as we do not want to function merely as a catalogue of language resources, but as their archive. There are two exceptions to this rule. First, if a corpus is mounted on the concordancers offered by the RI but the data cannot be included in the repository (typically because of copyright limitations that prevent download), nor does it have an associated web page showing its authors and other bibliographic information, then we include only the metadata of such a corpus in the repository, in order to enable its users to correctly cite

[2] This number does not include entries hidden for browsing in the repository. We typically hide entries of previous version of submitted resources, as they are not of interest for browsing and merely extend the list of resources. Counting older version of resources as well, the number is closer to 800 entries.

it and access a description of its properties. The second exception is the ELEXIS catalogue of digital dictionaries, as further discussed in Section 6.2.

In Figure 2 we give the number of entries published on the repository since its inception. The number of new entries per year is mostly stable, with about 50 new resources published each year. The exceptions occur in 2022 and 2023, when the ELEXIS collection, the language resources developed in the scope of the national project "Development of Slovene in a Digital Environment" (cf. Section 6.3) and the EU project MaCoCu (cf. Section 6.2), as well as the multilingual CLASSLA annotation models (cf. Section 4.2) were submitted to the repository.

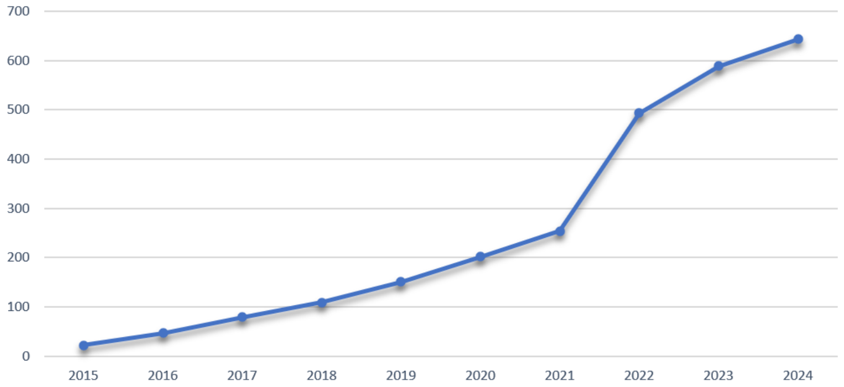


Figure 2: Number of CLARIN.SI repository entries over time.

The number of visits to the CLARIN.SI repository has consistently grown over the years, from about 3,000 in 2015 to 39,000 page views in 2024, although these numbers may include some bots.[3] The most frequently visited resources were the various versions of the ParlaMint corpora, cf. Section 6.5.

As evidenced by the above, CLARIN.SI thus plays a pivotal role in the deposition of open language resources and assists in their creation and description in Slovenia and the region. It has already made a significant contribution to the implementation of the concept of open, verifiable, repeatable and responsible science in the field of linguistic research in Slovenia. It also safeguards numerous language resources created within (Slovenian) research projects from disappearing and provides them with international visibility and influence.

4 Web Services

In addition to the repository, CLARIN.SI maintains several other online services, which are discussed below.

4.1 Concordancers

The most significant services are the concordancers, in particular noSketch Engine and KonText. Both rely on the same back-end program, Manatee (Rychlý, 2007), which enables fast querying of large and richly annotated corpora, the construction of subcorpora, frequency lexica, collocations, etc. To support Slovenian users, CLARIN.SI undertook the localisation of both their user interfaces.

NoSketch Engine is the open-source version of the commercial Sketch Engine concordancer (Kilgarriff et al., 2014), developed by the company Lexical Computing. This is our main concordancer, as it is regularly updated, and has the most advanced functionality and interface. As a matter of principle, we have so far[4] supported anonymous use of the concordancers, so the default installation of noSketch Engine does not require (and therefore support) log-in. This, however, comes with a price, as user-particular

[3] We use Matomo for tracking the use of the repository, and while this platform attempts to exclude bots, it might not always be successful.

[4] Note that this policy might change in the future, as it is becoming increasingly difficult to prevent aggressive harvesting by large companies, which degrades the service.

features are therefore not available (e.g. creation of subcorpora). We therefore recently introduced another instance of noSketch Engine, with self-registration.

The KonText concordancer was developed at the Czech National Corpus Department of Charles University in Prague (Machálek, 2020). It offers a narrower range of functions, but is more in line with the rest of the CLARIN infrastructure: it is used by LINDAT/CLARIAH-CZ, it supports AAI-based log-in, and the currently most recent version should also support FCS.

In addition to these three concordancers, we also still support the old (so-called "Bonito") version of noSketch Engine, though its development was stopped five years ago. However, a number of resources and some services are non-trivially linked directly to this concordancer, such as the glossaries of historical Slovenian linked to the IMP corpus (Erjavec, 2015), the Japanese-Slovenian learner's dictionary of Slovenian (Hmeljak & Erjavec, 2010), linked to several Japanese(-Slovenian) corpora, and the API of the Korpusnik service (cf. Section 4.3).

CLARIN.SI thus maintains four production concordancers, and, in addition, three development instances: old and new noSketch Engine, and KonText. This means that a new corpus has to be mounted on up to seven different virtual machines. To optimise this process, we developed a system of remote management scripts with which, from a remote machine, new (sets of) corpora can be added to any combination of concordancer instances. Additionally, only the development machines are used to index a corpus, which is pertinent for large and heavily annotated corpora, since it is a resource-intensive process that can last over a day. The production machines simply copy all the index files for a corpus to a temporary directory, and then rename it to the production one, minimising disruption of services.

CLARIN.SI concordancers mostly provide access to the same set of corpora, comprising over 200 corpora in 41 languages, including monitor, reference, many types of specialised, and parallel corpora. We here mention only the metaFida corpus (Erjavec, 2023), which combines 34 Slovenian corpora (4.5 billion tokens), making it the largest and most diverse Slovenian corpus available for on-line analysis.

The CLARIN.SI concordances are widely used in university study programmes, linguistic research, research projects, as well as by Slovenian translation companies. They serve a diverse user base, with users originating not only from Slovenia and other South-Slavic-speaking countries, such as Croatia, Serbia, Bulgaria, and North Macedonia, but also from more distant locations, including France, Canada, and Japan.

We have recently implemented tracking of the use of the corpora available on the concordancers. This not only provides us with some key performance indicators, but will also help us to focus on the corpora that are used the most in order to concentrate development where it will have the greatest impact. The analysis is made on the basis of Apache logs, with every effort made to exclude bots.

The analysis of the logs for the new version of noSketch Engine without login, covering the period from June (when tracking was implemented) until December 2024, revealed significant usage of the concordancer, with approximately 400,000 requests in total. Figure 3 presents the daily usage statistics for the recorded period, indicating a consistent demand throughout the year, with an average of 2,000 requests per day.

The analysis showed that users most frequently queried web corpora, particularly the recently developed CLASSLA-web corpora for South Slavic languages (Ljubešić & Kuzman, 2024) and the corpora from the WaCky Web Corpus family (Baroni et al., 2009). In total, web corpora accounted for 65% of all requests made to the top 50 corpora. Other frequently queried corpora included the Slovenian metaFida corpus (Erjavec, 2023); parliamentary corpora, notably those from the ParlaMint collection (cf. Section 6.5); news corpora, such as the Trendi monitor corpus of Slovenian (Kosem, 2022) and the ENGRI corpus of Croatian news portals (Bogunović et al., 2021); the OSS corpus of Slovenian scientific publications (K. Žagar et al., 2023); and the GOS corpus of spoken Slovenian (Zwitter Vitez et al., 2023).

4.2 CLASSLA Annotation Tool

As part of the CLASSLA K-Centre (cf. Section 5), CLARIN.SI also provides the CLASSLA Annotation Tool, an online service for automatic linguistic annotation of raw texts, either by pasting or uploading texts to the web interface.

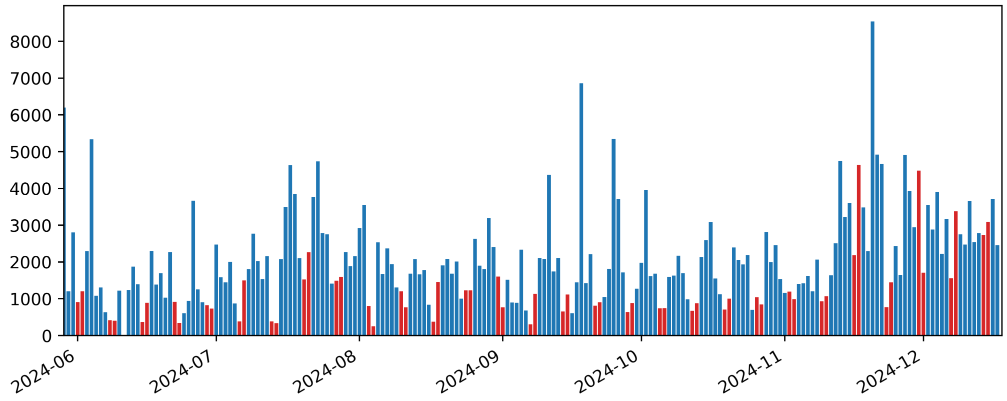


Figure 3: Requests per day through the noSketch Engine in the second half of 2024. Red columns represent requests during weekends.

The web interface, adapted from the CJVT Označevalnik service (Dobrovoljc, 2024b), uses the CLASSLA-Stanza pipeline (Ljubešić et al., 2024d), which is based on the Stanza pipeline (Qi et al., 2020). With the developed models (deposited in the CLARIN.SI repository) the service covers Croatian, Serbian, Slovenian, Macedonian, and Bulgarian, including non-standard (colloquial) varieties of the first three. The tool can annotate the levels of tokenisation, sentence segmentation, morphosyntactic features, lemmas, dependency syntax, named entities, and, for some languages, semantic roles. Additionally, the service enables users to view and download results in various formats, including tables (.xlsx), XML files, CONLL-U files, and graphical images (.svg).

4.3 Korpusnik and SENTA Tools

Partly due to CLARIN.SI-supported projects (cf. Section 5.2) two on-line tools were developed and installed at CJVT, and have been included among the CLARIN.SI services, namely Korpusnik and SENTA tools. Both tools are designed for the general public, with special emphasis on ensuring inclusivity and accessibility.

Korpusnik, a corpus summarizing tool for Slovenian (Kosem et al., 2023) provides a simple and intuitive overview of key information from five Slovenian corpora, selected for their relevance to the general public. This information includes collocations, example sentences, distribution by text type, year of publication, and source. The tool aims to compare and visualise corpus data in a user-friendly way, offering engaging insights into authentic language use.

SENTA (Sentence Simplification and Analysis) is a tool that transforms complex Slovenian sentences into simpler, more comprehensible ones while preserving the original meaning (A. Žagar et al., 2024). The system consists of two components: a neural classifier that identifies sentences requiring simplification and a Slovenian fine-tuned language model based on the T5 architecture to perform the simplification. The tool is accessible through a user-friendly interface, which also provides basic statistics on the texts before and after simplification.

4.4 WebAnno Annotation Platform

The manual corpus annotation platform WebAnno (Yimam et al., 2013), developed within CLARIN-DE, was first used in Slovenia in the context of several national projects, and then an instance was installed at CLARIN.SI, which also manages users, their access levels, as well as the project (annotation) definitions. In the scope of CLARIN.SI we also created scripts to convert TEI-encoded corpora into the WebAnno TSV3 format and to merge manual annotations back into the source TEI corpora (Erjavec et al., 2016).

The more important projects that used CLARIN.SI WebAnno were: “Linguistic analysis of non-standard Slovenian” (Fišer et al., 2020) (manual normalisation of word forms and for assigning lemmas and morphological annotation to Slovenian user-generated content); “Slovenian scientific texts: sources and description” (Erjavec et al., 2021) (marking bilingual terms); “Terminology and knowledge schemes in the interlinguistic space” (Vintar & Martinc, 2022) (marking term definitions in texts); “May ‘68 in Literature and Theory: The Last Season of Modernism in France, Slovenia, and the World” (marking named entities, and foreign and non-standard language words and phrases) (Žejn & Šorli, 2023); the EU project “INTAVIA: In/Tangible European Heritage, Visual Analysis, Curation & Communication”(marking up abbreviations in the Slovenian Biographical Lexicon) (Daza et al., 2022); and a series of projects aimed at manual syntactic parsing of written and spoken Slovene (Dobrovoljc, 2024a; Dobrovoljc et al., 2023).

4.5 Usage of Git

CLARIN.SI also leverages Git platforms for collaborative development of software and language resources. On GitHub, CLARIN.SI hosts over 100 open-source projects under its virtual organisation. For projects requiring local hosting or restricted access due to copyright concerns, CLARIN.SI also operates a GitLab installation. This platform hosts around 20 projects, both public (such as the already mentioned TEI conversion for WebAnno) and private (mostly related to CLARIN.SI services).

5 Support and Dissemination

Apart from the repository and web services offered by CLARIN.SI, the infrastructure also supports the user community via the knowledge centres it belongs to, and by an annual call for projects, as well as by disseminating its results, as discussed in this section.

5.1 Knowledge Centres

CLARIN.SI is active in promoting and encouraging the development of computational linguistics not only for Slovenian but also for other South Slavic languages (i.e., Croatian, Serbian, Bosnian, Montenegrin, Macedonian and Bulgarian), significantly increasing the international use of the RI.

Together with the Bulgarian CLARIN research infrastructure (CLADA-BG) and the Institute of Croatian Language, CLARIN.SI manages the CLARIN Knowledge Centre for South Slavic languages CLASSLA, which provides assistance in the use of language resources and technologies for the whole South Slavic language group. CLASSLA supports researchers with documentation on open language resources, tools for creating and processing text corpora, and other language technologies. In 2024, the centre also organised a series of tutorials called CLASSLA-Express, which is described in detail in Section 5.4.

In addition, the CLASSLA centre develops its own language technologies and corpora to meet the needs of South Slavic languages. Notable resources, developed by the knowledge centre, are the CLASSLA-Stanza models for linguistic annotation (Ljubešić et al., 2024d), discussed in Section 4.2, many linguistically annotated text corpora, including the first general linguistically annotated corpus of Macedonian, CLASSLA-web.mk (Ljubešić et al., 2024c); and the first large and open speech corpora for Croatian and Serbian called ParlaSpeech (Ljubešić et al., 2024b). In 2024, the CLASSLA knowledge centre set up a crawling infrastructure for the annual collection of web corpora for South Slavic languages. The first version of corpora, CLASSLA-web 1.0 (Ljubešić & Kuzman, 2024), comprises 11 billion words in 7 languages, which represent the largest general corpora for most South Slavic languages. The corpora, which are automatically annotated with linguistic and genre information, are available both on the CLARIN.SI concordancers and the CLARIN.SI repository, enabling linguistic research as well as development of language technologies on these languages.

In 2021, CLARIN.SI also became a member of the CLARIN Knowledge Centre for Processing User-Mediated Communication CKCMC, managed by Eurac Research in Bolzano, while in 2024 it established the CLARIN-ELEXIS Knowledge Centre for Lexicography, which operates as a distributed virtual centre supported by 17 institutions from 12 CLARIN National Consortia. It offers expertise and support in using open-access data, tools and services for lexicographers and is a follow-up of the ELEXIS (European Lexicographic Infrastructure) H2020 project.

5.2 Project Support

CLARIN.SI provides financial support for projects selected through an annual call open to consortium members. Since the initiative began in 2018, 36 projects have been successfully concluded, producing notable results, such as the first version of the corpus of parliamentary debates of the National Assembly of the Republic of Slovenia, siParl, now at version 4 (Pančur et al., 2024), the speech corpus Gos Video-Lectures (Verdonik et al., 2019), the LIST tool for analysis of Slovenian corpora (Krsnik et al., 2019), the SloBENCH evaluation framework for language technologies (Žitnik & Dragar, 2021), and teaching materials for corpus approaches to the research of parliamentary discourse (Fišer & Pahor de Maiti, 2021).

5.3 Organization of Events

CLARIN.SI supports events in the field of computational linguistics and related topics that take place in Slovenia, such as the recent 34th European Summer School in Logic, Language and Information (ESSLLI 2023, University of Ljubljana, Faculty of Computer and Information Science) and SyntaxFest 2025 (Faculty of Law, University of Ljubljana).

CLARIN.SI is one of the organisers of the International Conference on Language Technologies and Digital Humanities, a biennial event held in Ljubljana and the primary conference for the field in Slovenia. The conference, first held in 1998 (then called "Language Technologies"), was initiated by the SDJT association, as mentioned in Section 2.1, one of the members of the CLARIN.SI consortium. The conference features invited lectures, on-line proceedings, a student session and associated events, such as tutorials, panels and workshops.

Since 2005, SDJT has been organising a series of lectures, known as JOTA, where CLARIN.SI supports the recording and archiving of lectures on VideoLectures.NET. So far, 20 lectures have been recorded and viewed 15,000 times.

5.4 Promotion and Training

CLARIN.SI presents its activities and those of its knowledge centres at national and international workshops, conferences and events, such as the ESFRI and CLARIN conferences. The work of CLARIN.SI and the CLASSLA K-centre was presented as part of the "Tour de CLARIN" initiative in 2019 (Fišer et al., 2019).

The RI organises training sessions on compiling, depositing and using corpora and other language resources (e.g., using the noSketch Engine concordancer, WebAnno, and Git platforms). In particular, CLASSLA participated in a workshop on using corpora for analysis of the regional variation of gender marking in a language, and, in 2024, organised the CLASSLA-Express series of tutorials, which comprised hands-on exercises on using the CLASSLA-web corpora on the CLARIN.SI concordancers (Ljubešić et al., 2024a). Six CLASSLA-Express training sessions took place so far: Croatia (Zagreb and Rijeka), Serbia (Belgrade), North Macedonia (Skopje), Bulgaria (Sofia), and Slovenia (Ljubljana). The second iteration of CLASSLA-Express in 2025, focusing on comparing traditional corpus-based approaches to approaches using large language models, is visiting five cities in three countries, showing that the workshop series has potential to become a continuous event.

CLARIN.SI shares regular updates on the activities of the Consortium members and Knowledge Centres through its website and the CLASSLA mailing list, with about 100 subscribers from Slovenia and abroad, including Croatia, Serbia, Montenegro, Bulgaria, North Macedonia, Italy, Spain, France, Germany, and the USA.

The infrastructure has accounts and posts news on the social media platforms Discord (100 members), LinkedIn (150 followers) and X (270 followers).

6 Involvement in Projects

CLARIN.SI participates in national and European projects, thus promoting the utilisation of its resources, and enhancing its visibility.

6.1 European Cohesion Policy Funds

The Slovenian 2018–2021 cohesion projects financed upgrading the research equipment of ESFRI infrastructures. With these funds three CLARIN.SI consortium partners upgraded their computer equipment. JSI significantly upgraded its computer cluster, with high-end machines and plentiful disk storage, and likewise CJVT, while the University of Maribor acquired an Nvidia GPU server, which is used for research into deep learning of big data language processing. With these upgrades, CLARIN.SI provides the Slovenian research community with a state-of-the-art research infrastructure, attracting Slovenian partners to international research and innovation projects and supporting scientific excellence. For instance, the EU project MaCoCu (Bañón et al., 2022) used the CLARIN.SI cluster to collect and process large web corpora for over 10 European languages.

6.2 Involvement in European Projects

While CLARIN.SI has not been directly involved in EU projects, there were several language-related projects that were either led by Slovenian researchers or Slovenian teams participated in them, and these projects deposited the language resources they developed in the CLARIN.SI repository (e.g., the MaCoCu web corpora (Bañón et al., 2022), and the news corpora for several languages of the EMBEDDIA project).

As one of its goals, the EU https://elex.is/ELEXIS project had set the creation of a catalogue of online European dictionaries, which was deposited as a dedicated collection into the CLARIN.SI repository. The collection contains metadata and links to the on-line portals of 143 digital dictionaries.

6.3 Involvement in National Projects

CLARIN.SI directly or indirectly participated in numerous national projects, the largest being "Development of Slovenian in a Digital Environment". CLARIN.SI contributed by reviewing language resources created within the project, which were then deposited in its repository, and by defining schemas for the mark-up of Slovenian language resources.

6.4 Cooperation with other infrastructures and initiatives

CLARIN.SI works closely with its two sister RIs in Slovenia, namely DARIAH-SI, the national RI node for digital humanities, and ADP, the national RI node of the CESSDA, the Consortium of European Social Science Data Archives. As already mentioned in Section 2.1, both RIs are led by institutions that are members of the CLARIN.SI consortium.

CLARIN.SI has a long-standing collaboration with DARIAH-SI in the development of encoding and creation of parliamentary corpora, starting with the creation of Slovenian parliamentary corpora (Meden et al., 2024), and continued in the Parla-CLARIN and ParlaMint initiatives, as further explained in Section 6.5. With ADP we collaborated in the project "RDA Node Slovenia" (2019–2020), in the scope of which we established the national RDA Node, which acts as a long-term contact point between the Research Data Alliance and Slovenian data practitioners. In this context CLARIN.SI also reviewed and analysed Slovenian research data repositories (Meden & Erjavec, 2021).

CLARIN.SI is also a founding member of the Slovenian national supercomputer network SLING and of the recently established Slovenian Open Science Community.

6.5 Participation in the Work of CLARIN ERIC

CLARIN.SI plays an active role in the work of CLARIN ERIC, not only by participating in its various bodies, but also by its members directly contributing to the ERIC and by leading CLARIN projects, as evident by the various awards received by its members. In addition, representatives of the CLARIN.SI management committee participate in the CLARIN committees on Legal issues, Standardisation, User involvement, and Technical Centres.

Jakob Lenardič (INZ) received the CLARIN Steven Krauer Award for the young researcher of the year in 2019, also for his work (together with Darja Fišer) in establishing the CLARIN Resource Families initiative (Lenardič & Fišer, 2022a). Tomaž Erjavec received the Steven Krauer Award for CLARIN Achievements 2021 for his work on the ParlaMint project, Darja Fišer (INZ) and Kristina Pahor de Maiti Tekavčič (UL, INZ) received the Teaching with CLARIN Award in 2021 for the best teaching material related to the use of CLARIN resources, while Ajda Pretnar Žagar (UL, INZ), Kristina Pahor de Maiti Tekavčič (UL, INZ) and Darja Fišer (INZ) received the 2022 Teaching with CLARIN Award for their tutorial "What's on the Agenda? Topic Modelling Parliamentary Debates before and during the COVID-19 Pandemic". Kaja Dobrovoljc (UL, JSI) presented CLARIN.SI at the conference on the 20th anniversary of ESFRI in Paris in 2022.

Between 2016 and 2020, Darja Fišer was the CLARIN director for user involvement, and since 2023 she has been the executive director of CLARIN ERIC.

CLARIN.SI (JSI) led two CLARIN projects that included international workshops in 2016 (Ljubljana) and 2019 (Amersfoort). The latter, in cooperation with DARIAH-SI, was dedicated to the development of recommendations for the standardised coding of corpora of parliamentary debates under the name Parla-CLARIN (Erjavec & Pančur, 2022), which has become a popular choice for encoding parliamentary corpora. On this basis, CLARIN.SI acquired a key role in two major CLARIN Flagship projects, ParlaMint I (2020–2021) and ParlaMint II (2022–2023).

The ParlaMint projects created comparable, interpretable and uniformly coded corpora of parliamentary debates and an environment to compile, convert, and process the corpora. In ParlaMint I, CLARIN.SI led the collection and encoding of 17 corpora of national parliaments (Erjavec et al., 2022b). ParlaMint II expanded and enriched the existing corpora while also adding new ones, and resulted in the production of 29 corpora (Erjavec et al., 2024). CLARIN.SI members (co-)led four of the five work packages of the project. In both projects, the corpora were deposited to the CLARIN.SI repository and mounted on its concordancers. Each project released the corpora in three versions (pilot, project final, bug fix) and the corpora are available in several variants (text, linguistically analysed, and machine translated to English in the more recent versions). Additionally, ParlaSpeech corpora consisting of subsets of ParlaMint corpora with aligned speech (Ljubešić et al., 2024b) were released for four languages, Croatian (already in version 2), Serbian, Polish and Czech, spanning five thousand hours of recorded material. Currently, an extension of the ParlaSpeech corpora collection is being prepared, which includes the Slovenian, Bulgarian, Bosnian, Serbian, and Ukrainian ParlaMint dataset. The repository thus currently hosts 19 ParlaMint entries, while the concordancers (which have each country as a separate corpus + joint parallel original-MTed corpus + three ParlaSpeech corpora) mount 139 ParlaMint corpora.

The results of the ParlaMint project are now being used in two successor projects. The on-going OSCARS project ParlaCAP "Comparing agenda settings across parliaments via the ParlaMint dataset´´, which we lead, aims to enhance the speeches in ParlaMint corpora with information about sentiment expressed and topic discussed, add information about political parties from new resources, and especially open this dataset for political scientists and other researchers that rely on tabular data rather than on textual collections. The project PressMint "Interoperable corpora of historical newspapers" has been accepted for funding in the 2025 CLARIN Flagship project call, and will take the developed ParlaMint infrastructure and apply it to compiling a set of historical newspaper corpora. In this project we lead a work-package and several tasks and will contribute the Slovenian corpus.

7 Conclusions

The paper has presented the Slovenian CLARIN infrastructure in its tenth year of existence. The focus has been on the management of CLARIN.SI, its repository for language resources and tools, the web services it offers, its contributions to dissemination activities and support of the field in Slovenia, and its involvement in various projects and in the work of CLARIN ERIC. The overview shows that CLARIN.SI is an established infrastructure that covers a wide interdisciplinary field and supports both basic and applied research, as well as the development of resources and tools.

As regards further work, the general directions are given in the CLARIN.SI strategy 2024–2030, which was written and approved by the CLARIN.SI Management Committee at the end of 2023. The strategy follows the CLARIN ERIC Strategy 2024–2026 but is focused on the Slovenian node of CLARIN and on cooperation of CLARIN.SI with and within CLARIN ERIC. The CLARIN.SI Strategy is aligned with Slovenia's Research Infrastructure Roadmap 2030, and thus also covers a longer time period than the CLARIN Strategy.

Specifically, we plan in the next period, first and foremost, to maintain and support existing services, thus justifying our status as an infrastructure. Here the main challenge we face is upgrading our current repository platform to the new version being developed by LINDAT/CLARIAH-CZ, which will involve a complex migration procedure of the documentation and URLs.

Next, CLARIN.SI will continue to promote the production of FAIR research data as well as data reuse, especially among humanities researchers. This will necessitate strengthening user support, including education and training, especially as universities and research agencies are increasingly demanding from researchers in doctoral and research programs plans for handling research data and their permanent storage. We plan to adopt a more proactive approach to conducting lectures and workshops for students, lecturers and researchers (as piloted e.g., by the CLASSLA-Express series of tutorials), to provide assistance in creating a data management plan for students and researchers, and to extend our online documentation and tutorials. A survey is currently being prepared to gather feedback on user experience with the CLARIN.SI infrastructure. The survey aims to identify key issues and priorities as seen by the users, helping to harmonise future work with user needs.

There is also the growing importance of research infrastructures for capturing, storing and processing data from social networks and the web. CLARIN.SI has already paid special attention to such language resources, and in the future, it will continue these activities for all South Slavic languages within the CLASSLA knowledge centre.

It is, furthermore, important to instrumentalise and make accessible data and services relevant to individual communities. The CLARIN.SI consortium currently includes 12 members, which cover the majority of Slovenian stakeholders who either produce or use language resources and technologies, but not all of them. In the next period, CLARIN.SI will try to expand its consortium to also cover communities of potential users of the infrastructure that have not yet been included in its operation.

Last but not least, Slovenia's Research Infrastructure Roadmap 2030 states that Slovenia "plans to continue and strengthen activities within the framework of international CLARIN projects" (p. 60), acknowledges the existing cooperation with RI DARIAH-SI and CESSDA/ADP, and foresees connection with new RIs, namely OPERAS (Open scientific communication in the European research area for social sciences and humanities), which is managed in Slovenia by ZRC SAZU, and with PRACE (Partnership for Advanced Computing in Europe), managed by ARNES, the Academic and Research Network of Slovenia.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful suggestions. The presented work was supported by the Slovenian research agency ARIS in the scope of its funding for ESFRI research infrastructures.

References

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., et al. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. *23rd Annual Conference of the European Association for Machine Translation*, 301–302. https://aclanthology.org/2022.eamt-1.41/

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*, 209–226. https://doi.org/10.1007/s10579-009-9081-4

Bogunović, I., Kučić, M., Ljubešić, N., & Erjavec, T. (2021). *Corpus of Croatian news portals ENGRI (2014-2018)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1416

Daza, A., Fokkens, A., & Erjavec, T. (2022). Dealing with abbreviations in the Slovenian biographical lexicon. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8715–8720. https://aclanthology.org/2022.emnlp-main.596/

Dobrovoljc, K. (2024a). Extending the Spoken Slovenian Treebank. *Proceedings of the Conference on Language Technologies and Digital Humanities*, 113–143. https://doi.org/10.5281/ZENODO.13936394

Dobrovoljc, K. (2024b). Can't See the Forest for the Trees: Tools and Services for Investigating Slovene Dependency Treebanks. *Proceedings of the CLARIN Annual Conference*. https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2023). Universal dependencies za slovenščino. *Slovenščina 2.0*, *11*(1), 218–246. https://doi.org/10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2015). The IMP historical Slovene language resources. *Language Resources and Evaluation*, *49*, 753–775. https://doi.org/10.1007/s10579-015-9294-7

Erjavec, T. (2023). *Corpus of combined Slovenian corpora metaFida 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1775

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022a). Raziskovalna infrastruktura CLARIN.SI (The CLARIN.SI research infrastructure). *Proceedings of the Conference on Language Technologies and Digital Humanities*, 47–54. https://doi.org/10.5281/zenodo.14165471

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, *55*(2), 551–583. https://rdcu.be/b7GrB

Erjavec, T., Holdt, Š. A., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., & Zupan, K. (2016). Annotating CLARIN.SI TEI corpora with WebAnno. *Proceedings of the CLARIN annual conference*. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf

Erjavec, T., Javoršek, J. J., & Krek, S. (2014). Raziskovalna infrastruktura CLARIN.SI. *Zbornik Devete konference JEZIKOVNE TEHNOLOGIJE*. https://nl.ijs.si/isjt14/proceedings/isjt2014_03.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, Starkaður, 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., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2022b). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., & Pančur, A. (2022). The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative (Selected Papers from the 2019 TEI Conference)*, (14). https://doi.org/10.4000/jtei.4133

Fišer, D., Lenardič, J., Auziņa, I., Bernstein Ratner, N., De Smedt, K., Dobrovoljc, K., Dodé, R., Domeij, R., Dyvik, H., Erjavec, T., Gerassimenko, O., Hajič, J., Křen, M., Ljubešić, N., MacWhinney, B., Monachini, M., Nava, B., Navarreta, C., Nedyalkova, A., … Vider, K. (2019). *Tour de CLARIN Volume Two*. Zenodo. https://doi.org/10.5281/zenodo.3754164

Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, *54*, 223–246. https://rdcu.be/7RX4

Fišer, D., & Pahor de Maiti, K. (2021). “Prvič, sem političarka in ne politik, drugič pa…”: Korpusni pristop k raziskovanju parlamentarnega diskurza. *Contributions to Contemporary History*, *61*(1). https://doi.org/10.51663/pnz.61.1.07

Hmeljak, K., & Erjavec, T. (2010). The Japanese-Slovene dictionary jaSlo: its developments, enhancement and use. *Studia Kognitiva*, *10*, 211–224. https://nl.ijs.si/jaslo/bib/HmeljakErjavec2010.pdf

Jong, F. D., Maegaard, B., Smedt, K. D., Fišer, D., & Uytvanck, D. V. (2018). CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1515

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. (2022). Trendi - a monitor corpus of Slovene. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*, 230–239. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p1-21_Front-matter.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

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S., & Robnik-Šikonja, M. (2019). *Corpus extraction tool LIST 1.2*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276

Lenardič, J., & Fišer, D. (2022a). The CLARIN resource and tool families. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 343–372). https://doi.org/10.1515/9783110767377-013

Lenardič, J., & Fišer, D. (2022b). CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement. *Proceedings of the CLARIN Annual Conference*. https://www.clarin.eu/event/2022/clarin-annual-conference-2022

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282. https://aclanthology.org/2024.lrec-main.291/

Ljubešić, N., Kuzman, T., Petrović’ Filipović’, I., Parizoska, J., & Osenova, P. (2024a). CLASSLA-Express: a Train of CLARIN.SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings*, 31. https://doi.org/10.48550/arXiv.2412.01386

Ljubešić, N., Rupnik, P., & Koržinek, D. (2024b). The ParlaSpeech collection of automatically generated speech and text datasets from parliamentary proceedings. *International Conference on Speech and Computer*, 137–150. https://doi.org/10.1007/978-3-031-77961-9_10

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024c). *Macedonian web corpus CLASSLA-web.mk 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1932

Ljubešić, N., Terčon, L., & Dobrovoljc, K. (2024d). CLASSLA-Stanza: the next step for linguistic processing of South Slavic languages. *Proceedings of the Conference on Language Technologies and Digital Humanities*, 251–274. https://zenodo.org/records/13936406

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

Meden, K., & Erjavec, T. (2021). *Pregled slovenskih repozitorijev raziskovalnih podatkov* (tech. rep.). Jožef Stefan Institute. CLARIN.SI. https://www.clarin.si/info/services/projects/%5C#RDA_Node_Slovenia

Meden, K., Erjavec, T., & Pančur, A. (2024). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09746-8

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

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.48550/arXiv.2003.07082

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. https://nlp.fi.muni.cz/raslan/2007/papers/12.pdf

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A. (2019). *Spoken corpus Gos VideoLectures 4.0 (transcription)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1223

Vintar, Š., & Martinc, M. (2022). Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 129–156. https://doi.org/10.1075/term.21005.vin

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. https://aclanthology.org/P13-4001

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/

Žagar, K., Ferme, M., Ojsteršek, M., Jemec Tomazin, M., & Erjavec, T. (2023). *Corpus of scientific texts from the Open Science Slovenia portal OSS 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1774

Žejn, A., & Šorli, M. (2023). Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus. *Slovenščina 2.0*, *11*(1), 118–137. https://doi.org/10.4312/slo2.0.2023.1.118-137

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1469

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., Potočnik, T., Sepesy Maučec, M., Majhenič, S., Žgank, A., Bizjak, A., Gril, L., Dobrišek, S., Križaj, J., Bajec, M., Lebar Bajec, I., Jelovšek, T., Trojar, M., Bernjak, M., … Dobrovoljc, K. (2023). *Spoken corpus Gos 2.0 (transcriptions)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1771

Policy Domains and the Speakers' Gender in ParlaMint-DK 4.1

Costanza Navarretta
University of Copenhagen, Denmark
costanza@hum.ku.dk

Dorte Haltrup Hansen
University of Copenhagen, Denmark
dorteh@hum.ku.dk

Bart Jongejan
University of Copenhagen, Denmark
bartj@hum.ku.dk

Abstract

In this paper, we describe the ParlaMint-DK 4.1 corpus, which consists of the Danish parliament speeches from 2014 to 2022 annotated with 20 general policy domains mapped to the codebook of the Comparative Agendas Project. The policy domains were added to the speeches semi-automatically using the agenda titles under which the speeches occurred. In the paper, we also account for how some of the linguistic annotations of the corpus were improved using the Text Tonsorium and present some of our previous studies on parliament data. We also describe novel investigations, based on the policy domain annotations in ParlaMint-DK aimed at determining which domains are most frequently addressed in the speeches and the frequency by which policy areas are debated by female and male politicians during the various governments covered by the corpus.

1 Introduction

In this paper, we present the latest version of the ParlaMint-DK corpus, ParlaMint 4.1, focusing on the policy domain annotation, which it contains. ParlaMint-DK is the Danish part of the corpora of parliamentary speeches that were collected and annotated under the CLARIN's flagship project ParlaMint (Erjavec, Kopp, Ljubešic, et al., 2024)[1]. ParlaMint-DK 4.1 differs from the other ParlaMint corpora, because it contains the annotations of policy domains. These annotations are mapped to the categories in the Comparative Agendas Project[2]'s codebook. The annotation of general policy areas or domains in parliamentary debates has been addressed by political scientists for a long time, since these enable the analysis and comparison of how specific topics are dealt with by different parties or political wings, nationally or internationally, in various periods of time (Baumgartner et al., 2011; Ivanusch, 2024; Merz et al., 2016a; Ristilä & Elo, 2023; Yu et al., 2023; Zirn et al., 2016).

The policy domain annotations were first added to some of the Danish speeches in a pilot study described in (Hansen et al., 2019). The speeches used in this work covered the period from October 2009 to June 2017 and were released as the Danish Parliament corpus with subject annotations v.2 (Hansen & Navarretta, 2021) in 2021. This corpus, as it is also the case for ParlaMint-DK, was downloaded from the Danish Parliament (*Folketinget*)'s website[3] and is available in CSV format. The corpus was used for training and testing classification algorithms to identify the speeches' main policy domain automatically. Hansen et al. (2019) report a 0.8 F1-score when identifying 18 domains in a balanced data set. In other experiments, classifiers were trained to identify both primary and secondary policy domains in speeches that are annotated with two domains (Navarretta & Hansen, 2022). The policy domain annotations of the Parliament corpus v.2 with subject annotations were also used in other studies, for example, to investigate how *Immigration* and *Environment* have been handled by different left- and right-wing parties in the covered period (Navarretta & Hansen, 2023; Navarretta et al., 2022).

[1] The corpora are available at (Erjavec, Kopp, Ogrodniczuk, Osenova, Agerri, et al., 2024; Erjavec, Kopp, Ogrodniczuk, Osenova, Agirrezabal, et al., 2024)

[2] https://www.comparativeagendas.net/

[3] ftp://oda.ft.dk

Costanza Navarretta, Dorte Haltrup Hansen and Bart Jongejan 2024. Policy Domains and the Speakers' Gender in ParlaMint-DK 4.1. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 86–99. DOI: https://doi.org/10.3384/ecp216

Recently, we coded the policy domains of the speeches from June 2017 to June 2022 and added these to the speeches in the ParlaMint-DK corpus. In this paper, we use these policy domain annotations to determine which areas were most frequently discussed in the Danish parliament, and to find which policy domains were more often addressed by female and male politicians, and whether the frequency of speeches about the various policy domains by the two genders changes over time. ParlaMint-DK 4.1 also contains improved linguistic annotations compared to version 4.0, and in the present paper we also describe these improvements and how they were achieved.

The paper is organized as follows. In section 2, we present studies aimed at classifying policy areas in political discourse, and in section 3 we account for the classification system which we adopted, and the annotation method which we applied. In section 4, we describe how the linguistic annotations of ParlaMint-DK were improved while in section 5, we discuss the distribution of the main policy domains and the co-occurring policy areas in the corpus. In section 6, we account for which policy areas were most frequently addressed by female and male politicians during the governments covered by ParlaMint-DK. Finally, section 7 contains a short conclusion and discussion of future work.

2 Background work

Various classifications of policy domains have been proposed to cover different types of data. The classification systems most often used are the ones created by the Comparative Manifesto Project[4] (Merz et al., 2016b), and by the Comparative Agendas Project[5] (Baumgartner et al., 2011).

The Comparative Manifesto Project classification is fine-grained and comprises more than 550 categories used to annotate so-called quasi-sentences[6] in party election manifestos from many countries. The annotations distinguish positive and negative quasi-sentences and are produced manually.

The scheme used in the Comparative Agendas Project (CAP) builds on the classification applied in the Policy Agendas Project[7], whose aim was to structure the US policy data. The CAP scheme is a modified version of this classification to cover not only the policy activities of the US data, but also the policy domains of other countries (Baumgartner et al., 2011). The CAP classification scheme comprises 21 main domain categories and 192 subcategories. Danish researchers in political science from the University of Aarhus have manually annotated political data from 1953 to 2007 in the Danish Policy Agendas Project[8] using an adapted version of the CAP scheme. We have been inspired by their work.

3 The classification of Policy Areas in ParlaMint-DK

The classification scheme of policy domains which we have used in ParlaMint-DK 4.1, consists of 20 classes. 19 of these correspond to the areas of responsibilities in the Danish Parliament (spokesmanships) in the covered period, while the latter class, *Other*, was added to cover government operations. Table 1 shows the 20 policy domain classes, the corresponding areas of responsibility in the Danish parliament, the corresponding CAP codes and CAP areas.

3.1 The annotation method

The policy domain annotations were semi-automatically added to the speeches in The Danish Parliament Corpus (2009-2017) extracting them from the titles of the agenda items of the meetings. The method was described in (Hansen et al., 2019) where the first pilot annotations were presented. The method consists of the following steps: 1) extraction of the agenda titles 2) normalization, e.g., “Third reading of bill N: XYZ” becomes ”XYZ”, 3) manual annotation of the agenda titles with one or two policy areas, and 4) automatic assignment of the policy area(s) to each speech in the meeting covered by the relevant agenda titles. For example, for the title *Tax on saturated fat in food*, the domain *Agriculture*, which comprises the food subcategory, was assigned as the primary policy domain, while *Economy*, which comprises *Tax*,

[4] https://manifesto-project.wzb.eu/

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

[6] Quasi-sentences correspond to sentences in the majority of cases, but they can also indicate entities such as titles of films and books.

[7] https://liberalarts.utexas.edu/government/news/feature-archive/the-policy-agendas-project.php

[8] http://www.agendasetting.dk/.

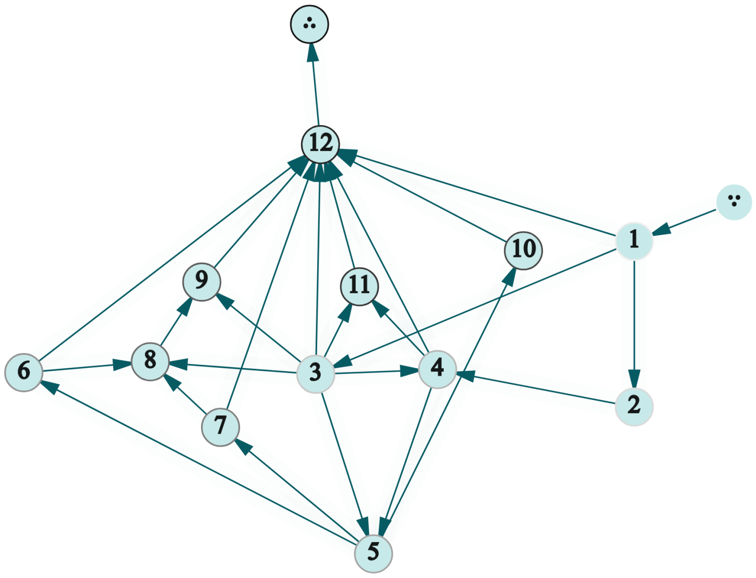
Policy Domain	**Area of Responsibility**	**CAP no.**	**CAP Areas**
Economy	Finance, Fiscal Affairs	1	Domestic Macroeconomic Issues
Health Care	Psychiatry, Health	3	Health
Agriculture	Animal Welfare, Fisheries, Food, Agriculture Consumer Policy	4 1525	Agriculture Consumer Policy
Labour	Labour market	5	Labour and Employment
Education	Higher Education and Research Education	6	Education
Environment	Environment	7	Environment
Energy	Energy Climate	8 705	Energy Air and Noise Pollution, Climate Change and Climate Policies
Immigration	Immigration and Integration, Alien Affairs, Naturalization	9	Immigration and Refugee Issues
Infrastructure	Transportation IT, Media	10 17	Transportation Space, Science, Technology and Communications
Justice	Legal Affairs Constitutional Matters	12 20	Law, Crime, and Family Issues Government Issues
Social Affairs	Children, Family, Social Services, Senior Citizens Gender Equality	13 2	Social Welfare Civil Rights, Minority Issues, and Civil Liberties
Housing	Housing	14	Community Development & Housing Issues
Local and Regional Affairs	Rural Districts and Islands Municipal Affairs	4 2001	Community Development & Housing Issues Local Government Issues
Business	Trade and Industry	15	Industrial and Commercial Policy
Defence	Defence	16	Defence
Foreign Affairs	Foreign Affairs, Development, Cooperation	19	International Affairs and Foreign Aid
European Integration	EU	1910	International Affairs and Foreign Aid
Territories	Faroe Islands, Greenland	2105	Dependencies and Territorial Issues
Culture	Cultural Affairs Ecclesiastical Affairs Sport	23 210 1526	Cultural Policy Issues The Danish National Church Sport and Gambling
Other	-	2000	Government Operations

Table 1: Policy domains, and corresponding responsibility areas, CAP numbers, and CAP areas in ParlaMint-DK

was chosen as the second domain. More difficult cases require knowledge of the content of specific bills. For example, the title of the agenda *1. Behandling af B 47 om Amager Fælled* (1. Treatment of B 47 about Amager Fælled) refers to a bill proposing to sell a green area in Copenhagen as building area. The speeches referring to this agenda item were classified with the main domain *Local and Regional Affairs* and the secondary domain *Housing*. Later, when the parliament discussed the legality of selling this area with respect to the Environment legislation, the speeches about *Amager Fælled* were annotated with the policy domain *Justice*.

5000 speeches that were coded with two policy areas were reviewed by two annotators independently. The two annotators did not find any errors in the assignment of the two policy areas, but in some cases they disagreed on which of the two annotated areas should be considered the primary (Navarretta & Hansen, 2022).

The extended annotations of policy domains covering speeches from 2017 to 2022 were performed according to the same methodology as in (Hansen et al., 2019), but, in the ParlaMint-DK annotations of policy domains, we decided to use an extra domain *Other*, which covers speeches about government operations and other issues. To add all the annotations to the ParlaMint-DK speeches, we first created a TEI taxonomy over the policy domains, and then the policy domain annotations were added to each



ID	Tool	ID	Tool
∵	input (TEI P5)	7	Anno-splitter (PoS tags)
1	TEI tokenizer (token & sentence boundaries)	8	PoS translator (UD → CST)
2	Sentence extractor (sentence boundaries)	9	CSTlemma (replaces UD lemmas)
3	Token extractor (defines token ID's)	10	Anno-splitter (syntax)
4	TEI-segmenter (defined as token ID ranges)	11	CSTner (named entities)
5	udpipe (PoS, lemmas, morphology, syntax)	12	TEI annotator (aggregates)
6	Anno-splitter (morphological features)	∴	output (TEI P5)

Figure 1: The Text Tonsorium workflow for the annotation of the Danish ParlaMint dataset.

speech as an @*ana* attribute in the *u*-element[9]. Since we had assigned a unique unifier to each parliament speech, and the unifier is based on the exact time and date of each speech, this step was trivial.

4 The linguistic annotations and Text Tonsorium

The linguistic annotations were performed as in the previous versions of ParlaMint-DK through Text Tonsorium[10] (TT) (Jongejan et al., 2021). TT is a workflow management system that not only executes linguistic annotation workflows, but can also compose workflows by combining different Natural Language Processing tools. The linguistic annotations of ParlaMint-DK were made using ten different tools in a workflow comprising twelve steps, see figure 1.

Evaluating the linguistic annotations of ParlaMint-DK 4.0, we found some systematic lemma annotation errors. We corrected them by taking morphology and word form into account when mapping between the Universal tag set output by UD-pipe and the CST tag set used by CSTlemma (Jongejan, 2016). Already for the first published version of ParlaMint-DK we decided not to use the UD-pipe software for delivering lemma annotations because, as a lemmatizer, UD-pipe performs worse than CSTlemma. Also, we required that the application of a lemmatization rule was conditioned on the word class of the input word, while UD-pipe often applies lemmatization rules that are not meant for the word classes

[9] This was done after consulting the head of the ParlaMint project, Tomaž Erjavec.

[10] Via CLARIN-DK website: https://dkclarin.dk/clarin.dk/ or directly: https://cst.dk/texton/, Source code: https://github.com/kuhumcst/texton

assigned by the UD-pipe software. Previously, the mapping from the UD-tag set onto the CST tag set was performed by CSTlemma itself, using a simple lookup table. Now, the mapping task is delegated to a separate tool, the PoS translator (tool 8 in figure 1). The tool combines information from tokenization, PoS-tagging and morphological analysis to provide the correct information to CSTlemma.

The NER annotations were also improved in this version of the annotated corpus. This improvement especially dealt with the abbreviations of parties' and organizations' names. The refined mapping between tag sets had the desired effect and the frequent, systematic lemmatization errors and many NER errors were gone.

The TT is also available via the CLARIN Language Switchboard. We are currently working on additions to its toolbox that can be interesting to users outside CLARIN-DK. We are integrating the Stanford CoreNLP tools to handle Chinese, English, French, German, Hungarian, Italian, and Spanish texts even better than we already do. We are also continuously extending the TT's capabilities and improving its user interface. A recent example of the latter is the graphical representation of workflows. A TT workflow is structured as an Acyclic Directed Graph (DAG) with potentially a few tens of nodes. In the new graphical representation, each tool is represented by a numbered circle. Data streams are drawn as arrows between these circles, with the arrow head indicating the direction towards the output. Input and output are represented as circles with ∵ resp. ∴ symbols (see figure 1 for an example). The TT consists of a toolbox and software that manages the tools and the metadata about the tools. The toolbox is in part filled with 3rd-party tools (UD-Pipe, CoreNLP, Tesseract and many others) and in part with our own tools. All software in the TT is free, open source and not depending on frameworks, DBMS'es, virtualization or new technologies with short expected lifespans. Although automated, the TT is not a black box; it can be opened and inspected. There is also an option to export a workflow as a list of command lines, so users with technical skills can recreate workflows even after the TT has gone out of service.

Compared to other workflow managers, the management of the tools by the TT is more advanced, since the TT computes workflows while other workflow managers depend on human workflow architects[11]

5 ParlaMint-DK 4.1

ParlaMint-DK 4.1 comprises the transcriptions of 398,610 speeches. The transcriptions were produced by the *Office of the Parliament Hansard*. They state that the transcribed speeches are reported literally, but with small editions whose aim is to adapt the speeches into a colloquial and syntactically coherent written language ensuring that the intentions of the speaker are clear. Factual errors and minor speech errors are corrected in the transcribed speeches. Moreover, punctuation marks have been added to the transcriptions, and spoken language characteristics, such as pauses, speech marks, hesitations, and self corrections have been left out (Hansen et al., 2018). One characteristic of the Danish parliament is that politicians must follow specific rules during the debates. The rules are spelled out in *The Standing Orders of the Danish Parliament*, which the Speaker (the chairman of the Parliament), the Chair, henceforth, enforces during the debates. According to these rules, expressions of approval or disapproval during the debates are considered disorderly, and the Chair can stop a politician who is judged to say something improper.

5.1 A quantitative analysis of the policy domains

In our analysis of policy domains in ParlaMint-DK 4.1, we removed the utterances of the Chair. These utterances have the same policy domain annotation as the speeches that are chaired in that section, if they occur under an agenda point with that policy domain. These utterances should not be considered when analyzing the content of policy domains since the Chair does not address the domains and only chairs the meetings, e.g. introducing the various speakers or enforcing the rules for the debates. The utterances by the Chair and speeches which do not address a policy domain[12] can be though interesting in studies

[11] Due to the complexity of the task, it would have been hard and time-consuming to implement the TT in main-stream programming languages such as Java, Python, or C++. We have used an internally developed programming language, Bracmat (https://github.com/BartJongejan/Bracmat), itself written in C, for all functionality not related to the handling of internet protocols. For the latter, we used Java and PHP.

[12] These speeches have no domain annotation.

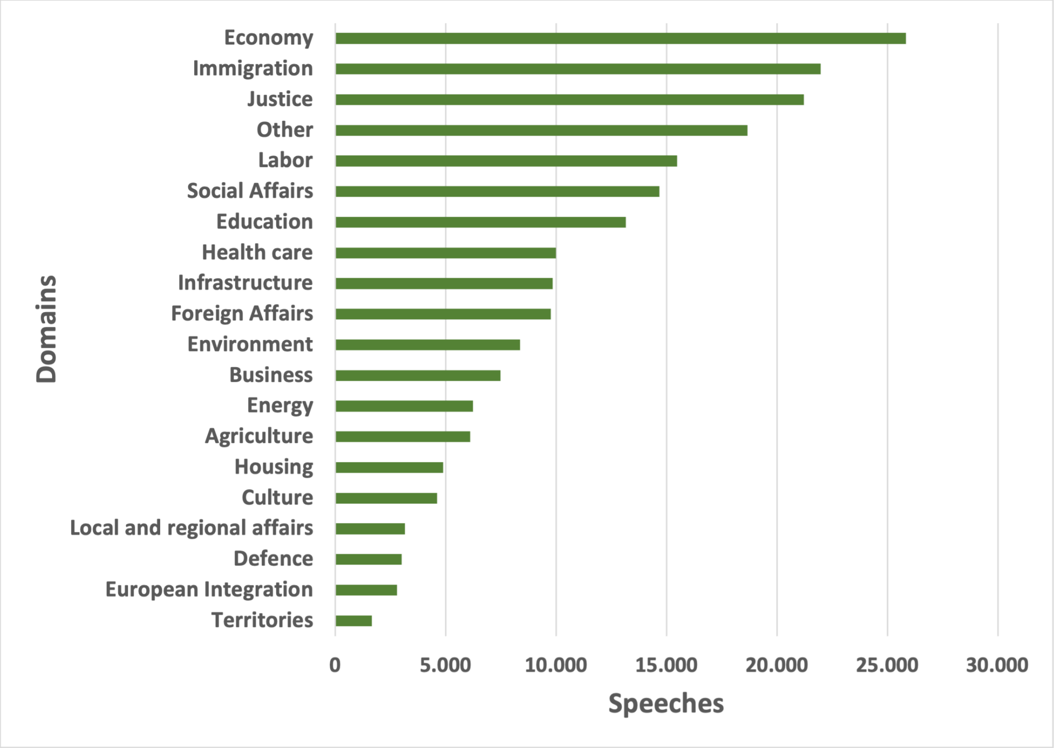


Figure 2: The distribution of policy domains.

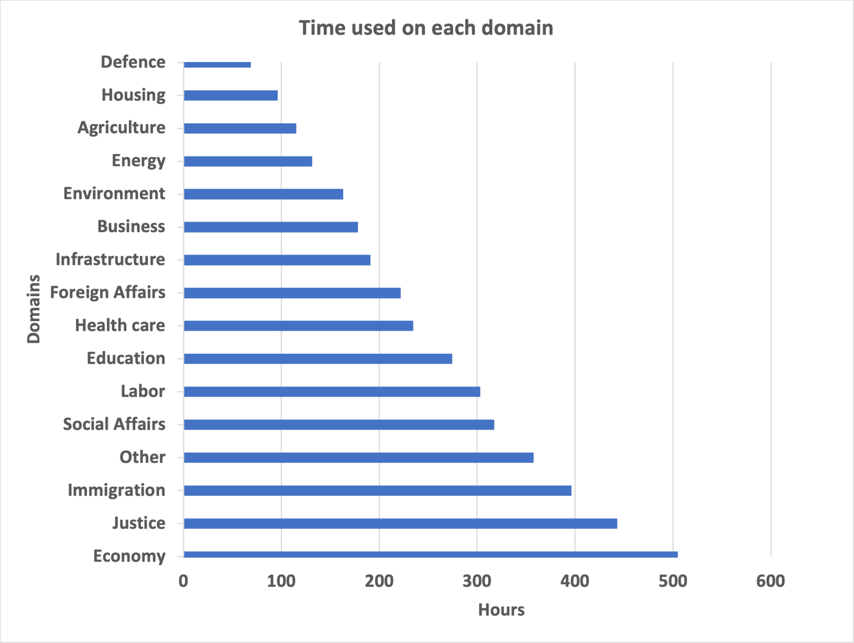


Figure 3: The Duration of speeches about policy domains.

addressing for instance the debates' dynamics. After removing the utterances of the Chair, we obtained 208,881 speeches with policy domains.

The distribution of the most frequently debated policy domains in ParlaMint-DK is given in figure 2. The policy domain that is most frequently addressed in the speeches is *Economy*, followed by *Immigra-*

tion, *Justice*, and *Labor*. The prominence of the *Economy* domain is not surprising, while the frequency of speeches about *Immigration* indicates the importance that this topic has had in Danish politics the past ten years. In one of our previous studies (Navarretta et al., 2022), we investigated how *Immigration* was addressed by seven Danish main parties not only in the speeches, but also in their manifestos.

The policy domain that is debated for the longest period of time is *Economy*, which relates to it being the most frequent speech domain. The second domain with respect to the duration of the speeches, which address it, is *Justice*, followed by *Immigration*, *Social Affairs* and *Labor*.

Approximately 18% of the speeches in ParlaMint-DK (38,425) are annotated with two policy domains. In Figure 4, the most frequently co-occurring domains in the corpus are displayed. They are the following: a) *Labor* and *Social Affairs*, b) *Immigration* and *Justice*, c) *Justice* and *European Integration*, d) *Economy* and *Regional affairs*.

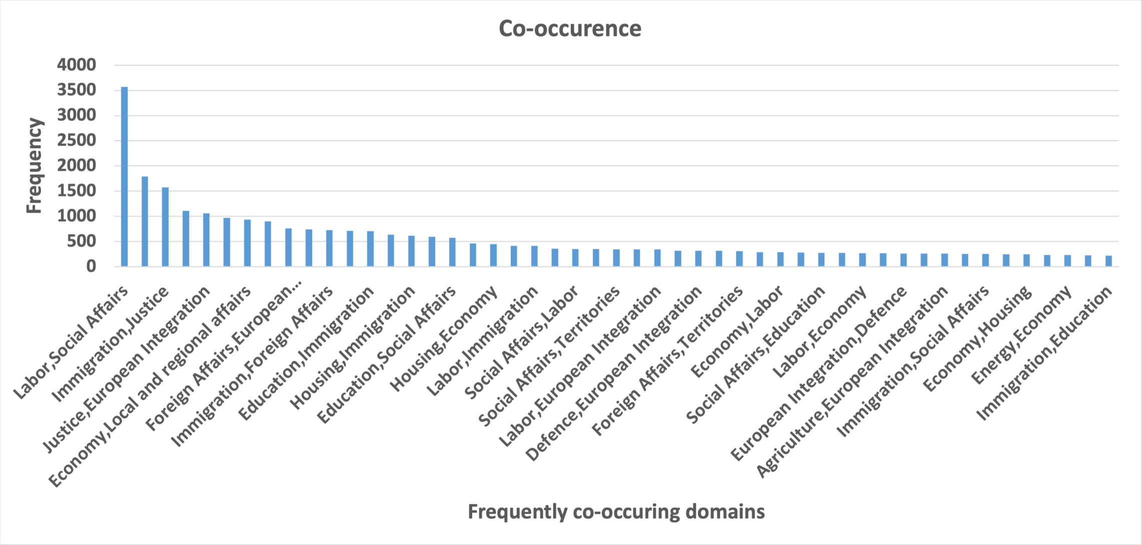


Figure 4: The Distribution of co-occurring policy domains.

6 Investigation of the policy domains most often addressed by women and men

6.1 Related studies

In this study, we used the annotations of policy domains in ParlaMint-DK 4.1 to investigate which areas were more often addressed by male vs. female politicians, and whether the most frequently debated domains by each of the two groups change over time. Gender differences in political speeches have been addressed in studies focusing on various aspects. For example, Paxton et al. (2007) analyze a number of articles describing the political participation of women in different countries. They conclude that even if all studies show that the number of female parliament members is low, the figures vary from country to country. The female parliament members are most numerous in the Scandinavian countries, while they are fewest in the Middle East. In 2005 their percentage was 38% in Scandinavia and 8% in the Middle East.

Dahllöf (2012) addresses the automatic identification of gender, age (young vs. old) and political affiliation (left or right wing) of politicians in the initial words of selected transcriptions of the Swedish parliament debates from 2003 to 2010. Words characterizing each class were used as features, and a support vector machine was applied to the data for classification. The accuracy for gender identification on balanced train and test sets was around 0.6 and gender identification was better for right-wing than for left-wing politicians. Mandravickaitė and Krilavičius (2017) compare the most frequent words in the transcripts of parliamentary speeches by female and male members of the Lithuanian parliament, and they apply hierarchical clustering to samples obtained on the basis of similarity measures. They find

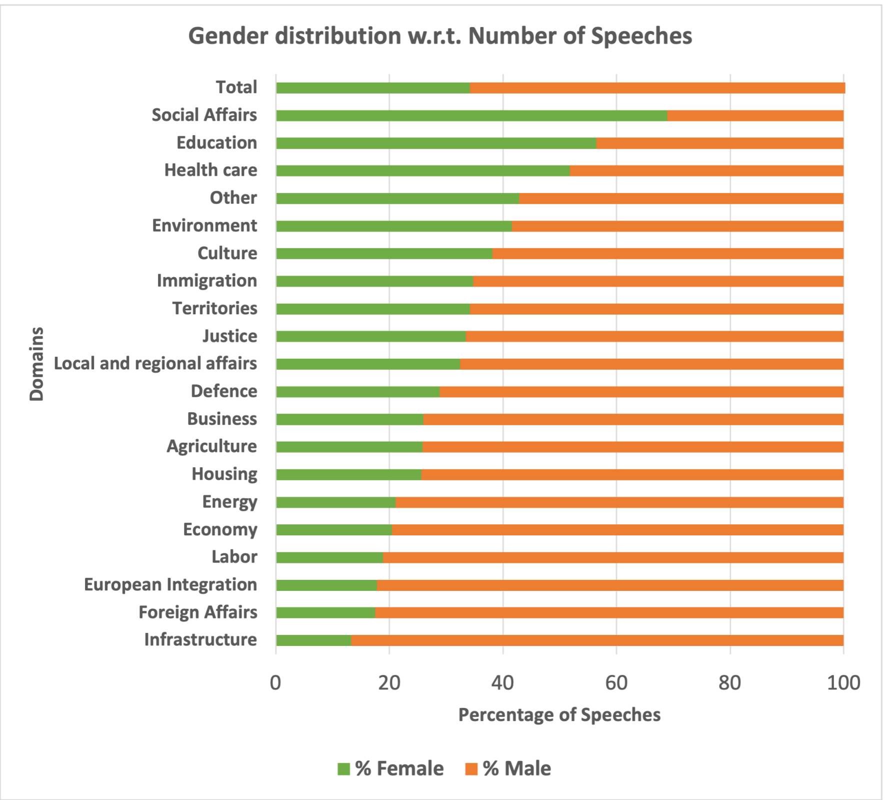


Figure 5: The distribution of speeches by female and male politicians.

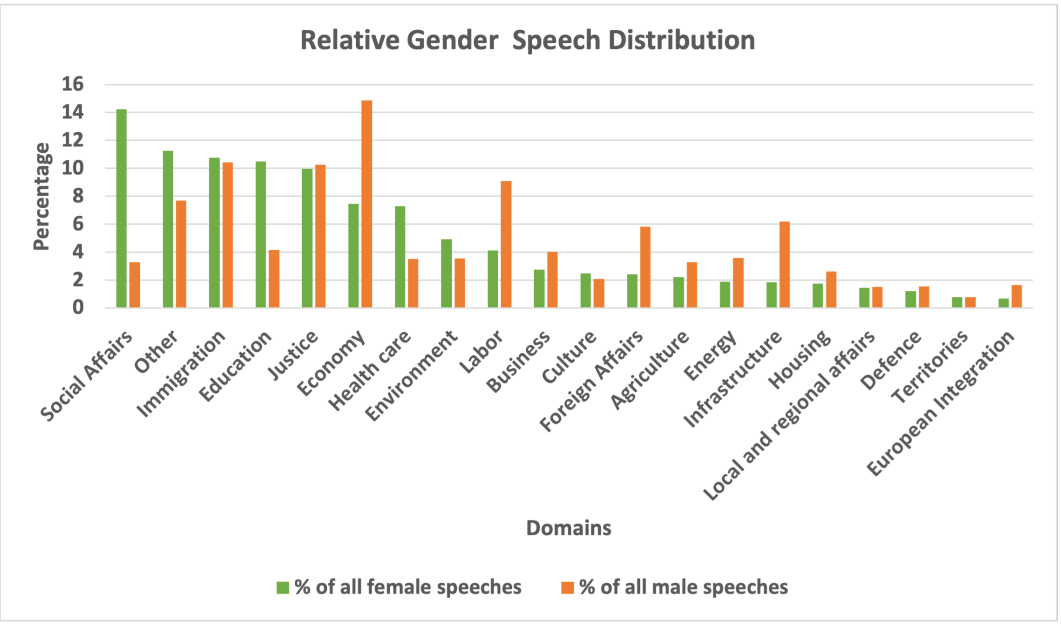


Figure 6: The relative distribution of speeches by female and male politicians.

differences in both the lexical items used by female and male politicians and in the stylometric figures for the speeches of the two groups.

Hansen et al. (2018) address gender differences in the Danish Parliament Corpus (2009-2017). They find statistically significant differences in the number and duration of speeches by male and female politicians, and their results show that the role of politicians in their party influences their participation in the debates. Hansen et al. (2018) also find that female ministers talk more in periods when the prime minister is a female than they do when the prime minister is a male. This indicates that female politicians are inspired by the presence of a female head of government. The positive influence of female ministers on the active participation of female parliament members in the British debates is underlined in a study by Blumenau (2021).

Bäck and Debus (2019) study the Swedish parliamentary debates from 2002 to 2010 and find that female politicians take the floor less often than male politicians, and they talk less about ”harder” policy areas such as energy, finance, macro economy, foreign affairs, and national security than their male colleagues.

Analyzing 200 debates of the UK's House of Commons between 1997 and 2006, Hargrave and Langengen (2021) find that female politicians refer more often to their personal experience, discuss policies in a more concrete way, and are less antagonistic than men in their speeches. Hargrave and Blumenau (2022) also find that the debating styles of the female members of the British parliament have changed over time (1997-2019) and that women have adopted stylistic traits, which traditionally have been considered masculine. The authors conclude that the division between female and male stylistic traits for politicians does not longer correspond to the reality and builds upon old stereotypes.

6.2 Our analyses

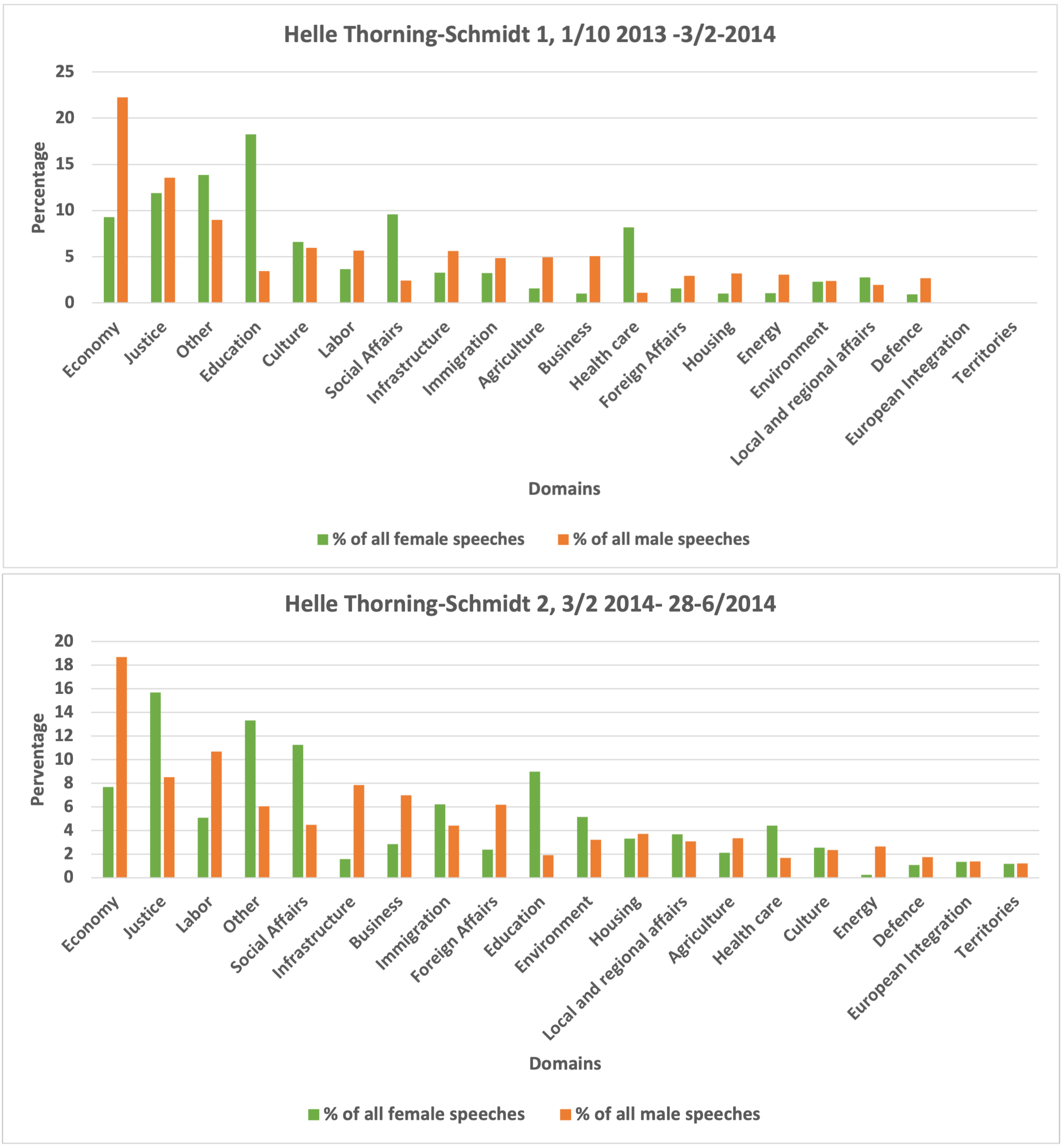
In the following, we investigate the frequency of speeches about the various policy domains addressed by female and male politicians in ParlaMint-Dk 4.1 and, inspired by Hargrave and Blumenau (2022), we also look at whether the frequency figures change over time. In the period covered by our corpus, the percentage of women in the Danish parliament is of 38.5%. More specifically, the percentage is 39.1% after the elections in 2011 and 2019, and 37.4% after the elections in 2015. The percentage of speeches by women in the same period is slightly lower (32.3%), which could be due to the fact that the majority of ministers have been men. Figure 5 shows the absolute frequency of speeches by women and men about each policy domain in ParlaMint-DK 4.1. The policy domains that are addressed by female politicians more frequently than by male politicians are *Social Affairs*, *Education*, and *Health Care*. The speeches regarding *Environment*, *Culture* and *Immigration* are also frequent. The policy domains that are debated by women less often than by men are *Labor*, *European Integration*, *Foreign Affairs* and *Infrastructure*.

According to the frequency numbers in figure 5, the women in the Danish parliament, similarly to the women in the Swedish parliament (Bäck & Debus, 2019), talk more often about ”soft” policy areas than men. This was also found in the Danish speeches from 2009 to 2017 by Hansen et al. (2018). They note that women are often responsible for these areas in their parties and are therefore also assigned them as ministers when they are part of the government.

In figure 6, we show the frequency of speeches in each policy domain by female and male politicians relative to the number of speeches made by each group, respectively. According to figure 6, female politicians talk frequently about *Social Affairs*, *Immigration*, *Education* and *Justice*, while male politicians talk most often about *Economy*, *Immigration*, *Justice*, *Labor*, and *Infrastructure*. The figure also shows that *Immigration* and *Justice* are prominent topics in the speeches of both groups.

Figure 7 shows the policy domains that have frequently been discussed by female and male politicians during the two governments under the social-democrat Helle Thorning-Schmidt (2013-2014). She was the first female prime minister in Denmark. In her first government 34% of the ministers were female, while in her second government the percentage of female ministers was 47%. Under Helle Thorning-Schmidt's first government, the policy domain which is most frequently addressed by female politicians is *Education* followed by *Justice*, *Social Affairs*, and *Economy*. In the same period, the domains that are addressed by male politicians most frequently are *Economy*, *Justice*, and *Culture*. In Helle Thorning-Schmidt's second government, women speak frequently about *Justice*, *Social Affairs*, and *Education* while the domains most often addressed by men are *Economy*, *Labor* and *Justice*.

Figure 8 shows the frequency of speeches made by female and male politicians during the two govern-


Helle Thorning-Schmidt 1, 1/10 2013 -3/2-2014

Helle Thorning-Schmidt 2, 3/2 2014- 28-6/2014

Figure 7: The relative distribution of speeches by female and male politicians under the first and second government under Helle Thorning-Schmidt.

ments under Lars Løkke Rasmussen (2014-2019), who was, at the time, the leader of the Liberal Party. In the first of these periods, the percentage of female ministers was 29% while in the second period the percentage was 41%. Figure 8 shows that women under the first Lars Løkke Rasmussen government talk most frequently about *Immigration*, while the second most frequently addressed area is *Justice*. The high number of speeches on *Immigration* is a consequence of the immigration crisis in Europe in September 2015, as also discussed in Navarretta et al. (2022). Male politicians also speak more often about *Immigration* in this period than earlier, but the most frequently addressed speech by them is still *Economy*. In the third government under Lars Løkke Rasmussen, women more often talk about *Social Affairs*, *Health Care*, and *Immigration* than men, while the domain most often addressed by men is *Economy* followed by *Labor* and *Justice*. These are the same domains that were most often debated by male politicians under Helle Thorning-Schmidt's second government. The difference between the number of speeches about *Immigration* by women and men in this period is remarkable and this can be due to the fact that

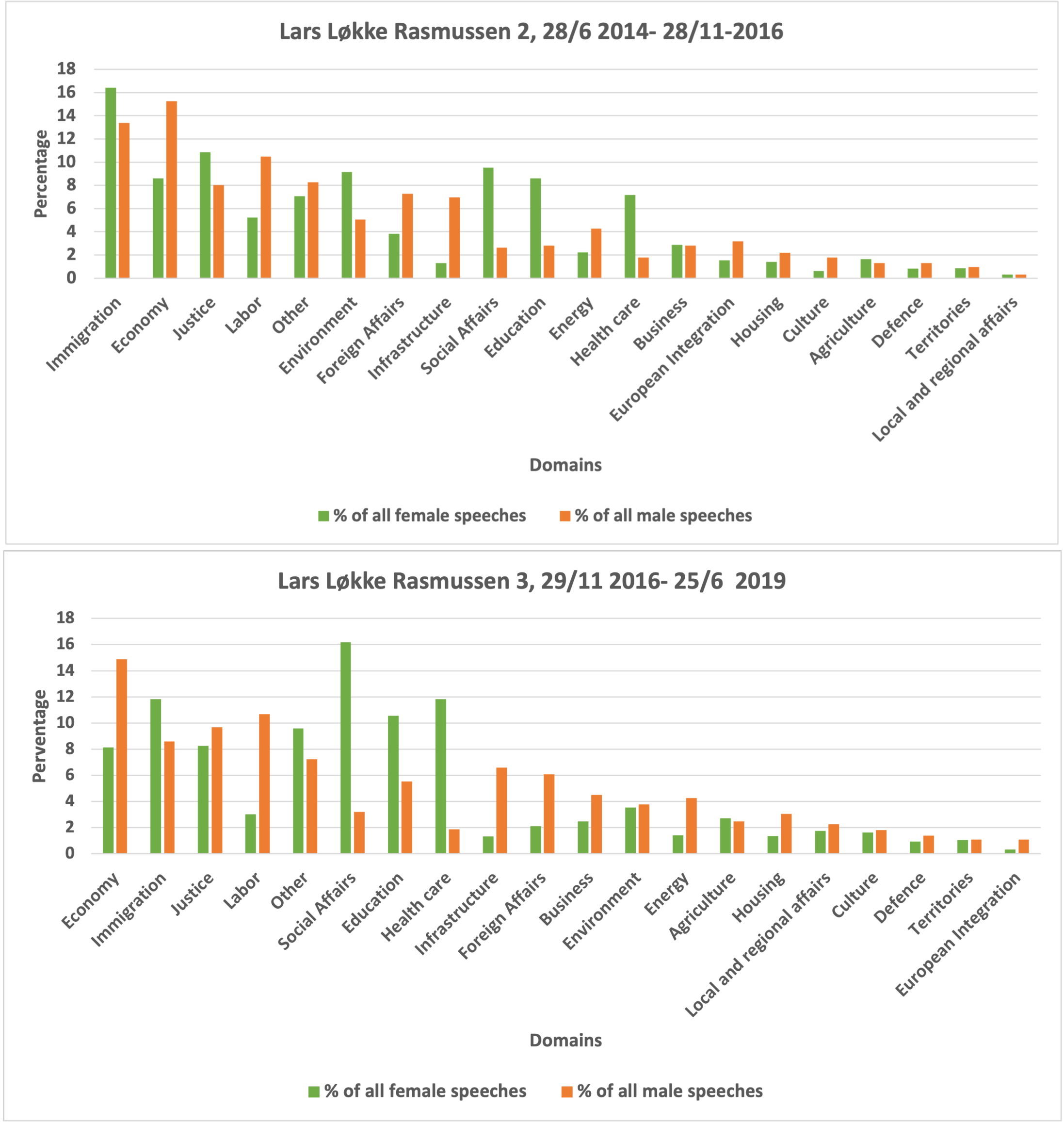


Figure 8: The distribution of speeches by female and male politicians during the second and third government under Lars Løkke Rasmussen.

the minister of immigration and integration in this period was a woman, Inger Støjberg.

In figure 9 the frequency of speeches by female and male politicians during the first government of the social democrat Mette Frederiksen (2019-2022) is shown. 29% of the ministers were women during this government. The policy domains most often addressed by female politicians between 2019 and 2022 are *Social Affairs*, *Education*, and *Immigration* while men most often spoke about *Immigration*, *Economy*, and *Justice*. Figure 9 clearly indicates that *Immigration* is a policy domain that continues to be a central topic in the debates. In this period, men talk more about *Health care* than women, and this can be due to the fact that the minister of health during the COVID crisis was a man, Magnus Heunicke.

The analysis of the frequency of speeches about various policy domains addressed by women and men in the Danish parliament shows that the frequency vary over time. Not surprisingly, it can be influenced by external events such as the immigration crisis in 2015 and the COVID-19 pandemic. In general, however, female politicians address ”soft areas” more often than their male colleagues. Another factor to

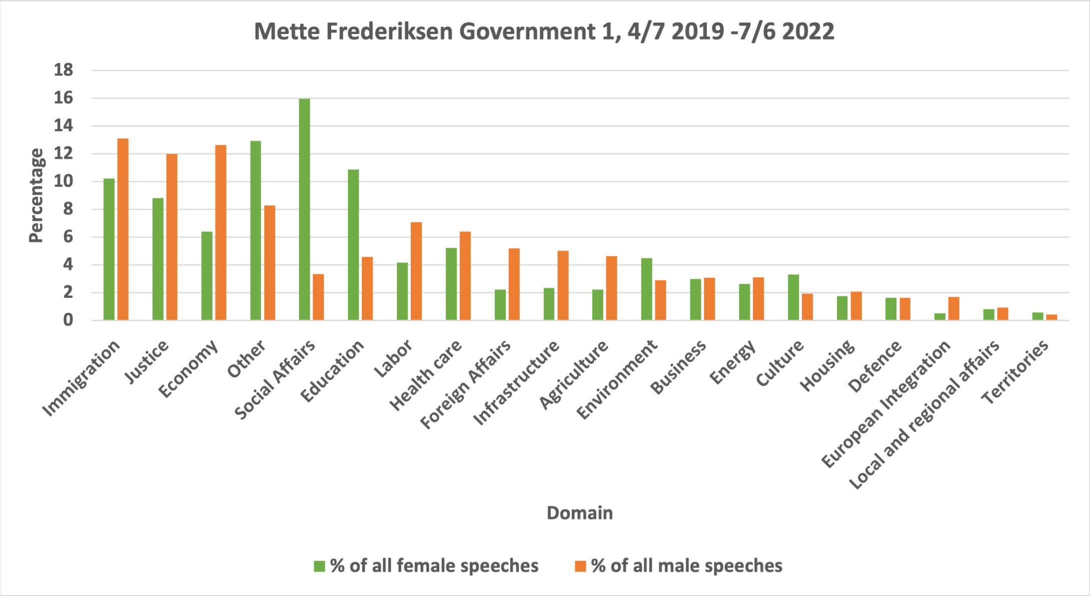


Figure 9: The relative distribution of speeches by female and male politicians during the first Mette Frederiksen's government.

take into consideration is the gender of the minister for a given domain since ministers often talk more than other politicians about that domain. It is also interesting that women often talked slightly more about *Environment* than men while men talked much more about *Energy*, even though the two domains are strongly related. The fact that female politicians often talk about ”soft” areas also reflects the job situation in Denmark since there are more women employed as social workers, nurses, and educators than men.

7 Conclusions and future work

In the paper, we have presented ParlaMint-DK 4.1 with policy domain annotations and improved lemma and NER annotations. We have also presented a quantitative analysis of the most frequently addressed policy domains in the corpus and of the most frequently co-occurring domains in it. Quantitative analyses of the most frequently addressed policy domains in the speeches made by female and male politicians under the five governments covered in the corpus are also presented. Our analyses confirm that Danish female politicians talk about ”soft” domains such as *Social Affairs*, *Education*, and *Health care* more often than their male colleagues do, and men talk about ”hard” areas such as *Economy* and *Labor* much more frequently than women. However, events such as the immigration crisis in 2015 and the COVID-19 pandemic from 2020 can change the relevance of specific policy areas. The gender distribution of the speeches on policy domains also reflects the gender of the ministers for those areas.

In the future, we will investigate how well the policy domain annotations can be used for automatically annotating new speeches, and to determine how different parties have dealt with the same domain over time.

References

Bäck, H., & Debus, M. (2019). When do women speak? a comparative analysis of the role of gender in legislative debates. *Political Studies*, *67*(3), 576–596.

Baumgartner, F. R., Jones, B. D., & Wilkerson, J. (2011). Comparative Studies of Policy Dynamics. *Comparative Political Studies*, *44*(8), 947–972. https://doi.org/10.1177/0010414011405160

Blumenau, J. (2021). The effects of female leadership on women's voice in political debate. *British Journal of Political Science*, *51*(2), 750–771.

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches — A comparative study of classifiability. *Literary and linguistic computing*, *27*(2), 139–153.

Erjavec, T., Kopp, M., Ljubešic, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, S., Bartolini, R., Bel, N., Pérez, M. C., Darģis, R., … Fišer, D. (2024). Parlamint ii: Advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*. https://doi.org/https://link.springer.com/article/10.1007/s10579-024-09798-w

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agerri, R., 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., … Fišer, D. (2024). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1911

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

Hansen, D. H., & Navarretta, C. (2021). The Danish Parliament Corpus 2009 - 2017, v2, w. subject annotation [CLARIN-DK-UCPH Centre Repository]. http://hdl.handle.net/20.500.12115/44

Hansen, D., Navarretta, C., & Offersgaard, L. (2018). A Pilot Gender Study of the Danish Parliament Corpus. *Proceedings of the ParlaClarin workshop at the Eleventh International Conference on Language Resources and Evaluation*, 67–72.

Hansen, D., Navarretta, C., Offersgaard, L., & Wedekind, J. (2019). Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus [https://cst.dk/DHN2019/DHN2019.html]. *CEUR Workshop Proceedings*, *2364*, 166–174.

Hargrave, L., & Blumenau, J. (2022). No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK. *British Journal of Political Science*, *52*(4), 1584–1601. https://doi.org/10.1017/S0007123421000648

Hargrave, L., & Langengen, T. (2021). The Gendered Debate: Do Men and Women Communicate Differently in the House of Commons? *Politics & Gender*, *17*(4), 580–606. https://doi.org/10.1017/S1743923X20000100

Ivanusch, C. (2024). Where do parties talk about what? party issue salience across communication channels. *West European Politics*, *0*(0), 1–27. https://doi.org/10.1080/01402382.2024.2322234

Jongejan, B. (2016). Implementation of a Workflow Management System for Non-Expert Users. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 101–108.

Jongejan, B., Hansen, D., & Navarretta, C. (2021). Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus. *CLARIN Annual Conference 2021 Proceedings*, 70–73.

Mandravickaitė, J., & Krilavičius, T. (2017, April). Stylometric analysis of parliamentary speeches: Gender dimension. In T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, & R. Yangarber (Eds.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (pp. 102–107). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1416

Merz, N., Regel, S., & Lewandowski, J. (2016a). The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. https://doi.org/10.1177/2053168016643346

Merz, N., Regel, S., & Lewandowski, J. (2016b). The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. https://doi.org/10.1177/2053168016643346

Navarretta, C., Haltrup Hansen, D., & Jongejan, B. (2022, June). Immigration in the manifestos and parliament speeches of Danish left and right wing parties between 2009 and 2020. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the LREC22 Workshop ParlaCLARIN III* (pp. 71–80). ELRA. https://aclanthology.org/2022.parlaclarin-1.11

Navarretta, C., & Hansen, D. H. (2023, September). According to BERTopic, what do Danish parties debate on when they address energy and environment? In C. Klamm, G. Lapesa, V. Gold, T. Gessler, & S. P. Ponzetto (Eds.), *Proceedings of the 3rd KONVENS Workshop on Computational Linguistics for the Political and Social Sciences* (pp. 59–68). Association for Computational Lingustics. https://aclanthology.org/2023.cpss-1.6

Navarretta, C., & Hansen, D. H. (2022). The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification. *Proceedings of LREC 2022*.

Paxton, P., Kunovich, S., & Hughes, M. M. (2007). Gender in Politics. *Annual Review of Sociology*, *33*(1), 263–284. https://doi.org/10.1146/annurev.soc.33.040406.131651

Ristilä, A., & Elo, K. (2023). Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling. *Parliaments, Estates and Representation*, 1–28.

Yu, H.-C., Rehbein, I., & Ponzetto, S. P. (2023). Policy domain prediction from party manifestos with adapters and knowledge enhanced transformers. *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, 229–244.

Zirn, C., Glavas, G., Nanni, F., Eichorst, J., & Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016)*, 88–93.

Managing Access to Language Resources in a Corpus Analysis Platform

Eliza Margaretha Illig
Department of Digital Linguistics
IDS Mannheim, Germany
margaretha@ids-mannheim.de

Nils Diewald
Department of Digital Linguistics
IDS Mannheim, Germany
diewald@ids-mannheim.de

Paweł Kamocki
Department of Digital Linguistics
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Marc Kupietz
Department of Digital Linguistics
IDS Mannheim, Germany
kupietz@ids-mannheim.de

Abstract

Corpus query tools are crucial to CLARIN's mission of facilitating the sharing and use of language data for research. It is a huge challenge for online corpus platforms to manage user access rights for large corpora with complex licenses and heterogeneous restrictions on access methods and purposes. This paper presents an approach to maximize user access to corpus data while protecting rights holders' legitimate interests. Query rewriting techniques and authorization procedures allow for modeling license terms in detail, enabling broader applications. This offers an alternative to methods that only model a greatest common denominator of licenses, thereby limiting the possibilities for using the data. Our approach constitutes a flexible and extensible corpus license and user rights management component applicable for other language research environments.

1 Introduction

The core value of CLARIN is to accommodate the reuse of language data and tools for research. In pursuing this objective, CLARIN, and linguistics in general, face the challenge that its research data are typically affected by the rights of third parties. One approach to dealing with this is to use technical measures that, on the one hand, ensure that the interests of the rights holders are not infringed and, on the other hand, restrict the use of the data as little as possible. This is typically done using an online corpus query system, which only allows indirect access to the data.

Provided that uniform licenses are available, corpus concordancers, for example, allow authenticated users who have agreed to their terms of use to view keywords in context (KWICs) without allowing full-text reconstruction. The situation becomes more difficult when – as is often unavoidable – different licenses and rights exist for different parts of the data and different groups of users, or when close reading of KWICs is not the only use case. Our paper presents the KorAP (Diewald et al., 2016) approach to making very large corpora, such as the German Reference Corpus DeReKo (Kupietz & Lüngen, 2014), which is affected by more than 200 partly heterogeneous licenses and is used in very different contexts, as usable as possible, while safeguarding the legitimate interests of rights holders.

1.1 Corpus Licenses

Providers of corpus analysis platforms typically do not hold any property rights in language resources which they allow the analysis of. They usually obtain the rights to use the resources via contractual arrangements with the rights holders, which can generally be referred to as licensing agreements. These agreements allow the provider to grant limited access to the resources to the platform users, or certain groups thereof.

Language corpora represent significant economic value (which has only increased with the advent of LLMs, and is bound to increase even further), and require substantial investment (of time, qualified effort and money) to build. In order to protect this value and investment from (perceived) 'free riders' (Olson, 1965) and potential competitors, rights holders often choose to restrict access by certain categories of users, or restrict certain uses (e.g., downloading *n*-grams).

Licensing agreements define by whom, for what purposes and how a resource can be used. They can be divided into those limited to academic uses, i.e. teaching and research, sometimes only in certain

Eliza Margaretha Illig, Nils Diewald, Pawel Kamocki and Marc Kupietz 2024. Managing Access to Language Resources in a Corpus Analysis Platform. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 100–111. DOI: https://doi.org/10.3384/ecp216

fields (such as linguistics, as opposed to journalism or media science) and those allowing commercial uses. ‘Academic’ licenses are generally less costly to obtain, but much more restrictive. For example, they may allow access only to users affiliated with a research organization, with an authenticated account (this also applies to automated access on behalf of the user). Further restrictions in both types of licenses may include access only via a dedicated platform or API, or even access only from a specific physical location or via a specific network.

Popular public licenses, such as Creative Commons (CC), do not discriminate between groups of users allowing resources to be available on online corpus analysis platforms without authentication. However, even CC licenses contain restrictions on re-use, ranging from simple acknowledgement of the source (BY), through the requirement to retain the original license in any modified versions (SA, or ‘share-alike’), to the prohibition of making any modifications (ND, or ‘no-derivatives’), to the prohibition of any commercial use (NC).

Statutory exceptions in applicable copyright legislation can also be interpreted as ‘licenses’ *sui generis*, i.e. permissions granted not by the rights holder, but directly by the legislator. The statutory exceptions for Text and Data Mining (TDM), harmonized at the EU level by the Digital Single Market (DSM) Directive (2019/790), deserve special attention. Article 3 of the DSM Directive allows research organisations to build corpora for TDM purposes, which (at least in certain EU jurisdictions such as Germany) can subsequently be shared with research partners.

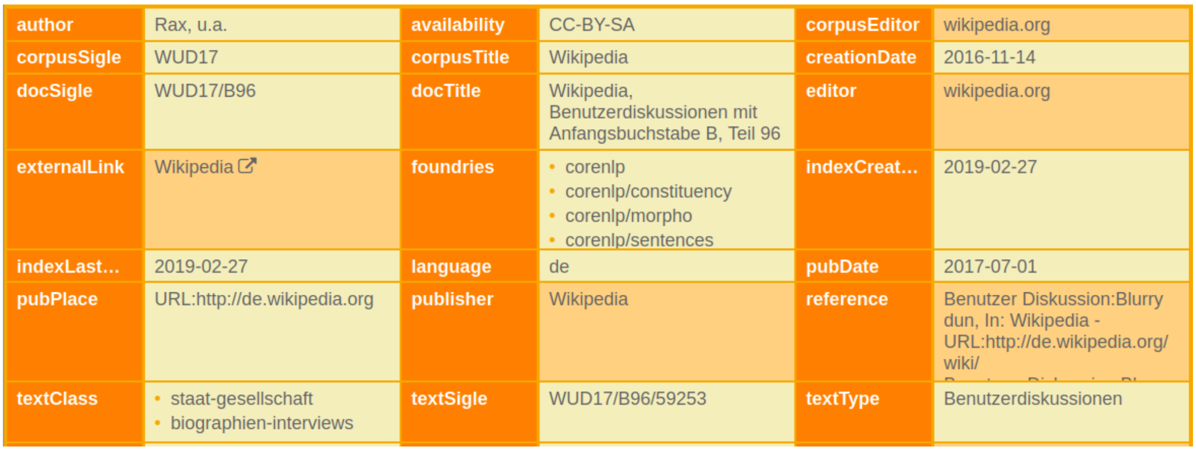
Certain metadata and annotations can be licensed. Even though in general metadata would rarely attract copyright protection, they may still be protected by the *sui generis* database right. Moreover, some metadata may contain elements such as abstracts which typically qualify for copyright protection. Since the perceived economic value of metadata is low, licenses for metadata tend to be more liberal (oftentimes, a waiver of rights such as CC0 is used). On the other hand, many rightholders are unaware of the possibility to license metadata separately, and hence the metadata of many resources are not accompanied with any licenses.

1.2 User Rights Management

Corpus licenses determine which users have which access rights to which parts of primary data, metadata, and annotation data (the latter being determined by software licenses as well). The rights of a user therefore have to be managed in addition by a corpus platform and can be matched with the licenses after authentication and before any data can be delivered to an account. As previously introduced, these licenses determine not only whether, but also in which ways a user can access the data. This requires that access rights must not only be compared statelessly and statically, but also take into account the temporal and local contexts. For example, if licenses only allow short excerpts from texts (e.g., KWICs), there is a reasonable concern of licensors that the original full-text can be reconstructed from the search or analysis results by cleverly formulating follow-up search queries.[1] In order of prohibiting such use, it may be necessary to monitor an account’s search queries over time to detect and/or prevent misuse.

In addition to static permissions of an account, further factors such as location and time can initiate dynamic restrictions. Some corpus licensors also limit the availability of their data to sites within a specific location or network. The user rights management of an online corpus platform must therefore be able to match the IP address of the account with the address space allowed for access. If licensors make licenses available to users only for a limited period of time, the user rights management system must log the initial access and check with each access whether the approved time frame has not yet been exceeded.

Users are permitted to utilize corpus data shared under TDM exceptions for text and data analysis in non-commercial research. To ensure this prerequisite is met, it is reasonable to implement a policy that restricts access to these resources on a request-only basis. In straightforward scenarios, for instance when the data is bound to a particular project, it is adequate and appropriate to host these data in a separate instance of a corpus system, ensuring that only users with approved requests can log in and access them.



author	Rax, u.a.	availability	CC-BY-SA	corpusEditor	wikipedia.org
corpusSigle	WUD17	corpusTitle	Wikipedia	creationDate	2016-11-14
docSigle	WUD17/B96	docTitle	Wikipedia, Benutzerdiskussionen mit Anfangsbuchstabe B, Teil 96	editor	wikipedia.org
externalLink	Wikipedia	foundries	• corenlp • corenlp/constituency • corenlp/morpho • corenlp/sentences	indexCreat...	2019-02-27
indexLast...	2019-02-27	language	de	pubDate	2017-07-01
pubPlace	URL:http://de.wikipedia.org	publisher	Wikipedia	reference	Benutzer Diskussion:Blurry dun, In: Wikipedia - URL:http://de.wikipedia.org/ wiki/
textClass	• staat-gesellschaft • biographien-interviews	textSigle	WUD17/B96/59253	textType	Benutzerdiskussionen

Figure 1: Metadata fields of a text in DeReKo can be used to create a virtual corpus in KorAP. The availability field is particularly used to enforce access policies for DeReKo.

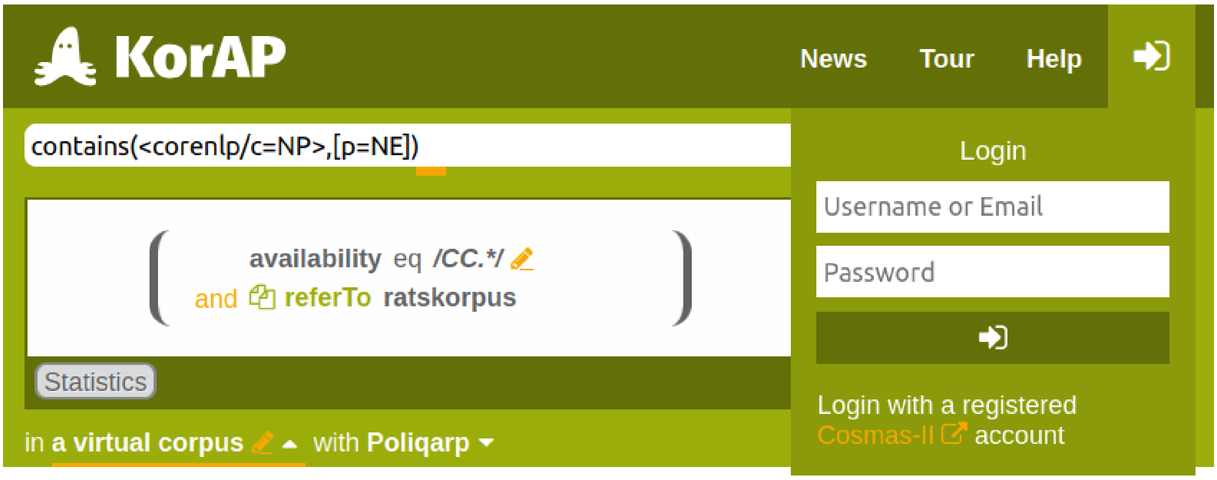
2 Related Work

Some corpus platforms allow users to build and work on their own corpora, for instance *Sketchengine* (Kilgarriff et al., 2014) provides corpus building tools to upload or find texts from the web. KorAP provides pre-defined virtual corpora, that are collections of texts dynamically assembled based on certain criteria such as a list of text identifiers (`textSigle`). Additionally, KorAP supports functionalities enabling users to create their own virtual corpora on the fly by filtering corpus metadata such as author name or publication year. Figure 1 presents some metadata fields of a DeReKo text. One or more virtual corpora can be used in a search by adding a reference to their identifiers, e.g. `referTo ratskorpus` illustrated in Figure 2.

In digital rights management, Open Digital Rights Language (ODRL; Iannella and Villata, 2018) is commonly used to represent policies on the use of digital content and services. It is used to define permissions, prohibitions, and obligations between parties (resource owners, users) and assets (resources), actions on assets, and constraints. It can be expressed in various serialization formats including XML and JSON-LD (Sporny et al., 2014), a lightweight Linked Data format that enables semantic definitions of the data. Comparable to an ODRL model, KorAP uses a JSON-LD based representation to describe user queries on virtual corpora (assets) on which the queries are applied, and access policies (contraints) on them. The constraints are dynamically adjusted by means of query rewriting according to authentication, authorization and access location (see Section 3). Authentication is the process of verifying a user's identity, while authorization happens after authentication to determine which resources and actions the user have access to. Furthermore, authorization scopes define access permissions that can be granted to an application. In our case, permissions on assets are not included in the representation but incorporated as authorization scopes. While ODRL focuses constraints on parties, assets or actions, our approach emphasizes constraints based on licenses required to access DeReKo.

In authentication and authorization management, Shibboleth (Cantor & Scavo, 2005) is a common identity and access management (IAM) system used in academic and research communities including CLARIN. It supports Single Sign-On (SSO) typically used to allow academic users to use their institutional logins to access corpora with academic licenses provided by corpus platforms such as OpenSoNaR (Reynaert et al., 2014). Besides, Lightweight Directory Access Protocol (LDAP; Howes and Smith, 2006) provides a central location to store information such as user details, in a hierarchical structure (directory), and defines a way to verify user credentials and determine user permissions to access systems or resources. While Shibboleth and LDAP provide foundational authentication mechanisms, they are inherently limited

[1] This is also a concern for corpus use in language models.



contains(<corenlp/c=NP>,[p=NE])

availability eq /CC.*/
and referTo ratskorpus

Login with a registered Cosmas-II account

Figure 2: The web UI Kalamar displays the virtual corpus menu with two criteria defining a virtual corpus: the *availability* metadata field for various CC license types, and a reference to a pre-defined virtual corpus named *ratskorpus*. A pen icon indicates that query rewriting has been performed. A login form is provided to facilitate user authentication.

to static and predefined user configurations. Thus, they do not cover all access control requirements in KorAP, particularly to support access control for third-party applications. In addition to LDAP for authentication, KorAP makes use of OAuth 2.0 for authorization (see Section 3.3). OAuth 2.0 (Hardt, 2012) is an authorization framework that allows users to grant specific permissions to third-party applications, for instance to access their data in KorAP and perform a search on their behalf, without requiring them to share their credentials.

By means of authorization, KorAP is able to provide numerous linguistics resources of DeReKo to third parties. It has been integrated into the CLARIN Federated Content Search (FCS) enabling access DeReKo. However, since FCS lacks support for an authorization mechanism, that is a prerequisite to access protected DeReKo resources (see Section 3.1), only a small amount of free resources are accessible. Nonetheless, KorAP permits unauthorized requests to search the metadata of all resources, including protected ones, and can report the number of matches found. If FCS were to support the display of such results, it would significantly enhance the user experience.

In access control management, Keycloak (Thorgersen & Silva, 2021) provides a comprehensive admin console, that allows a wide range of authentication and authorization management, for example, to customize an authentication flow and manage access based on user roles. It supports login with social networks that are not limited to institutions like Shibboleth, authorization using OAuth 2.0, and user federation enabling access to external user data stores such as LDAP. The BlackLab corpus search engine (de Does et al., 2017) has been carrying an ongoing development on authentication using Keycloak. Besides, Google Zanzibar (Pang et al., 2019) is a global authorization system to store and manage access controls across a vast range of applications. It replicates authorization data to rapidly determine user access and permissions on resources over hundreds of applications with various access control policies. While Keycloak and Google Zanzibar allow managing access based on user roles and groups, KorAP also requires access control based on licenses (see Section 3.1).

3 KorAP Approach

The challenge of a corpus license and user rights management system is to find technical solutions for mapping rights and licenses, and restricting data access accordingly. The system must be performant and adaptable to changes in rights and new license forms. To maintain flexibility and independence from

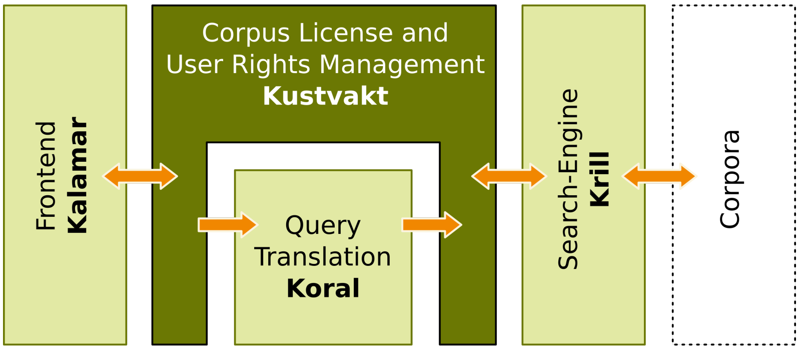


Figure 3: KorAP consists of multiple independent components. The corpus license and user rights management component Kustvakt is amiddleware broker service between API requests and the search engine.

underlying data and user interfaces, we develop a separate server-based component called Kustvakt (Illig et al., 2025)[2], that is responsible for the corpus license and user rights management in KorAP.

Figure 3 depicts the architecture of KorAP that consists of multiple independent small components. Kustvakt plays the role of middleware managing communications between the components and their tasks. It manages KorAP web service APIs that allow the web front-end Kalamar (Diewald et al., 2019) and other clients to communicate with KorAP. By administering user authentication and authorization processes, it verifies whether a request is authenticated or authorized, and applies access policies accordingly to return appropriate responses. Upon receiving (authorized) API requests, it uses Koral to translate user and corpus queries to *KoralQueries* (Bingel & Diewald, 2015). Subsequently, Kustvakt performs query rewriting (Bański et al., 2014) on KoralQueries respecting corpus license and user rights, forwards them to the search engine Krill (Diewald & Margaretha, 2017) and returns the responses to the requesting entity.

3.1 Access policies

IDS has been concluding license agreements for the use of texts for linguistics for around 50 years. Over this period, unavoidable variations have arisen beyond the classes of license conditions, mentioned above. It is typically too expensive and too risky or impossible to reopen any of the legacy agreements. This is the typical situation of institutions that offer large corpora.

To ensure compliance with licensing agreements, we have to deal with the most significant access limitations. Firstly, a great number of DeReKo resources are only available for academic or research purposes, whereas commercial use is explicitly prohibited. Therefore, users must register and agree to our terms of use to access all of DeReKo resources through KorAP. User data is stored and managed through a centralized LDAP for multiple services at IDS Mannheim including KorAP. Secondly, access to DeReKo must be restricted through a query system, such as KorAP, particularly designed to prevent the downloading and reconstruction of original texts from search results. In KorAP, we impose a limit to the size of the match context, and employ a timeout mechanism to restrict search duration thereby enhancing system responsiveness. Thirdly, users must authenticate to access protected resources.

To address all these limitations, we use the six license categories: CC (Creative Commons) variants, ACA-NC (academic, non-commercial), ACA-NC-LC (license contract also required), QAO-NC (query-analysis-only, non-commercial), QAO-NC-LOC:ids (only accessible through IDS network), QAO-NC-LOC:ids-NU:1 (only one user at a time), that are described in detail in Kupietz and Lüngen (2014). Each text of DeReKo is annotated with an `availability` metadata field representing its license category. Figure 1 presents some metadata fields of a Wikipedia text in DeReKo including the availability field with license category CC-BY-SA.

Based on license categories, login and network location, we define three types of corpus access policies

[2] https://github.com/KorAP/Kustvakt

Corpus Access	Regex Patterns of License Categories	Login Required	Access Location
Free	CC.*	no	anywhere
Public	CC.*, ACA.*, QAO-NC	yes	anywhere
All	CC.*, ACA.*, QAO.*	yes	IDS

Table 1: To comply with DeReKo license agreements, three types of access policies are defined in KorAP based on license categories, login and access location.

for KorAP: 1. *Free* access on corpora under CC licenses, that are accessible from anywhere without login, e.g., Wikipedia; 2. *Public* access on free and academic corpora, that requires login; 3. *All* corpora access, that requires login and access through IDS network or Virtual Private Network (VPN) providing a secure connection to the IDS network. By login, we mean not only user authentication but also authorization given to a third-party application (see Section 3.3). We use IP address ranges to determine the access location of requests. Table 1 summarizes the access policies and describes the regular expression patterns of license categories corresponding to each corpus access policy. Kustvakt determines and grants access to a request according to these policies.

Corpus Access	Rewrite Rules
Free	availability = CC.*
Public	availability = CC.* | ACA.* | QAO-NC
All	availability = CC.* | ACA.* | QAO.*

Table 2: Rewrite rules for corpus access are defined using *availability* metadata and regular expression patterns of license categories.

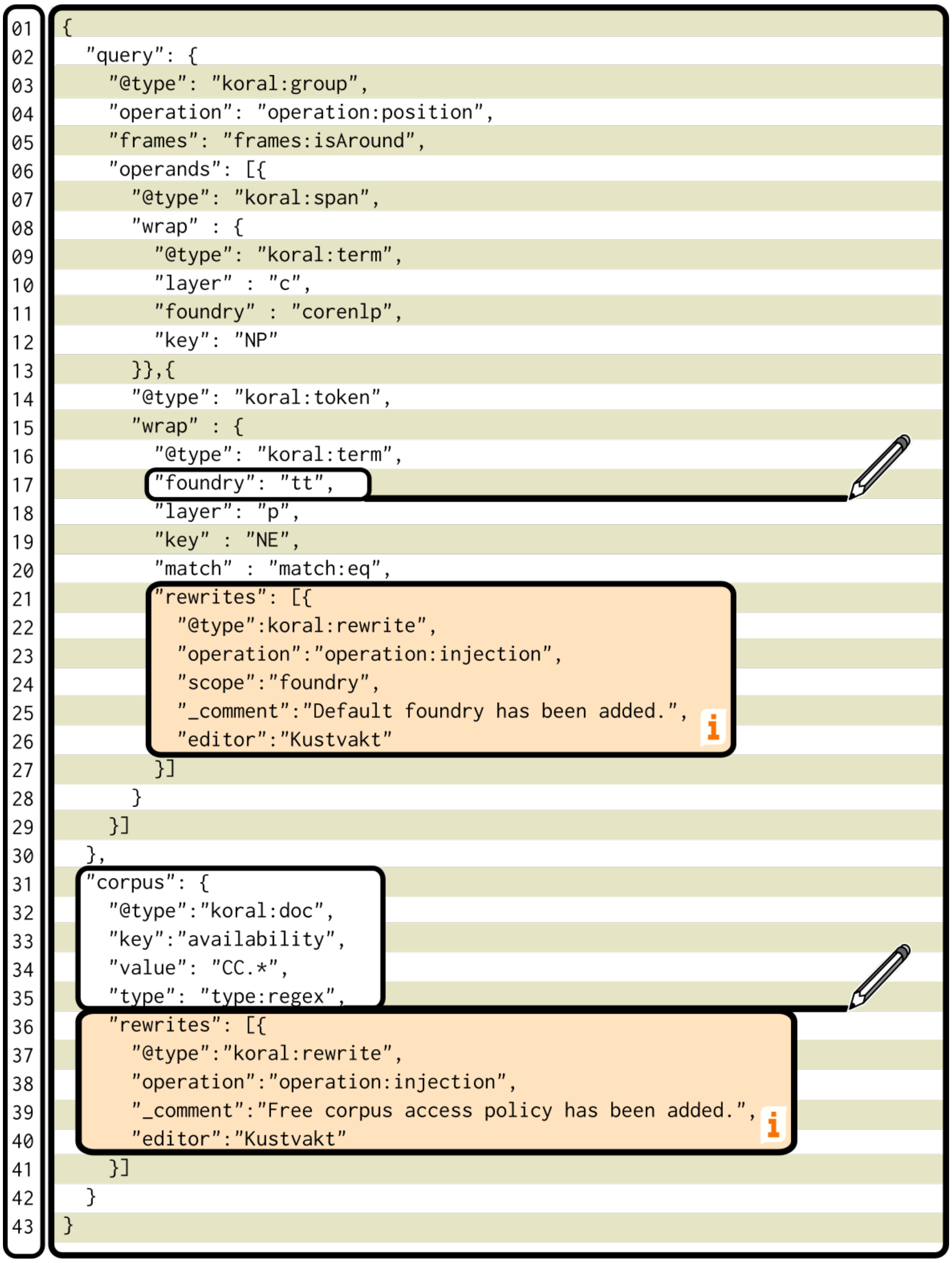
The access policies are enforced through query rewriting described in the following section. Table 2 presents the rewrite rules derived from the availability field and the regular expression patterns of the license categories. The rewrite rules define a virtual corpus (a subset of DeReKo) accessible under each specific access policy. After determining the appropriate access for a request, Kustvakt dynamically modifies the user query according to the rewrite rules. A search request from an non-authenticated user, for example, would be granted free access and executed on a virtual corpus containing all texts whose availability field specifies a CC license category variant. Figure 4 illustrates the search request and the free access granted through the virtual corpus definition at lines 31-40.

The access policies are also applied to pre-defined virtual corpora, as well as those created by users on the fly. The size of the virtual corpora accessible on request may vary according to login and access location constraints. However, metadata of all corpora is freely available regardless of access restrictions on corpus content.

3.2 Query Rewrites

To manage access to a resource in terms of both licenses and user rights while granting the user the greatest possible amount of liberty, an approach based on *query rewriting* was chosen. In this approach, a resource request (see Fig. 5) is reformulated via a central component (Kustvakt) to correspond to the access rights of the requesting entity and can be answered by the database without further knowledge of licenses and user rights (cf. Rizvi et al., 2004).

To achieve this, restrictions in the form of metadata constraints are encoded at the individual text level and can thus be excluded directly during a search or analysis on the corpora. In principle, it is possible to take any metadata into account in the rewrite process, for example the identification of a license (as in Figure 4 via the metadata field `availability`, line 33), but also corpus labels or author names. Additional rights are added to the query as additional constraints; in this sense, the approach is fundamentally



```
{
  "query": {
    "@type": "koral:group",
    "operation": "operation:position",
    "frames": "frames:isAround",
    "operands": [{
      "@type": "koral:span",
      "wrap" : {
        "@type": "koral:term",
        "layer" : "c",
        "foundry" : "corenlp",
        "key": "NP"
      }},{
      "@type": "koral:token",
      "wrap" : {
        "@type": "koral:term",
        "foundry": "tt",
        "layer": "p",
        "key" : "NE",
        "match" : "match:eq",
        "rewrites": [{
          "@type":koral:rewrite",
          "operation":"operation:injection",
          "scope":"foundry",
          "_comment":"Default foundry has been added.",
          "editor":"Kustvakt"
        }]
      }
    }]
  },
  "corpus": {
    "@type":"koral:doc",
    "key":"availability",
    "value": "CC.*",
    "type": "type:regex",
    "rewrites": [{
      "@type":"koral:rewrite",
      "operation":"operation:injection",
      "_comment":"Free corpus access policy has been added.",
      "editor":"Kustvakt"
    }]
  }
}
```

Figure 4: The corpus query ‘Return all nominal phrases NP annotated in the corenlp foundry and contain a named entity NE’ is rewritten by Kustvakt to use a default foundry annotation tt for the part-of-speech layer p and to restrict access to free corpora licensed under Creative Commons.

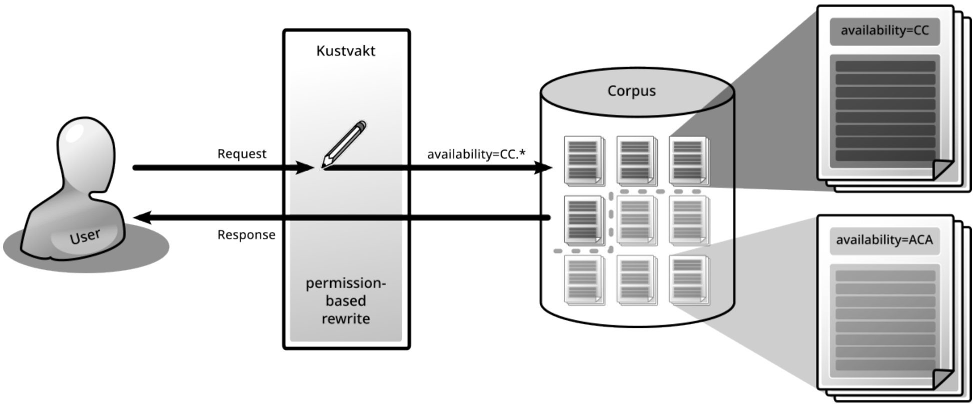


Figure 5: Resource request flow of an authenticated user: The request is rewritten to inject a constraint regarding the metadata field `availability` based on the user's access permissions.

“additive”.

The basis for this query rewriting is KoralQuery (Bingel & Diewald, 2015), an implementation of CQLF (Bański et al., 2016) in form of a unified JSON-LD-based (Sporny et al., 2014) representation of an abstract corpus query. It is comparable to SQL for relational data queries (cf. Figure 4) and independent of any corpus query language such as CQP, Annis-QL, Poliqarp, and Cosmas-2-QL supported by KorAP. By adding or changing constraints, a new corpus query can be formulated that satisfies all requirements, and can be passed to a database. Figure 4, lines 31–35, illustrates a constraint that restrict a virtual corpus to all texts with a Creative Commons license.

In addition to restrictions on text access, it is also possible to exclude query options. For example, rules can be formulated to exclude the search in certain annotations or set defaults for queries on annotations. Figure 4, line 17, describes a constraint restricting the search of part-of-speech *NE* to the default source of annotation data (called *foundry*) *tt*, that is the TreeTagger annotation.

Any modification to the query is marked (see Figure 4, lines 21–26 and 36–40), and can be used by clients to inform users or to reconstruct the original query. Figure 2 presents an example where the web UI Kalamar displays a pen icon next to the virtual corpus menu, indicating query rewriting that has been performed on the corpus query level. This may be necessary, as it is the only way to ensure transparency and provide users with feedback on requested resources they actually have access to.

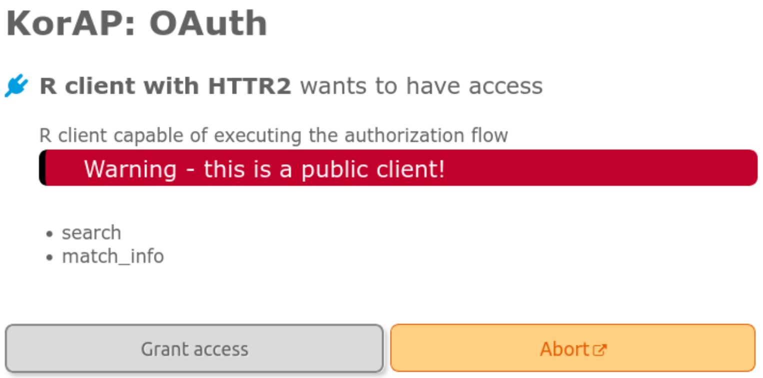
Query rewriting is independent of the corpus and the user size, therefore it scales and performs well with a growing database. It is only dependent on the different restrictions that need to be lifted in the case of permissions granted to the users.

3.3 Authorization

An API enables client applications to communicate with a server-based corpus platform like KorAP allowing technically skilled users to integrate web-services supported by the platform into their own applications. For instance, using RKorapClient (Kupietz et al., 2020) library, users can send search and annotation requests from R to Kustvakt, which is the API provider of KorAP, and then extract and visualize the results in R. In terms of reusability and scalability, API allows Kustvakt to be easily set up for other research environments, extended to meet specific needs, and combined with other front-ends.

As described in the previous section, only corpora with free licenses and metadata in DeReKo are accessible without user authentication. To access licensed corpora via the KorAP API, third party applications must obtain authorization, that is permission granted by users to act on their behalf. Note that location-based access restrictions still apply.

KorAP supports the authorization framework OAuth 2.0 that defines communication protocols with client applications to grant them authorizations in forms of access tokens. Access tokens are bound to



KorAP: OAuth

R client with HTTR2 wants to have access

R client capable of executing the authorization flow

Warning - this is a public client!

- search
- match_info

Figure 6: Kalamar displays an authorization request originating from an R client using httr2, that requests the *search* and *match_info* scopes to be able to perform search and annotation requests on behalf of the authenticated user. The user is given the option to either grant or decline access.

particular authorization scopes specifying to what extent these applications may act on user behalf or access user data. Kustvakt acts as an authorization server, that manages communications to clients via API and issues and manages access tokens (Kupietz et al., 2022). In addition to that, Kalamar acts as a front-end to the API, that provides web user interface for user authentication depicted in Figure 2.

To obtain access tokens, we support authorization code grant flow suitable for server-based client applications. This flow involves sending and processing HTTP requests multiple times to enhance security ensuring that access tokens are not leaked to intermediate user-agents such as browsers, but directly sent to clients. When a user would like to use a client to perform a search and acquire match annotations in KorAP, the client sends an authorization request including the necessary scopes (*search* and *match_info* in this example) to the KorAP authorization server. If the user is not authenticated to KorAP yet, it would ask him/her to login, and then to grant access to the client to perform search and retrieve annotation requests on behalf of him/her (see Figure 6). When granted, the KorAP authorization server would redirect the user to the client redirect URI including an authorization code. The client can subsequently exchange the authorization code with an access token by sending a token request.

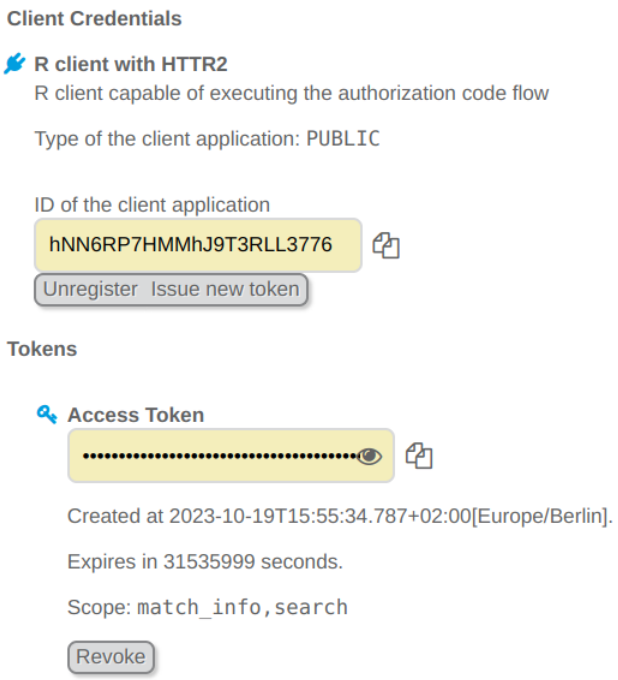
For non-server-based clients such as desktop applications, that are incapable of handling HTTP requests, we provide a feature to obtain access tokens from the web UI Kalamar as shown in Figure 7. Alternatively, local web-servers or libraries can be utilized to facilitate the authorization code flow, for example RKorAPClient uses httr2 (Wickham, 2023) that also supports OAuth 2.0. To enable this workflow, the KorAP authorization server permits localhost as a client redirect URI.

Client Type	Can store secrets	Access token Validity	Refresh Token Validity
Confidential	Yes	Short-lived	Long-lived
Public	No	Long-lived	Not Available

Table 3: To reduce the risk of token compromise, the time validity of access tokens and the issuance of refresh tokens are adjusted depending on the ability of clients to securely store secrets.

To protect users from potential authorization abuse by malicious software, we implement the following measures. Firstly, client registration is required to use the KorAP authorization APIs. Figure 7 displays a screenshot of the web UI Kalamar illustrating a registered OAuth2 client in KorAP. Kalamar provides details about the client including its type, identifier, and active access tokens.

Secondly, the time validity of access tokens is deliberately limited depending on the ability of clients to keep credentials. It is crucial to limit the time validity of access tokens to reduce the period of time



Client Credentials

R client with HTTR2

R client capable of executing the authorization code flow

Type of the client application: PUBLIC

ID of the client application

hNN6RP7HMMhJ9T3RLL3776

Tokens

Access Token

Created at 2023-10-19T15:55:34.787+02:00[Europe/Berlin].

Expires in 31535999 seconds.

Scope: match_info,search

Figure 7: Kalamar provides a web user interface for managing OAuth clients and tokens. It provides client details, supports client unregistration, and shows access tokens info including expiry and authorization scopes. It also allows users to issue access tokens for public clients, particularly for non server-based client, and to revoke existing tokens.

in which a stolen access token can be exploited by an attacker. Thus, issuing short-lived access tokens is a reasonable and necessary strategy to minimize the potential impact of token theft. Since new access tokens can be issued using a special token, called refresh token, the security implications are even more significant when a refresh token is compromised. The client's ability to securely store secrets also plays a critical role in reducing the risk of token theft. According to the OAuth 2.0 specification, clients are categorized into 2 types: *confidential clients*, that can store secrets and authenticate securely, and *public clients*, that cannot. Confidential clients are issued short-lived access tokens along with a long-lived refresh token, that allows them to obtain new access tokens without requiring re-authorization. On the other hand, since public clients are more vulnerable to token compromise, they are not provided with refresh tokens. They are issued long-lived access tokens because they cannot obtain new access tokens without a refresh token. This distinction is summarized in Table 3.

Thirdly, KorAP supports token revocation (Lodderstedt & Scurtescu, 2013) to invalidate tokens before they expire. It is especially beneficial when users suspect unauthorized or malicious use of their access tokens. This functionality is available via both the API and the web UI Kalamar depicted in Figure 7.

4 Extensibility

KorAP has already covered all the access rights requirements for DeReKo. Some extensions can be profitable as follows.

Metadata is generally freely available regardless of access restrictions on corpus content. However, when certain metadata fields are restricted due to licensing terms, access to those fields can also be limited, similar to the restrictions on corpus content. Moreover, in exceptional cases, metadata such as titles (e.g., newspaper headlines) may reveal information about natural persons, which may also justify restrictions on access.

In addition to policy enforcement, the protocol-based approach enables the integration of other query rewriting methods, such as query expansion (cf. Baeza-Yates & Ribeiro-Neto, 2010, ch. 5) independent of the user and corpus base. This approach allows applying cascading rewrites to a query with policy enforcement at the end to prevent unintended expansion of permissions.

Following query rewriting, *response rewriting* can also be performed. In this case, the result set from the search engine is filtered according to certain criteria before it is returned to the account. However, since response rewriting usually requires more data to be requested from the resource than can finally be processed, this variant of rewriting is only suitable for small result sets for performance reasons. For example, it is well-suited for the individual shortening of text snippets that are displayed to the account (when certain text licenses allow longer contexts than others). Response rewriting is currently being implemented to enrich data results with external information (specifically mappings for universal dependency annotations). There are currently no efforts and no need to use this mechanism for the access-based filtering of results, which is why we do not discuss it further in this article.

Using Shibboleth, CLARIN enables distributed access to protected resources through SSO across organizational boundaries. However, this alone is not sufficient to access DeReKo's protected resources, as users must also explicitly agree to our terms of use, as described in Section 3.1. Shibboleth can be implemented in Kustvakt as an alternative to LDAP for authentication, particularly when serving academic corpora that do not require the same user agreement as DeReKo. Since KorAP is independent of specific resources, it can operate as a standalone instance serving a wide range of corpora beyond DeReKo.

5 Conclusion

Directly integrating policy enforcement at the protocol level through query rewriting and abstract authorization mechanisms allows for a great deal of transparency and flexibility for efficient and detailed access control to corpus resources with complex licenses. Our approach facilitates maximum access and usage of corpora while ensuring compliance with complex licenses. The currently applied rule set in our implementation is based on the needs of the different licenses of DeReKo, so the full flexibility is not yet exhausted. Our query rewriting approach is developed as programming APIs allowing easy integration of new rules for other applications. Simple new rewrite rules can be introduced by minor changes to the configuration, while more complex rules can be added as filters (currently between 30 and 200 lines of code). The largest application using our approach is currently a corpus query system that serves a corpus of 87 million texts for an average of 6000 queries per day. Kustvakt is open source and in conjunction with KoralQuery universally applicable for resource control in corpus analysis applications.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (2010). *Modern Information Retrieval: The Concepts and Technologies behind Search* (2nd ed.). Addison-Wesley.

Bański, P., Diewald, N., Hanl, M., Kupietz, M., & Witt, A. (2014). Access control by query rewriting: The case of KorAP. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 3817–3822. http://www.lrec-conf.org/proceedings/lrec2014/pdf/743_Paper.pdf

Bański, P., Frick, E., & Witt, A. (2016). Corpus Query Lingua Franca (CQLF). *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 2804–2809.

Bingel, J., & Diewald, N. (2015). KoralQuery - a General Corpus Query Protocol. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*.

Cantor, S., & Scavo, T. (2005). Shibboleth architecture. *Protocols and Profiles*, *10*(16), 29.

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating Research Environments with BlackLab. *CLARIN in the Low Countries*, 245–257.

Diewald, N., Barbu Mititelu, V., & Kupietz, M. (2019). The KorAP user interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique. On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, *64*(3), 265–277. http://www.lingv.ro/images/RRL%5C%203%5C%202019%5C%2006-%5C%20Diewald.pdf

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., & Witt, A. (2016). KorAP Architecture - Diving in the Deep Sea of Corpus Data. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 3586–3591.

Diewald, N., & Margaretha, E. (2017). Krill: KorAP search and analysis engine (M. Kupietz & A. Geyken, Eds.). *Journal for language technology and computational linguistics (JLCL)*, *31*(1), 73–90. http://www.jlcl.org/2016_Heft1/Heft1-2016.pdf

Hardt, D. (2012, October). The OAuth 2.0 Authorization Framework. https://doi.org/10.17487/RFC6749

Howes, T., & Smith, M. C. (2006, June). Lightweight Directory Access Protocol (LDAP): Uniform Resource Locator. https://doi.org/10.17487/RFC4516

Iannella, R., & Villata, S. (2018). ODRL Information Model 2.2. *W3C Recommendation*, *15*.

Illig, E. M., Diewald, N., Kupietz, M., Hanl, M., & Bodmer, F. (2025, March). *Kustvakt* (Version 0.76). Zenodo. https://doi.org/10.5281/zenodo.15044768

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). JSON-LD 1.0. A JSON-based Serialization for Linked Data. http://www.w3.org/TR/json-ld/

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on [Communicated by Yukio Tono.]. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kupietz, M., Diewald, N., & Margaretha, E. (2020). RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC '20)*, *12*, 7015–7021.

Kupietz, M., Diewald, N., & Margaretha, E. (2022). Building paths to corpus data. A multi-level least effort and maximum return approach. In D. Fišer & A. Witt (Eds.), *CLARIN. The Infrastructure for language resources* (pp. 163–189). de Gruyter. https://doi.org/10.1515/9783110767377-007

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. *Proceedings of the 9th conference on international language resources and evaluation (LREC'14)*, 2385.

Lodderstedt, T., & Scurtescu, M. (2013, August). *OAuth 2.0 Token Revocation* (RFC No. 7009). RFC Editor. https://tools.ietf.org/html/rfc7009

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press.

Pang, R., Caceres, R., Burrows, M., Chen, Z., Dave, P., Germer, N., Golynski, A., Graney, K., Kang, N., Kissner, L., Korn, J. L., Parmar, A., Richards, C. D., & Wang, M. (2019). Zanzibar: Google's Consistent, Global Authorization System. *2019 USENIX Annual Technical Conference*.

Reynaert, M., van de Camp, M., & van Zaanen, M. (2014). OpenSoNaR: User-Driven Development of the SoNaR Corpus Interfaces. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 124–128.

Rizvi, S., Mendelzon, A., Sudarshan, S., & Roy, P. (2004). Extending Query Rewriting Techniques for Fine-Grained Access Control. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 551–562. https://doi.org/10.1145/1007568.1007631

Thorgersen, S., & Silva, P. I. (2021). *Keycloak-identity and access management for modern applications: harness the power of Keycloak, OpenID Connect, and OAuth 2.0 protocols to secure applications*. Packt Publishing Ltd.

Wickham, H. (2023). *httr2: Perform HTTP Requests and Process the Responses* [https://httr2.r-lib.org, https://github.com/r-lib/httr2].

Choosing the Right Tool for You: Informed Evaluation of Text Analysis Tools

Angel Daza
Netherlands eScience Center
`j.daza@esciencecenter.nl`

Antske Fokkens
Vrije Universiteit Amsterdam
`antske.fokkens@vu.nl`

Abstract

Natural Language Processing (NLP) research showcases many promising tools and methods for text analysis. Researchers from diverse fields who want to use NLP for their research are confronted with a wide availability of ready-to-use models that claim excellent performance on standard benchmarks. Consequently, choosing an appropriate tool has become a task on its own. Our goal is to exemplify a methodology that stimulates critical evaluation and detailed analysis of automatic outputs of NLP tools. Particularly, we analyze the case of choosing the best Named Entity Recognition (NER) tool for a corpus of Dutch biographies. Our use case is an example of how to make informed decisions by considering different aspects of custom datasets at the instance and aggregated levels, improving the outcomes of the original research question.

1 Introduction

Recent developments in NLP have seen great progress towards one of CLARIN's primary goals: providing (relatively) easy-to-use language technology. This has led to an increase in the use of these technologies in various domains, such as Digital Humanities (DH) (Colavizza et al., 2021; Ehrmann et al., 2023; Schweter et al., 2022). However, the same rapid advancement of NLP has left the area with a weak spot regarding detailed and careful evaluation. Standard benchmarks and metrics often do not provide insight at the right level of detail for users to establish which tool works best for their specific use case, or whether a tool is appropriate for their methodological set-up at all (Fokkens et al., 2014). In our view, choosing an appropriate tool has become a task on its own, and supporting users in this task is an essential part of CLARIN's mission of making tools available to researchers. Therefore, in this paper we exemplify a methodology that stimulates critical evaluation and detailed analysis of automatic outputs of NLP tools, which can help external researchers select the right tool for their specific needs. Specifically, we propose a visually aided assessment of the strengths and weaknesses of using different models for a specific use case.

Although the methods proposed here can be applied directly to any span-based NLP tasks, for clarity, we describe a specific Digital Humanities application of a classic NLP task. Specifically, we assume the perspective of a historian who wants to perform NER to build networks of people (`PER`), organizations (`ORG`), and places (`LOC`) from digitized biographies. To achieve this, they must first select an NER model for their corpus, a dataset of thousands of biographies written in Dutch between the 18th and the 21st centuries. We propose combining a *distant evaluation* (global metrics) with a *close evaluation* of NER outputs (instance-level inspection) to make more informed choices when selecting a tool and to gain insight into how such tools will behave when we apply them to our data. This way, the user can ensure they behave as desired on the instances that matter the most for answering the original research question.

In the remainder of the paper, we start by briefly describing related work on NER classifiers and evaluation (Section 2), then we describe the dataset that we use to illustrate why careful evaluation is needed when picking a tool for a custom use case (Section 3.1). Next, we list four different state-of-the-art NER systems for Dutch (ready to be used out of the box) that we apply to our data to compare their

Angel Daza and Antske Fokkens 2024. Choosing the Right Tool for You: Informed Evaluation of Text Analysis Tools. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 112–126. DOI: https://doi.org/10.3384/ecp216

performance (Section 3.2). In Section 4 we describe metrics that allow us to visually explore each of the models' predictions on specific documents, spotting edge cases where major disagreement exists. This analysis can be performed even if we do not have gold data available, which is often the case for non-NLP practitioners. Section 5 then dives into a scenario where we do have a human-labeled dataset; however, in this case, too, different ways of evaluating (on the same dataset) can lead to different outcomes, showing the need for grounding the model selection in the research goals that have been established before the start of the technical implementation. Finally, Section 6 discusses our main findings, and Section 7 shows some of the limitations of the approach that we took here.

2 Related Work

The techniques for identifying named entities in texts have evolved with the general state of NLP and have become less interpretable with time. Initially, gazetteers were used to identify entity occurrences in a text directly. Later, statistical learning and Machine Learning classifiers were used to identify sequences of labels as a token classification task with some loose notion of what instances each entity category should comprehend (Nadeau & Sekine, 2007). After that, neural networks offered the option of including sentence context to predict a label for each token (Lample et al., 2016), using LSTMs with a classifier on top (Panchendrarajan & Amaresan, 2018) or using wider contexts by fine-tuning pre-trained language models such as BERT (Devlin et al., 2019) to perform structured prediction (Akbik, Bergmann, & Vollgraf, 2019; Yamada et al., 2020), this includes the classifiers used by Stanza, Flair and XLM-R presented in this paper (for more examples see Yadav and Bethard, 2018). Lastly, with the advent of LLMs, NER has been re-framed to use prompt engineering where a model such as GPT-x could be queried to generate as a response the series of entities found in a text, such as Ashok and Lipton, 2023 and Wang et al., 2023. In this paper, we also test GPT3.5's capabilities for spotting entities and trying to provide the spans where they were initially found. We argue that the fact that NER is being constantly re-framed to work with different techniques, such as prompt engineering, calls for a closer look at how evaluation is done, since comparison across models keeps getting more difficult given the variety of conceivable experimental setups. This constant evolution of LLMs related to NER can be seen in even more recent publications such as (Jiang et al., 2024; Kim et al., 2024; Picco et al., 2023; Tong et al., 2025).

Recently, more attention has been paid to expanding NER into other specific domains (De Los Reyes et al., 2021; Li et al., 2022). This includes a wide variety of annotated datasets to train task-specific classifiers (Arnoult et al., 2021), pre-trained language models with historic corpora (Manjavacas Arevalo & Fonteyn, 2022; Schweter et al., 2022) and general techniques to identify entities beyond the standard core labels and resources (Luthra et al., 2023; Tedeschi & Navigli, 2022).

There is also important work on evaluation for NER beyond just applying scores. For example, Ushio and Camacho-Collados (2021) show a tool for directly comparing different flavors of transformer models fine-tuned in several well-known benchmark datasets. Closer to our approach, Fu et al. (2020) present a tool that shows a behavioral overview of model outputs by bucketing entities according to their attributes. Their tool generates an HTML report that helps users interpret at the corpus level where the NER models fail the most. In contrast, our approach focuses on instance-level inspection of errors, where a researcher can explore errors based on edge cases or instances of particular interest.

3 Methodology

3.1 Case Study: NER for Dutch Biographies

The Biography Portal of the Netherlands[1] (BPN) is an online collection containing several biographical dictionaries written in the country through the years. It includes 25 different existing collections with biographical information on the inhabitants of the Netherlands, with more than 75,000 biographical entries. We are interested in applying NER to the BPN texts, which comprise biographies written between the 18th and 21st centuries; thus, they can be partly seen as a type of historical text (Romary, 2014). This domain poses its own set of challenges, such as: Language variety and dynamic changes through the cen-

[1] http://www.biografischportaal.nl

turies, a mixture of record typologies (each collection contains a specific style and idiosyncrasies when describing people), significant divergence in biography length, a prominent presence of abbreviations and rare entities that do not necessarily exist nowadays, to mention a few.

We created a human-labeled dataset containing a subset of 346 biographies of various lengths (some biographies have only dozens of tokens, and others have thousands). This subset was generated by stratified sampling to keep the original dataset's source distribution, ensuring we have examples from each collection. Annotators were asked to follow the guidelines for labeling three core entities: Persons (PER), Locations (LOC), and Organizations (ORG), as well as Dates (TIME), works of art (WOA), and miscellaneous (MISC). For this paper, we will focus only on the three core entities. Of the 346 biographies, 50 were triply annotated, for which we obtained an inter-annotator agreement of 78.3 Krippendorff's Alpha,[2] had a round of discussions, and proceeded to annotate the rest with a partial overlap to ensure agreement remained high. We also asked annotators to manually correct tokenization errors and re-split or merge sentences when necessary to end up with a clean pre-tokenized dataset in conll format, as is customary for standard NER evaluation. We describe the annotated subset in Table 1, where we also include the mean, median, standard deviation and maximum of entities annotated in the documents to showcase the diversity of the documents.

Category	Count	Mean	Median	Max	StdDev
PER	5,743	17	10	92	18.7
LOC	3,879	11	8	77	12.0
ORG	2,196	6	2	58	9.8
ALL	11,818	34	21	164	35.8
Tokens	189,507	548	245	3,126	613.8
Sents	8,210	23	21	210	13.5
Docs	346	-	-	-	-

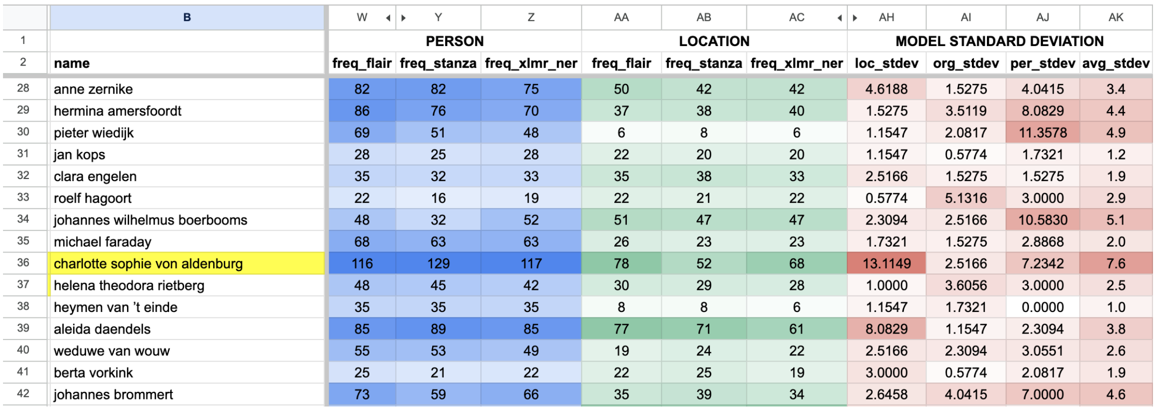
Table 1: General statistics of our manually annotated corpus. We include the average, maximum, and median of occurrences per document to illustrate the heterogeneity of the dataset at the document level.

3.2 NER Models

We consider four candidate models that deliver (at least) the three desired NER labels, PER, LOC, and ORG. We focus only on those three because their definition is less controversial than other entity categories, yet disagreement still emerges across model (Nadeau & Sekine, 2007). Three of the four models have a similar architecture with only minor variants on the top layers of the classifiers, which could make us think there will not be much difference in the final results. Importantly, all models are readily available to use out-of-the-box, which is very attractive for a non-NLP researcher, primarily if the models are published with a high F1 score (around 90 points) associated with them. Commonly, the reported scores are in one of the most famous benchmarks for NER, in our case, the Dutch portion of conll-02 Shared Task (Tjong Kim Sang, 2002). The systems we will benchmark here are:

- **Flair NLP:** This open-source NLP framework (Akbik, Bergmann, Blythe, et al., 2019) contains very strong models for the NER task. The latest model is based on FLERT (Schweter & Akbik, 2020), an optimized method for NER in different languages, including Dutch. The basic architecture is the cross-lingual version of RoBERTa (Y. Liu et al., 2019), with a CRF layer on top to improve the structured prediction. This model was fine-tuned to account for document-level features using the Conll-02 dataset and reports a global F1 score of 94.5 on the Dutch test set.
- **Stanza:** The stanza NER model for Dutch is based on a concatenation of BERT (Devlin et al., 2019) pre-trained vectors and a character-based bidirectional LSTM with a CRF layer on top for the structured prediction. The corpus used for training the Dutch model is from Nothman et al. (2013).

[2] We use the NLTK implementation with binary distance to compute the agreement of labels.



	B	W	Y	Z	AA	AB	AC	AH	AI	AJ	AK
1		PERSON			LOCATION			MODEL STANDARD DEVIATION			
2	name	freq_flair	freq_stanza	freq_xlmr_ner	freq_flair	freq_stanza	freq_xlmr_ner	loc_stdev	org_stdev	per_stdev	avg_stdev
28	anne zernike	82	82	75	50	42	42	4.6188	1.5275	4.0415	3.4
29	hermina amersfoordt	86	76	70	37	38	40	1.5275	3.5119	8.0829	4.4
30	pieter wiedijk	69	51	48	6	8	6	1.1547	2.0817	11.3578	4.9
31	jan kops	28	25	28	22	20	20	1.1547	0.5774	1.7321	1.2
32	clara engelen	35	32	33	35	38	33	2.5166	1.5275	1.5275	1.9
33	roelf hagoort	22	16	19	22	21	22	0.5774	5.1316	3.0000	2.9
34	johannes wilhelmus boerbooms	48	32	52	51	47	47	2.3094	2.5166	10.5830	5.1
35	michael faraday	68	63	63	26	23	23	1.7321	1.5275	2.8868	2.0
36	charlotte sophie von aldenburg	116	129	117	78	52	68	13.1149	2.5166	7.2342	7.6
37	helena theodora rietberg	48	45	42	30	29	28	1.0000	3.6056	3.0000	2.5
38	heymen van 't einde	35	35	35	8	8	6	1.1547	1.7321	0.0000	1.0
39	aleida daendels	85	89	85	77	71	61	8.0829	1.1547	2.3094	3.8
40	weduwe van wouw	55	53	49	19	24	22	2.5166	2.3094	3.0551	2.6
41	berta vorkink	25	21	22	22	25	19	3.0000	0.5774	2.0817	1.9
42	johannes brommert	73	59	66	35	39	34	2.6458	4.0415	7.0000	4.6

Figure 1: We display the table of metrics in a spreadsheet. With conditional formatting, we can display it as a heatmap, where the rows with more discrepancy stand out because of the contrast of colors. This allows us to spot right away which documents show key differences in model behavior.

They report a macro F1 score of 89 in conll-02 and 94.8 on the WikiNER test (Ghaddar & Langlais, 2017).

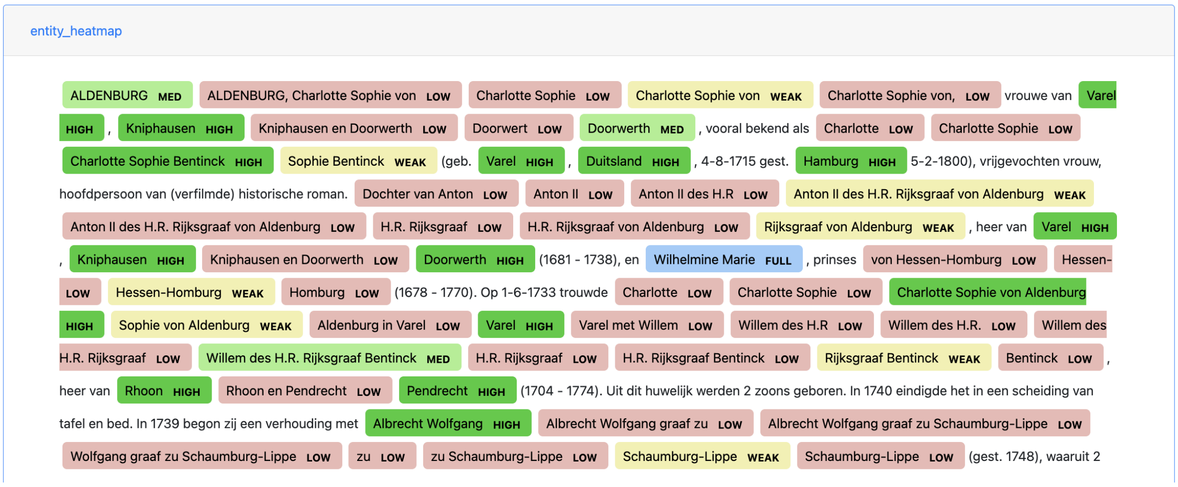
- **Fine-tuned XLM-R:** this is a model[3] published in the HuggingFace repository and is described as a Named Entity Recognition model for ten high-resourced languages (Arabic, German, English, Spanish, French, Italian, Latvian, Dutch, Portuguese and Chinese) based on the Cross-lingual RoBERTa base model (Conneau et al., 2020). It has been fine-tuned with a neural linear layer on top that has been fine-tuned to recognize `LOC`, `ORG`, and `PER`. Given how easy it is to run models published on HuggingFace and that the model card claims to have been trained in the most important NER resources, it is feasible that someone will base their research on the information published there.

- **gpt3.5-turbo:** Finally, given the recent popularity of LLMs with the claims that they can provide very strong performance in almost any task (Min et al., 2023), we also explore the alternative of prompting gpt3.5 (Brown et al., 2020) for obtaining NER labels. Since our focus is not on prompt engineering (P. Liu et al., 2023), we measure the performance as a zero-shot setting with an intuitive prompt and visualize how it compares to applying NER-specific models on our dataset. Specifically, we provide the following prompt: *Identify and Label (PERSON, ORGANIZATION, TIME, LOCATION, ARTWORK, MISC) all of the Named Entities in the following text. Return also the character spans in which these entities appear. Return the results in TSV Format with Columns: [Entity, Label, Span Start, Span End]. Text: <ANSWER>*

4 Inspection of Model Behavior

We can spot interesting behavior in specific instances by comparing raw model outputs in parallel. An instance can be a sentence, paragraph, or document, depending on the required granularity. By looking closely at relevant instances, we can investigate which errors or tagging biases occur with the different models and make decisions accordingly. It would be impossible to inspect every single instance closely. Therefore, we propose to use the predictions of all models to compute straightforward metrics as a proxy for spotting edge cases, that is, to highlight the documents that will be the most useful to examine closely for understanding model behavior and anticipating whether that behavior is optimal for the intended use case. This is possible even when there is no human-annotated data available at all.

[3] https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl



entity_heatmap

ALDENBURG MED ALDENBURG, Charlotte Sophie von LOW Charlotte Sophie LOW Charlotte Sophie von WEAK Charlotte Sophie von, LOW vrouwe van Varel
HIGH , Kniphausen HIGH Kniphausen en Doorwerth LOW Doorwert LOW Doorwerth MED , vooral bekend als Charlotte LOW Charlotte Sophie LOW
Charlotte Sophie Bentinck HIGH Sophie Bentinck WEAK (geb. Varel HIGH , Duitsland HIGH , 4-8-1715 gest. Hamburg HIGH 5-2-1800), vrijgevochten vrouw,
hoofdpersoon van (verfilmde) historische roman. Dochter van Anton LOW Anton II LOW Anton II des H.R LOW Anton II des H.R. Rijksgraaf von Aldenburg WEAK
Anton II des H.R. Rijksgraaf von Aldenburg LOW H.R. Rijksgraaf LOW H.R. Rijksgraaf von Aldenburg LOW Rijksgraaf von Aldenburg WEAK , heer van Varel HIGH
, Kniphausen HIGH Kniphausen en Doorwerth LOW Doorwerth HIGH (1681 - 1738), en Wilhelmine Marie FULL , prinses von Hessen-Homburg LOW Hessen-
LOW Hessen-Homburg WEAK Homburg LOW (1678 - 1770). Op 1-6-1733 trouwde Charlotte LOW Charlotte Sophie LOW Charlotte Sophie von Aldenburg
HIGH Sophie von Aldenburg WEAK Aldenburg in Varel LOW Varel HIGH Varel met Willem LOW Willem des H.R LOW Willem des H.R. LOW Willem des
H.R. Rijksgraaf LOW Willem des H.R. Rijksgraaf Bentinck MED H.R. Rijksgraaf LOW H.R. Rijksgraaf Bentinck LOW Rijksgraaf Bentinck WEAK Bentinck LOW ,
heer van Rhoon HIGH Rhoon en Pendrecht LOW Pendrecht HIGH (1704 - 1774). Uit dit huwelijk werden 2 zoons geboren. In 1740 eindigde het in een scheiding van
tafel en bed. In 1739 begon zij een verhouding met Albrecht Wolfgang HIGH Albrecht Wolfgang graaf zu LOW Albrecht Wolfgang graaf zu Schaumburg-Lippe LOW
Wolfgang graaf zu Schaumburg-Lippe LOW zu LOW zu Schaumburg-Lippe LOW Schaumburg-Lippe WEAK Schaumburg-Lippe LOW (gest. 1748), waaruit 2

Figure 2: We can visualize which entities are preferred by most models and which ones cause more disagreement among them.

We also use a set of visualizations built on top of everyday tools to inspect the instances and make the process more user-friendly. We specifically use spreadsheets to show what we call *divergence matrices* (see Section 4.1 and Figure 1). Additionally, we created a small Flask web app[4] that integrates visualizations from `displaCy`[5] to see the spans in the text; and Google Charts[6] to show aggregated statistics.

4.1 Divergence Matrix

A divergence matrix Z is defined as a collection of P instances $\{p_0, p_1, \ldots, p_P\}$ (for our use case hee we chose a document as the instance unit), a set of N systems (models) denoted as $\{s_0, s_1, \ldots, s_N\}$, and M evaluation metrics $\{m_1, m_2, \ldots, m_M\}$, therefore Z has dimensions $P \times (S * M)$. In this context, a row represents an instance p_i, and each column $Z_{i,j,k}$ signifies the performance score of system s_j when evaluated using metric m_k on instance p_i with $i = 0, 1, \ldots, P;\ j = 0, 1, \ldots, N;\ k = 0, 1, \ldots, M$. Consequently, the matrix encapsulates the evaluated models' performance across all instances and metrics (See Figure 1). Metrics can be defined as required per use case, here we show 3 example metrics:

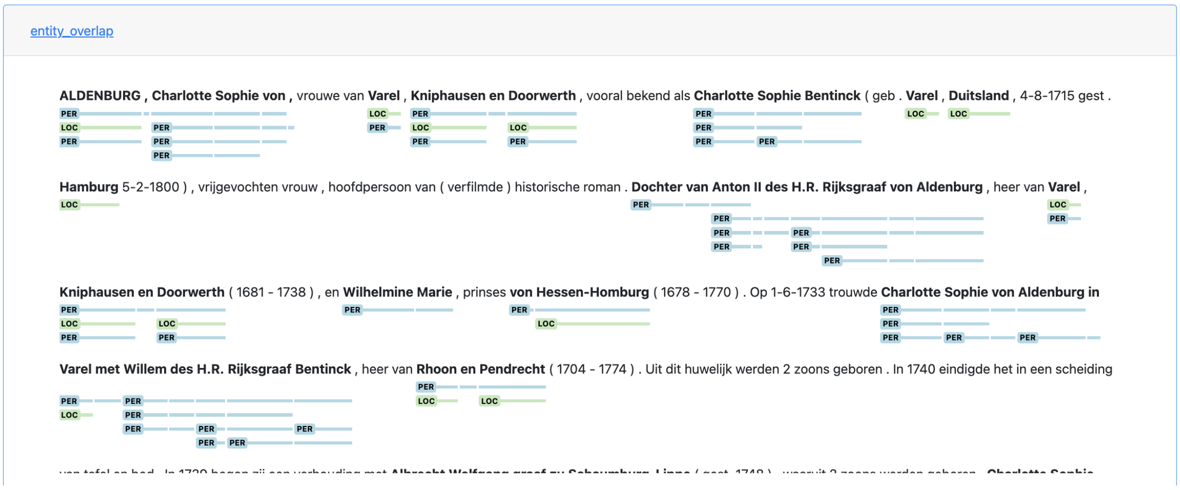
Entity Frequency: get the raw counts of Named Entities found in each document according to each model. These are counted per category (`PER`, `LOC`, `ORG`), and also globally (the sum of all categories).

Entity Density: In this case, we divide the entity frequency by the number of tokens in the instance to get a weighted metric and be less biased towards extensive documents.

Entity Divergence: We define this metric by considering an array of frequencies from $model_0$ to $model_N$, where each element is the entity frequency predicted by a $model_n$. Then, we compute the divergence as the standard deviation of this array. The reasoning is that models will obtain the same amount of labels for easy instances (all models will agree), whereas instances with complex cases or cases with noise will have a high divergence.

To avoid falling again into the trap of looking at these metrics only globally, we can display the divergence matrix as a heatmap, where higher numbers ger darker colors and lower numbers approach to no-color. This way, it is visually easy to spot problematic instances (where color intensity changes dramatically across columns). For example, in Figure 1, the matrix is displayed in a spreadsheet, which has a low threshold for users. In this example, the instance that stands out the most is the biography of

[4] Available here: https://github.com/angel-daza/bios-dutch
[5] https://demos.explosion.ai/displacy-ent
[6] https://developers.google.com/chart/



entity_overlap

ALDENBURG , Charlotte Sophie von , vrouwe van **Varel , Kniphausen en Doorwerth** , vooral bekend als **Charlotte Sophie Bentinck** (geb . **Varel , Duitsland** , 4-8-1715 gest .

Hamburg 5-2-1800) , vrijgevochten vrouw , hoofdpersoon van (verfilmde) historische roman . **Dochter van Anton II des H.R. Rijksgraaf von Aldenburg** , heer van **Varel** ,

Kniphausen en Doorwerth (1681 - 1738) , en **Wilhelmine Marie** , prinses **von Hessen-Homburg** (1678 - 1770) . Op 1-6-1733 trouwde **Charlotte Sophie von Aldenburg in**

Varel met Willem des H.R. Rijksgraaf Bentinck , heer van **Rhoon en Pendrecht** (1704 - 1774) . Uit dit huwelijk werden 2 zoons geboren . In 1740 eindigde het in een scheiding

van tafel en bed . In 1729 begon zij een verhouding met **Albrecht Wolfgang graaf zu Schaumburg-Lippe** (gest. 1748) , waaruit 2 zoons werden geboren . **Charlotte Sophie**

Figure 3: We can also visualize the predicted labeled-span divergences of all models in parallel. Some errors are systematic and can be later counteracted by post-processing.

Charlotte Sophie von Aldenburg, where the matrix shows that there is substantial disagreement between models in LOC and PER assignments.

4.2 Entity Agreement Heatmap

If we dig into an instance (document) with high divergence and visualize the label overlap of all models in it, we can spot where in the text they converge and where they disagree. Figure 2 illustrates the example of Charlotte Sophie von Aldenberg. We classify text spans (entity candidates) into five buckets according to the certainty of correctness using models in a voting mechanism: the entities that have the votes of N models are highlighted in blue (FULL), in dark green are the entities that $N-1$ models identified (HIGH), in light green $N-2$ (MED), in yellow $N-3$ (WEAK) and the rest of entities are labeled in red (LOW) meaning that the certainty of them being the right span is low since most models voted differently. For example, the last name von Aldenburg is obviously derived from the fact that Charlotte came from the nobility related to the Aldenburg location, hence the disagreement.

4.3 Parallel Span Comparison

Another example where we can see this LOC-PER confusion is in the entity overlap (Figure 3). Take the span Varel, Kniphausen en Doorenwerth, which is basically an enumeration of places that Charlotte is a *lady of*. However, some models interpreted that this segment could mean she is the *wife of* and thus classifies the subsequent Entities as people (in this case, stanza was the model that committed this confusion consistently, whereas flair got them right). Because none of the NER systems are trained to deal with these cases, we see a higher disagreement in how they are labeled and, hence, the high divergence across PER and LOC in this document. If we should pick a tool based on this aspect, we could deduce that the flair model behaves more closely to what we would expect with this kind of case.

These simple visualizations already give us a deeper insight into how models behave with our specific examples of interest without the need to spend time generating labeled data. Nevertheless, if the model is to be used at scale, relying on a handful of examples is not sufficient. The need for evaluating the model outputs vs. human annotations is still the desirable scenario. In the next section, we explore different ways of doing this.

	PER			LOC			ORG			MICRO			MACRO		
SYSTEM	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
stanza_iob	77.0	85.0	81.0	82	85.0	83.0	58.0	52.0	55.0	76.0	79.0	77.0	72.0	74.0	73.0
flair_iob	81.0	89.0	85.0	86.0	92.0	89.0	69.0	65.0	67.0	81.0	85.0	83.0	79.0	82.0	**80.0**
stanza_span	61.1	89.7	72.7	64.0	93.7	76.0	42.9	75.3	54.7	59.2	89.1	71.1	56.0	86.2	**67.8**
flair_span	53.4	83.9	65.3	66.9	93.4	78.0	47.5	78.9	59.3	56.7	86.4	68.4	55.9	85.4	67.5

Table 2: Comparison of the same models when feeding pre-tokenized data vs. feeding raw text

5 Evaluation Modes

When you intend to use automatic outputs to support your research, the ideal case is to have more than one human annotator and create a labeled subset with information on inter-annotator agreement. This labeled set can be used to compute evaluation scores that provide information on the quality of the automatic tools when applied to your data. NLP researchers tend to report results of classifiers with mainly three metrics: Precision, Recall, and F1 Score. These metrics were inherited from the field of information retrieval and adopted for other tasks (Powers, 2011). In the next section, we show that even in this standard scenario, it is essential to be careful with how we evaluate the performance scores. It remains important to keep the ultimate research goal in mind when determining which model is the *best model* for the scenario at hand. Different models will rank as *the best* according to different criteria. Often, scores are taken for granted, and differences that appear to be subtle when calculating such scores could be hiding behavior that is particularly harmful to a specific use case (Fokkens et al., 2014).

5.1 Tokenization Matters

Starting with the CoNLL-02 (Tjong Kim Sang, 2002) and CoNLL-03 shared tasks (Tjong Kim Sang & De Meulder, 2003), NER has been approached in NLP as a sequence labeling task, where each token is assigned one label in the IOB format (or related) (Ramshaw & Marcus, 1995). Some of the best-known benchmarks for NER have also taken this approach (Balasuriya et al., 2009; Tedeschi & Navigli, 2022). The conll format already presents a pre-tokenized text, which is also split into sentences. The reasoning behind this is to focus only on evaluating NER performance. The problem is that this approach can give the false impression that tokenization and sentence splitting are trivial tasks when, in fact, differences in scores are expected when applying models to untokenized text (Daza et al., 2022). This is particularly problematic, because the most common scenario for non-NLP users is to apply an out-of-the-box tool to raw text.

We run two parallel experiments: we evaluate the NER task in its default format (feeding the models with the clean and tokenized sentences) vs. the performance of the same models when processing raw text into tokenized text with Named Entities. We call the first evaluation mode *IOB match*. We performed this evaluation using the widely used Python module `seqeval`.[7] In the second approach, which we call *Span match*, we evaluate identified character spans identified by the tools in the raw text. We used character span matches to focus on the Named Entities to avoid noise in the results while comparing sentences of different sizes or tokenization divergences (each model produces different tokens and sentences). In Table 2, we see that the evaluation of raw text (*Span match*) gives significantly lower scores. Secondly, we highlight the fact that in the *IOB match* mode, the best model is flair, whereas, in the *Span match* mode, stanza is the best-performing model.

We consider the *Span match* mode to be closer to what external users of tools need; therefore, the following evaluations will be using the *Span match* mode. Within this approach, we can distinguish two submodalities: to consider a true positive only those entities whose labeling has an exact match with respect to the gold labels, or to be more lenient and consider as a true positive an entity whose span partially overlaps with the gold label (and has the same label, of course). A third evaluation modality emerges with

[7] https://github.com/chakki-works/seqeval

	PER			LOC			ORG			MICRO			MACRO		
SYSTEM	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
stanza	61.1	89.7	72.7	64.0	93.7	76.0	42.9	75.3	54.7	59.2	89.1	71.1	56.0	86.2	67.8
flair	53.4	83.9	65.3	66.9	93.4	78.0	47.5	78.9	59.3	56.7	86.4	68.4	55.9	85.4	67.5
xlmr_ner	59.8	86.5	70.7	65.0	90.2	75.6	48.8	75.7	59.3	59.8	86.2	70.7	57.9	84.1	**68.5**
gpt-3.5	0.3	8.2	0.6	0.6	32.2	1.1	0.1	11.1	0.1	0.3	11.9	0.6	0.3	17.2	0.6

Table 3: Strict evaluation NER. We measure how many predicted spans completely match the gold spans, in terms of span start, end and assigned label.

	PER			LOC			ORG			MICRO			MACRO		
SYSTEM	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
stanza	64.3	92.6	75.9	65.6	88.7	75.4	50.6	70.5	58.9	62.4	87.6	72.9	60.1	83.9	70.1
flair	59.7	97.4	74.0	68.6	95.9	80.0	54.0	80.8	64.8	61.4	94.1	74.3	60.8	91.4	**72.9**
xlmr_ner	63.6	90.9	74.8	67.4	88.2	76.4	56.0	71.0	62.6	63.6	86.6	73.3	62.3	83.4	71.3
gpt-3.5	3.8	6.0	4.6	1.7	1.5	1.6	0.5	0.4	0.5	2.8	3.5	3.1	2.0	2.7	2.2

Table 4: Partial Evaluation NER

the advent of LLMs. Some papers started to ignore spans when reporting scores by evaluating results as what we call here *bag of entities* (Ashok & Lipton, 2023), where the set of predicted entities is compared versus the set of gold entities in the text.[8] All evaluation setups could be considered valid in their own contexts; however, it is important to avoid comparing numbers obtained with different setups as they are not measuring the same signals. Next, we describe each evaluation scenario and show a table of scores when applying each setup. Note that the systems were trained to detect labeled spans (except GPT-3.5); we are only changing the way we evaluate the models, we are not re-training nor changing anything to the models. The observed performance differences thus will be solely because of the assumptions behind each way of evaluating.

5.2 Full Match

In this setting, an entity is considered correct only if it completely matches the gold span and the label. In Table 3, we can see several dimensions when comparing different models. If one only analyzes system-level scores, one would conclude that xlmr_ner is the best-performing system on the test set. However, when we inspect the performance more closely, we gain some insight into what each model is doing: When evaluating with the exact match setup, GPT 3.5's performance is definitely poor, and the reason is that with the prompt that was given, even though it gets several entities correctly, it generates pseudo-random spans. For the traditional systems, we see that Stanza performs much better for detecting `PER` than xlmr_ner and flair, which is very relevant if, for example, we are primarily interested in recognizing people. `LOC` is the easiest category for all models, suggesting fewer unknown entities were found (compared to the entities seen during training) in the biographical dataset; this is possibly because most `LOC` entities are nowadays still popular city and country names that have not changed in the last three centuries. Flair appears to be much better if we are interested in detecting `ORG`s. This category is also the weakest, perhaps because the names of organizations are more sparse, and the definition of an organization has changed through the years compared to the other two categories.

5.3 Partial Match

This mode behaves quite similarly to the full match. The main difference is that the criteria for considering a True Positive is looser. The strict match policy is too harsh for cases where the classifier almost got the whole entity, certainly with the correct label, but missed a couple of tokens compared to the gold. For example, if a model labeled `Charlotte Sophie` instead of `Charlotte Sophie von`

[8] Note that these F1 scores are compared directly to the IOB scores from previous papers, without highlighting the fact that LLMs are not predicting spans in text.

	PER			LOC			ORG			MICRO			MACRO		
SYSTEM	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
stanza	55.1	79.8	65.2	65.9	80.9	72.6	43.2	48.7	45.8	56.2	73.3	63.6	54.7	69.8	61.2
flair	45.4	72.4	55.8	70.5	86.9	77.8	47.0	57.5	51.7	53.1	74.0	61.8	54.3	72.3	61.8
xlmr_ner	54.9	77.6	64.3	66.2	80.6	72.7	48.4	51.9	50.1	57.3	73.0	64.2	56.5	70.0	**62.4**
gpt-3.5	68.9	49.4	57.5	83.5	58.3	68.6	41.5	30.5	35.1	67.6	48.3	56.3	64.6	46.1	53.8

Table 5: Evaluation of NER as bag-of-entities. This setup only considers unique entities per document and does not consider spans where they appear, nor repetitions of the same entity.

Aldenburg as PER, it is still usable information. In this example, the exact match will penalize twice the same mistake: the predicted span is a false positive since it does not exist in gold, and also, the gold span was not predicted, which counts as a false negative. This is not a recommended way of evaluating, but it is used for some cases (e.g. Luthra et al., 2023), so we keep it here for comparison purposes. Table 4 shows the results when evaluating the models on the same dataset but with the partial-match policy. As expected, numbers go up compared to the strict match. In this case, there is also a swap in the model ranking: the best model according to strict match macro F1 was xlmr, and according to partial match macro F1, the best model overall would be flair.

5.4 Bag of Entities

We could argue that in some oversimplified cases, for example, if the last goal is to build a network, we only need to encounter the right entities in the text once and do not care about their place in the text. In this case, it would not matter if *Amsterdam* is mentioned 20 times in a biography, it only matters that we recognize it once as a LOC to draw an edge between this person and this location. Nevertheless, note that this looses track of the instances where *Amsterdam* appears in the text as an ORG (for example, a metonymy) or PER (for example if *Amsterdam* was a last name). Assuming you are willing to take such risks, by evaluating NER as a bag of entities (the set of all unique SurfaceText_Label mentions in an instance), we can drop the requirement of detecting the spans, drawing more attention to the importance of recovering key entities at the expense of losing their original context. We show in Table 5 the precision, recall, and F1 when evaluating this way. At first glance, the numbers go down to the low 60s. Based on the Macro F1, we would pick again xlmr_ner as the best-performing system according to the test, but by a much narrower margin. Notably, without the Span requirement, GPT 3.5 outputs are now useful (which might make this evaluation setup tempting), giving a closer performance related to the NER-specific systems. The same trends stand at the Entity type level: LOC is the easiest category, followed by PER and ORG. If we look at PER metrics: Precision, Recall, and F1 on StrictMatch (Table 3 vs BoE in Table 5), we see a constant decrease in performance (Stanza F1 drops from 72 to 65; Flair F1 drops 63 to 56). This suggests that, since we are only considering Bag of Entities, the globally frequent entity names, which are very often predicted correctly, have less weight *inflating* the performance of the models. This is why the scores from BoE are lower than the Strict Match in general.

5.5 Precision vs Recall

Another critical aspect to determine is whether we would rather have a system that is biased toward False Positives (FP) or towards False Negatives (FN). Systems with a higher proportion of FPs translate into low precision, meaning that more noisy cases will appear in our categories, and systems with a high ratio of FNs, translate into low recall, indicating that many useful cases will be ignored. We can visualize the FPs and FNs of each instance separately so a close inspection can show what kind of mistakes are happening. Importantly, because the NER task is span identification and span labeling, there are different causes of error: i) The span was correctly identified, but the label is wrong, ii) The span is wrongly identified, but the label is correct, iii) The span was wrongly identified and the label is wrong, iv) The labeler did not tag the entity at all. By packing this in a single P, R, or F1 score, we lose the ability

SYSTEM	PER	LOC	ORG	TOTAL
stanza	55.10	65.86	43.22	56.22
flair	45.42	70.47	47.01	53.14
xlmr_ner	54.93	66.21	**48.44**	57.33
gpt-3.5	**68.85**	**83.51**	41.46	**67.55**

Table 6: Table of **Precisions** for each NER label, when evaluated as Bag of Entities

SYSTEM	PER	LOC	ORG	TOTAL
stanza	**79.79**	80.87	48.65	73.28
flair	72.39	**86.93**	**57.46**	**73.95**
xlmr_ner	77.62	80.64	51.87	72.95
gpt-3.5	49.43	58.27	30.45	48.28

Table 7: Table of **Recalls** for each NER label, when evaluated as Bag of Entities

to analyze the nature of the errors. However, visualizing them at the instance level can provide more explanations of the models' behavior.

If we evaluate the systems based on their precision (Table 6), the highest precision is for GPT-3.5. Notably, if we are only interested in recognizing important organizations (and do not care about recall), then it seems like xlmr is the best-performing system in the test data. On the other hand, if what we need is to maximize recall, results in Table 7 show that flair is the highest-performing model. We can also see that GPT3.5's recall is dramatically worse compared to the other systems, suggesting that (at least with the current prompt) GPT will systematically ignore several entities, even though it is precise when generating them (but, once again, this only happens when ignoring spans in the text).

5.6 Micro vs Macro(s)

Evaluating NLP classifiers with Precision, Recall, and F1 is customary. However, it is sometimes surprisingly hard to identify whether the reported metrics are done at the macro level or micro level. Normally, macro metrics are preferred because they average category-level scores. This alleviates the bias that exists across classes (although Opitz and Burst (2019) warn that sometimes different formulas are also used when computing macro scores, resulting in different numbers). Looking at Table 3, it is not surprising that stanza is the best model according to the micro F1, but xlmr is the best model according to the macro F1. The reason for this change in ranking occurs because there is a bigger proportion of `PER` entities in the corpus, a category for which stanza is much better at labeling, obtaining a higher micro; however, the macro counteracts this bias and gives a fairer score, showing the system that performs the best across categories. Hence, we reiterate that an underperforming system could be picked if one does not carefully make the difference between these two.

6 Findings and Discussion

This section provides an overview and discussion of our most important findings. As mentioned in Section 4, we built the visualizations on top of ready-to-use libraries such as Microsoft Excel, DisplaCy and Google charts. These visualizations need at most a couple of lines of basic code to run. We aim for a low technical threshold and the technical skills required to use these visualizations are comparable to those needed for running the tools. The visualizations we propose can help identify where the source of divergences is and remind us that we should check for the following aspects when evaluating. Depending on which aspects are more relevant to a specific use case, different systems can come up as the best solution for our specific use case. The following subsections highlight our main findings.

Tokenization Matters NER has traditionally been approached as a sequence labeling task, where each token is assigned one label in the IOB format (or related) (Ramshaw & Marcus, 1995). Notably, the most common scenario for non-NLP users is to apply an out-of-the-box tool to raw text. Our experiments show that evaluation on raw text gives significantly lower scores than the evaluation assuming tokenization as a given. We consider the *Span match* mode that starts from raw text to be closer to what external users of tools need. Similar gaps between results obtained in "clean" experimental settings for standard benchmarks and real world scenarios may occur for other NLP tasks as well. It is therefore worthwhile to find out what the experimental setup that led to reported results looked like exactly to get a better feeling of what may be expected in your own use case. These verifications provide additional information next to other known factors that may determine results such as differences in genre, domain or time period in which the data was created.

At the same time, not all errors may be equally problematic. It is therefore worthwhile to consider how the output of the tools will be used to see how to weigh different kinds of errors. The following paragraphs highlight relevant aspects in approaching this question.

Full Match vs Partial Match Only entities that match entirely the gold span label are considered correct in a full match setting. This strict match policy can be too harsh for cases where the classifier almost got the whole entity with the correct label but missed a couple of tokens compared to the gold (which can also be fixed with a simple post-processing rule). For example, if a label `Charlotte Sophie` instead of `Charlotte Sophie von Aldenburg` as `PER` in her own biography can easily be mapped to the correct person leading to a fully correct outcome for making networks, thus partial match evaluation could be enough for this case.

Bag of Entities In some cases, e.g. identifying relevant documents or creating networks based on loose connections, we only need to encounter the correct entity in the text once. It then does not matter if *Amsterdam* occurs many times in the document, we only need to recognize it once as a LOC to draw an edge between this person and *Amsterdam*. Here, evaluating without the need of validating spans can be enough.

Precision vs Recall A related aspect is whether high precision (not many false positives) or high recall (not many items missed, or few false negatives) is more important. Because the NER task is span identification and span labeling, there are different causes of error: i) The span was correctly identified, but the label is wrong, ii) The span is wrongly identified, but the label is correct, iii) The span was wrongly identified and the label is wrong, iv) The labeler did not tag at all the entity. If this information remains packed in a single P, R, or F1 score, we lose the ability to analyze the errors. Visualizing the FPs and FNs of individual documents separately can show what kind of mistakes are made so we can act accordingly to fix them.

7 Limitations

This paper compares four specific NER tools for Dutch that can be applied out of the box. We are aware that comparison among more tools could provide more decisive conclusions. The tools that we used as an aid for the instance-level evaluation in principle apply to other languages and text domains, given that the tool handles "labeled spans" regardless of the nature of the labels and the expected spans; however, this paper did not provide experiments on other tasks. We focused on NER only to make a more straightforward point that even an *easy task* can produce relevant disagreement across pre-trained models.

As is widely known, the landscape of Large Language Models is evolving incredibly fast, and several far more powerful models have appeared since this research started. We want to bring attention to the fact that evaluation techniques should be fair across model architectures and task definitions. More importantly, one should proceed with caution when reading about loosely defined experiments, where LLMs are given more leniency on their *inner methods* used to produce results, and an assessment for the specific use case at hand should always be done before assuming someone else's prompt is what will work best.

Finally, we also omitted the scenario where a custom model is fine-tuned for the specific corpus. The reason for this is to emulate the scenario for a Digital Humanities researcher who wishes to use already available tools instead of creating their own models. These two options are not mutually exclusive, as an analysis such as the one shown in this paper can motivate the researcher to train their own models if no available model satisfies the goals of their task at hand.

8 Conclusion

In this paper, we call for a more detailed evaluation of NLP span classification tasks. We apply several out-of-the-box NER models to a corpus of Dutch Biographies and compare different options for evaluation. We aim to illustrate the importance of inspecting the output of models at various levels when investigating whether their output provides the information that is needed with sufficient reliability. We also aim to show that a higher F1 score in a benchmark does not necessarily mean that the model will also be the best choice for our specific use case. We highlight the importance of carefully looking at the outputs to know whether the selected models are working in the way we intend to, going beyond just trusting global metrics in the abstract. Finally, we shared the code where we performed all these analyses as to encourage researchers from various fields to use these methods for their own use cases.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59.

Akbik, A., Bergmann, T., & Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 724–728.

Arnoult, S. I., Petram, L., & Vossen, P. (2021). Batavia asked for advice. pretrained language models for named entity recognition in historical texts. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 21–30. https://doi.org/10.18653/v1/2021.latechclfl-1.3

Ashok, D., & Lipton, Z. C. (2023). Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., & Curran, J. R. (2009). Named entity recognition in Wikipedia. *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, 10–18. https://aclanthology.org/W09-3302

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and ai: An overview of current debates and future perspectives. *J. Comput. Cult. Herit.*, *15*(1).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Daza, A., Fokkens, A., & Erjavec, T. (2022). Dealing with abbreviations in the Slovenian biographical lexicon. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8715–8720. https://doi.org/10.18653/v1/2022.emnlp-main.596

De Los Reyes, D., Barcelos, A., Vieira, R., & Manssour, I. (2021). Related named entities classification in the economic-financial context. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 8–15. https://aclanthology.org/2021.hackashop-1.2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, *56*(2).

Fokkens, A., Ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., et al. (2014). Biographynet: Methodological issues when nlp supports historical research. *LREC*, 3728–3735.

Fu, J., Liu, P., & Neubig, G. (2020). Interpretable multi-dataset evaluation for named entity recognition. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6058–6069. https://doi.org/10.18653/v1/2020.emnlp-main.489

Ghaddar, A., & Langlais, P. (2017). WiNER: A Wikipedia annotated corpus for named entity recognition. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 413–422. https://aclanthology.org/I17-1042

Jiang, G., Luo, Z., Shi, Y., Wang, D., Liang, J., & Yang, D. (2024, May). ToNER: Type-oriented named entity recognition with generative language model. 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. 16251–16262). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.1412/

Kim, H., Kim, J.-E., & Kim, H. (2024, November). Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 8653–8670). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.492

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270. https://doi.org/10.18653/v1/N16-1030

Li, Y., Nair, P., Pelrine, K., & Rabbany, R. (2022). Extracting person names from user generated text: Named-entity recognition for combating human trafficking. *Findings of the Association for Computational Linguistics: ACL 2022*, 2854–2868. https://doi.org/10.18653/v1/2022.findings-acl.225

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, *55*(9). https://doi.org/10.1145/3560815

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luthra, M., Todorov, K., Jeurgens, C., & Colavizza, G. (2023). Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, *80*(5), 1080–1105.

Manjavacas Arevalo, E., & Fonteyn, L. (2022). Non-parametric word sense disambiguation for historical languages. *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 123–134. https://aclanthology.org/2022.nlp4dh-1.16

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, *56*(2). https://doi.org/10.1145/3605943

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*, 3–26. https://api.semanticscholar.org/CorpusID:8310135

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, *194*, 151–175.

Opitz, J., & Burst, S. (2019). Macro F1 and macro F1. *CoRR*, *abs/1911.03347*. http://arxiv.org/abs/1911.03347

Panchendrarajan, R., & Amaresan, A. (2018). Bidirectional LSTM-CRF for named entity recognition. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. https://aclanthology.org/Y18-1061

Picco, G., Martinez Galindo, M., Purpura, A., Fuchs, L., Lopez, V., & Hoang, T. L. (2023, July). Zshot: An open-source framework for zero-shot named entity recognition and relation extraction. In D. Bollegala, R. Huang, & A. Ritter (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 3: System demonstrations)* (pp. 357–368). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-demo.34

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Ramshaw, L., & Marcus, M. (1995). Text chunking using transformation-based learning. *Third Workshop on Very Large Corpora*. https://aclanthology.org/W95-0107

Romary, L. (2014). Natural Language Processing for Historical Texts Michael Piotrowski (Leibniz Institute of European History) Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 17), 2012, ix+157 pp; paperbound, ISBN 978-1608459469. *Computational Linguistics*, *40*(1), 231–233. https://doi.org/10.1162/COLI_r_00180

Schweter, S., & Akbik, A. (2020). FLERT: document-level features for named entity recognition. *CoRR*, *abs/2011.06993*. https://arxiv.org/abs/2011.06993

Schweter, S., März, L., Schmid, K., & Çano, E. (2022). Hmbert: Historical multilingual language models for named entity recognition. *Experimental IR Meets Multilinguality, Multimodality, and Interaction - Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2022)*.

Tedeschi, S., & Navigli, R. (2022). MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). *Findings of the Association for Computational Linguistics: NAACL 2022*, 801–812. https://doi.org/10.18653/v1/2022.findings-naacl.60

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://aclanthology.org/W02-2024

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. https://aclanthology.org/W03-0419

Tong, Z., Ding, Z., & Wei, W. (2025, January). EvoPrompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 5136–5153). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.345/

Ushio, A., & Camacho-Collados, J. (2021). T-NER: An all-round python library for transformer-based named entity recognition. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 53–62. https://doi.org/10.18653/v1/2021.eacl-demos.7

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Yadav, V., & Bethard, S. (2018, August). A survey on recent advances in named entity recognition from deep learning models. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 2145–2158). Association for Computational Linguistics. https://aclanthology.org/C18-1182/

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523

Unlocking the Corpus: Enriching Metadata with State-of-the-Art NLP Methodology and Linked Data

Jennifer Ecker[1], **Stefan Fischer**[2], **Pia Schwarz**[1], **Thorsten Trippel**[1],
Antonina Werthmann[1], **Rebecca Wilm**[1]
[1]Leibniz Institute for the German Language (IDS), Mannheim, Germany
{ecker,schwarz,trippel,werthmann,wilm}@ids-mannheim.de
[2]Saarland University, Saarbrücken, Germany
stefan.fischer@uni-saarland.de

Abstract

In research data management, metadata are indispensable to describing data and are a key element in preparing data according to the FAIR principles. Metadata in catalogues and registries are usually recorded either by archivists or subject matter experts, i.e. researchers involved in the creation or assembling of the data, or provided in the data preparation workflow. Extracting metadata from textual research data is currently not part of most metadata workflows, even more so if a research data set can be subdivided into smaller parts, such as a newspaper corpus containing multiple newspaper articles. If we look at descriptive metadata from a large corpus of newspapers, the basic metadata may consist of information, for example, about the title, or year of publication. Our approach is to add semantic metadata on the text level to facilitate the search over data. We show how to enrich metadata with three methods: named entity recognition, keyword extraction, and topic modeling. The goal is to make it possible to search for texts that are about certain topics or described using certain keywords or to identify people, places, and organisations mentioned in texts without actually having to read them.

1 Introduction

Enriching the information extracted from corpora to find more relevant parts of a corpus for deeper analysis is the overall aim of this contribution. Newspaper corpora or other large collections contain a multitude of texts of various topics, timespans, authors, etc. Some of the properties are already available in form of metadata, hence they can be used to select partitions of these corpora. If these properties are not provided in the metadata, the corpus can basically only be used as is, which may still be a valid and useful application. To enhance usability of the data according to the FAIR principles (Wilkinson et al., 2016), rich metadata as a meaningful representation of the data are a key element. In this contribution, we explore options based on a large reference corpus to enrich the metadata not only for the whole corpus but on the level of subportions, such as articles in a newspaper corpus. For developing the processes, we select a small section of the corpus. With this approach we provide options to find parts of the corpus that may be more relevant for specific tasks, for example to create a subcorpus for specific topics, on given individuals, places or organisations.

2 Motivation

The German Reference Corpus *DeReKo* is the largest linguistically motivated collection of German language material (Kupietz & Keibel, 2009; Kupietz et al., 2010, 2018). The corpus is an example of a national corpus, with all legal restrictions of modern data. Although the sample corpus is not called a national corpus, DeReKo serves the same purposes as national corpora for other languages. It contains multiple newspapers, books, transcriptions, etc. For the purpose of this contribution, the authors received access to a number of data files in their native XML structures. For developing the methodology we focus on one file in DeReKo's native XML format with all issues of one newspaper of one year.

Jennifer Ecker, Stefan Fischer, Pia Schwarz, Thorsten Trippel, Antonina Werthmann and Rebecca Wilm 2024. Unlocking the Corpus: Enriching Metadata with State-of-the-Art NLP Methodology and Linked Data). *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 127–140. DOI: https://doi.org/10.3384/ecp216

Specific research questions may focus on parts of such a corpus. However, there is no general criterion for substructuring a corpus, as this is highly dependent on the research questions. For someone interested in the style of specific authors, the substructuring of such a corpus would be best if all articles or contributions were clustered by their author; for someone interested in specific topics, the clustering should be by topic, for specific timeframes by dates, etc. For a reference corpus that is intended for multiple uses, no such clustering seems to be a universal way to structure a corpus, and the granularity – the archival unit – is not fixed.

Usually each archival unit receives a persistent identifier, hence there is a clear relation between the archival units and PIDs. For PIDs, criteria have been recommended to determine the granularity of archival objects. *ISO 24619:2011* (2011) recommends four different procedures for determining the granularity of objects to receive a persistent identifier which can also be used to determine the granularity of archival objects. These four options in the standard are: (1) use the granularity of an existing PID schema, if such exists, otherwise (2) the PID should be assigned if a resource is complete within one file; if this is not the case (3) it should be assigned to a unit that exists autonomously outside a larger context, and else (4) the PID should be assigned to a unit that should become citable.

For DeReKo, there are files that are used for extending the corpus, such as books represented according to the DeReKo native format I5 (Lüngen & Sperberg-McQueen, 2012), which is an XML schema defined according to recommendations of the Text Encoding Initiative (TEI P5; *TEI P5*, 2021). Most newspapers are added into the reference corpus on a yearly basis, i.e. every issue of the newspaper in its digital form is part of an archive of a specific year. Sources such as Wikipedia with their discussions and history functions are ingested based on these different functions, i.e. one file for all articles, one for the discussions and one for the history. Hence, for archiving and sustainable preservation, DeReKo currently relies on the ingest files, which are also the base for assigning PIDs, and the ingest files constitute the archival object's unit. Each of these collections is represented by a metadata file which is publicly available even if the data file itself is only available under certain restrictions. The internal structure of the original corpus data, however, also allows for other partitions, such as the identification of all newspaper articles published on a specific date, belonging to specific sections, etc. These structures can be identified by their internal unique text *sigles*, which are part of the XML representations of the data. Hence, a sigle is a unique identifier of a subpart of a corpus, either on a corpus, document, or text level, the latter for example for individual newspaper articles. Using the sigle for identifying parts of the corpus requires access to the underlying files.

For accessing linguistic structures, searches on the word, phrase or sentence level are possible with the specialised corpus query tool *KorAP* (Bański et al., 2013; Diewald & Margaretha, 2016; Diewald et al., 2016; Kupietz et al., 2017). Someone who has (legal) access to the full source file can utilise available information and create their own selection of units from the reference corpus based on for example the text sigle as well. Hence, it becomes feasible to create arbitrary subcorpora based on all available attributes, for example clustering subparts by persons with a specific role, place, date, topic, or keyword, if these attributes can be identified for a specific unit.

In the current standard representation, there are only limited options due to sparseness of properties available in the metadata representation. Consequently, it seems very relevant to extract properties of units on various granularity levels. Due to the sheer size of the reference corpus, it will be impossible – and is indeed out of any reasonable suggestion – to add these properties manually: NLP methods need to automate the enrichment of the data.

3 Approach

As a starting point to enrich DeReKo with semantic metadata, we focus on extracting topics, keywords, and three types of named entities: academic institutions, research areas, and persons with an academic background. The experiments are described in Section 5.1 for topic modeling, Section 5.2 for keyword extraction and Section 5.3 for named entity recognition (NER). We believe that these might be useful entry points for researchers to partition the reference corpus to fit their particular research questions. Additionally, these NLP methods give important insights into the corpus and facilitate the search over data.

At the same time, applying three different NLP methods allows us to explore how we can implement a metadata extension – see Section 3.1 – which captures semantic metadata besides the existing catalog metadata. This extension also includes the possibility to record potential links between the extracted semantic information and other external data sets, further described in Section 3.2. There are other established tools for the processing of corpora provided by CLARIN consortia, such as WebLicht (Hinrichs et al., 2010) developed in CLARIN-D and a web-based natural language processing workflow (Walkowiak & Piasecki, 2015) within CLARIN-PL. Due to legal restrictions, our data must remain within our organisation during processing. For this reason, we were unable to utilise external tools that require data to be processed on other servers.

3.1 Metadata Profile Extension

Before enriching the metadata with semantic information, we considered different approaches to accomplish our aim straightforwardly. There were four main options:

1. We can enhance the metadata by adding semantic information into the metadata header provided in the I5 files. These files are based on TEI standards, rendering them not only extensible but also adaptable to our specific requirements. The advantage of this approach is that the text and the newly enriched information are stored within the same file. The major disadvantage of this approach is that it necessitates modifications to the primary I5 data. This can lead to issues if something goes wrong during the enrichment process and requires repairs or modifications. Additionally, this approach of enriching data results in the I5 files becoming larger and more complex. Any subsequent changes or extensions to the I5 data demand greater processing capacity to process the entire DeReKo data set. Furthermore, modifying the I5 files also requires adjustments to the I5 scheme, adding to the overall workload. The update here also poses an issue with regards to long term archiving, as the integrity of the archived files has to be ensured and a change in the underlying data format may change results based on the previously archived files. However, this can be addressed by providing different versions of the files, which in itself causes additional obstacles by additional memory requirements, etc.

2. Enriched metadata can be stored in separate metadata files, e.g. in CMDI[1] (Broeder et al. 2012; *ISO 24622-1:2015* 2015; *ISO 24622-2:2019* 2019) or JSON-LD[2] (*JSON-LD 1.1* 2020) format for each individual sigle of the I5 file. The advantage of this approach is that semantic information can be sequentially extended without affecting the I5 files. However, there are certain disadvantages to consider: The persistent identification for each sigle, using a PID with the sigle ID as part of the identifier, is necessary. This requires technical adjustments within the existing repository. In addition, the heightened complexity of the I5 data structure may lead to reduced clarity. A single I5 file can encompass numerous sigles, requiring the generation of corresponding semantic metadata files for each of them. This can result in increased data complexity, making the data less user-friendly and potentially more challenging to interpret.

3. We can generate a CMDI file for each individual I5 file. The advantages of this approach are: CMDI files containing descriptive metadata can be automatically generated from I5 files and then subsequently expanded with semantic metadata. Both descriptive and semantic information are stored in the same file, and they can be extended and modified at a later stage without requiring changes to the primary I5 files. Changing the CMDI schema is also possible at any time. Additionally, the CMDI files can be converted to alternative formats, such as JSON-LD or HTML, in the future. However, there are some disadvantages to consider: Enriching semantic CMDI metadata requires intricate data modeling and interpretation. Furthermore, the size of a CMDI file depends on the size of the corresponding I5 file and can become quite large. In cases with numerous sigles, the CMDI file may become increasingly complex and challenging to understand.

[1] https://www.clarin.eu/content/component-metadata
[2] https://www.w3.org/TR/json-ld11/

4. We can generate semantic information through real-time analysis. The advantages of this approach are: The I5 files remain unaltered, and there is no need to create additional metadata files. However, it is essential to take into account the significant technical and result-related disadvantages: Intensive real-time processing places a significant burden on computational resources. Such tasks require substantial processing power, leading to higher infrastructure costs and potential bottlenecks in data processing pipelines. The regular changing, modifying, and updating of the tools with the latest technology after deployment can be exceptionally expensive. Changing the licenses of the used NLP applications and models can also have unpleasant consequences because, in the worst case, we may no longer be able to use one or the other tool. Processing queries on huge text data is time-consuming, making real-time query responses practically impossible or excessively difficult. Additionally, the results of queries are reliant on the tools used. If these tools are updated or modified, replicating the results becomes challenging, if not impossible.

After careful consideration, we have chosen to pursue option (3) as this way of adding semantic stand-off annotation is comparatively easy to maintain in case of any changes, whether they concern the original I5-formatted data or the semantic annotation itself. On top of that, legal restrictions of the data are not specified on the metadata. With CMDI files, we can make our findings available, even though the original data can only be accessed by specific users of the corpus. The extracted metadata can be shared under open licenses (such as CC-0 or CC-BY), offering reference to the original data and thereby promoting accessibility and transparency.

3.2 Linking Data to External Knowledge Bases

There are a number of available knowledge bases that the named entities, topics, and keywords that we extracted from our corpus could be linked to. These resources differ with respect to size, domain, and annotation quality. In part, this can be explained by their different creation processes: While some resources are created manually, others are generated automatically. In addition, resources can be subject to strict curation processes, or they might be limited to a specific domain.

Wikidata[3] (Vrandečić & Krötzsch, 2014) is a multilingual knowledge graph that is part of the Wikimedia project. Like Wikipedia, it is created by volunteers and available under a permissive license. Each Wikidata item has a persistent identifier (QID) and is explained by a label and short description. Additional information is provided in structured key–value statements. As the knowledge graph is based on the collaborative editing effort of volunteers, records might be inaccurate or not up-to-date. With more than 100,000,000 items, which are not restricted by topic, Wikidata provides large coverage and can serve as a “hub” that connects identifiers from different authority files. Wikidata has been used for the development of named entity linkers (Delpeuch, 2020; Möller et al., 2022; Sakor et al., 2020).

The Integrated Authority File[4] (Gemeinsame Normdatei; GND; Behrens-Neumann & Pfeifer, 2011) provides a catalogue of entities with unique and reliable identifiers. The GND covers persons, corporate bodies, events, place names, subject headings as well as cultural and academic works. It comprises roughly 10,000,000 records, which can only be edited by participating organisations. In comparison to Wikidata, the GND provides high-quality entries from an authoritative source at the cost of lower coverage.

Although the aforementioned resources cover a broad range of domains, they might not provide sufficient coverage for rather specific use cases, e.g. academic named entity recognition (see Section 5.3). However, there are domain-specific databases providing information about both research organisations and individual researchers: the Research Organization Registry[5] (ROR; Lammey, 2020) and the Open Researcher and Contributor ID[6] (ORCID; Haak et al., 2012). The ROR is a community effort to provide persistent identifiers and metadata about research organisations. The ROR contains entries for more than 100,000 international organisations. The registry is searchable via an API and can be downloaded freely.

[3] https://www.wikidata.org

[4] https://explore.gnd.network

[5] https://ror.org

[6] https://orcid.org

ORCID is a platform that assigns persistent identifiers to participating researchers. After registration, a researcher can provide information about their publications, affiliations or grants. The ORCID database can be downloaded freely and the records are also searchable via an API.

GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) is a lexical-semantic network for German. In this network, lexical units are organised into concepts (*synsets*) whose members share the same meaning. Both lexical units and synsets are identified by unique IDs. GermaNet provides lexical-semantic information and a conceptual network for adjectives, adverbs, nouns, and verbs. However, named entities referring to persons are not part of its database. In our use case, the hierarchical relations between synsets might be of interest. For example, one could determine (co-)hyponyms of a detected named entity in order to connect texts that cover the same concept but use different surface forms. The GermaNet database can be used freely for academic research.

In order to link extracted topics, keywords, and named entities, the respective identifiers (QIDs, GND IDs, IDs from ROR and ORCID, GermaNet IDs) of corresponding items from all these external knowledge bases can be encoded per text sigle in the semantic stand-off annotation files. So far, we have not implemented this linking of the data, partly because the problem of word sense disambiguation between the extracted items in our corpus and possible counterparts in the different knowledge bases has not yet been solved.

4 Data

For our experiments we selected a data sample from DeReKo, the 2020 volume of the newspaper corpus Mannheimer Morgen (M20), published under the QAO-NC license (Kupietz & Lüngen, 2014) allowing for query-and-analysis only academic and non-commercial use. According to DeReKo's structure, M20 is a single I5-formatted XML file containing several individual newspaper articles each identified by their text sigle (e.g. M20/APR.00192), which consists of the corpus identifier M20, the document identifier, corresponding to the month in which the article was released, and a five-digit text identifier. In total the M20 subcorpus comprises 44,383 texts.

The selection of a subset for developing the process has not been chosen arbitrarily. We chose the subset based on the following criteria:

- The size of the data set is substantial. Running metadata extraction processes on a test set should provide sufficient information on runtime and hardware use to evaluate whether these processes could be scaled to the full data set, i.e. some thousand additional subcorpora as input.
- The size of the data is small enough to allow experimenting with the data in short periods of time and with limited computing power. The overall goal was to make sure that individual processes on such a corpus would not take too long. The duration of such a process influences the development, as restarts and new tries may be required frequently. In addition, the CO_2 footprint of failed experiments would not be too large.
- The data and its TEIish serialisation are prototypical for many other data sets that are part of the corpus, e.g. other newspapers and magazines, but also books and Wikipedia articles.

As such a metadata enrichment process is not a real-time application, initial processes have to be fast enough and modest in hardware requirements to scale in a way that the full process could be run without HPC environments in a reasonable amount of time.[7] The M20 subcorpus had the right size and format for experimenting.

5 Experiments

Using NLP for keyword extraction, named entity recognition and topic modeling is a rather typical task in the development of these methods. Enriching metadata based on a large set of linguistically annotated

[7] In the current development state, we experience runtimes without specific optimisations that could easily lead to durations in the magnitude of years on the full data set of DeReKo.

Number	Generated Name	Topic Words
0	Musik ‘music’	Songs, Album, Sound, Pop, Song, Musiker, Hits, Gitarren, Schlagzeuger, Gitarrist, ... ‘songs’, ‘music album’, ‘sound’, ‘pop’, ‘song’, ‘musician’, ‘hits’, ‘guitars’, ‘drummer’, ‘guitarist’
1	Religion ‘religion’	Kirche, Gläubigen, Pfarrer, Gottesdienst, Kirchen, Gottesdienste, Gebet, Andacht, Gläubige, beten, ... ‘church’, ‘the faithful’, ‘pastor’, ‘service’, ‘churches’, ‘services’, ‘prayer’, ‘devotions’, ‘the faithful’, ‘to pray’
88	Entfesselung ‘unleashing’	schlank, Zeugen, Jacke, trug, Hinweise, Hose, Täter, bekleidet, Fahndung, flüchtete, ... ‘slim’, ‘witnesses’, ‘jacket’, ‘wore’, ‘evidence’, ‘trousers’, ‘offender’, ‘dressed’, ‘tracing’, ‘escaped’

Table 1: Topics named with the help of a Llama model, two of them correctly labeled (number 0 and number 1) and one of them incorrectly labeled (number 88).

corpora has not been applied to the DeReKo corpus. As the authors are not the maintainers of the corpus, it is obvious that the representation of the added information will have to follow stand-off procedures, such that the archived corpus may not be modified or tampered with. Hence, our approach is a standard procedure in NLP by using independent processes on the data to be investigated in order to extract the wanted set of information and to see if these methods can beneficially be applied to this corpus, as well as to figure out how the extracted information could best be represented in a productive system. In the following, we present current work from topic modeling, keyword extraction and named entity recognition.

5.1 Topic Modeling

One exception for existing metadata of the DeReKo corpus is that each text sigle is annotated with a topic as described in Weiß (2005): First, a human annotator constructs a thematic taxonomy based on specific guidelines and the training data. These guidelines include clusters generated by a document clusterer and an external ontology (i.e., the Open Directory Project). Afterwards, corpus texts are automatically classified using the training data. However, this method has significant deficiencies: the taxonomy, developed partly bottom-up through clustering, no longer covers all the domains needed, the granularity is inadequate, the base taxonomy from the Open Directory Project is no longer in use, and the classifier is outdated due to its almost 20-year-old training data. Therefore, a new approach of assigning topics is required.

For our data set, we employ topic modeling to group the articles into categories. The categories are not predefined with the help of an ontology, but learned by a topic model. We use the Top2Vec model[8] (Angelov, 2020) to assign topics to the articles. Before deciding to use Top2Vec, we also experimented with other topic modeling approaches like Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and BERTopic (Grootendorst, 2022). After manually reviewing the results, we found that the results from Top2Vec best met our requirements. Furthermore, Top2Vec proved to be the easiest approach to implement. With Top2Vec, we selected the doc2vec embedding model for application to our dataset, configuring the speed parameter to deeplearn while using the default settings for all other parameters. Top2Vec divides the articles automatically into 348 topics and assigns a similarity to every article. The cosine similarity between the article vector and the topic vector depicts this semantic similarity. After a manual inspection, we use hierarchical topic reduction to

[8] https://github.com/ddangelov/Top2Vec

reduce the topics from 348 to 150 to circumvent that about half the topics are semantically too close to each other.

The output of the topic model is a number of unnamed topics and corresponding topic words for each topic. Unnamed is meant in the sense that there is no word that describes or sums up the topic words. Instead, numbers are assigned to differentiate between the topics. There are different approaches available for labeling the topics. While some take the highest ranked topic word as the name, others only provide the first n words as a description. Another alternative is to use graph-based labeling with the help of a knowledge resource to automatically label the topics (Ecker, 2024). Manually labeling the topics is also an option, but we decide to prompt a large language model[9] to do the labeling. The used model is based on Meta's Llama 2 model[10]. The top twenty topic words calculated and ranked by the model form the basis for computing the label (see Table 1 for an example with a few topic words). The topic name can be one of the topic words or a word derived from them. If the generated label is not suitable to describe the topic words, we define a label manually (see Section 6 for examples).

5.2 Keyword Extraction

Up to ten uni- or bigram keywords are extracted for each article by combining YAKE! (Campos et al., 2018a, 2018b, 2020), a state-of-the-art unsupervised keyword extraction method that assigns a score to each possible keyword based on statistical features, with a filter based on spaCy (Honnibal et al., 2020) part-of-speech tags[11] to exclude any parts of speech other than nouns and proper nouns. In order to avoid inflected forms such as 'Bundesfinanzministeriums' ('Federal Ministry of Finance's'), the resulting keywords are lemmatised using spaCy.

5.3 Academic NER

In order to recognise academic named entities, we fine-tune a German BERT$_{\text{BASE}}$ model as, to our knowledge, no German NER model with the required entity types academic person, academic organisation, and research area exists. Sentences were filtered out of 10,000 randomly selected texts from DeReKo (mainly newspaper articles and press releases) with the help of an off-the-shelf NER model from the Stanza NLP package (Qi et al., 2020) and word lists containing prototypical mentions for each of the entity types. The word list for the entity type academic person (PER-RES) lists academic titles such that a string match with an item from the list combined with a detected person entity from the Stanza NER model results in a candidate entity. For academic institutions (ORG-RES) and research areas (AREA-RES), the word lists were created using official lists from the German Research Foundation[12] and the Federal Government[13] to detect candidate entities using simple string matching. During post-processing, candidate entities were manually reviewed, which resulted in a data set of 4,928 sentences with a total of 7,199 tags. The data was split into a training, development, and test set with a ratio of 70/20/10. Table 2 provides an overview of the entity type distribution across data splits.

Entity Type	No. of Tags in Train / Dev / Test	P	R	F1
PER-RES	2,942 / 858 / 423	93.68	97.19	95.4
ORG-RES	1,624 / 484 / 192	89.58	88.66	89.12
AREA-RES	450 / 147 / 79	89.47	80.0	84.47
Overall	5,016 / 1,489 / 694	92.12	92.78	92.45

Table 2: Tag distribution of the entity types and resulting model performance measured in precision, recall, and F1-score in percent. Overall scores are micro-averaged.

[9] https://huggingface.co/TheBloke/Llama-2-13B-chat-GGML

[10] https://huggingface.co/meta-llama

[11] The 'de_core_news_lg' model is used for spaCy part-of-speech tagging and lemmatisation.

[12] https://www.dfg.de/dfg_profil/gremien/fachkollegien/faecher/index.jsp

[13] https://www.bundesbericht-forschung-innovation.de/de/Liste-der-Einrichtungen-1790.html

Using the spaCy transformer library[14], we fine-tune the model *de_dep_news_trf*[15] on a single Tesla P4 GPU. With respect to the hyperparameters for model training, spaCy's default settings were used: a batch size of 128, a dropout rate of 0.1, the Adam optimiser with an initial learning rate of 10^{-5}, and early stopping based on the F1 score. Model evaluation was done on the 489 sentences of the test split, yielding an overall micro-F1 score of 92.45%. One observation that can be made already are that the individual F1 scores for the entity types increase according to the amount of tags per type (see Table 2). However, these numbers do not necessarily need to correlate, and there might be other reasons why for the model some entity types are more difficult to recognize than others. For more details about the NER model refer to Schwarz (2024), which also includes experiments with LLMs compared to the BERT approach: The evaluation on the test data shows that the fine-tuned BERT model yields better results than any of the LLMs. However, given the very dynamic developments of LLMs, we can certainly not exclude that later models can keep up with the fine-tuned BERT model. The approach of fine-tuning an LLM with the given NER task was not tested and might be subject of further experiments.

6 Results and Discussion

Enriched metadata is crucial for usability, accessibility, and overall value of corpora for users and must therefore be of high standards. Information about the whole corpus can be accessed without the need to read each text and therefore users can quickly locate pertinent information about the topic of the text, relevant keywords, and people, places, and organisations named in the text. In this section, we present the results of the application of the three methods and we discuss which improvements could be made.

6.1 Topic Modeling

For the method of topic labeling, the majority of the generated names for the topics are suitable. Nevertheless, we changed the label in 39 out of 150 cases (26 percent). Table 1 includes an example of an incorrectly labeled topic (number 88). The German word 'Entfesselung' ('unleashing') is not fitting, and a more appropriate name for the topic would be 'Fahndung' ('tracing'), which is also included in the topic words. Besides, there are other mistakes, such as pseudowords, which are similar to a German word that is suitable as the name of the topic but contain a wrong letter, or English words, which we then translated from English to German, because the generated English word is generally suitable. An example for a pseudoword is 'Fahrverböt' instead of 'Fahrverbot' ('suspended driving licence'), and an example for an English name is 'Digital Life' instead of 'digitales Leben'. To improve the topic words, one option is to apply lemmatisation before or after the topic model. In this case, 'Gottesdienst' ('service') and 'Gottesdienste' ('services') would only be listed once as 'Gottesdienst' (see Table 1 for topic number 1). The list of topic words (in topic number 88) would also not include conjugated verbs such as 'trug' ('wore'), which would then be represented with the infinitive form 'tragen' ('to wear'). With lemmatisation, the number of topics correctly labeled by the Llama model could increase, but again, incorrectly lemmatised words may impede this. Furthermore, experiments with other Llama models should be carried out to compare the outcome of different models.

6.2 Keyword Extraction

While no quantitative evaluation is performed for keyword extraction due to a lack of gold-standard data, table 3 presents the results obtained for three different articles, showcasing some remaining difficulties. For all three articles, the keywords make it easy to guess what they may be about (in combination, at least), thus fulfilling their general purpose. However, some of the keywords are not ideal due to part-of-speech tagging and lemmatisation mistakes. Keywords such as 'jugendliche' ('young') and 'Ungewöhnlichst Buchtitel' ('most unusual book title') were selected although all constituent tokens were supposed to be restricted to be nouns or proper nouns. 'Gedankengänge' ('trains of thought') was lemmatised to 'Gedankengäng' rather than the correct 'Gedankengang', a form that does not actually exist in German and would be unlikely to be searched for. It may be possible to improve the results by experimenting with

[14] https://spacy.io/api/transformer
[15] https://github.com/explosion/spacy-models/releases/tag/de_dep_news_trf-3.7.2

other part-of-speech taggers and lemmatisers. Most notably, the respective annotations that are already available for DeReKo seem worth exploring.

Further difficulties arise from the parameters chosen for YAKE!. As demonstrated by the keywords ‘Saturday Night’ and ‘Night Live’, which would be more adequately represented by the trigram keyword ‘Saturday Night Live’, there are cases in which the approach of only taking unigram and bigram keywords into consideration falls short. With no gold keywords available, however, it is difficult to determine whether the advantages of increasing the value of n would outweigh the drawbacks. The same holds true for the number of keywords extracted per article. For the text sigle M20/APR.00002, the surname ‘Sträter’ appears in three of the extracted keywords, but the combination of the comedian’s first name and surname, ‘Torsten Sträter’ (arguably an especially important keyword since it is one that users would seem likely to search for), which does appear in the original text, is missing. While increasing the number of keywords extracted per article would lead to more relevant keywords being found, more irrelevant keywords would also be extracted. It is not clear what the ideal number or cutoff value in terms of YAKE! score would be.

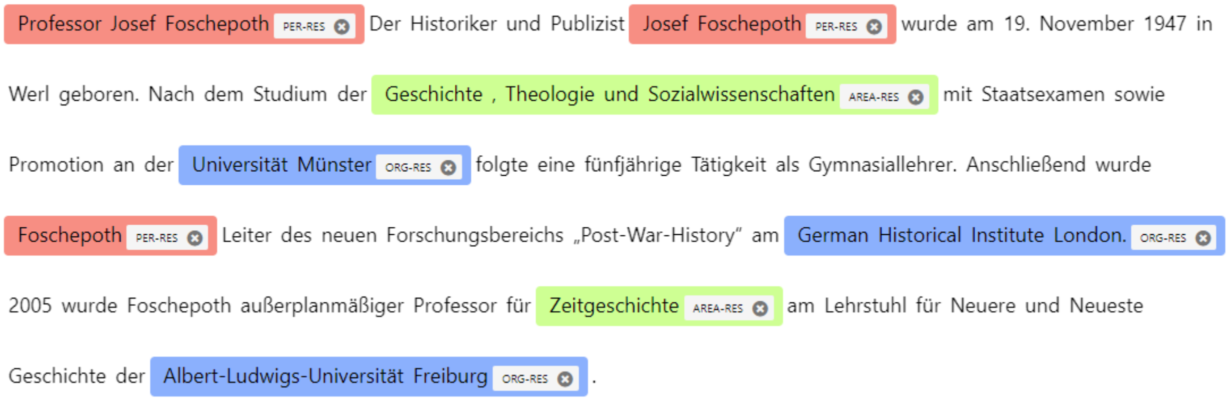
Finally, it may be worth exploring more modern, LLM-based approaches to keyword extraction. Unfortunately, given that no gold keywords are available for the given articles, it is difficult to compare the performance of different models on the given task. We decided against manually annotating the data with gold keywords since the task would be very time-consuming and highly subjective.

Article	Keywords
M20/JAN.00004	Jugendförderung, Zeltlager, Vorbereitung, Toskana, Möglichkeit, Viernheim, jugendliche, Stadtteilbüro Ost, Ferienfreizeit, Anmeldeformular ‘youth empowerment’, ‘camp’, ‘preparation’, ‘Tuscany’, ‘possibility’, ‘Viernheim’, ‘young, ‘district office East’, ‘holiday camp’, ‘registration form’
M20/JAN.00060	Saturday Night, Night Live, Sender NBC, Howard Shore, – Aviator, Panic Room, Kanada, Bild, Ton, Howard ‘Saturday Night’, ‘Night Live’, ‘channel NBC’, ‘Howard Shore’, ‘– Aviator’, ‘Panic Room’, ‘Canada’, ‘picture’, ‘sound’, ‘Howard’
M20/APR.00002	Sträter Gedankengäng, Gedankengäng, Zuschauer, lachen, Luft, Bühne, Sträter, Ungewöhnlichst Buchtitel, Thalia Buch, Sträter Verhältnis ‘Sträter train of thought’, ‘train of thought’, ‘spectator’, ‘laughing’, ‘air’, ‘stage’, ‘Sträter’, ‘most unusual book title’, ‘Thalia book’, ‘Sträter relationship’

Table 3: Keyword extraction results for three articles.

6.3 Academic NER

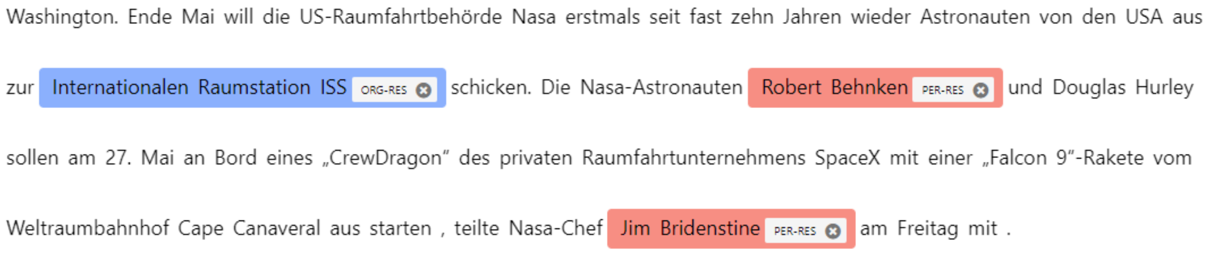
When applying the fine-tuned NER model to the M20 subcorpus, at least one academic named entity is tagged in almost 40 percent of the 44,383 newspaper articles. Most of the tags, over 20,000, fall upon the type PER-RES, almost 10,000 items are tagged with ORG-RES, and a bit more than 3,000 with the entity type AREA-RES. Sentences from two randomly selected articles illustrate good and bad example output of the NER model. In Figure 1, the person ‘Josef Foschepoth’ is detected as a researcher three times. In the first occurrence, the preceding academic title makes it an obvious choice. For the second and third occurrence, however, no such title is present, but the model is still able to correctly assign a PER-RES tag based on the context, unlike in the last sentence. Although through the context it is obvious for a human reader to identify ‘Foschepoth’ as an person with academic background, the model leaves this occurrence untagged. Regarding the detected entities of type AREA-RES, the first one contains three research areas for which it would have been preferable to have each research field tagged as an individual entity. Whereas the second detected entity of the type AREA-RES is fine, the model fails to tag the sequence ‘Neuere und Neueste Geschichte’ (‘recent and modern history’) in the last sentence. The entities of the type ORG-RES are all tagged correctly. This is not the case for the sentences in Figure 2, where the International Space Station ISS is erroneously tagged as ORG-RES and two NASA members



Professor Josef Foschepoth PER-RES Der Historiker und Publizist Josef Foschepoth PER-RES wurde am 19. November 1947 in Werl geboren. Nach dem Studium der Geschichte , Theologie und Sozialwissenschaften AREA-RES mit Staatsexamen sowie Promotion an der Universität Münster ORG-RES folgte eine fünfjährige Tätigkeit als Gymnasiallehrer. Anschließend wurde Foschepoth PER-RES Leiter des neuen Forschungsbereichs „Post-War-History“ am German Historical Institute London. ORG-RES 2005 wurde Foschepoth außerplanmäßiger Professor für Zeitgeschichte AREA-RES am Lehrstuhl für Neuere und Neueste Geschichte der Albert-Ludwigs-Universität Freiburg ORG-RES .

Figure 1: Mostly correct predictions of named entities in sentences from article M20/APR.00264 tagged with the fine-tuned NER model (with tags PER-RES in red, AREA-RES in green, and ORG-RES in blue).

are tagged as PER-RES although there is no evidence in the text for their academic background. Due to the sheer size of the subcorpus, no detailed qualitative analysis of the output is made, but it might be worth to invest some effort into finding error patterns in the results and test if targeted additional training data would diminish the amount of incorrectly tagged data. A certainly helpful feature for the NER model would be some kind of score accompanying the tags to indicate the model's confidence in the respective tags.



Washington. Ende Mai will die US-Raumfahrtbehörde Nasa erstmals seit fast zehn Jahren wieder Astronauten von den USA aus zur Internationalen Raumstation ISS ORG-RES schicken. Die Nasa-Astronauten Robert Behnken PER-RES und Douglas Hurley sollen am 27. Mai an Bord eines „CrewDragon“ des privaten Raumfahrtunternehmens SpaceX mit einer „Falcon 9“-Rakete vom Weltraumbahnhof Cape Canaveral aus starten , teilte Nasa-Chef Jim Bridenstine PER-RES am Freitag mit .

Figure 2: Incorrect predictions of named entities in sentences from article M20/APR.02952 tagged with the fine-tuned NER model (with tags PER-RES in red and ORG-RES in blue).

6.4 Technical Challenges

It applies to all three methods that there are still improvements to be done regarding run-time and memory consumption. For the selected M20 subcorpus, the three processes were run individually and took between six hours and three days. When processing DeReKo as a whole, this would be multiplied by several thousand times and needed to be optimised, e.g. through parallel computing. The same issue holds for temporary files created during the preprocessing of large I5 files before the three NLP methods can be applied.

7 Conclusion and Future Work

Our experiments show that extracting metadata from linguistically motivated large corpora is possible. The usefulness of this metadata will have to be proven in the future based on possible tools (e.g. for corpus analysis) making use of these bits of information to identify subcorpora.

The results can be applied to all types of national corpora and other large corpora including those that have legal restrictions. The sample corpus is not called a national corpus, but DeReKo serves similar purposes as national corpora for other languages. The method described and implemented here will be usable across all languages and corpora with a similar structure, e.g. containing newspaper texts, articles etc. Though the method was only applied to a defined substantial subset of the corpus, it was created to finally analyse the whole corpus. To apply this to other national corpora, a requirement is the existence of appropriate models for the respective languages and that the corpora are available or can at least be preprocessed to obtain the appropriate structures suitable to the task (topic modeling, NER, keyword extraction), e.g. a segmentation into individual articles, sentences etc.

With the application of this methodology to a corpus such as DeReKo and the perspective of applying it to national corpora, there is also potential for its use within other areas of international research data infrastructures such as CLARIN. Enriching the metadata for textual corpora allows for additional functionalities, including preparing linked data applications based on the metadata involved.

One issue left open so far is the integration of linked data sources for the identified topics, keywords and named entities. In the future we will use GermaNet for words and concepts, ontologies for topics, and authority files for named entities. This will allow us to connect the metadata to the Semantic Web and other forms of knowledge graphs.

The described methods were applied to only a subset of the corpus. One of the next steps is to scale these experiments to the full data set. Running the processes several thousand times will require stable processes and fully automated metadata enrichment. The initial tests on the limited data set were successful.

Another set of intended experiments will look at different types of corpora. For example, a corpus of endangered languages such as DOBES[16] contains many different languages lacking existing NLP models tailored to their specific linguistic characteristics. However, multi-tier annotations, for example for spoken language, often contain a gloss in another language such as English, German or French, for which such models are available. Additionally, metadata often contain a textual description of the data which should also be valuable sources for NLP processes for topic modeling, keyword extraction, and named entity recognition. Hence, we will explore how multi-tier annotated corpora can utilise the same technique for enriching metadata as well.

Acknowledgements

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

References

Angelov, D. (2020). Top2Vec: Distributed representations of topics. https://doi.org/10.48550/arXiv.2008.09470

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C., & Witt, A. (2013). KorAP: The new corpus analysis platform at IDS Mannheim. In Z. Ventulani & H. Uszkoreit (Eds.), *Proceedings of the 6th Language and Technology Conference (LTC'13)* (pp. 586–587). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3261/file/Banski_KorAP_2013.pdf
The following values have no corresponding Zotero field:tertiary-authors: Vetulani, Z., and H. Uszkoreittertiary-title: Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conferencepages: 586-587.

[16] https://dobes.mpi.nl/

Behrens-Neumann, R., & Pfeifer, B. (2011). Die Gemeinsame Normdatei - ein Kooperationsprojekt (Deutsche Nationalbibliothek, Ed.). *Dialog mit Bibliotheken*, *23*(1), 37–40.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., & Trippel, T. (2012, May). CMDI: A component metadata infrastructure. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini, & T. Trippel (Eds.), *Proceedings of the workshop describing language resources with metadata: Towards flexibility and interoperability in the documentation of language resources (LREC'12)* (pp. 1–4). European Language Resources Association. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-108677

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018a). A text feature based automatic keyword extraction method for single documents. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 684–691). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_63

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018b). Yake! collection-independent automatic keyword extractor. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 806–810). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_80

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Delpeuch, A. (2020, November). OpenTapioca: Lightweight entity linking for Wikidata (short paper). In L.-A. Kaffee, O. Tifrea-Marciuska, E. Simperl, & D. Vrandečić (Eds.), *Proceedings of the 1st Wikidata Workshop (Wikidata 2020)*. CEUR-WS.org.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., & Witt, A. (2016). KorAP Architecture - Diving in the Deep Sea of Corpus Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 3586–3591). Paris: European Language Resources Association (ELRA) 2016.

Diewald, N., & Margaretha, E. (2016). Krill: KorAP search and analysis engine (M. Kupietz & A. Geyken, Eds.). *Corpus Linguistic Software Tools. Journal for Language Technology and Computational Linguistics (JLCL)*, *31*(1), 73–90.

Ecker, J. (2024, May). Labeling results of topic models: Word sense disambiguation as key method for automatic topic labeling with GermaNet. 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. 10014–10022). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.875/

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. https://doi.org/10.1087/20120404

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. https://aclanthology.org/W97-0802

Henrich, V., & Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2228–2235. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf

Hinrichs, M., Zastrow, T., & Hinrichs, E. (2010, May). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA). https://aclanthology.org/L10-1184/

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). SpaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

Language resource management – Persistent identification and sustainable access (PISA) (International Standard). (2011, May). International Organization for Standardization (ISO). Genf. https://www.iso.org/standard/37333.html

Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model (International Standard). (2015, January). International Organization for Standardization (ISO). Geneva.

Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language (International Standard). (2019, July). International Organization for Standardization (ISO). Geneva. https://www.iso.org/standard/64579.html

JSON-LD 1.1 – A JSON-based Serialization for Linked Data (W3C Recommendation 16 July 2020). (2020, July). World Wide Web Consortium (W3C). https://www.w3.org/TR/json-ld/

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (pp. 1848–1854). European Language Resources Association (ELRA) 2010.

Kupietz, M., Diewald, N., Hanl, M., & Margaretha, E. (2017). Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In M. Konopka (Ed.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung. Jahrbuch des Instituts für Deutsche Sprache 2016* (pp. 319–329). de Gruyter.

Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi (Ed.), *Workings Papers in Corpus-based Linguistics and Language Education* (pp. 53–59, Vol. 3). Tokyo University of Foreign Studies 2009.

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2378–2385. http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018, May). The German Reference Corpus DeReKo: New Developments – New Opportunities. 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)* (pp. 4353–4360). European Language Resources Association (ELRA).

Lammey, R. (2020). Solutions for identification problems: A look at the Research Organization Registry. *Science Editing*, *7*(1), 65–69. https://doi.org/10.6087/kcse.192

Lüngen, H., & Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, *Issue 3*. https://doi.org/10.4000/jtei.508

Möller, C., Lehmann, J., & Usbeck, R. (2022). Survey on English entity linking on Wikidata: Datasets and approaches. *Semantic Web*, *13*(6), 925–966. https://doi.org/10.3233/SW-212865

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Sakor, A., Singh, K., Patel, A., & Vidal, M.-E. (2020). Falcon 2.0: An entity and relation linking tool over Wikidata. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3141–3148. https://doi.org/10.1145/3340531.3412777

Schwarz, P. (2024). Semiautomatic data generation for academic named entity recognition in German text corpora. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, & M. Wiegand (Eds.), *Proceedings of the 20th conference on natural language processing (konvens 2024)* (pp. 173–181). Association for Computational Linguistics. https://aclanthology.org/2024.konvens-main.20

P5: Guidelines for electronic text encoding and interchange (version 4.2.1. last updated on 1st march 2021, revision 654a5c551) (tech. rep.). (2021). Text Encoding Initiative. Retrieved October 9, 2023, from https://guidelines.tei-c.de/en/html/index.html

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Walkowiak, T., & Piasecki, M. (2015). Web-based natural language processing workflows for the research infrastructure in humanities. *5th Conference of the Japanese Association for Digital Humanities*, 61–63.

Weiß, C. (2005). Die thematische Erschließung von Sprachkorpora. *Mannheim: Institut für Deutsche Sprache. OPAL-Online publizierte Arbeiten zur Linguistik, 1/2005*. https://pub.ids-mannheim.de/laufend/opal/pdf/opal2005-1.pdf

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*. https://doi.org/https://doi.org/10.1038/sdata.2016.18

FAIR Tool Discovery: an automated software metadata harvesting pipeline for CLARIAH

Maarten van Gompel
KNAW Humanities Cluster
Amsterdam, the Netherlands
proycon@anaproy.nl

Menzo Windhouwer
KNAW Humanities Cluster
Amsterdam, the Netherlands
menzo.windhouwer@di.huc.knaw.nl

Abstract

We present the Tool Discovery pipeline, a core component of the CLARIAH infrastructure in the Netherlands. This pipeline harvests software metadata from the source, detects existing heterogeneous metadata formats already in use by software developers, and converts them to a single uniform representation based on schema.org and codemeta. The resulting data is then made available for further ingestion into other user-facing catalogue/portal systems.

1 Introduction

Software is indispensable in a lot of modern-day research, including in sectors such as the Humanities and Social Sciences that may have traditionally been less focused on information technology. It is also appreciated more and more as valid research output, alongside more conventional output such as academic publications, presentations, and datasets. Scholars often have a need for research software to do their research efficiently.

For scholars it is therefore important to be able to find and identify tools suitable for their research, we call this process *tool discovery*. We define *tool* here and throughout this paper to broadly refer to any kind of software, regardless of the interface it offers and the audience it targets. The scholar's requirement to find tools is reflected in the letter F for *Findable* in the ubiquitous acronym FAIR [1] that has received a lot of attention in recent years in academic circles. The term is often adopted to promote quality and sustainability in research software (Jiménez et al., 2018). In order to find tools, researchers must have access to catalogues that relay *accurate* software metadata.

There is no shortage in existing initiatives in building such catalogues; many research groups, projects or institutes have some kind of website featuring their tools. Aggregation of software metadata from multiple sources is also not new. CLARIN itself already does this in the CLARIN Virtual Language Observatory[2] (van Uytvanck et al., 2010), and DARIAH in the SSHOC Open Marketplace[3] . However, the system we describe in this paper is not an attempt to build another catalogue. We developed a generic pipeline that harvests software metadata from the software's source, leveraging various existing metadata formats, and converting those to a uniform linked open data representation. This data can then be used to feed catalogues.

2 The need for high-quality metadata

Unlike most digital data, software is uniquely characterised as a constantly moving target rather than a static deliverable entity. Releases at different points in time address bugs, security vulnerabilities, or add new features. Moreover, software lives not in isolation, but in connection to other software; its dependencies. Updates are needed to adapt to changes in its runtime environment.

For software metadata to be informative in this dynamic setting, it needs to reflect this moving target and explicitly link to a particular version of the software. This also facilitates provenance keeping and

[1] Findable, Accessible, Interoperable and Reusable

[2] https://vlo.clarin.eu

[3] https://marketplace.sshopencloud.eu/

Maarten Van Gompel and Menzo Windhouwer 2024. FAIR Tool Discovery: an automated software metadata harvesting pipeline for CLARIAH. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 141–150. DOI: https://doi.org/10.3384/ecp216

scientific reproducibility. Furthermore, metadata should convey information about the stage of development the software is in and the level of support an end-user may expect. The user would be wise to exercise caution in adopting software that is unmaintained and unsupported. In practice we often find this information lacking and come across catalogues that were manually compiled once but rarely updated since.

The need for accurate up-to-date metadata goes hand-in-hand with the need for *complete* metadata. If vital details are missing, the end-user may not be able to make an informed judgment.

A common pitfall we have observed in practice is that metadata is often manually collected at some stage and published in a catalogue, but never or rarely updated or revised. In best case, the software has moved on and the metadata covers a mere subset, in worst case, the software or the entire catalogue is unmaintained and out of date.

3 Bottom-up harvesting from the source

What we propose is a *fully automated* pipeline where software metadata is kept at the source, i.e. alongside the software source code, and *periodically* harvested from there. This is in contrast to approaches where metadata primarily resides in an intermediate system that is manually constructed or curated, which is a common approach for many software catalogues[4]. Our approach has a number of important advantages:

1. Source code is often already accompanied by software metadata in existing schemas because many programming language ecosystems already either require or recommend this. Our aim is to avoid any duplication of metadata and *reuse* these existing sources to the maximum extent possible.

   Consider for example `pyproject.toml` or `setup.py` for Python projects, `package.json` for javascript/npm/nodejs projects, `pom.xml` for Java/Maven and `Cargo.toml` for Rust. Aside from these, valuable machine parsable metadata may be extracted from other conventional files such as a `LICENSE` file or a `README.md` file. The latter often contains machine-interpretable badges. Badges are small images often included on top of the README to express certain properties of the software, such as links to documentation, continuous integration services, development status, packaging status. Research software developers also often include a `CITATION.cff`[5] file which we can automatically parse for metadata. All these different sources may be present and can be recombined in our harvesting process.

2. Source code is typically held in a version control system (usually git) and published in forges such as Github, Gitlab, Bitbucket, Codeberg or Sourcehut. This solves versioning issues and ensures metadata can exactly describe the version alongside which it is stored. It also enables the harvester to properly identify the latest stable release, provided some kind of industry-standard versioning system like semantic versioning is adhered to. Software forges themselves may also provide an Application Programming Interface (API) that may serve as an extra source to find software metadata (e.g. descriptions, keywords, links to issue trackers and/or continuous integration services).

3. The developers of the tool have full control and authorship over their metadata. There are no middlemen.

4. Software forges were designed precisely for collaboration on open source software development, so mechanisms for any third party to amend or correct the metadata are already in place (e.g. via a pull/merge request or patch via e-mail). So while developers retain full authorship, this does not mean outside contribution and curation is not possible.

We do not harvest any metadata from intermediaries. By that we mean that we do not use other catalogues as sources (e.g. via the aforementioned OAI-PMH endpoints), only the software source itself.

[4] for example, https://research-software-directory.org/ offers such a platform. Metadata can often be exported via OAI-PMH.

[5] https://citation-file-format.github.io/

Using intermediaries would defeat our philosophy. We do have one extra input source for harvesting: In case the tool in question is Software as a Service (SaaS), i.e. a web-application, web-service, or website, we harvest not only its software source code, but also its web endpoint and attempt to automatically extract metadata from there. We make a clear distinction between the software source code, software instances (executables) you can run locally, and software instances offered as a service via the web. Formally, the software source code has no knowledge when, where, and by whom it may be deployed, neither locally on some user's computer nor as a service on some server. This link is therefore established at an independent and higher level. In the resulting metadata, there will be an explicit link between the source code and *applications* of that source code[6]. The sources for harvesting source repositories and web endpoints (both effectively just URLs) are the only input that needs to be manually provided to our harvesting system, we call this the *source registry*. This is the higher level we referred to earlier. We keep the source registry in a simple git repository containing very minimalistic configuration files (one yaml file per tool). This is also the only point in our pipeline where there is the possibility for a human curator to decide whether or not to include a tool.

Usage of such a manually curated source registry means that, for this project, automatic discovery of tools is not in scope. That is, we do not actively crawl the web in search for tools that might or might not fit a certain domain. Some interpret the term 'tool discovery' to also include such functionality, but we do not. Such a step, however, can be envisioned as a separate step prior to execution of our pipeline.

4 A unified vocabulary for software metadata

The challenge we are facing is primarily one of mapping multiple heterogeneous sources of software metadata to a unified vocabulary. Fortunately, this is an area that has been explored previously in the CodeMeta project[7]. They developed a generic vocabulary for describing software source code, extending schema.org vocabulary and contributing their efforts back to them. Moreover, the CodeMeta project defines mappings, which they call *crosswalks*, between their vocabulary and many existing metadata schemes, such as those used in particular programming languages ecosystems or by particular package managers.

Schema.org and CodeMeta are both linked open data (LOD) vocabularies[8], and codemeta is canonically serialised to a JSON-LD[9] file which makes it easily parsable for both machine and human alike. This `codemeta.json` file can be kept under version control alongside a tool's source code. A rather minimal example of a such a file is shown below:

```json
{
    "@context": [
        "https://w3id.org/codemeta/3.0",
        "http://schema.org",
    ],
    "@id": "https://example.org/mysoftware",
    "@type": "SoftwareSourceCode",
    "identifier": "mysoftware",
    "name": "My Software",
    "author": {
        "@type": "Person",
        "givenName": "John",
        "familyName": "Doe"
    },
    "description": "My software does nice stuff",
    "codeRepository": "https://github.com/someuser/mysoftware",
    "license": "https://spdx.org/licenses/GPL-3.0-only",
    "developmentStatus": "https://www.repostatus.org/#active",
    "thumbnailUrl": "https://example.org/thumbnail.jpg"
}
```

[6] Applications are instances of the source-code in executable form after build and deployment and may also refer to availability as a service over a network

[7] https://codemeta.github.io

[8] i.e. building upon RDF and being retrievable over HTTP

[9] https://www.w3.org/TR/json-ld/

The developer has a choice to either run our harvester and converter themselves and commit the resulting codemeta file, or to not add anything and let the harvester dynamically reconstruct the metadata every harvest cycle.

The convention to add a `codemeta.json` file alongside the source code was established by the CodeMeta project. In addition to this, we define another method that is specific for our metadata harvester: Developers can add a `codemeta-harvest.json` file instead of `codemeta.json`. Whereas `codemeta.json` by definition contains the complete metadata, the `codemeta-harvest.json` file contains an arbitrary subset and is used to supplement any automatically harvested metadata. This allows developers to rely on the harvester for most fields, without having to run it themselves, but still allows them to provide additional manual metadata. All these different options ensure that developers themselves can choose precisely how much control to exert over the metadata and harvester. It allows us to accommodate both projects that aren't even aware they're being harvested, as well as projects that want to fine-tune every metadata field to their liking, effectively rendering most of our periodic harvester out of work at run-time.

4.1 Additional Vocabularies

We link to various other linked open data vocabularies, listed below. Most of these are formulated as SKOS[10] vocabularies.

- **repostatus.org**[11] – *Development Status* – The repostatus.org project allows developers to express usability and development/support status of a project. A LOD (SKOS) version of this vocabulary was developed in the scope of this project and contributed back to the repostatus project.

- **SPDX**[12] – *Open-source software licenses* – The Software Package Data Exchange project is a Linux Foundation project that defines, amongst other things, open source software licenses. It is widely used, e.g. by package managers.

- **TaDiRaH**[13] – *Research activities* – The TaDiRAH vocabulary "classifies and categorizes the activities that comprise digital humanities" (Borek et al., 2016) with the aim to help scholars group and identify projects that share certain commonalities. We adopt this vocabulary to describe the research activities a software tool can be used for. Example of some top-level categories in this vocabulary are 'Analyzing', 'Capturing', 'Creating', 'Enriching'. An example of a deeper-level category that is particularly common in language resources such as those seen in CLARIN is for example 'Enriching → Annotating → Named Entity Recognition'.

- **NWO Research Domains**[14] – NWO is the Dutch Research Council. They define several research fields that are used for official grant applications. As CLARIAH is a Dutch project, we use this vocabulary to express the research domain a tool is used for. A LOD (SKOS) version of this vocabulary was developed in the scope of this project.

The first two vocabularies are generic enough to be applicable to almost all software projects, we strongly recommend their usage. The latter two may be more constrained to research software as developed in CLARIAH and CLARIN. In your projects, you can adopt whatever you find suits your needs best, the power to mix and match is at the heart of linked open data after all.

Moreover, we formulated some of our own extensions on top of codemeta and schema.org:

- **Software Types**[15] – Software comes in many shapes and forms, targeting a variety of audiences with different skills and needs. We want software metadata to be able to accurately

[10] https://www.w3.org/TR/skos-reference/

[11] https://repostatus.org

[12] https://spdx.dev

[13] https://vocabs.dariah.eu/tadirah/

[14] https://www.nwo.nl/en/nwo-research-fields

[15] https://github.com/SoftwareUnderstanding/software_types

express what type(s) of interface their software provides. The schema.org vocabulary distinguishes `softwareApplication`, `WebApplication`, `MobileApplication` and even `VideoGame`. This covers some interfaces from a user-perspective, but is not as extensive nor as fine-grained as we would like yet. Interface types from a more developer-oriented perspective are not formulated. We therefore define additional classes such as `DesktopApplication` (software offering a desktop GUI), `CommandLineApplication`, `SoftwareLibrary` and others in this add-on vocabulary.

- **Software Input/Output Data**[16] – This minimal vocabulary defines just two new properties that allows for software metadata to express what kind of data it consumes (e.g. takes as input) and what kind of data it produces (e.g. output). It does not define actual data types because schema.org already has classes covering most common data types (e.g. `AudioObject`, `ImageObject`, `VideoObject`, `TextDigitalDocument`, etc...) and properties like `encodingFormat` to tie these to MIME-types or `inLanguage` to tie it to natural languages.

  Do note that describing a full API is explicitly out of scope for our project. A full API description would describe exactly which function or web-endpoints take and return what data. Although this too can be considered a type of metadata, such functionality goes beyond what we consider the primary software metadata which end-users need to make a informed decision regarding the suitability of a tool for their ends. Other existing projects such as the OpenAPI Initiative[17] delve into this realm for Web APIs. For software libraries there are various existing API documentation generators[18] that derive documentation directly from the source code in a formalised way. We do not intend to duplicate those efforts.

5 Architecture

The full architecture of our pipeline is illustrated schematically in Figure 1. Although we demonstrate this in the context of the CLARIAH project, the underlying technology is generic and can also be used for other projects.

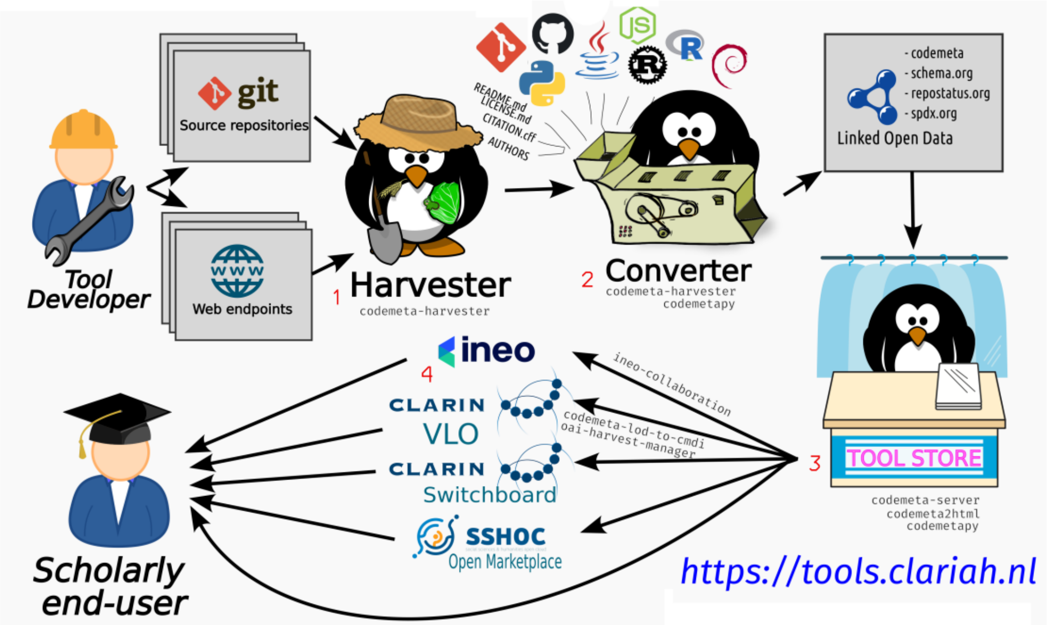


Figure 1: The architecture of the CLARIAH Tool Discovery pipeline. Key steps are numbered in red and referenced in the text.

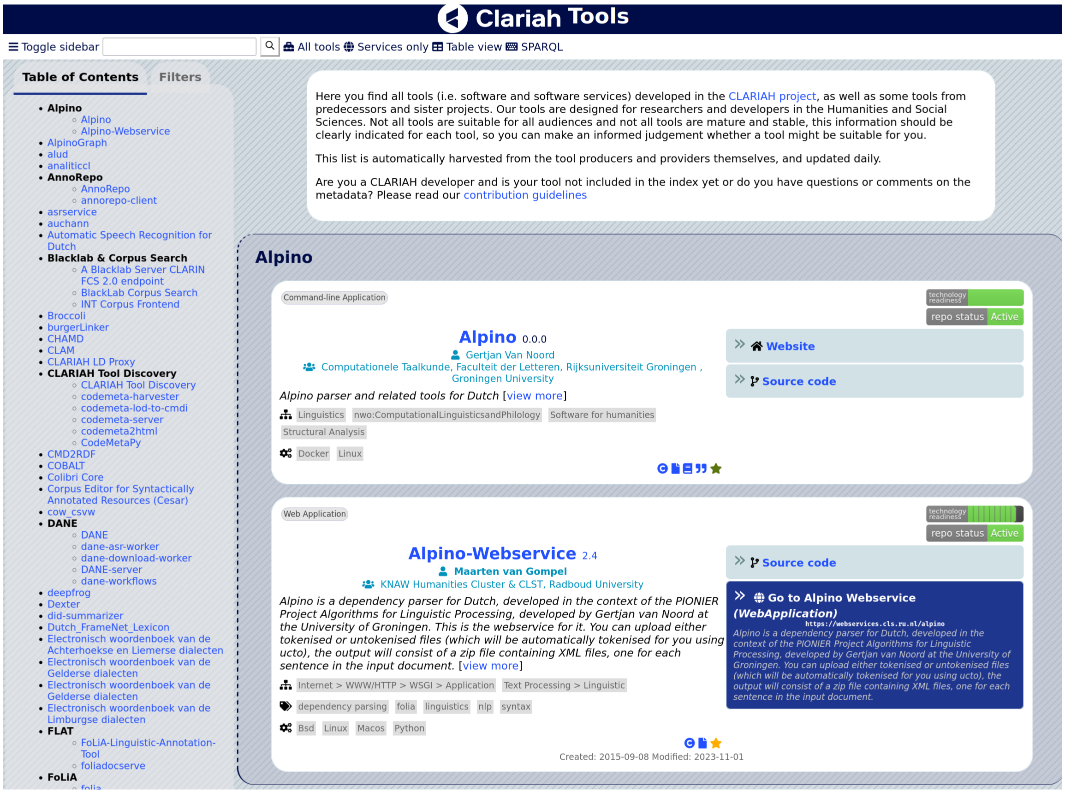
[16] https://github.com/SoftwareUnderstanding/software-iodata

[17] https://www.openapis.org

[18] e.g. doxygen, sphinx, rustdoc, etc...

Using the input from the source registry, our *harvester*[19] (1) (van Gompel et al., 2024) fetches all the git repositories and queries any service endpoints. It does so at regular intervals (e.g. once a day). This ensures the metadata is always up to date. When the sources are retrieved, it looks for different kinds of metadata it can identify there and calls the converter (2) powered by codemetapy[20] (van Gompel, 2024) to turn and combine these into a single codemeta representation. This produces one codemeta JSON-LD file per input tool.

All of these together are loaded in our *tool store* (3), powered by codemeta-server[21] (van Gompel, 2023) and codemeta2html[22]. This is implemented as an RDF triple store and serves both as a backend to be queried programmatically using SPARQL, as well as a simple web frontend to be visited by human end-users as a catalogue. The frontend for CLARIAH is accessible as a service at https://tools.clariah.nl and shown in Figures 2 and 3. At the time of writing, there are 114 registered source repositories and 34 web endpoints.



- **Alpino**
  - Alpino
  - Alpino-Webservice
- AlpinoGraph
- alud
- analiticcl
- **AnnoRepo**
  - AnnoRepo
  - annorepo-client
- asrservice
- auchann
- Automatic Speech Recognition for Dutch
- **Blacklab & Corpus Search**
  - A Blacklab Server CLARIN FCS 2.0 endpoint
  - BlackLab Corpus Search
  - INT Corpus Frontend
- Broccoli
- burgerLinker
- CHAMD
- CLAM
- CLARIAH LD Proxy
- **CLARIAH Tool Discovery**
  - CLARIAH Tool Discovery
  - codemeta-harvester
  - codemeta-lod-to-cmdi
  - codemeta-server
  - codemeta2html
  - CodeMetaPy
- CMD2RDF
- COBALT
- Colibri Core
- Corpus Editor for Syntactically Annotated Resources (Cesar)
- cow_csvw
- **DANE**
  - DANE
  - dane-asr-worker
  - dane-download-worker
  - DANE-server
  - dane-workflows
- deepfrog
- Dexter
- did-summarizer
- Dutch_FrameNet_Lexicon
- Electronisch woordenboek van de Achterhoekse en Liemerse dialecten
- Electronisch woordenboek van de Gelderse dialecten
- Electronisch woordenboek van de Gelderse dialecten
- Electronisch woordenboek van de Limburgse dialecten
- **FLAT**
  - FoLiA-Linguistic-Annotation-Tool
  - foliadocserve
- **FoLiA**
  - folia

Here you find all tools (i.e. software and software services) developed in the CLARIAH project, as well as some tools from predecessors and sister projects. Our tools are designed for researchers and developers in the Humanities and Social Sciences. Not all tools are suitable for all audiences and not all tools are mature and stable, this information should be clearly indicated for each tool, so you can make an informed judgement whether a tool might be suitable for you.

This list is automatically harvested from the tool producers and providers themselves, and updated daily.

Are you a CLARIAH developer and is your tool not included in the index yet or do you have questions or comments on the metadata? Please read our contribution guidelines

Alpino

Command-line Application

technology readiness

repo status Active

Alpino 0.0.0

Gertjan Van Noord

Computationele Taalkunde, Faculteit der Letteren, Rijksuniversiteit Groningen , Groningen University

Alpino parser and related tools for Dutch [view more]

Linguistics · nwo:ComputationalLinguisticsandPhilology · Software for humanities · Structural Analysis

Docker · Linux

» Website

» Source code

Web Application

technology readiness

repo status Active

Alpino-Webservice 2.4

Maarten van Gompel

KNAW Humanities Cluster & CLST, Radboud University

Alpino is a dependency parser for Dutch, developed in the context of the PIONIER Project Algorithms for Linguistic Processing, developed by Gertjan van Noord at the University of Groningen. This is the webservice for it. You can upload either tokenised or untokenised files (which will be automatically tokenised for you using ucto), the output will consist of a zip file containing XML files, one for each sentence in the input document. [view more]

Internet > WWW/HTTP > WSGI > Application · Text Processing > Linguistic

dependency parsing · folia · linguistics · nlp · syntax

Bsd · Linux · Macos · Python

Created: 2015-09-08 Modified: 2023-11-01

» Source code

» Go to Alpino Webservice *(WebApplication)*

`https://webservices.cls.ru.nl/alpino`

Alpino is a dependency parser for Dutch, developed in the context of the PIONIER Project Algorithms for Linguistic Processing, developed by Gertjan van Noord at the University of Groningen. You can upload either tokenised or untokenised files (which will be automatically tokenised for you using ucto), the output will consist of a zip file containing XML files, one for each sentence in the input document.

Figure 2: Screenshot of the CLARIAH Tool Store showing the index page

5.1 Propagation to Software Catalogues

Our web front-end is not the final destination; our aim is to propagate the metadata we have collected to other existing portal/catalogue systems (4), such as the CLARIN VLO, the CLARIN Switchboard, the SSH Open Marketplace, and CLARIAH's Ineo[23]. The latter has already been implemented, the VLO export will be done via a conversion from codemeta to CMDI, and the Marketplace conversion has started in collaboration with DARIAH.

Propagation of software metadata can be visualised as a simple input/output process where the input side connects to our tool store, either via our SPARQL endpoint or by simply obtaining the entire (or

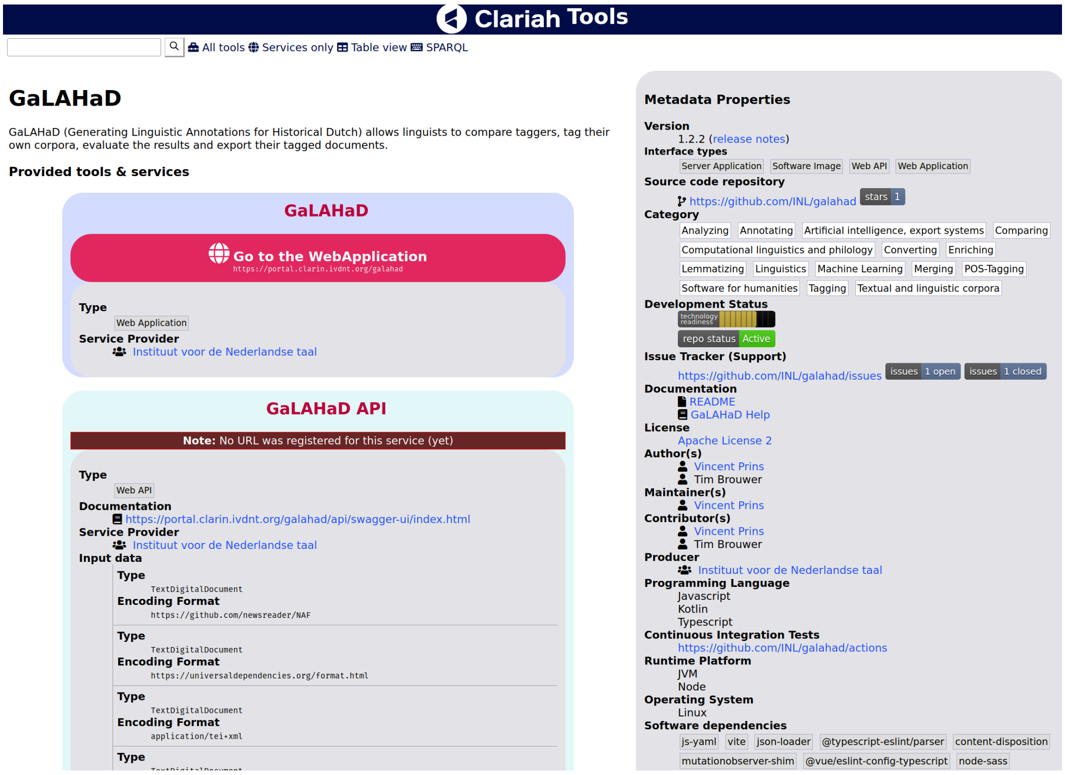
[19] https://github.com/proycon/codemeta-harvester, marked with a red 1 in Figure 1

[20] https://github.com/proycon/codemetapy

[21] https://github.com/proycon/codemeta-server

[22] https://github.com/proycon/codemeta2html

[23] https://vlo.clarin.eu/, https://switchboard.clarin.eu/, https://marketplace.sshopencloud.eu/, https://ineo.tools



GaLAHaD

GaLAHaD (Generating Linguistic Annotations for Historical Dutch) allows linguists to compare taggers, tag their own corpora, evaluate the results and export their tagged documents.

Provided tools & services

GaLAHaD

Type
Web Application
Service Provider
Instituut voor de Nederlandse taal

GaLAHaD API

Note: No URL was registered for this service (yet)

Type
Web API
Documentation
https://portal.clarin.ivdnt.org/galahad/api/swagger-ui/index.html
Service Provider
Instituut voor de Nederlandse taal
Input data
Type
TextDigitalDocument
Encoding Format
https://github.com/newsreader/NAF
Type
TextDigitalDocument
Encoding Format
https://universaldependencies.org/format.html
Type
TextDigitalDocument
Encoding Format
application/tei+xml
Type

Metadata Properties

Version
1.2.2 (release notes)
Interface types
Server Application · Software Image · Web API · Web Application
Source code repository
https://github.com/INL/galahad
Category
Analyzing · Annotating · Artificial intelligence, export systems · Comparing · Computational linguistics and philology · Converting · Enriching · Lemmatizing · Linguistics · Machine Learning · Merging · POS-Tagging · Software for humanities · Tagging · Textual and linguistic corpora
Development Status
Issue Tracker (Support)
https://github.com/INL/galahad/issues
Documentation
README
GaLAHaD Help
License
Apache License 2
Author(s)
Vincent Prins
Tim Brouwer
Maintainer(s)
Vincent Prins
Contributor(s)
Vincent Prins
Tim Brouwer
Producer
Instituut voor de Nederlandse taal
Programming Language
Javascript
Kotlin
Typescript
Continuous Integration Tests
https://github.com/INL/galahad/actions
Runtime Platform
JVM
Node
Operating System
Linux
Software dependencies
js-yaml · vite · json-loader · @typescript-eslint/parser · content-disposition · mutationobserver-shim · @vue/eslint-config-typescript · node-sass

Figure 3: Screenshot of the CLARIAH Tool Store showing the metadata page for a specific tool

a specific part of the) graph in JSON-LD. This may be a periodic query or even a real-time query. The output side connects to either a catalogue-specific API or directly to some kind of database underlying the catalogue system. The process itself consists of conversion from our codemeta representation to whatever representation is suited for the catalogue system.

The connection to the SSHOC Open Marketplace is still ongoing work[24]. For this conversion, we load the JSON-LD graph into an in-memory triple store, iterate over specific triples, and then perform API calls to the SSHOC Open Marketplace API.

In the case of CLARIN's VLO the codemeta schema has been translated into a CMDI profile (Windhouwer & Goosen, 2022). The VLO's harvester has been extended to, next to the traditional OAI protocol, allow other "protocols" to be plugged in[25]. In this case the plugin takes the JSON-LD dump from the tool store and converts the records to equivalent CMDI records compliant with the profile. The changes we in CLARIAH made to the VLO's harvester are currently being tested by CLARIN. Once a new stable version of this harvester has been released the harvesting cycle of CLARIN will be extended to harvest the metadata.

Finally, CLARIAH's Ineo also has a harvesting cycle, which transforms the JSON-LD records into the JSON expected by Ineo's update API. These transformations are minimal as the tool metadata has been designed with this target catalogue in mind.

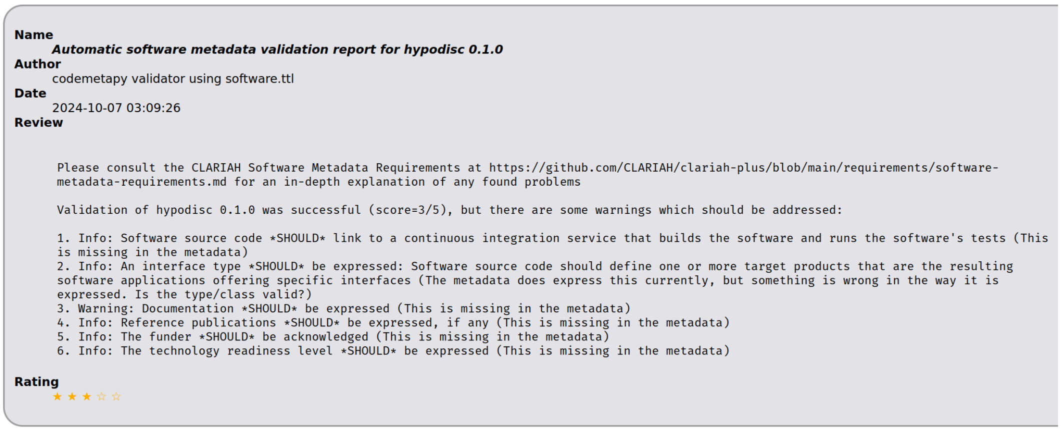
6 Validation & Curation

Having an automated metadata harvesting pipeline may raise some concerns regarding quality assurance. Data is automatically converted from heterogeneous sources and immediately propagated to our tool store, this is not without error. In absence of human curation, which is explicitly out of our intended scope, we tackle this issue through an automatic validation mechanism. This mechanism provides feedback for the developers or curators.

[24] An initial prototype can be found at https://github.com/proycon/codemeta2mp

[25] https://github.com/clarin-eric/oai-harvest-manager, https://github.com/CLARIAH/oai-harvest-manager

The harvested codemeta metadata is held against a validation schema (SHACL) that tests whether certain fields are present (completeness), and whether the values are sensible (accuracy; it is capable of detecting various discrepancies). The validation process outputs a human-readable validation report which references a set of carefully formulated *software metadata requirements* [26]. These requirements state exactly what kind of metadata we expect for software in the CLARIAH project, using normative keywords such as MUST, SHOULD and MAY in accordance with RFC2119 (Bradner, 1997). These requirements provide instructions to developers about how they can provide this metadata in their `codemeta.json` or `codemeta-harvest.json` if metadata can not be automatically extracted from existing sources. The validation schema and requirements document are specific to the CLARIAH project, but may serve as an example for others to adapt and adopt. An example of a validation report referencing the metadata requirements is shown in Figure 4.



Name
Automatic software metadata validation report for hypodisc 0.1.0
Author
codemetapy validator using software.ttl
Date
2024-10-07 03:09:26
Review

```
Please consult the CLARIAH Software Metadata Requirements at https://github.com/CLARIAH/clariah-plus/blob/main/requirements/software-
metadata-requirements.md for an in-depth explanation of any found problems

Validation of hypodisc 0.1.0 was successful (score=3/5), but there are some warnings which should be addressed:

1. Info: Software source code *SHOULD* link to a continuous integration service that builds the software and runs the software's tests (This
is missing in the metadata)
2. Info: An interface type *SHOULD* be expressed: Software source code should define one or more target products that are the resulting
software applications offering specific interfaces (The metadata does express this currently, but something is wrong in the way it is
expressed. Is the type/class valid?)
3. Warning: Documentation *SHOULD* be expressed (This is missing in the metadata)
4. Info: Reference publications *SHOULD* be expressed, if any (This is missing in the metadata)
5. Info: The funder *SHOULD* be acknowledged (This is missing in the metadata)
6. Info: The technology readiness level *SHOULD* be expressed (This is missing in the metadata)
```

Rating
★ ★ ★ ☆ ☆

Figure 4: Screenshot of a validation report for a particular tool, viewed from the tool store

Using this report, developers can clearly identify what specific requirements they have not met. The over-all level of compliance is expressed on a simple scale of 0 (no compliance) to 5 (perfect compliance), and visualised as a coloured star rating in our interface. This evaluation score and the validation report itself becomes part of the delivered metadata and is something which both end users as well as other systems can filter on. It may even serve as a kind of 'gamification' element to spur on developers to provide higher quality metadata.

We find that human compliance remains the biggest hurdle and it is hard to get developers to provide metadata beyond what we can extract automatically from their existing sources. The metadata compliance rankings for CLARIAH are shown in Figure 5.

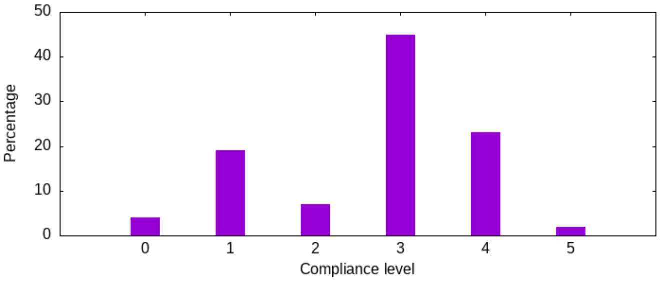


Figure 5: Histogram of metadata compliance ranking in CLARIAH (114 tools), the ranking is to the number of stars given (0 to 5, where 5 is perfect compliance)

For propagation to systems further downstream, we set a threshold rating of 3 or higher. Downstream systems may of course posit whatever criteria they want for inclusion, and may add human validation

[26] https://github.com/CLARIAH/clariah-plus/blob/main/requirements/software-metadata-requirements.md.

and curation. As metadata is stored at the source, however, we strongly recommend any curation efforts to be directly contributed back upstream to the source, through the established mechanisms in place by whatever forge (e.g. GitHub) they are using to store their source code.

7 Discussion & Related Work

We limit automatic metadata extraction to those fields and sources that we can extract fairly reliably and unambiguously. In certain cases, it is already a sufficient challenge to map certain existing vocabularies onto codemeta and schema.org, as concepts are not always used in the same manner and do not always map one-to-one.

We do extract certain information from README files, but that is mostly limited to badges which follow a very standard pattern that is easy to extract with simple regular expressions. Extracting more data from READMEs is something that was done in Kelley and Garijo, 2021 and predecessor Mao et al., 2019; they analyse the actual README text and extract metadata from it. They use various methods to do so, including building supervised classifiers to identify common section headers and mapping those to a metadata category such as 'description', 'installation', 'license', etc.... Their classifiers, however, only produced adequate results for four common categories, so they diverted to alternative methods such as exploration/detection of other files (like `LICENSE`), using regular expressions to capture badges, and calling APIs like GitHub's. All of those techniques we have implemented as well in our harvesting pipeline. In line with their findings, we did not expect much from supervised classification (measured against the effort that goes into labelling data) so did not pursue that.

With the advent of Large Language Models in recent years, we can also envision these playing a role in metadata extraction. We would, however, caution restraint here as their innate nature to hallucinate and lack of transparency is at odds with the objective to extract accurate metadata. Our extraction pipeline focusses on using relatively simple techniques to quickly get high precision results and on re-using already existing metadata schemes. We do think it is good practise to have developers manually provide metadata, we just want to ensure they only need to do it once alongside their own source-code, using schemas they use anyway, and not duplicate the effort for every software catalogue or package manager.

We also want to draw a quick line of comparison with the Research Software Directory (Cahen et al., 2024; Spaaks, 2018). This is an open-source content management system for software catalogues, so a different beast than our metadata extraction pipeline. They do, however, offer some integrations with third party services such as GitHub, Zenodo, ORCID, etc... to automatically extract or autocomplete certain metadata. It illustrates there are hybrid approaches possible where a content management system is available for human metadata curation, but with key parts automated to reduce both the human workload as well as the common pitfalls we addressed in section 2.

8 Conclusion & Future Work

We have shown a way to store metadata at the source and reuse existing metadata sources, recombining and converting these into a single unified LOD representation using largely established vocabularies. We developed tooling for codemeta that is generically reusable and available as free open source software[27]. We hope that our pipeline results in metadata that is accurate and complete enough for scholars to assess whether certain software is worth exploring for their research. We think this is a viable solution against metadata or entire catalogues going stale, in worst case unbeknownst to the researcher who might still rely on them. Quality assurance can be addressed, in part, via automated validations against carefully formulated validation rules. Furthermore, we also showed that the metadata we collect can be propagated to other downstream software catalogue systems.

Future work will focus on keeping in sync with vocabulary developments in CodeMeta and schema.org, as well as on working on the automatic propagation of harvested metadata into catalogue systems such as the SSHOC Open Marketplace.

[27] GNU General Public Licence v3

Acknowledgements

The FAIR Tool Discovery track has been developed as part of the CLARIAH-PLUS project (NWO grant 184.034.023), as part of the Shared Development Roadmap.

Connectivity to the SSHOC Open Marketplace is being continued in the SSHOC-NL project as part of Task 1.1.

References

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016). Tadirah: A case study in pragmatic classification. *Digit. Humanit. Q.*, *10*(1). http://dblp.uni-trier.de/db/journals/dhq/dhq10.html#BorekDPS16

Bradner, S. (1997). *IETF RFC 2119: Key words for use in RFCs to Indicate Requirement Levels*. http://www.ietf.org/rfc/rfc2119.txt

Cahen, E. J., Mijatovic, D., Garcia Gonzalez, J., Maassen, J., Jong, M., Meeßen, C., Rüster, M., Hanisch, M., & Ziegner, N. (2024). *Research Software Directory (as a service)*. Zenodo. https://doi.org/10.5281/ZENODO.14243099

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., Capella-Gutiérrez, S., Hong, N. P. C., Cook, M., Corpas, M., Flannery, M., García, L. J., Gelpi, J. L., Gladman, S. L., Goble, C. A., Ferreiro, M. G., González-Beltrán, A. N., Griffin, P., Grüning, B. A., . . . Crouch, S. (2018). Four simple recommendations to encourage best practices in research software. https://api.semanticscholar.org/CorpusID:214915242

Kelley, A., & Garijo, D. (2021). A Framework for Creating Knowledge Graphs of Scientific Software Metadata. *Quantitative Science Studies*, 1–37. https://doi.org/10.1162/qss_a_00167

Mao, A., Garijo, D., & Fakhraei, S. (2019). SoMEF: A Framework for Capturing Scientific Software Metadata from its Documentation. *2019 IEEE International Conference on Big Data (Big Data)*, 3032–3037. https://doi.org/10.1109/BigData47090.2019.9006447

Spaaks, J. (2018). *The Research Software Directory and how it promotes software citation* [Accessed: 2025-01]. https://blog.esciencecenter.nl/the-research-software-directory-and-how-it-promotes-software-citation-4bd2137a6b8

van Gompel, M. (2023). *codemeta-server* (Version 0.4.1). Zenodo. https://doi.org/10.5281/zenodo.10204020

van Gompel, M. (2024). *codemetapy* (Version 2.5.3). Zenodo. https://doi.org/10.5281/zenodo.11656553

van Gompel, M., de Boer, D., & Broeder, J. (2024). *codemeta-harvester* (Version 0.4.0). Zenodo. https://doi.org/10.5281/zenodo.11472618

van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Lrec*. European Language Resources Association. http://dblp.uni-trier.de/db/conf/lrec/lrec2010.html#UytvanckZBWG10

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/doi:10.1515/9783110767377

CLARIAH-EUS: A Strategic Network Helping Basque Country Researchers to Participate in European Research Infrastructures

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna,
Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga,
Jose Mari Arriola
HiTZ Center - Ixa
University of the Basque Country, Spain
jon.alkorta,aritz.farwell,joseba.fernandezdelanda,begona.altuna,
ainara.estarrona,mikel.iruskieta,xabier.arregi,xabier.goenaga,
josemaria.arriola@ehu.eus

Inma Hernáez
HiTZ Center - Aholab
University of the Basque Country
Spain
inma.hernaez@ehu.eus

David Lindemann
Diachronic Linguistics, Typology, and
the History of Basque Research Group
University of the Basque Country
Spain
david.lindemann@ehu.eus

Abstract

CLARIAH-EUS is a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN and DARIAH, Europe's leading digital research infrastructures for the humanities, arts, and social sciences. Focused on Basque or Basque culture-related research in these fields, CLARIAH-EUS offers scholars digital tools and resources. Distinct from other nodes, CLARIAH-EUS serves a language (Basque) rather than a specific territory, making the infrastructure transnational. This article outlines the rationale behind establishing CLARIAH-EUS, its development process, ongoing projects, and future plans.

1 Introduction

Academic research in fields driven by technology is often characterized by rapid change and the constant application of new methods. Disciplines traditionally less dependent on technology, however, have historically experienced more measured technological integration. Although it may be argued that this has generally been true for the humanities, arts, and social sciences, a "digital turn" over the past two decades is driving new modes of research and lines of inquiry in these areas. Digital tools, methods, and data are casting new light on complex social patterns and providing innovative techniques to interpret cultural heritage (Crawford et al., 2014; Terras, 2011). Moreover, the rise of digital humanities has blurred the boundaries between disciplines, spurring interdisciplinary collaborations in research, teaching, and publishing (Burdick et al., 2016).

Language technology, tools, and resources tailored specifically for the social sciences and humanities play a pivotal role in this pioneering work. Yet, much of this technological support is designed for use with English, creating an imbalance between techniques that can be applied when conducting English-language research and those that are available for research in other languages. This is especially true for languages spoken by smaller populations (Arzoz, 2015). Basque, one of these languages, has fortunately made significant strides in language technology due to deliberate efforts to foster the sociolinguistic conditions necessary for its successful development and dissemination. This includes sustained and proactive collaboration between research groups, foundations, industry clusters, and regional institutions (Gonzalez-Dios & Altuna, 2022; Sarasola et al., 2023). Nevertheless, Basque still faces challenges in

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna, Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga, Jose Mari Arriola, Inma Hernaez and David Lindemann 2024. CLARIAH-EUS: A Strategic Network Helping Basque Country Researchers to Participate in European Research Infrastructures. *Selected papers from the CLARIN Annual Conference 2024. Ed. by Vincent Vandeghinste and Thalassia Kontino.* Linköping Electronic Conference Proceedings 216, pp. 151–161. DOI: https://doi.org/10.3384/ecp216

terms of research maturity and availability of the wide-ranging resources needed to fully support social science and humanities projects.

The CLARIAH-EUS consortium was established to overcome these existing limitations. On the one hand, it seeks to encourage the use of language technology among researchers in Basque-related humanities, arts, and social sciences. On the other, it strives to strengthen and facilitate collaboration between these researchers, enabling them to share ideas and innovative approaches more effectively. To do so more effectively, CLARIAH-EUS took the strategic decision to orient itself towards language rather than geographical boundaries, making it transnational in scope. Furthermore, as a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN ERIC and DARIAH ERIC, CLARIAH-EUS is aligned with Europe's leading digital research platforms for the humanities, arts, and social sciences.

2 Objectives

As highlighted above, one of CLARIAH-EUS's objectives is to support language technology for Basque humanities, arts, and social sciences research. This effort translates into two key areas. The first is to build a repository that contains digital resources specifically for Basque. These resources will be integrated into the wider CLARIN and DARIAH infrastructures, ensuring that the Basque-focused tools that are developed become readily accessible to researchers. The second is to empower researchers by offering them dedicated services and training. We plan to provide users who are creating or utilizing Basque language technology for their projects with the resources to work autonomously in the digital domain.

By cultivating these two areas, CLARIAH-EUS intends to foment a vibrant research community that is dedicated to advancing Basque language technology for the humanities, arts, and social sciences. This focus on collaboration is designed to 1) open doors to greater participation in international projects by leveraging shared expertise to create more impactful outcomes and 2) nourish an environment that sparks groundbreaking approaches to Basque digital humanities and language technology.

3 Institutional Funding

CLARIAH-EUS prioritizes securing financial backing to ensure the viability of research initiatives across the short-, medium-, and long-term. This approach guarantees the ongoing usability and value of any resources that are created. The focus on sustainability has resonated with several public funding bodies, who have pledged support to the infrastructure. Currently, CLARIAH-EUS is supported by the Basque Government through its Department of Culture and Linguistic Policy,[1] the Provincial Council of Gipuzkoa,[2] and the University of the Basque Country (UPV/EHU). Backing by the UPV/EHU comes from the Vice-Rectorate of Basque, Culture and Internationalization,[3] and from HiTZ, the Basque Center for Language Technology.[4] HiTZ, in addition to providing financial support, also houses the infrastructure's administrative office. Furthermore, several of its members sit on the CLARIAH-EUS steering committee, contributing their guidance and expertise. Thanks to the support of these institutions, CLARIAH-EUS has assembled a team of four staff members, who play a crucial role in ensuring the smooth operation of both the CLARIAH-EUS infrastructure and the shared administrative office with CLARIAH-ES (also overseen by HiTZ).

4 Origins and Growth

To date, the evolution of CLARIAH-EUS has included a design phase (2021-2023) (see sections 4.1 and 4.2) and an implementation and consolidation phase (2023-present) (see sections 4.3 and 4.4).

4.1 First Workshop: Needs and Manifesto

CLARIAH-EUS's first workshop,[5] *Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen* (*Designing the CLARIAH-EUS infrastruc-*

[1] https://www.euskadi.eus/eusko-jaurlaritza/kultura-hizkuntza-politika-saila/

[2] https://www.gipuzkoa.eus/eu/

[3] https://www.ehu.eus/eu/web/nazioarteko-harremanak

[4] https://www.hitz.eus/eu

[5] https://www.clariah.eus/eu/1-workshop

ture to develop language technology for Basque in the humanities and social sciences) was organized by HiTZ on November 26, 2021 and laid the foundation for the future infrastructure.

The workshop aimed to foster discussion about opportunities and needs across various research areas related to Basque language and culture. It featured several activities: 1) the presentation of a collection of use cases and posters depicting digital projects relevant to Basque studies, which provided a platform for researchers to share their work; 2) collaborative breakout sessions focused on identifying the strategic resources most crucial for Basque research across different disciplines; and 3) engaging and building bridges between researchers that encouraged active participation in CLARIAH-EUS's future.

The event drew participants from nine institutions and thirty-four researchers representing twenty distinct research groups. Fourteen projects were presented and 134 organizations and individuals signed a manifesto.[6] This collective voice underscored the widespread demand for a dedicated digital humanities infrastructure for Basque research.

4.2 Assembling the Network

Between 2021 and 2023, CLARIAH-EUS's goal was to pursue the backing of several organizations and research groups. This was procured from seventeen entities: HiTZ (UPV/EHU), Basque Summer University (UEU),[7] Iker research group,[8] Elhuyar,[9] Gogo Elebiduna research group (UPV/EHU),[10] Elebilab research group (UPV/EHU),[11] Aholab research group (UPV/EHU),[12] Ixa research group (UPV/EHU),[13] Soziolinguistika Klusterra,[14] Diachronic Linguistics, Typology and the History of Basque (DLTB) research group (UPV/EHU),[15] Basque Research Group of Theoretical Linguistics (HiTT) (UPV/EHU),[16] Badalab,[17] Gizapedia,[18] Tralima-Itzulik research group (UPV/EHU),[19] General Directorate of Linguistic Equality (Prov. Council of Gipuzkoa),[20] the Public University of Navarre (UPNA),[21] and the UNESCO Chair in Human Rights and Public Powers (UPV/EHU).[22] Sixteen of these institutions and research groups are based in the southern Basque Country and one (Iker) is from the northern Basque Country (see Figure 1). During this time, CLARIAH-EUS's position within CLARIAH-ES in Spain, and CLARIN and DARIAH at the European level, was further solidified.

4.3 Second Workshop: Community and Organization

CLARIAH-EUS held its second workshop[23] in November 2023. We presented the CLARIAH-EUS infrastructure and existing Basque digital humanities projects. The workshop marked a significant milestone in the form of a kickoff ceremony for the founding members. The focus shifted from initial brainstorming to outlining the infrastructure's future. Key discussions centered on CLARIAH-EUS's organizational structure and a road map, which put an emphasis on strategic directions for the next five years. Two invited speakers shared their expertise and twenty-one research groups presented posters. A selection of these, along with details about the participating research groups, will be featured in a forthcoming publication, offering a valuable glimpse into the Basque digital humanities landscape.[24]

[6] https://www.clariah.eus/eu/manifestua

[7] https://www.ueu.eus/

[8] https://iker.cnrs.fr/?lang=en

[9] https://www.elhuyar.eus/en

[10] https://www.ehu.eus/HEB/

[11] https://www.ehu.eus/en/web/elebilab/

[12] https://aholab.ehu.eus/aholab/

[13] http://ixa.ehu.eus/

[14] https://soziolinguistika.eus/en/

[15] https://ekoizpen-zientifikoa.ehu.eus/grupos/24732/detalle?lang=en

[16] http://www.hittlinguistics.eus/en/

[17] https://badalab.eus/

[18] https://gizapedia.org/

[19] https://www.ehu.eus/en/web/tralimaitzulik/home

[20] https://www.gipuzkoa.eus/web/council

[21] https://www.unavarra.es/home

[22] http://katedraddhh.eus/en/

[23] https://www.donostiakultura.eus/eu/ikastaroak/clariah-eus-euskararako-ikerketa-azpiegitura-eraikitzen

[24] https://www.clariah.eus/eu/2-workshopa-azpiegitura-eraikitzen

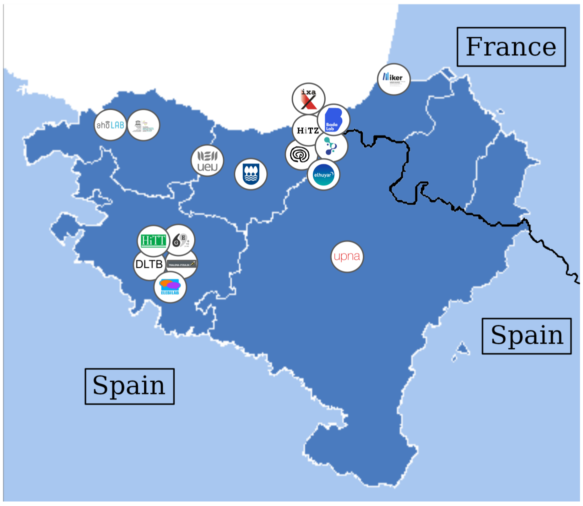


Figure 1: The CLARIAH-EUS Consortium

4.4 Third Workshop: A Vibrant Community, a Practical Infrastructure

For its third workshop,[25] which took place in November 2024, CLARIAH-EUS focused on strengthening the ties between those involved in local digital humanities projects and on highlighting ways the infrastructure can provide practical support. The workshop opened with a talk by invited speaker Mikko Tolonen, who shared Finland's experience in erecting FIN-CLARIAH and discussed how humanities-driven AI in the social sciences and humanities might be shaped. The event also included a roundtable discussion on technology and the social sciences, an introduction to CLARIAH-EUS's new B-centre, two presentations on CLARIN and DARIAH, a hands-on demonstration of Basque LLMs, and a poster session dedicated to digital humanities research currently being carried out in the Basque Country.

5 Projects, Tools, and Resources

As previously underscored, a core objective of CLARIAH-EUS is to empower researchers with the tools and resources[26] they need to excel in the digital humanities and social sciences. These resources fall into two categories: 1) resources that existed before CLARIAH-EUS was established, but which are now integrated into the network to maximize their reach and usability for the research community; and 2) newly developed resources created by CLARIAH-EUS members. The following section includes examples of both types, as well as a brief survey of several recent and ongoing projects that reflect current work being done on Basque-related topics.

5.1 Projects

5.1.1 CLARIAH-EUS-gArA

CLARIAH-EUS-gArA, funded by the Provincial Council of Gipuzkoa, is constructing a trustworthy conversational assistant for Basque news and research in the digital humanities. More specifically, the project explores how Latxa,[27] currently the largest and best performing LLM family for Basque, may be utilized in concert with retrieval-augmented generation (RAG) techniques to help guarantee the fidelity of responses to queries concerning news in Basque. The currently intrinsic tendency of LLMs to invent responses is an obstacle that CLARIAH-EUS-gArA also addresses. Accordingly, the project seeks to ensure that generated responses extracted from in-depth interactions with knowledge repositories include

[25] http://www.clariah.eus/eu/3.workshopa
[26] https://www.clariah.eus/eu/baliabideak_sailkapena
[27] https://huggingface.co/HiTZ/latxa-70b-v1.1

detailed annotations on the provenance of data so that researchers may fact-check, verify information, and assess reliability through accurate citations that provide direct access to original sources.

This ability to validate source material is not only a critical factor in restoring confidence in online content, but also a fundamental aspect of SSH research. To reach its objectives, CLARIAH-EUS-gArA is establishing methods to permanently crawl news as it is released and collect past news articles, offer synchronic and diachronic perspectives on events and opinions, and allow for multilingual comparison of Basque news with reporting in other languages. The technology developed for the project will be a step forward in this regard and for Basque DH in general because it will 1) create computational tools for Basque; 2) offer new means to browse news in Basque; 3) advance how Basque-language data are processed; and 4) support DH researchers in verifying sources and citations when searching for answers to complex research questions.

5.1.2 HarilkAI: Prototyping AI applications in the Social and Communication Sciences

HarilkAI's objective is to investigate the application of LLMs and AI tools to the field of teaching and education in the social and communication sciences.[28] To do so, a socio-technical infrastructure will be designed to experiment with AI tools and investigate how they may be better adapted to this area of study. HarilkAI addresses three main issues:

- the practice and epistemology of social and communication sciences. Participants will be asked to interact with AI tools, report on their effectiveness, and propose protocols, programs, and prototypes.
- digital sovereignty based on open-source philosophy and practices. This includes encouraging collaborative research and finding alternatives to proprietary software and platforms.
- the use of tools created from, and in, Basque within the context of LLMs.

Researchers and teachers from various research areas will be able to access HarilkAI's laboratory, which will adapt to their needs in an experimental and collaborative way. This includes developing software, implementing and configuring hardware, and constructing virtual spaces to test AI tools.

5.1.3 BIM/SAHCOBA

The projects BIM (*Basque in the Making: A Historical Look at a European Language Isolate*)[29] and SAHCOBA (*Syntactically Annotated Historical Corpus in Basque*) focused on developing a morphosyntactically annotated historical corpus of the Basque language (Estarrona et al., 2022). The interdisciplinary initiatives brought together experts in linguistics and natural language processing. BIM was dedicated to gathering key Basque texts from the fifteenth to mid-eighteenth centuries (Archaic and Old Basque), while the SAHCOBA project expanded this corpus to cover the period from the mid-eighteenth to the mid-twentieth centuries (Early and Late Modern Basque), coinciding with the emergence of standardized Basque. The corpus includes annotations for both parts of speech and syntax, along with extensive metadata. The database enables users to search the annotated corpus by word, lemma, grammatical category, sequences of grammatical categories, and specific syntactic structures. For example, the canonical order for negation in contemporary Basque is *ez da etorri* (has not come), that is ”negation particle (*ez*) + auxiliary verb (*da*) + verb in participle (*etorri*). But we know that in ancient texts the order *etorri ez da* appears. Thus, if we wanted to study the evolution of this syntactic structure over time, we could use the interface to search for the structure ”participle verb + negation particle (*ez*) + auxiliary verb” and we would retrieve examples of the type *etorri ez da*.

5.1.4 ZITERAUZI

Semantic web technologies (Linked Open Data) have enabled new opportunities for recording, exhibiting, and querying publication collections. The ZITERAUZI project (Astigarraga et al., 2025) aims to take advantage of this shift by creating a tool chain for citation extraction from scientific articles published

[28] http://www.clariah.eus/sites/default/files/posterrak/Poster_HarilkAI.pdf

[29] http://bim.ixa.eus/search

in Basque in order to track the use of the language within the academic sphere. Inguma, a Basque scientific production database, has served as a starting point. In collaboration with the Digital Humanities Center in Errenteria, the ZITERAUZI team enriched the scientific publication metadata on articles from the IkerGazte conference series currently found in Inguma by representing extracted citation relations. The project's objectives include representing these relationships in a directed graph. This proposed graph will serve as an infrastructure for different use cases, which together will advance the study of scientific production in Basque.

5.1.5 Historical Texts in Wiki-platforms as Linked Data

This project proposes a data model for storing Basque historical texts in a database following the Linked Data paradigm by utilizing Wikibase software as infrastructure (Lindemann & Alonso, 2025). On the one hand, the project models entities, describing corpus tokens and token spans on top of the preset Wikibase data model. On the other hand, it builds a Standard Basque dictionary, deploying the Ontolex-Lemon vocabulary on the same Wikibase instance. This model allows for linking tokens and token spans to the dictionary's entries on a lexical entry (lemma) level, lexical sense level, and inflected form level. At the same time, entities at these three levels may carry additional descriptions and other links. This extends a traditional way of morphosyntactic annotation with literal values towards linking dictionary elements as entities. In addition to this corpus-lexicon interface, other kinds of annotations are also modeled, such as philological and semantic annotations.

5.2 Tools and Resources

5.2.1 CORPErrore

CORPErrore is a resource that enriches the HABE-IXA corpus[30] by labeling errors in Basque over the corpus using the INCEpTION annotation tool (Klie et al., 2018). Through the CORPErrore website,[31] searches may be conducted by error or suberror, errors can be queried by tiers, and text searches may be performed.

5.2.2 ETEL

The ETEL system[32] is designed to analyze texts and obtain linguistic complexity measures in a simple way for researchers and research groups. ETEL allows research teams to collaborate on the same project with different user profiles. The system offers four main functionalities: 1) text analysis (linguistic complexity indicator selector and results visualizer), 2) text suggestion for selected text level, 3) corpus management tools (file manager and complexity-level manager), and 4) a user administrator. Various aspects of textual complexity may be analyzed using ETEL, including lexical phenomena (distinct words or lemmas), syllabic measurements, and PoS information, such as the number of verbal lemmas. ETEL's multifaceted analysis system compares analyzed text with the results of a pre-classified and analyzed corpus and graphically displays its position in relation to the corpus measurements already performed for each analyzed phenomenon.

5.2.3 IGARRITZ

IGARRITZ (Iruskieta et al., 2024) is an adapted web environment that employs AI techniques for Basque text prediction.[33] It is designed to facilitate the creation of texts in Basque for secondary school students with cerebral palsy. To achieve this goal, the project developed a web interface that is based on the HiTZ/roberta-eus-euscrawl-base-cased language model. This was retrained with an educational Basque corpus sourced from texts that appear in Gizapedia, Wikipedia, and the Basque newspaper *Berria*. Igarritz has been incorporated into the CLARIAH-EUS B-centre.

[30] https://b2share.eudat.eu/records/81433fddcd06405f8505c7606b29ff99

[31] https://corperrore.clariah.eus/

[32] https://etel.clariah.eus/etel/Analyzer

[33] https://igarritz.clariah.eus/

5.2.4 Contemporary Basque Student Handwritten Model

The Contemporary Basque Student Handwritten Model[34] is a Basque AI model in Transkribus designed to transcribe learners' handwriting in Basque. The dataset consists of school-based texts written by adolescent students aged 12–16. Original errors in the handwriting were preserved and transcribed verbatim. The model was trained on a corpus of 51,195 words in Basque, collected from various schools in the Basque Autonomous Community in 2023.

5.2.5 HiTZketan

Launched by HiTZ, HiTZketan (lit. *in conversation*) has developed a speech-to-speech translation system for Basque and Spanish. The system receives a speech signal in one of the two languages, translates it to the other, and then delivers this translation using speech with a personalized voice that imitates the original speaker. Accordingly, HiTZketan enables Basque-language support for state-of-the-art technologies in automatic speech recognition, machine translation, and personalized speech synthesis.[35]

5.2.6 Parlamint-ES-PV 4.0

ParlaMint 4.0 is a collection of comparable corpora[36] featuring transcripts of parliamentary discussions from twenty-nine European nations and autonomous regions, primarily spanning from 2015 to mid-2022. Each corpus contains between nine million and 126 million words, with the entire compilation exceeding 1.1 billion words. CLARIAH-EUS has developed the Basque and Spanish corpus (Alkorta & Iruskieta, 2022) using data and metadata sourced from the Basque Parliament.

5.2.7 Computational Social Science Resources

At least three datasets related to social media analysis are available for research and tool development in Basque: 1) the Heldugazte dataset,[37] which focuses on identifying the writing style of a given text sequence (Fernandez de Landa et al., 2019); 2) the Heldugazte-Age dataset,[38] designed to determine the age group of Basque social media users, classifying them as either minors or adults (Fernandez de Landa & Agerri, 2021); and 3) VaxxStance,[39] which analyzes social media posts to assess opinions on vaccines (Agerri et al., 2021). This dataset categorizes tweets based on their stance as AGAINST, IN FAVOR, or NEUTRAL regarding a predefined topic.

5.2.8 BERnaT: Modeling the Diversity of the Basque Language

BERnaT applies research done on linguistic diversity in the creation of language models to Basque.[40] Its objective is to develop Basque models that take into account the diversity of the language. By utilizing the Basque corpus EusCrawl as a foundation, along with the best-performing discriminative model (Artetxe et al., 2022), multiple Basque models have been created that incorporate varying levels of linguistic diversity. Four datasets were used to achieve this:

- EusCrawl, a clean and standardized Basque corpus.
- the Latxa corpus (Etxaniz et al., 2024), currently the largest available Basque corpus.
- a spontaneous corpus created specifically for this study, consisting of thousands of tweets from Basque users.
- a combined dataset integrating all the aforementioned corpora.

The newly trained models have been evaluated using the BasqueGLUE (Urbizu et al., 2022) benchmark. BERnaT aims to demonstrate that models trained with more diverse linguistic corpora can be beneficial for tasks involving more spontaneous language or greater linguistic variation.

[34] https://www.transkribus.org/model/contemporary-basque-student-handwritten
[35] http://www.clariah.eus/sites/default/files/posterrak/Hitzketan(2)Clariah.pdf
[36] https://www.clarin.si/repository/xmlui/handle/11356/1859
[37] https://github.com/joseba-fdl/heldugazte-corpus
[38] https://github.com/joseba-fdl/heldugazte-age-corpus
[39] https://vaxxstance.github.io/
[40] http://www.clariah.eus/sites/default/files/posterrak/Clariah-eus2024_posterra_bernat.pdf

5.2.9 C1 Automatic Evaluator for Basque

HiTZ has developed an automatic evaluator that determines whether Basque-language writings meet the C1 level or not. To develop the system, essays from candidates who took the HABE (Institute for Adult Literacy and Basque Relearning)[41] C1 exams were utilized, taking into account the grades assigned by examiners. Ten thousand automatically transcribed texts and around 600 manually transcribed compositions were obtained through an agreement between the IKERGAITU project[42] and HABE-HiTZ. The system is based on language models with a neural foundation. Experiments were conducted with different types of language models to identify the most suitable configuration: monolingual or multilingual, and encoder or decoder. Accuracy rate was utilized for automatic evaluation metrics. Results demonstrate that models based on the Latxa language model (Etxaniz et al., 2024) performed the best, with the top model achieving an overall accuracy rate of 79%. A demo has been created based on the most suitable system.[43]

5.2.10 Resources and Tools for the Development of High-Level Academic Texts

Researchers at HiTZ have developed various resources and tools to stabilize and promote the use of academic registers in Basque, without which the Basque linguistic community risks losing the specialized language that is essential within an academic context. These include: the Garaterm corpus,[44] the TZOS (Online System for the Service of Terminology)[45] terminological database, and the bilingual Academic Text Writing Support Tool (HARTA/TAILA).[46] These resources represent part of HiTZ's efforts to create a Basque work environment and foster collaboration through interoperable data sharing and a dynamic network of experts.

5.2.11 Children's Literature Corpus

The Basque Language Institute[47] and the research group Gogo Elebiduna (Bilingual Mind)[48] at the University of the Basque Country (UPV/EHU) have created a children's literature corpus to provide a resource for speech therapists, audiologists, and language teachers who are working with children with speech and language difficulties. In addition to the corpus, based on literature for children aged 0-8, the goal of the initiative is to develop a search system for utilizing the corpus, where professionals can select language materials according to various criteria.[49] The initial version of the corpus contains 428 books, categorized into two age groups (0-4 and 5-8), and a web interface that enables users to perform simple searches and phonemic queries. Future plans to expand the corpus include enriching it with a broader range of language materials, introducing visual media, and refining the search engine to allow for searches that encompass several phonological, lexical, semantic, morphosyntactic, and grammatical levels.

5.2.12 Basque PhD Abstracts and Abstracts Corpus Collection

The data collected in PhD theses completed at the University of the Basque Country, along with the tools created to analyze these data, are currently dispersed across university repositories, hindering their potential reuse. One of CLARIAH-EUS's objectives is to mitigate this issue by facilitating the searchability of thesis-produced data, enhancing their visibility and reusability for other researchers. To do so, CLARIAH-EUS aims to upload PhD abstract data into EuDat, subsequently making it accessible through CLARIN's VLO for wider utilization by researchers under the CC BY-NC 4.0 license. This exercise will include the development of novel methods and tools to generate and evaluate PhD abstracts.

[41] https://www.habe.euskadi.eus/hasiera/

[42] https://www.hitz.eus/iker-gaitu/

[43] https://huggingface.co/spaces/HiTZ/C1_sailkapen_demoa

[44] http://garaterm-corpusa.ixa.eus/

[45] https://tzos.ehu.es/?setuilang=eu

[46] https://harta.ixa.eus/

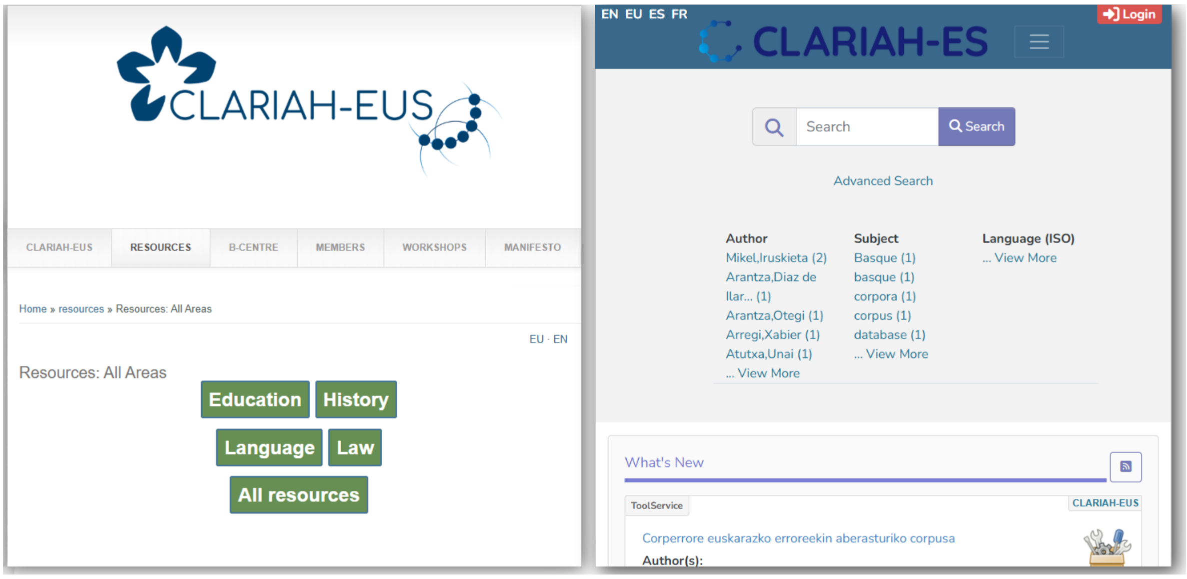
[47] https://www.ehu.eus/en/web/eins/

[48] https://www.ehu.eus/HEB/

[49] https://www.ehu.eus/ehg/08corpusa/

6 CLARIN B-Centre

The CLARIAH-EUS B-centre at HiTZ is part of the CLARIAH-ES infrastructure and, as such, will serve that larger community. Nonetheless, the need to establish a repository that can deliver digital language technology to those mainly engaged in Basque-language research and Basque Studies is a driving motivation behind the creation of the B-centre at HiTZ. Our hope is that the repository, currently in the final stages of construction, will help meet CLARIAH-EUS's desire to foster a collaborative space for those working with Basque or on Basque-related projects. With this in mind, the following brief survey of the CLARIAH-EUS B-centre highlights the initial steps taken in its ongoing construction.



Author
Mikel,Iruskieta (2)
Arantza,Diaz de Ilar... (1)
Arantza,Otegi (1)
Arregi,Xabier (1)
Atutxa,Unai (1)
... View More

Subject
Basque (1)
basque (1)
corpora (1)
corpus (1)
database (1)
... View More

Language (ISO)
... View More

Corperrore euskarazko erroreekin aberasturiko corpusa
Author(s):

Figure 2: CLARIAH-EUS B-Centre Webpages

6.1 Requirements

Several requirements must be fulfilled to be certified as a CLARIN B-centre and obtain compliance with CoreTrustSeal standards (CoreTrustSeal Standards and Certification Board, 2022). The latter provide a framework for building and maintaining trustworthy digital repositories that ensure the long-term accessibility and usability of digital data and metadata. HiTZ is currently working to put these guidelines and requirements in place. To begin with, we have defined our main objectives and mission in keeping with the goals of CLARIAH-EUS. Second, we have assigned staff to oversee the repository's construction and future management, with the expectation that we will hire and train new staff members in the near future. Qualified technicians from HiTZ are currently building the repository's infrastructure, housed at the University of the Basque Country's Department of Computer Science in San Sebastián, Spain. Third, and with respect to more technical questions, we are in the process of gathering information about data management, rights, legality, quality, suitability, and reliability in order to determine which measures and protocols are best suited to ensure CoreTrustSeal standards are met.

6.2 The B-Centre Server

The B-centre server's infrastructure must be configured to withstand accidental and temporary outages or crashes. This allows users and applications to continue operating without interruption and to access data and services. In our current setup, HiTZ has two servers available that we will configure in failover mode: one active and one passive, ready to take over if the first fails. However, to obtain a better High Availability (HA), we plan to integrate a third server for enhanced resilience and to strengthen our ability to withstand disruptions. Leveraging technology that links three servers, such as Ceph, can offer superior reliability compared to using only two servers.

A virtual environment manager that functions with infrastructures based on two servers must also be selected. Ideally, the VE should allow for straightforward management and, out of several options, we have chosen Proxmox. This choice ensures that all team members responsible for handling the group's infrastructure can easily manage the system. However, HiTZ has decided that one individual will be designated to oversee the process. After selecting the OS, we proceeded with the installation process and configured the cluster following these steps: 1) created the cluster and connected the servers; 2) configured Corosync to work with two nodes; 3) connected the servers with a straight cable at ten Gbps; 4) configured a zfs PoolStorage; 5) installed the guest machine VM (Rocky Linux); 6) implemented replication; and 7) enabled High Availability to migrate the guest VM on failure.

Each CLARIN B-centre is required to maintain data and software, necessitating the establishment of a repository. While various Digital Resource Managers are available, CLARIN does not mandate a specific choice. A common option among centers is the utilization of open source platforms like DSpace. We have chosen to utilize the CLARIN DSpace implementation[50] to meet CLARIN requirements and to facilitate connection to the Virtual Language Observatory (VLO). CLARIN DSpace 7[51] has been installed.

7 Future Steps for CLARIAH-EUS

In the near future, CLARIAH-EUS will complete the final phase in the implementation of its CLARIN B-centre, which will be added to its already operational CLARIN K-centre. This will allow us to offer technical services as well as valuable instructional guidance to researchers. As this process unfolds, three key criteria will guide CLARIAH-EUS's immediate development:

- **Building Resources**. CLARIAH-EUS will prioritize creating or adapting resources and services that are readily accessible to researchers through the CLARIAH-EUS node.
- **Strategic Focus**. The infrastructure will target resources and services that strategically address the needs of the Basque research community.
- **Collaboration**. CLARIAH-EUS will create or adapt resources and services that seamlessly integrated with CLARIN and DARIAH.

CLARIAH-EUS is focused on making an immediate impact by adapting existing resources and incorporating them into our B-centre. Specifically, we hope to include the Analhitza tool (Otegi et al., 2017), the Euscrawl system (Artetxe et al., 2022), and ParlaMint. Additionally, we plan to develop new resources and provide various corpora, including literature, historical texts, and social network data. Looking ahead, we intend to create tools and resources for sociology, journalism, literature, and history, ideally in alignment with GLAM-related initiatives.

Acknowledgments

We thank the Basque Government and its Department of Culture and Linguistic Policy, the Provincial Council of Gipuzkoa, the Vice-Rectorate of Basque, Culture and Internationalization at the University of the Basque Country (UPV/EHU), and the HiTZ center for their support, as well as the CLARIAH-ES Network (RED2022-134527-E), funded by the Spanish Ministry of Science, Innovation and Universities.

References

Agerri, R., Centeno, R., Espinosa, M., Fernandez de Landa, J., & Rodrigo, A. (2021). VaxxStance@IberLEF 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, *67*, 173–181. 10.26342/2021-67-15

Alkorta, J., & Iruskieta, M. (2022). Adding the Basque Parliament Corpus to ParlaMint Project. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 107–110.

[50] CLARIN DSpace implementation: https://github.com/ufal/clarin-dspace

[51] CLARIN DSpace installation https://github.com/ufal/clarin-dspace/wiki

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., & Soroa, A. (2022). Does Corpus Quality Really Matter for Low-Resource Languages? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7383–7390).

Arzoz, X. (2015). The Impact of Language Policy on Language Revitalization: The Case of the Basque Language. *Cultural and Linguistic Minorities in the Russian Federation and the European Union: Comparative Studies on Equality and Diversity*, 315–334.

Astigarraga, A., Lindemann, D., & Bidaguren, M. (2025). Ziterauzi: The tool chain for citation extraction from basque academic texts. *CLARIAH-EUS: Zientzia Sozialak eta Humanitate Digitalak gaur egun*.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2016). *Digital_humanities*. MIT Press.

CoreTrustSeal Standards and Certification Board. (2022, September). CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 Extended Guidance. https://doi.org/10.5281/zenodo.7051096

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*, *8*(0), 1663–1672.

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, *37*(2), 391–404.

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., & Soroa, A. (2024). Latxa: An open language model and evaluation suite for basque. https://arxiv.org/abs/2403.20266

Fernandez de Landa, J., & Agerri, R. (2021). Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, *0*(0), 1–15.

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, *10*(6).

Gonzalez-Dios, I., & Altuna, B. (2022). Natural Language Processing and Language Technologies for the Basque Language. *Cuadernos Europeos de Deusto*, (04), 203–230.

Iruskieta, M., de la Iglesia, I., Atutxa, U., & Ortiz, L. (2024). IGARRITZ: euskarazko testu iragarpenerako web ingurune egokitua. *Ekaia*, *Ale berezia*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Lindemann, D., & Alonso, M. (2025). Historical texts in wiki-platforms as linked data. *CLARIAH-EUS: Zientzia Sozialak eta Humanitate Digitalak gaur egun*.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., & Uria, L. (2017). ANALHITZA: A tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58), 77–84.

Sarasola, K., Aldabe, I., Díaz de Ilarraza, A., Estarrona, A., Farwell, A., Hernáez, I., & Navas, E. (2023). Language Report Basque. In *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 95–98). Springer.

Terras, M. (2011). Quantifying digital humanities. *UCL Centre for Digital Humanities*.

Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., & Soroa, A. (2022, June). BasqueGLUE: A natural language understanding benchmark for Basque. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1603–1612). European Language Resources Association.