



Selected papers from the
CLARIN Annual Conference 2020
Virtual edition



CLARIN

like m. it
way
selected lang
form fo
sel
reference
kno
workflow
neural
original
linguistic
parser
transcrip
de
ty

Selected Papers from the
CLARIN Annual Conference 2020

Virtual Event, 2020, 5-7 October

edited by Costanza Navarretta and Maria Eskevich



Front Cover Illustration:

Picture Composition by CLARIN ERIC

Licensed under Creative Commons Attribution 4.0 International:

<https://creativecommons.org/licenses/by/4.0/><https://creativecommons.org/licenses/by/4.0/>

Linköping Electronic Conference Proceedings

180

eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)

2021

ISBN 978-91-7929-609-4

Introduction

Franciska de Jong

Executive Director CLARIN ERIC
Universiteit Utrecht, The Netherlands
f.m.g.dejong@uu.nl

Costanza Navarretta

Programme Committee Chair
University of Copenhagen
Copenhagen, Denmark
costanza@hum.ku.dk

This volume presents the highlights of the 9th CLARIN Annual Conference 2020. The conference was held in the virtual format on 5th —7th October 2020 because of the COVID-19 pandemics. CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform for 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 driven by language data.

The infrastructure is run by CLARIN ERIC¹, a consortium of participating countries and institutes that since it was established in 2012 has grown in size considerably. Currently there are 21 member countries, 3 observers, and more than 100 associated research institutions who are all encouraged and supported to be represented at the annual conference which is meant to be a central event for CLARIN community and which 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 use meet, in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. The conference covers a wide range of topics, including the design, construction and operation of the CLARIN infrastructure, the data, tools and services that are or could be on offer, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure. The aim is to attract researchers from all the various SSH fields who work with language materials, i.e. the people who are the *raison d'être* for CLARIN. Early in 2020 a call² was issued for which 40 abstracts were submitted. The authors of the submissions to the main conference session represented 19 countries, all of them from CLARIN ERIC countries with the exception of one paper from Spain. A few papers were written in cooperation by authors from different countries and institutions, but the number of the cross-country submissions is lower than in the previous conference editions. This can be due to the fact that CLARIN members, as the rest of the world, have not been able to meet each other face to face in most of 2020. Moreover, we did not receive contributions outside Europe, which again could be a negative effect of the pandemic restrictions.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 40 submitted abstracts 36 submissions were accepted for presentation at the conference (acceptance rate 0.9). The submissions were grouped in the following subjects:

- Annotation and Visualization Tools
- Data Curation, Archives and Libraries
- Metadata and Legal Aspects
- Research Cases
- Repositories and Workflows
- Resources and Knowledge Centres for Language and AI Research

¹<http://www.clarin.eu>

²<https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2020>

The accepted contributions were published in the online Proceedings of the Conference³.

Following the well received student poster session that was part of the programme of the 2018-2019 editions of the CLARIN Annual Conference, a PhD-session was organised with 7 presentations by PhD-students. One of the PhD-presenters represented a non-CLARIN country. The abstracts of the student presentations were published in the online CLARIN 2020 Book of Abstracts⁴.

The 2020 edition of the CLARIN Annual Conference was shaped as an online event. The virtual format enabled us to share quality content with almost 500 registered participants, including attendants of previous editions as well as newbies with an interest in getting familiarised with what CLARIN is about. The conference programme contained both traditional conference elements, and novel items better suited for the virtual set-up:

- **Invited talk** by Dr. Antske Fokkens (Faculty of Humanities, Vrije Universiteit Amsterdam) gave a talk entitled “Language Technology & Hypothesis Testing”. In this talk, she has highlighted that both the quality and accessibility of language technology has drastically increased over the last decade. Generic language models and deep learning have led to impressive results and both models and code for creating and using them is often made available. As such, an increase of these technologies are seen to be used in industry and various research disciplines outside of computational linguistics. Despite sometimes impressive results, however, currently developed technologies are still far from perfect and much is still unknown about how well those models work for specific use cases. Eventually, she has argued for the importance of going back to the foundations and ground research in hypotheses, both for studying language technology itself as well as for applying it in other research domains.
- **Panel on Artificial Intelligence, Language Data and Research Infrastructures** moderated by Ben Verhoeven with the following experts:
 - Prof. Jan Hajic, full professor of Computational Linguistics and the deputy head of the Institute of Formal and Applied Linguistics at the School of Computer Science, Charles University in Prague;
 - Dr Vukosi Marivate, ABSA UP Chair of Data Science at the University of Pretoria;
 - Prof. Marie-Francine Moens, full professor at the Department of Computer Science at KU Leuven, Belgium; director of the Language Intelligence and Information Retrieval (LIIR) research lab, a member of the Human Computer Interaction group, and head of the Informatics section;
 - Prof. dr. Malvina Nissim, Professor in Computational Linguistics and Society at the University of Groningen, The Netherlands; coordinator of the Computational Linguistics Group of the Center for Language and Cognition Groningen.
- Three **Special Appetizers** during the lunch breaks
 - CLARIN Café: “This is CLARIN. How can we help you?” with the aim to give an overview of CLARIN in a nutshell.
 - Social Networking Lunch
 - Improbotics - Improvised Theatre Show
- **Sessions of accepted conference papers** were organised as moderator-led discussions, and followed by poster-style discussions during which session participants could visit the individual paper authors and engage into discussions.
- During the **CLARIN Student session**, PhD-students presented their work in progress. The aim of the session was to share the next generation of researchers supported by or contributing to the CLARIN infrastructure and enable them to receive feedback on their work from CLARIN experts.

³<https://office.clarin.eu/v/CE-2020-1738-CLARIN2020ConferenceProceedings.pdf>

⁴<https://www.clarin.eu/content/clarin2020-book-abstracts>

- The **CLARIN in the Classroom session** invited university lecturers who had used CLARIN resources, tools or services in their courses to present their experience and suggest future steps that could help facilitate and accelerate the further integration of CLARIN into university curricula. (The slides of both sessions can be found in the conference programme).
- The **CLARIN Bazaar** provided as usual an informal setting for conversations with CLARIN people and a space to showcase ongoing work and exchange ideas.
- Each day was finished a **wrap-up session** that combined both personal highlights of two experts in the field and an illustration by a professional sketch artist.

In addition, on the event page⁵ CLARIN published a rich set of materials related to the conference:

- The complete conference programme and most of the slides presented: <https://www.clarin.eu/content/programme-clarin-annual-conference-2020>
- Recordings of keynote, panel, and CLARIN Café that are available on the CLARIN YouTube channel: link to be added.

After the conference, the authors of the accepted papers and student submissions, as well as participants of the CLARIN in the Classroom session were invited to submit full versions of their papers to be considered for the post-conference proceedings volume. The papers were anonymously reviewed, each by three PC members. We received 27 (including 1 student paper and 1 paper by the group of lecturers) full length submissions, out of which 23 were accepted for this volume. All the main topics addressed at the conference are covered in the papers.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Maria Eskevich from CLARIN Office for her indispensable support in the process of preparing these proceedings, and our colleagues at the Linköping University Electronic Press, who have ensured that the digital publication of this volume came about smoothly. In order to support the programme chair and the programme committee in the organisation of reviewing and programme planning, a programme subcommittee was established starting from CLARIN 2020. With respect to the establishment of the programme subcommittee, it was decided that the programme chair from the preceding year's conference is one of members in order to ensure continuity from one year's conference to the following one. The members of the 2020 PC subcommittee were Eva Hajičová, Monica Monachini, Kiril Simov, Inguna Skadiņa, and Martin Wynne.

Members of the Programme Committee for the CLARIN Annual Conference 2020:

- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Koenraad De Smedt, University of Bergen, Norway
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Eva Hajičová, Charles University Prague, Czech Republic
- Martin Hennelly, South African Centre for Digital Language Resources, South Africa
- Erhard Hinrichs, University of Tübingen, Germany
- Marinos Ioannides, Cyprus University of Technology (CUT), Cyprus
- Nicolas Larrousse, Huma-Num, France

⁵<https://www.clarin.eu/event/2020/clarin-annual-conference-2020-virtual-event>

- Krister Lindén, University of Helsinki, Finland
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
- **Costanza Navarretta, University of Copenhagen, Denmark (Chair)**
- Jan Odijk, Utrecht University, The Netherlands
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- Eiríkur Rögnvaldsson, University of Iceland, Iceland
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Marko Tadič , University of Zagreb, Croatia
- Jurgita Vaičėnienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Kadri Vider, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewers of this volume:

- Iulianna van der Lek-Ciudin, CLARIN ERIC, The Netherlands
- Riccardo Del Gratta, ILC “A. Zampolli” CNR Pisa, Italy

Contents

Introduction	i
<i>Franciska de Jong and Costanza Navarretta</i>	
Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data	1
<i>Timofey Arkhangel'skiy, Hanna Hedeland and Aleksandr Riaposov</i>	
CMDI Explorer	8
<i>Denis Arnold, Ben Campbell, Thomas Eckart, Bernhard Fisseni, Thorsten Trippel and Claus Zinn</i>	
Signposts for CLARIN	16
<i>Denis Arnold, Bernhard Fisseni and Thorsten Trippel</i>	
Studying Emerging New Contexts for Museum Digitisations on Pinterest	24
<i>Axelsson, Daniel Holmer, Lars Ahrenberg and Arne Jönsson</i>	
“Tea for two”: the Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure	37
<i>Federico Boschetti, Riccardo Del Gratta, Monica Monachini, Marina Buzzoni, Paolo Monella and Roberto Rosselli Del Turco</i>	
Extending the CMDI Universe: Metadata for Bioinformatics Data	47
<i>Olaf Brandt, Holger Gauza, Steve Kaminski, Mario Trojan, Thorsten Trippel and Johannes Werner</i>	
Community-based Survey and Oral Archive Infrastructure in the Archivio Vi.Vo. Project	55
<i>Silvia Calamai, Niccolò Pretto, Maria Francesca Stamuli, Duccio Piccardi, Giovanni Candeo, Silvia Bianchi and Monica Monachini</i>	
A Two-OCR Engine Method for Digitized Swedish Newspapers	65
<i>Dana Dannélls, Lars Björk, Torsten Johansson and Ove Dirdal</i>	
PoetryLab as Infrastructure for the Analysis of Spanish Poetry	75
<i>Javier De La Rosa, Salvador Ros, Álvaro Pérez, Aitor Díaz, Laura Hernández, Mirella De Sisto and Elena González-Blanco</i>	
Contagious “Corona” Compounding by Journalists in a CLARIN Newspaper Monitor Corpus	83
<i>Koenraad De Smedt</i>	
Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources	93
<i>Hanna Hedeland</i>	
Integrating TEITOK and Kontext/PMLTQ at LINDAT	104

<i>Maarten Janssen</i>	
The CLARIN-DK Text Tonsorium <i>Bart Jongejan</i>	111
When Size Matters. Legal Perspective(s) on N-grams <i>Paweł Kamocki</i>	122
Sharing is Caring: a Legal Perspective on Sharing Language Data Containing Personal Data and the Division of Liability between Researchers and Research Organisations <i>Aleksei Kelli, Krister Lindén, Kadri Vider, Paweł Kamocki, Arvi Tavast, Ramūnas Birštonas, Gaabriel Tavits, Mari Keskküla, Penny Labropoulou, Irene Kull, Age Värvi, Merle Erikson, Andres Vutt and Silvia Calamai</i>	129
The Literary Irony in the Works of Juliusz Słowacki <i>Anna Medrzecka</i>	148
Digitizing University Libraries - Evolving from Full-Text Providers to CLARIN Contact Points on Campuses <i>Manfred Nölte and Martin Mehlberg</i>	155
Towards Semi-Automatic Analysis of Spontaneous Language for Dutch <i>Jan Odijk</i>	165
Stimulating Knowledge Exchange via Transnational Access – the ELEXIS Travel Grants as a Lexicographical Use Case <i>Sussi Olsen, Bolette Pedersen, Tanja Wissik, Anna Woldrich and Simon Krek</i>	176
An internationally FAIR Mediated Digital Discourse Corpus: Improving Knowledge on Reuse <i>Rachel Panckhurst and Francesca Frontini</i>	185
Complementing Static Scholarly Editions with Dynamic Research Platforms: Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) for the Wittgenstein Nachlass <i>Alois Pichler</i>	194
LABLASS and the BULGARIAN LABLING CORPUS for Teaching Linguistics <i>Velka Popova, Radostina Iglíkova and Krasimir Kordov</i>	208
A Pipeline for Manual Annotations of Risk Factor Mentions in the COVID-19 Open Research Dataset <i>Maria Skeppstedt, Magnus Ahltop, Gunnar Eriksson and Rickard Domeij</i>	214

Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data

Timofey Arkhangelskiy
QUEST

Universität Hamburg, Germany
timofey.arkhangelskiy
@uni-hamburg.de

Hanna Hedeland
QUEST

Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Aleksandr Riaposov
QUEST

Universität Hamburg, Germany
aleksandr.riaposov@uni-hamburg.de

Abstract

This paper presents the QUEST project and describes concepts and tools that are being developed within its framework. The goal of the project is to establish quality criteria and curation criteria for annotated audiovisual language data. Building on existing resources developed by the participating institutions earlier, QUEST also develops tools that could be used to facilitate and verify adherence to these criteria. An important focus of the project is making these tools accessible for researchers without substantial technical background and helping them produce high-quality data. The main tools we intend to provide are a questionnaire and automatic quality assurance for depositors of language resources, both developed as web applications. They are accompanied by a knowledge base, which will contain recommendations and descriptions of best practices established in the course of the project. Conceptually, we consider three main data maturity levels in order to decide on a suitable level of strictness of the quality assurance. This division has been introduced to avoid that a set of ideal quality criteria prevent researchers from depositing or even assessing their (legacy) data. The tools described in the paper are work in progress and are expected to be released by the end of the QUEST project in 2022.

1 Introduction

The QUEST¹ project is one of twelve projects funded by the German Federal Ministry of Education and Research across all disciplines with the aim of enhancing research data quality and re-use. As the full title, "Quest: Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data", suggests, the focus is on one particular resource type, for which reliable quality standards and curation criteria will be developed. The project, which runs from 2019 to 2022, was based on the existing cooperation within the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)² (Hedeland et al., 2018). The CKLD partners involved in the application were the Data Center for the Humanities (DCH)³ and the Department of Linguistics (IfL)⁴ (both Cologne), the Endangered Language Archive (ELAR)⁵ and the SOAS World Languages Institute

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html>

²<http://ckld.uni-koeln.de/>

³<https://dch.phil-fak.uni-koeln.de/>

⁴<https://ifl.phil-fak.uni-koeln.de/en/>

⁵<https://www.soas.ac.uk/elar/>

(SWLI)⁶ (both London), the Hamburg Centre for Language Corpora (HZSK)⁷ and the long-term project INEL⁸ (both Hamburg) and the Leibniz Centre General Linguistics⁹ (ZAS, Berlin). For the QUEST project, the CKLD members were joined by the German Sign Language Corpus project (DGS-Korpus)¹⁰ in Hamburg and the Archive for Spoken German (AGD)¹¹ at the Institute for German Language (IDS) in Mannheim, who brought in their respective expertise. With the focus on annotated audiovisual language data, the aim of the project is twofold. On the one hand, it is to develop generic quality criteria valid regardless of intended usage scenarios. On the other hand, it aims to establish specific curation criteria tailored to certain re-use scenarios related to individual disciplines and/or research methods. To enable researchers to adhere to such criteria, these must be both adequate and not conflicting with research. Additionally, there must be comprehensive support for researchers with little technical background in applying them to their data, which is another important part of the project's goals.

After a brief review of previous work in this area in section 2, we will describe the conceptual project work briefly in section 3 and focus on the development of the various parts of a quality assurance system in section 4.

2 Background

The conceptual parts of QUEST regarding the definition of criteria draw on the expertise gathered within all project members' institutions and other relevant organisations. For the implementation of the quality assurance system, previous efforts by the data centres AGD (the Archive for Spoken German) and the HZSK (the Hamburg Centre for Language Corpora), which are both CLARIN B Centres, play a major role. One such existing resource we build upon is the assessment guidelines for legacy data (Schmidt et al., 2013), which were developed to set minimal standards for data deposits and make decisions regarding data curation transparent. Both the AGD and the HZSK were curating deposited resources to make them comply with internal quality requirements necessary for the integration into digital infrastructure and software solutions provided by the centres. The curation of audiovisual language resources is however a very time-consuming task that at the same time requires an advanced understanding of this particular data type. The need to handle the increasing amount of incoming resources with more efficiency and transparency at the Hamburg Centre for Language Corpora led to the development of another resource relevant to the QUEST project, the HZSK Corpus Services (Hedeland and Ferger, 2020). The HZSK Corpus Services are a conceptual and technological framework for collaborative data curation and quality control, originally based on the EXMARaLDA system (Schmidt and Wörner, 2014) and the version control system Git¹² combined with the project management system Redmine¹³. The framework enabled efficient collaborative resource creation in the long-term project INEL, which is based on the HZSK technical infrastructure and expertise, and have been developed further within this context as Corpus Services¹⁴ and LAMA¹⁵.

Other relevant approaches not related to the QUEST project include what is referred to as the "Open Source analogy for research data curation"¹⁶ and applied in the collaborative workflows of the Cross-Linguistic Linked Data (CLLD) project (Forkel, 2015). The work on continuous quality control and reproducibility within the CONQUAIRE (Continuous quality control for research data to ensure reproducibility) project (Cimiano et al., 2015) focuses on other resource types than the ones central to the QUEST project, but the methods and technology in use are very similar. To some extent, the DoorKeeper functionality of the FLAT repository at the Max Plank Institute for Psycholinguistics (MPI) in Nijmegen

⁶<https://www.soas.ac.uk/world-languages-institute/>

⁷<https://corpora.uni-hamburg.de/hzsk/en>

⁸<https://www.slm.uni-hamburg.de/inel/>

⁹<https://www.leibniz-zas.de/en/>

¹⁰<https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html>

¹¹http://agd.ids-mannheim.de/index_en.shtml

¹²<https://git-scm.com/>

¹³<https://www.redmine.org>

¹⁴<https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services>

¹⁵<https://gitlab.rrz.uni-hamburg.de/corpus-services/lama>

¹⁶<https://clld.org/2015/02/03/open-source-research-data.html>

(Trilsbeek and Windhouwer, 2016), is also a relevant example of data quality assessment, though it is focused on archivability rather than content-related resource quality or reproducibility. Related to the increasing importance of the FAIR principles (Wilkinson and others, 2016) for research data management, several projects have developed and provided means of assessing the level of data FAIRness manually and/or automatically. A comprehensive overview of "resources to measure and improve FAIRness" can be found at the educational website of the FAIRsharing project, FAIRassist¹⁷. However, all of the approaches based on the FAIR principles are aimed at assessing research data in general and do not provide specific criteria for individual resource types and/or disciplines.

3 Resource Types, Data Formats and Data Maturity Levels

The aim of the QUEST project is not to standardize the creation of audiovisual language resources but rather to take stock of the existing heterogeneity and promote such standards and formats in use that lend themselves to (preferably automatic) quality control. This, to a certain degree, includes machine-understandability, which is crucial for true semantic interoperability between various formats, standards and conventions. Another important aim is to find means to implement functionality for quality assessment and control. A first step is however to review and describe variation in existing resources both regarding resource structure, i.e. all relevant, partly abstract, data types and relations, and also regarding resource content. On the content level, the various file formats and data models in turn come with different macro-structures of tiers and speaker contributions and for one single file format or data model there are also possibly a wide range of different micro-structures based on various annotation schemes and transcription conventions (Schmidt, 2011). Following an inventory of QUEST associated and other relevant (CLARIN) data centres, an initial set of linguistically relevant data types based on their role within a resource was defined as the basis for meaningful recommendations on file formats. This set includes audio and video recordings, transcription/annotation data, lexical databases, additional relevant written or image material, contextual (meta)data on sessions and participants, documentation, catalogue and detailed metadata, and settings files. For generic data types such as audio, video, image and unstructured text files used for documentation there is little controversy regarding good practices for archival formats¹⁸. However, for the file formats used for transcription/annotation data and contextual data, the situation is far more complex. Schmidt (2011) provides a comprehensive overview still valid today.

While a few widely used and interoperable formats (such as ELAN (Sloetjes, 2014) or EXMARaLDA) are accepted across all centres, the level of structuredness, machine-readability and comprehensibility of resources created with these formats differs widely. This depends to a large extent on the research methods employed, especially as qualitative approaches do not rely on machine-readable data. While the original research might not profit from structured and machine-understandable data, discoverability and the options for future re-use scenarios depend on these aspects. Reliable preservation including possible migration into future file formats are further reasons for the aim to curate all deposits at the AGD and HZSK centres. Data curation is however a very costly endeavour and only partly possible for orphaned resources, and at the same time the numbers of deposits are growing steadily due to funders' recent requirements on researchers. Thorough curation within a reasonable time frame for the publication of deposits will thus become impossible without increasing the number of employed staff members accordingly, which is not an option. Still quality assessment and documentation are necessary to comply with the Core Trust Seal requirements (CoreTrustSeal Standards and Certification Board, 2019), which all certified CLARIN B service centres have to do. This is described by the requirement on data quality: "R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.". The requirement further states: "Data, or associated metadata, may have quality issues relevant to their research value, but this does not preclude their use if a user can make a well-informed decision on their suitability through provided documentation.". To achieve more transparency regarding the suitability of

¹⁷<https://fairassist.org>

¹⁸This applies for archives focusing on linguistic data, while "true" audio-visual archives recommend uncompressed formats of little relevance to the research context of digital language resources and linguistic fieldwork

resources than a simple distinction between curated and non-curated resources, a division in three main data maturity levels was developed. The aim of data curation would then rather be to comply with the next possible level, even if this is not enough for full integration into digital infrastructure solutions and services. This approach also allows researchers to comply with well-defined quality criteria for a lower level of data maturity instead of failing when strict quality criteria become obstacles preventing deposits of valuable data. There are three data maturity levels, which will be refined throughout the project's funding period, and in particular the names should be considered placeholders, since there are of course other descriptions or definitions of these terms. The three levels are:

- "deposits", which only feature a minimal set of legal, administrative and descriptive metadata and need not fulfill any criteria regarding the resource content,
- "collections", with additional requirements regarding resource structure, i.e. that the included data types and relations between all individual objects are described and consistent, and
- "corpora", in which the requirements on structure and consistency also pertain to the resource content, i.e. the contents of individual transcription/annotation files such as the use of tier types, annotation schemes and transcription conventions, but also contextual data such as valid participant identities across the resource.

Based on this division, an adequate evaluation of the resource quality becomes possible.

4 Quality Control for QUEST Data Centres and Users

For the implementation of quality control functionality within QUEST, data quality requirements will be harmonized across centres where possible. However, the main goal is to create an adjustable common diagnostic framework compatible with varying requirements, while also including existing validation functionality for e.g. EXMARaLDA and ELAN resources.

4.1 A Planning and Evaluation Tool for Depositors

The depositors' questionnaire (Schmidt et al., 2013) was originally developed at the HZSK and the AGD as a generic initial checklist for deposits and possible curation of legacy data. It has to be somewhat enhanced so that it can be used as a pre-ingest or pre-evaluation checklist at the participating centres. The questionnaire also has to be adapted to the data maturity levels defined within QUEST and to accommodate further information required to perform automatic quality assessment.

The content of the depositors' questionnaire was migrated to a new technical solution and partly extended according to the QUEST context. The questionnaire is now implemented as a web application and serves as the initial step of the quality control pipeline.

Unlike the original questionnaire, the updated one can be used in two scenarios, which contain different sequences of questions. In the first scenario, the user is planning a project and does not have the actual data at hand. In this case, they answer questions regarding their prospective data, e.g. whether they are going to have morphological annotation. At the end, the questionnaire generates templates for transcription or annotation files tailored to the user's needs that can be used throughout the project. At the moment, supported formats are ELAN template files and EXMARaLDA stylesheets. Both can be used for creating new empty transcription or annotation files in the respective software. This ensures that the data will have consistent annotation (e.g. consistent tier structure representing the annotation layers), thus reducing curation workload after the project is complete. In this scenario, the questionnaire app has overlapping functionality with data management planning software (i.e. making the user think about their data in advance and ensure its reusability).

In the second scenario, it is assumed the data has already been collected and processed, and the user would like either to deposit it to a QUEST center, or just to make sure it conforms to the quality requirements as defined by QUEST. In this scenario, the distinction between the three data maturity levels described in 3 is made. Depending on the data maturity level selected by the depositor at the beginning of the questionnaire, some of the questions may be skipped. If the user's responses indicate problems

that prevent their data from undergoing further quality control, such as lack of informed consent, the questionnaire app lists them together with the tips that could help resolve them. If no such problem is found, the user receives a machine-readable settings file with the summary of their responses and control settings, which can later be submitted to the second stage of quality control. These settings turn certain checks on or off, as well as provide parameter values (such as transcription tier name) to some checks.

Scenario	Initial stage of the project	Final stage of the project
Function	Planning tool	Questionnaire for data providers
Goal	Help to organize the project	Find potential issues with the data
Result	Templates and schemas for data and metadata	Settings for quality control tool

Figure 1: Planning and evaluation tool

4.2 A Flexible Quality Control Framework

There are two main directions in which the HZSK Corpus Services framework is extended in order to make it more universally applicable.

First direction is the usability. Corpus Services are a Java application that can only be run from the terminal; additionally, the user has to pass numerous arguments to switch particular tests on or off. In projects working with the software, this is done by using batch scripts customized for individual resources. Since this is beyond limits to most ordinary linguists, a web application was developed to make the testing process accessible to a wider audience. The front end is a web page that allows the user to upload an archive with the corpus to be tested, along with settings. In order to achieve smoother user experience and add extra flexibility for more advanced users, settings parameters may be passed to the server in three different ways:

- Upload a settings file generated by the questionnaire (section 4.1);
- Automatically generate settings based on the last valid questionnaire input received earlier in the session. If no settings data are available, i.e. the user has not completed the questionnaire before trying to upload their files for testing, this option is turned off;
- Manually choose the checks to be performed from a list of validators. This option is advantageous for users who want better control over the testing process (e.g., it can be selected to run only one specific check on the data for time-saving purposes).

The back end unpacks the archive in a temporary folder on the server and runs Corpus Services with arguments defined in the settings file. After the test is complete (which may take minutes or even hours), the corpus files are removed from the server. The HTML report generated by Corpus Services is then sent to the user via email. It can also be accessed afterwards on the server through a unique URL generated at upload time and shown to the user. Although this solution cannot be applied to corpora that are too large to be uploaded, we believe it will still cover the majority of cases.

Second, the contents of the framework is extended according to the QUEST context, since currently, only EXMARaLDA data can be validated. First and foremost, this means adding the ability to process the EAF format of the ELAN software used by the centres in London, Cologne and Berlin, and preferably also the FOLKER format used at the Archive for Spoken German and, possibly, other formats.

Also, many more checks/services should be added for generic and specific criteria developed within the QUEST project. This part of the extension is in its initial stage now.

The Quality Control Framework currently supports the following types of checks implemented in the Corpus Services:

- XML validators compare the files against the relevant XSD schema and XSLT stylesheet;
- coverage validators check if the links and references found in the data lead to existing files;
- structure validators inspect the files for anomalies (e.g. empty events in the transcription);
- string validators look for forbidden symbols/characters or problematic information such as absolute paths and other user-related information in the data;
- file and tier naming validators issue warnings if there are mismatches between names of the same entity or if the naming does not abide by a specific convention;
- segmentation validators check if the transcription data can be divided into linguistic segments and tokenized according to the transcription conventions;
- annotation validators examine the files for existence of overlapping annotations of the same type and compatibility with the annotation scheme.

More information on the checks is available in the Corpus Services documentation¹⁹. Further extension of validators included in the framework is schematized in figure 2 below:

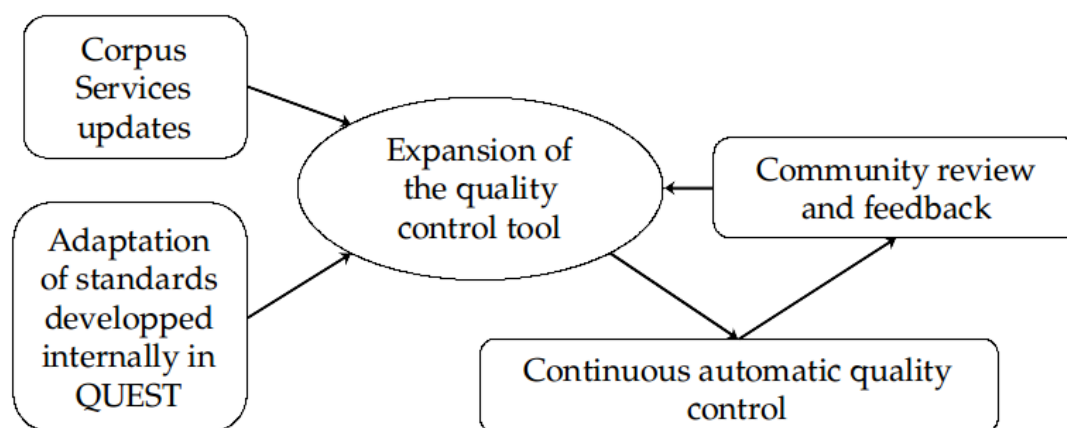


Figure 2: Development of the framework

4.3 A Common Knowledge Base

In order to facilitate adherence to the quality criteria established in QUEST, they should be formulated as simple instructions, recommendations and explanations accessible to an ordinary linguist. This is why a Knowledge base was added to the QUEST web services. Its purpose is to contain such recommendations, as well as definitions of the notions used in the questionnaire and Corpus Services reports, such as resource classification (section 3). The knowledge base is multilingual by design; ideally, all texts should be available in major lingua francas alongside English. The texts are stored in reStructuredText format²⁰, which makes it easy to track changes in version control and generate output HTML files.

¹⁹<https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services>

²⁰<https://docutils.sourceforge.io/rst.html>

5 Discussion

Since common widely accepted best practices and support in adhering to them are still lacking for researchers working with audiovisual language data, the work within the QUEST project can hopefully gain impact and applicability beyond original QUEST centres through the CLARIN Knowledge Sharing Infrastructure connection. It could also provide valuable input for the creation of Domain Data Protocols for audiovisual annotated language resources as suggested by Science Europe (Science Europe, 2018), which might be a way of providing quality criteria to users in a transparent and applicable manner.

The experiences with continuous quality control within the INEL project imply that without the Corpus Services and the staff members specifically responsible for their development (and for the support of non-technical staff using the software), the output of the project would not have been achieved in terms of data quality and quantity. By adapting the existing Corpus Services for the requirements of further QUEST partners and improving overall usability, the benefits of continuous quality control would become available for other projects. Providing various diagnostic tests for audiovisual resources that can be used at deposit but also during resource creation to external projects will allow these to prepare for data deposit and make this process more transparent, resulting in more high quality resources becoming available for interdisciplinary re-use within existing and emerging digital research infrastructures for the humanities and social sciences.

References

- Philipp Cimiano, John McCrae, Najko Jahn, Christian Pietsch, Jochen Schirrwagen, Johanna Vompras, and Cord Wiljes. 2015. CONQUAIRE: Continuous quality control for research data to ensure reproducibility: an institutional approach, September.
- CoreTrustSeal Standards and Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022, November.
- Robert Forkel. 2015. Cross-Linguistic Linked Data: Dateninfrastruktur für Diversity Linguistics. In *Forschungsdaten in den Geisteswissenschaften (FORGE) 2015, (Hamburg, 5-18 September, 2015)*, pages 10–12, Hamburg.
- Hanna Hedeland and Anne Ferger. 2020. Towards continuous quality control for spoken language corpora. *International Journal for Digital Curation*, 15(1).
- Hanna Hedeland, Timm Lehmberg, Felix Rau, Sophie Salfner, Mandana Seyfeddinipur, and Andreas Witt. 2018. Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation. In Nicoletta Calzolari et al., editors, *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 2340 – 2343, Paris, France. European language resources association (ELRA).
- Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.
- Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.
- Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.
- Science Europe. 2018. Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management, January.
- Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.
- Paul Trilsbeek and Menzo Windhouwer. 2016. FLAT: A CLARIN-compatible repository solution based on Fedora Commons. In *Proceedings of the CLARIN Annual Conference 2016*. CLARIN ERIC.
- Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.

CMDI Explorer

Denis Arnold

Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
arnold@ids-mannheim.de

Ben Campbell

Eberhard Karls Universität
Tübingen, Germany
ben.campbell@uni-tuebingen.de

Thomas Eckart

Universität Leipzig
Leipzig, Germany
teckart@informatik.uni-leipzig.de

Bernhard Fisseni

Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
fisseni@ids-mannheim.de

Thorsten Trippel

Eberhard Karls Universität
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Claus Zinn

Eberhard Karls Universität
Tübingen, Germany
claus.zinn@uni-tuebingen.de

Abstract

We present CMDI Explorer, a tool that empowers users to easily explore the contents of complex CMDI records and to process selected parts of them with little effort. The tool allows users, for instance, to analyse virtual collections represented by CMDI records, and to send collection items to other CLARIN services such as the Switchboard for subsequent processing. CMDI Explorer hence adds functionality that many users felt was lacking from the CLARIN tool space.

1 Motivation

A scientific resource often comprises many different parts. A proper description of such a resource with metadata according to CLARIN standards will yield a rich metadata record that lists each of the significant parts with detailed information. Consider the following example. A doctoral project that investigates the acquisition of language in small children might involve a number of experiments where babies are exposed to various visual and auditive stimuli, where eye tracking and other sensor data is used to observe their reactions, and where various Python and R scripts are employed to manipulate and analyse such data automatically. To describe such study, the doctoral candidate will attach, for instance, the media type to each stimulus, describe the nature of the sensor data, or refer to each of the processing scripts and the order they need to be executed. Rich metadata makes it easier for others to follow-up on research, say, when trying to reproduce research results, or to build a follow-up project on existing work, say by conducting a meta-study where the work of our doctoral student is taken to be one of many similar studies. A proper description of the meta-study, in turn, will yield a yet more complex metadata record, now describing the meta-study and how it has used the individual studies in the amalgamation.

Reading and processing complex metadata is no trivial matter. In this paper, we propose a tool that supports researchers in working with highly structured metadata and its associated research data.

2 Background

With CMDI being the de-facto standard for metadata in the CLARIN world, our community has built a good range of tools that process, in some way or another, CMDI metadata (Broeder et al., 2012).

The CLARIN *Virtual Language Observatory* (VLO; <https://vlo.clarin.eu>) gives researchers access to hundreds of thousands of language-related resources via their metadata descriptions (Van Uytvanck et al., 2012). At regular intervals, its back-end engine harvests CMDI-based metadata from many different metadata providers. It needs to analyse these CMDI records, which adhere to many

This work is licenced under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>

different metadata profiles, for content to correctly fill the various facets (*e.g.*, *language*, *resource type*, *modality*, *format*) that users will use to conduct faceted search in the VLO front-end. Once users have found a resource of interest, its individual metadata records are presented in a tabbed user interface, which includes a listing of its constituent parts via a simple hierarchical representation (*cf.* Fig. 1). Many descriptions in the VLO, however, are highly structured CMDI records. Navigating such metadata in such a tabbed environment, where a tabular entry points to a complex structure, involves following persistent identifiers attached to substructures *manually*. Hence, it can take some time to identify a sub-tree's leaves where, say, the auditive stimuli of a study of interest can be found.



Figure 1: A simple structured CMDI record in the VLO.

The CLARIN *Virtual Collection Registry* (VCR; <https://collections.clarin.eu>) enables scientists to assemble resources of interest into a virtual collection via persistent identifiers (PIDs) that refer to their individual metadata (Elbers, 2017). With virtual collections themselves being referred to by PIDs, scientists can easily create collections that have references to simple elements (such as single publications), and to complex elements (such as other virtual collections). It is hence easily possible to construct highly-structured virtual collections. The VCR is the primary entry point for scholars to create new virtual collections, share them with others, and browse through shared collection records. The portal offers some basic search functionality; it also provides a lean presentation of associated resources (*cf.* Fig. 2), which however, fails short at mirroring the potentially hierarchical structure of a collection.

The CLARIN *Language Resource Switchboard* (<https://switchboard.clarin.eu>) makes it easy for users to identify and invoke software tools that can process a language-related resource in one way or another (Zinn, 2018). The Switchboard's tool space, however, is geared towards, so to speak, the leaves of CMDI record trees, the actual scientific resources such as their text or audio files. The Switchboard cannot, for instance, handle a CMDI file that, say, describes a *plain* set of text files, which users will want to batch-process one by one with the same chosen tool.

Both the VLO and the VCR stop analysing CMDI files when it comes to resolving hierarchical structures marked by `ResourceProxyLists` of type `Metadata`. Unaware of the deep hierarchical structures behind a CMDI file, both VCR and VLO fail to offer users the crucial capability to navigate through them. Hence, users cannot easily explore those structures, select parts of them, say, to download them for off-line processing, or to send them to the Switchboard for a further analysis.

Resources		
Type	Reference	Actions
Resource	Peer Gynt (in English) An untraced edition, apparently from 1875.	...
Resource	HDL 2027/njp.32101068574639	...
Resource	HDL 2027/wu.89035754027	...
Resource	HDL 2027/uc1.32106002253281	...
Resource	HDL 2027/mdp.39015005058386	...
Resource	HDL 11245/1.146820 Dutch PhD thesis (with English abstract) by R.G.C. van der Zalm	...
Resource	HDL 1874/250346 Dutch BA thesis by L.G. de Jong	...

Figure 2: A simple structured CMDI record in the VCR.

Other tools face similar deficits. The CLARIN community offers a number of converters from CMDI to bibliographic metadata standards such as Dublin Core or MARC 21 (Zinn et al., 2016). These converters either show similar shortcomings when it comes to processing highly structured CMDI files, or are tailored to specific CMDI profiles where hence structural complexities are known in advance.

There are a couple of tools that go the extra mile of processing highly structured CMDI files: *SMC Browser* (Đurčo, 2013), see <https://clarin.oeaw.ac.at/smc-browser/index.html>) and *Curation Module* (King et al., 2016; Ostojic et al., 2017), see <https://curate.acdh.oeaw.ac.at/>) Both applications focus on a computer-assisted quality assessment of CMDI files. The Curation Module aims at providing statistical information relevant to the evaluation of the usability of a metadata instance. This includes link resolution checks and an evaluation of a record's adequacy for faceted search in the VLO. The focus on quality assurance makes these tools mostly suited for metadata creators and publishers, but not for the general user.

In sum, the VLO, VCR, and the Switchboard would profit from software that crosses navigational boundaries. The CLARIN *CMDI Explorer* aims at complementing (and supporting) the CLARIN tool space with the much-needed functionality of handling complex CMDI metadata. It provides a simple way of accessing all data files associated with a resource described by a CMDI metadata file. For this, it accesses the `ResourceProxyList` of a (possibly recursive) CMDI file and provides a navigable tree overview of all files associated with a collection. Each individual file can then either be downloaded directly, or it can be send to the Switchboard for further processing. *CMDI Explorer* also allows users to download all data files (or a selection thereof), depending on license restriction, for off-line usage.

3 CMDI Explorer

Consider the fragment of a CMDI file given in Fig. 3. The fragment has been taken from a CMDI file that describes data associated with a PhD dissertation¹ (Dima, 2019), and which will be our running example for the remaining part of this paper. Let us consider the nine `ResourceProxy` children in the fragment in more detail. There are eight resources of type `Metadata` and one resource of type `LandingPage`. The first child contains a self-reference to the CMDI file itself while the second child

¹See <http://hdl.handle.net/11022/0000-0007-CFE2-1>.

```

<ResourceProxyList>
  <ResourceProxy id="metadata0000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFE2-1@CMDI.xml</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="landingpage0000-0007-CFE2-1">
    <ResourceType>LandingPage</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFE2-1</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res10000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFD8-D</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res20000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFD9-C</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res30000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFD7-E</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res40000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFDC-9</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res50000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFDB-A</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res60000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFDA-B</ResourceRef>
  </ResourceProxy>
  <ResourceProxy id="Res70000-0007-CFE2-1">
    <ResourceType mimeType="application/x-cmdi+xml">Metadata</ResourceType>
    <ResourceRef>https://hdl.handle.net/11022/0000-0007-CFD6-F</ResourceRef>
  </ResourceProxy>
</ResourceProxyList>

```

Figure 3: A CMDI fragment, where references to resources are listed.

informs readers that the metadata description of Dima’s PhD has a landing page. Readers can copy the respective handle and paste it in their browser to view the landing page’s content. In our case, the landing page points to the TALAR research data repository. Here, users can enjoy to read the CMDI file in a more user-friendly manner. The remaining seven `ResourceProxy` children refer to the actual research data sets used in Dima’s thesis. Users need to navigate to another fragment of Dima’s CMDI file (`ResourceProxyListInfo`) to get more, albeit sparse, information about these data objects, see Fig. 4. To obtain more detailed information about each dataset, users will need to navigate back to the first fragment (see Fig. 3) to take the handle given in the `ResourceProxyList` and resolve it in the browser (which takes care of multiple redirects). In our running example, each handle points to the TALAR repository that holds the corresponding dataset. In the web interface of TALAR showing the

```

<cmdp:ResourceProxyListInfo>
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res10000-0007-CFE2-1">
    <cmdp:ResProxItemName>German Noun-Noun Compounds Dataset for Compositionality Tests</cmdp:ResProxItemName>
    <cmdp:ResProxFileName>deu-comp-nn-only</cmdp:ResProxFileName>
  </cmdp:ResourceProxyInfo>
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res20000-0007-CFE2-1"> [3 lines]
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res30000-0007-CFE2-1"> [3 lines]
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res40000-0007-CFE2-1"> [3 lines]
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res50000-0007-CFE2-1"> [3 lines]
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res60000-0007-CFE2-1"> [3 lines]
  <cmdp:ResourceProxyInfo xmlns:ns1="http://www.clarin.eu/cmd/1" ns1:ref="Res70000-0007-CFE2-1"> [4 lines]
</cmdp:ResourceProxyListInfo>

```

Figure 4: A CMDI fragment, where resource references are further described.

dataset’s description, users can then, if interested, navigate to the section “Data files” to download the data streams associated with the data set one by one. Users will then need to repeat the process for the other six datasets, which is a rather tedious enterprise.

In the given example, this enterprise is tedious but relatively smooth: each handle associated with a dataset resolves to the metadata resource that is nicely displayed in the TALAR user interface; and each time, users only need to navigate through the same user-friendly interface to get download access to the research datasets. In general, datasets may refer to landing pages that are hosted on different servers and where landing pages may then differ significantly from each other, hence making it harder for the user to download the datasets of interest.

CMDI Explorer aims at giving users easy access to such complex collections of datasets. Our new software automatically follows the `ResourceProxyList` children of a CMDI metadata file, accumulates relevant information, and presents the resulting tree structure to users. Users can then navigate through the tree structure to identify and download research data they find relevant.

In line with the other pillars of the CLARIN infrastructure, CMDI Explorer is implemented as a web-based application. Similar to the Switchboard, users have three options to enter their CMDI metadata: they can enter a PID that resolves to the metadata, upload a metadata file from their local computer, or copy and paste metadata they have found elsewhere. Similar to the Switchboard, CMDI Explorer will have an open communication channel with the VLO and the VCR. That is, VLO and VCR users will get the option to send a CMDI file to CMDI Explorer for further analysis.

CMDI Explorer is designed to analyse highly recursive CMDI files and to provide a tree-based representation and visualisation of its entire structure, both in the browser and as a downloadable HTML and JSON file. CMDI Explorer can also work with resources composed of constituents that are described using different CMDI profiles: As it only relies on the contents of `ResourceProxyList`, it will just assemble all files referenced there irrespectively of their role in the CMDI file. As it is designed to work with collections, CMDI Explorer can operate across repositories, provided the metadata is freely available, *i.e.*, not behind an Authentication and Authorization Infrastructure (AAI) wall.

The implementation CMDI Explorer uses `Java` for the back-end and `react-js` for the front-end.

Back-end. The algorithm for retrieving the data associated with a collection is as follows: first, the CMDI XML data associated with the PID of the collection is retrieved. The `ResourceProxyList` is then extracted and each resource is analyzed. Resources with the resource type `Resource` are downloaded and saved as files. For resources with the resource type `Metadata`, the CMDI XML data is retrieved and the process is repeated recursively until all information for all files associated with the collection has been obtained. All information is then stored in a tree structure corresponding to the structure of the collection with the node names based on the names of the various collections and files, with the corresponding PID being added as a prefix to each file name in order to avoid any name collisions.

There were a number of challenges involved with the extraction of the data. Firstly, it is not always straightforward to extract the CMDI XML data associated with a PID. For each PID URL, there are a number of redirects to go through before arriving at the final URL.² Moreover, the path that the redirects take is also affected by the value of the `Accept` header. It was found that this needed to be set to `application/x-cmdi+xml`, `application/xml+cmdi`, `application/xml` for both following the redirects as well as extracting the data from the URL to cover the different possible Media Types the CMDI XML data could have as its `Content-Type`.

There are also a number of CMDI metadata resources in certain collections that are mislabelled as `Resource` as their resource type, when the resource type should be `Metadata`. In order to solve this, all resources labelled as `Resource` with an XML, CMDI, or no Media Type are analyzed and checked if they can be analyzed as a CMDI file, *i.e.*, it is checked whether the file contains a `ResourceProxyList` which is non-empty. If so, then the resource is analyzed as CMDI metadata.

²For example, the handle `http://hdl.handle.net/11022/0000-0007-CFE2-1` redirects to `https://weblicht.sfs.uni-tuebingen.de/PidResolver/erdora/SFB833/A03/Compositional-Model-Sem/comp-models-sem-interp`, which in turn redirects to `https://talar.sfb833.uni-tuebingen.de/erdora/cmd/SFB833/A03/Compositional-Model-Sem/comp-models-sem-interp`.

Another problem is that some resources are inaccessible for various reasons such as the need for authentication. Inaccessibility of resources is indicated in the data tree and the corresponding HTML or JSON files. Moreover, some collections are ‘circular’, *i.e.*, a resource in the CMDI metadata of one resource will eventually lead back to the original resource. CMDI Explorer keeps a record of resources encountered so far, and any resource that has already been analyzed is simply not included in the tree that is being constructed.

Front-end. Fig. 5 shows the main interface, presenting users the three options to enter their input. Once the metadata has been submitted, CMDI Explorer’s back-end attempts to extract its underlying

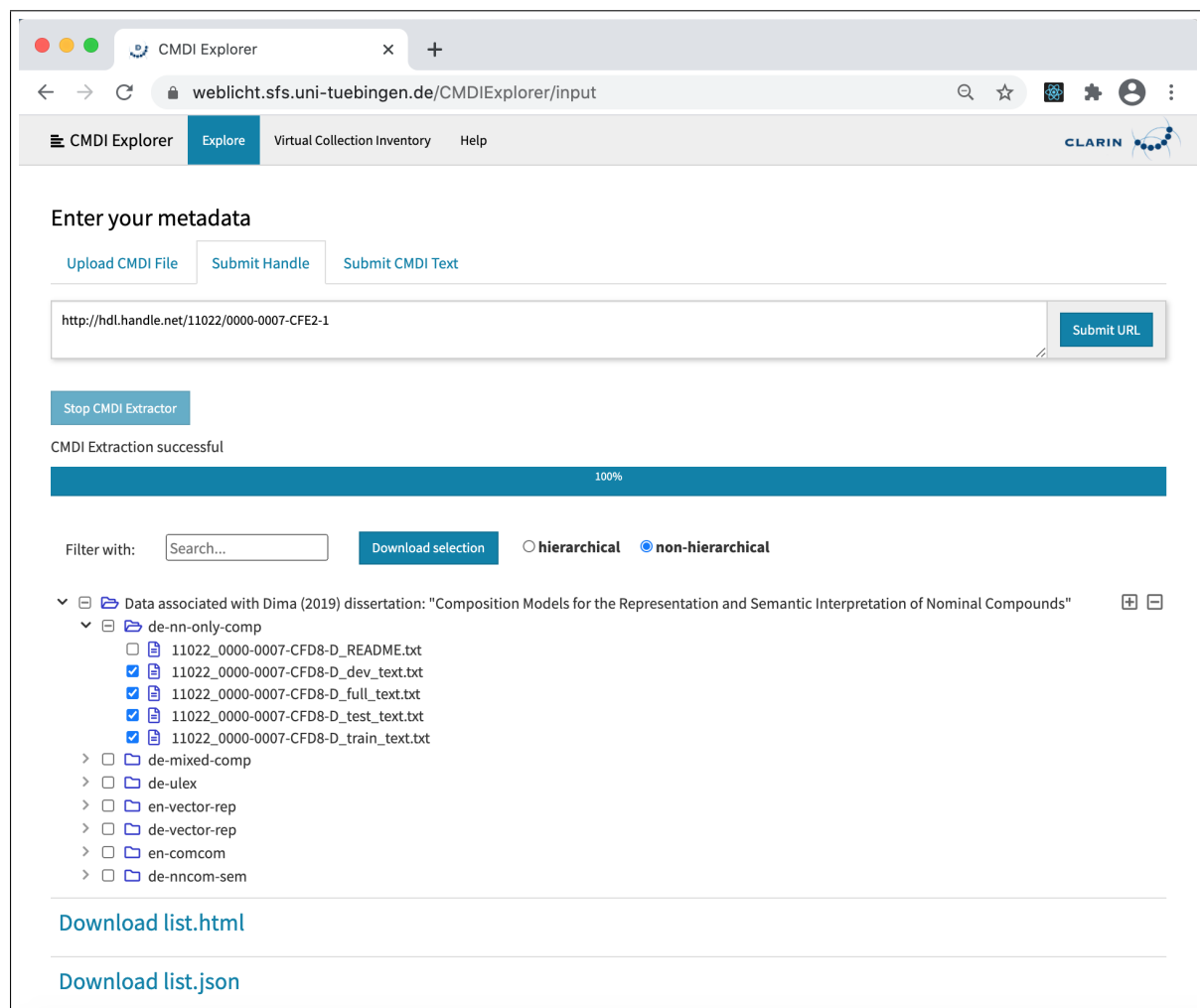


Figure 5: A CMDI record displayed in CMDI Explorer

tree. During extraction, users see a progress bar and live updates to the emerging tree visualisation. Fig. 5 shows a top node together with its immediate children. Each branch of the tree can be unfolded individually, and there are also two controls (+, –) to unfold or fold the entire tree. Leaf nodes are actionable. When users click on a leaf node, a pop-up window appears, see Fig. 6. The window shows some metadata about the selected resource such as its name, size and mediatype. It also gives users three follow-up actions to choose from: (i) copy the handle to the clipboard, (ii) click on the handle so that the respective resource resolves in the browser, and (iii) send the handle to the Switchboard for further processing.

Users can choose to download the entire tree, either in a structure-preserving or in a flatly structured HTML or JSON format (note the bottom two links in Fig. 5). Moreover, users can select nodes (subtrees, leaf nodes) individually. Selected nodes will be added to the ‘download basket’; the corresponding data resources of their selection are made available as a ZIP archive file.

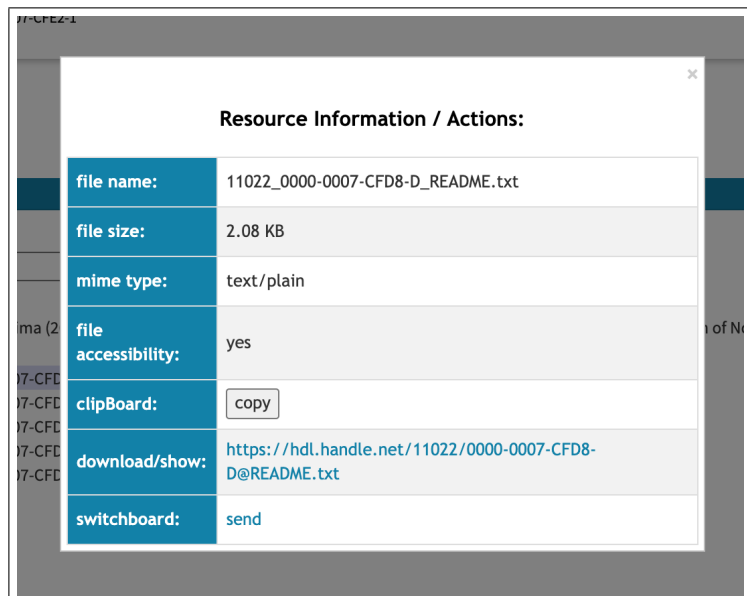


Figure 6: Actions attached to leaf nodes.

4 Current State and Future Work

We have built a prototype of CMDI Explorer that implements its core functionality and which is now open for beta testing at <https://weblicht.sfs.uni-tuebingen.de/CMDIExplorer/>. We invite readers to explore the tool and encourage their feedback. Given CMDI Explorer’s lineage from the Switchboard’s codebase, we expect users to easily grasp its user interface and the functionality it offers.

During testing, we were confronted with the fact that there is a limitation on the size of the files which can be downloaded by the user. Due to a limit in the size of array buffers in web browsers, the maximum file size which can be downloaded is about 2.14 GB, corresponding to the maximum 32-bit signed integer value. To overcome this limitation, we will need to enhance CMDI Explorer’s back-end to deliver such files in piecewise fashion, and CMDI Explorer’s front-end to reconstruct such files from the pieces delivered to the user.

The assemblage of resources from archives becomes a complex task when resources are protected by usage rights or licenses, and hence by AAI protocols. We are aware of the issue but are unsure which path to take as we do not want to move CMDI Explorer behind a Shibboleth wall.

We are also considering to use CMDI Explorer to hold an inventory of existing virtual collections (see the top navigation bar item “Virtual Collection Inventory” in Fig. 5). Here CMDI Explorer would simply show a manually curated catalogue of virtual collections that may be of interest to a wider CLARIN community. It would give a short description of a given collection together with a handle to access it. When clicking on the handle, CMDI Explorer would then show the tree structure of the corresponding collection.

CMDI Explorer has already been included in the test version of the Switchboard as a metadata processing tool. That is, when users give the Switchboard a CMDI file to process, the Switchboard identifies CMDI Explorer as applicable tool, which can then be directly invoked to show the possibly complex resource tree described by the CMDI metadata. We expect CMDI Explorer to be connected to the VLO and the VCR as well so that users can easily invoke it wherever they find complex CMDI metadata they need to explore further.

Acknowledgements

Our work was funded by the German Federal Ministry of Education and Research (BMBF), the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK), and CLARIN-D. Emanuel Dima, Willem Elbers, Dirk Goldhahn, Marie Hinrichs and Dieter Van Uytvanck participated in the initial discussion and contributed to the conceptualisation of the explorer.

References

- Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.
- Corina Dima. 2019. *Composition Models for the Representation and Semantic Interpretation of Nominal Compounds*. Ph.D. thesis, University of Tuebingen.
- Willem Elbers. 2017. Virtual collection registry v2. Technical report, CLARIN ERIC.
- Margaret King, Davor Ostojic, Matej Ďurčo, and Go Sugimoto. 2016. Variability of the facet values in the VLO – a case for metadata curation. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015*, pages 25–44, Linköping. Linköping University Electronic Press.
- Davor Ostojic, Go Sugimoto, and Matej Ďurčo. 2017. Curation module in action – preliminary findings on VLO metadata quality. In *Proceedings – CLARIN Annual Conference 2016*.
- Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari et al., editor, *Proceedings of LREC'12*, Istanbul, Turkey. ELRA.
- Claus Zinn, Thorsten Trippel, Steve Kaminski, and Emanuel Dima. 2016. Crosswalking from CMDI to Dublin Core and MARC 21. In Nicoletta Calzolari et al., editor, *Proceedings of LREC'16*, Portorož/Paris. ELRA.
- Claus Zinn. 2018. The Language Resource Switchboard. *Computational Linguistics*, 44(4):631–639.
- Matej Ďurčo. 2013. *SMC4LRT – semantic mapping component for language resources and technology*. Ph.D. thesis, Technische Universität Wien.

Signposts for CLARIN

Denis Arnold

Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany
arnold@ids-mannheim.de

Bernhard Fisseni

IDS and
Universität Duisburg-Essen, Germany
bernhard.fisseni@uni-due.de

Thorsten Trippel

Eberhard Karls Universität
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Abstract

An implementation of CMDI-based signposts and its use is presented in this paper. Arnold, Fisseni et al. (2020) present signposts as a solution to challenges in long-term preservation of corpora. Though applicable to digital resources in general, we focus on corpora, especially those that are continuously extended or subject to modification, e.g., due to legal injunctions, but also may overlap with respect to constituents, and may be subject to migrations to new data formats. We describe the contribution signposts can make to the CLARIN infrastructure, notably virtual collections, and document the design for the CMDI profile.

1 Introduction

The current paper presents an implementation of the concept of *signposts* (Arnold, Fisseni et al. 2020) which is based on the Component Metadata Infrastructure (CMDI, see Broeder et al. 2012), and explains how signposts can contribute to the overall CLARIN infrastructure. The contribution concerns the use of persistent identifiers (PIDs) for resources, and the handling of data removal, data migrations and versioning as well as deduplication.

A **signpost** is a metadata file for a leaf on the tree of resources, for instance a single text or an audio recording. Using terminology from the area of long-term archival, we distinguish *conceptual object* (CO) from *logical object* (LO) (see chapter 9 by Stefan Funk in Neuroth et al. 2009).¹ A CO can be realized in different LOs, for instance an audio recording (CO) can be realized in files of different audio formats (LO). **A signpost represents a conceptual object (CO), and also refers to logical objects (LOs, typically at least one) belonging to it.**

The most important point about signposts is that they change the idea what a PID refers to when providing data: While traditionally, PIDs may point directly to data files (LOs),² it is suggested here that PIDs only refer to signposts (COs), and to leave it to signposts to point to files. The reason for this change is that LOs may be volatile, even if the represented information stays the same. By adding signposts as a layer of indirection, we can achieve an acceptable trade-off between the necessity of modifying data and the demands of long-term archival on the one hand and Open Science as well as reproducibility on the other.

Signposts are first motivated below with examples regarding *growing corpora*, i.e. large corpora that are constantly extended and contain material where the conglomerate of commercial interests, intellectual property rights and privacy rights constitutes a non-trivial problem. However, all aspects signposts address are relevant to other kinds of corpora as well, generally to a different degree. Moreover, not only customary, but also virtual collections in particular benefit from signposts, as will be discussed in section 4. In case of small and ‘legally’ permanent corpora, signpost information may be included in the corpus metadata. Signposts replace the concept of tombstones, which are less flexible than signposts (see Arnold, Fisseni et al. 2020).

2 Motivating Signposts

The motivation for signposts comes from the impermanence of logical objects, specifically three aspects: the necessity of *deletions* due to legal actions, (conceptual) *deduplication* and data *migration*.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹Funk (chapter 9 in Neuroth et al. 2009) also distinguish the level of *physical object* which, however, is not immediately relevant for our current discussion.

²PIDs are also used to refer to datasets (see, e.g., De Smedt, Koureas and Wittenburg 2020 for a suggestion on how to structure datasets). However, we focus on data here.

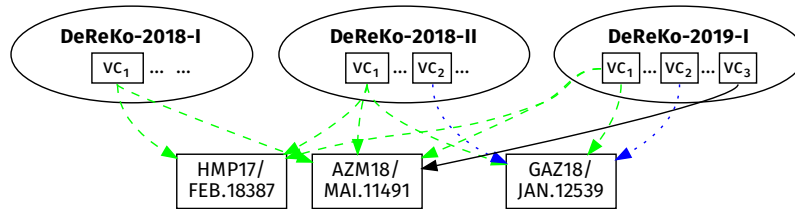


Figure 1: Relationship between DeReKo releases, virtual sub-corpora and texts (from Arnold, Fisseni et al. 2020). Texts may be part not of one, but many (virtual) corpora, and may belong to different versions of corpora.

Long-term Preservation vs. Legal Necessity of Deletions. In the realm of long-term preservation, we assume that original data, in the sense of COs, will always be retained. However, e.g. when building corpora from newspapers, it may become necessary to remove data from COs due to injunctions or revocation of licenses.

Migration. File formats may fall out of use, so that data must be converted to new formats, which in the OASIS model (CCSDS 2012) is called *migration*. Anecdotally consider the German Reference Corpus (DeReKo, see, e.g. Kupietz et al. 2010) compiled at the Leibniz Institute for the German Language (IDS). Between 1999 and 2005, SGML (ISO8879:1986 1986) / CES were used as its data format, then DeReKo was converted to XML (for the history and the decisions involved, see Lungen and Sperberg-McQueen 2012), based on the TEI’s P3 and later P5 recommendations (Sperberg-McQueen and Burnard 1999; Burnard and Bauman 2020). Similar conversions occurred in the IDS’ oral corpora. Even if we assume that we retain the original LOs, which goes beyond the OASIS model, we would want to add new ones as time progresses. For instance, we want to provide XML files conforming to P5 today rather than P3. It may then be a good idea to retire the intermediate versions to avoid storage cost. With the traditional approach to metadata, these changes mean that we have to change the metadata in each of these steps. With signposts, we only change the signpost.

Complex Corpus Structures. Especially growing corpora may have intricate structures, e.g. overlapping with respect to COs. If information were recorded in the metadata of the parent structures of the leaf COs, the metadata records would have to be changed for several corpora, while with signpost only the latter must be adapted. Figure 1 shows the relationships between the DeReKo corpus releases and virtual corpora vc_1, \dots, vc_3 , and three texts. Based on release DeReKo-2018-I, vc_1 was defined,³ already containing the texts HMP17/FEB.18387 and AZM18/MAI.11491. DeReKo-2018-II added GAZ18/JAN.12539 to vc_1 . Based on DeReKo-2018-II, vc_2 was defined, containing the text GAZ18/JAN.12539. vc_3 was defined initially on DeReKo-2019-I, also containing AZM18/MAI.1149. This shows that texts in DeReKo may belong to many different corpora. In this case, removal becomes a complex matter.

In the next releases of both corpora in the IDS repository, we plan to implement signposts to avoid manually editing thousands of files in cases of conversion and legal issues. We will report on the first corpus successfully using signposts in section 5.

3 Signposts for CLARIN: A CMDI Profile for Signposts

Reusing existing CMDI components, we developed a metadata profile for signposts, the signpost profile has the identifier `clarin.eu:cr1:p_1587363818266`⁴ in the component registry.

The CMDI profile reuses existing components and intends to include technical information that can be automatically extracted based on a file. This information can be gained by the File Information Tool Set (FITS)⁵ or other readily available tools to facilitate processing. The collected information includes the original filename, media type, file size in bytes, various checksums and cryptographic hashes. Besides this basic technical information on a LO, the signpost should also include information on the status of each LO, i.e.

³For the importance of virtual corpora in DeReKo’s *primordial sample* design and extensionally or intensionally defined virtual corpora see Kupietz et al. (2010).

⁴see https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1587363818266/1.2/xsd, which is still in the development state, but accessible

⁵See <http://fitstool.org>

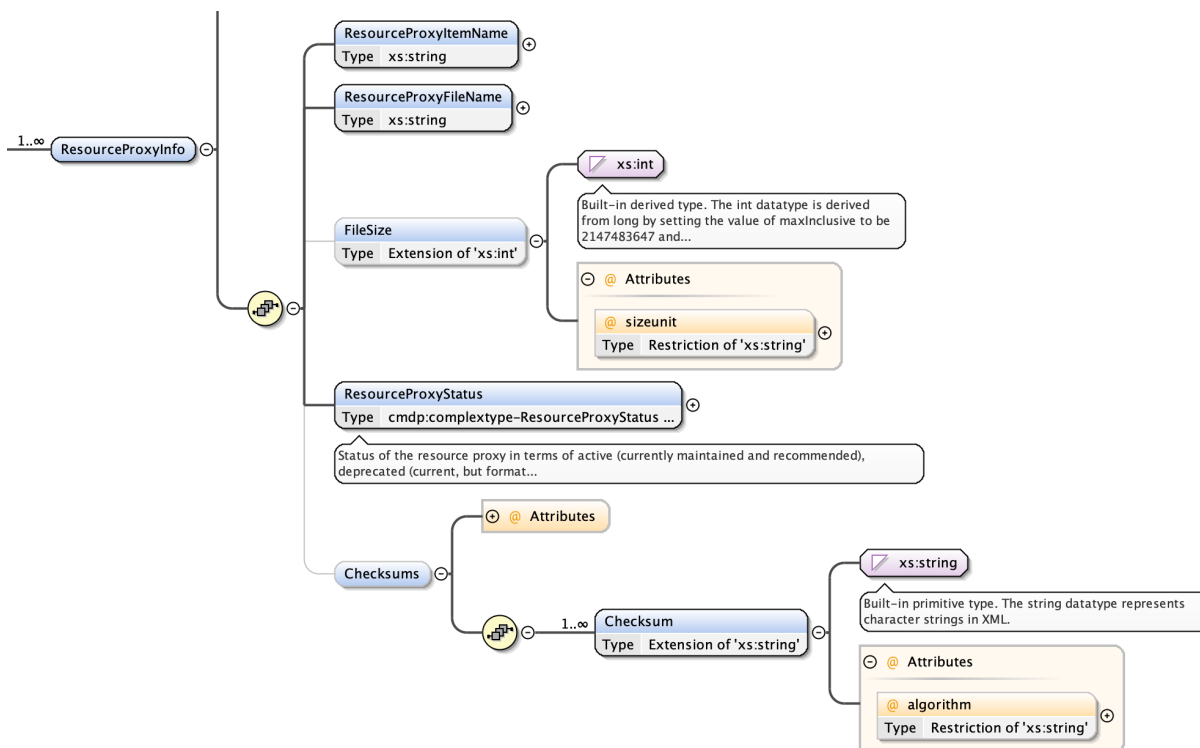


Figure 2: Visualisation of the structure for the metadata provided for each resource proxy in the CMDI Profile *signpost* (clarin.eu:cr1:p_1587363818266)

whether this is the currently maintained version, whether the use of the LO is deprecated (for example in the case of a migration to other data formats) or whether a file is no longer available. Reasons can be given in the provenance information (see next subsection). The schema also allows referring to a LO by a specific name that is not its filename, which is occasionally necessary.

The nature of the *signpost* profile is different from other profiles, in the sense that it is not intended to provide a meaningful description of a resource and does not foster findability by search engines such as the VLO. Hence it does – by design – neither cater for the VLO facets, nor respect quality criteria which are automatically evaluated by the Curation Module⁶. As the media type is already provided in the resource proxy list of the CMDI file, there is no component including it. For each LO in the CMDI’s resource proxy list, the profile allows the provision of various and multiple checksums, each specifying the algorithm in an attribute.

3.1 Provenance

The CMDI 1.2 specification (CMDI Taskforce 2016) implies that provenance information is not to be included in the CMDI file, but instead in one or more separate file(s) called journal files. One reasoning behind this is that provenance information is not (necessarily) to be machine-interpretable and should be directed to human readers. We implement the journal file in HTML with microformats, as this approach allows formatting for human consumption by means of a web browser and aims at semantic interoperability. We include information in the journal file which is useful both for ensuring reproducibility of research and for keeping track of the development of a resource.

We include a log of every modification to the CO described by the *signpost*. These **changes** contain the following: creation, ingest, injunction, and migration. Moreover, all changes are dated with a **time stamp** and include a short human-readable **log message**. We suggest to also include modifications of LOs, i.e. **object changes**. Changes of the LO are marked as an **addition**, a **replacement**, or a **removal**. An `xml:id` (Marsh, Veillard and Walsh 2005) attribute can be used in the *signpost* to identify LOs. This way, the log allows to determine the lifespan of a LO in a machine-readable way.

⁶<https://curate.acdh.oeaw.ac.at/>


```

<h1>Log for <a href="http://PID-1">Conceptual Object
<code>http://PID-1</code></a></h1>
<ul class="sign_post_log">
  <li class="creation_entry">
    <span class="timestamp">2021-05-15T02:00:00+02:00</span> <span class="log_message">object created</span></li>
  <li class="ingest_entry">
    <span class="timestamp">2021-07-07T02:00:00+02:00</span> <span class="log_message">File ingested into IDS LTA</span>
  <ul class="object_changes">
    <li class="change_addition"><a href="http://PID-1#lo_1">Element</a> added</li></ul></li>
  ...

```

3.2 PIDs

The usage of persistent identifiers differs significantly from the current usage in repositories. Traditionally, care is taken to assure that links to logical objects remain available and persistent; COs are not necessarily represented. We reverse this: Not the LO (file) but the CO (signpost) is primary. This means, only the signpost is granted a persistent identifier, and the access to logical objects is through URLs for which the archive does not give any guarantees. As this is currently unconventional, we must alert the user to the impermanence of LO URLs. We have considered the following strategies:

We can implement a notice at the **presentation layer** of the repository: e.g. using a link text line *temporary download link*. This would inform human users, but is of no consequence for machine-processing of CMDI records. We suggest this is not a grave problem, for two reasons: First, as long as URLs for logical objects are not reused, tools relying on the `ResourceProxyList` and even caches will have no problems. This means that for a tool like the CMDI Explorer (Arnold, Campbell et al. 2020) currently developed by CLARIN-D and the CLARIN ERIC, and which will recursively process chains of CMDI records, nothing changes, except there is one link more in the chain for each conceptual object. Moreover, in case of multiple active LOs, the CMDI Explorer would have to avoid downloading them all. Secondly, we assume that by the time signposts are in widespread use, tool authors will have been made aware of the concept, and will take care not to download LOs blindly, but rely on signposts. It may be advantageous to integrate licensing information per LO, potentially in tandem with access control lists. Crawling of resources rather than metadata (including signposts), etc., can be prohibited in the `robots.txt`.

Alternatively or additionally, one could take care to generate temporary links to logical objects and hence force users to not rely on their URLs. We assume that this strategy generally wastes resources and should only be the last resort.

4 Signposts for CLARIN: Making Virtual Collections Future-Proof

Virtual collections are an important asset of the CLARIN infrastructure. They allow to recombine existing collections and individual resources into new collections, which is done in CLARIN's Virtual Collections Registry (<https://collections.clarin.eu/>, see Elbers 2017). In that sense, they correspond to the case discussed as 'complex corpus structures' above, but with the additional twist that they are potentially defined across repositories. While not a goal in the design of signposts, signposts come in handy when devising virtual collections.

The goal is that such collections are from the user's point of view indistinguishable from non-virtual collections. Tools like the CMDI Explorer allow to work with virtual collections, and generally have to resolve the components of a virtual collection according to the `ResourceProxyList` and the relationships codified in the Component Metadata of the collection.

Tools using virtual collections have even less control over what happens to constituents. In that sense, the benefit of using signposts is even greater. Consider the traditional use of PIDs in virtual collections as depicted in figure 3: The PIDs used in the virtual collection point to resources directly, also to parts of existing collections. When an item becomes unavailable (red item), the PID will just cause an error. It will not even be easy to decide whether this is just a temporary network issue or a change in the repository.

Now, we introduce signposts. Figure 4 shows what a collection looks like in the traditional way and then what it looks like using signposts. In figure 5, it is shown how using signposts simplifies working with virtual collections: As the single resources are protected behind signposts, the error can be caught immediately, and deliberate changes to the resource will be traceable in time due to the log associated with the signpost. Similar, format migrations can be taken advantage of using the multiple logical objects pointed to by the signpost in such a case.

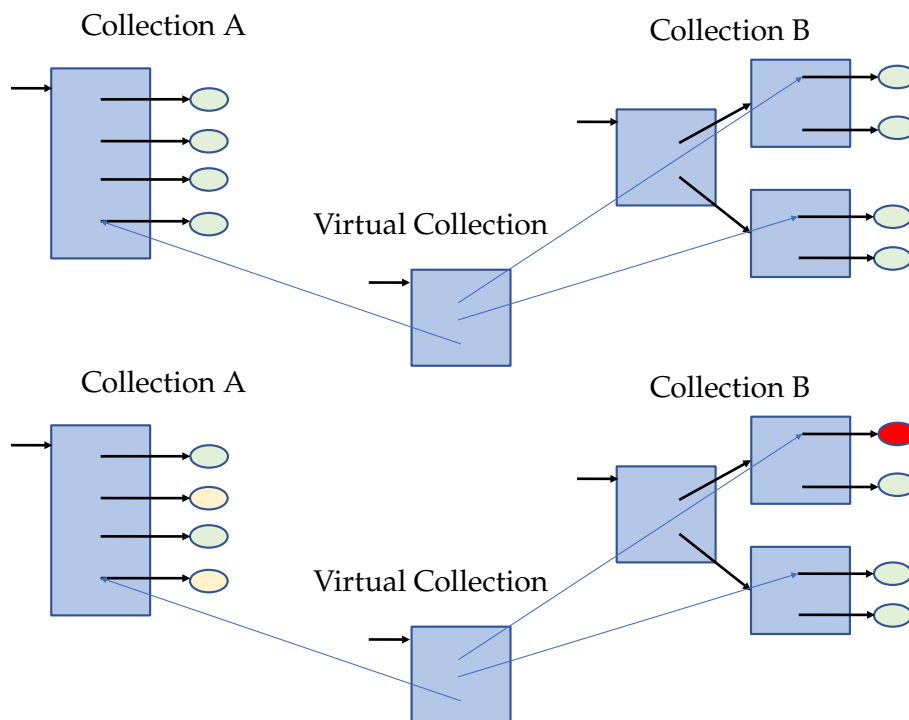


Figure 3: Virtual collection; big boxes are (virtual or non-virtual) collections, ovals are single resources, black arrows are links through PID. The lower collection contains a item (red) which is unavailable.

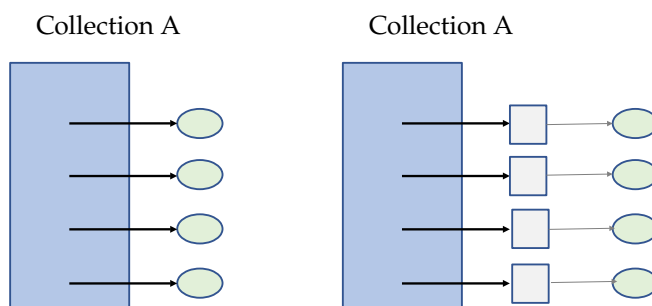


Figure 4: Collection without signposts (left) and using signposts (right); big boxes are (virtual or non-virtual) collections, ovals are single resources, black arrows are links through PID. Small boxes are signposts, light arrows are feeble non-permanent links.

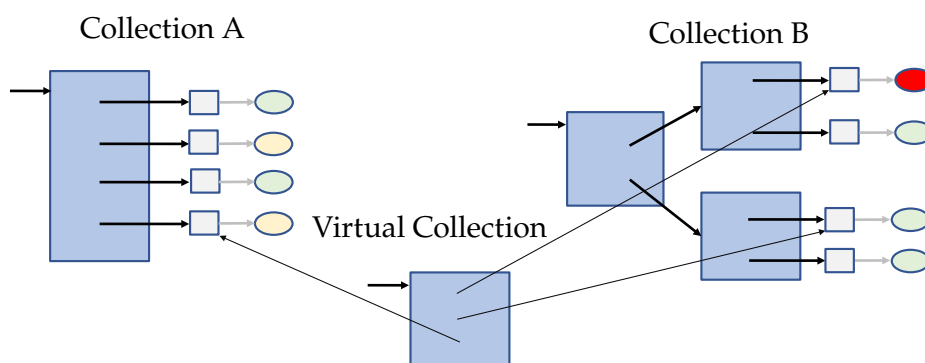


Figure 5: Virtual collection using signposts; big boxes are (virtual or non-virtual) collections, ovals are single resources, black arrows are links through PID. Small boxes are signposts, light arrows are feeble non-permanent links. The collection contains a item (red) which is unavailable.

Earlier, when referenced data was required to undergo any form of editing, the resources of the virtual collection became unavailable without notice to a user of the virtual collection, besides maybe a webserver error message, but even that would not have been available when a tombstone was in place. Signposts on the other side provide information to a user of a virtual collection that is meaningful, pointing to the actual data or informing them on new versions or added restrictions. This method is scalable in contrast to technical solutions automatically verifying the persistent availability of resources that are part of virtual collections. It also extends to those resources with restricted access, which would not allow an automatic evaluation, for example by comparing checksums of data streams part of a virtual collection.

5 Signposts in CLARIN: A Corpus and A Virtual Collection

Virtual Collection Registry Browse Create Help Login CLARIN

Signpost Demo Collection – Bonner Gerontologische Längsschnittstudie (BOLSA), all recordings of subject 2608

General

Name: Signpost Demo Collection – Bonner Gerontologische Längsschnittstudie (BOLSA), all recordings of subject 2608

Type: EXTENSIONAL

Creation date: 2021-01-25

Description: This virtual collect study *Bonner Gerc*

REPO FORSCHUNGSDATEN-REPOSITORIUM DIGITALES LANGZEITARCHIV IDS LEIBNIZ-INSTITUT FOR DEUTSCHE SPRACHE

Purpose: **IDS Repository** **SignPost for Conceptual Object**
SAMPLE <http://hdl.handle.net/10932/00-0537-F1FB-1BA9-6301-2>

Reproducibility: **Description**
INTENDED **Terms of Use**
Browse
Search
About
OAIprovider

Persistent identifie **Collection:** Leibniz-Institut für Deutsche Sprache, CLARIN-D-Zentrum, Mannheim
HDL 11372/VCF **Conceptual Type:** audio
DOI 10.34733/v **Log:** log file
Logical Objects: **Logical Object #1:** BLSA_E_26081_A_01.WAV
Status: active
Media Type: audio/wav
Size: 1,531,795,408 bytes

Keywords:
• signposts
• demo
• BOLSA
• signposts
• demo

Figure 6: Screenshots from the Virtual Collection Registry showing the virtual collection of all recordings of subject 2608.

The first collection that was ingested into a CLARIN repository using signposts was the Bonn Gerontological Longitudinal Study (*Bonner Gerontologische Längsschnittstudie*, acronym BOLSA, see Lehr and Thomae 1987), which can be viewed in the IDS repository at <http://hdl.handle.net/10932/00-0537-F2C9-1120-E101-9>. This is the audio part of BOLSA, the further data and background information are available at the University of Halle (<https://bolsa.uni-halle.de/>) It contains interviews with the same 222 subjects (born between 1890 and 1895 and between 1900 and 1905, respectively) between 1965 and 1984.

The data is structured as follows: The whole audio corpus is structured into recording events which are grouped into ‘waves’ and also associated with test subjects. Recording events contain metadata about the recording (date, place, topics and keywords) as well as the test subject and interviewer. This specific metadata is kept as payload, but all metadata corresponding to facets used by the Virtual Language Observatory are made available as Component Metadata. However, only the whole data set is displayed in the VLO, not the recording events.

The collection and the recording events have been assigned Component Metadata as usual, pointing to their parts through PIDs. All single recordings and all original metadata files have been assigned PIDs which point

to signposts. These signposts point to the data. Currently, unfortunately, the data cannot be accessed publicly yet for legal reasons.

We extensionally defined a virtual collection (see DOI 10.34733/vcr-1036) according to metadata criteria for all recordings of subject 2608. Screenshots from the Virtual Collection Registry can be found in figure 6.

6 Conclusion

We proposed the notion of signpost for addressing data removal, migration and deduplication in long-term archival of resources, with a specific focus on growing corpora. We also presented the added value to virtual collections and provided an implementation of the concept in the CLARIN infrastructure. The concept of signposts fits to the overall architecture of managing resources in distributed environments, including decentralized provision of metadata endpoints for harvesting, adapted metadata schemas according to ISO 24622-1 (CMDI) and persistent identification. In addition to these features it also allows for consistent integration within Virtual collections as for example implemented in the Virtual Collection Registry (VCR) of CLARIN.

We illustrated the usage of signposts within CLARIN building a virtual collection based on the first collection ingested into a CLARIN repository using signposts.

It is important to note that both from the depositor's and from the end user's point of view, signposts need not be taken into account once repositories and tools implement them. Generating signposts is best done during the ingest of data into the repository, as the are tied to the preservation policy of the repository.

We welcome feedback on the concept and on the implementation. Future work will concern the adaptation of the format, and the integrations with tools, as outlined above.

Acknowledgements

The work reported here was funded by the German Federal Ministry of Education and Research (BMBF), the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK), Project Management Agency German Aerospace Centre (DLR), and CLARIN-D.

We thank the anonymous reviewers for helpful comments that have allowed us to sharpen the text.

References

- Arnold, Denis, Ben Campbell, Thomas Eckart, Bernhard Fisseni, Thorsten Trippel and Claus Zinn (2020). 'The CMDI Explorer'. In: *Proceedings of CLARIN Annual Conference 2020.05 – 07 October 2020, Online Edition*. Ed. by Costanza Navarretta and Maria Eskevich, pp. 157–161. URL: https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.
- Arnold, Denis, Bernhard Fisseni, Paweł Kamocki, Oliver Schonefeld, Marc Kupietz and Thomas Schmidt (2020). 'Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora'. In: *Proceedings of the LREC 2020 Workshop 'Challenges in the Management of Large Corpora' (CMLC-8)*. Marseille, France.
- Marsh, Jonathan, Daniel Veillard and Norman Walsh (Sept. 2005). *xml:id Version 1.0*. W3C Recommendation TR xml-id. The World Wide Web Consortium. URL: <https://www.w3.org/TR/xml-id/>.
- Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen and Thorsten Trippel (2012). 'CMDI: a component metadata infrastructure'. In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*. Vol. 1.
- Burnard, Lou and Syd Bauman, eds. (2020). *Guidelines for Electronic Text Encoding and Interchange. TEI P5*. version 1.0.0 2007; latest release 4.0.0 on 2020-02-13. Chicago, New York: Text Encoding Initiative.
- CCSDS (2012). *Reference model for an open archival information system (OAIS)*. CCSDS 650.0-M-2. 2nd ed. Washington: CCSDS. URL: <https://public.ccsds.org/pubs/650x0m2.pdf>.

- CMDI Taskforce (2016). *Component Metadata Infrastructure (CMDI): Component Metadata Specification, version 1.2*. Tech. rep. CLARIN ERIC. URL: https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf.
- De Smedt, Koenraad, Dimitris Koureas and Peter Wittenburg (2020). ‘FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units’. In: *Publications* 8.2.
- Elbers, Willem (2017). *Virtual Collection Registry v2*. Tech. rep. CLARIN ERIC.
- ISO8879:1986 (1986). *Information processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. Standard No. ISO 8879:1986. International Organization for Standardization.
- Kupietz, Marc, Cyril Belica, Holger Keibel and Andreas Witt (2010). ‘The German Reference Corpus DEReKo: A Primordial Sample for Linguistic Research’. In: *Proceedings LREC’10*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner and Daniel Tapias. Valletta/Paris: European Language Resources Association (ELRA), pp. 1848–1854.
- Lehr, Ursula and Hans Thomae, eds. (1987). *Formen seelischen Alterns. Ergebnisse der Bonner gerontologischen Längsschnittstudie (BOLSA)*. German. Stuttgart: Enke.
- Lüngen, Harald and Christopher Michael Sperberg-McQueen (2012). ‘A TEI P5 Document Grammar for the IDS Text Model’. In: *Journal of the Text Encoding Initiative* 3, pp. 1–18. URL: <http://jtei.revues.org/508>.
- Neuroth, Heike, Achim Oßwald, Regine Scheffel, Stefan Strathmann and Mathias Jehn, eds. (2009). *nestor Handbuch. eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.0 [3/2010]. nestor.
- Sperberg-McQueen, Christopher Michael and Lou Burnard, eds. (1999). *Guidelines for Electronic Text Encoding and Interchange. TEI P3*. initial release 1994-05-16; last version dated May 1999. Chicago, New York: Text Encoding Initiative.

Studying Emerging New Contexts for Museum Digitisations on Pinterest

Bodil Axelsson

Department of Culture and Society
Linköping University, Sweden
bodil.axelsson@liu.se

Daniel Holmer

Dept. of Computer and Information Science
Linköping University, Sweden
daniel.holmer@liu.se

Lars Ahrenberg

Dept. of Computer and Information Science
Linköping University, Linköping, Sweden
lars.ahrenberg@liu.se

Arne Jönsson

Dept. of Computer and Info. Science
Linköping University, Sweden
arne.jonsson@liu.se

Abstract

In a SweClarin cooperation project we apply topic modelling to the texts found with pins in Pinterest boards. The data in focus are digitisations of Viking Age finds from the Swedish History Museum and the underlying research question is how they are given new contextual meanings in boards. We illustrate how topic modelling can support interpretation of polysemy and culturally situated meanings. It expands on the employment of topic modelling by accentuating the necessity of interpretation in every step of the process from capturing and cleaning the data, to modelling and visualisation. The paper concludes that the national context of digitisations of Viking Age jewellery in the Swedish History Museum's collection management system is replaced by several transnational contexts in which Viking Age jewellery is appreciated for its symbolical meanings and decorative functions in contemporary genres for re-imagining, reliving and performing European pasts and mythologies. The emerging contexts on Pinterest also highlight the business opportunities involved in genres such as reenactment, neo-paganism, lajv and fantasy. The boards are clues to how digitisations serve as prototypes for replicas.

1 Introduction

For more than one decade, the digitisation of collections and archives has been a major tool for heritage institutions to make their holdings widely accessible and fulfil cultural policy goals related to democracy. Open collection management systems, digital heritage portals such as Digitalt Museum and Europeana invite people outside institutions to share and interpret what they like. The advent of social media platforms has furthered this development and digital images of museum objects now extensively circulate online. Many museums now fully embrace the changes, encourage all kind of uses of their holdings and are curious about emerging interpretations outside the frame of museum knowledge.

In line with this development, digitisations, that is digital images of museum objects, from all kinds and sizes of museums now appear on the content sharing platform Pinterest. Major international museums link images to the platform, which promotes itself as a visual discovery engine. It invites users to create themed collections called boards, either by linking images from other websites or by selecting among the images circulating on the platform, for an example see Figure 1. The platform serves an increasing number of users with images out of a growing bank of pins (at the end of 2020, over 440 million of users and 200 billion of images). Once on Pinterest, images are set in motion by machine learning algorithms that present users with grids of images that change with each subsequent click.

The aim of this paper is to investigate new contextual meanings of digitisations of Viking Age Jewellery from the Swedish History Museum (SHM) on Pinterest. The museum does not post digitisations on the platform, still images of objects in its collection are numerous on the site. They are linked directly from the museum's open collection management system or from the museums Flickr account. But more often they have been linked to personal websites and blogs before inserted into Pinterest large bank of images (Wilson, 2016).

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

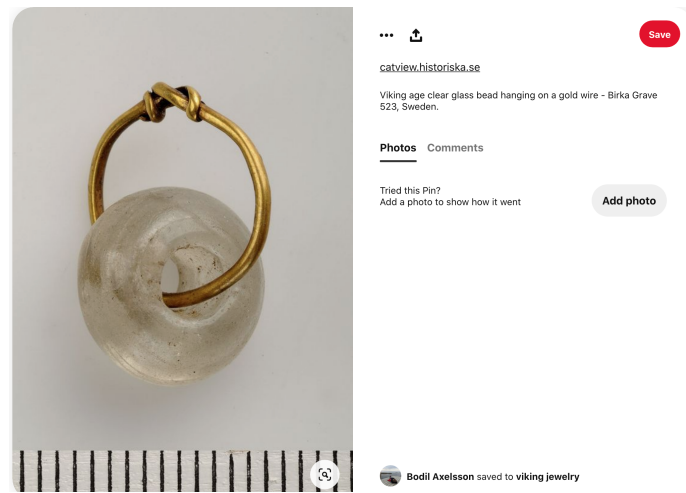


Figure 1: Typical pinterest entry used in this study.

Museum objects and their digital images are polysemic (Cameron and Robinson, 2007). Their meaning depends on available techniques and genres for interpretation, framing and contextualisation (Kirshenblatt-Gimblett, 1998). On Pinterest, they circulate as digital images in a space designed for consumption (Kidd, 2014). Here, their meaning depends both on how they are described in words and how they are displayed in boards. Digitisations from SHM that find their way to Pinterest are often recognisable as sourced from this museum. Moreover, they are depicted according to conventions for picturing the museum's objects, a view from above against a light greyish background, sometimes with a measure stick. In the museum's database, digitisations are embedded in knowledge developed for collection management and the discipline of archaeology such as inventory number, find location (place, parish, region and country), estimated dating, substance, category, keyword and type. When a user saves a digitisation on a board it is sourced with url and a brief text description of the user's choice. Consequently, only a selection of the museum's metadata is transferred to Pinterest. When on Pinterest, the platform identifies the digitisation with a signature that is common for all subsequent repinnings to additional boards and links are created between all the boards that contain the sourced digitisation (Liu et al., 2017). Each board to which the pin is subsequently saved provides a new context (Hall and Zarro, 2012).

2 Topic modelling and interpretation

The first author made contacts with the Swe-Clarín K-centre at Linköping University to discuss methods for analysing the text descriptions. We agreed that topic modelling would be a suitable technique for the purposes of the project, at the same time furthering the K-centre's skills and the CLARIN goals of supporting scholars "who want to engage in cutting edge data-driven research". Topic modelling is one of the language analysis tools provided by Clarín's Language Resource Switchboard and this paper expands on the employment of this tool by accentuating the necessity of acknowledging interpretation. We take from Drucker (2017) that data is always captured, counted, and represented by someone. Human decision making is involved in every step of topic modelling, from capturing and cleaning the data, to modelling and visualisation (Bechmann and Bowker, 2019). By reflecting on the role of human interpretation in the employment of topic modelling, we approach the "meaning problem" in digital humanities (Bode, 2019; Drucker, 2017; Liu, 2013).

At first, it seemed counter-intuitive to analyse digitisations on a site for visual discovery with a method for analysing texts. However, Pinterest API does not allow for downloading images and semantic information is vital for the platform's recommendation system. It is easily extracted and readable to both

humans and machines as clues to what images represent (Zhang and Lapata, 2017). Thus, the users' brief text descriptions contextualise digitisations and other images on the platform and add to the polysemy of digitisations.

Opinions vary as to the extent to which topic modelling supports the analysis of polysemy. Proponents of the method argue that it takes a relational approach to meaning. Relying on specific probabilistic assumptions, the method groups co-occurring words and identifies keywords for each topic. The different topics represent different contexts in which one and the same word can take on different meanings (Bode, 2019; DiMaggio et al., 2013). Jeffrey M. Binder instead argues that the method lures the analyst to assume that words have one single meaning because it dislocates words from their immediate linguistic contexts such as sentences and modality. Consequently, the analysis risks getting trapped in hegemonic word meanings and miss nuances. To counter methodological and linguistic naivety, interpretation emerges as a key issue in the application of the method (Binder, 2016). As concluded by Brett (2012), and Schmidt (2012) topic modelling indexes what the corpus is about, but a solid analysis builds on prior knowledge of the corpus and critical engagement with both the words that make up the topics and the presumable neutrality of probabilistic distribution.

In this study, the interpretation of words and topics are informed by an extensive engagement with the data. In fact, engaging with the data already in the phase of data capturing turned out to be a necessity. Pinterest's API did not provide any means to identify all boards that displayed SHM digitisations. An alternative sampling strategy had to be implemented. The strategy detailed below provides no way of estimating the representativeness, validity and reliability of the data in line with the parameters set up within quantitative social sciences, cf. Lomborg and Bechmann (2014). Instead, it secured that the interpretation of the topics was informed by a rich understanding of the data and its limitations.

The approach taken can be described as data-intensive heritage ethnography with a mix of human centred interpretation and automated data-capturing, cf. Bonacchi and Krzyzanska (2019). The starting point for the creation of data in this study, was a qualitative immersive experience with the aim of understanding how digitisations from the Swedish History Museum circulate on Pinterest. After entering the query "Swedish History Museum" in the platform's search bar, the recommendation service suggested the topic Viking Jewellery, a term that exists on Pinterest, but not in any museum database. The first author created the corpus by systematically scrutinising boards that included digitisations from SHM, suggested by the platform's recommendation system. She exclusively tapped on pins with brooches, pearls, rings, and necklaces from SHM that appeared in her home-feed, saved them on a board of her own and followed the boards where they were pinned. A second sampling strategy entailed selecting boards suggested by daily e-mail notifications with phrases such as "if you like Viking Jewellery [name of a board] you might also like" or "people like you were looking for Viking Women" [guided search category]. Simultaneously she took reflexive notes and screen shots. Thus, conditioned on the one hand by the researcher's ability to recognise objects from SHM and relevant boards, and, on the other hand, the platform's recommendation system, the collection of data took place between March 2018 and October 2018. It stopped when the platform's recommendation service started to suggest boards already collected.

Data from 480 boards created in interaction with the platform was fetched by using Pinterest's developer API. The dataset comprises a total of 329,999 entries. From this we filtered out both description and picture duplicates, and entries with empty description fields, giving us a dataset of 107,165 unique entries. Data was tokenized with the NLTK (Bird et al., 2009) tokenizer for English. The majority but not all descriptions are in English, we also identified Swedish, Russian, Norwegian, German and Dutch, but the English tokenizer was used for all languages. Thus, for dividing text into meaningful units we favoured English and did not treat all languages on equal terms and hence may have missed nuances in other languages cf. Bechmann and Bowker (2019). Furthermore, we filtered out words using the NLTK lists of stop words, with some domain specific additions such as image, search, and show, words that probably are auto generated by Pinterest or museums' websites. The texts were then lemmatized using the NLTK lemmatizer, and multi-word units such as Thor hammer were identified using bigrams and the *phrases* functionality in Gensim (Řehůřek and Sojka, 2010). With additional functionality in Gensim, we also filtered the dataset by pruning the words appearing in more than 90% of the boards. In most cases,

the threshold of the pruning of frequent words are set at higher levels (Maier et al., 2019), but due to some boards consisting entirely of languages other than English, the threshold had to be lowered for it to have any effect.

The human choices involved in our employment of topic modelling and visualisation of the topics are detailed below. As emphasised by Drucker (2017) and Drucker (2018), even though visualisations commonly are used for presenting data, they are not identical with the data but representations thereof. They are the outcome of curation and cleaning, parameterisation, the chosen model and the algorithms applied for its display. Therefore, Drucker (2018) proposes that visualisations should be environments for modelling. We instead draw the conclusion that the topics as displayed in the visualisation need to be carefully validated and interpreted. The first author made use of the ways in which topic modelling allows for identifying relevant documents for each topic. She returned to Pinterest to scrutinise signature boards, that is the boards with the highest probability for each topic (minimum 99 percent). She also returned to her field-notes to learn why and how these boards were chosen, inspected their grid of images and looked into descriptions and source links. Furthermore, she traced the possible meanings of the keywords generated for each topic in relation to how they were contextualised with pins, boards, the overall context of the platform, as well as in relation to how Vikings are conceived in archaeology and popular culture. The interpretation started from the proposition that meaning is dependent on shifting contexts, and that these contexts are reconstructed by the methods for detecting them cf. Seaver (2015). In this study then, contextualisation is on the one hand based in the processing of data in abstract mathematical space and then represented in the visualisation, and, on the other hand, something distinctly rooted in the situatedness and interpretive work of the first author, her knowledge about the boards and the role of Vikings Age jewellery in archaeology and various strands of popular culture.

3 Topic modelling

The purpose of topic modelling is to reveal thematic patterns in a collection of documents. The method enables extraction of knowledge from large collections of texts, that would otherwise be near impossible to analyse. In this study, the aim was to discover thematic patterns across a collection of boards all of which related to Viking jewellery one way or another. These thematic patterns are, after being extracted by a topic modelling method, represented as separate collections of keywords giving an overview of the most prevalent words in each topic.

We used Latent Dirichlet Allocation (Blei et al., 2003) to perform the topic modelling. Latent Dirichlet Allocation (LDA) is a generative probabilistic model where each document is represented as a mixture of latent topics, and each topic constitutes a multinomial distribution of words. The words with the highest probability in each topic are assumed to be the most probable representation of its content. To implement the LDA-model, Gensim was used. Gensim is a Python library that provides several tools for semantic modelling, one of which is a full implementation of the LDA-algorithm, namely the `LdaModel` class¹.

One of the main challenges when creating a good topic model is to determine the number of topics, which have to be specified in advance of its creation. To aid in this process, a coherence score shown to be largely correlated with human evaluations was used to assess the semantic quality of the topics (Newman et al., 2010). The assumption is that a model with a high overall topic coherence score produces topics that make more sense to a human, than a model with a low overall topic coherence score. Gensim implements the framework proposed in Röder et al. (2015) to calculate coherence scores, and this was also used in our study. The number of topics to use was decided after calculating the coherence score on a wide range of numbers of topics. In the end, 13 was found to give the highest overall score.

Other parameters used by the Gensim `LdaModel` were tuned according to the results of manual testing. The alpha value (α), that influences the prior distribution of topic weights in the documents, and the beta value (β), that influences the prior distribution of the word weights in the topics, are of great importance for the final results of the topic model. Earlier studies like Maier et al. (2018) suggest using the default β -value as implemented in Gensim, while altering the α -value. This approach was a starting point for

¹For documentation, see: <https://radimrehurek.com/gensim/models/ldamodel.html>

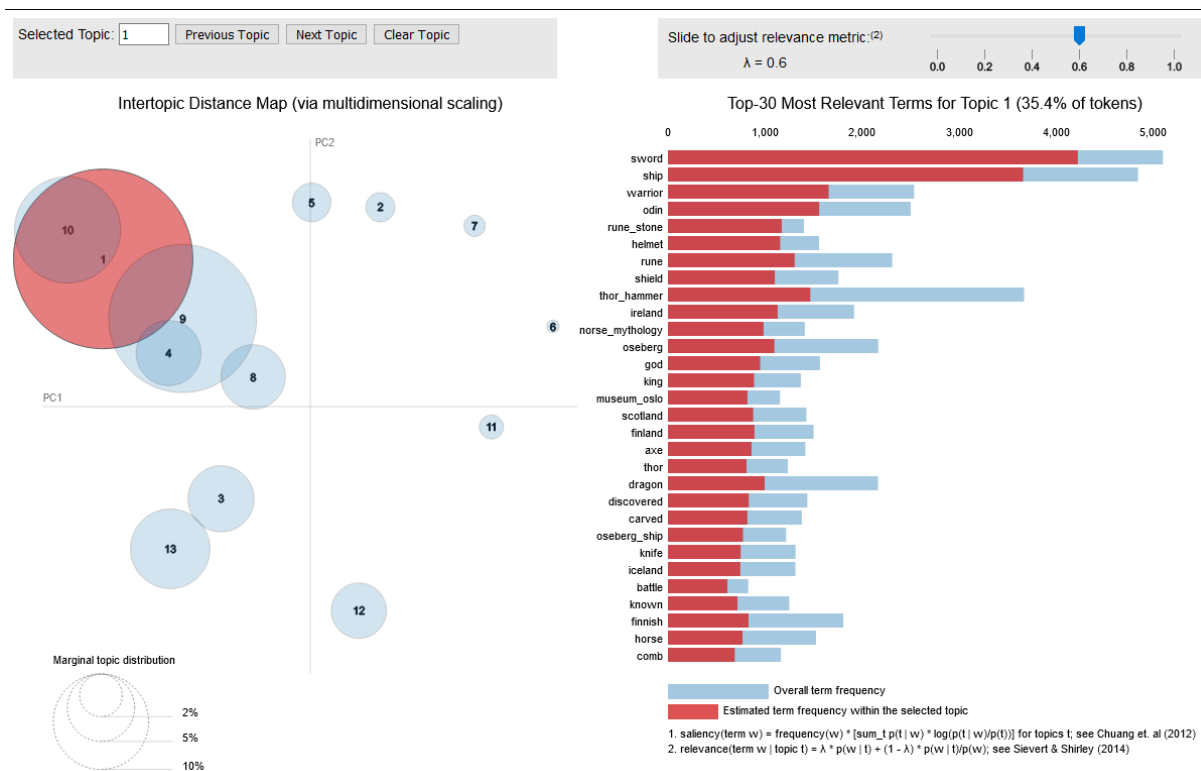


Figure 2: Visualisations of the topic models. Topic labels: 1. Norse Culture 2. Rings 3. Pearls 4. Pre-Christian Europe Cultures 5. Birka 6. Jewellery and trade 7. Brooches 8. Lajv and fantasy 9. Reenactment 10. Viking Jewellery 11. Metal work 12. Shiniies 13. Antique Jewellery.

our testing, and for the final topic model, the α -value was set to *symmetric* (which entails a value of $1/k$, where k is the number of topics).

The output of a LDA topic model using Gensim is a multinomial distribution over the topics and their most prevalent words. To aid the interpretation of the distributions we used a Python implementation of LDAvis (Sievert and Shirley, 2014), which through a web interface visualise how prevalent each topic is, as well as how the contents of different topics relate to each other, see Figure 2. The topic clusters are projected as circles in a two-dimensional plane, where the relative sizes and distances between the circles represent prevalence and similarity of the topics. This means that a topic with a large circle is seen more frequently in the entire collection of boards, and circles with a closer proximity share more features than circles that are projected further apart. In addition to the projection of the topics, there is also a bar chart of the most prevalent words in each topic. These are shown together with the frequency of the word in the entire corpus, allowing for a better understanding of the importance of the word in the current topic. LDAvis also introduces the term relevance, which is a way of ranking the words within the topics.

By adjusting the relevance metric closer to 0, it is possible to filter out words that are globally frequent, and assigning higher weights to words that are unique to the topic, while a relevance metric closer to 1 uses the standard ranks of the LDA model. The assumption is that globally frequent words might be too common and not accurately reflect what differs between the different topics. However, being too strict with the filtering comes with a drawback; the unique words are innately rare, which often makes the topics hard to interpret. That is, a word that is relevant to a topic will be undervalued and ignored, if it appears in another topic simultaneously. This is of particular importance for this study, where the descriptions all revolve around Vikings, and many relevant words are shared between topics. Sievert and Shirley (2014) therefore suggest a somewhat balanced relevance metric, about 0.6, which was also

the value we found produced the most meaningful topics². Because the aim of the study is to explore polysemy and the contextualisations of digitisations, rather than to study hierarchies between words in each topic or between topics we did not consider an analysis of conceptual hierarchies. Instead, the fact that keywords appear across many topics is taken as the starting point for interpreting the many points of contacts between the topics.

4 Keywords, topics and cultural frameworks

The following analysis situates the keywords and topics in various strands of popular culture and fashion. The analysis evolves through each of the topics and considers how they are related in the visualisation. It starts from the visualisation's top right corner with broad topics relating to various appropriations of Viking Age Jewellery. It then proceeds to the top left corner of the visualisation in which the topics pertain to a specific archeological find location and particular types of objects. Finally, the analysis land in the lower part of the visualisation to discuss the ways in which jewellery associated to Vikings become part of jewellery collections with wider scopes.

The heart of the data is represented in the topic **Viking Jewellery**³. The topic's first keyword thor_hammer, is associated with an object with varying symbolic meanings, Figure 3. In Norse mythology, Thor's hammer or Mjölmir, is a magical weapon "that nor would fail or miss, nor would fly so far as not to return" (Skáldskaparmál I, 42, lines 20–34, cited in Knutson (2019, p. 39). Today replicas are worn by white supremacists and performers and fans of black metal. In the latter context, it associates to masculinity, strength and violence and sometimes implies a rejection of Christianity (Thompson, 2018, pp. 144-145).

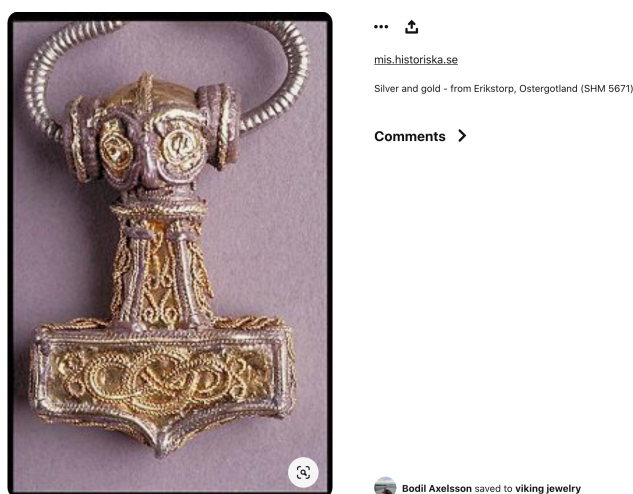


Figure 3: Thor hammer.

However, the topic Viking Jewellery presents a different context, more closely aligned with the ways in which Thor's hammer comes to matter as a playful prop or historicising accessory for Viking enthusiasts and crafters at festivals and heritage sites, cf. Burlingame (2020), and Ashby and Schofield (2015). Among the other keywords are finger_ring, fibula, arm_ring, oval_brooch, tortoise_brooch, disc_brooch, trefoil_brooch, penannular_brooch. They are common grave finds and interpreted by archaeologists in terms of function or as representing regional styles, gender, social or cultural status of the carrier or different phases in the Viking era (Myrberg Bjurström, 2015). For female Viking enthusiasts and reenactors they are essential accessories (Kobialka, 2013; Price, 2019).

A row of keywords points to techniques for producing jewellery: *cast*, *mount*, *plate*, *gilded*, *twisted*, *filigree*. Others point to a measure for weighting objects (*gram*), a mineral for gemstones (*garner*): or

²The visualisation can be found at: <https://www.ida.liu.se/projects/sweclarin/Pinterest-topics/>

³In the following analysis, topics are bolded, keywords are in italics and the three most frequent keywords in each topic are underlined.

figurative motives such as *raven*, *dragon* and *animal-head*. One keyword points to a style (*Borre-style*), and three to cultural or geographical attribution: *anglo_saxon*, *saxon* and *finnish*. Finally, three keywords point to sources for the images (*fotoportalen-unimus*, *Sweden_shm*), and one to an e-commerce site where private persons and businesses exchange goods (*e-bay*).

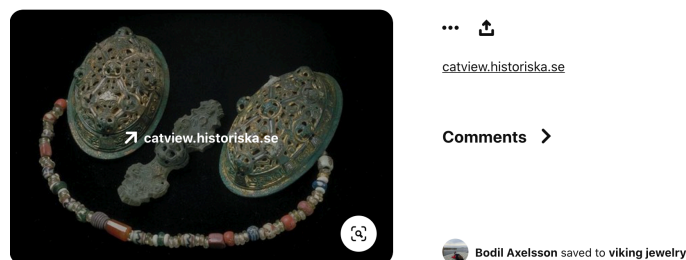


Figure 4: Brooches and pearls.

Imagewise, the topic is characterised by close-ups of pieces of jewellery, giving away ornamentations, shapes and details. Neither the range of different jewellery nor the width of the sources is reflected by the topic modelling method. In addition to the above mentioned types of jewellery the boards display pins, keys, amulets, religious objects, chains with household tools, and sets of Viking jewellery comprised of a pair of brooches and pearls, Figure 4. The artefacts range from simple round needle-pins to figurative objects. Digitisations from SHM are displayed side by side with digitisations from museums and heritage centres in the Nordic countries, Russia and the Baltics. Other sources are The British Museum and the Metropolitan Museum of Art, as well as a Tumblr account dedicated to Russian private collections. Furthermore, images are sourced from auction sites and web shops specialised in museum replicas or web pages that gather images and pieces of information about Vikings, jewellery or clothing.

In the visualisation, **Viking Jewellery** clusters with the topic that appears across the largest number of boards: **Norse Culture**. The keywords in this topic resonate with the masculine warrior cult associated with Vikings in popular culture (*sword*, *ship*, *warrior*, *odin*, *helmet*, *shield*, *thor_hammer*, *axe*, *thor*, *knife*, *battle*). Furthermore, the keywords *rune_stone* and *rune* point to memorials of travel, achievements or ownership (Källström, 2016) or magic (Bäckvall, 2019). Some keywords point to the grounds for knowledge production (*discovered*, *known*), to a Norwegian find location where the remains of two women were found in a ship decorated with wood carvings and several *horses* (*oseberg_ship* and *Oseberg*) and the museum that displays the ship (*museum-oslo*). Others point to locations (*ireland*, *scotland*, *finland*, *finnish*, *iceland*). Finally, the keywords *norse mythology*, *god* and *dragon* add a mythological context.

The images of the signature boards for this topic nuance a strong emphasis of masculine warrior cult that comes forth in the keywords. In the signature boards for this topic, images of artefacts from a range of museums in Scandinavia and the UK mix with images of open graves from archaeological digs, lush green landscapes, national or regional monuments, household utensils as well as interiors and exteriors of recreated buildings such as festive halls. One signature board displays memes comprised of aphorisms or brief facts from Norse Mythology illustrated by present day imagery in the style of the fantasy genre: giants, Valkyrias and warriors as well as mythical creatures such as snakes and dragons. Another signature board instead displays memes created by an American jewellery designer with brief historical “facts”. Because these memes turn text into stylised graphical images, the entirety of their content was not picked up by topic modelling. Consequently, the memes contribute to a wider contextualisation of museum digitisations in the intersection of historical knowledge, mythology and popular culture. All in all, the images suggest that authentic archaeological objects mix and mingle with conceptions of Vikings in romantic fiction, studies of runes and Icelandic sagas (Ward, 2000), and not the least how historical Scandinavians are portrayed as Vikings in heritage tourism, fantasy inspired video games or series and popular history shows on pay television networks (Birkett and Dale, 2019). Importantly, the ways in which the musealisation of archaeological objects, the creation of monuments and heritage sites have been influenced by national sentiments are downplayed in favour of a transnational mix of motives.

The second most prevalent topic is labelled **Re-enactment** to reflect the way words and images here place Viking jewellery in the practice of recreating everyday life of the Viking Era. The most frequent words allude to female coded garments such as *dress* and *apron-dress*, but other keywords point to gender neutral clothing (*tunic* and *coat*) or accessories (*shoes*, *belt*), materials for garments (*wool*, *linen*, *silk*, *textile*), techniques for dress-making (*embroidery*, *trim*, *stich*). Some tokens refer to find locations (*oseberg*, *hedeby*) whereas *ru* refers to Russia where there are active reenactors and Viking finds are prevalent in certain areas. Keywords such as *ship*, *helmet* and *rune* testify to overlaps with the topic **Norse Culture**.

Taken together keywords such as *costume*, *outfit*, *garb*, *replica*, *reconstruction* and *tutorial* allude to the practice of re-enactment. Re-enactors meet at festivals and heritage sites to display and relive the past by immersion in dynamic webs of sensuous experiences (Axelsson, 2010). Material culture such as costumes, jewellery and props is conjured up from historical sources such as archeological finds and it is key to how re-enactors produce a sense of authenticity and immersion into an imagined historical past (Daugbjerg, 2013; Holtorf, 2013; Kobińska, 2013). In this context, the keyword *love* may be interpreted as an expression for admiration of the Viking Age, the care for crafting or the care-taking of men (Karpinska, 2019).

Imagewise this topic is characterised by a blend of digitisations of all sorts of Viking jewellery, images of predominantly young women dressed in long dresses with aprons decorated with brooches and necklaces, men dressed in Viking garb, household utensils, patterns or sketches, closeups of details in the outfits or hairstyles, and not the least scenes from reenactment festivals with women and men cooking, crafting, playing or just posing for the camera in their outfits. The scenes are sourced from both personal blogs and websites for crafters manufacturing and selling reproductions of reenactment gear.

In the visualisation, the topic **Re-enactment** encloses and overlaps the topic **Pre-Christian Europe**. Many keywords in the latter topic point to people, tribes, kingdoms or more broadly speaking cultures that populated, migrated and at times ruled across Europe from the late iron ages to the early middle ages (*celtic*, *saxon*, *celt*, *merovingian*, *téne*, *tene*, *ostrogothic*, *gaul*, *visigothic*, *slavic-kijevian*, *frankish*, *irish*, *ireland*, *culture_frankish*). Similar to Vikings and the Viking Era, contemporary knowledge of these cultures is elusive and rests on archaeological finds, medieval texts and folkloric documents. Taken together the keywords *celtic*, *celts*, *torque* and *celtic_knot* point to the centrality of this culture for the topic. Again, some keywords refer to types of metal artefacts (*fibula*, *torque*, *belt-buckle*, *buckle*), jewellery production (*plate*), or material (*garnet*), including also a Russian term for antique metal collectables *виолити антиквариат*. There are also keywords that refer to sources such as a Norwegian collection database (*fotoportalen-unimus*), a UK-based database for archaeological objects found by members of the public in England and Wales (*antiquity_scheme*, *recorded_portable*), and a St Petersburg-based workshop specialised in replicas of Scandinavian, Celtic and Slavik ancient jewellery (*ruyan_ruyanworkshop*).

Two text tokens – *pagan* and *mjolnir* – point to how Vikings and Celts today nourish contemporary paganism or heathenry, religious beliefs practiced mainly in the US and Europe that take inspiration from Germanic Pre-Christian European religions and adopt them to the present (Strmiska, 2017). In this context, Mjolnir, or Thor's Hammer, often is a symbol for expressing religious identity or faith (Cragle, 2017).

The imagery of the signature boards for this topic displays closeups of artefacts from Viking grave finds with pictures of predominantly Celtic material remains. They display a variety of metal artefacts such as pendants, brooches, and figurines. This means that the ornamental styles associated with Vikings are set side by side with the decorative style connected to the Celtic cultural sphere, for example the torques, a stiff twisted or decorated neck ring. In addition to content highlighted by topic modelling, images display figurative heads and bodies made in sandstone or metal artefacts in the shapes of animals and deities associated with Norse and Celtic mythologies. The context of paganism or heathenry hence comes forth in the motives of some of the displayed artefacts.

In the topic **Lajv and fantasy** keywords like *make*, *deviant_art*, *larp*, *cosplay*, *fantasy* point to several genres for imagining the Viking Age that differs from the search for historicity and authenticity in re-enactment. For example, *lajv* encompasses sewing costumes, arranging props and creating char-

acters for the improvising of a role play set out in a fantasy world (Lundell, 2014, pp 14-15). Similar to cosplay and larp, the genre is indifferent to principles of reality to instead rely its own set of rules often inspired by transmedial storyworlds from video-games and television series (Vu, 2017). In the topic **Lajv and Fantasy**, keywords referring to sources (*thecasparart_deviantart* and *ArtStation*) further points to these contexts, and so do keywords like *diy*, *inspiration*, *recipe*, *armor*, *leather_armor*, *skirt*. Nevertheless, this topic shares many keywords with the four previous topics. The keywords *odin*, *rune*, *helmet*, *warrior*, *dragon* are shared with **Norse Culture**. *Celtic* is shared with **Pre-Christian Europe** and *make*, *deviantart*, *costume*, *dress*, *embroidery*, *leather*, *tutorial*, *tunic* and *love* with **Reenactment**. However, the ranking of the keywords differs.

Imagewise, this topic displays similarities with the topic **Re-enactment** with numerous images of young women dressed in long dresses, aprons and jewellery sets, sourced from personal blogs, Tumblr and Flickr accounts as well as from webpages for roleplay societies or craft entrepreneurs. However, digitisations of viking jewellery or present-day replicas are scarce. This topic instead displays a greater variety of styles, including men in armoury, women in medieval dresses, figures in post-apocalyptic style as well as female shamans and fantasy style drawings of women in warrior outfits.

In addition to these broader topics, the method singled out rings, brooches, pearls, and the Swedish archaeological site Birka by the specificity of their aboutness. The topic **Brooches** mainly displays digitisations of various types of oval brooches from museums in Scandinavia. The topic **Rings** displays neckrings, armrings as well as fingerrings pictured by museums and auction houses together with replicas for sale on etsy. **Birka** is one of many archeological find locations represented in the data from which textiles and jewellery finds have become the basis for reenactors' outfits, cf. Karpinska (2019). Birka, on the island of *Björkö* in the parish of *Adelsö*, has long been of interest to Swedish archeologists (*holger_arbman*; *greta_arwidsson* and *agnes_geijer*). Recent digs have been documented on a webportal run by SHM, thus given the site and re-interpretations of graves and finds increased Internet presence⁴.

The method also singled out a topic that focused the craft of making jewellery. The topic **Metal work** is represented by keywords such as *tutorial*, *knit*, *wire_wrapped*, *diy*, *make*, *scroll_saw*, *рукодел* (needleworker), *making_daily*, *chasing_tool*, *wire_wrapping*, *проволоки* (wire), *fretwork_pattern*, *making*, *технике_wire* (teqnique_wire), *fireplace_screen*; *materials aluminium*, *metal*, *stained_glass*, *beading_gem*, *copper* as well as products *ear_cuff*, *tiara*, *knit_bracelet*, *crown*, and motives *elf_ear*, *elvenear*, *seahorse*, *dragon*. One keyword points to a source (*Youtube*). The signature board testify to that this topic stretches beyond Vikings. It displays a variety of jewellery side by side with tools and procedures for making them.

The topic **Pearls** comes out as more varied than **Brooches** and **Rings**. A row of keywords refers to specific gemstones, materials or techniques for manufacturing pearls (*carnelian*, *mosaic_glass*, *agate*, *blue_glass*, *eye*, *bead_motive*, *melon_bead*, *blue-white phoenecian_glass*). Other keywords attribute pearls to archeological find locations (*Ribe* and *Kaupang*) or styles and cultures (*vikingtid_sted*, *roman*, *phoenecian*, *phoenecian_carthagian*, *islamic*). A second cluster of keywords points to Norwegian museum data bases (*kari_bestillingsnr_lisens*, *fotoportalen_unimus*, *nedre_fotograf*, *gjenstand_perler*, *hammer_åse*), a private not-for-profit museum dedicated to glass and glass making in New York (*corn-ing_museum*), and Hunterian, University Museum Glasgow (*glahm_bead*). The keywords *string* and *stk-halskjedje* refer to how pearls are mounted and *hon_skatten* refers to a large hoard of pearls and gold artefacts held by Oslo Museum of Cultural History. Finally, *roman_mosaics* have been used as a source to how pearls have been worn and the keyword *trade* refers to the economical role of pearls.

In terms of images, the topic mixes digitisations of single monochrome or poly-chrome beads with displays of groups of pearls, pearls on strings to form necklaces or bracelets or pearls in jewellery sets. Many are sourced from museum data bases, but there are also links back to blogs conveying historical knowledge, news-sites and retail websites. The presence of catalogue sheets or sample cards make this topic stand out from all other topics. These cards give an overview of pearls in different shapes and they are sourced from museum data bases as well as from retailers selling replicas of pearls from archaeological digs.

⁴<https://historiska.se/birka/>

Finally, there are three topics that include occasional digitisations of Viking jewellery from SHM in boards that have different geographical and periodical scopes than the European Middle Ages. The topic **Ancient Jewellery** borders the **Pearl** topic and share the tokens *roman* and *pearls*. However, the topics are clearly differentiated in terms of the range of jewellery types, for **Ancient Jewellery** (*earring, cameo, finger_ring, hoop, signet_ring, fibula, badge*), materials (*garnet, pearl, emerald, amethyst, bezel*), and associated techniques (*engraved, enamel, enamelled and cabochon*). Two keywords refer to dating (*circa_century, bce*) and a cluster of keywords attribute the jewellery pieces to what is commonly referred to as the classical Antiquity or Mediterranean empires (*roman, byzantin, hellinistic, etruscan*) and related contemporary nation states (*italy, france, french, greece, greek*). In addition, some pieces of jewellery are attributed to the *scythians*, a nomadic people involved in trade networks connecting Greece, Persia, India and China. The images in the board that best represent this topic display exquisite golden artefacts from museums, auction houses and new sites. The dominance of gold gives away the impression of wealth and exclusivity.

The topic **Shines** instead places occasional SHM digitisations in the context of contemporary jewellery design. Materials and techniques refer to jewellery pieces that are accessible for a broader public (*sterling_silver, copper, sterling, oxidised*). With the exception of diamond, the keywords that refer to gemstones point to cheaper materials (*turquoise, aquamarine, pearl, labradorite, opal, quartz, rose, moonstone*). The focus is on what one can order (*etsy, via_etsy*). But importantly, keywords such as *handmade, handcrafted, artizan_jewellery, wire_rapped* and *hammered* refer to crafted rather than mass-produced goods. In this topic, jewellery is contextualised by keywords associating to fashion styles such as *boho, artisan* and *vintage* and described in contemporary vocabulary such as *ear_ring, dangle_earring* and *bangle*.

The final, and smallest, topic **Jewellery and trade** is placed in the periphery of the top left corner in the visualisation. Drawing together keywords for attribution such as *yemen, tibetan, hilltribe, tibet, antique_african, indian, african* and *morocco* it points to places outside Europe and North America. Furthermore, keywords point to materials in brown, yellow or variety of nuances of red (*dzi_bead, pema_raka, cinnabar, coral*) and other materials such as *marble* and *celluloid*, and a Japanese bead for fastening cords often produced of ivory (*asian_ojime*). Some of the keywords point to contemporary jewellery stores and workshops that specialise in antique jewellery and gemstones (*lewis_clark, columbia_river, mitchell, dorje_design*), a Canadian retail business group (*hudson_bay*), as well as to how antique beads today are for sale (*sample_card*) and are ideal gifts (*card, blank_card*). Finally, the topic includes keywords that places beads and pearls in the context of colonialism and the exploitation of African resources (*antique_venetian, african_trade, venitian_trade, trade, venitain_sample*). Glass beads produced in Venice, an early center for Asian-European trade, were used as a cheap and easily produced currency in the exchange for African raw materials and slaves⁵. Thus the topic associates to historical transactions between Europe, Asia and Africa at the same time as it testify to contemporary aesthetisation and commodification of heritage pieces and replicas thereof.

5 Conclusion

The interpretation of new contexts for Viking Age Jewellery sourced from SHM presented in this paper relies not only on how meaning emerge as relations between each topic's keywords, but also on the ways in which images are collected in boards and how keywords and images can be situated culturally. It turned out that the topics and the keywords provided a valid tool for exploration and further interpretations of new contexts for SHM digitisations. In fact, the way in which topic modelling enabled a reduction and mapping of only a share of an abundantly rich empirical material – the users' descriptions – provided a necessary first step for further interpretative work. Topic modelling was especially valuable for investigating polysemy as many topics shared some keywords. It should also be noted, and in line with the method (DiMaggio et al., 2013), many boards mix topics. This could be taken as an evidence of unexplored additional polysemy. It should also be noted that the interpretation of keywords required clustering them thematically, rather than relying on their probabilistic distribution within the topic. Be-

⁵<http://www.vam.ac.uk/content/articles/t/trade-bead>

cause pinners only to a less extent acknowledge sources in their descriptions, the method did not fully account for the variety of sources within each topic. This may be further investigated by analysing the images' url:s more systematically. Nevertheless, the highlighted sources direct the attention to the fact that the contexts for SMHs digitisations include various museum databases, re-enactors' blogs and websites, news sites, and not the least websites of a variety of businesses selling both replicas and original jewellery.

The inspection of the signature boards reveals that there are some discrepancies in the relations of the topics in the visualisation and how relations between topics come forth in the cultural analysis. Taken together the cluster of large topics displayed at the top left corner of the visualisation singles out **Viking Jewellery** as a consistent topic first and foremost overlapping with the topic Norse Culture. But this does not do justice to the centrality of Viking jewellery for the topic of **Re-enactment**; the many depictions of **Viking jewellery** in the signature boards and the ways in which digitisations of grave finds serve as models for replicas in this genre for performing past times. Furthermore, culturally, there are more overlaps between the topics **Norse Culture, Pre-Christian Europe and Lajv and Fantasy** than the visualisation reveals. They are connected by influences of early Germanic mythologies, in particular the various symbolical values attached to Thor's hammer and mythological creatures. Thus, genres for imagining the past as well as contemporary conceptions of historical cultures and mythologies leak. It may also be more relevant to discuss **Rings, Brooches, Pearls and Birka**, all very specific topics singled out by the method and projected as separate in the visualisation, as subtopics, also present in the larger topics as these also display these types of objects. The visualisation of **Ancient Jewellery, Shinies and Jewellery and trade** as detached from the topics in the visualisation's left right corner is consistent with the cultural analysis. However, the political and economical perspectives that come forth in the analysis of **Jewellery and trade** may also be relevant to apply to all the other topics. This very small topic, which at first, may seem an odd context for Viking jewellery, serves as a reminder of the fact that the materials for Viking jewellery historically were part of Early-Medieval relations involving trade and conquest.

To sum up, on Pinterest, the national context of digitisations of Viking Age jewellery in SHM:s collection management system is replaced by several transnational contexts. Archaeological frameworks give way for contexts in which Viking Age jewellery is appreciated for its symbolical meanings and decorative functions in contemporary genres for reimagining, reliving and performing European pasts and mythologies. The emerging contexts on Pinterest also highlight the business opportunities involved in genres such as re-enactment, neo-paganism, lajv and fantasy. The boards are clues to how digitisations serve as prototypes for replicas.

References

- Steven P. Ashby and John Schofield. 2015. 'hold the heathen hammer high': representation, re-enactment and the construction of 'pagan' heritage. *International Journal of Heritage Studies*, 21(5):493–511.
- Bodil Axelsson. 2010. History on the web: museums, digital media and participation. In Anders Ekström, Solveig Jülich, Frans Lundgren, and Per Wisselgren, editors, *History of participatory media: politics and publics*, pages 158–171. London: Routledge.
- Anja Bechmann and Geoffrey C Bowker. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, January–June:1–11.
- Jeffrey M. Binder. 2016. Alien reading: Text mining, language standardization, and the humanities. In Matthew K. Gold and Lauren F. Klein, editors, *Debates in the Digital Humanities 2016*, pages 201–217. Minneapolis: University of Minnesota Press.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Tom Birkett and Roderick Dale. 2019. *The Vikings Reimagined. Reception, recovery, engagement*. Berlin: Walter de Gruyter.
- David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

- Katherine Bode. 2019. *A world of fiction. Digital collections and the future for literary history*. Ann Arbor: University of Michigan Press.
- Chiara Bonacchi and Marta Krzyzanska. 2019. Digital heritage research re-theorised: Ontologies and epistemologies in a world of big data. *International Journal of Heritage Studies*, 25(12):1235–1247.
- Megan A. Brett. 2012. Topic modelling: a basic introduction. *Journal of Digital Humanities*, 2(1).
- Catherine Burlingame. 2020. *Dead landscapes – and how to make them live*. Lund: Lund University.
- Maja Bäckvall. 2019. “pick up the rune”: the uses of runes in digital games. In Tom Birkett and Roderick Dale, editors, *The Vikings Reimagined. Reception, recovery, engagement*, pages 201–213. Berlin: Walter de Gruyter.
- Fiona Cameron and Helena Robinson. 2007. Digital knowledgescapes: cultural, theoretical, practical, and usage issues facing museum collections databases in a digital epoqe. In Fiona Cameron and Sarah Kenderdine, editors, *Theorizing digital cultural heritage: a critical discourse*, pages 165–191. Cambridge, Mass: MIT Press.
- Joshua Marcus Cragle. 2017. Contemporary germanic/norse paganism and recent survey data. *Pomegranate*, 19(1):77–116.
- Mads Daugbjerg. 2013. Patchworking the past: materiality, touch and the assembling of ‘experience’ in american civil war re-enactment. *International Journal of Heritage Studies*, 20(7-8):724–741.
- Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modelling and the sociological perspective on culture: application of newspaper coverage of u-s governments arts funding. *Poetics*, 41(6):570–606.
- Johanna Drucker. 2017. Why distant reading isn’t. *PMLA*, 132(3):628–35.
- Johanna Drucker. 2018. Non-representational approaches to modeling interpretation in a graphical environment. *Digital Scholarship in the Humanities*, 33(2):248–263.
- Catherine Hall and Michael Zarro. 2012. Social curation on the website pinterest. com. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–9.
- Cornelius Holtorf. 2013. The time travellers’ tools of the trade: some trends at lejre. *International Journal of Heritage Studies*, 20(7-8):782–797.
- Klaudia Karpinska. 2019. Women in viking reenactment. In Tom Birkett and Roderick Dale, editors, *The Vikings Reimagined. Reception, recovery, engagement.*, pages 69–88. Berlin: Walter de Gruyter.
- Jenny Kidd. 2014. Museums in the new mediascape: Transmedia, participation. *Ethics*.
- Barbara Kirshenblatt-Gimblett. 1998. *Destination Culture: Tourism, Museums, and Heritage*. Univ of California Press.
- Sara Ann Knutson. 2019. The materiality of the myth: divine objects in norse mythology. *Temenos*, 55(1):29–53.
- Dawid Kobialka. 2013. The mask(s) and transformers of historical re-enactment: material culture and contemporary vikings. *Current Swedish Archeology*, 21:141–161.
- Magnus Källström. 2016. Viking age runes. In Gunnar Andersson, editor, *We call them Vikings*. Stockholm: Historiska museet.
- David C. Liu, Stephanie Rogers, Raymond Shiau, Dmitry Kislyuk, Zhigang Zhong Kevin C. Ma, Jenny Liu, and Yushi Jing. 2017. Related pins at pinterest: The evolution of a real-world recommender system. In *International World Wide Web Conference Committee (IW3C2), Perth, Australia*.
- Alan Liu. 2013. The meaning of the digital humanities. *PMLA*, 128(2):409–423.
- Stine Lomborg and Anja Bechmann. 2014. Using apis for data collection on social media. *Information society*, 30(4):256–265.
- Erika Lundell. 2014. *Förkroppsligad fiktion och fiktionaliserade kroppar. Levande rollspel i Österjöregionen*. Stockholm: Acta Universitatis Stockholmiensis.

- Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. 2018. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118.
- Daniel Maier, Andreas Niekler, Gregor Wiedemann, and Daniela Stoltenberg. 2019. How document sampling and vocabulary pruning affect the results of topic models. *OSF Preprints*, 20 Nov. 2019, 11.
- Nanouschka Myrberg Bjurström. 2015. Things of quality: possessions and animated objects in the scandinavian viking age. In Alison Klevnäs and Charlotte Hedenstierna-Jonson, editors, *Own and be owned: archaeological approaches to the concept of possession*, pages 23–48. Stockholm: Department of Archaeology and Classical Studies.
- David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 100–108, USA. Association for Computational Linguistics.
- Neil Price. 2019. My vikings and real vikings: drama, documentary, and historical consultancy. In Tom Birkett and Roderick Dale, editors, *The Vikings Reimagined. Reception, recovery, engagement*, pages 28–43. Berlin: Walter de Gruyter.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Benjamin M. Schmidt. 2012. Worlds alone: dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1).
- Nick Seaver. 2015. The nice thing about context is that everyone has it. *Media, Culture & Society*, 37(7):1101–1109.
- Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June. Association for Computational Linguistics.
- Michael F. Strmiska. 2017. Paganism and politics: a view from central-eastern europe. *Pomegranate*, 19(2):166–172.
- Christopher Thompson. 2018. *Norges våpen. Cultural memory and uses of history in Norwegian Black Metal*. Studia Historica Upsaliensia. Uppsala: Acta Universitatis Upsaliensis.
- Ryan Vu. 2017. Fantasy after representation: D&D, game of thrones, and postmodern world-building. *Extrapolation: A journal of science fiction and fantasy*, 58(2-3):273–301.
- Andrew Ward. 2000. *The Vikings and the Victorians: inventing the Old North in nineteenth-century Britain*. Cambridge: Brewer.
- Fiona Wilson. 2016. Queens of collection and curation: Pinterest use in the society for creative anachronism. Master's thesis, School of Information Management, Wellington University, New Zealand.
- Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

“Tea for two”: the Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure

Federico Boschetti
ILC “A. Zampolli” CNR, Pisa
& VeDPh, Venezia, Italy
federico.boschetti
@ilc.cnr.it

Riccardo Del Gratta
ILC “A. Zampolli” CNR
Pisa, Italy
riccardo.delgratta
@ilc.cnr.it

Monica Monachini
ILC “A. Zampolli” CNR
Pisa, Italy
monica.monachini
@ilc.cnr.it

Marina Buzzoni
ALIM, Università Ca’ Foscari
Venezia, Italy
mbuzzoni
@unive.it

Paolo Monella
ALIM, Sapienza Università
di Roma, Italy
paolo.monella
@uniroma1.it

Roberto Rosselli Del Turco
ALIM, Università degli
Studi di Torino, Italy
roberto.rossellidelturco
@unito.it

Abstract

This paper aims at showing how integrating the Archive of the Italian Latinity of the Middle Ages (ALIM) into the ILC4CLARIN repository can provide mutual benefits. Making ALIM available to a large community of scholars and researchers, on the one side, represents the first step to reduce the lack of resources for Medieval Latin in CLARIN and, on the other side, constitutes an unprecedented contribution to not only linguistic investigations, but also to the studies of the culture and science at the basis of the Western European society. The paper describes the adopted approach aiming to keep intact the structure of the archive and its metadata, which are both accurately mirrored into the ILC4CLARIN repository in order to maintain existing access practices of the users. This structure can be found in exactly the same state within the CLARIN VLO. Finally, the paper illustrates the advantages of experimenting with some ALIM data, once introduced within the CLARIN Language Resource Switchboard service: first results are shown from the analysis of some texts with the UDPipe tool suite and the distant reading tool Voyant.

1 Introduction

The Archive of the Italian Latinity of the Middle Ages – in Italian, Archivio della Latinità Italiana del Medioevo (ALIM) – is an Italian national research project aimed to provide free online access to a large number of Latin texts produced in Italy during the Middle Ages. ALIM makes an unprecedented contribution to the study not only of Latin, but also of the culture and science at the basis of the Western European society. For several centuries, in fact, Latin represented the only language in which many of the major creations of thought, science, and literature of the Middle Ages were expressed. Even when national languages imposed themselves in written form, Latin never lost its role and prestige as a transnational language – until the end of the Middle Ages and beyond.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Federico Boschetti, Riccardo Del Gratta, Monica Monachini, Marina Buzzoni, Paolo Monella and Roberto Rosselli Del Turco 2021. “Tea for two”: the Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 37–46.

The general aim of this paper is to place ALIM within the framework of CLARIN-IT and CLARIN at large. Section 2 shows how ALIM may contribute to fill an important gap in the availability of literary and historical sources for Latin: query searches run on the Virtual Language Observatory for Latin-related resources demonstrate that no resource with the features and potentialities of ALIM is currently available. The internal structure and the metadata of the Archive are presented in Section 3, while the strategy for the integration of ALIM into the ILC4CLARIN repository is discussed in Section 4. Finally, ALIM's contribution in strengthening and widening the research directions in CLARIN-IT and its advantages for the CLARIN community are presented.

2 Latin resources in CLARIN-IT

The Italian CLARIN (CLARIN-IT) consortium¹ (Nicolas et al., 2018) has a strong interest in the field of Digital Classics, which is still affected by a shortage or restricted availability of language resources for historical languages such as Ancient Greek and Latin. To this end, the consortium aims to make some of the existing digital resources for Ancient Greek and Latin available through its national repository – ILC4CLARIN (Pisa).

Within the CLARIN-IT consortium, the collaboration between the Centre for Comparative Studies “I Deug-Su”, Department of Philology and Literary Criticism at the University of Siena – DFCLAM² and the ILC4CLARIN data center mostly concerns the study of methods and the development of services to offer online secure access to some digital archives of literary and historical texts. ALIM, currently hosted by the University of Siena³, is the largest digital library of the Italian Latinity, including both literary and documentary sources.

Evidence shows that the CLARIN data centers do not offer resources such as ALIM. The faceted-search functionality of the Virtual Language Observatory (VLO), performed combining *Latin text resource* and *Middle Ages*, returns 53 records, while 124 records are returned by the query *Latin* combined with the adjective *medieval*. These data are essentially images of manuscripts; no XML-TEI texts seem to be available by using these search keys. A further query with XML as ‘free text’, Latin as ‘language’, and text or corpora as ‘resource type’ returns about 1300 records, mostly consisting in Treebanks or in documents coming from the EUROPEANA platform⁴.

ALIM represents the first step to reduce the lack of resources for Medieval Latin in CLARIN-IT and eventually in CLARIN. On the one hand, ALIM will offer high-quality data since: (i) the resources are curated by domain experts; (ii) a strong organization is dedicated to maintenance; (iii) the resources cover a broad historical period; (iv) the resources are TEI encoded. On the other hand, the Language Resource Switchboard and the Weblicht workflow engine will use the texts provided by ALIM to produce both engaging visualizations and interesting linguistic analyses.

3 ALIM: history, goals, structure

ALIM is an archive of medieval Latin texts composed in the Italian area between the 8th and 15th centuries. It originated as a UAN (Unione Accademica Nazionale) project in the Nineties and was later supported by the national Ministry of Education. Its original aim was twofold: to make medieval Latin literature texts openly available and to provide a textual corpus serving as a basis to create a new dictionary of medieval Latin in its Italian variety. The latter goal explains a unique feature of ALIM: it does not only include literary sources, but also collections of documentary texts. ALIM is, therefore, divided into two sections: “Fonti letterarie” and “Fonti documentarie”. While the majority of texts are drawn from printed editions, some are new, born-digital editions⁵.

¹The composition of the Italian Consortium is available at <http://clarin-it.it/en/content/consortium>.

²Prof. Francesco Vincenzo Stella.

³<http://alim.unisi.it/il-progetto/>

⁴<http://www.europeana.eu>

⁵More information on the history and scientific objectives of ALIM, with further bibliography, are in (Alessio, 2003); (Buzzoni and Rosselli Del Turco, 2016, par. 7.1.2); (Ferrari, 2017) and (D'Angelo and Monella, 2019).

3.1 From ALIM1 to ALIM2, from ALIM2 to CLARIN: text and metadata

Until 2016, ALIM was hosted by the servers of the University of Verona, Italy⁶ and its texts were annotated with procedural markup, based on simple HTML markers. We shall refer to this version as “ALIM1”.

In 2016, the current version of the archive (“ALIM2”) was launched. The migration process involved the following tasks: (1) building a new open source software TEI XML-based digital library infrastructure and publishing it on the servers of the University of Siena⁷; (2) re-encoding text markup and metadata in TEI XML P5.

Task 1 was realised in collaboration with the external IT company Net⁸ and completed in 2016/17, when the ALIM2 website was launched. Task 2 involved a longer process, still ongoing, curated by the “équipe di codifica” (Ferrarini, Monella and Rosselli Del Turco) to gradually improve the level of formalisation and the granularity of text markup and metadata.

In the current version of the archive, each literary text is encoded as a TEI XML P5 file with a <TEI> root element, while in the documentary section, each TEI XML file includes a whole volume of a documentary collection⁹, has a <teiCorpus> root element and includes each individual document in a separate <TEI> element. In the latter case, both <teiCorpus> and <TEI> have their own <teiHeader> with metadata respectively regarding the whole collection and the individual document.

In the ALIM2 TEI-XML files for literary texts deriving from the initial export from ALIM1 (labeled as “encoding level ALIM2_0”), much metadata was still included in unstructured <note> elements of the TEI. Also, most texts lacked any TEI structural markup such as <div>. In 2017/18, literary texts were gradually upgraded to “encoding level ALIM2_1”, thanks to the work of ALIM collaborator Chiara Casali on metadata integrity and of Jan Ctibor on metadata encoding and structural markup. Ctibor’s activity was brought forth in the framework of a collaboration agreement between ALIM and the *Corpus Corporum*¹⁰, the largest full-text repository for Latin (163 M words). The current policy of ALIM requires that all new texts included in the archive must be encoded at “level ALIM2_2”: this includes markup of work titles, quotes, speeches, persons, or place names.

The archive also includes born-digital scholarly editions directly based on handwritten medieval witnesses, whose encoding level is labeled as ALIM2_3¹¹.

The ALIM project provided CLARIN-IT with the TEI headers of the XML files in the archive, at the highest available encoding level, to extract metadata from them.

4 ALIM in CLARIN-IT

4.1 Structure for ALIM data into ILC4CLARIN repository

As described in Section 3, the ALIM digital library is arranged into two complementary sections: *Fonti Letterarie (Literary Sources)* and *Fonti Documentarie (Documentary Sources)*. The former is a collection of single documents (about 350), while the latter is a collection of 50 corpora that groups about 6455 texts. Since ALIM keeps these two resources separated, we decided to mirror this structure in the ILC4CLARIN repository. We created two collections, *Literary Sources* and *Documentary Sources*, under the *OPEN* community¹². This structure is important for, at least, two reasons. The first underlying motivation for this partially conservative choice was that this decision would provide philologists, linguists, and historians with a user experience on the VLO consistent with the navigation in the original ALIM environment. The second reason is directly con-

⁶<http://www.alim.df11.univr.it/>

⁷<http://alim.unisi.it/>

⁸<https://www.netseven.it/>

⁹E.g.: *Codex diplomaticus Cavensis*, volume 1: http://alim.unisi.it/dl/fonte_documentaria/7381.

¹⁰<http://www.mlat.uzh.ch/MLS/>

¹¹See <http://alim.unisi.it/collection/nuove-edizioni-editiones-principes-e-prime-trascrizioni/> for a list of such editions. In general, on markup levels see the *Manuale di codifica dei testi ALIM in TEI XML* in <http://alim.unisi.it/documentazione/>

¹²Given that ILC4CLARIN uses the clarin-dspace repository, we have used the terminology community and collections. For clarity, collections are nested into communities.

nected with the VLO. In section 2, we briefly mentioned the faceted-search of the VLO. One of such facets is the collection (in the original repository) the data come from. The ALIM data are retrieved from the VLO using either “fq=collection:ALIM+Literary+Sources&fqType=collection” or “fq=collection:ALIM+Documentary+Sources&fqType=collection”¹³.

4.2 Population of the repository with ALIM data

The about 350 *Literary Sources* have complete descriptive metadata, although period, author and title are often debated in the scholarly community and, therefore, tentative in the collection. Author names have two issues: the actual authorship attribution and alternative Latin spellings of the name. Titles too are not always standardised, and the very identification of the “work”, as well as of the composition period, is problematic. However, each of these metadata fields has a value in ALIM (for the author, it can also be “Anonimo”). The 50 corpora of *Documentary Sources* group 6455 small documents. For these small documents the metadata set differs from *Literary Sources*, since they do not represent a creative work by an author. For example, private documents are actually written by a notary, but their “author” is the stakeholder (the person who buys, sells etc.), while charters are created by a public institution.

As a consequence, we decided to completely import *Literary Sources* metadata into the repository, but, at the same time, to describe only the 50 corpora of *Documentary Sources*, without importing the whole amount of data (even if technically possible).

The ratio behind this decision is related to the ALIM organization again. As noted in Section 3.1, the TEI version of each document in literary sources has its own `<teiHeader>`, corresponding to the TEI root element, that can be parsed. While for documentary sources the most informative `<teiHeader>` is extracted from `<teiCorpus>`, for literary sources metadata are extracted from the header of each files’ `<TEI>` element.

Given the large number of items to describe in the repository, we decided to use the import functionality of the repository¹⁴ to batch-load the items. Since this procedure is unsupervised, as far as the content of the items is concerned, we decided to manually create a prototypical item, export it, and automatically clone it. In this way, every item is syntactically correct and can be safely imported into the repository. More in detail: (i) we took one document from literary sources and one from documentary works and kept them as prototypes; (ii) we carefully created a submission, mapping the elements of the `<teiHeader>` into the fields of the submission form of the repository; and (iii) once the internal workflow of metadata quality is passed, we exported the item.

The exported item is an archive which contains the following metadata files: **metadata_local.xml**, **dublin_core.xml**, and **metadata_metashare.xml**. All of them are populated with data extracted from elements of the `<teiHeader>`. The different metadata files combine to create the descriptive items in the repository. The ALIM research team checked sample metadata from the CLARIN archive and verified that they correspond to those included in the TEI headers of the ALIM XML files and to the general project information pertaining to the archive. It is important, here, to notice that the official URL of the ALIM project (in our case, <http://it.alim.unisi.it/>) is contained in the **dublin_core.xml** files, while **metadata_local.xml** files contain the *demo URL*, that is to say where the resource stays in the ALIM digital library.

This mapping enforces our decision to describe the `<teiCorpus>` instead of describing every single document in the corpus. Literary Sources have a clear URL where the document resides: for example, the “Dialogus” by Gerius Aretinus is available at <http://it.alim.unisi.it/dl/resource/194>. By contrast, Documentary Sources point to URLs that report the whole corpus. For example, the “Codex diplomaticus Cavensis - 01” is available at http://it.alim.unisi.it/dl/fonte_documentaria/7381. On the web page, a JavaScript function allows the user to jump to the desired documents, such as the 27th document, whose internal URL is http://it.alim.unisi.it/dl/fonte_doc

¹³At the time of writing, only the *Literary Sources* have been imported into the ILC4CLARIN production repository. The items are available at <https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-c0-111/130>.

¹⁴<https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simpl e+Archive+Format>

umentaria/7381#doc_27. Unfortunately, ‘#’ is a reserved character¹⁵ which separates information sent to server from client side actions, and no data transmitted as part of the URL must contain it. The complete mapping guide, the scripts, and XSLT style sheets are available at <https://github.com/cnr-ilc/alim2clarin-dspace>.

As an example, Figure 1 shows the different elements in the descriptive item mapped to their sources in the TEI header.



Figure 1: Repository item and its sources in the TEI Headers.

Before concluding this section, let us provide some information on the licence of the data. The original data, contained in the ALIM digital library, are released under the Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND4.0). To stay compliant with the original license, the loading procedures adds this information to the descriptive items, as reported in Figure 2.

dc.rights.uri	http://creativecommons.org/licenses/by-nc-nd/4.0/
dc.rights.label	PUB

Figure 2: Additional metadata inform about licences and IPR.

4.3 Versioning

The ILC4CLARIN repository implements the versioning of the described items. Indeed, it is always possible to add to the repository a new item as “new version of” a previous one. The versioning of the items on the repository should be consistent with the one on the ALIM digital library. The latter allows contributors to replace the XML-TEI file of a literary work or documentary collection with a new one, including changes in the text or in the metadata. The ALIM2 digital library keeps all previous XML files available in the backend but only makes the last one (and the derivative HTML, PDF, and plain text files) available to the user.

¹⁵<https://www.urlencoder.io/learn/>

To make the versioning of the ILC4CLARIN repository consistent with that of ALIM, we decided to remove the *demo URL* from the old versions. In this way, users access the latest version of the document from the repository and, if they still need older data, they can contact ALIM and request them.

5 CLARIN Services and ALIM

5.1 Available analysis tools

In this section, we show the results of an analysis of an ALIM sample text with tools made available through the CLARIN infrastructure.

The Language Resource Switchboard (LRS)¹⁶ (Zinn, 2018) has been used to connect the input data with suitable and available tools. Figure 3 below shows the suggested tools for the input file “Historia Langobardorum”¹⁷, which consists of about 38000 words.

The screenshot displays the LRS interface. At the top, under 'Resources', there is a card for 'Paulus Diaconus - Historia Langobardorum(1).txt' (264.08 KIB). To the right, 'Mediatype' is set to 'text/plain' and 'Language' is set to 'Latin'. Below this, the 'Matching Tools' section is visible, with a 'Group by task' checkbox checked and a 'Search for tool' input field. Three tool categories are expanded: 'Dependency Parsing' with 'UDPipe' (LINDAT logo), 'Distant Reading' with 'Voyant Tools' (Voyant logo), and 'Text Analytics' with 'WebLicht Advanced Mode' (WEBLIGHT logo).

Figure 3: “Historia Langobardorum” and connected tools.

The LRS lists a distant reading tool, Voyant¹⁸ and a dependency parsing tool, UDPipe (Straka and Straková, 2020; Straka and Strakova, 2017).

Figure 4 displays the *Cirrus* and *TermsBerry*, while Figure 5 provides some textual statistics on the examined text¹⁹.

¹⁶<https://switchboard.clarin.eu/>

¹⁷<http://hdl.handle.net/20.500.11752/OPEN-152>.

¹⁸<https://voyant-tools.org/>

¹⁹We invite the interested readers to replicate the experiment by providing the URL <http://alim-admin.unisi.it/download.txt?id=201> to the LRS and eventually use *Voyant tools*.

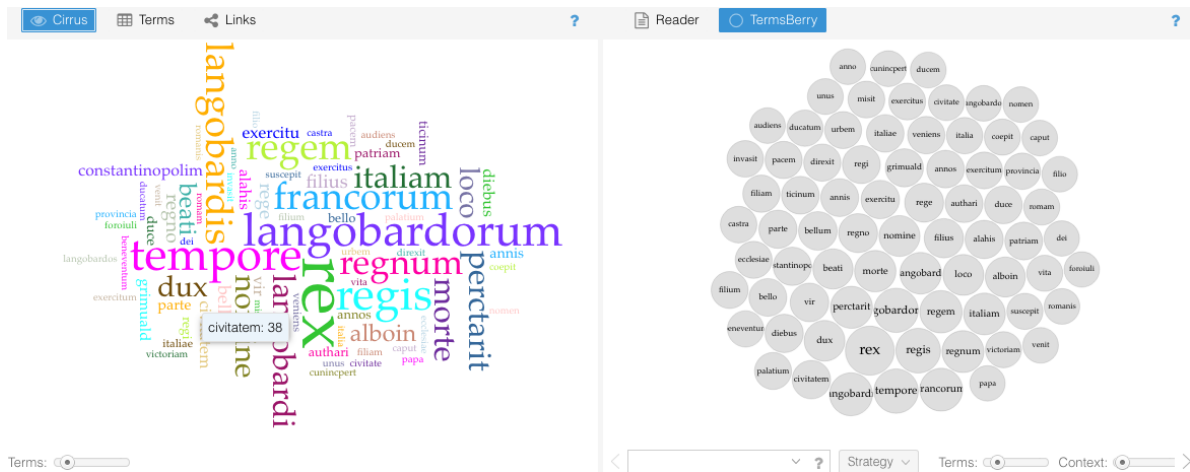


Figure 4: Cirrus and TermsBerry



Figure 5: Statistics of the input texts.

The dependency parser tool, UDPipe, is both available as a service²⁰, from LINDAT/CLARIAH-CZ²¹, and as an integrated service into the WebLicht workflow engine²² (Hinrichs et al., 2010). The following figures show the analysis of the sentence “Inter haec moritur Godehoc, cui succedit Claffo, filius suus”²³, extracted from the same text, performed using the UDPipe parser from LINDAT/CLARIAH-CZ. The model used is the *latin-llct-ud-2.6-200830*.

²⁰<https://lindat.mff.cuni.cz/services/udpipe/info.php>.

²¹<https://lindat.mff.cuni.cz>.

²² Again here, we recommend the interested readers at replicating the experiment, focusing on the different access modalities.

²³ With more complex sentences the results might not be completely correct, both in terms of lemmatization and syntactic trees.

# text = Inter haec moritur Godehoc, cui successit Clafoo, filius suus									
1	Inter	inter	ADJ	AAXXX----1A----	Degree=PoslForeign=YeslPolarity=Pos	0	root	_	TokenRange=0:5
2	haec	haec	NOUN	AAXXX----1A----	Degree=PoslForeign=YeslPolarity=Pos	1	flat:foreign	_	TokenRange=6:10
3	moritur	moritur	NOUN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNumber=SinglPolarity=Pos	1	flat:foreign	_	TokenRange=11:18
4	Godehoc	Godehoc	PROPN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNameType=GivlNumber=SinglPolarity=Pos	1	flat:foreign	_	SpaceAfter=NoI TokenRange=19:26
5	,	,	PUNCT	Z:-----	_	7	punct	_	TokenRange=26:27
6	cui	cui	PRON	PQ--4-----	Animacy=InanlCase=AcclPronType=Int,Rel	7	obl:arg	_	TokenRange=28:31
7	successit	successit	VERB	VpYS---XR-AA---	Aspect=PerflGender=MascplNumber=SinglPolarity=PoslTense=PastlVerbForm=PartlVoice=Act	3	acl:reicl	_	TokenRange=32:41
8	Clafoo	Clafoo	PROPN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNameType=GivlNumber=SinglPolarity=Pos	7	nsubj	_	SpaceAfter=NoI TokenRange=42:48
9	,	,	PUNCT	Z:-----	_	10	punct	_	TokenRange=48:49
10	filius	filius	NOUN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNumber=SinglPolarity=Pos	8	appos	_	TokenRange=50:56
11	suus	suus	NOUN	NNIS1-----A----	Animacy=InanlCase=NomiGender=MascplNumber=SinglPolarity=Pos	10	nmod	_	SpaceAfter=NoI TokenRange=57:61

Figure 6: A sample UDPipe table.

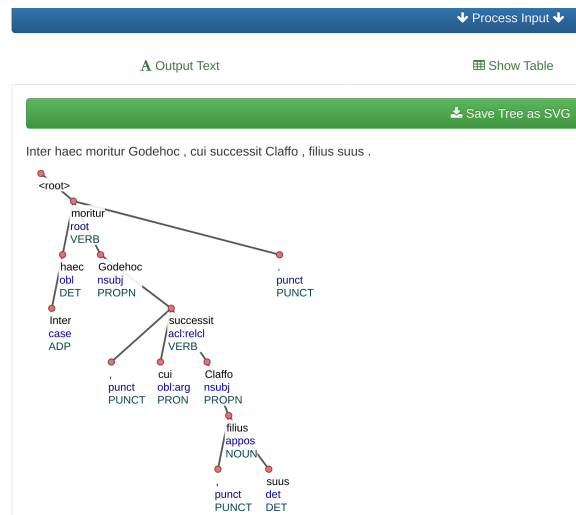


Figure 7: A sample UDPipe tree.

A few words on the use of WebLicht for the ALIM texts. As a workflow engine, WebLicht can run as many tools as needed. Figure 8 reports the analysis chain which can be run on the same sentence. The chain consists of three modules: a tokenizer, a tagger, and a parser. The tokenized text can be further used in the Federated Content Search (Stehouwer et al., 2012), cf. Section 6.

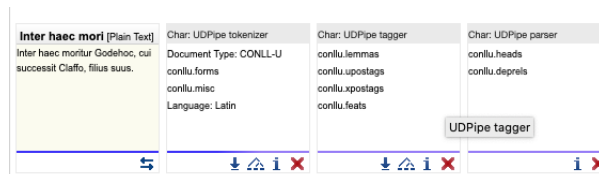


Figure 8: WebLicht analysis chain.

5.2 Are similar resources available?

The “Historia Langobardorum” was already available from the VLO as a digital facsimile of the medieval manuscript provided by e-codices, the Virtual Manuscript Library of Switzerland (<https://>

[//www.e-codices.ch/en/csg/0635/1/0/](http://www.e-codices.ch/en/csg/0635/1/0/)). Even if the direct source of the text of “Historia Langobardorum” in ALIM is the digitization of the nineteenth-century edition by L. Bethman - G. Waitz²⁴, the manuscript available from the VLO is one of its witnesses. Indeed, the increased use of the IIF (www.iiif.io) framework to publish digital facsimiles of medieval manuscripts means that for many of the ALIM texts there are already one or more digitized witnesses freely available on the Web. Not only that, this opens the way for a future connection of an ALIM text to those facsimiles: for the most advanced ALIM critical editions, those belonging to level 3 of the encoding, i.e. a full use of the elements of the Critical Apparatus TEI module (cf. Section 3.1), this could also mean generating automatically the corresponding witness and linking it to the digitized images of the manuscript. s

6 Concluding Remarks

The DFCLAM committed itself to offering data and free online access to some digital archives of literary and historical texts: one of them is ALIM (the Archive of the Italian Latinity of the Middle Ages), the largest digital library of the Italian Latinity including both literary and documentary texts, encoded in TEI XML from philologically checked printed editions or published directly from manuscripts produced in Italy during the Middle Ages, in new born-digital scholarly editions. Strategies for importing the metadata of ALIM in the ILC4CLARIN repository through a shared TEI header are under study, as well as procedures for delivering dedicated tools for textual and linguistic analysis through the CLARIN channels. This would allow meta-queries and cross-queries on semantic items which could connect Latin and modern European languages derived from Latin and allow to develop semantic trees and networks of lexical derivations at the very heart of the European shared vocabulary.

ALIM complements the Latin resources in CLARIN by providing access to a large corpus of medieval literary and documentary Latin texts with granular curated metadata. On the other hand, participating in CLARIN provides ALIM with a valuable opportunity in terms of sustainability, long term preservation, persistent identification, format migration, and visibility for the ALIM research outputs. The VLO makes the resources produced and described in the ILC4CLARIN repository, including ALIM metadata, available to a wider audience in the SSH community, while the CMDI model ensures high quality metadata curation. Also, CLARIN offers ALIM the possibility to use technology and text analysis tools available at CLARIN data centers to deal with multilingual data. For example, Weblicht allows to combine web services so as to handle and exploit textual data, while the Language Resource Switchboard can connect the ALIM texts to visualization tools such as Voyant. A further reciprocal advantage is that CLARIN contributed in enhancing the ALIM strategies on Open Access and open source policies by supporting ALIM in planning the actions necessary to provide FAIR (Findable, Accessible, Interoperable, Reusable) data (de Jong et al., 2018).

Finally, ILC4CLARIN is an endpoint of the Federated Content Search (FCS)²⁵, a tool to query data distributed across local collections available at the various CLARIN centres. At the time of writing, the only CLARIN collection including Latin texts (`lat_wikipedia_2012_100K`, about 1.5M tokens) is available at the FCS aggregator from the Automatische Sprachverarbeitung - Universität of Leipzig. The addition of further sources to the aggregator will be of fundamental importance to increase the number of high-quality Latin texts available through the Federated Content Search.

²⁴MGH SS rer. Lang., Hannover 1878, pp. 45-18

²⁵<https://contentsearch.clarin.eu/>.

References

- Gian Carlo Alessio. 2003. Il progetto alim (archivio della latinità italiana del medioevo). In Francesco Santi, editor, *In Biblioteche elettroniche. Letture in Internet: una risorsa per la ricerca e per la didattica*, volume 1, pages 73–81. SISMEL - Edizioni del Galluzzo.
- Marina Buzzoni and Roberto Rosselli Del Turco. 2016. Evolution or revolution? digital philology and medieval texts: History of the discipline and a survey of some italian projects. In *Mittelalterphilologien heute. Eine Standortbestimmung. Band 1: Die germanischen Philologien*, pages 265–294. Königshausen und Neumann.
- Edoardo D’Angelo and Paolo Monella. 2019. ALIM (Archivio della Latinità Medievale d’Italia). Storia, attualità, prospettive di una banca-dati di testi mediolatini. In Roberto Gamberini, Paolo Canettieri, Giovanna Santini, and Rosella Tinaburri, editors, *La Filologia Medievale. Comparatistica, critica del testo e attualità. Atti del Convegno (Viterbo, 26-28 settembre 2018)*, volume 3 of *Filologia Classica e Medievale*. L’Erma Di Bretschneider.
- Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica Digitale*, 1:7–17.
- Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.
- Lionel Nicolas, Alexander König, Monica Monachini, Riccardo Del Gratta, Silvia Calamai, Andrea Abel, Alessandro Enea, Francesca Biliotti, Valeria Quochi, and Francesco Vincenzo Stella. 2018. CLARIN-IT: State of Affairs, Challenges and Opportunities. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, Linköping electronic conference proceedings (Print), pages 1–14.
- Herman Stehouwer, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated search: Towards a common search infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Milan Straka and Jana Strakova. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. pages 88–99, 01.
- Milan Straka and Jana Straková. 2020. Udpipes at evalatin 2020: Contextualized embeddings and treebank embeddings. *arXiv preprint arXiv:2006.03687*.
- Claus Zinn. 2018. The language resource switchboard. *Comput. Linguist.*, 44(4):631–639, December.

Extending the CMDI Universe: Metadata for Bioinformatics Data

Olaf Brandt, Holger Gauza, Steve Kaminski,
Mario Trojan, Thorsten Trippel, Johannes Werner
Eberhard Karls Universität
Tübingen, Germany
firstname.lastname@uni-tuebingen.de

Abstract

The Component Metadata Infrastructure (CMDI) is a discipline independent metadata framework, though it is currently mainly used within CLARIN and by initiatives in the humanities and social sciences. CMDI allows flexible modelling of metadata schemas that are adjusted to the type of data. The model has built in functionality for semantic interoperability based on inventories providing persistent identifiers for definitions. In this paper we investigate, if and how CMDI can be used in bioinformatics for metadata modelling and describing the research data. For this purpose we embed CMDI based metadata in METS containers. Two sample schemas are developed the first based on a bottom up process and the second one takes the requirements of data publishing portals as the starting point of development.

1 Introduction

Data management in bioinformatics projects requires a very diverse and flexible set of metadata to accommodate for different scientific, organisational, and technical needs. Data categories must provide for the workflows of various types of experiments in the field of OMICS research (*genomics*, *proteomics* etc.), including workflows of researchers, laboratories archives, public repositories¹ and third party suppliers such as sequencing labs. Most laboratory working groups use individual, table based metadata for their projects, which are neither semantically described, nor interoperable with established workflows in data archival or data analysis. Within the project *BioDATEN*² funded by the state of Baden-Württemberg in Germany, subject matter experts meet to develop an environment that facilitates data storage and collaboration of different bioinformatics working groups and archives. BioDATEN combines expertise in data management, archiving, library science, bioinformatics and related scientific workflows.

Part of ensuring the interoperability and semantic interpretation of metadata is the discussion of a common description of metadata. Though there are specific metadata schemata in the bioinformatic community like the PRIDE schema for proteomics and approaches like qPortal³ there is no recognized gold standard for metadata handling in this subfield of OMICS research let alone the broader field of bioinformatics. On the other hand, there are well established standards outside bioinformatics that are used in the archiving and library community, such as METS/MODS⁴, PREMIS⁵, MARC 21⁶, etc. The variety of research data, research questions, methods and workflows require additional flexible and research specific schemata, that can be adjusted to the needs of the concrete projects' and working groups' context. Here, the ISO standardised ISO 24622-1 and -2 XML based CMDI framework is going to be explored as a candidate for representing the metadata in this project.

2 Motivation

Collaboration in bioinformatics is becoming increasingly important, by sharing information about genetic sequencing and data for reproducing results, applying different algorithms and workflows. Sharing primary data

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹A prominent example of a public repository is the *National Center for Biotechnology Information* (NCBI, USA, <https://www.ncbi.nlm.nih.gov/>).

²<http://www.biodaten.info>

³Mohr et al. 2018.

⁴for example <METS>Version 1.6.

⁵Caplan 2009 and <http://www.loc.gov/standards/premis/>.

⁶<https://www.loc.gov/marc/>.

becomes more and more widespread within the last 20+ years, but is still comparatively new to the field. This is contrasted by the fact that prices for genetic sequencing are constantly dropping, resulting in an exponential growth of data production and availability. Due to the increased amount of data, no single data centre is and will be able to store and provide access to all data, even if repositories specialize for species, etc. This results in the need for a distributed infrastructure in which research data is provided in a FAIR⁷ way.

Distributed environments providing data require a clear idea of the required levels of descriptions of research data. In the context of the European infrastructure initiative CLARIN, this has a long tradition with metadata being available and searchable with tools such as the Virtual Language Observatory⁸, though CLARIN addresses primarily research in the humanities and social sciences. In bioinformatics, similar methods and tools have been developed⁹, accompanied by strong market influences of large archives and publishers. In this paper, we try to elaborate on the technology used within CLARIN to see if the methods applied there are applicable for the BioDATEN project in bioinformatics.

2.1 Structured documentation of research data

The idea of sharing research data implies the distributed nature of research. Often more than one working group is interested in specific research questions to be addressed with the help of specific data sets. The diversity of research questions, size of groups, and distribution of interested parties results in the need for detailed descriptions that are necessary to understand the data. The internal documentation of each group such as code books, Read-Me files, laboratory books etc. are part of this documentation. This is not only true for labs working in natural language processing (NLP), but independent of the research discipline.

Within the BioDATEN project, various partners provided samples of metadata they use in their respective laboratories and for publication purposes. In bioinformatics, we noticed that there is metadata documentation available. However, to our knowledge there is no established and bioinformatics-wide schema. There are attempts to use representations based on `schema.org` within the `bioschemas.org` project, but these are not sufficient: First, there are no profiles provided that fit OMICS data. Second, those initiatives provide a number of data category definitions that could be utilized, but the categories extending `schema.org` do not have an identifier that can be used for references and it is unclear if they are stable enough for long time use. For example, the Gene Profile¹⁰ is targeted at life sciences, including diseases, and omits processing information, while the Biosample type¹¹ does not provide identifiers for some data categories that could be used. However, where possible, the concepts used by `bioschemas.org` will be reused here as well.

As a science data centre (SDC) of the state of Baden-Württemberg in Germany, the BioDATEN project also has ties to other initiatives in the field, which includes contacts to ELIXIR, the German Network for Bioinformatics infrastructure (De.NBI), etc. Consequently, in the development of metadata schemas we monitor the developments, planning for interfaces between these state infrastructures and larger initiatives. This also includes the use of existing ontologies where they fit to the needs of the OMICS community, avoiding duplication of work. However, to date ontologies of these communities are - as for the `bioschemas.org` representations - not sufficient for the OMICS community that uses the services of BioDATEN.

The metadata provided by project partners can be subdivided according to (1) their descriptive function, (2) specific information for the community, (3) process oriented information, and (4) technical metadata. Descriptive metadata is used to describe the data in an archive for citation purposes, such as the DataCite standard¹², but some data categories are not applicable in the context of bioinformatics, e.g. the concept of *Author* of raw DNA sequencing data. Process and workflow-oriented information¹³ provides the background and origin of data, as well as information about the tools and experimental techniques that have been used to generate the research data¹⁴. Technical information contains file information often provided in terms of PREMIS. Community specific information is often provided to allow specific keywords and structures in the search process.

When investigating the sample metadata provided by project partners, it became obvious that existing ontologies and taxonomies are often not applied in the concrete laboratory situation, where ad hoc or laboratory

⁷e.g. Wilkinson et al. 2016.

⁸van Uytvanck, Stehouwer and Lampen 2012.

⁹e.g. BLAST <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

¹⁰<https://bioschemas.org/profiles/Gene/0.7-RELEASE/>

¹¹https://bioschemas.org/types/BioSample/0.1-RELEASE-2019_06_19/

¹²DataCite 2019.

¹³for example De Nies 2013

¹⁴for an example of the workflow with its documentation, see Mohr et al. 2018.

specific spreadsheets are used to identify and document the research data. The lack of standard metadata formats thwarts the exchange and searchability of data with tools.

The interoperability of individual laboratory data with existing portals for sharing research data thus requires tailored solutions for each lab and portal despite the fact that labs often use similar testing machinery and testing procedures. Besides the rather informal definition of metadata, the public data repositories are often used for enhancing publications and sharing data. Data publication becomes even more imminent as publication of research data is mostly required for funding purposes and publication of scientific results. This is very similar to disciplines in the humanities such as (linguistic) annotation in fieldwork, corpus linguistics, etc.

2.2 The unit of description: the granularity of the research data

Infrastructures in the humanities and for OMICS research both show a great variety. The variety in the infrastructures for humanities results from different data types ranging from lexical resources, corpora, to data matrices. These data sets are serialized according to various conceptual models, such as graphs in RDF, XML annotations, table structures, etc.

In contrast to that, OMICS research operates on data sets that structurally have much more in common. However, the variety is still huge with for example different species, recurrence of analysis, size of cohorts, geographic variation, number of cells investigated, etc. Besides this variety, there are different workflows, also influenced by laboratory hardware. As the diverse data sets in the humanities require a flexible metadata model, the same applies to the area of bioinformatics. The archiving of research data, as well as searching and retrieving data units relies on descriptions by means of metadata and the assignment of a persistent reference to these units. For this reason, it is essential to have a solid understanding of the data unit to be described, usually termed the *granularity* of data. Granularity in this sense is the unit of data to be stored, archived and referenced in the research process.

ISO 24619 recommends on the granularity to use existing granularities, complete files, resource autonomy, and the requirement for a unit to be citable as criteria for selecting the underlying unit. This standard has been applied in CLARIN for assigning PIDs. In bioinformatics, there are some obvious candidates for archival objects. Inherent *atomic* units could be a base pair of nucleobases, a gene, a chromosome or an entire genome of an individual. From a computer science perspective, it could also be a single data file that is created in the process, such as FASTQ, FASTA, BAM, VCF, Excel or CSV files. Another natural unit would be a package of all files in an experiment, or all files that relate to a publication.

For bioinformatics applications it turned out that the granularity is implicitly given by the *sample*, i.e. the unit of a physically extracted sample of material, for example drawn with a needle. In bioinformatics workflows, these samples do only occur initially, afterwards other units will be referred to, such as sequencing information or experiments. For archiving, the sample often remains the common unit, but sometimes multiple samples are packaged into a study. It is noteworthy that raw data produced by a sequencing lab (DNA, RNA etc.) is nearly always transformed, trimmed, cleaned etc. This pre-processing is necessary to allow deeper analysis. The pre-processing is very similar to the processing and selection of corpus data in the humanities.

2.3 Automatic metadata extraction requirements during a data creation workflow

Metadata creation is often seen as a burden for researchers creating data. Due to the lack of standardised processes and project management software, archiving metadata is often created manually, based for example on the headers of TEI files¹⁵, or partly automatized by language processing applications and workflow engines such as WebLICHT¹⁶. The quality and completeness of the metadata in the archiving process is a major issue, for which automatic metadata enrichment processes are seen as a major step forward. This could mean to enrich metadata by authority file references, keyword extraction from textual resources, technical information extraction such as file size, checksums, dates, etc.

In bioinformatics processes, samples are analysed and processed in complex workflows. Many of these workflows are run on high performance clusters (HPC) or cloud infrastructure, are automatized and require only little intervention, hence the manual creation of metadata is even more problematic. The creation of metadata, especially of process and technical metadata, can partly be automatized, as the workflow engines on the infrastructure use, collect and provide process information during the process. Additionally, the technical metadata can be generated easily with appropriate software tools. Larger parts of the descriptive

¹⁵TEI P5 2020.

¹⁶M. Hinrichs, Zastrow and E. Hinrichs 2010.

and community specific information tend to be very similar in specialized labs, working with specific species, controlled conditions, health environments, etc. These can partly be defined in templates to be post-edited by the researcher. Again, this is similar to fieldwork situation or within large annotation projects in the humanities, though here this is often a manual process. For large NLP tasks, the metadata related processes are comparable to the bioinformatics workflows.

3 Specification and serialization options

In bioinformatics, researchers have to adhere to requirements by sequencing labs and scientific publishers. For sequencing labs, metadata descriptions contain details about the arrangement and preparation of samples to attribute the reads to the samples, treatments etc. The information may be lost in the resulting raw DNA sequences. Researchers have to define their lab processes to ensure that the DNA sequences are attributed to the sample, and in turn to the treatment or experimental condition. For publishing articles, the publication of data sets is often a requirement, the publication portals requiring various bits of information about the underlying sample. Hence a metadata schema needs to cater for the third party and laboratory internal requirements. To avoid redundancy in the metadata, the metadata categories need to be mapped onto each other, identifying common concepts and allowing transformation.

BioDATEN interdisciplinarily explores options for serializing the metadata with the full flexibility of metadata schemas required. Options using different data models such as RDF are left out. However, converting the metadata into RDF and offering it, possibly enriched by ontologies and authority data is seen as a valid option for the integration into the linked data cloud. In the following we discuss PREMIS, METS and CMDI serialization.

3.1 PREMIS

Implemented and used by archives and libraries, the PREMIS standard¹⁷ is meant to support the long-term preservation of digital objects via metadata. In the BioDATEN project, PREMIS will be primarily utilized for the storage of technical and rights metadata, as well as for the recording of events like data format conversion, checksum validation or changes in the related metadata records. The PREMIS data dictionary offers comprehensively controlled vocabularies allowing pointers with persistent identifiers. The description of scientific workflows denotes a clear limitation of the PREMIS standard. Hence PREMIS will be used for interoperability, but alone it is not sufficient for meeting all requirements.

3.2 METS

In order to manage the different metadata schemas used to describe research data, it is useful to collect them in a container format. Having multiple metadata records for one digital object should be avoided. One solution would be to use a container format such as the XML based METS format to combine different schemas. METS is described by an XML schema and is almost exclusively serialized as XML. As a container format it is able to integrate other XML schemas without loss of information via so called extension schemas. A decisive reason to choose METS is the integration with PREMIS, which is described in detail in the literature. The different building blocks offered by the METS standard can be used to store the variety of metadata schemas needed for research data. These schemas can be registered in METS profiles¹⁸, which also allow for a comprehensive documentation and therefore re-usability of metadata in the METS container format. However, as a container format, METS does not provide the required metadata schemas in itself.

3.3 CMDI

Another option for modelling the metadata is by using the Component Metadata Infrastructure (CMDI, ISO 24622-1 and ISO 24622-2), which is applied in the CLARIN community. As an XML based serialization, many tools for editing and maintaining exist, archival systems implement ways of storing the data. Transformation into other XML based formats is easily conducted using XSLT or similar technologies. CMDI offers flexible modelling options. For example, each lab can create their own metadata profile, assembling all necessary data categories required in their respective workflow. At the same time, they can reuse parts of the metadata profiles that match the requirements of portals and service providers, archives and other partners. Using these common components, the target data format can easily be generated by a simple transformation. In

¹⁷for example Caplan 2009

¹⁸METS Profiles 2018

fact, due to the definition of the CMDI components with concept links, for example, referring to definitions in the CLARIN Concept registry or in persistent ontologies, a high degree of semantic interoperability is achieved. For the serialization, the envisaged problems are similar to those already known in the CLARIN community: if labs define their own profiles and components, a certain degree of fragmentation is bound to be the result. Additionally, there is currently no fixed set of data categories, and the CLARIN Concept Registry does not contain the required definitions for bioinformatics data sets beyond interdisciplinary metadata categories. Another potential problem is that neither the bioinformatics archives and portals nor the external service providers natively support CMDI, hence a transformation is required at each step in the workflow, if CMDI were used. However, metadata generated in an automatic workflow can successively be added to the metadata file, which supports the required flexibility.

4 Implementation

Bearing in mind the established metadata workflows of the data centres and publishers based on METS, we envision the integration of CMDI in METS-containers, also integrating PREMIS metadata. Based on previous work within CLARIN we created two CMDI Profiles, the BioDATEN Profile ([clarin.eu:cr1:p_1588142628378](https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1588142628378))¹⁹ and the BioDATEN Minimal Profile ([clarin.eu:cr1:p_1610707853515](https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1610707853515))²⁰. Figure 1 shows the preliminary integration of the mentioned schemes in the form of a simplified section of a METS-XML file.²¹

Using this procedure, we can combine CMDI's flexibility for modelling while keeping the data interoperable with the archives and service providers.

4.1 The BioDATEN profile

The CMDI profile reuses components previously defined in their newest developmental version, especially

- the GeneralInfo component for general information, which is Dublin Core inspired
- the optional Project component for information on the project
- the optional Publications component to provide information on associated publications
- the Creation component with information on the creation of the resource. This component was enriched by a new ethics component providing information on obligations by ethics commissions, etc. As this also becomes more relevant for other disciplines, this should be a general recommendation for future releases of the creation component.
- the optional Documentations component for available documentation that is not part of the publications
- the Access component to provide information on accessing the resource
- the ResourceProxyListInfo component providing information on each data stream, including checksums, size, and original file name.

The tailored component SequencingInfo provides specific information on OMICS data beyond the creation process. For selecting data categories here, we were able to use Excel files used for managing metadata and provided by some partner laboratories. The schema is defined with its extension in mind, especially during a consolidation phase in which the community tests the schema. By planning for the extension, it is possible to add fields requested by researchers. The intention was to provide a bottom-up design of a metadata profile.

Currently, we evaluate the mapping of laboratory internal metadata storage to this profile and assess if the integration of this metadata framework, including CMDI, METS and PREMIS in the Invenio²² repository system, used within the BioDATEN project. However, to our knowledge even open repository systems such as Invenio or Fedora-Commons require additional work when used with tailored metadata schemas. Additionally, it is still essential to investigate, which transformations are required from a lab internal metadata set to interoperable metadata sets used by archives and portals.

The selection of data categories was bottom-up, starting with the partner laboratories and researchers of the project. The schema turned out to be very detailed with 99 fields, most of them optional. The interaction with public repository platforms proved to be problematic, as they required other metadata fields. Additionally

¹⁹https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1588142628378/xsd

²⁰https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1610707853515/xsd

²¹The complete sample metadata file and the schemas discussed in this paper have the DOI 10.5281/zenodo.4506354 and can be viewed and downloaded from the following url: <https://doi.org/10.5281/zenodo.4506354>.

²²<https://github.com/inveniosoftware/invenio>

```

<mets:mets>
  <mets:metsHdr ID="e6879deaa-2f64-48d7-bfc9-21cd77fb9571"
    CREATEDATE="2020-08-13T11:28:51"
    LASTMODDATE="2020-08-13T11:28:51" RECORDSTATUS="NEW">
    <mets:metsDocumentID TYPE="UUID">
      e6879deaa-2f64-48d7-bfc9-21cd77fb9571</mets:metsDocumentID>
    </mets:metsHdr>
  <mets:dmdSec
    ID="dmdSecGeneralDataCite_6879deaa-2f64-48d7-bfc9-21cd77fb9571">
    <mets:mdWrap MDTYPE="OTHER">
      <mets:xmlData>
        <cmd:CMD CMDVersion="1.2">
          ...
          <cmdp:BioDatenProfile>
            <cmdp:GeneralInfo>
              <cmdp:ResourceName xml:lang="en">Sample data set</cmdp:ResourceName>
              <cmdp:ResourceTitle xml:lang="en" />
              <cmdp:ResourceClass>OMICS data</cmdp:ResourceClass>
              <cmdp:Version xml:lang="en" />
              <cmdp:LifeCycleStatus>archived</cmdp:LifeCycleStatus>
              <cmdp:dateCreated>2020-08-08</cmdp:dateCreated>
              <cmdp:LegalOwner xml:lang="en" />
              <cmdp:FieldOfResearch>Bioinformatics</cmdp:FieldOfResearch>
              ...
            </cmdp:GeneralInfo>
            ...
          </cmdp:BioDatenProfile>
          ...
        </cmd:CMD>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  ...
</mets:mets>

```

Figure 1: Simplified extract of the preliminary integration of the schemes CMDI and PREMIS into a METS-XML container.

by reusing preexisting components such as the ones for general information, the resulting metadata showed redundancies between the CMDI section and other parts of the metadata in the METS container.

4.2 The BioDATEN Minimal Profile

Based on the BioDATEN-Profile in conjunction with a survey of various archives and portals targeting OMICS data we created a second profile. In order to cover the variety of the OMICS research on the one hand and the demand for a minimal set of metadata to be supplied by the scientists on the other hand we based our profile on the *Minimum Information about any (x) Sequence* (MIxS) schema.²³ At the same time we concluded that a perspective focused solely on samples (and environments) is too narrow and classified the subject-specific metadata in groups named study, experiment, sample, environment, run, and data. This process approach has been inspired mainly by the Extracellular RNA Atlas²⁴ and the MOD-CO schema²⁵. The new schema comprises the absolutely required fields for data portals plus additional fields suggested by expert users, resulting in 21 OMICS specific fields. We expect to see a demand for more metadata fields beyond our minimal set. In order to guide this in a manageable way we plan to offer many optional fields from the MIxS schema, currently the schema offers about 30 of these. Furthermore additional information can be added without structural restrictions. By this procedure we hope to collect a feedback about necessary metadata apart from the minimal set. The current implementation does not implement the full set of vocabularies that are intended to be used in the schema. The integration of the vocabularies based on various ontologies is still pending.

For review of the schema, an HTML based input form was generated, using the Comedi editor²⁶. Unfortunately, this editor still implements CMDI 1.1; integration into METS containers in our case requires CMDI 1.2. Different editor generation tools based on the XSchema are still being tested such as the open source tool XSD2HTML2XML²⁷, but this tool is of limited use for schemas embedded in containers and shows some usability issues when using controlled vocabularies in the schema.

5 Future Work

Automatic metadata retrieval from bioinformatics workflows still remain challenging. Several approaches have already been developed, however a standard independent of specific research disciplines is not yet existent. This requires APIs in the workflow engines to extract the appropriate metadata in the process where they are present. The adaptation of the enriched metadata to specific modelling environments such as CMDI would result from this.

Based on the automatic generation of an HTML form to edit the metadata instances, we plan to test the schema with researchers from OMICS research and test, if the information can be fed into data publication portals with standard APIs.

Within BioDATEN, the development points in a direction of not only supporting one metadata schema, but a variety of schemas that fit to the needs of the users in the specific domains. Initially, BioDATEN will support a limited number of schemas providing appropriate converters to DataCite, Dublin Core and other relevant formalisms. Users of the BioDATEN infrastructure will be required to restrict themselves to these schemas at first, but an extension to a general framework for registering metadata schemas might be required for scalability reasons. This is very similar to the development in CLARIN and the CMDI infrastructure with the component registry and a tool for mapping the data categories to a faceted search comparable to the VLO. However, as various schemas, also outside of the CMDI universe might be required, a generalized framework might be required that allows also the registration of other schemas, including schemas that are not defined for XML. Due to the support within the tools of BioDATEN, a proliferation can still be avoided, as no unsupervised schema development is part of the process. The modelling and implementation of this will be part of future developments.

Acknowledgements

The work reported here was funded by the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK). We would also like to thank the anonymous reviewers for helpful comments.

²³<https://gensc.org/mixs/>

²⁴<http://exrna-atlas.org>

²⁵<https://www.mod-co.net>

²⁶Lyse, Meurer and De Smedt 2015

²⁷<http://www.linguadata.nl/> and <https://github.com/MichielCM/xsd2html2xml>

References

- Priscilla Caplan. 2009. Understanding premiss.
- DataCite. 2019. DataCite Metadata Schema 4.3, 08.
- Tom De Nies. 2013. Constraints of the prov data model. W3C Recommendation.
- Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. Weblicht: Web-based lrt services in a distributed e-science infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 05. European Language Resources Association (ELRA).
- ISO 24619:2011(E). 2011. Language resource management — Persistent identification and sustainable access (PISA). Standard, International Organization for Standardization, Geneva, CH, 01.
- ISO 24622-1:2015(E). 2015. Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model. Standard, International Organization for Standardization, Geneva, CH.
- ISO 24622-2:2019(E). 2019. Language resource management — Component metadata infrastructure (CMDI) — Part 2: Component metadata specification language. Standard, International Organization for Standardization, Geneva, CH, 07.
- Gunn Inger Lyse, Paul Meurer, and Koenraad De Smedt. 2015. COMEDI: A component metadata editor. *Selected Papers from the CLARIN 2014 Conference*, 8(116):82–98, 08.
- METS 1.6. 2010. <METS> metadata encoding and transmission standard: Primer and reference manual. version 1.6 revised. Technical report.
- METS Profiles. 2018. METS profiles. Technical report.
- Christopher Mohr, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czernel, Oliver Kohlbacher, and Sven Nahnsen. 2018. qportal: A platform for data-driven biomedical research. *PLOS ONE*, 13(1):1–18, 01.
- PRIDE. n.d. Guide to generate PRIDE XML files.
- TEI P5. 2020. TEI P5: Guidelines for electronic text encoding and interchange. TEI Recommendation.
- Dieter van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034, Istanbul, Turkey, 05. European Language Resources Association (ELRA).
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018).

Community-Based Survey and Oral Archive Infrastructure in the *Archivio Vi.Vo.* Project

Silvia Calamai
Siena University
Arezzo, Italy
silvia.calamai
@unisi.it

Niccolò Pretto
University of Padova
Padova, Italy
niccolo.pretto
@dei.unipd.it

Maria Francesca Stamuli
Ministry of Culture and
Tourism, Florence, Italy
mariafrancesca.stamuli
@beniculturali.it

Duccio Piccardi
Siena University
Arezzo, Italy
duccio.piccardi
@unisi.it

Giovanni Candeo
ILC, CNR
Pisa, Italy
giovanni.candeo
@ilc.cnr.it

Silvia Bianchi
Siena University
Arezzo, Italy
bianchi.silvia
@unisi.it

Monica Monachini
ILC, CNR
Pisa, Italy
monica.monachini
@ilc.cnr.it

Abstract

Audio and audiovisual archives are at the crossroads of different fields of knowledge, yet they require common solutions for both their long-term preservation and their description, availability, use and reuse. *Archivio Vi.Vo.* is an Italian project financed by the Region of Tuscany, aiming to: (i) explore methods for long-term preservation and secure access to oral sources, and (ii) develop an infrastructure under the CLARIN-IT umbrella offering several services for scholars from different domains interested in oral sources. This paper describes the project's infrastructure and its methodology through a case study on Caterina Bueno's audio archive.

1 Introduction

Audio and audiovisual archives¹ are scattered all over the Italian peninsula, from researchers' private houses to universities and research centres, from cultural institutions (e.g., Istituti per la Resistenza) to State institutions, such as State Archives and Libraries. The pilot survey made by Galatà and Calamai in 2018 has emphasised the status of precariousness, instability, and insecurity that affects audio and audiovisual archives available at different communities of Italian researchers. Almost half of the resources listed in the survey (49.6%) were barely accessible. Only 9.2% of the resources was accessible and available, 4.6% was partially accessible, 35.1% was available upon request, 1.5% was available upon request but only for selected parts. As for the resources which were declared to be accessible, the access policies were as follows: only 9.2% of these resources was freely accessible online (with no authentication), 7.6% was accessible online via authentication, and 29% was accessible onsite (i.e. where the resources are physically stored). As for the long-term maintenance and preservation, the answer receiving the highest number of responses was no-one (43%), followed by reference institutes, such as associations, foundations, libraries and their archives (17%), reference universities (16%), and the owners/individuals themselves (15%)². Several research projects in recent years have aimed to disseminate audio and audiovisual archives, which have been collected over the years by both researchers and amateur fieldworkers. Some Italian³ examples are, among others, *Grammo-foni. Le soffitte della voce*, also referred to as Gra.fo (Calamai and Biliotti, 2017), *Voci, parole e testi della Campania*⁴, *I granai della*

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹Archives containing speech and sounds are differently labelled according to the fields of expertise of the researchers (audio archives, oral archives, speech archives). Among linguistics, the label *corpus/corpora* is also used. In the present paper, the label "oral archives" is considered a synonym of "audio archives".

²Further details available in (Galatà and Calamai, 2019)

³A detailed review of similar projects from other countries goes beyond the scope of this paper. We limit ourselves to indicating the Telemeta (Fillon et al., 2014) project and the Endangered Languages Archive (ELAR - www.elararchive.org), which share some similarities with the work presented here. At the following URL oralhistory.eu/collections/clarin-eric an extensive list of European oral archives and related research projects can be found.

⁴www.archivicampani.unina.it/archivi_campani_dev Last visited February 4th, 2021

*memoria*⁵, and *Circolo Gianni Bosio Audio Archives*⁶. Nevertheless, fragmentation and lack of common and shared standards are often the features of certain initiatives, whose duration over time crucially appears to be dependent on the duration of external funding, if any. Moreover, a researcher working with audio archives is not necessarily competent in long-term preservation of audio data and data management. Co-incidentally, not all the research projects dealing with audio archives receive financing for all the different professional profiles involved in their preservation, managing and valorisation.

Given this picture, it appears urgent and no longer postponable to provide an infrastructure offering: i) a long-term preservation service for audio archives, ii) a shared set of metadata compliant with the main international standards and FAIR principles, and iii) an access interface which takes into account the peculiarities of the audio modality and which is able to support researchers in different disciplines. This paper presents how the *Archivio Vi.Vo.* project tackles these problems, illustrating the overall methodology adopted and infrastructure that is being developed.

2 The *Archivio Vi.Vo.* Project

2.1 The Project

In 2019, the Region of Tuscany decided to support a project entitled *Archivio Vi.Vo.*, which aims to catalogue and disseminate oral archives. The partners involved are: Siena University (Silvia Calamai), CNR-ILC & CLARIN-IT (Monica Monachini), Soprintendenza Archivistica e Bibliografica della Toscana (Maria Francesca Stamuli) and Unione dei Comuni del Casentino (Pierangelo Bonazzoli). In order to reach the above-mentioned ambitious objectives, *Archivio Vi.Vo.* has concentrated most of its efforts on the design and development of an architecture, hosted by CLARIN-IT, the Italian consortium of the CLARIN research infrastructure, which could be used by several other projects concerning audio archives. A crucial step towards this aim concerns the definition of the metadata set(s) used to describe the data. This set has to be compliant with international archival standards, such as ISAD(G) and ISAAR, as well as several others chosen from different disciplines (cf. Section 4.1).

2.2 The Case Study

The architecture (cf. Section 3) is in the process of being validated on a specific audio archive, namely Caterina Bueno's audio archive, which appears to be rather challenging, for the following reasons: (i) it has a complex archival history, (ii) it is in a very poor conservation condition, and (iii) it contains highly heterogeneous audio material.

Caterina Bueno (San Domenico di Fiesole, IT, 2nd April 1943 – Florence, IT, 16th July 2007) was an Italian ethnomusicologist and singer (Giorgi et al., 2013). Her work as a researcher has been held in high regard for its cultural value, as it brought together many folk songs from Tuscan and central Italy that had been orally passed down from one generation to the next until the 20th century (when this centuries-old tradition started to vanish). Her work as a singer was always oriented towards research. At the age of twenty, she started travelling through the Tuscan countryside and villages recording Tuscan peasants, artisans, common men and women singing any kind of folk songs: lullabies, *ottave* (rhyming stanzas sung during improvised contrasts between poets), *stornelli* (monostrophic songs), narrative songs, social and political songs, and much more. These were the same songs that she sang in her performances, making them well-known and appreciated both in Italy and abroad in the second half of the 20th century, when she was at the pinnacle of her career. Caterina Bueno's sound archive is composed of about 476 analogue carriers (audio open-reel tapes and compact cassettes), corresponding to more than 700 hours of recordings, and it was digitised during the previous Gra.fo project. The analogue recordings were located with two different owners: part of them were stored at Caterina's heirs' house, while the rest was kept by the former culture counsellor of the Italian Municipality of San Marcello Pistoiese, in the *Montagna Pistoiese*, where a multimedia library was supposed to be set up. Unfortunately, disagreements and misunderstandings between the two parties have so far left the archive fragmented and inaccessible to the community. Both owners, independently, have turned to Silvia Calamai for the reassembly of the

⁵www.granaidellamemoria.it Last visited February 4th, 2021

⁶www.circologiannibosio.it/archivio.php Last visited February 4th, 2021

whole archive in the digital domain, so as to respect the artist's wishes. After being digitised, the carriers were returned to their owners.

In several cases, the original carriers were devoid of all the contextual information (place and date of recordings, speakers involved in the recordings). In other cases, the open-reel tapes were recorded at different speeds and using different track-head configurations, thus making the digitisation process and the creation of access copies rather complex. In this respect, Caterina Bueno's audio archive represents an extreme test case where different levels of complexity call into question different professional profiles and skills. In addition, such an archive may be of interest to different fields of research; from this point of view it would have to meet different needs according to different types of users.

3 User Needs in Oral Archives

In order to figure out the different needs and the different requests of those using audio archives in their research, an online survey was launched. The aim of the questionnaire was to gather together exploratory data on i) the profiles of Italian oral archive users; ii) their research routines (e.g., the type of document they are interested in in their searching strategies); iii) the features that they would like to add to an oral archive infrastructure. Our intent is to tailor the development process of the *Archivio Vi. Vo.* infrastructure (see below, Section 5) to the needs of the target population of users. The 56 responses will be thoroughly analysed here.

Materials. The questionnaire was anonymous and written in Italian. At the beginning, the participant was informed of the context and purposes of the survey; the expected completion time was explicitly stated and set at approximately 2/3 minutes. Then, the participant was given the questionnaire, which consisted of 11 items and could be divided into 3 thematic sections.

i) The first explored the respondents' profiles by asking them to state their main field of interest ("Linguistics", "Oral history", "Sociology", "Anthropology", "Ethnomusicology", "Sound and music computing", "Other"), number of years of experience ("0-5", "6-10", "10+"), age and gender.

ii) The second group of items pertained to their research routines, asking them to state their driving motivation for the use of oral archives ("Study and research", "Work - for Institutions or Archives", "Work - of different nature", "Leisure and Hobby", "Other"), their frequency of use of online oral archives (on a 1-4 Likert scale from "Never" to "Always"), the oral archives which were most familiar (open-ended), the frequency of their searches for specific type of documents ("Audio files", "Summaries of audio files", "Catalog cards", "Video files", "Other", on a 1-4 Likert scale from "Never" to "Always"), the frequency of their use of specific searching strategies ("By author", "By keywords", "By topic", "By genre", "By dialect/language", on a 1-4 Likert scale from "Never" to "Always") and their levels of perceived usefulness of specific searching strategies ("By genre", "By topic", "By keyword", "By language/dialect", "By abstract", "Other", on a 1-4 Likert scale from "Useless" to "Very useful").

iii) The last section consisted of a single open-ended question: the participants were directly asked to provide their suggestions for the development of a digital oral archive.

Procedure. The questionnaire was imported into Google Forms and distributed online via 8 mailing lists of the major Italian associations and their pertinent fields of research, such as Associazione Italiana Scienze della Voce⁷ (AISV), Associazione Italiana di Storia Orale⁸ (AISO), Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)⁹. The survey was available from June 18th, 2020 to August 27th, 2020.

Analysis. Descriptive statistics of the demographic features of our respondents will be provided. Then, bar plots will be generated in R (R Core Team, 2021; Lüdtke, 2021) in order to present a compact visualisation of the Likert responses of section ii. An exploratory correlation matrix will also be built (cor-

⁷www.aisv.it Last visited February 4th, 2021

⁸www.aisoitalia.org Last visited February 4th, 2021

⁹www.aiucd.it Last visited February 4th, 2021

relation package: (Makowski et al., 2020); method set to “auto”) with the aim of underlining significant patterns in the research routines of the respondents. Given the high number of tests (105), type I errors were controlled for with the Holm method; only the significant results will be reported and discussed here. Moreover, with the aim of proving that the involvement of researchers with diverse disciplinary backgrounds is not a mere exercise in academic demography, we run a second batch of point-biserial correlations between a dummy variable representing the expressed expertise in either oral history or linguistics (1 vs. 0, decided by coin flip) and each of the 15 series of numerical responses provided in the second thematic section of the questionnaire. This procedure will highlight the needs of the users with specific research interests as a reflection of their work routines. We chose to compare oral historians with linguists because of their representativeness in our pool of responses (see below); moreover, we may safely advance some preliminary hypothesis, such as the preference on the part of linguists for searching strategies by language or dialect. A Holm adjustment of p values was applied to this analytical batch as well. Lastly, a qualitative commentary on the answers to section iii will be discussed.

Results: Demographics (i). 32 female (57.1%) and 22 male (39.3%) respondents took part in our survey. Two participants did not specify their gender. Their age ranged from 27 to 74 (mean: 47.6, st. dev.: 13.7). Oral historians (22: 39.3%) and linguists (11: 19.6%) formed the most numerous groups of respondents. Sound and music engineers (6: 10.7%), ethnomusicologists (4: 7.1%) and anthropologists (3: 5.4%) followed with moderate numbers of participants. The remaining group of 10 interviewees was extremely fragmented, presenting one individual per category (sociologist, psychologist, archaeologist, mixed competences etc.), proving the importance of oral archives in several disciplines. Twenty-seven participants (48.2%) stated they had more than 10 years of experience in their respective field, followed by 19 (33.9%) with 0-5 years of experience and 10 (17.9%) with 6-10 years of experience.

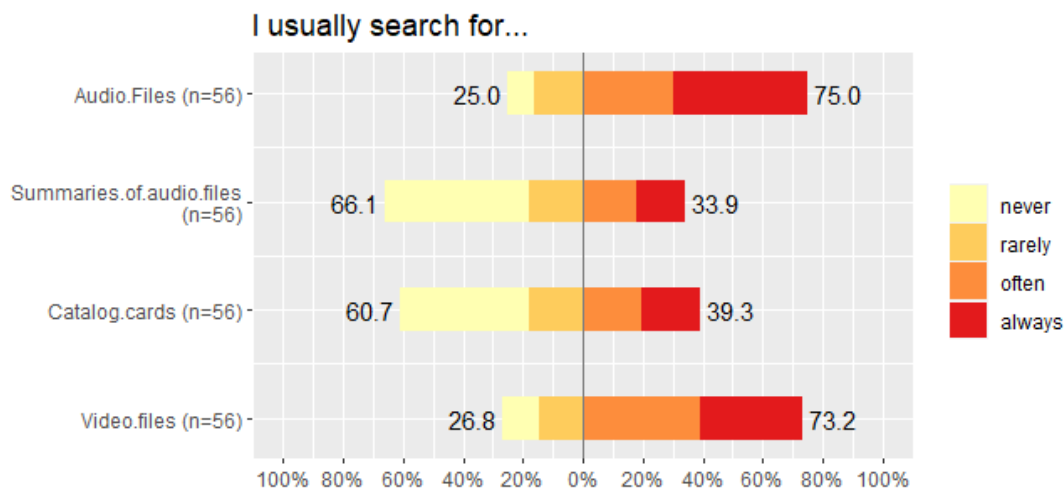


Figure 1: Bar plot of the perceived frequency of searching for specific document types

Results: Research routines (ii). Most of the respondents (42: 75%) refer to oral archives because of research and study activities. Eight individuals (14.3%) work for institutions pertaining to archives, while the remaining 6 made fragmentary references to the other categories outlined above. The distribution of the responses concerning the perceived frequency of use of online oral archives forms a left-skewed bell, with 13 (23.2%) “never”, 24 (42.9%) “rarely”, 13 (23.2%) “often” and 6 (10.7%) “always”. The majority of our interviewees (27: 48.2%) work (or worked) on a single online oral archive, while 10 of them (17.9%) mentioned more than one resource, and 19 (33.9%) failed to specify the name of specific archives. Curiously, Youtube was referred to as an oral archive repository in two responses. Figures 1 and 2 show the rate of responses to the questions about the perceived frequency of searching for specific document type and using specific search strategies. The plots clearly show that our participants search for

audio/video files more frequently than they do for summaries or catalog cards. The preferred searching strategies are by keywords and topic. Other materials they searched for included transcriptions (3 occurrences), complementary materials, metadata, time stamps, preprocessed data for linguistic analysis, etc. Figure 3 illustrates the visualisation of the perceived usefulness of specific searching strategies.

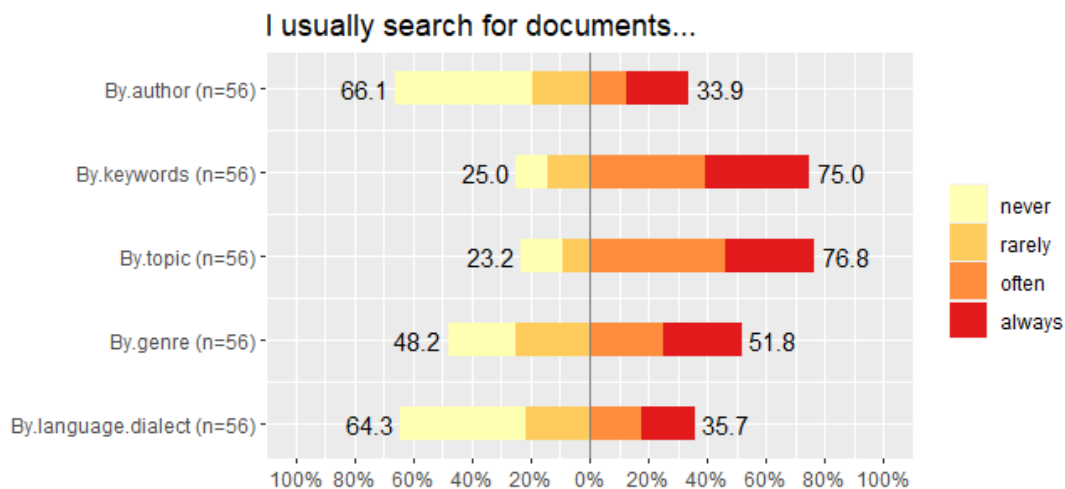


Figure 2: Bar plot of the perceived frequency of specific searching strategies

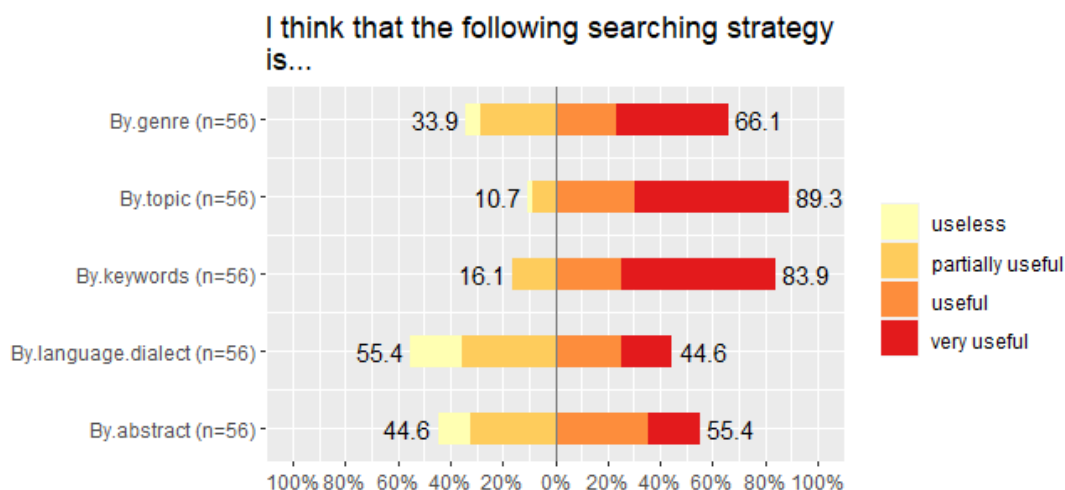


Figure 3: Bar plot of the perceived usefulness of a specific searching strategy

The four categories shared by the plots in Figures 2 and 3 follow very similar distributions. The searching strategies which are perceived as very useful are by topic and keywords. Indeed, all the correlations between these four pairs of items are positive and significant (by keywords: $r(54) = .47$, $p = .026$; by topic: $r(54) = .5$, $p = .01$; by genre: $r(54) = .5$, $p = .01$; by language/dialect: $r(54) = .53$, $p = .003$). While these correlations are conceptually self-evident, they signal a certain level of internal coherence and that, presumably, the participants maintained a sufficient level of attention to the questionnaire. A number of searching strategies were also suggested by our interviewees, which focused mainly on the time and the place in which the original recording was performed. Other significant correlations unveil subtler nuances of data distribution. The search for oral materials by keywords goes hand in hand with the strategy by topic ($r(54) = .62$, $p = .001$), which may imply some level of conceptual or functional overlap between the two categories. By the same token, there seems to be some sort of connection between the search by language/dialect and by genre. The usefulness ratings of the two strategies are positively correlated ($r(54) = .51$, $p = .005$), and those who search by language/dialect more often also give higher usefulness

ratings to the strategy by genre ($r(54) = .48, p = .016$). The search for summaries of audio documents correlates with the use of keywords ($r(54) = .5, p = .008$) and topic ($r(54) = .58, p < .001$), stressing the need for an effective reduction to only a few focal concepts. Indeed, keywords are also used more frequently by those who search for video files more often ($r(54) = .64, p < .001$). Lastly, the reliance on written summarisations can also be noted at other levels of analysis. Overall, those who refer to online oral archives more often search for written summaries ($r(54) = .49, p = .013$) and catalog cards ($r(54) = .46, p = .031$) more frequently.

In our pool of responses, the language/dialect searching criterion lacked behind the other options both in terms of frequency of use and perceived usefulness. However, this population pattern may hide the preferences of specific subsets of respondents. A series of point-biserial correlations between the expressed expertise in oral history (22 observations) or linguistics (11 observations) and the numerical answers to the 15 questions presented in this section suggests that this is indeed the case. As we anticipated, the only correlations which survived the Holm adjustment concern the language/dialect searching criterion (frequency of use: $r(31) = -.5, p = .045$; perceived usefulness: $r(31) = -.71, p < .001$). Contrary to the general trend, these coefficients confirm the paramount importance of this strategy for the research routines of linguists. Overall, the presence of strong disciplinary specificity in the expressed preferences may suggest the need for the development of additional access strategies tailored around the research background of the user.

Results: Suggestions (iii). Unfortunately, only less than half of the participants (26: 46.4%) submitted suggestions on the development of an online oral archive. In line with the results of the previous section, the most common recommendation (5 comments) concerned the development of a written counterpart to the audio documents. While a single participant asked for abstracts, four of them suggested the implementation of transcriptions, which should be aligned with the audio file and made searchable. Four participants were concerned with the quality of audio materials, which should be at least 24 bit and refreshable. A search engine by audio quality is also desirable. Others mentioned the need for supplementary materials, which could help in the search for relevant documents (e.g., by image). Lastly, other topics of discussion were granularity of metadata and terminology, cataloguing standards, accessibility, reusability and networking with other archives.

4 Building a Home for Audio Archives

4.1 Data and Metadata

In order to preserve and provide access to analogue audio recordings (e.g., the compact cassettes or the open-reel tapes), it is essential to digitise them. The result of the digitisation process is the digital *preservation copy*, which is composed of the audio content as well as other information about the carrier (such as the photo of the carrier itself, or its box)¹⁰. As the name suggests, the preservation copy is the “means” for safeguarding the content of the audio documents and it can be considered as the new digital master for long-term preservation. The preservation copy is only one element of the data workflow, since it is inextricably linked to the *archival unit*. In audio and audiovisual archives, it is defined as a set of data and documents pertaining to the very same communicative event, per unit of time and place. The archival unit is the outcome of a meticulous process involving listening, analysis and comparison. In the domain of oral archives, it is not infrequent that the content of a carrier needs to be re-organised. For example, an archival unit might be composed by content that is stored in several physical carriers (and, therefore, in several preservation copies), or vice versa, several archival units might be stored in the same physical carrier¹¹. Given the absence of a one-to-one relationship between the physical carrier (i.e., compact-cassettes, open-reels) and the archival unit, the preservation copies are kept separately from the archival units (Mulè, 2003; Calamai et al., 2014; Stamuli, 2020).

¹⁰An extended description of the preservation copy is available in (Bressan and Canazza, 2013)

¹¹Several other kinds of transformation could be performed, but their description goes beyond the scope of this paper.

This approach leads to a very complex set of metadata, organised along three different layers:

- (i) metadata for the description of the preservation copy,
- (ii) metadata for the description and managing of oral sources as items of an (audio) archive (archival unit),
- (iii) metadata expressing the relationship between the preservation copy metadata and the digital archive metadata.

In *Archivio Vi.Vo.*, a customised set of metadata has been defined for (i), inspired by other international standards for audio material description, in particular the one proposed by the Association of Sound and Audiovisual Archives (IASA Technical Committee, 2009). The project adopted ISAD(G) and ISAAR standards for the archival units (ii), encoding the information about archival material with Encoded Archival Description (EAD) and Encoded Archival Context (EAC) standard data models. One of the main challenges is to make these metadata structures interoperable with the CLARIN VLO infrastructure component which is part of CLARIN's Component Metadata Infrastructure and can cope with many different metadata descriptions, as long as they are implemented through (or converted to) the Component Metadata framework. The metadata structure for expressing the relationship between the preservation copy and the archival unit (iii) is based on the methodology described below.

4.2 From Preservation Copies to Archival Units

The methodology formalised and adopted in *Archivio Vi.Vo.* is composed of several steps. All the operations performed during these steps and the information inserted by audio technicians, researchers and/or cataloguers are stored and duly described through a set of appropriate metadata, thus maintaining the relation between preservation copies and archival units. The methodology starts with the creation of the preservation copy of a carrier. This phase has proved to be very delicate and time-consuming.

Digitised audio recording is often barely accessible, due to, e.g., different speeds, configurations or digitisation errors (Pretto et al., 2020). In these cases, if necessary, researchers or audio technicians recur to *clip*¹² in order to separate parts with different speeds, channels with different recordings or recordings in different directions. In *Archivio Vi.Vo.* a *clip* is defined as a duplicate of an audio segment extracted from a preservation copy. One or more clips can be extracted from a preservation copy. In some cases, the clips are the result of a restoration operation, which is necessary for accessing the sound content. The process of creating (and restoring) the clips cannot modify the preservation copy. The resulting clips will be accessible and allow the single researcher/cataloguer to listen, analyse and describe their contents. If some parts of the very same clip belong to different events (and, therefore, to different archival units, see Section 4.1), they will be segmented accordingly and new sub-clips will be created (archival unit clips). In some cases, some archival unit clips, derived from different preservation copies, would be part of the same event, and therefore they will also be part of the same archival unit. During the analysis, the researcher/cataloguer may decide to discard some clips in case she/he believes the content is not related to the archive or not relevant for the users (e.g., several minutes of silence without recording due to empty tape at the end of the digitised recording). It is essential that these choices be formalised and preserved in the metadata maintaining the history of the documents. Since the questionnaire (see Section 3) outlines the need for written summarisation for some of the participants, the *Archivio Vi.Vo.* platform includes a complex abstract (It. *regesto*) field. In the project, the abstract of an archival unit is divided into several segments related to the different parts that compose a single event. These segments are recognised and described by researchers and/or cataloguers during their analysis. Each segment is characterised by two temporal instants (the beginning and the end of the segment, respectively) and the description. The segments' length could be equal to or smaller than an archival unit clip (in the second case, the audio file will not be trimmed). As soon as all the archival unit clips are put into order, and all the missing metadata required by ISAD(G) is added, the archival unit will be created and available through the access interface.

¹²The concept of *clip* is commonly used to indicate data of either video or audio that has been clipped out (copied) from a larger carrier such as a reel or a video tape.

5 The Infrastructure

As for the infrastructure, the CLARIN-IT national data centre hosted in Pisa, ILC4CLARIN (Monacchini and Frontini, 2016), will implement new experimental approaches to preservation, management and access to audio data and metadata. The experimental activity aims to adopt the model and the high-performance computing and archiving services of the new GARR network infrastructure, built along the Cloud paradigm¹³. The project will also exploit the federated identity service of the CLARIN infrastructure, in order to manage users' access. A robust system for managing authentication is essential for audio and audiovisual archives because of the frequent privacy, ownership, and copyright issues concerning their content (Kelli et al., 2019; Kelli et al., 2020). Several classes of users are taken into consideration, each of them with different access grants.

The infrastructure consists of two different parts. The first one provides a data and metadata entry interface for archivists, archive owners or, in general, researchers. The system is highly complex; it must be able to manage various international standards and several kinds of specific functionalities. Considering the complexity of the project, the infrastructure could hardly be developed from scratch. Therefore, as a first step, ten types of archival software were evaluated on the basis of several features and technologies (standards, programming languages, frameworks, DBMS, license, etc.). The software selected was the open-source platform xDams¹⁴. Three main characteristics influenced the adoption of the software: (i) the completeness of its coverage of standards, (ii) its extensible no-sql database as well as (iii) the open-source license. The second part of the infrastructure consists of an access interface that can support researchers of different disciplines in discovering and studying audio or audiovisual documents.

In order to study the interaction with the software, two mock-ups were developed for studying and testing the interfaces for inserting and cataloguing the digitised documents (Figure 4a) and for accessing their content (Figure 4b), respectively. The two mockups have been developed with the Vue.js and Bootstrap frameworks, respectively, Web Audio API, as well as Peak.js and Audiowaveform, two libraries developed by the BBC¹⁵.

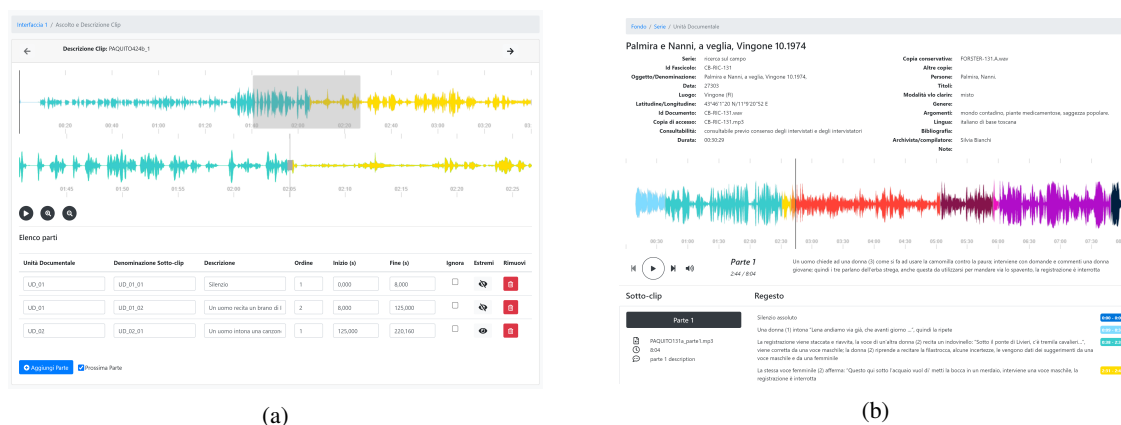


Figure 4: (a) section of the interface for creating archival units from preservation copies: the clips derived from the duplicate of the preservation copy are segmented, described, ordered and assigned to an archival unit; (b) the interface for accessing the archival units' contents.

The first mock-up interface is intended to provide the researchers and/or cataloguers with a web tool for listening to and describing the audio content of the clips extracted from the preservation copy. In addition, the user should be able to assign the described audio segments to different archival units, if necessary. The interface is composed of two main elements: the audio player and the section for the segmentation of the audio content. The audio player is complemented by the visualisation of the generated waveform using Peak.js and Web Audio API. The waveform is displayed both in its entirety and in a zoomed view,

¹³cloud.garr.it Last visited February 4th, 2021

¹⁴www.xdams.org Last visited February 4th, 2021

¹⁵github.com/bbc/peaks.js and github.com/bbc/audiowaveform Last visited February 4th, 2021

giving the user a very accurate time interval selection. In the section below the player, the data binding of Vue.js combined with the Segment API of Peak.js is exploited to allow the user to create, describe and sort the segments and assign them to the correct archival units.

The second mock-up interface aims to provide the user with an access tool that takes into the account the available contextual information in order to facilitate an adequate analysis of the archival unit. This interface can be divided into different areas: an upper area that displays the metadata of the archival unit, a middle area that contains the audio player with the waveform of the audio clip that is playing, and a bottom area containing the clips with their metadata, along with the abstract. In this latter area, the clips are grouped into Bootstrap card components. On the left side of the card, one can view the metadata and play/pause the clip. On the right side, there are the segments that compose each archival unit clip. The user can interact with this list by clicking on the different segments which are bound to the temporal instant of the audio clip. In order to facilitate user interaction with the interface, the segments are colour-coded with the time intervals to which they correspond in the waveform view. In this mockup interface, unlike the previous one, the waveform visualisation is created using a previously generated DAT file of the audio file in question (using Audiowaveform). This design choice was made because using Peak.js with Web Audio API requires downloading the entire audio file for waveform generation and this would result in a major slowdown in the web application. The use of pre-generated waveform files is very beneficial in our case since it is not necessary to load the complete audio files in the initial page loading. This allows the user to interact and view the waveform almost instantly (when the page is loaded, only the metadata of the audio files are loaded), taking advantage of audio streaming and thus avoiding long wait times due to the download of large audio files.

6 Final Remarks

Archivio Vi.Vo. constitutes a pilot case study within CLARIN-IT to experiment with methods for long-term preservation and secure access to oral archives and offer targeted services for both specialists and the general public interested in these data. *Archivio Vi.Vo.* aims not only to develop an infrastructure for preservation, description and use of audio archives, but, more ambitiously, to define and develop a model for the management, protection and enhancement of intangible archival heritage which can be replicated, even outside the context of Tuscany.

References

- Federica Bressan and Sergio Canazza. 2013. A systemic approach to the preservation of audio documents: Methodology and software tools. *Journal of Electrical and Computer Engineering*, pages 1–21.
- Silvia Calamai and Francesca Biliotti. 2017. The Gra.fo project: from collection to dissemination. *Umanistica Digitale*, 1:85–103.
- Silvia Calamai, Francesca Biliotti, and Pier Marco Bertinetto. 2014. Fuzzy archives. what kind of an object is the documental unit of oral archives? In Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 777–785, Cham. Springer International Publishing.
- Thomas Fillon, Joséphine Simonnot, Marie-France Mifune, Stéphanie Khoury, Guillaume Pellerin, Maxime Le Coz, Estelle Amy de la Bretèque, David Doukhan, Dominique Fourer, Jean-Luc Rouas, Julien Piquier, Julie Mauclair, and Claude Barras. 2014. Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–8.
- Vincenzo Galatà and Silvia Calamai. 2019. Looking for hidden speech archives in Italian institutions. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, number 159, pages 46–55. Linköping University Electronic Press.
- Pamela Giorgi, Fabiana Spinelli, and Serena Masolini. 2013. *Caterina Bueno: inventario del fondo documentario*. Firenze: Consiglio Regionale della Toscana.

- IASA Technical Committee. 2009. *Guidelines in the Production and Preservation of Digital Audio Objects: standards, recommended practices, and strategies*. IASA TC-04. International Association of Sound and Audio-visual Archives, 2nd edition.
- Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavriilidou, Pavel Stranák, et al. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, Italy, 8-10 October 2018*, pages 72–82. Linköping University Electronic Press.
- Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värvi, Pavel Straňák, et al. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. In *Selected Papers from the CLARIN Annual Conference 2019, Leipzig, Germany, 30 September, 2019*, pages 53–65. Linköping University Electronic Press.
- Daniel Lüdecke, 2021. *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.7.
- Dominique Makowski, Mattan S. Ben-Shachar, Indrajeet Patil, and Daniel Lüdecke. 2020. Methods and algorithms for correlation analysis in r. *Journal of Open Source Software*, 5(51):2306.
- Monica Monachini and Francesca Frontini. 2016. CLARIN, l’infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2):11–30.
- Antonella Mulè. 2003. Le fonti orali in archivio. un approccio archivistico alle fonti orali. *Archivi per la storia*, 16(1):111–129.
- Niccolò Pretto, Alessandro Russo, Federica Bressan, Valentina Burini, Antonio Rodà, and Sergio Canazza. 2020. Active preservation of analogue audio documents: A summary of the last seven years of digitization at csc. In *Proceedings of the 17th Sound and Music Computing Conference, SMC20, Torino*, pages 394–398, Torino.
- R Core Team, 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Maria Francesca Stamuli. 2020. Fonti orali, documenti e archivi: riflessioni e proposte per la nascita di un ‘archivio vivo’. In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Studi AISV 6*, pages 95–109, Milano. Officinaventuno.

A Two-OCR Engine Method for Digitized Swedish Newspapers

Dana Dannélls

Språkbanken Text, Dept. of Swedish
University of Gothenburg, Sweden
dana.dannells@gu.se

Lars Björk

Kungliga biblioteket
Stockholm, Sweden
lars.bjork@kb.se

Ove Dirdal

Zissor
Oslo, Norge
ove@zissor.com

Torsten Johansson

Kungliga biblioteket
Stockholm, Sweden
torsten.johansson@kb.se

Abstract

In this paper we present a two-OCR engine method that was developed at Kungliga biblioteket (KB), the National Library of Sweden, for improving the correctness of the OCR for mass digitization of Swedish newspapers. To evaluate the method a reference material spanning the years 1818–2018 was prepared and manually transcribed. A quantitative evaluation was then performed against the material. In this first evaluation we experimented with word lists for different time periods. The results show that even though there was no significant overall improvement of the OCR results, some combinations of word lists are successful for certain periods and should therefore be explored further.

1 Introduction

The process of converting images into digitized editable text is called Optical Character Recognition (OCR). OCR techniques have been applied since the late 90s and their performances have been improved significantly during the last decade with the advances of neural networks (Amrhein and Clematide, 2018; Nguyen et al., 2020). However, OCR processing of historical material, especially newspapers, remains a challenge because of low paper and print quality, variation in typography and orthography, mixture of languages and language conventions (Gregory et al., 2016; Chiron et al., 2017).

Kungliga biblioteket (KB), the National Library of Sweden, is the central source for digitized Swedish newspapers, offering access to more than 25 million pages via the web service “Svenska dagstidningar”.¹ The accuracy of the OCR system is therefore an important factor in order to maximize the access and usability of the digitized collections. To address this, KB, in collaboration with the Norwegian software company Zissor,² has implemented a novel OCR technique for combining two OCR engines: Abbyy and Tesseract. The two-OCR engine method has so far only been used as an internal testbed, and was only evaluated manually, awaiting automatic evaluation and suggestions of possible improvements.

In 2019, KB embarked on an infrastructure project together with Språkbanken Text,³ the Swedish Language bank at the University of Gothenburg, which is the coordinating Swedish CLARIN (Swe-CLARIN) node and CLARIN B Center, with the aim of evaluating and improving the results of the method (Dannélls et al., 2019).

In this paper we describe the two-OCR engine method, how we selected and prepared the ground truth material, spanning the years 1818–2018, and report the first quantitative evaluation. We further describe

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://tidningar.kb.se/>

²<https://zissor.com/>

³<https://spraakbanken.gu.se/en>

our attempt to improve the OCR performance for this period by increasing the OCR engines build-in vocabulary. By experimenting with different word lists for different time periods we take diachronic aspects into consideration and thereby, hope to adhere to linguistic change in the course of time (Springmann and Lüdeling, 2017). To our knowledge, this is the first use case study of evaluating and improving the OCR accuracy of Swedish newspaper texts over a period of 200 years.

2 Related Work

Newspapers are challenging material because of their mixture of typeface, size and complex layout. Gregory et al. (2016) present some of the biggest challenges in working with digitization of the British Library's nineteenth century newspaper collection. They emphasise the importance of knowing the source material and looking into the original data. Their work provides indication of developing methods and solutions that are tailored to the original text segments. In this project we follow their recommendations, but instead of text segments we are focusing on smaller units, namely on paragraph levels.

Earlier work on historical English demonstrated the challenges with combining multiple OCR engines (Lund et al., 2011). They reported an improvement over the word accuracy rate using voting and dictionary features. Reul et al. (2018) applied a confidence voting scheme between OCR models that were trained on a single engine. Their evaluation on Latin books showed a relative improvement over the character accuracy rate.

OCR errors are often classified into two groups: non-word errors and real-word errors (Mei et al., 2016; Nguyen et al., 2019). Real-word errors are more challenging because they require human inspections. Correcting real-word errors by increasing the engine's vocabularies has been studied by several authors who have proven the usefulness of lexicons, among other successful strategies for improving the OCR accuracy (Kissos and Dershowitz, 2016; Schulz and Kuhn, 2017; Nguyen et al., 2018). Authors have shown that a lexicon-based approach is competitive if it is adapted to certain domains or time periods. Our assumption here is that curated word lists are suitable to experiment with for Swedish material that spans over hundred years since a great number of changes in orthography and morphology occurred during this time, in particular around 1906. Real-word errors can also be caused by insertion, replacement or deletion of characters, for example *föreda* 'provide' became *breda* 'sprea' because "fö" was merged to "b". Several successful approaches have been proposed to detect these types of errors, ranging from Levenshtein distance (Samanta and Chaudhuri, 2013) to Finite-State (Silfverberg et al., 2016) and n-gram (Eger et al., 2016) methods. We have so far not explored any of these, but have plans to incorporate Levenshtein distance in the next phase of our work.

The work presented by Koistinen et al. (2017) aims to calculate the accuracy of the OCR system by comparing the Abbyy average confidence score that is assigned to each processed documents automatically. In this work we do not take this score into consideration because it was proven unreliable in an internal study which we conducted after the method was implemented.

Clematide and Ströbel (2018) discuss how to improve the quality of newspapers texts. An important outlook from their work is to understand how the performance of the OCR system varies in relation to studying how often mixture between Antiqua and Blackletter occurs. One conclusion was that high number of Blackletter articles often results in low OCR accuracy, therefore it is important to check the distribution of Antiqua and Blackletter. Something will address when we evaluate the complete material.

The results we report here are almost as accurate as the results reported for Finnish and Swedish newspaper texts (Drobac et al., 2019). However, they are not directly comparable because Drobac et al. experiments were done on a smaller selected set from 1771 until 1874, which we at the time of writing do not have any access to.

3 Two-OCR Engine Method

The two-OCR engine method was developed in 2017 in cooperation between KB and Zissor. The method was designed to enable adjustment and control of some key parameters of the post-capture stage of the OCR process, including dictionaries and linguistic processing, to match typical features of the newspaper as a printed product, characteristics that in a historic perspective change over time, such as layout,

typography, and language conventions. The working principle of the method is based on the evaluation and comparison between the results from two separate OCR engines: the proprietary system Abbyy FineReader, version 11.1.16,⁴ and the open source system Tesseract, version 4,⁵ developed by Google and is based on Long Short Term Memory (LSTM) Recurrent Neural Networks (Smith, 2007). The pre-trained models provided by each system were used to OCR the Swedish newspaper material. Abbyy and Tesseract have been integrated in such a way that allows to choose between one of them or run both in parallel. Figure 1 provides an overview of the two-OCR engine architecture.

When Abbyy and Tesseract run in parallel the results from the two engines are analysed and the best results on the word level are prioritised to create a new combined ALTO XML file. This process is accomplished by comparing the two ALTO XML files, word by word. The comparison and process of selecting between the results of the OCR systems relies on a scoring model that is based on the internal dictionaries of each system (see Section 3.1).

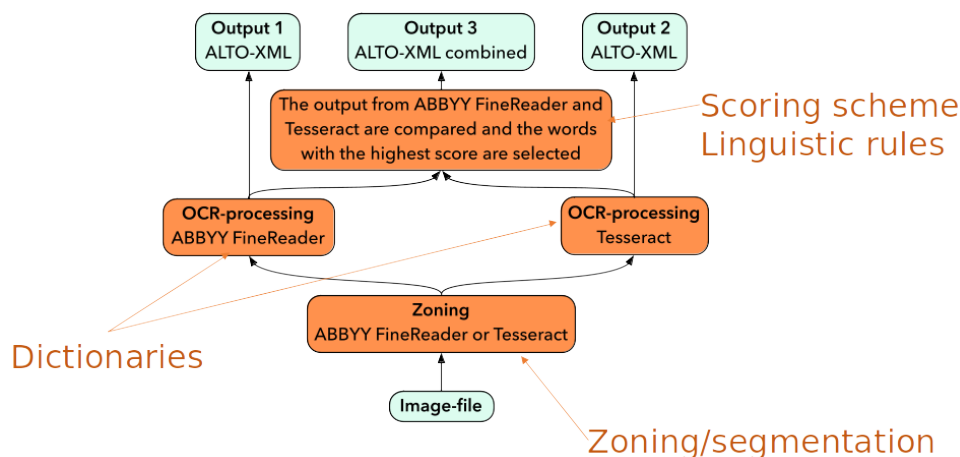


Figure 1: The design of the two-OCR engine method

3.1 Scoring Scheme

Abbyy and Tesseract apply different OCR techniques for analysis, recognition and segmentation of the scanned image. Their internal dictionaries also vary with respect to size and how confidence scores are calculated for each OCRed word. In a series of experiments Zissor analysed the OCR results generated by Abbyy and Tesseract and examined the differences between them in more details (Zissor, 2017). The aim of the experiments was to determine how to combine the OCR results from both systems to yield the most accurate results. The results of the experiments showed Tesseract has more overlapping zones compared to Abbyy – a feature which makes article segmentation difficult. Tesseract is also more sensitive to image noise, something that affects both zoning and the OCR quality. On the other hand, Tesseract seems to handle larger text fonts better, and in general greater variance in font compared to Abbyy. Each of the systems generates a confidence score for the OCRed words, but because this score is calculated differently for Abbyy and Tesseract, a comparison based on the OCR confidence scores is not straightforward. Character and word coordinates are the same for Abbyy and Tesseract which implies the same word can be taken from both systems for in-depth character and word comparisons.

Following the results of the experiments a rule-driven scoring model was implemented to determine automatically how to prioritise between the systems when disagreement on word level occurs. Each individual word is either verified (if confirmed by both dictionaries) or falsified (if rejected by one or both dictionaries) and subjected to further comparison, according to the rule set in the scoring model. Figure 2 demonstrates the scoring scheme in its present version. The results from the automatic comparison is a combined ALTO XML file containing words from each system.

⁴<http://finereader.abbyy.com/>, accessed via Server Web Services API

⁵<https://github.com/tesseract-ocr/>

Step	Rule	Sequence/consequence and choice of word
1.	Both words are equal = ABBYY	If ABBYY's suggestion and Tesseract's are equal: ABBYY's word is selected. [IF NOT: STEP 2]
2.	ABBYY's suggestion is blank = no word is given	If ABBYY's suggestion is blank no word is given. [IF NOT: STEP 3]
3.	Both words are equal after the removal of "noise characters" = ABBYY	If ABBYY's suggestion and Tesseract's suggestion are equal after the removal of one "noise character" in the beginning and end of the given word: ABBYY's word is selected. [IF NOT: STEP 4]
4.	The word is found in ABBYY's dictionary = ABBYY	If ABBYY's suggestion is found in ABBYY's dictionary or in a customised dictionary: ABBYY's word is selected. [IF NOT: STEP 5]
5.	ABBYY's suggestion is a numeral = ABBYY	If ABBYY's suggestion is a numeral that numeral is selected. [IF NOT: STEP 6]
6.	Tesseract's suggestion is blank = ABBYY	If Tesseract's suggestion is blank: ABBYY's word is selected. [IF NOT: STEP 7]
7.	If Tesseract's suggestion is a single character = ABBYY	If Tesseract's suggestion is a single character: ABBYY's word is selected. [IF NOT: STEP 8]
8.	If Tesseract's suggestion is found in Tesseract's dictionary or in a customised dictionary = Tesseract	If Tesseract's suggestion is found in Tesseract's dictionary or in a customised dictionary: Tesseract's word is selected. [IF NOT: STEP 9]
9.	If Tesseract's suggestion is a numeral = Tesseract	If Tesseract's suggestion is a numeral: Tesseract's numeral is selected. [IF NOT: STEP 10]
10.	ABBYY word = ABBYY	If none of the previous steps has rendered a word: ABBYY's word is selected.

Figure 2: The scoring schema underlying the voting principles between Abbyy and Tesseract

3.2 Segmentation Tool

In addition to the ALTO XML files produced by the OCR process, statistics about the word errors and applied corrections are generated as Excel-files for each system. In this way, the process can be manually monitored and adjusted using Zissor's article segmentation tool as illustrated in Figure 3.

3.3 External Swedish Word Lists

As mentioned above, Abbyy and Tesseract have their own internal dictionaries incorporated in the systems for improving the OCR word accuracy. The size of Abbyy's internal dictionary is unknown and the dictionary in Tesseract is rather small, containing around one million entries that have been compiled from unknown time periods. One possible way to improve the accuracy of the systems is therefore by increasing the internal vocabulary with external word lists for specific periods in the post-processing step. For this purpose we compiled four word lists from dictionaries and lexical resources from different time periods:

- Dalin, a full form lexicon for the 19th century, covering the morphology of late modern Swedish (Borin and Forsberg, 2011), containing 509,924 entries;
- Saldo, a full form modern lexicon (Borin et al., 2013), containing 1,704,718 entries;⁶
- Saol-hist,⁷ a subset of the Swedish Academy historical lexicon, containing only base forms, amounting to 128,720 entries;
- Fem, a word list over name entities that was compiled from five lexical sources at KB, containing in total 311,481 entries.

⁶Both Dalin and Saldo are part of CLARIN lexical resources.

⁷<http://spraakdata.gu.se/saolhist/>

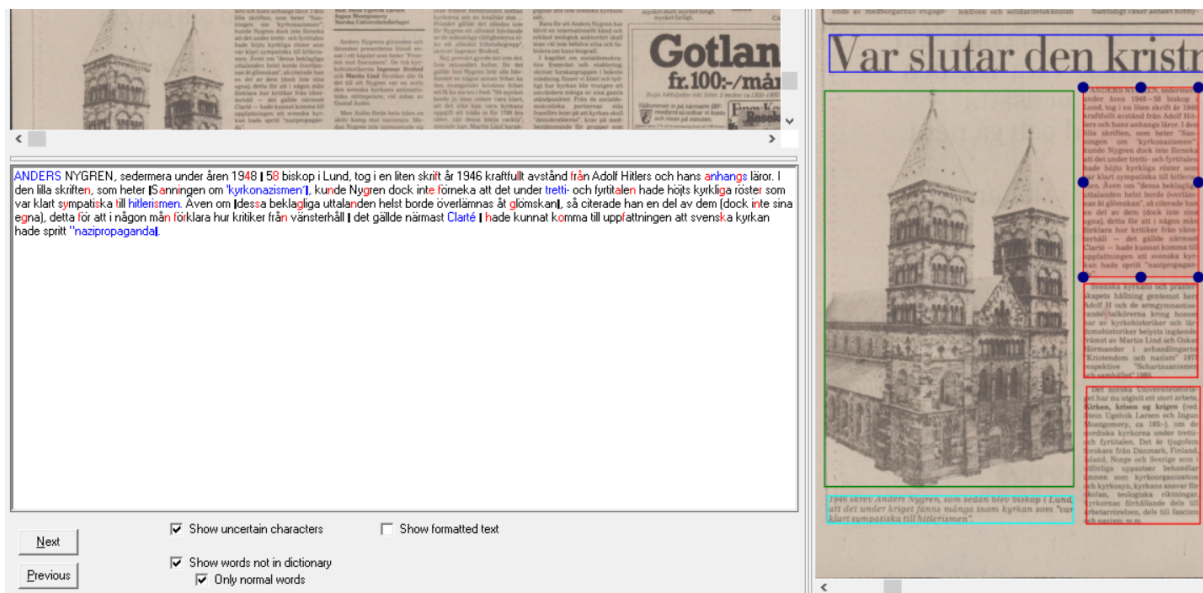


Figure 3: Zissor’s article segmentation and OCR tool with Abbyy. Selected Zone; Abbyy OCR text. Blue words are words that are not included in the Abbyy standard dictionary or in additional imported Swedish dictionaries used in the OCR process. Red letters are letters that Abbyy is unsure of during the OCR process.

The first three word lists (Dalín, Saldo and Saol-hist) were extracted from the original sources. Fem was extracted from a selected set of corpora. While Saol-hist and Fem only contain plain vocabulary lists, Dalín and Saldo contain morphological gazetteers.

To get an estimate of the coverage of each word list, we counted the number of tokens in our ground truth material (see Section 4) and compared each token with the words in each word list. For each word list we also kept a count of the tokens without any match. We summarise the results periodically, grouped by frequency in Figure 4. As Figure 4 shows the total amount of tokens in the whole material is 1,112,996 of which 521,816 tokens appear in newspapers between 1818-1906, 512,182 appear in newspapers between 1907-1996 and 78,998 tokens in 1997-2018. We can observe that for each period about 50% of the words are covered in Saldo and Dalín, and about 20% are covered in Saol-hist and Fem. When we inspected the list of words that were not found in any of the word lists (category “None” in Figure 4) we found that the majority of these words are numbers, place names, and proper names.

4 Reference Material and Processing

Our reference material consists of 400 pages, selected from 200 newspapers spanning the years from 1818 until 2018. Pages were carefully chosen to reflect typical variations in layout and typography. Two pages were selected from each newspaper; the second and the fourth. The underlying assumption for this decision was that there are generally less advertisements and pictures on the second and fourth pages.⁸

Each page in the reference material was segmented down to paragraph level, and each paragraph was marked with an ID number. This segmentation scheme is kept as a matrix that can be reused for the comparison between the reference material and the corresponding section in the OCR processed material. The resulting ground truth material amounts to 43823 IDs.⁹

⁸We acknowledge that advertisements and pictures are important mediums to study our cultural heritage, but for the purpose of this study where text is the primary focus, we chose to minimize these.

⁹A selection of the material is freely available under open source license from Språkbanken Text <https://spraakbanken.gu.se/en/resources#refdata>.

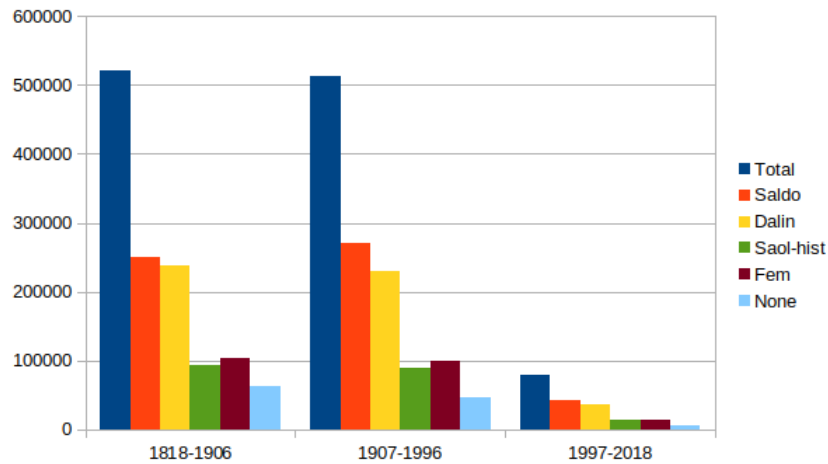


Figure 4: The frequency of tokens for three time periods, their total amount in the ground truth material, their coverage across four word lists (Saldo, Dalin, Saol-hist and Fem) and the number of tokens that does not appear in any of the word lists (marked with the “None” category).

The selected reference material was sent to Grepect, a transcription company who specialises in manual transcription of older material.¹⁰ Based on our inspections of the material we defined the guidelines for the transcription which contained instructions for typeface, size and location changes.¹¹

To produce our baseline we run the material through the two-OCR engine system using Abbyy’s and Tesseract’s internal dictionaries exclusively. Next, we run the material in the two-OCR engine system four times, each run with a different word list (see Section 3.3). For each run the system delivers three results: one result for Abbyy, one for Tesseract and one calculated/verified Abbyy-Tesseract.

5 Experiments and Evaluation

For evaluation we used the OCR frontier toolkit (Carrasco, 2014). The method calculates the results of the OCR errors by measuring character accuracy rate (CAR) and word accuracy rate (WAR).

First we run the evaluation against the prepared ground truth material of the 400 newspaper pages, once without external word lists (our baseline) and once with external word lists. Table 1 shows the evaluation results of these runs. At first glance we note that the best performing system on the character level is achieved with the combined method (91.64%), but the improvement is minor compared to Abbyy and Tesseract respectively. On the word level we observe a similar tendency. Surprisingly, the accuracy of the Abbyy improves with the external word lists, as opposed to the other systems whose results actually decrease when run with the external word lists.

OCR engine	Without word lists		With four word lists	
	CAR (%)	WAR (%)	CAR (%)	WAR (%)
Abbyy	91.54	83.39	91.45	83.6
Tesseract	91.39	87.37	91.39	85.53
Abbyy-Tesseract	91.64	84.21	91.5	83.9

Table 1: Evaluation results of 400 pages for the whole time period 1818-2018, run with three types of OCR engine setups. Without word lists (our baseline) on the left and with all word lists on the right.

¹⁰<http://www.grepect.de/>

¹¹The resources that are developed in this project, covering the years up to 1909 will be made freely available for download through <https://vlo.clarin.eu>

Second, we calculated the CAR and WAR results for each run on the same material, this time separately with each word list and divided into three time periods. Table 2 shows the evaluation results of our runs without external word lists (baseline) and with the external word lists both separately and all combined.¹²

OCR engine	1818-1906		1907-1996		1997-2018	
	CAR (%)	WAR (%)	CAR (%)	WAR (%)	CAR (%)	WAR (%)
Abbyy baseline	87.38	71.47	94.22	92.11	93.34	92.58
Tesseract baseline	88.14	81.26	92.48	89.93	92.84	91.17
Abbyy-Tesseract baseline	87.46	72.66	93.36	90.05	93.43	90.11
Abbyy Dalin	87.31	71.87	94.05	91.98	93.31	92.69
Tesseract Dalin	87.86	76.74	93.32	91.44	92.55	91.07
Abbyy-Tesseract Dalin	87.32	72.49	93.13	89.7	93.34	89.92
Abbyy Saldo	87.19	71.33	93.02	89.15	93.38	89.72
Tesseract Saldo	87.89	76.73	92.46	89.39	92.95	90.21
Abbyy-Tesseract Saldo	87.18	71.96	93.05	89.42	93.27	89.78
Abbyy Saol-hist	87.37	71.96	93.13	89.44	93.47	89.83
Tesseract Saol-hist	87.84	76.75	92.46	89.38	92.95	90.19
Abbyy-Tesseract Saol-hist	87.41	72.73	93.19	89.82	93.42	90.03
Abbyy Fem	87.1	70.81	92.81	88.62	93.15	89.2
Tesseract Fem	87.88	76.72	92.45	89.37	92.96	90.2
Abbyy-Tesseract Fem	87.19	71.61	92.89	89.03	93.09	89.4
Abbyy All	87.36	71.88	93.09	89.52	93.4	89.76
Tesseract All	88.03	77.09	92.57	89.41	92.97	90.16
Abbyy-Tesseract All	87.41	72.35	93.13	89.75	93.34	89.82

Table 2: Evaluation results of 400 pages divided into three time periods. First run without external word lists (baseline), four runs with external word lists, one run for each word list and last run with all four word lists. Each run was performed with three engines. The results marked in bold highlight successful runs that outperform the baseline.

As can be observed in Table 2, Abbyy shows a small improvement on the word level with Dalin, Saol-hist and all word lists for 1818-1906. Interestingly, it also shows a small improvement on the word level with Dalin for 1997-2018. There is some improvement for Tesseract both on the character and word levels, with Dalin for the later period 1907-1996, and minor improvement with all word lists for the same period. Neither of the runs for 1907-1996 has improved over the baseline for Abbyy, Tesseract or Abbyy-Tesseract. This could be explained by the fact that the systems are eager to find a lexical match in the external word lists, and since the external lists get higher priority, wrong words are being replaced. Thus, there is a higher percentage of words that are replaced with incorrect ones for that particular content. Consequentially, CAR is also decreasing. Another explanation of the low performance is the high ratio of out-of-word vocabulary for this period as seen in Figure 4.

Further, Saldo, Saol-hist and all word lists improve the character accuracy of Abbyy and Tesseract for the period 1997-2018. Tesseract shows a minor improvement with Fem on the character level.

6 Summary and Conclusions

We explored the effect of adding curated word lists to improve the two-OCR engine system when digitizing Swedish newspapers from 200 years. We found that the addition of word lists in combination with the two-OCR engine system did not provide the expected significant improvement of the OCR result, even though some combinations proved more successful than others.

¹²The differences in time for the different runs with the two-OCR engine were marginal. It took approximately 157 seconds to process one page. To process the whole material, we run 3 OCR processes simultaneously, which took 6 hours. With additional OCR processes, this time could be reduced accordingly.

One explanation of the OCR results reported in this study might stem from the fact that our method for matching whether a word appears in the word list is rather naive. It simply applies string matching. To make the best out of the word lists, matching should be combined with a Levenshtein-distance method. Another problem could be grounded in the correlation between specific types of OCR errors and images as well as graphical elements in the printed page, resulting in occurrences of segments with no referable word because the system mistook an image for being a word. When we examined the effect of word lists using the segmentation tool we found that external word lists could improve the word accuracy if we consider the confidence score as a factor in deciding whether a word is correct or not. Consequently, words that receive a higher confidence level when found in the word lists, might lead to better word prioritisation and in turn improve the quality of the text as a whole. Our manual inspection showed the system had to “guess” when deciding on the correct word, hence resulting in a more or less random output. Furthermore, a closer look on the source material showed 30% of the material has low print quality. When we studied the source material to get an estimate of the distribution of Antiqua and Blackletter by going through the images manually we learned that for 1818-1906 25% is Antiqua and 75% is Blackletter, for 1907-2018 75% is Antiqua and 25% is Blackletter. Therefore more thorough qualitative analysis combined with quantitative analysis of the relation between OCR errors and the source material could provide the basis for a taxonomy of error types which will further contribute to the development of benchmarks and quality indicators. Segmentation precision also plays a significant role for the amount of character and word errors. In the complete material we identified around 100 segments with considerable difference in sequential tokens between the ground truth and the OCRred data caused by inaccurate segmentation. To address this problem, preprocessing techniques are being tested by Zissor to improve the accuracy of the image processing.

There are however some unexpected variations in the accuracy results that have to be examined in detail in relation to the specific word lists and the given time period of the sample. The quality control of the two engine system verifies that the word lists are taken into consideration during the OCR process but their apparent unpredictable effect on the results have to be further analysed. The external word lists had noticeable effect on the internal confidence values of the OCR programs. The effect of these combined with the possibility to include the rate of correspondence between the results from the two-OCR engine (on the word level), could be used as a variable to indicate the quality of the results, something we will explore later in the project. Another important variable seems to be the scoring scheme which is governed by a set of rules for deciding on which of the systems processed the correct word. The impact of the rule set will be analysed once the consequences of the use of word lists is further examined.

The preliminary results reported here are based on the first quantitative evaluation against the complete ground truth. By studying the results manually we could observe some correlation between specific types of OCR errors and images as well as graphical elements in the printed page. Future work aims to combine these findings with a substantial qualitative analysis to address possible sources of errors resulting from degradation of paper, bad print quality and complexity of layout and typography. The final results will be reported in future publications but these preliminary findings indicate some areas that will receive a more detailed focus as the project progresses.

Acknowledgements

The research presented here is funded by the Swedish Research Council (the project *Evaluation and refinement of an enhanced OCR-process for massdigitisation* grant: IN18-0940:1). It is also supported by Språkbanken Text and Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant 821-2013-2003).

References

Chantal Amrhein and Simon Clematide. 2018. Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods. *Journal for Language Technology and Computational Linguistics (JLCL)*, 33(1):49–76.

- Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer, Berlin, Germany.
- Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet’s yang. *Language Resources and Evaluation*, 47(4):1191–1211.
- Rafael C. Carrasco. 2014. An open-source OCR evaluation tool. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH, pages 179–184, NY, USA. Association for Computing Machinery.
- Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE.
- Simon Clematide and Phillip Ströbel. 2018. Improving OCR quality of historical newspapers with handwritten text recognition models. In *Workshop DARIAH-CH*, Neuchâtel. University of Neuchâtel.
- Dana Dannélls, Lars Björk, and Torsten Johansson. 2019. Evaluation and refinement of an enhanced ocr process for mass digitisation. In *Proceedings of Digital Humanities in the Nordic Countries*, pages 112–123, Copenhagen, Denmark. CEUR-WS.org.
- Senka Drobac, Pekka Kauppinen, and Krister Linden. 2019. Improving OCR of historical newspapers and journals published in Finland. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 97–102, Brussels, Belgium. ACM.
- Steffen Eger, Tim vor der Brück, and A. Mehler. 2016. A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. *The Prague Bulletin of Mathematical Linguistics*, 105:77–99.
- Ian Gregory, Paul Atkinson, Andrew Hardie, Amelia Joulain-Jay, Daniel Kershaw, Catherine Porter, Paul Rayson, and C.J. Rupp. 2016. From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection. *Journal of Siberian Federal University, Humanities and Social Sciences*, 9(4):994–1006.
- Ido Kissos and Nachum Dershowitz. 2016. OCR error correction using character correction and feature-based word classification. In *12th IAPR Workshop on Document Analysis Systems DAS*, pages 198–203, Santorini, Greece. IEEE.
- Mika Koistinen, Kimmo Kettunen, and Tuula Pääkkönen. 2017. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur and Antiqua Models and Image Preprocessing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA*, Gothenburg, Sweden. Association for Computational Linguistics.
- William B. Lund, Daniel D. Walker, and Eric K. Ringger. 2011. Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines. In *International Conference on Document Analysis and Recognition*, pages 764–768, Beijing, China. IEEE.
- Jie Mei, Aminul Islam, Yajing Wu, Abidrahman Mohd, and Evangelos E Milios. 2016. Statistical learning for OCR text correction. *arXiv preprint*, abs/1611.06950.
- Thi-Tuyet-Hai Nguyen, Mickaël Coustaty, Doucet Antoine, and Nhu-Van Nguyen. 2018. Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. In *20th International Conference on Asia-Pacific Digital Libraries, ICADL*, Hamilton, New Zealand. Lecture Notes in Computer Science.
- Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *Proceedings of the 18th Joint Conference on Digital Libraries, JCDL*, pages 29–38, Champaign, IL, USA. IEEE.
- Thi-Tuyet-Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, and Antoine Doucet. 2020. Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *JCDL: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Virtual Event*, pages 333–336, New York, NY. Association for Computing Machinery.
- Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

- Pratip Samanta and Bidyut B. Chaudhuri. 2013. A simple real-word error detection and correction using local word bigram and trigram. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 211–220, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).
- Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored ocr post-correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.
- Miikka Silfverberg, Pekka Kauppinen, and Krister Lindén. 2016. Data-driven spelling correction using weighted Finite-State methods. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 51–59, Berlin, Germany. Association for Computational Linguistics.
- Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633, Curitiba, Brazil. IEEE.
- Uwe Springmann and Anke Lüdeling. 2017. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11(2).
- Zissor. 2017. Zissor Content System. Implementation of dual OCR motors, Phase II. Technical report, Kungliga biblioteket (KB), Stockholm, Sweden.

PoetryLab as Infrastructure for the Analysis of Spanish Poetry

Javier de la Rosa

LINHD

UNED

Madrid, Spain

versae@linhd.uned.es

Álvaro Pérez

LINHD

UNED

Madrid, Spain

alvaro.perez@linhd.uned.es

Laura Hernández

LINHD

UNED

Madrid, Spain

laura.hernandez@linhd.uned.es

Aitor Díaz

Control and Communication Systems

UNED

Madrid, Spain

adiazm@scc.uned.es

Salvador Ros

Control and Communication Systems School of Human Sciences and Technology

UNED

Madrid, Spain

sros@scc.uned.es

Elena González-Blanco

IE University

Madrid, Spain

egonzalezblanco@faculty.ie.edu

Abstract

The development of the network of ontologies of the ERC POSTDATA Project brought to light some deficiencies in terms of completeness in the currently available European poetry corpora. To tackle the issue in the realm of the Spanish poetic tradition, our approach consisted in designing a set of tools that any scholar could use to automatically enrich the analysis of Spanish poetry. The effort crystallized in the PoetryLab, an extensible open source toolkit for syllabification, scansion, enjambment detection, rhyme detection, stanza identification, and historical named entity recognition for Spanish poetry. We designed the system to be interoperable, compliant with the project ontologies, easy to use by tech-savvy and non-expert researchers, and requiring minimal maintenance and setup. Furthermore, we propose the integration of the PoetryLab as a core functionality in the tool catalog of CLARIN for Spanish poetry.

1 Introduction

The main goal of the ERC-funded POSTDATA Project (Curado Malta and González-Blanco, 2016)¹ was to formalize a network of ontologies capable of expressing any poetic expression and its analysis at the European level, thus enabling scholars all over Europe to interchange their data using Linked Open Data. The POSTDATA Project bridges the digital gap between traditional cultural assets and the growing world of data. It is focused on poetry analysis, classification, and publication, applying Digital Humanities methods of academic analysis –such as XML-TEI encoding (Dombrowski and Denbo, 2013; Flanders and Hamlin, 2013)– in order to look for standardization, as well as innovation by using semantic

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) funded by the European Research Council (ERC, <https://erc.europa.eu>) under the research and innovation program Horizon2020 of the European Union: <http://postdata.linhd.uned.es/>

web technologies (Cigarrán-Recuero et al., 2014) to link and publish literary datasets in a structured way in the linked data cloud. The advantages of making poetry available online as machine-readable linked data are threefold: first, the academic community would have an accessible digital platform to work with poetic corpora and to contribute to its enrichment with their own texts; second, this way of encoding and standardizing poetic information will be a guarantee of preservation for poems published only in old books or even transmitted orally, as texts will be digitized and stored; third: datasets and corpora will be available and open access to be used by the community for other purposes, such as education, cultural diffusion or entertainment.

However, varied research interests result in corpora that might not share the same facets of an analysis. To alleviate this concern and foster the completeness of the interchanged corpora, our team set to build a software toolkit to assist in the analysis of poetry. This paper details the first iteration of the PoetryLab, an extensible open source toolkit for syllabification, scansion, enjambment detection, rhyme detection, stanza identification, and historical named entity recognition for Spanish poetry, that achieves state of the art performance in the tasks for which reproducible alternatives exist.

2 PoetryLab

Despite a long and rich tradition (Bello, 1859; Navarro Tomás, 1991; Caparrós, 2014), not many computational tools have been created to assist scholars in the annotation and analysis of Spanish poetry. With ever increasing corpora sizes and the popularization of distant reading techniques, the possibility to automate part of the analysis became very attractive. Although solutions exist, they are either incomplete, e.g., scansion of fixed-metre poetry (Agirrezabal et al., 2016; Navarro-Colorado, 2017; Gervas, 2000; Agirrezabal et al., 2017), not applicable to Spanish (Agirrezabal et al., 2017; Hartman, 2005), or not open or reproducible (Gervas, 2000). Moreover, disparate input and output formats, operating system requirements and dependencies, and the lack of interoperability between software packages, further complicated the limited ecosystem of tools to analyze Spanish poetry. These limitations guided the design of the PoetryLab as a two layer system: a REST API that operates as middleware connecting the different tools, and a consumer web-based UI that exposes the functionality to non-experts users. All tools are released as independent Python packages with their own command line interface applications (where appropriate), and are ready to produce RDF triples compliant with the POSTDATA Project network of ontologies. Figure 1 shows a diagram of the general architecture of the system.

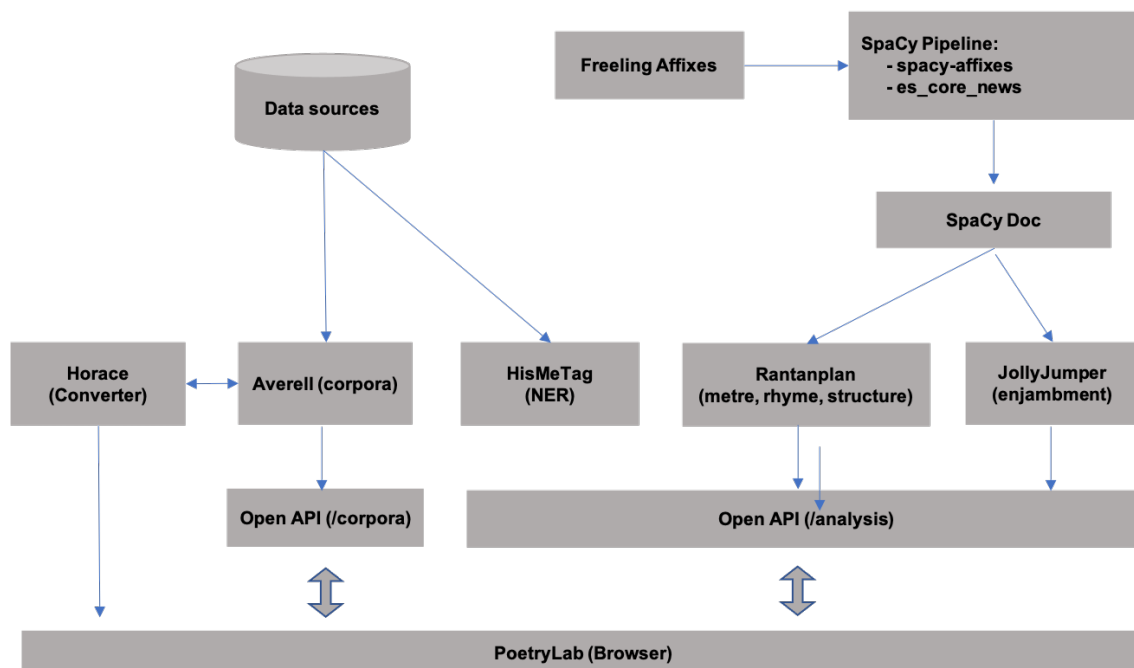


Figure 1: General architecture of the PoetryLab.

This granular design allows for each component of the PoetryLab to be used and deployed as a set of Docker images, which makes managing the different tools lifecycle and versioning a less problematic issue. We tested this approach by using *ouroboros*², a service to automatically update running docker containers with the newest available image, and the demo site of the PoetryLab has been running without major incidents over a year now³. We feel hosting the PoetryLab as one of the tools in the catalog of software tools available in CLARIN would be a good addition to its ecosystem, since it requires little effort to setup and the maintenance of the different tools is deferred to their own maintainers, as it usually happens in the Open Source ecosystem, making it easy for hot-replacement when new versions become available. Moreover, the use of Docker containers as deployment strategy and the fact that the tools are stateless, allow the use of lambda architectures to minimize the running costs.

2.1 PoetryLab API

At its core (see Figure 1), the PoetryLab API provides a self-documented Open API (OpenAPI Initiative, 2017) that connects the independent packages together and exposes their outputs in different formats. Two main endpoints provide functionality to analyze texts uploaded by an user (`/analysis`), and to work with a catalog of existing corpora (`/corpora`)⁴.

2.1.1 Endpoint `/analysis`

The first endpoint of the PoetryLab API, `/analysis`, leverages three tools to perform several aspects of the analysis of a poem: scansion and rhyme identification, enjambment detection, and named entity recognition (i.e. Rantanplan, Jollyjumper, and Hismetag). AnCora (Taulé et al., 2008), the corpus spaCy is trained on for Spanish, splits most affixes thus losing the multi-token word information and causing some failures in the part of speech tags it produces. To circumvent this limitation and to ensure clitics were handled properly, we integrated Freeling’s affix rules via a custom built pipeline for spaCy. The resulting package, *spacy-affixes*⁵, splits words with affixes so spaCy can handle their part of speech correctly (Padró and Stanilovsky, 2012). Getting this information right was crucial to identify the stress of some monosyllabic and disyllabic words, and to find a special kind of enjambment called *sirrematic* in which a grammatical unit is divided in two lines (see Table 1 for a summary of the performance of our scansion system). The outputs of these two tools are then transformed to accommodate to the definitions given in the network of ontologies developed within the POSTDATA Project.

Method	Accuracy
(Gervas, 2000)	88.73
(Navarro-Colorado, 2017)	94.44
(Agirrezabal et al., 2017)	90.84
Rantanplan (ours)	96.23

Table 1: Scores on Navarro-Colorado’s fixed-metre 1,400 verses corpus. Best scores in bold.

Lastly, the PoetryLab API provides a pluggable architecture that allows for the integration of external packages developed in languages other than Python. This is the case for our named entity recognition system, *HisMeTag* (Platas et al., 2021), developed in Java and connected to the PoetryLab API through an internal REST API exposed via Docker. The only requirement is to consume raw plain text and to produce both a JSON output and RDF triples compliant with the POSTDATA Project network of ontologies.

2.1.2 Endpoint `/corpora`

The second available endpoint, `/corpora`, aims to facilitate working with existing repositories of annotated poetry. *Averell*, the tool that handles the corpora, is able to download an annotated corpus and

²<https://github.com/pyouroboros/ouroboros>

³<http://postdata.uned.es/poetrylab>

⁴A demo with the Open API user interface is available at <http://postdata.uned.es:5000/ui/>.

⁵<https://github.com/linhd-postdata/spacy-affixes/>

reconcile different TEI entities to provide a unified JSON output and RDF triples at the desired granularity. That is, for their investigations some researchers might need the entire poem, poems split line by line, or even word by word if that is available. Averell allows to specify the granularity of the final generated dataset, which is a combined JSON or RDF with all the entities in the selected corpora.

Name	Size	Docs	Words	Granularity	License
Disco V2	22M	4,088	381,539	stanza, line	CC-BY
Disco V3	28M	4,080	377,978	stanza, line	CC-BY
Sonetos Siglo de Oro	6.8M	5,078	466,012	stanza, line	CC-BY-NC 4.0
ADSO 100 poems corpus	128K	100	9,208	stanza, line	CC-BY-NC 4.0
Poesía Lírica Castellana del Siglo de Oro	3.8M	475	299,402	stanza, line, word, syllable	CC-BY-NC 4.0
Gongocorpus	9.2M	481	99,079	stanza, line, word, syllable	CC-BY-NC-ND 3.0 FR
Eighteenth Century Poetry Archive	2.4G	3,084	2,063,668	stanza, line, word	CC BY-SA 4.0
For Better For Verse	39.5M	103	41,749	stanza, line	Unknown
Métrique en Ligne	183M	5,081	1,850,222	stanza, line	Unknown
Biblioteca Italiana	242M	25,341	7,121,246	stanza, line, word	Unknown

Table 2: Available corpora in Averell

Each corpus in the catalog must specify the parser to produce the expected data format. At the moment, there are parsers for five corpora, all using the TEI tag set (see Table 2). For corpora not in our catalog, the researcher can define her own or reuse one of the existing ones to process a local or remote corpus.

Moreover, for plain text local corpora Averell allows to post-process the raw texts with Rantanplan to enrich poems with their metrical and structural information as detected by the tool. The result of this process can still be combined seamlessly with the existing corpora in the catalog.

2.2 PoetryLab UI

The PoetryLab API is then used to provide with functionality to a React-based web interface that non-technical scholars can use to interact with the packages in a graphical way (see Figure 2). The frontend gives the option to download the generated data in both JSON and POSTDATA Project RDF triples formats⁶.

The web interface is run entirely in the browser as a stateless application. However, the collection of analyzed poems are saved to the browser local storage which persists between sessions and restarts. Unfortunately, it lacks a user management system that could provide with persistent storage in a backend.

2.3 PoetryLab tools

Most tools included in PoetryLab are generally available to the public as standalone libraries or applications. Rantanplan, JollyJumper, HisMeTag and Horace are all orchestrated to be used by tech-savvy and non-expert researchers, through the Poetrylab UI or as standalone applications. Moreover, Averell and spacy-affixes are auxiliary packages that the rest of the toolkit builds upon.

2.3.1 Rantanplan

Rantanplan⁷ (de la Rosa et al., 2020) is a Python library for the automated scansion of Spanish poetry. Scansion is the measurement of the rhythm of verses of a poem. It is comprised of four modules that work together to perform scansion of both fixed-metre as well as mixed-metre poetry: part-of-speech (PoS) tagger, syllabification, stress assignment, and metrical adjustment (see Algorithm 1). Rantanplan is fast and accurate as it is built using SpaCy while leveraging Freeling rules to handle clitics and other

⁶<http://postdata.uned.es/poetrylab>

⁷<https://github.com/linhd-postdata/rantanplan/>

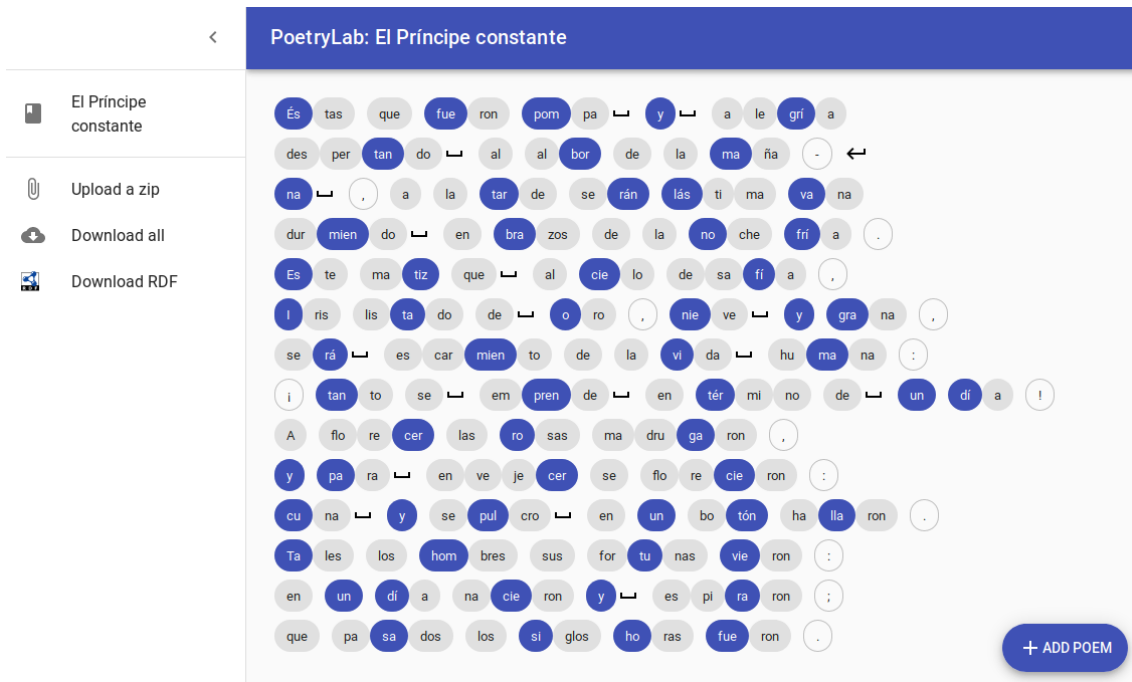


Figure 2: PoetryLab showing stressed syllables (blue), sinalefas (◡) and enjambments (◀).

nuances of the Spanish language through spacy-affixes. Rantanplan is the current state-of-the-art both in terms of speed and accuracy on the reference corpus used by Navarro-Colorado (Navarro-Colorado et al., 2016), yielding a 96.23% of accuracy on a fixed-metre corpus of hendecasyllabic verses.

Rantanplan is also able to identify up to 45 different types of the most significant Spanish stanzas (see Figure 3). Stanzas are structural units formed by lines of verses, and therefore they are related to the author style and even historical preferences that make identifying them a complicated task. Rantanplan first analyzes the verses that comprise a stanza to gather information about their lengths, rhyme pattern and rhyme type. With this information, it checks a set of rules crafted and sorted by domain experts in order to propose a stanza type. Rantanplan is able to correctly identify 78.63% of the stanza types in a corpus of over 5000 stanzas manually annotated.

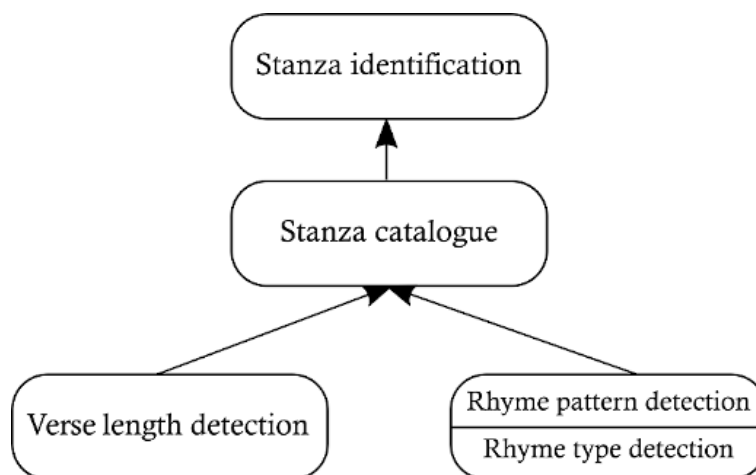


Figure 3: Stanza detection architecture

Algorithm 1: Rantanplan main algorithm

Input: Poem as a sequence \mathcal{L} of lines $\langle l_1, l_2, \dots, l_n \rangle$,
each with a sequence \mathcal{W} of words $\langle w_1, w_2, \dots, w_n \rangle$

Output: Data structure with metrical information

```
for  $l_i \in \mathcal{L}$  do
  for  $w_i \in \mathcal{W}$  do
     $tag_i \leftarrow \text{pos}(w_i)$ 
     $syllables_i \leftarrow \text{syllabify}(w_i)$ 
     $stresses_i \leftarrow \text{stress}(syllables_i, tag_i)$ 
  end
   $groups \leftarrow \text{phonological}(syllables, stresses)$ 
   $pattern \leftarrow \text{transform}(groups)$ 
  if  $length$  then
    while  $|pattern| < length$  do
       $g \leftarrow \text{generate\_phonological}(\mathcal{W})$ 
       $pattern \leftarrow \text{transform}(g)$ 
    end
  end
   $patterns \leftarrow \text{push}(pattern)$ 
end
 $rhyme \leftarrow \text{extract}(patterns)$ 
 $stanza \leftarrow \text{identify}(pattern, rhyme)$ 
return  $patterns, rhyme, stanza$ 
```

2.3.2 JollyJumper

Jollyjumper⁸ is our enjambment detection Python library for Spanish (see Figure 4). Enjambment is the metric phenomenon that occurs when there is a disagreement between the syntactic unit and metric unit, that is, when the syntactic unit exceeds the verse pause and overflows into the next verse, or when elements of the unity of sense which constitutes the next verse are anticipated at the end of a verse. Automatic detection allows for large scale quantitative analyses on the phenomenon. As an example, we ran the system on approx. 9,000 sonnets from the 15th to the early 20th century, examining patterns of evolution in the distribution of the use of enjambment according to line-position in the sonnet.

2.3.3 HisMeTag

HisMeTag⁹ is a Java tool for the identification and tagging of place names in Medieval Spanish texts. It combines lexical, syntactic, and semantic analysis with NLP technologies. This task involves specific challenges: the complex morphosyntactic characteristics in proper-noun use in medieval texts, the lack of strict orthographic standards, and the diachronic and geographical variations in Spanish from the 12th to the 15th century. The system is also integrated in Poetrylab API as a Docker image which then exposes its functionality using the same common API.

2.3.4 Horace

Horace¹⁰ is a translation tool between the natively produced PoetryLab JSON format consumed internally and the semantic formats of the POSTDATA Project. This tool is capable of reading the outputs of most of the tools in PoetryLab to build knowledge graphs in RDF format compliant with the POSTDATA set of ontologies. The step is crucial as it allows exposing the information produced by the toolkit through a SPARQL endpoint, thus enabling the interoperability and sharing of both the data included in the different public corpora and the automated annotations produced by our tools.

⁸<https://github.com/linhd-postdata/jollyjumper>

⁹<https://github.com/linhd-postdata/hismetag>

¹⁰<https://github.com/linhd-postdata/horace>

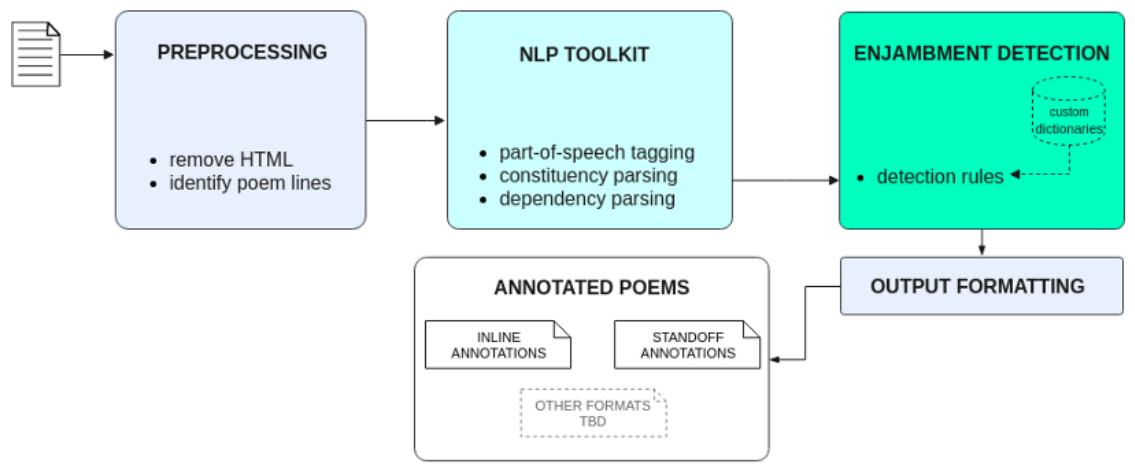


Figure 4: Jollyjumper General architecture

2.3.5 Averell

This tool is a one-stop command line interface application to gather existing corpora. Averell¹¹ is able to download annotated poetic corpora from different sources, parse them, and turn them into a single JSON or CSV file ready for analysis. This is especially useful when setting benchmarks, since it reduces the burden of cleaning and parsing from the researchers. Since Averell makes file translation internally for each corpus in its catalog, it also allows researchers to create a corpus that is the combination of several other corpora, selecting sets of poems that meet a specific set of conditions (e.g., only corpora in Italian with manually verified annotations of metrical patterns). Moreover, the granularity at which this selection can be made is to the preference of the researcher, being able to choose between poem, stanza, verses, words, and even syllables (whenever available in the specific corpus, see Table 2).

3 Conclusion

The PoetryLab has proven useful in that it provides an integrated set of tools for Spanish poetry scholars. It might become useful at several stages of the research cycle. Averell helps build ad-hoc corpora, which may include metrical information generated by Rantanplan, rhetorical devices as detected by JollyJumper, and even historical named entities as recognized by HisMeTag. It also produces machine readable and interoperable data suitable to be ingested into a triple store compliant with the POSTDATA Project network of ontologies (i.e., Horace). In fact, this approach is already being tested as we export the analysis of poems and feed them into a Virtuoso Universal Server that integrates with the POSTDATA Project network of ontologies to produce repertoires knowledge graphs.

The PoetryLab will be eventually integrated into the larger POSTDATA Project public website, making working with European repositories of poetry a more pleasant task, and assisting whenever possible with the metrical and rhetorical side of the analysis. Moreover, more attention needs to be put into building a friendly web user interface useful for different user profiles.

Acknowledgements

Research for this paper has been achieved thanks to the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) obtained by Elena González-Blanco. This project is funded by the European Research Council (<https://erc.europa.eu>) (ERC) under the research and innovation program Horizon2020 of the European Union.

¹¹<https://github.com/linhd-postdata/horace>

References

- Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4.
- Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A comparison of feature-based and neural scansion of poetry. *arXiv preprint arXiv:1711.00938*.
- Andrés Bello. 1859. *Principios de la ortología i métrica de la lengua castellana..* la Opinión.
- José Domínguez Caparrós. 2014. Teoría métrica del verso esdrújulo. *Rhythmica: revista española de métrica comparada*, 12:55–96.
- Juan Cigarrán-Recuero, Joaquín Gayoso-Cabada, Miguel Rodríguez-Artacho, María-Dolores Romero-López, Antonio Sarasa-Cabezuelo, and José-Luis Sierra. 2014. Assessing semantic annotation activities with formal concept analysis. *Expert Systems with Applications*, 41(11):5495–5508.
- Mariana Curado Malta and Elena González-Blanco. 2016. Postdata. towards publishing european poetry as linked open data. In *International Conference on Dublin Core & Metadata Applications*. DCMI.
- Javier de la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, and Elena González-Blanco. 2020. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.
- Quinn Dombrowski and Seth Denbo. 2013. Tei and project bamboo. *Journal of the Text Encoding Initiative*, 5.
- Julia Flanders and Scott Hamlin. 2013. Tapas: building a tei publishing and repository service. *Journal of the Text Encoding Initiative*, 5.
- Pablo Gervas. 2000. A logic programming application for the analysis of spanish verse. In *International Conference on Computational Logic*, pages 1330–1344. Springer.
- Charles O Hartman. 2005. The scandroid 1.1. *Software available at <http://oak.conncoll.edu/cohar/Programs.htm>*.
- Borja Navarro-Colorado, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.
- Borja Navarro-Colorado. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.
- Tomás Navarro Tomás. 1991. Métrica española. *Reseña histórica y descriptiva*, 50.
- OpenAPI Initiative. 2017. Openapi specification. Retrieved from *GitHub*: <https://github.com/OAI/OpenAPI-Specification/blob/master/versions/3.0>, 1.
- Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.
- M^a Luisa Díez Platas, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*, 72(2):224–238. [.eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24399](https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24399).
- Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *International Conference on Language Resources and Evaluation*.

Contagious “Corona” Compounding by Journalists in a CLARIN Newspaper Monitor Corpus

Koenraad De Smedt

University of Bergen, Norway

desmedt@uib.no

Abstract

Newspaper monitor corpora, which incorporate new materials on a regular basis, are particularly useful for tracking linguistic changes spurred by current developments. The COVID-19 pandemic prompted a case study in the Norwegian Newspaper Corpus. The corpus was mined for productive compounds with the stems “corona” and its alternative spelling “korona”, tracing their frequencies and dates of first occurrence during the first wave of the pandemic. The quantitative analysis not only monitored the daily volume and variation of such compounds, but also the dynamics of vocabulary growth, and a change in their preferred spelling. The paper concludes with reflections on methodology and data sources.

1 Introduction

The COVID-19 pandemic, which started to spread around the world in the spring of 2020, has quickly become the subject of much research, not just in medicine, but also in the social sciences and humanities. For various research purposes, corpora containing specific types of discourse have been compiled, from scientific articles (Lu Wang et al., 2020) to tweets (Dimitrov et al., 2020). Furthermore, large *monitor* corpora, which are regularly updated from a wide range of sources, are allowing lexicographers and others to detect linguistic changes in almost real time (OED Editorial, 2020; Paton, 2020).

Newspaper corpora are essentially time-stamped journalistic descriptions of daily events. Newspaper *monitor* corpora are moreover regularly updated; thereby they are not only a window into the course of current events, but they also provide up-to-date data samples of journalistic language. Such corpora are unfortunately scarce. In fact, the only monitor corpus that I could identify in the CLARIN resource family overview of newspaper corpora¹ is the Norwegian Newspaper Corpus (Andersen and Hofland, 2012) at the CLARINO Bergen Centre (De Smedt et al., 2016). This large resource, containing two billion words and growing, has been useful in earlier studies of neologisms, loan words and other vocabulary expansion (Andersen, 2012).

The pandemic provided an exceptional opportunity to further demonstrate the use of this monitor newspaper corpus. It is a rare experience to observe a sudden dramatic increase in the vocabulary in a very short period of time. Events related to the outbreak and pandemic were extensively discussed in the media all over the world. This stimulated the coining of new words in many languages. Particularly striking in Norwegian was the productivity of compounds, such as *testkø* (“testing queue”), *hjemmeisolering* (“home isolation”), *kommunekarantene* (“municipal quarantine”), *smittesporingsapp* (“contagion tracing app”) and *flokkimmunitetsstrategien* (“the herd immunity strategy”). Like other Germanic languages, Norwegian has indeed very productive compounding, and compounds are normally written as one word.

In this context, *corona/korona* stands out. From January 2020, the word by itself quickly became a common term for both the virus, the disease and the epidemic. It also became by far the most frequent initial part of compounds. What is unique about *corona/korona*, moreover, is that practically all its compounds

This work is licensed under a Creative Commons Attribution 4.0 International License. License details:
<http://creativecommons.org/licenses/by/4.0/>

¹<https://www.clarin.eu/resource-families/newspaper-corpora>

were completely new. Before 2020, *corona* occurred in only a handful of relevant compounds, such as *coronavirus*, (“corona virus”), *koronafamilien* (“the corona family”), and *corona-vaksiner* (“corona vaccines”).² During the first wave of the pandemic, there was an explosion of new compounds, such as *koronatelefon* (“corona telephone”), *koronadødsfall* (“corona death”), *koronafrykten* (“the fear of corona”), *corona-cruiset* (“the corona cruise”) and *coronatider* (“corona times”). In contrast to *virus*, the term *corona/korona* is more specific and eye-catching, something that appeals to newspaper editors. This may explain why its use seemed contagious in the journalistic sphere.

The current work is a case study showing the possibilities of mining a newspaper monitor corpus accessible through CLARIN. Its primary objective is to trace the productivity of compounds with *corona* and its alternative spelling *korona* during the first wave of the pandemic. The hypothesis was not only that an evolution in the tempo of vocabulary expansion had occurred, but also an evolution in the ratio between types and tokens, so the goal was to measure the extent and speed of these trends. Another objective was to trace spelling change in terms of changing proportions of the two variant spellings. There were indications that the normalization by the Language Council near the end of January 2020 had influenced journalists’ spelling, but the extent of the change had not been quantified previously.

2 Data and Method

The Norwegian Newspaper Corpus (Andersen and Hofland, 2012) was the data source for the present study. It is updated every night by harvesting publicly accessible articles from ten major Norwegian online newspapers.³ At every automatic update, boilerplate is removed so that nearly clean text is left, and every article is tagged with the date and the source. This corpus is accessible in two ways.

One way to use the corpus is through an instance of the IMS Corpus Workbench (CWB; Evert and Hardie, 2011).⁴ In this system, the corpus is split up in different sections, most or them covering one year. Searches can be specified by regular expressions and can be limited to a year, a month or a date. Some disadvantages of the CWB version are that search can only be performed in one section at a time, and that it is not possible to specify arbitrary start and end dates. Another disadvantage is that the system does not have a download function, so that relevant items must be extracted from the HTML encoding of the search result pages.

The corpus is also accessible through the Corpuscle corpus management and search system⁵ (Meurer, 2012) at the CLARINO Bergen Center. This system has a better interface and a more powerful and efficient query system (Meurer, 2020). It allows the specification of arbitrary start and end dates in queries. It also offers download of matching strings, with optional annotation features, to a file with tab-separated values. Unfortunately, this version of the corpus is updated less regularly than the CWB version. Both versions were consulted, but the data from Corpuscle, which was up to date until March 8, 2021, are the basis for the present study.

The query "[ck]orona.*" %c :: year = "2020|2021]" was used in Corpuscle to retrieve all occurrences of words starting with *corona* or *korona*, in uppercase or lowercase, from the Bokmål⁶ section of the corpus, tagged with the year 2020 or 2021. All matches were downloaded as a tab-separated file with keywords, newspaper codes and dates. The first observation was on January 9, 2020, and the last one on March 8, 2021. The period with observations thus spans a year and two months, or 425 days to be precise.

The base forms *corona/korona* and their inflected forms were removed, as well as obvious spelling errors and unrelated words such as *koronar* og *coronal*. The cleaned word list has 167957 tokens, which are all compounds, with or without hyphens. Preprocessing, analysis and plotting was performed with a shell script that called programs in Awk, Python and R.

²Before 2020 these referred to viruses other than SARS-CoV-2, primarily SARS-CoV and MERS-CoV.

³The following newspapers, with their codes, are represented in the corpus: Adresseavisen (AA, Trondheim), Aftenposten (AP, Oslo), Bergens Tidende (BT, Bergen), Dagsavisen (DA, Oslo), Dagbladet (DB, Oslo), Dagens Næringsliv (DN, Oslo), Fædrelandsvennen (FV, Kristiansand), Nordlys (NL, Tromsø), Stavanger Aftenblad (SA, Stavanger) and Verdens Gang (VG, Oslo).

⁴<http://korpus.uib.no/avis/bokm.html>

⁵<http://clarino.uib.no/korpuskel>

⁶The corpus also has a separate Nynorsk section, which is much smaller and was not used in this study.

3 Analysis

3.1 Spelling

From October 1998 until the end of 2019, the few existing compounds with *corona*, in senses related to the virus, occur in the Norwegian Newspaper Corpus only with initial *c-*, whereas *koronautbrudd* (“corona outbreak”) with *k-* was only used in the sense of “solar flare.” An n-gram search of both spellings in the digital newspaper collection of the National Library of Norway, which goes further back in time, confirms this practice, as shown in Figure 1. In January 2020, the spelling with *c-* was still very dominant in the Norwegian Newspaper Corpus.

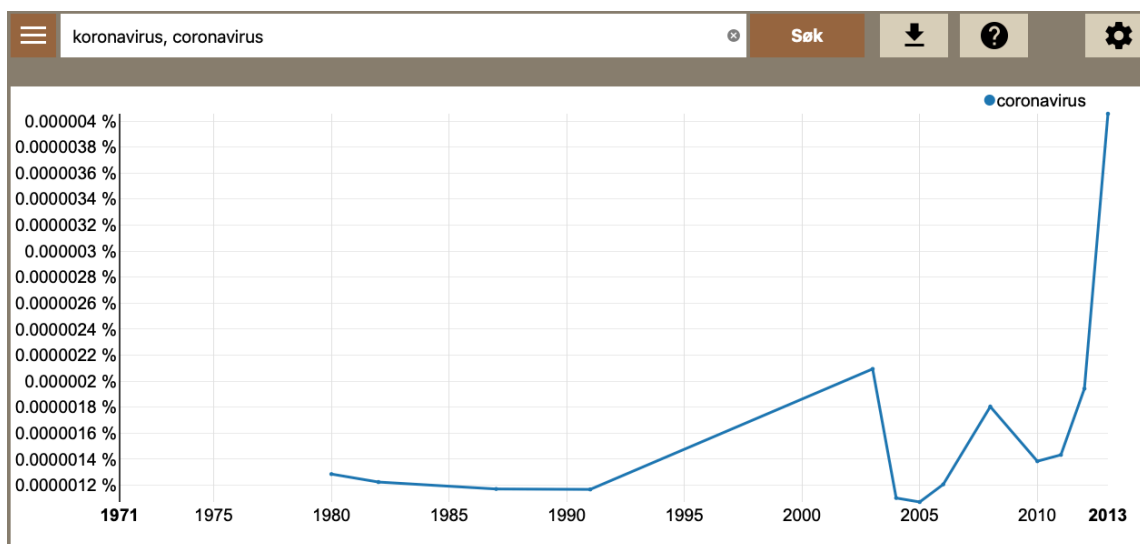


Figure 1. Relative frequencies in the newspaper collection of the National Library of Norway: *coronavirus* occurs from 1980 until (at least) 2013, *koronavirus* does not occur in this period.

In an online article on January 28,⁷ however, the Language Council of Norway stated that the word is to be spelled with *k-*, thereby effectively normalizing the spelling for the first time and doing so in a way that went against the commonly practiced spelling. The present study is probably the first quantitative assessment of the effect of that normalization. After a brief period of fluctuation between spellings, the use of *k-* in a majority of cases was observed after the middle of February, as shown in Figure 2. However, after more than a year since the spelling change, there is no further convergence towards the new spelling. This seems due to the fact that not all newspapers adopted the newly normalized spelling. Figure 3 shows the variation per newspaper, revealing some clear discrepancies between newspapers as regards the choice between *c-* and *k-*. A final note on spelling is that among the 167957 tokens there were 18168 written with a hyphen, which is normally unnecessary, except to avoid the collision of two *as*, such as in *korona-app*, or in combinations with a number, such as *korona-17.mai* (“corona 17th of May”, Constitution Day in Norway).

3.2 Frequency, Variation and Productivity

The number of tokens per day is shown in Figure 4. The earliest occurrences of relevant compounds in the Norwegian Newspaper Corpus in 2020 were *coronavirus* (indef. sg.) and *coronaviruset* (def. sg.), on January 9, 2020. The use of these and other compounds remained modest for over a month, but on February 26, 2020, when the virus was detected in Norway, a marked increase can be seen. The maximum token count was 1765 on a single day.

The token counts are generally somewhat lower in the weekends when the volume of articles is lower. In that respect it might have been useful to count normalized frequencies on the basis of the volume of

⁷<https://www.sprakradet.no/Vi-og-vart/hva-skjer/Aktuelt-ord/koronavirus/>

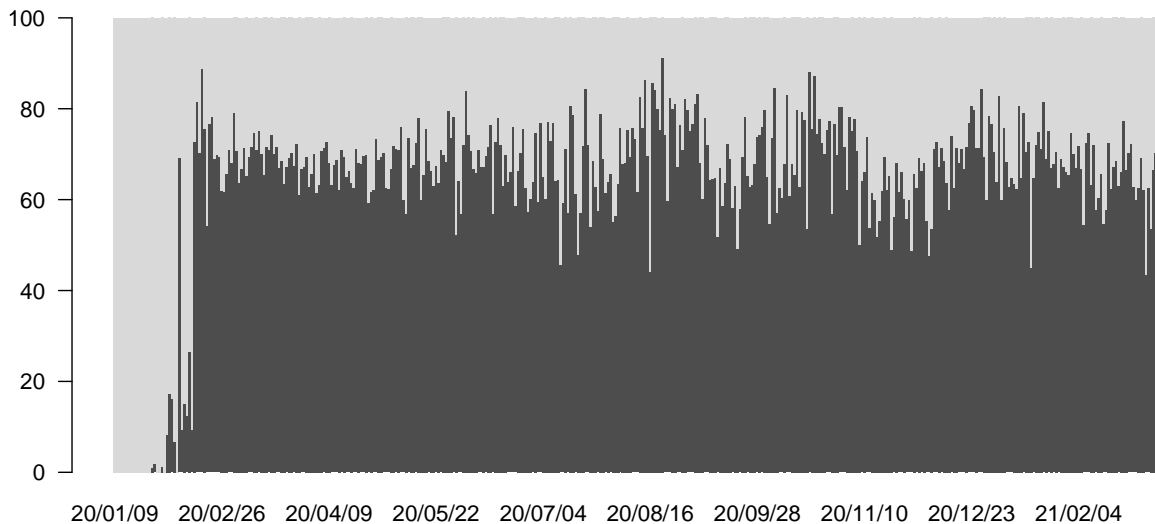


Figure 2. Distribution of *c-* (light) and *k-* (dark) over time. No bars are shown for days without any occurrences of either.

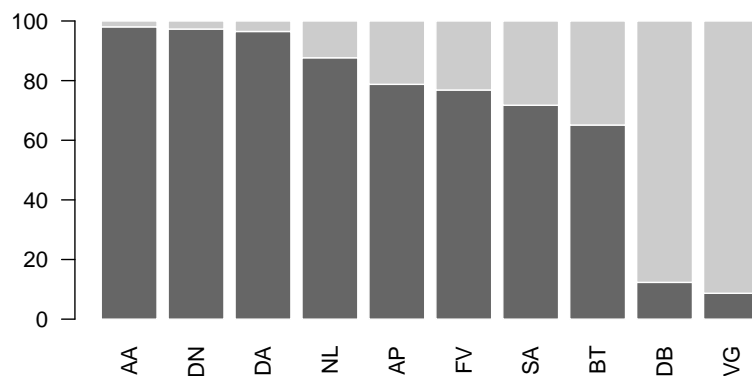


Figure 3. Distribution of *c-* (light) and *k-* (dark) per source; AA=Adresseavisen, AP=Aftenposten, BT=Bergens Tidende, DA=Dagsavisen, DB=Dagbladet, DN=Dagens Næringsliv, FV=Fædrelandsvennen, NL=Nordlys, SA=Stavanger Aftenblad and VG=Verdens Gang.

harvested words per day, but unfortunately the daily volumes are not provided by the corpus interface. Token counts in themselves are however not the primary focus of the present investigation.

Whereas the volume of tokens indicates how much is written about a topic in general, the breadth of the discussion in terms of subtopics may rather be revealed by looking at the number of distinct types. For this purpose, normalization was applied by deleting the first part of the compound so that the above-mentioned spelling variation and the possible use of a hyphen are disregarded. The remaining wordforms were lemmatized.⁸ Lemmatization resulted in a few errors, most of which were corrected with a manually constructed script. Furthermore, lemmatization was not entirely consistent, e.g., some deverbal adjectives were reduced to the verb lemma, whereas others were not. Also, some ambiguities may not have been correctly resolved, because the lemmatizer was run on a simple list of wordforms, which does not provide any helpful context, as compared to applying the lemmatizer to running text. A better solution would obviously be to lemmatize the whole corpus, but that was not a realistic option at the time of this study. Nevertheless, the lemmatizer output was in general useful and its minor imperfections do not seem to have distorted the general picture.

⁸Lemmatization was done by means of the model *nb_core_news_md-2.3.0* in *spaCy*, https://github.com/explosion/spacy-models/releases/tag/nb_core_news_md-2.3.0, details at https://spacy.io/models/nb#nb_core_news_md.

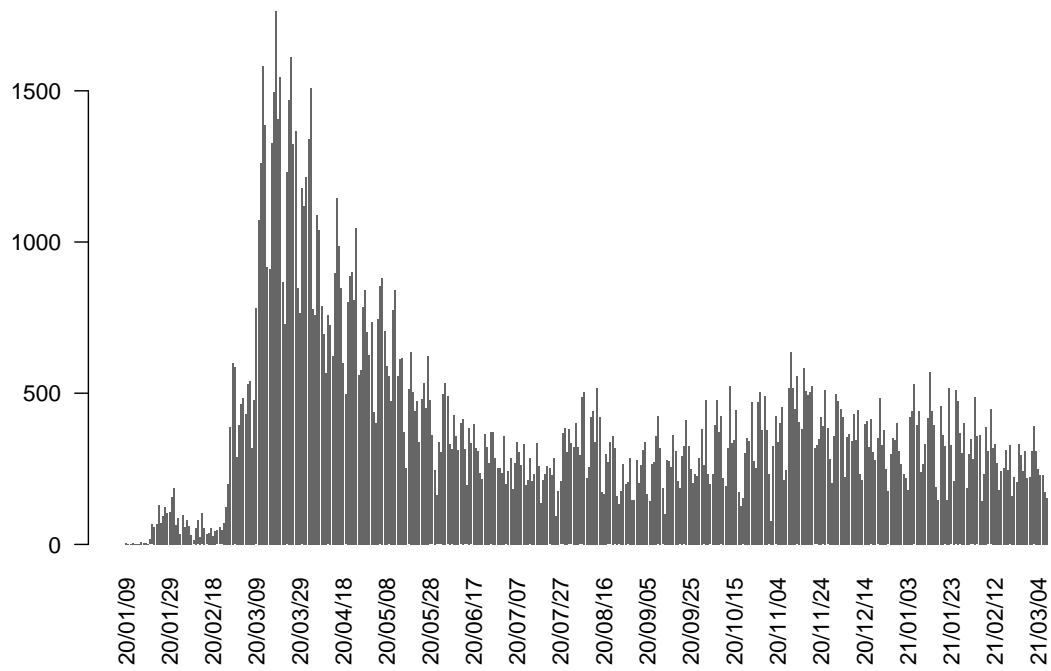


Figure 4. Number of occurrences observed per day.

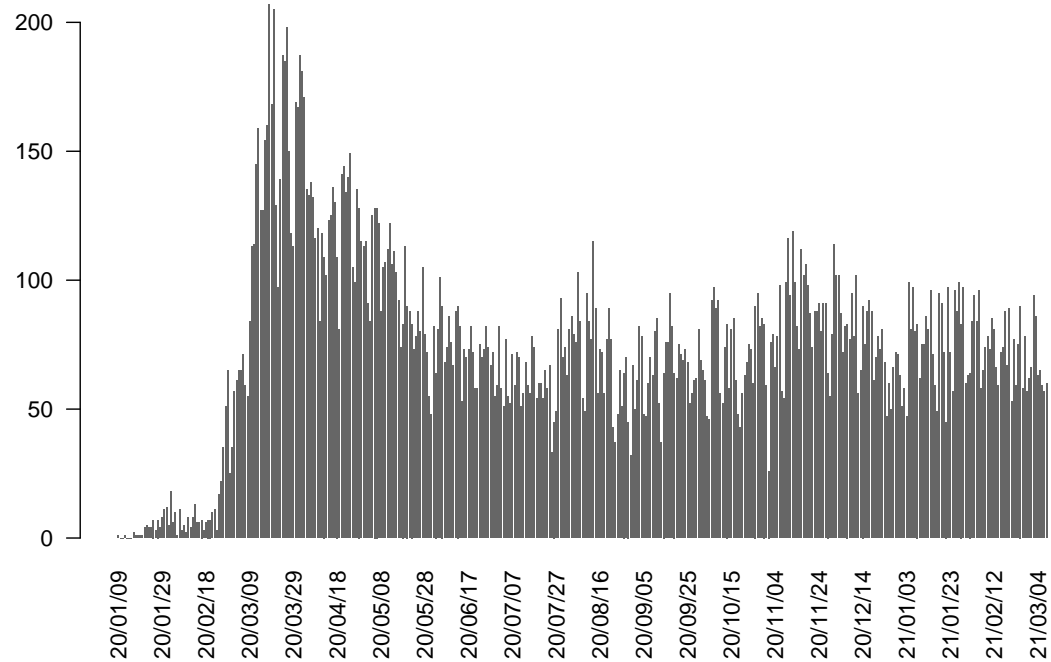


Figure 5. Number of types (lemmatized) observed per day.

Normalization and lemmatization meant, for instance, that the word forms *koronatiltakene* (“the corona measures”) and *Corona-tiltak* (“corona measure(s)”) were both reduced to the same lemma *tiltak*. Alternative spellings of lemmas such as *oppmyking* and *oppmykning* (“softening”) remain however separate items. In the end, the original 167957 tokens, consisting of 3012 distinct word forms, were reduced to 2133 lemma types. A frequency list was made of all the types, showing a typical Zipf distribution.⁹ Compounds containing *virus* make up close to half of the total number of tokens.

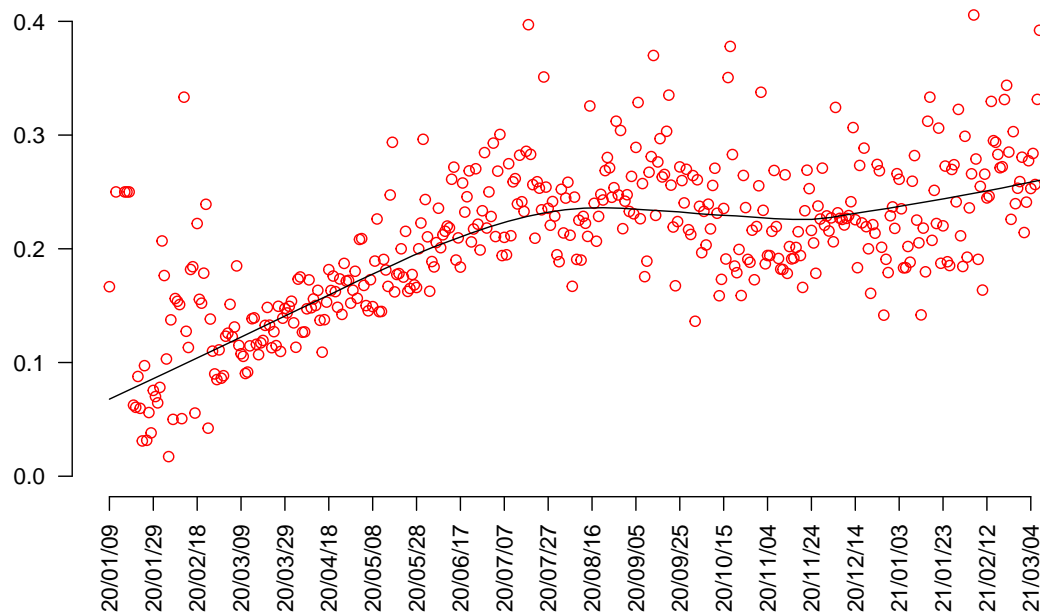


Figure 6. Variation (types / tokens) per day.

The type count per day, as shown in Figure 5, at first sight seems to roughly follow the increase in the token count. However, as Figure 6 shows with a trend line (fitted with local polynomial regression), the ratio of types to tokens per day is not constant, but increasing, with some flattening from the late summer of 2020. Initially the ratio was around 0.1, which means that every word was used on average ten times per day. Near the end of the studied period, the ratio had risen to around 0.25, which means that every word was used on average only about four times per day. This evolution suggests that the variation in subtopics in the discourse not only increased markedly during the initial few months of the pandemic, but also remained high until the end of the observation period.

Another measurement is the number of *new* words per day, i.e. types which had not been recorded on earlier dates during this period (and not even before 2020, for practically all the words). A list was made of all types, with their first date of occurrence and the first newspaper in which they were observed. Cumulative counts of these new types, per day, shown in Figure 7, show the speed of the vocabulary growth. In January and February 2020, the number of new compounds increased very slowly, but a sharp acceleration can be observed around February 26, 2020, when the virus had reached Norway. This steep increase continues throughout March 2020, before it flattens out slightly in April and a bit more in May, but after May, the vocabulary growth remains remarkably strong and linear until the end of the period with observations. This can be seen as an indication that the discourse needed more and more descriptive words as the effects of the pandemic continued to affect more and more aspects of our society.

As expected, most compounds were nouns, e.g. *koronapsyken* (“the corona psyche”), some were verbs, e.g. *koronastenge* (“close down due to corona”), some were adjectivally used participles, e.g. *korona-stanset* (“stopped by corona”) and some were adjectives, notably including *koronafast* (“stuck due to

⁹A frequency list and a list of items by date of first occurrence can be accessed at <https://github.com/clarino/corona>. In these lists, types are reduced to their final parts.

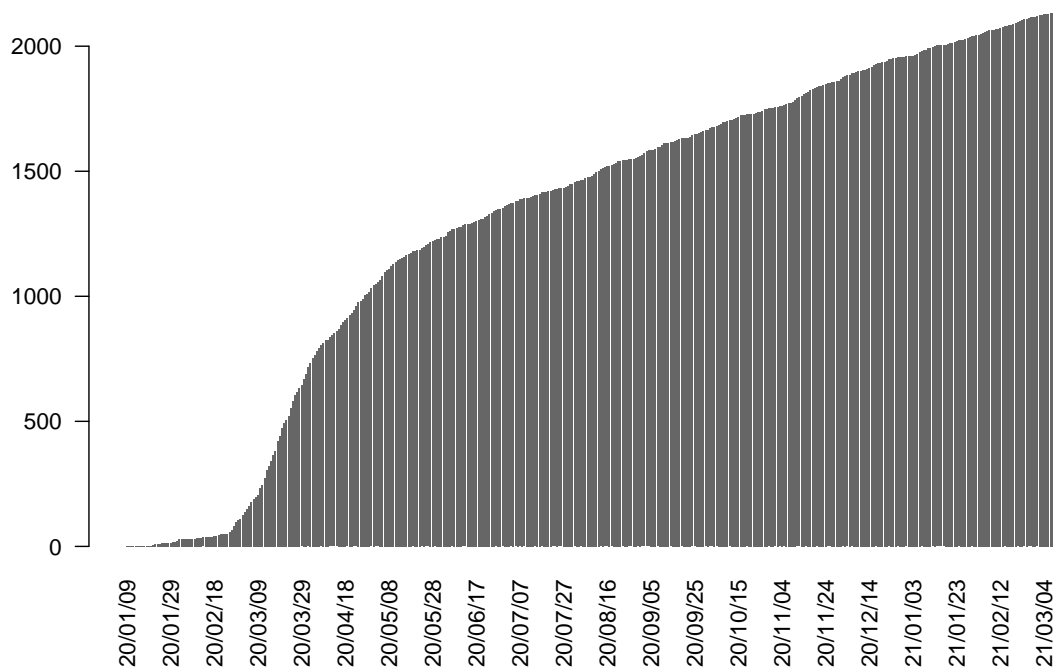


Figure 7. Cumulative increase of the *corona* compound vocabulary.

corona”), the latter modeled after the existing *værfast* (“stuck due to bad weather”) and the relative newcomer *askefast* (“stuck due to the ash cloud”) in 2010 (De Smedt, 2012).

Most of the new compounds that appeared in the current study are not quite transparent. Indeed, before 2020 it would have been difficult to interpret, for instance, *korona-telt* (“corona tent”), *koronautsettelsene* (“the corona postponements”), *coronalov* (“the corona law”) and *corona-kompensasjon* (“corona compensation”). Taken as a whole, the semantic contribution of *corona/korona* in such compounds is a broad contextualization of the meaning and can be paraphrased as “related to the virus, the disease, the epidemic, or the measures to combat all of these.” Several of the compounds are metaphorical and have emotional connotations, such as the final parts *knekken* (“the breakdown”), *knipen* (“the pinch”), *spøkelset* (“the ghost”), *tsunamien* (“the tsunami”) and *tabu* (“taboo”).

4 Discussion

This paper presents a timely use case demonstrating the potential of a newspaper monitor corpus, i.e. a corpus which is regularly updated with fresh newspaper articles, for the purpose of tracing changes to the language in almost real time. In particular, it tracks and analyzes new compounds with *corona/korona* in the Norwegian Newspaper Corpus. An earlier study (in Norwegian) with a similar objective and method covered a period of only 139 days, ending on May 26, 2020 (De Smedt, 2020). In contrast, the present study represents a considerable extension in time, as it covers 425 days, i.e. a period more than three times as long, stretching until March 8, 2021, when the current data were collected.

It was found that the majority spelling had changed in the course of about one month, although further convergence on the new standard, which could have been expected, is not borne out by the data. The huge number of occurrences (tokens) of compounds with *corona/korona* in the studied period may not be entirely surprising, but the dynamics of vocabulary growth and diversity are noteworthy. The sharp acceleration in the creation of compounds from about February 26, 2020 started to slow in April and May 2020; if that slowing had continued, the vocabulary might have flattened out at around 1500 words. Instead, the vocabulary growth continued in a surprisingly linear climb, adding at least 600 more words from June 2020. Another noteworthy result was that the variation of compounds in use, measured as the ratio of types vs. tokens per day, not only increased in the initial phase, as already observed by De Smedt (2020), but remained high during the entire period. These dynamics are certainly driven by the continuing

need to report on a widening range of situations and events that were consequences of the pandemic, but perhaps just as much by journalists' willingness to continue exploiting the salience of a new word and bombard their readership with attention-drawing compounds at an average pace of about five new ones per day.

The present findings show similarities and differences with a previous study on compounds with *aske* ("ash") following the volcanic eruption in Iceland in 2010 (De Smedt, 2012). That study found a sharp increase in variation but it was less broad in scope and it flattened out after half a month. In comparison, the presently reported increase in variation did not rise as quickly, but accelerated after about a month. Also, the creation of new compounds did not decrease quickly, but was sustained until the end of the studied period. A possible explanation for these differences might be that the effects of the current pandemic are not only lasting longer, but are also having more widespread and long-lasting effects on society. Nevertheless, both in the earlier study of *aske* and the current study of *corona/korona*, creative compounding seems to be contagious among journalists who appear intent on outdoing each other with ever more creative neology.

There are other systems for tracking news or tracing new words. Among news trackers, the European Media Monitor¹⁰ has a long-standing reputation. However, its interface is oriented towards topic tracking and alerts rather than detecting neologisms. The use of the Norwegian newspaper archive Atekst Retriever¹¹, based on daily harvesting from even more media sources than the Norwegian Newspaper Corpus, was briefly considered. However, Retriever is less suitable for linguistic research, as it returns webpages with summaries, from which it is more difficult and less reliable to extract keywords and relevant information (i.e. at least the date and source). Furthermore, the earliest mention of a relevant compound with *corona/korona* in Retriever was December 28, 2019, which turned out to be a mistake in dating: in reality it was an article from February 28, 2020. It must be added that also the Norwegian Newspaper Archive had some problems with dates, so that two occurrences dated January 2, 2020 were considered unreliable and were removed by the script. More generally, such issues show that although correct dating is paramount in studies like these, there is always a risk that errors go under the radar; for a related discussion of hidden dangers in digitized text, see Nunberg (2009).

For the detection of neologisms as such, several systems are useful in their own right, such as The Word Spy¹² for English and Die Wortwarte¹³ for German, the latter developed in the context of CLARIN-D and using monitor corpora. However, these sites seem to offer neither regular expression search, nor output as a complete list of observations tagged with sources and dates. For those reasons, they are less suitable for the kind of data aggregation and analysis of compounds on a time line as presented here. In fact, the Norwegian Newspaper Corpus also features a separate automated system which every day identifies new words that have been added to the corpus. However, the goal of the present study is not so much to spot new words, but rather to trace both the creation and the protracted use of compounds based on a specific stem through a given period.

The dynamics exposed in this study may, on the one hand, serve to illustrate collective journalistic practice and contagious tendencies towards salient and eye-catching terms. On the other hand, the dynamics may also provide clues as to how fast, how broadly and for how long salient events are affecting our society, whether an event may initially have been underestimated, and so on. This kind of information may in turn be used in applications, such as the automatic detection of significant "bursts" of words in information streams (Kleinberg, 2002, e.g.), which, in cases like the current one, would benefit from compound analysis.

Further studies, including investigations of *corona* creativity across languages, might be interesting. However, despite the advantages of the availability of many newspaper corpora through CLARIN, the above-mentioned CLARIN resource family overview of newspaper corpora shows a lack of up-to-date monitor corpora. Almost all of the newspaper corpora in the CLARIN list consist of fairly dated materials and their different periods do not always overlap. Furthermore, the corpora are not easily interoperable;

¹⁰<https://emm.newsbrief.eu>

¹¹<https://web.retriever-info.com/services/archive>, accessed March 17, 2020

¹²<https://www.wordspy.com>

¹³<https://wortwarte.de>

they do not share the same annotation and formatting, and they are not searchable through the same interface. Addressing these issues might call for a kind of multilingual Federated Content Search on this CLARIN resource family, and by the compilation of more monitor corpora that allow the study of vocabulary linked to current events.

A final remark is that the Norwegian Newspaper Corpus may not be sustainable in its current form. Current agreements with the newspaper publishers allow scraping of the public newspaper websites, but more and more articles are being hidden behind paywalls. Over the years, the annual corpus accrual of 100 million words for the ten major newspapers has sunk to about 50. Some newspapers that provided materials in an early phase of corpus collection have no open articles at all now. Furthermore, changing webpage formats affect the quality of materials obtained by scraping. Clearly, it would be better to obtain complete newspapers, appropriately encoded and licensed, directly from the publisher. This requires new agreements, which are currently being discussed with Norwegian media corporations, through cooperation with The Norwegian Language Bank at the National Library of Norway, with the aim of building a new Norwegian Media Corpus.

Acknowledgements

Thanks to Knut Hofland for implementing and maintaining the Norwegian Newspaper Corpus and to Paul Meurer for importing the corpus in Corpuscle. Thanks to Mikkel Ekeland Paulsen, Carina Nilstun, Margunn Rauset, Victoria Rosén, Sturla Berg-Olsen and anonymous reviewers for information and comments that were helpful in preparing this paper.

References

- Andersen, G. 2012. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia.
- Andersen, G. and Hofland, K. 2012. Building a Large Corpus Based on Newspapers from the Web. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia, 1–28.
- De Smedt, K. 2012. Ash Compound Frenzy: A Case Study in the Norwegian Newspaper Corpus. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia, 241–255.
- De Smedt, K. 2020. Smittsomme Koronaord. *Oslo Studies in Language* 11(2):59–73.
- De Smedt, K., Samdal, G. I. L., Kyrkjebø, R., Al Ruwehy, H. A. H., Gjesdal, Ø. L., Rosén, V., and Meurer, P. 2016. The CLARINO Bergen Centre: Development and Deployment. *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*. Linköping Electronic Conference Proceedings. Linköping University Electronic Press, 1–12.
- Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zsu, X., Zloch, M., and Dietze, S. 2020. TweetsCOVID-19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic [Preprint]. 29th ACM International Conference on Information & Knowledge Management (CIKM2020), Resource Track. Association for Computing Machinery.
- Evert, S. and Hardie, A. 2011. Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.
- Kleinberg, J. 2002. Bursty and Hierarchical Structure in Streams. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101.
- Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Xin Ru Wang, N., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. 2020. CORD-19: The Covid-19 Open Research Dataset. arXiv: 2004.10706.

- Meurer, P. 2012. Corpuscle – a New Corpus Management Platform for Annotated Corpora. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. *Studies in Corpus Linguistics* 49. John Benjamins, Amsterdam/Philadelphia, 31–49.
- Meurer, P. 2020. Designing Efficient Algorithms for Querying Large Corpora. *Oslo Studies in Language* 11(2):283–302.
- Nunberg, G. Aug. 31, 2009. Google’s Book Search: A Disaster for Scholars. *The Chronicle of Higher Education*.
- OED Editorial. Apr. 15, 2020. *Corpus Analysis of the Language of Covid-19*. OED blog. URL: <https://public.oed.com/blog/corpus-analysis-of-the-language-of-covid-19/> (visited on 09/14/2020).
- Paton, B. Apr. 9, 2020. *Social Change and Linguistic Change: The Language of Covid-19*. OED blog. URL: <https://public.oed.com/blog/the-language-of-covid-19/> (visited on 09/14/2020).

Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources

Hanna Hedeland

Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Abstract

Though digital infrastructures such as CLARIN have been successfully established and now provide large collections of digital resources, the lack of widely accepted standards for data quality and documentation still makes re-use of research data a difficult endeavour, especially for more complex resource types. The article gives a detailed overview over relevant characteristics of audiovisual annotated language resources and reviews possible approaches to data quality in terms of their suitability for the current context. Conclusively, various strategies are suggested in order to arrive at comprehensive and adequate definitions of data quality for this specific resource type and possibly for digital language resources in general.

1 Introduction

The successful development of international digital research infrastructures such as CLARIN has enabled the sharing and re-use of language resources across geographic and, partly, disciplinary boundaries. This has led to a shift in focus from the technical means of data sharing towards the data itself and in particular its quality and fitness for re-use. However, while e.g. in Germany, the German Council for Scientific Information Infrastructures (RfII) states in the latest of its published recommendations that "securing and improving data quality is a fundamental value of good scientific practice" (RfII, 2020), widely acknowledged and adequate definitions of data quality for the various types of language resources provided through digital infrastructures are yet to be defined. Generic approaches such as the FAIR Principles (Wilkinson and others, 2016) or even the FAIR Metrics (Wilkinson et al., 2018) and similar approaches based on a comprehensive assessment and evaluation of the FAIR Principles do not provide detailed guidance for research data management for specific resource types or research methods related to specific disciplines. For example, regarding Reusability, the FAIR Metrics, as the FAIR Principles, only refer to resources that "meet community standards", the FAIRsFAIR Data Object Assessment Metrics (Devaraju et al., 2020) to standards and formats "recommended by the target research community" and the Data Maturity Indicators of the FAIR Data Maturity Model WG (RDA FAIR Data Maturity Model Working Group, 2020) to data that "complies with a (machine-understandable) community standard". It is not possible to formulate more specific criteria for the data within these generic metrics, this task is delegated to the communities and those providing research data management services for them. For archives or research data centres aiming to make the deposited and hosted data FAIR, the existing frameworks could be used to design corresponding metrics or indicators for specific communities or resource types. Focusing on the resource type rather than the communities might be advisable, since as Bahim et al. (2021, p. 5) report after surveying "communities" on topics related to FAIR data assessment, "[d]espite being widely mentioned in the RDA context, the term "community" remains unclear. The respondents are still questioning who these communities are and who are the stakeholders constituting them."

Research data quality calls for adequate and comprehensive definitions, but this raises several – often overlooked – fundamental questions. Suitable quality criteria need to be transparent and operationalized, but also reflect the complexity of the subject matter, in our case: audiovisual annotated language data. A

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

first step is therefore a review of this resource type, before various approaches to defining data quality criteria can be evaluated in terms of their applicability.

2 Taking Stock of Audiovisual Annotated Language Data Resources

The various resource types subsumed under "audiovisual annotated language resources" are highly heterogeneous but have in common that they comprise several data types and display a complex structure of abstract entities and data objects with different types of relations. A comprehensive description of these resources and the variation within the group is therefore an important first step. In figure 1, relevant data types and processing stages are pictured in detail in an approach similar to that proposed by Himmelmann (2012). For a specific data type, each stage or higher level can be based on all available data on the level(s) below, the most relevant input is however placed directly beneath it. Data types in filled dashed boxes are usually superseded by their counterparts on the next level and only retained for archival or reproducibility purposes. Data types in transparent boxes represent further processing and analysis stages based (primarily) on non-audiovisual sources, i.e. images or textual data. These data types are usually not created for audiovisual annotated language corpora, which rather only include annotations of audiovisual sources, i.e. recordings. In other types of research data within the humanities and social sciences, the data model expects coding to be performed on several sources of various modalities. While data models for these different types of annotations all include references to the prepared sources, the previous processing steps are only rarely recorded as (provenance) metadata. On each level, the notion of data quality has different implications, which takes the relevant research activities into account.

A better understanding and description of this resource type is one goal of the QUEST¹ project, which was based on the existing cooperation of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation² – including DCH/IfL (Cologne), ELAR/SWLI (London), HZSK/INEL (Hamburg) and ZAS (Berlin) – and extended by the German Sign Language Corpus project (Hamburg) and the Archive for Spoken German at IDS (Mannheim), adding their complementary expertise on sign language data and German data, respectively. ELAR and the Cologne Language Archive (LAC) allow self-deposit of resources with basic requirements on file formats and metadata, whereas the AGD and (previously) the HZSK curate resources to comply with data models and data consistency requirements. The data deposited with the AGD and the HZSK is mainly from projects working with qualitative methods only, for which the requirements regarding data modelling and consistency play a subordinate role. This is also reflected in the data to a varying extent. The resources in all four centres differ along several dimensions, which can be described as structural, methodological and content-based heterogeneity.

2.1 Structural Heterogeneity

Abstract data models for language resources such as EXMARaLDA (Schmidt and Wörner, 2014) or the DGD data model (Schmidt et al., 2013a) provide explicit requirements not only on the overall resource structure of abstract entities and data objects, but also regarding the resource content, i.e. consistency in tier structure and the conventions and schemes used for transcription and annotation and the identity of speakers. Even without an explicit data model, the resource structure is also defined by contextual data, including structurally relevant entities such as recording sessions, and by metadata on included files and their relations. However, not all resources found in archives today exploit such elaborate models or metadata formats. Some are limited to a set of audio and video recordings and individual transcripts, in some cases with no explicit information on the internal structure and only minimal metadata.

When it comes to the explicit or implicit data models describing collections of audiovisual linguistic research data, there are several differences as outlined in Figure 2, which shows various data models and metadata standards. In the figure, filled boxes represent the modelling of real world entities, e.g. participants, as entities, i.e. an independent object being referred to via a unique ID. Transparent boxes on the other hand represent complex or simple data types, which also have IDs, but these are not used for reference to avoid redundant listing of e.g. a participant with all its information for each session. The

¹<https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html>

²<http://ckld.uni-koeln.de/>

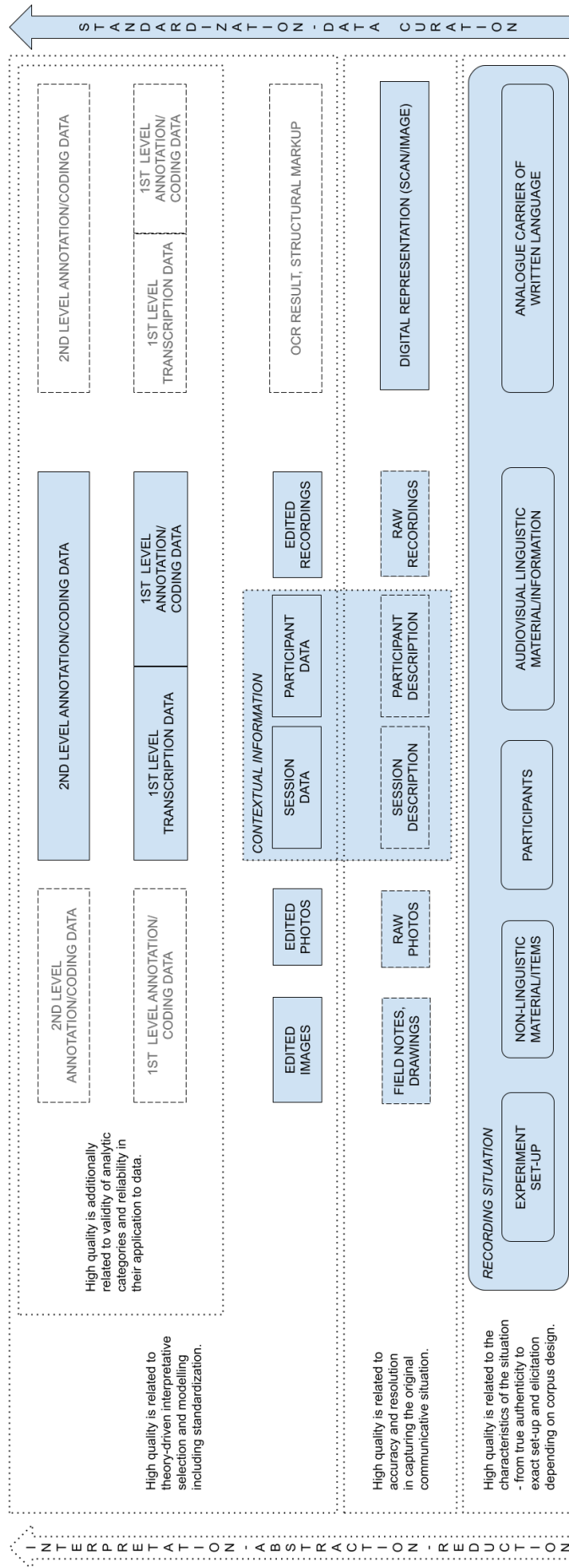


Figure 1: Data of various types are derived from existing data of a research project through various research activities during data creation and evaluation.

way participants are modelled is an example of which kind of information can be specified and how consistency can be controlled. The EXMARaLDA Corpus Manager models communications and Speakers as independent entities, and can therefore only model globally valid participant information. The IMDI and TEI formats, on the other hand, only model session specific participant information as participants do not exist independently of sessions. This allows for inclusion of relevant session specific information but makes data consistency more challenging. While participants can be listed in the `<teiHeader>` of the `<teiCorpus>` for TEI, there is no designated mechanism of referring to these instead of declaring them on the document level. The data model of the DGD also views Speakers as entities and can relate Speakers to Speech events, but the model is also capable of accommodating session specific information (such as the participant’s role in the interaction or age at the date of recording) through the ”Speaker in Speech Event” relation. Another problematic aspect for interoperability across resources is the level(s)

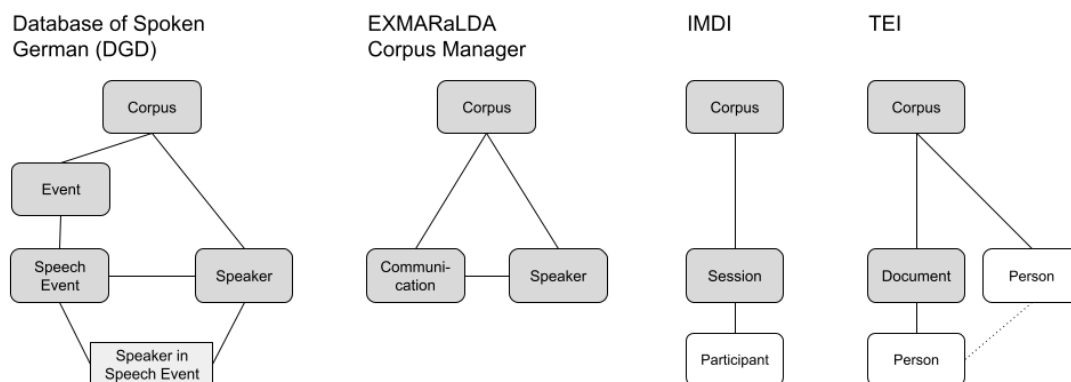


Figure 2: Various data models and metadata formats make explicit and implicit statements regarding resource structure.

below the top resource level. If we disregard corpus design specific sublevels, e.g. ”L2 Speakers” vs. ”L1 Speakers” we still find differences and in many cases ambiguities regarding the meaning of structural units such as sessions. Basically, there are three relevant dimensions in this case: The time-based *recording session*, specific linguistically determined *communicative events*, possibly contained in a common recording session, and the *partition based on storage media*. While structuring a collection according to storage media (e.g. unit ”Tape 005, side A”) will most likely never be intended as a design choice, it is often encountered in legacy data, where analogue carriers have been digitized but not yet further processed and described. Due to the existing heterogeneity, it is not possible to decide on a certain recommended alternative, but the information should be included in the metadata. This also applies for cases where for technical reasons (e.g. handling of large files) communicative events have been arbitrarily split into separate units.

Below the session level the heterogeneity is even greater, as the file formats used to encode transcription and annotation data vary quite a lot. In particular, they are either unstructured, i.e. provided in some (plain) text format, or (semi-)structured, as most XML transcription/annotation tool formats. Such structured formats show differences both in the macro-structure of tiers and speaker contributions and annotations and the micro structure related to annotation schemes and transcriptions conventions as explained in detail in Schmidt (2011). A comprehensive discussion is beyond the scope of this paper.

2.2 Methodological Heterogeneity

Differences on the transcript micro-level (Schmidt, 2011) depend on the research methods employed, especially in terms of qualitative or quantitative approaches. Annotations thus range from comprehensively applied systematic tags to selectively applied free comments. Since transcription conventions select and foreground certain aspects of language (Ochs, 1979), they also differ regarding units such as utterances or intonation phrases and the amount of linguistic information integrated into the basic transcription. Furthermore, not all transcription and annotation schemes lend themselves to automatic checking.

2.3 Content-related Heterogeneity

The content-related resource design plays a major role when it comes to visible differences due to choices regarding geographical and temporal coverage, and the selection of participants, topics, (multi)linguality types etc. for the data collection. Furthermore the amount and categories of contextual data describing recording sessions and participants also differ accordingly. The importance of complementary data types beyond recordings, annotations and contextual data, such as written or image material present in the recording situation also depend on the research question and resource design, i.e. the content.

3 Approaches to Data Quality and Possible Applicability for Language Resources

Since audiovisual annotated language resources are research data, which is a specific type of data in general, more generic approaches to data quality can provide valuable insights. These are therefore reviewed while evaluating the need to complement them with further more specific criteria.

3.1 Generic Approaches to Data Quality

Generic approaches do not restrict the types of data they are applicable to and thus recommendations remain general and abstract. (Wang and Strong, 1996) distinguish fundamental dimensions: intrinsic, contextual, representational and accessibility data quality, pertaining to the data itself, a particular usage context, and the systems providing data, respectively. This distinction between inherent and system-dependent data quality is also reflected in ISO/IEC 25012 - The Data Quality Model³. The W3C provide relevant input in their Best Practices for Data on the Web⁴, both regarding the recommendations and the system used to disseminate them. However, these generic approaches do not provide directly applicable resource specific recommendations.

3.2 Approaches to Research Data Quality

Today, the main requirement for research data is to be FAIR. In terms of the generic data quality dimensions, not all of the FAIR principles and corresponding metrics or criteria are directly related to intrinsic characteristics of the data, but also to the infrastructure required to make data findable and accessible, and thus not under the control of the data creators. As outlined above, approaches aiming to operationalize the well-known principles also only refer to community-specific standards. The FAIRification process (Jacobsen et al., 2020) also still needs resource type specific requirements and workflows, but is a starting point to redefine data curation processes in line with FAIR concepts.

The concept of Data Maturity on the other hand, seems to be a suitable way of avoiding the word "quality" altogether for dimensions related to data structuredness and machine-understandability, which might not be considered to be closely related to data quality at all by some humanities scholars. An important aspect beyond the scope of this paper, but which must be considered at all times, is the quality of research data as an artefact of research. The artefact can only be as good as the research (and vice versa). While this quality aspect can only be assessed with thorough documentation of the data and its provenance, relevant theoretic frameworks, research methods and analytic categories used, for metadata and data to be machine-readable or even machine-understandable is not a relevant requirement.

3.3 Resource Type Specific Approaches to Data Quality

Within CLARIN, there is work in progress to collect recommendations from all CLARIN B centres on standards and formats accepted for deposit⁵. Apart from the participants of the QUEST project, some centres providing detailed recommendations for audiovisual data are e.g. The Language Archive at the MPI in Nijmegen⁶ and the Bavarian Archive for Speech Signals⁷. Furthermore, the German funder DFG

³<https://www.iso.org/standard/35736.html>

⁴<https://www.w3.org/TR/dwbp/>

⁵<https://www.clarin.eu/content/standards-and-formats>

⁶<https://archive.mpi.nl/tla/accepted-file-formats>

⁷<https://www.phonetik.uni-muenchen.de/Bas/BasInfoStandardsTemplateseng.html>

has published recommendations for technical standards⁸ collected through discussions within the relevant research communities. And still highly relevant after almost twenty years, (Bird and Simons, 2003) have described several aspects relevant for the long-time preservation and re-use of language documentation data. These recommendations are valuable resources for comprehensive definitions of resource type specific data quality.

4 Step One: Defining Data Maturity Levels for Audiovisual (Annotated) Language Resources

Considering the heterogeneity, it would be inappropriate to measure quality without regarding the conscious choices and trade-offs made by researchers that affect the machine-readability and consistency of the data, i.e. aspects of data maturity that might not be of direct value considering the chosen research method. The AGD and the HZSK have defined guidelines for deciding whether to perform data curation (Schmidt et al., 2013b). While curation of deposited data increases the re-use potential, the increasing deposit numbers without corresponding growth of capacities must necessarily run into a dead end. By allowing controlled variation, re-users know what to expect from language resources provided by others and more adequate goals for evaluation and curation can be defined. While the focus here is on audiovisual data, the following categories could also be applied to written language resources. The exact set of criteria for each level is still work in progress, as is the common system of quality and curation criteria for the QUEST project, but for now, three examples of criteria that have to be met are included for each level below. Some of these can be tested automatically, while others require manual (human) evaluation. The category names refer to prototypical resource subtypes and are not an integral part of the framework.

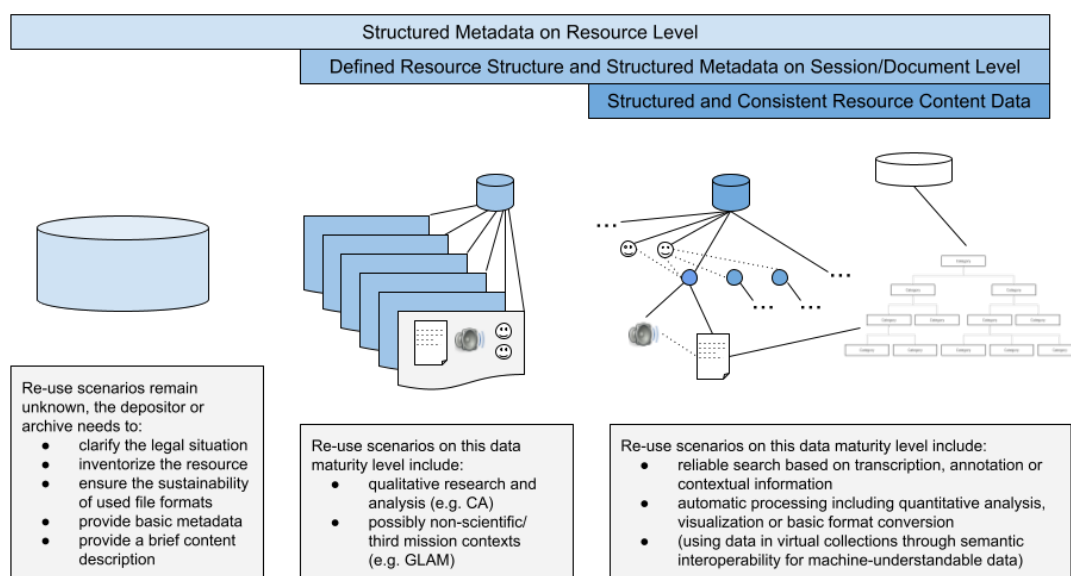


Figure 3: The three levels of data maturity developed for audiovisual annotated language resources allow for adequate assessment of data quality and realistic data curation plans.

4.1 Deposits

Since research data centres will always be confronted with orphaned legacy data, there have to be minimal requirements for data which is by no means FAIR, but still, especially in the case of endangered languages or oral history data, is undoubtedly worth being archived. A deposit is thus a data set with minimal metadata clarifying the legal situation and providing basic information on the content.

⁸https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

- The licensing information provided is sufficient to decide which re-use is possible.
- There is a complete inventory describing all files included in the resource.
- All files come in well-documented formats that can be read by non-proprietary software.

4.2 Collections

On the next level, Collections comply with additional requirements on the resource structure level, including the completeness and consistency of metadata and relations between all resource parts. The requirement on completeness of metadata only pertain to standardized cataloguing metadata describing the linguistic resource and its provenance to make the data discoverable and comprehensible. Contextual data, e.g. information on participants, on the other hand can not be standardized, since this would interfere with research design. The textual language data of Collections is usually provided in various unstructured text formats suitable for human manual analysis, but there are no requirements on the completeness or even existence of transcripts.

- There is basic metadata for individual recording sessions including participants.
- The deposited materials share thematic characteristics.
- All files come in well-documented formats that can be read by non-proprietary software.

4.3 Corpora

Corpora fulfill all requirements of Collections and are additionally structured and consistent on the resource content level, i.e. in the use of tier structure, annotation schemes and transcription conventions, but also regarding contextual data such as participant identities across the resource. While the Corpus data is machine-readable and suitable for reliable automatic analysis, definitions of for instance tier content or annotation schemes are often not machine-understandable and interoperability is thus generally limited to syntactic aspects.

- Tier (annotation level) definitions are consistent across the resource.
- (Relevant) participants can be identified across the resource.
- There is a clear design principle for the resource and its parts.

5 Step Two: Data Curation as FAIRfication

Since important aspects of research data quality are reflected by the FAIR principles and various metrics based on them, data curation can also be considered FAIRification, the process resulting in FAIR data. In this process, beyond syntactical correctness, the semantic information needs to be made explicit; we need to "define the semantic model". The differences in the level of data maturity described above are relevant for this process, since for Deposits and Collections, which do not have structured transcription/annotation data, machine-readable definitions enabling the data to become machine-understandable and further semantic enrichment and linked open data features are only possible for metadata. For Corpora on the other hand, the structured data allows for the semantic model to be defined more fine-grained, by linking tiers and annotations to controlled vocabularies and ontologies, but this option has rarely been used, e.g. the option to reference ISO Data Categories available in ELAN (Sloetjes, 2014) seems to play no role in the ELAN annotation data (EAF) currently found in archives (von Prince and Nordhoff, 2020).

As described in Schmidt (2011), the ISO standard for Transcription of Spoken Language⁹ provides more semantic information on units and information types as part of the underlying data model than most widely used formats for transcription/annotation data, which do not define the notation of e.g. participants' contributions, noise or pauses. As the standard was developed with this idea in mind, conversion into this format (Schmidt et al., 2017) would be one step towards semantic interoperability, while still not enforcing any semantics for the theory-dependent micro-structure. This is an important aspect,

⁹<https://www.iso.org/standard/37338.html>

since standardization efforts based on defining the micro-structure, such as the CHAT format, are bound to be restricted to the specific theoretical frameworks and usage scenarios they reflect.

Although the benefits of this approach and the required next step of providing reliable parsers for various micro-structures were outlined already in Schmidt (2011), almost ten years later, despite undeniable progress in digital research data management for audiovisual language resources, some basic modules and functionality are still lacking beyond conversion of widely used tool formats into the ISO/TEI format for this type of data to become truly FAIR. For this, agreement is necessary on how to describe both the information types in various annotation levels and tiers, and the rules for parsing the micro-structure in a machine-understandable way. Though alternatives such as OLiA¹⁰ exist, there is still no designated widely used method to include machine-readable references for tiers or individual annotations in TEI-based formats. Additional conventions would allow for a proper definition of the semantics of individual data sets and increase the options for re-use, especially for automatic processing and enrichment. Another aspect purposefully not treated in depth in the ISO/TEI approach is semantic interoperability on the level of the resource, or even means of encoding basic contextual data in a standardized manner. The generic TEI standard was not primarily developed for spoken corpora and the TEI Corpus is a set of documents. It would be possible to simply include relevant parts of existing formats¹¹, but these are also not interoperable, even though, as shown in 2, they share basic characteristics in the same way as transcription data does, simply because these are models of a common reality.

6 Step Three: Adding the "Fit for Purpose" Dimension

While the aspects of FAIRification (from data structuredness to semantic enrichment and linking) are generic, data quality is to a great extent a question of the data being fit for particular purposes or usage scenarios – and not all usage scenarios improve directly by using more structured data, i.e. a higher level of data maturity. Since it is not feasible for research projects creating language resources to consider all possible re-use scenarios, explicit and formalized definitions of re-use scenarios would allow projects to comply with specific re-use scenarios. Re-users would also be able to quickly judge whether the data is suitable for their purpose, which is often difficult to tell today, especially in the case of interdisciplinary re-use, e.g. between linguistics and education sciences, partly also due to the use of different terminology.

The definition and implementation of criteria for such interdisciplinary re-use scenarios are further important goals of the QUEST project, complementing the technical and intrinsic aspects of data quality. Within the QUEST project, four main re-use scenarios are being investigated and systematically described on various levels ranging from the general legal situation to the interoperability with specific data formats and the use of certain annotation schemes or transcription conventions. For example, to enable re-use of research data from linguistic research projects within third mission contexts, e.g. as audiovisual augmentation in museums, the legal situation must allow (parts of) the data to be made available to the public, and specific linguistic information will have to be removed from transcripts to make them readable to laymen.

When considering the data maturity levels of audiovisual resources described above, it also becomes clear how they enable various forms of re-use. While audio files might be available in any case, reliable metadata on individual recording level, as required for Collections, is necessary to make a selection. With structural speaker assignment and alignment of the transcripts with the audio, more options to tailor the material become available and only structured data, as required for Corpora, can be automatically enriched in a reliable manner, converted, aggregated or visualized to suit the needs of the re-using institution.

Within the QUEST project, two re-use scenarios are used as pilot cases for semantic interoperability using the ISO/TEI standard. The first is the kind of qualitative data created within many areas of non-linguistic humanities research by so-called CAQDAS, computer assisted qualitative data analysis software. Though such resources bear resemblance to the resource pictured in Figure 1, and often include coding on images and textual content in addition to audiovisual sources, the transcript data of such

¹⁰<http://www.acoli.informatik.uni-frankfurt.de/resources/olia/>

¹¹E.g. by using the xenoData element of the teiHeader.

resources is rarely structured. This also applies for the new open standard in development, REFI-QDA¹², which will serve as an exchange format between widely used proprietary tools such as ATLAS.ti¹³ or NVivo¹⁴. Since this format is only targeted at plain text transcripts, a proof-of-concept conversion scenario for this type of data to the ISO/TEI standard was implemented based on a TEI format developed at the UK Data Service QualiBank¹⁵. TEI as a structured transcription format is allowed in their CAQDAS exchange format, QuDEX¹⁶ developed several years ago.

The second pilot case is the morphosyntactic description of interlinear glossed text in language documentation data. While there are several conventions in use, many are very similar and only have slight variation on the level of symbols used to denote various concepts. Within the INEL project and the HZSK, an extension for this type of data was developed (Arkhangelskiy et al., 2019), however, the TEI class `att.linguistic`¹⁷ (Bański et al., 2018) could possibly provide a simpler solution. This TEI class is used for token-based annotation in the DGD, the MTAS-based platform developed within the ZuMult project¹⁸ and also for written language, e.g. in the DTA Basis format¹⁹.

Both pilot cases share their relevance beyond communities merely interested in the linguistic features of the data, since they can often both be considered useful for oral history related research. The combination of TEI and RDF or linked data is on the one hand common but has on the other hand been solved in various ways, with a uniform solution based on RDFa currently being developed²⁰.

7 Discussion

Though seemingly trivial, fundamental questions regarding the structure and content of annotated audio-visual language resources created as research data within various disciplines have yet to be thoroughly discussed and answered. The characteristics of such resources need to be systematically described in order to define suitable criteria for data quality. When defining such criteria, we need to acknowledge that data maturity might seem irrelevant to many humanities researchers and in some cases really is²¹. The characteristics of the research data is the result of the research methods, hence a research data centre should be able to handle different data maturity levels while providing comprehensive examples of how increasing the level of data maturity has benefits for the individual researcher. As discussed, standardization can only apply to certain aspects of the data, the metadata and the documentation. Machine-understandable expression and documentation of the micro-structure would allow for conversion into the ISO/TEI format, which would increase re-use options and allow for the use of TEI compatible services and tools where this makes sense from the researcher's perspective. However, this is not a task for humanities researchers but requires the expertise of research data managers or data stewards in close cooperation with the research projects.

Providing generic quality criteria applicable to resources with various data maturity levels is one aim of the QUEST project, another is to provide additional criteria for formalized re-use scenarios. To allow data creators to comply with these criteria, the project will also provide software solutions to evaluate various types of resources according to such generic and specific criteria (Arkhangelskiy et al., 2020). This evaluation will ideally be performed continuously during data creation as a part of the project's data quality assurance methods. In combination, the definitions and evaluation mechanisms developed within

¹²<https://www.qdasoftware.org/about/>

¹³<https://atlasti.com/>

¹⁴<https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>

¹⁵<https://ukdataservice.ac.uk/get-data/explore-online/qualibank/qualibank.aspx>

¹⁶<https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/qudex/>

¹⁷<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.linguistic.html>

¹⁸<http://zumult.ids-mannheim.de>

¹⁹<https://www.deutschestextarchiv.de/doku/basisformat/>

²⁰<https://github.com/TEIC/TEI/issues/1860>

²¹Cf. RDA FAIR Data Maturity Model Working Group (2020, p. 10) "[D]ata coming from humanities fields, especially from outside of Digital Humanities, will often not be expressed in a machine understandable knowledge representation (RDF, SKOS or LOD) by nature but instead, it is often expressed in natural language, even if encoded using machine readable methods (e.g. TEI). Therefore, it becomes quite clear that the indicator treating machine-understandable knowledge representation will be less relevant according to the humanities."

the QUEST project will hopefully make data depositing and re-use more transparent and fruitful within and across disciplines.

References

- Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124, Tartu, Estonia, January. Association for Computational Linguistics.
- Timofey Arkhangelskiy, Hanna Hedeland, and Aleksandr Riaposov. 2020. Evaluating and assuring research data quality for audiovisual annotated language data. In *Proceedings of the CLARIN Annual Conference 2020*. CLARIN ERIC.
- Christophe Bahim, Makx Dekkers, Edit Herczog, Keith Russell, and Shelley Stall. 2021. Survey on bridging the gap between funders and communities — perspectives on benefits and challenges of fair assessments. v1.0.
- Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight grammatical annotation in the TEI: New perspectives. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 1795 – 1802, Paris, France. European language resources association (ELRA).
- Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582, September.
- Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L’Hours, Joy Davidson, and Angus White. 2020. FAIRsFAIR Data Object Assessment Metrics. October.
- Nikolaus P. Himmelmann. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation*, 6:187–207.
- Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. 2020. A generic workflow for the data FAIRification process. *Data Intelligence*, 2:56–65.
- Elinor Ochs. 1979. Transcription as theory. In E. Ochs and B.B. Schieffelin, editors, *Developmental pragmatics*, pages 43–72. Academic Press, New York.
- RDA FAIR Data Maturity Model Working Group. 2020. Fair data maturity model: specification and guidelines. June.
- RfII. 2020. The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn.
- Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.
- Thomas Schmidt, Sylvia Dickgießer, and Joachim Gasch. 2013a. Die datenbank für gesprochenes deutsch - DGD2.
- Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013b. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.
- Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in clarin. In *Selected papers from the CLARIN Annual Conference*, pages 113–130, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.
- Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.
- Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.
- Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France, May. European Language Resources Association.

- Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.
- Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.
- Mark Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:180118, 06.

Integrating TEITOK and KonText/PMLTQ at LINDAT

Maarten Janssen

Institute of Formal and Applied Linguistics

Charles University, Czech Republic

janssen@ufal.mff.cuni.cz

Abstract

In this paper we describe how the TEITOK corpus platform was integrated with the KonText and PML-TQ corpus platforms at LINDAT to provide document visualization for both existing and future resources at LINDAT. TEITOK is an online platform for searching, viewing, and editing corpora, where corpus files are stored as annotated TEI/XML files. The TEITOK integration also means LINDAT resources will become available in TEI/XML format, and searchable in CWB on top of existing tools at the institute. Although the integration described in this paper is specific for LINDAT, the method should be applicable to the integration of TEITOK or similar tools into an existing corpus architecture.

1 Introduction

LINDAT/CLARIAH-CZ is a Czech centre for data providing certified storage and natural language processing services. The LINDAT repository provides a direct implementation of the core objective of the CLARIN ERIC to advance research in humanities and social sciences by giving researchers unified single sign-on access to a platform which integrates language-based resources and advanced tools. But although the repository makes the raw data accessible, without an online interface for searching, that only makes the data usable for a limited group of researchers. That is why LINDAT aims to gradually make as many of the corpora in its repository as possible available via KonText (Josífko, 2014) and, in the case of treebanks, PML-TQ (Pajas et al., 2009), two corpus search interfaces.

However, many of the corpora in the LINDAT repository, as well as many (Czech) corpora that are not currently in the repository, are corpora with a solid footing in the digital humanities, such as historical corpora and learner corpora. And for such corpora, a good part of the users will be more interested in visualizing the individual documents in the corpus in a readable form, than they are in KWIC lists or search statistics. Since KonText and PML-TQ do not provide a graphical interface to view entire documents, but only a view on the direct context in terms of corpus tokens, the decision was made to integrate TEITOK (Janssen, 2016) as a way to make corpora available in a manner more fit to documents for the digital humanities. In TEITOK, corpus files are stored in the TEI/XML format, and adopting this well-established standard will further improve the interoperability of the LINDAT corpora. In this presentation, we will show how the integration between TEITOK and KonText, as well as PML-TQ, was designed at LINDAT. The combined interface can be found online at: <http://lindat.mff.cuni.cz/services/teitok/>.

The two integrations take a different approach, and beyond explaining how the integration is done at LINDAT, we hope this paper illustrates how different corpus tools can be integrated within a single corpus architecture, and how TEITOK could be used by virtually any corpus tool for document visualization.

2 TEITOK, KonText and PML-TQ

This section will first give a short description of the three corpus tools involved in the integration, before turning to a description on how they were integrated. The description of the integration avoids too much technical detail, and attempts to use terms applicable to any corpus environment.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

2.1 KonText

KonText is an advanced corpus query interface and corpus data integration platform built around the open source version of the corpus search engine Manatee (Rychlý, 2007), maintained by the Institute of the Czech National Corpus. It allows searching corpora in a (slightly modified version of) the Corpus Query Language (CQL), which was initially designed for the Corpus WorkBench (Evert and Hardie, 2011), and in addition several simplified search options are provided. It has the same overall design as most other online corpus interfaces: it has a landing page where you can select a corpus. And once a corpus has been selected, you can run search queries to obtain lists of results. The default visualization is as a KWIC list, while other views are available as well. And in the KWIC list, you can click on results to see the data about the document the result was found in, or the direct context for the result.

For a query result, you can then obtain various statistical analysis, including a list of collocates and various types of frequency lists. Queries can be stored and shared, and the search results can be downloaded in various file formats for off-line treatment. And there is support for both spoken corpora and parallel corpora in KonText.

2.2 PML-TQ

The PML-Tree Query (PML-TQ) is a powerful open-source search tool for all kinds of linguistically annotated treebanks. The tool works natively with treebanks encoded in the Prague Mark-up Language (PML) data format. The tool can be queried using a search API, and there is also an online search interface at LINDAT.

The online interface for PML-TQ has the same overall design as KonText: there is a landing page where you can choose a dataset, after which you can query that dataset using the PML-TQ Query Language. The result can be either a list of frequencies or a list of results, depending on the query. Result lists are not shown as KWIC lists like in KonText, but rather the first resulting sentence is shown, with the dependency tree for it below it, and a button to go to the next result. The dependency trees are drawn by a web service that produces SVG images for any sentence.

PML is a multi-layered format, which can contain multiple tree layers. For instance for the Prague Dependency Treebank (the latest version is PDT-C, (Hajič et al., 2020)) there are two different layers: the surface syntax layer (a-layer, containing dependency trees), and the deep syntax layer (t-layer). In the query, you always have to indicate the layer you want to search in, and the results show the trees from the relevant layer.

2.3 TEITOK

TEITOK is an online platform to view and edit corpus files stored in the TEI/XML format. The base TEI/XML files can be viewed in a range of different ways depending on the type of XML file, including facsimile-aligned text views, original or normalized text rendering, wave-form aligned views, etc. Dedicated visualization modules can be added when needed. For an explanation of the visualization features of TEITOK, see for instance Janssen (2016) or Janssen (2018).

Corpora in TEITOK are made searchable using the Corpus WorkBench (CWB). TEITOK uses a dedicated program to directly write CWB files from the TEI/XML files. During that CWB export, TEITOK also writes the filename of the XML file as an attribute, and keeps the byte-offset of the token in the original file in a separate file. With these byte-offsets, the CWB corpus becomes an index over the XML files, where the CWB results can be used to directly look up the corresponding XML fragments. The CWB index is regenerated frequently to make sure the byte-offsets reflect any possible changes in the XML files.

How a TEITOK corpus is exported to CWB is defined in the corpus settings - a central configuration file for the corpus in TEITOK that defines, apart from the CWB export, a myriad of other things like which token attributes should be editable, which items should appear in the menu bar, what the fixed fields in the `teiHeader` are, etc. The export settings are explicitly not kept in the `teiHeader`, since they belong to the corpus, and not to the XML files; and also because the same XML file is sometimes used in multiple corpus projects, each with its own settings. Since TEI files are edited in TEITOK, keeping

copies of the same file with a different header would lead to inconsistencies.

The TEITOK search results are rendered as an XML fragments: the CWB results are used to lookup the corresponding XML using the byte-offsets written during the export. Therefore, all the information in the XML file is present in the search result - including information that is not in the CWB corpus. This can include XML regions that were not exported to the indexed corpus (such as bold face or italics); elements that fall between tokens and as such cannot be exported to CWB (such as line breaks or notes); elements that fall below the level of the token and as such cannot be exported to CWB (such as morphemes); and elements that should not be exported as tokens since that would interrupt the token sequences if they were (such as deleted words). All such elements are often highly relevant to correctly interpret the context. Also, the XML fragment will represent contractions as contractions, and not split into grammatical tokens as they are in CWB. So a text containing *wanna* or the Spanish *del* (of+the), will render the original text faithfully, instead of having it changed into *want to* and *de el*. This is important for linguists who want to use corpora to find example to copy-paste into an exercise or article.

2.4 Combining TEITOK and KonText

As mentioned before, the decision was made at LINDAT to adopt TEITOK for document visualization. KonText and TEITOK are comparable frameworks in the sense that both are online platforms to search corpora using the Corpus Query Language (CQL). And intuitively, it would seem to be the ideal solution to fully integrate the two platforms in a single platform by either incorporating the relevant parts of TEITOK into KonText, or by rewriting TEITOK to use the search engine using in KonText, Manatee, as a backbone instead of CWB. But the problem with both these options is that they constitute a major overhaul of either TEITOK or KonText, since they share their query language, but not their general set-up or storage format. And neither of which is used exclusively for LINDAT, but used in a growing number of projects around the world. And a major rewrite would inevitably lead to compatibility problems at other projects using these tools. The only feasible option would be to drop KonText altogether in favour of TEITOK, but not only are most users of LINDAT more used to the KonText interface than they are to their TEITOK counterparts, but also there are various projects at LINDAT that rely on the KonText interface.

Therefore, a more modest integration was selected in which the two platforms are kept as independent interfaces, with links leading from one to the other. That set-up has the added advantage that it allows users that are more familiar with the CWB flavour of CQL than they are with the KonText implementation have the option to use that by doing their searches using TEITOK rather than KonText. Because although the base query language is the same, they start to diverge for more complex queries, such as queries involving global restrictions or multi-valued attributes, as well as small differences such as the names of structural attributes (`text_year` vs. `text.year`).

The way combined corpora are created is as follows: TEITOK is used to directly create a CWB corpus from TEI/XML files, using the standard TEITOK set-up. Once complete, the CWB corpus is exported to VRT (the one-word-per-line format used in both CWB and Manatee) using the CWB tools, after which it is loaded into Manatee. The registry file for both platforms is written by TEITOK using the aforementioned corpus settings. All this is done by a single script that can be run from within the TEITOK interface.

In order to link the two platforms, a small addition was made to both platforms: in KonText an option was added to allow the context of a token to be provided by an external REST service. And in TEITOK a module was added that can render the XML context of any corpus token as an HTML page. Combining these two modules makes it possible to click on a token in the KonText search result, and in the pop-up window that shows the context see the TEITOK rendering of the original XML with all its attributes. The fragment also comes with a link to the visualization of the full document in TEITOK. An example of a TEITOK context from a Czech text from Skript 2015¹ is given in Figure 1, where the word *cvičtho* (practicing) is deleted in the original and hence not present in the CWB or Manatee corpus, as can be seen in the KWIC line. It is visible in the TEITOK context since it is in the original XML file.

¹<http://lindat.mff.cuni.cz/services/teitok/skript2015/>

<input type="checkbox"/>	vra_ka_129_01_1_1	maminka tam má svoji kočku Zrzku . Chodíme spolu na procházky	do lesa . A táta je doma a pracuje na
<input type="checkbox"/>	ki9apanzuz_1	Milana2 Nováka2 a Obchodní školu Milana3	. Ve volném čase si ráda čtu knížky ,
<input type="checkbox"/>	vra_jt_148_01_1_1	S tátou jezdím na trénink nebo chodím s mámou na procházky	. Mám rád oba dva rodiče . Jezdím s tátou
<input type="checkbox"/>	AR_Mare_006_12_1_1	je velký takový zrzavý chodím sním na procházky	a krmýho cvičím ho a mám ho rád
<input type="checkbox"/>	ki9apanzuz_1	, které jsem dostala k Vánocům . Chodím ven na procházky	s babičiným psem Maxem , jehož rasa je pudl a
<input type="checkbox"/>	ho5dhajluc_1	do Anglie za tatkou . Když jsme se vrátili z procházky	, byli puštěni pejsci Dak a Bady . Bady si
<input type="checkbox"/>	VRA_LC_037_01_1_1	bíl . 4 . Rád si s ním chodím na procházky	a rád si s ním hraju . Tato osoba ,
<input type="checkbox"/>	cl8bpaledv_1		psem Argem . V těchto teplých dnech na
<input type="checkbox"/>	vra_km_130_01_1_1		e s babí a dědou k lapáku .
<input type="checkbox"/>	cb1achrmil_02_1		je jet na kole vykoupat do Rudy
<input type="checkbox"/>	cb1akumzuz_01_1_1		uměl vyprávět tak skvělé vtipy !
<input type="checkbox"/>	vra_lc_142_01_1_1		hrajeme . O víkendu chodíme s
<input type="checkbox"/>	REZ_HAB_069_01_1_1		Ještě jsem si vzpoměla že
<input type="checkbox"/>	cl8bspipet_1		jde . Ahoj . Já jsem tvé

Figure 1: Example of a TEITOK context in KonText

Apart from the search integration, some additions were made or are being made to TEITOK to make it more compatible with an infrastructure like LINDAT. Two improvements that stand out in this respect are the following: (1) the inclusion of server-wide settings: in its original set-up, TEITOK corpora are completely independent, and each corpus has to define its own characteristics. But for an infrastructure with many corpora, it is necessary to be able to define a shared set of definitions and styles for all corpora. And (2) the option to generate static versions of TEITOK corpora: since TEITOK offers the option to edit corpora, TEITOK corpora are by default not fixed sets of data. But for a repository like LINDAT, fixed datasets are needed to be able to attribute them an object handle, and to allow users to quote reproducible data. To account for this, TEITOK now offers the option to create named corpus versions for inclusion in the repository.

2.5 Combining TEITOK and PML-TQ

In principle the same integration as was implemented from KonText could also be used for PML-TQ: have the web interface of PML-TQ display an externally retrieved context, which can then hence show the HTML output of TEITOK for the result sentence and a link to the textual context. The procedure for PML-TQ would have to be a bit more complex than that for KonText, since the PML-TQ corpus is not generated from the same TEI/XML files (and hence do not share their token IDs), but still possible since both the TEITOK and the PML-TQ corpus are generated from the same PML files. However, for the moment this has not been done, and we will see how people use the different tools whether this integration would be useful.

However, since the brunt of the work in PML-TQ is done by a web service, a deeper integration was made: in the TEITOK interface, you can type in a PML-TQ query. Upon submitting, that query is sent to the web service to retrieve a list of matching sentences (or a table of results). And rather than showing trees one by one, the system looks up the XML fragments for sentences in TEITOK that correspond to the first 100 results, and displays them as a KWIC list, with all the mouse-over token information TEITOK typically provides. From that KWIC list, it is possible to see the dependency tree generated by the PML-TQ API (the printserver), or to jump to the text display in TEITOK for that sentence (see Figure 2). For PDT corpora, there is furthermore a link that show the raw data from the PML files from which the corpus was generated, containing information from all the layers for the result sentence.

This approach of course needs a dedicated TEITOK module to deal with the search API of the target tool, in this case PML-TQ. But such a module is not very involved, and it makes for a full integration of the two tools, where the PML-TQ and CQL query languages work in very similar ways within TEITOK.

3 Adding resources

Having an integration between the three platforms is not sufficient for an infrastructure: it is also needed to get corpora into the hybrid system. When adding corpora to the integrated TEITOK/KonText infras-

Corpus Search (PML-TQ)

```
1  b-node $a := [  
2    sibling a-node [  
3      depth-first-follows $a,  
4      afun = $a.afun  
5    ] and afun = "ExD"  
6  ]
```

Execute Query Show treebank options

Results

Showing 0 - 49 of more than 10000 - next

[pml](#) [tree](#) [context](#) Kapitola 2 : Metody vyučování čtení v angličtině

[pml](#) [tree](#) [context](#) Děkuji , dobře , a vy ? Ano , shání se tady těžko .

faust_2010_07_es_02-SCzechA-p0142-s1-root

[pml](#) [tree](#) [context](#) všichni hrajou fotbal , jen já ne

Figure 2: Example of a PML-TQ search in TEITOK

tructure, one has to distinguish between corpora that are already in KonText, existing corpora that are not yet in KonText, and newly planned corpora.

The corpora that are already in KonText can be converted automatically, although they have to be regenerated in order to create the byte-offset files used by TEITOK. The existing corpus is exported to VRT, from which it is converted to a collection of TEI/XML files, one for each document. Some corpus specific action is needed, for instance, it is necessary to indicate the correct XPath according to TEI for each of the metadata in the corpus in order to end up with correct TEI/XML documents; but once these settings are provided, the process is fully automatic. The result is not a full-fledged TEITOK corpus, since information that was not in the KonText corpus will not be available in the TEITOK corpus either. This typically includes typesetting, as well as word spacing and contraction decomposition. But it is a direct and automatic way to add corpora.

For the class of existing corpora that are not yet in KonText, TEI/XML files are generated from the raw data (in whichever format the corpus came in), keeping as much of the original information as possible. In order to facilitate this, we are working on a set of conversion tools from popular corpus formats including ELAN, FoLiA, PagesXML, etc. Even for some of the corpora that were already in KonText, the corpus is being recreated from source, since some of the information was lost in the conversion. An example of this is the Prague Dependency Treebank, where the original PML files contain more morphological information than the KonText corpus that was previously created from it. The richer TEI/XML structure of TEITOK makes it possible to incorporate all this information in the hybrid framework.

And finally, there are those corpora that are still under development or planned for the future. For such corpora, the various options provided by TEI/XML and TEITOK make it possible to encode richer information in the corpus than a traditional corpus based on a one-word-per-line architecture would allow. An example of this is the upcoming ParCzech corpus (Hladká et al., 2020) which was designed from the start as a spoken corpus in TEITOK, and it searchable as a KonText corpus using the hybrid TEITOK/KonText infrastructure. In some cases, existing corpora are being re-planned to make use of the new options. An example of that is CzeSL (Štindlová et al., 2012), a learner corpus where not all the information present in the source material was encoded, but which is now in part being redone to include the deletions, corrections, normalizations, etc. that were previously impossible to include.

TEITOK provides a wide range of visualization options. Therefore, for all corpora, whether they are

automatically converted from KonText, rebuilt from the source, or completely planned in TEITOK, it is worth while to see whether the options provided by TEITOK cannot make more use of the corpus data. For instance, the ACL RD-TEC corpus (QasemiZadeh and Schumann, 2016) is a corpus providing manually annotated terms. This corpus is much better served by adding an interface that presents the collection of terms from the corpus as a terminology.

3.1 Unresolved issues

The integration between KonText and TEITOK works seamlessly for basic corpus features. But on features that are added on to the base corpus architecture, different tools take a different approach. A case in point is the treatment of dependency trees: as the value for a head attribute, KonText expects a relative ordinal number: -2 to indicate that the head occurs 2 positions to the left. While TEITOK expects the ID of the head token. In this particular case, the decision was made to extend KonText to accept different types of head attributes, so that KonText can draw dependency trees from the data provided by TEITOK.

But there are other advanced features where we are still working on the best way to handle them, most crucially speech alignment and parallel corpora. These issues of course only arise because we want the same corpus to work in both systems. If TEITOK were only used for text visualization, these would not be an issue.

For speech corpora, the problem is similar to that for dependency trees: KonText expects the aligned segments in the corpus (typically the utterance) to have an attribute that names the sound-file containing the corresponding speech segment. While TEITOK uses entire sound files marked out higher up in the document (typically in the metadata), and the aligned segment is expected to have attributes marking the start and end time of the corresponding speech segment in that sound file.

For parallel corpora, KonText uses different corpora for different languages, with a table linking two corpora indicating which are the corresponding elements. While TEITOK uses several strategies for parallel corpora, typically with all languages in a single corpus, and using shared attributes to mark out corresponding elements.

For such advanced features, it is necessary to adapt at least one of the two integrated tools to accept the solution provided by the other, or alternatively to have the corpus contain both solutions. For both speech corpora and parallel corpora, we are still exploring all available options.

4 Conclusion

In this paper, we have shown how TEITOK was integrated in the existing workflow based on KonText at LINDAT. The additional options of the combined TEITOK/KonText workflow make it possible to provide a richer document view and context view that are more in line with the requirement of the digital humanities. The TEITOK platform has proven its appeal to the DH community by a growing number of projects of diverse nature, including historical corpora like PostScriptum (CLUL, 2014), spoken corpora like NURC (Oliviera Jr, 2016), and learner corpora like COPLE2 (Mendes et al., 2016). The TEI/XML based design of TEITOK allows those corpora to maintain their domain specific characteristics while at the same time adhering to a standard NLP pipeline.

It is our hope that the combined workflow will attract more corpora into the LINDAT infrastructure that would otherwise not have been made available as a CLARIN resource, with the preliminary indications looking promising in newly established collaborations towards this end. The combined TEITOK/KonText workflow provides a lot of potential for the future, and the fact that TEITOK is now maintained at LINDAT makes it much easier to expand the framework to account for newly arising demands, such as making the corpora accessible via the Federated Content Search (FCS), something that is relatively easy given that there are existing solutions for CWB corpora.

The conversion of the existing corpora in KonText to the TEITOK/KonText hybrid has not been completed yet, not because a straight-forward port would be difficult, but because in many cases it seems that by rebuilding the corpus and thinking of additional interface options, more value could be extracted from the source data. And the spoken corpora as well as the parallel corpora have not yet been ported since the discussion on how to best deal with those in the TEITOK/KonText hybrid has not been finalized, but

is expected to come to a conclusion soon.

It is still too early to say much about the user appreciation of the hybrid solution, but the internal reception for all the corpora added to the hybrid solution has been very positive.

The integration described of TEITOK with an established corpus workflow is not specific to KonText: the changes in KonText are minimal, and the XML context provided by TEITOK can be called from any corpus tool - the only requirement is to have shared identifiers between the different versions of the corpus. And this easy to establish integration could add three important things to any corpus environment: a rich XML view on the context from the environment, a link to a full document visualization, and the option to allow corpus editors to easily maintain and edit their own corpora.

References

- CLUL. 2014. P.S. Post Scriptum. arquivo digital de escrita quotidiana em portugal e espanha na Época moderna.
- Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.
- Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2020. Prague dependency treebank - consolidated 1.0 (PDT-c 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Barbora Hladká, Matyáš Kopp, and Pavel Straňák. 2020. ParCzech PS7 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.
- Maarten Janssen. 2018. Adding words to manuscripts: From PagesXML to TEITOK. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, *Digital Libraries for Open Knowledge*, pages 152–157, Cham. Springer International Publishing.
- Michal Josífko. 2014. KonText web demo. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- Miguel Oliviera Jr. 2016. NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nure). *CHIMERA: Romance Corpora and Linguistic Studies*, 3(2):149–174.
- Petr Pajas, Jan Štěpánek, and Michal Sedlák. 2009. PML tree query. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Pavel Rychlý. 2007. Manatee/Bonito - a modular corpus manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, pages 65–70.
- Barbora Štindlová, Svatava Škodová, Jirka Hana, and Alexandr Rosen. 2012. CzeSL – an error tagged corpus of Czech as a second language. pages 21–32, 01.

The CLARIN-DK Text Tonsorium

Bart Jongejan

Department of Nordic Studies and Linguistics

University of Copenhagen, Denmark

bartj@hum.ku.dk

Abstract

The Text Tonsorium (TT) is a workflow management system (WMS) for Natural Language Processing (NLP). The software implements a design goal that sets it apart from other WMSes: it operates without manually composed workflow designs. The TT invites tool providers to register and integrate their tools, without having to think about the workflow designs that new tools can become part of. Both input and output of new tools are specified by expressing language, file format, type of content, etc. in terms of an ontology. Likewise, users of the TT define their goal in terms of this ontology and let the TT compute the workflow designs that fulfill that goal. When the user has chosen one of the proposed workflow designs, the TT enacts it with the user's input. This untraditional approach to workflows requires some familiarization. In principle, the TT cannot predict which of the proposed workflow designs is most appropriate, because the text may have peculiarities that are as yet uncharted. The user has to make the choice. In this paper, we reflect on the experiences with providing, testing and using workflows aimed at annotating transcripts of parliamentary debates. We propose possible improvements of the TT that can facilitate its use by the wider CLARIN community.

1 Introduction

Many developments in workflow management systems (WMS) are aimed at bringing down workflow execution time and handling ever bigger amounts of data. In the user community that CLARIN addresses, on the other hand, ease of use, especially for users with a nontechnical background, and adaptability to special needs, are often more important than speed and data size.

Our aim is to let small and medium scale scholarly projects benefit from an easy to use and open WMS that manages a well-maintained collection of state of the art NLP tools. This WMS is the Text Tonsorium (TT). It was constructed by the CLARIN-DK staff (Offersgaard et al., 2011), but it has been away from the clarin.dk web site for some years. After many technical improvements, it is again part of CLARIN-DK, this time with a new interface.

Traditionally, workflow management systems require (expert) users for the construction of workflow designs. A good example of such a system is WebLicht¹ (Hinrichs et al., 2010). The characteristic that sets the TT apart from traditional WMSes is that workflow designs are computed automatically. Central to this computation is an ontology that is used to describe all data objects involved in workflows. Tools, too, are described by this ontology, since the TT, when computing workflow designs, only needs to know what kind of data enters a tool and what kind of data comes out. A technical description that explains how the TT computes workflow designs is in Jongejan (2016). More about the user perspective of the TT, as conceived during the DK-CLARIN project, is in Offersgaard et al. (2011) and in Jongejan (2013).

Users experience the TT as a very different service, compared to traditional workflow management systems, because the TT dispenses with the expert user who creates and shares workflow

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

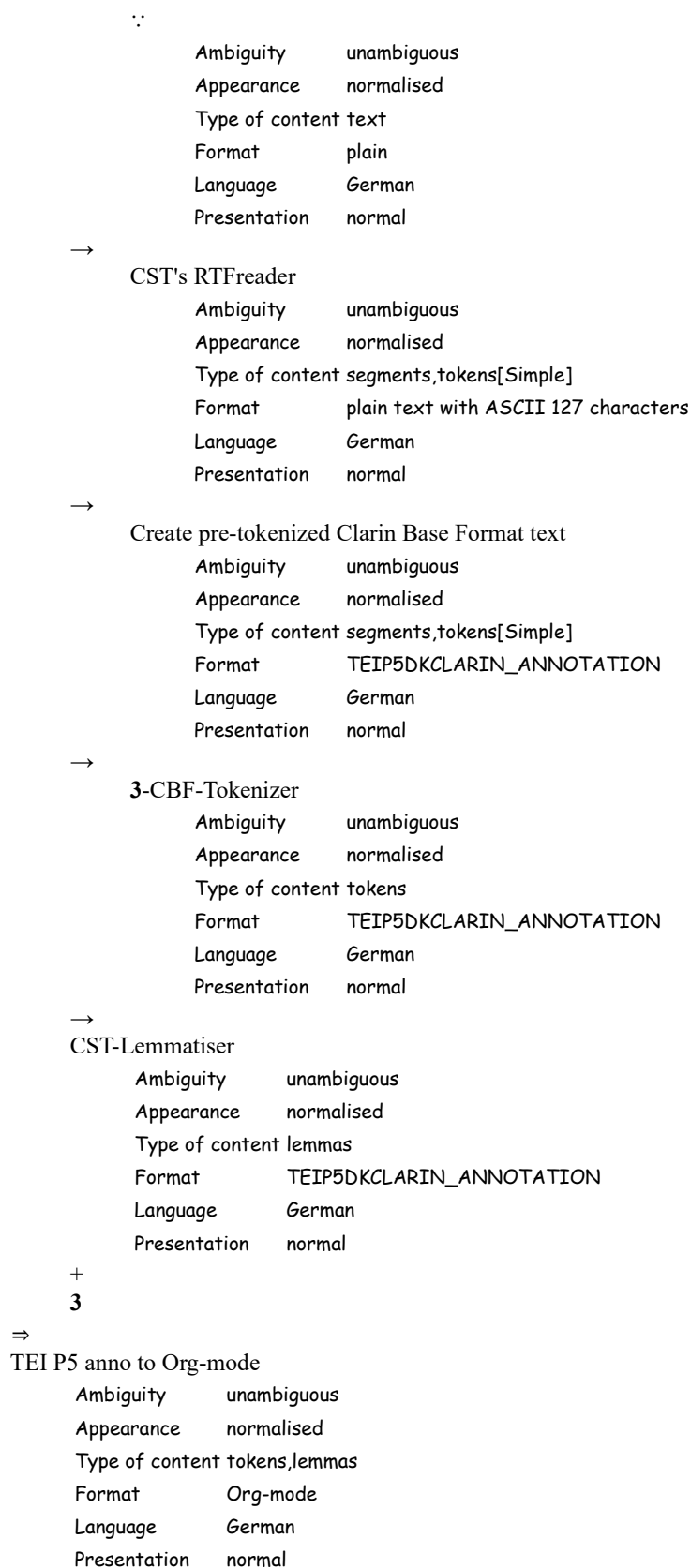


Figure 1: Graph of a workflow design with five tools. The flow is from top to bottom. The output from tools marked with a number is sent to the next tool and to any tool that has the same number mentioned as input. Here, the 3rd tool sends output to the 4th and the 5th tool.

designs with other less expert users. True, users do not have to think about the technical aspects of workflows, but users must decide for themselves which of the automatically computed workflow designs to choose. This choice can be difficult and time consuming, and would have been made once and for all, if a human expert user had created a workflow design for them. It is therefore necessary to offer guidance to inexperienced users. On the one hand, metadata about tools and datatypes must be presented in an understandable way and in the places where they can be helpful. On the other hand, knowledge and experiences must be shared between users, first and foremost by way of named, descriptive bookmarks that can be attached to workflow designs that have been tried out and used.

The structure of this paper is as follows. Section 2 is about the current state of the TT. Section 3 presents a condensed technical outline of the TT. Section 4 explains how the data is described by metadata and how the tools are described in terms of the metadata characterizing their input and output data. Section 5 is a case study of the design, implementation and use of the Danish ParlaMint workflow design. In Section 6 we outline an improvement that some users have proposed, but that cannot be realized. To make good for this, Section 7 is an exposition of our plans for making an interface to the TT that can be easy to use by the larger CLARIN community. Section 8 tells where the TT can be found on the internet. Finally, in Section 9, concluding remarks follow.

2 Current State

There are currently over 40 tools² integrated in the TT, spanning from tokenization to syntactic parsing and from PDF-to-text conversion to text-to-speech transformation. Some tools were developed to address the needs of a single project and later generalized to make them useful for a wider segment of researchers. An example of spin-off from two consecutive and unrelated projects is that the TT is able to annotate a wide class of TEI P5 formatted texts with lemmas, part-of-speech (PoS) tags and syntactic dependencies.

Many of the tools are multi-lingual. The CST lemmatizer (Jongejan and Dalianis, 2009) (CSTlemma), for example, lemmatizes 28 languages³. The Danish linguistic resources for CSTlemma cover three historical periods: medieval, late modern and contemporary. A few tools, such as the Named Entity Recognizer, work only for Danish.

3 Technical Outline

3.1 Software Components

The TT is a web application that consists of a hub and a webservice for each tool that is orchestrated by the TT. The integrated tools communicate with the hub by the HTTP protocol, but do not communicate with each other. Users interact only with the hub.

3.2 Tool Integration

The TT offers an easy way of embedding a tool in an ecosystem of already existing tools. First, a tool provider visits the administration page of the TT and enters boiler plate metadata (ToolID, ContactEmail, Version, Title, ServiceURL, Publisher, ContentProvider, Creator, InfoAbout, Description) as well as metadata that describes the input and the output of the tool in terms of language, file format, and a few other dimensions. The TT then creates a program stub in the PHP language for a web service that is already tailored to the tool to be integrated.

3.3 Workflow Composition

The TT uses dynamic programming with memoization to compute all workflow designs that combine tools such that the output will be in agreement with the user's specifications, given the

²See <https://cst.dk/texton/help>

³Some of training data sets with which CSTlemma was trained, to wit the MULTEXT East free and non-commercial lexicons (Erjavec et al., 2010a; Erjavec et al., 2010b), were found in the Slovenian CLARIN portal.

user's input. Computation of these designs, pruning unlikely designs and removing irrelevant details from the presentation of the list of remaining candidates, is relatively fast, given that there may be thousands of viable designs to sift through. This process can take anywhere from a few seconds to a couple of minutes.

Fig. 1 shows the full details of a single workflow design. In this case, the TT had recognized the input as plain text and the user had defined the goal as German lemmas. The shown workflow design is one out of 50 designs, a number that would have been reduced to ten if the user also had mentioned that the output had to be in ORG-mode format.

3.4 Workflow Enactment

When the user has chosen one of the proposed candidate workflow designs, the TT enacts that workflow with the data that the user has uploaded as input. The input, intermediary results and final results are temporarily stored on the computer on which the TT runs.

The results from each step in a running workflow can be inspected as soon as the step has been executed. The results can be downloaded as a zipped archive. The user can choose whether or not to include all input and intermediary results in the zip-file.

3.5 Data Formats

The TT handles Office documents, TEI P5 documents, HTML, PDF, images of text, plain text, JSON, ORG-mode tables, CONLL (14 column CONLL-2009 as well as 10 column CoNLL-U), CWB (Corpus Workbench), bracketed (Lisp-like) text, and WAV sound files. Some formats are only available on the input or on the output side.

The TT is primarily designed for annotation tools that output their result without also copying the input to the output. By allocating stand-off annotations in files separate from the input and from other annotations, the TT has the freedom to send intermediary results from earlier processes in any combination as inputs to later processes, thus circumventing a need to have data definitions for each possible combination.

Since users normally require output that contains results from several workflow steps combined, some data definitions for combinations of text and/or annotations have been made, employing expressive formats such as JSON, ORG mode tables, CONLL, CWB and TEI P5. Fig. 1 illustrates this: tokens and lemmas are in separate files, the first as the output of a tokenizer and the latter as the output of a lemmatizer. These two intermediary results are both sent to the final tool, which combines the tokens and the lemmas in a table with two columns, using the ORG-mode formatting mode.

Complex data types are not restricted to the final result of a workflow. If there is a tool that accepts a complex data type, workflow designs can be computed that take user input of that type. Such data types can also occur as intermediary result.

4 Metadata

Seven dimensions are used to describe data: the *Type of Content*, the *Language*, the *Format*, the *Historical Period*, the *Appearance*, the *Assemblage*, and the *Ambiguity*, see Fig 2.

The TT treats all dimensions on an equal footing: it does not care whether a tool transforms data between two *Languages*, or between two *Formats*, or between two *Types of Content*.

Each of these dimensions needs to be populated with values. For example, this is the current list of values that *Type of Content* can take: **text, tokens, sentences, segments, paragraphs, PoS tags, lemmas, word classes, (syntactic) dependencies, tagged terms, named entities, morphemes, noun phrases, repeated phrases, N-gram frequencies, keywords, multiple word terms, lexicon, and head movements**. These *Types of Content* are primitive. There are also some complex *Types of Content* that combine two or more primitive *Types of Content*. For an example of their use, see Section 3.5 and Fig. 1, where we see two examples of complex *Types of Content* that combine more primitive *Types of Content*: **segments,tokens** and **tokens,lemmas**.

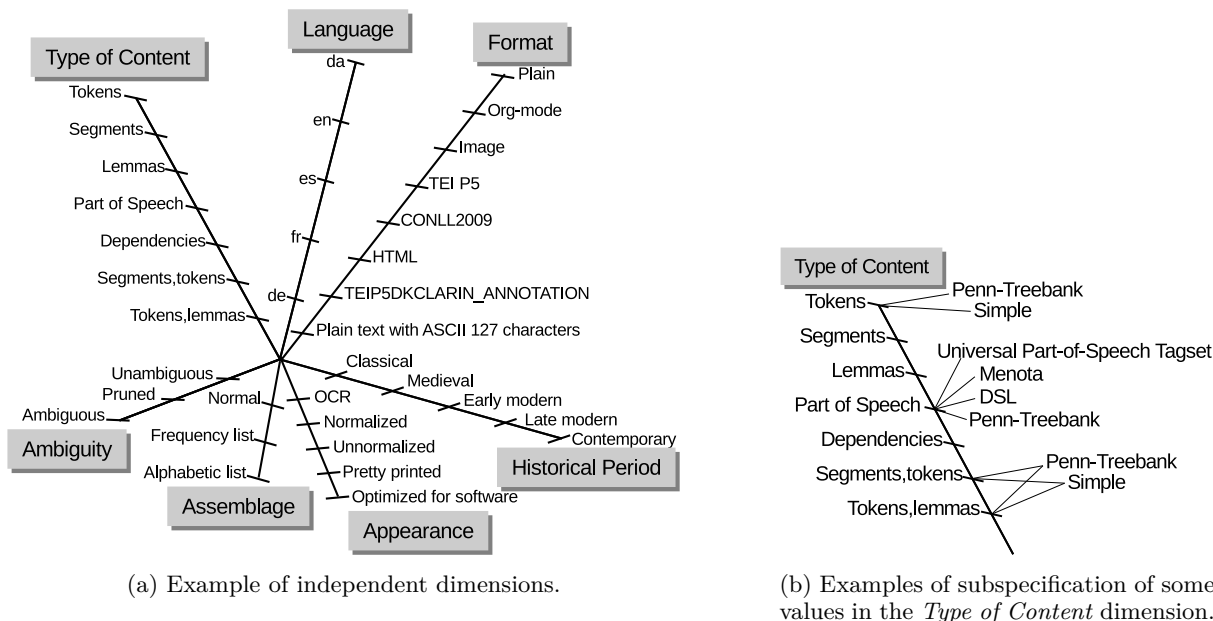


Figure 2: Dimensions and subspecifications of values along one of the dimensions.

The TT handles one more level of specification. In this extra level it is possible to discern variants of a given value in a dimension. For example, there may be a need to discriminate between Universal PoS tags and the PoS tag set used in the Penn Treebank, which are specifications of the value **PoS tags** in the *Type of Content* dimension. Similarly, it might be useful to distinguish between JPEG, PNG and SVG, which are specifications of the value **image** in the *Format* dimension. See Fig 2. The purpose of this extra level of information is to make matching tools with each other, with input data and with output requirements, more forgiving. It also helps to keep technical details away from the user.

5 Use case: Annotation of a Danish Parliamentary Corpus

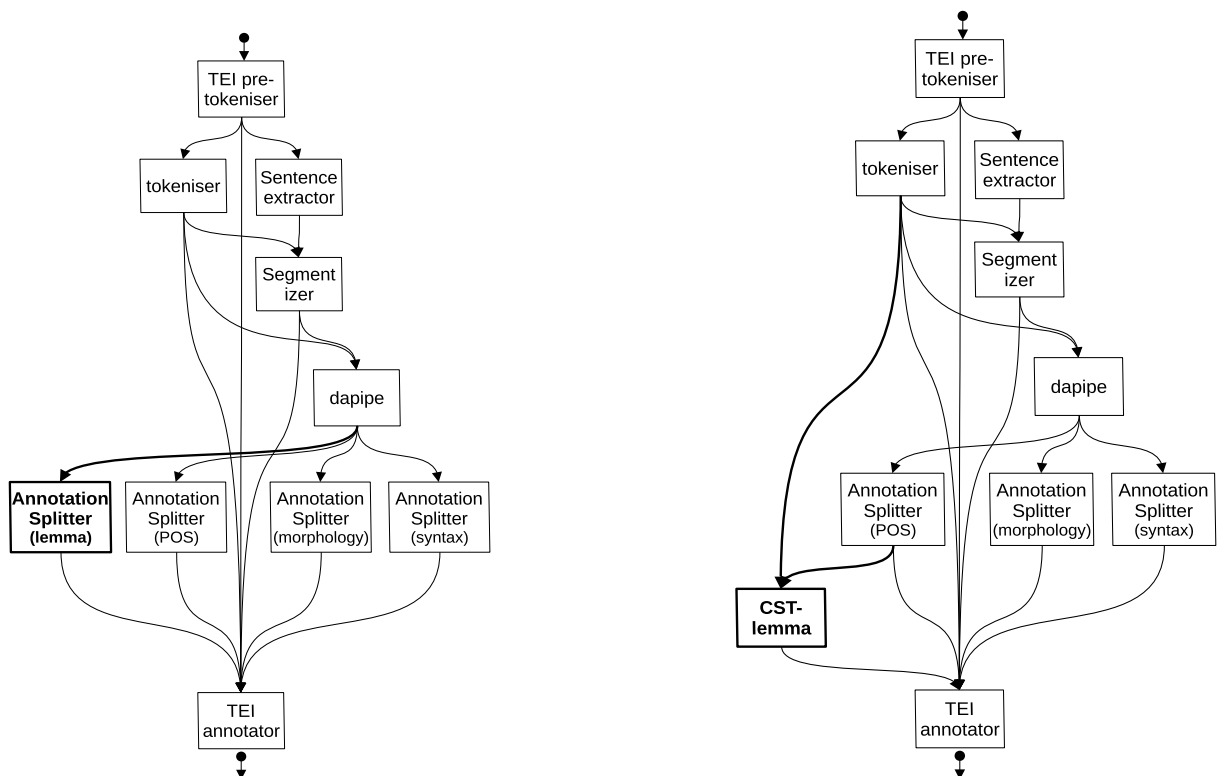
The ParlaMint project⁴ has the aim to make national parliamentary data in several countries available in a uniform format and enriched with linguistic annotations. The chosen format follows the TEI P5 guidelines. We have decided to annotate the Danish parliamentary data with morpho-syntactic descriptions (**msd**), lemmas and syntactic dependency relations.

For handling the linguistic annotation process of the Danish ParlaMint data, the choice fell upon the TT, since it already had several tools for transformation and annotation of TEI P5 data.

A tool for annotation of Danish (plain) text with **PoS tags**, **lemmas**, and syntactic **dependency** relations was already integrated in the TT. This tool, **dapipe**⁵, is based on UD-pipe (Straka and Straková, 2017) and trained by colleagues at the IT University of Copenhagen. Per default, **dapipe** expects plain text and does everything necessary: segmentation, tokenization, PoS tagging, morphological analysis, lemmatization, and syntax analysis. Our first experiment was to iterate over all **<seg>** elements in each input file, for each element (1) extracting the content, (2) saving it in a plain text file, (3) running **dapipe** with that file as input, and (4) transform **dapipe**'s output to follow the TEI P5 standard. This was the most straightforward way to use **dapipe** in the ParlaMint project, but it caused a considerable overhead, since **dapipe** had to be started and initialized with the Danish language model for every single utterance. Then we discovered that there is a possibility to feed **dapipe** with an already tokenized and seg-

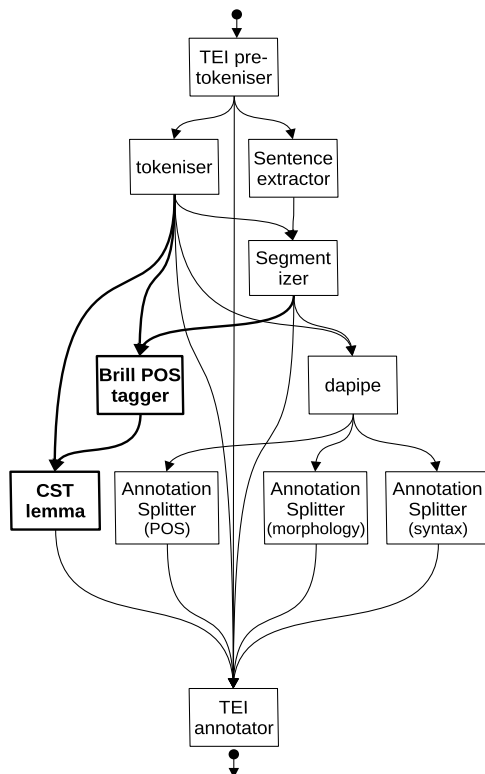
⁴ParlaMint: Towards Comparable Parliamentary Corpora (<https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>)

⁵<https://github.com/ITUnlp/dapipe>



(a) Workflow using dapipe for all linguistic annotations.

(b) Workflow using CSTlemma for better lemmatization. CSTlemma using dapipe's POS as hint.



(c) Workflow that uses an alternative tagger to provide hints to CSTlemma.

Figure 3: Three candidate workflow designs (out of 21) that were tried for annotating transcripts of parliamentary debates in Denmark. (b) was chosen because it was the only design guaranteeing congruence between lemmas and PoS tags.

mentized input. We decided to choose another, more complex but more efficient solution: first tokenize and segmentize the input, using TEI elements to indicate the **tokens** and **segments**, and then send these preprocessed data through *dapipe* in one go. *Dapipe*'s output could then be matched to the proper **tokens** and **segments**.

The other four processes (PoS tagging, morphosyntactic analysis, lemmatization, and syntax analysis) remained unseparable, however. To integrate *dapipe* in the TT, we therefore created an aggregate content type and a 'splitter' tool that does nothing but outputting just one of the four annotation layers in the aggregated content type: **PoS tags**, **morphology** (later in the process to be combined into **msd**), **lemmas** or syntactic **dependencies**. Now we could segmentize and tokenize a TEI document, while keeping its TEI structure, feed it to *dapipe*, and retrieve the annotation layers as `<spanGrp>` elements in output TEI files. These annotation layers could then be merged into the input. This workflow design is illustrated in Fig. 3(a).

On inspection, the lemmas produced by *dapipe* were not very good. Not as good, we believed, as the lemmas that another integrated tool, *CSTlemma*, would have produced. We mapped the Universal PoS tagset, which *dapipe* employed, to the PoS tag set employed by *CSTlemma*. When that had been implemented in the TT, workflow designs that contained both *dapipe* and *CSTlemma* were shown in the list of proposed workflow designs. One of these is shown in Fig. 3(b). So now we had a workflow design using *dapipe* for all annotation layers and another workflow design that ignored *dapipe*'s lemma output, using *CSTlemma* lemmas instead. Comparing the two outputs, it became clear that *CSTlemma* produced better lemmas than *dapipe*. We discovered another difference between the lemma predictions: whereas *CSTlemma* computed the lemma of a word according to rules dictated by the PoS tag assigned to that word, *dapipe*, when computing the lemma, seemed unaware that another part of *dapipe* was predicting the PoS tag of that word. The result was that a word could be PoS tagged as a verb, but lemmatized as though it was a noun.

It is in such experimental phases of projects that the TT shows its advantage over hand-made workflow designs: without a considerable investment of human working hours, we were able to compare two workflow designs before deciding which one to use to annotate all documents in the Danish corpus. On the one hand we had results produced by *dapipe* only, with lemmas and PoS tags predicted independently, and on the other hand we had results produced by *dapipe* and *CSTlemma* in unison, with lemmas created conditioned by the PoS tags produced by *dapipe*. On beforehand, it was impossible to know which result would be the best one. *Dapipe*'s lemmas could be erroneous, but would be erroneous independently of any errors in *dapipe*'s PoS tags. *CSTlemma*'s lemmas could be more or less erroneous than *dapipe*'s. If there were lots of PoS tagging errors, *CSTlemma*'s results would very likely be worse than *dapipe*'s, but if there were few PoS tagging errors, the lemmas would be of higher quality. No programmatic method would be able to predict which approach would be the best one. Only experimentation could tell.

We did a third experiment. We reasoned that if *dapipe* makes many PoS tagging errors, and if there is an alternative PoS tagger that makes fewer errors, than we should feed *CSTlemma* with PoS tags created by the alternative PoS tagger. Skimming through the list of workflow designs that the TT proposed, we found a workflow design that incorporated the Brill tagger, see Fig. 3(c). We run the same test document through all three workflows and compared the lemmas. As expected, workflow 3(a) and workflow 3(c) produced lemmas that not always were congruent with the PoS-tags predicted by *dapipe*. Furthermore, workflow 3(b) showed few lemmatization errors, compared to the other two, and many of the errors were due to PoS-errors. Since PoS errors percolate to the the syntax analysis, it seemed reasonable to also let the PoS errors percolate to the lemmatization, especially if the original *dapipe* lemmas had more errors than *CSTlemma* lemmas. After comparing the three workflow designs, we selected the workflow design depicted in Fig. 3(b).

The Danish Parliamentary corpus contains 688 xml files and cover a period from October 2014 until September 2020. We will annotate these files in groups comprising one year of debates at

a time, which amounts to about 100 documents per group. We expect that each group takes about 2 hours to be completely processed.

During the test phase of the workflow, we were in need of a visualisation of the results. The correctness of the syntactic annotation in the TEI P5 formatted workflow output was very hard to check, so we added an extra tool to the TT that transforms the hard-to-read TEI P5 format to a plain, easier to read CONLL-U format. To make the validation even more easy, we added another tool to the TT, a tool that converts the CONLL-U format to the Penn Treebank bracketed list format, which displays the syntactic dependencies as a tree structure.

Given that the TT was to have a workflow that enriches ParlaMint TEI P5 textual input with PoS, morpho-syntactic, lemma and syntax annotations, it was interesting to see that the same workflow designs could be applied to non-ParlaMint texts. From an earlier project, we knew that there are users who use the TT to annotate TEI P5 documents with PoS tags and lemmas, while retaining many sorts of TEI P5 tags attached to words in the input file, such as <add>, <app>, , <ex>, <lem>, and <rdg>. Now these users in addition can enrich their texts with syntax annotations.

6 An Improvement that is not coming

There is an improvement that is hard, if not impossible, to deliver. The Text Tonsorium has the technical expertise to construct workflow designs that ... work, but it does not have the expertise to tell which workflow design is the best in a given situation.

In other workflow management systems, users rely on the expert knowledge of an experienced colleague who manually picks the tools that, together, constitute the best workflow design for a given task. End-users do therefore not have to choose between lemmatizer A and lemmatizer B. Not only that makes life easy for users, a hand-made workflow design can be used again and again, given a name, described and shared with other users. There are however some pitfalls:

1. Some tools may have changed, delivering output that is not quite the same as when the workflow was manually designed. The quality label may lose its credibility over time.
2. Some tools may improve so much that the same expert user would choose that tool instead of the one that was chosen as part of the manually edited workflow design. So end-users do not reap the fruits of technological progress.
3. Tools may reach end-of-life and stop being executable. Workflow designs that depend on such tools stop functioning as well. Reproducibility of earlier experiments suddenly stops, and an expert user (the same or a new one) has to be called in to design a new workflow.
4. Workflow designers may for whatever reason choose to design several workflows that attain the same goal. The end-user would then have to choose the workflow that fits the actual situation best. That, of course, weakens the utility of having expert human users whose task it is to make choices on behalf of end-users.
5. Expert users may shun away from parts of the solution space for various reasons, but those reasons need not always be valid in situations that end-users encounter:
 - (a) An expert user may disregard some tools for reasons that have little to do with their function, just to be on the safe side. For example, a tool may be known to be excessively resource gobbling, or slow, or unable to handle large input, or in an unstable beta phase of development. However, such problems may be of no importance for an end-user, hardware resources may become better and software may become available in a more stable version.
 - (b) An expert workflow designer has to make assumptions about the input data that perhaps do not apply in the end-user's setting. For example, the expert may assess the quality of tools by running them with a text corpus that is an homogeneous mix of

many text types as input, while the end-user may work with text types with peculiar characteristics, such as bad spelling, OCR-errors, social media jargon, sociolects, or transcribed speech.

- (c) The expert just does not think that very complex workflow designs are worth testing. The idea that simple, short workflow designs are to be preferred can steer a human expert towards suboptimal solutions.

The TT displays all viable workflow designs that fulfill the user's goal. There is little that can be done about this. The accumulation of metadata about the tools that together constitute a workflow design does not sum up to an overall quality assessment of the workflow design. The same tool can be considered a strong link in one workflow design, but a weak link in another one, so there is not a 'tool quality measure' that can be used in the computation of an 'workflow quality measure'. The only way to improve on this is to add metadata that encompasses two or more tools at the same time.

7 Coming Improvements

The TT has the potential to generate thousands of useful workflow designs. In order to make this an advantage over WMSes that require manual composition of workflow designs, it is necessary to guide users when they state their goal, so that they are not overwhelmed by too many designs to choose from. There are only a few drop down lists that the user has to consider, but specifying a goal can still be hard, since some of the values to choose from are relatively arcane.

The leading idea of the improvements we envision is that the specification of the goal can be done in easy steps. For each choice to be made, users will be guided by explanations and examples. Users who are already acquainted to the TT can skip this guidance, if they like.

Firstly, we want to find equivalents of the concepts that are used in the TT in the CLARIN Concept Registry (CCR). The TT can use the CCR descriptions for those concepts for which approved CCR equivalents exist. If we find that a candidate CCR concept would have been the perfect choice, we will communicate that finding to the Concept Registry Coordinators. If the TT needs a concept for which no CCR equivalent exist, we will try to make that a CCR candidate ('medieval', 'OCR', 'blackletter', 'pruned').

Secondly, we want to construct an assortment of output samples that we can show to users, so they can make better informed choices.

The TT is not able to compute workflow designs for all goal specifications that the user can make. Whether or not any workflow designs exist for a given input and goal can only be found out by letting the TT try to find those candidates. As a third improvement to the user experience, we need to inform users that an empty list of workflow designs is not a shortcoming and does not mean that there is a bug in the TT. An informed user whose goal cannot be fulfilled by the TT, might even appreciate that he or she did not have to spend hours on trying to manually compose a workflow design that cannot be composed, given the current set of available tools.

A fourth improvement will be the possibility to bookmark a workflow design. To bookmark a selected workflow design, the user will have to provide both a name and a reason for marking the workflow design, if possible comparing it with other already bookmarked workflow designs. Shared bookmarks make it easier to cooperate in a project and are useful as a reminder for future use. Bookmarks cannot be guaranteed to exist forever, however. Changes in the registered metadata of a tool can potentially invalidate a bookmarked workflow design. If that is the case, the bookmark must be marked obsolete and deleted after a grace period of for example two years. When a user has uploaded input to the TT, only bookmarked workflow designs will be shown that can be applied to that input.

An alternative to a bookmark is simply freezing and saving the workflow design itself. Technically, that would not be very different from bookmarking a workflow design, but conceptually there is a difference: a saved workflow design exists until someone deletes it. It can continue to exist, even if the TT has evolved so much that the saved workflow design no longer can be

executed. A bookmark is merely a pointer with some metadata attached to it. From their experience with bookmarks to websites, users do not expect that bookmarks to workflow designs can be used in all eternity. Since the TT is meant to be expanding and improving over time, bookmarks are more useful than frozen and saved workflow designs.

Researchers and students from the Faculty of Humanities at the University of Copenhagen will be involved in testing and improving the new interface to the TT.

8 Availability

There is an online version of the TT at the address <https://cst.dk/texton/> that can also be reached via the clarin.dk website (<https://clarin.dk/clarindk/tools-texton.jsp>). An identical version can be downloaded from GitHub⁶. Some of the tools in the GitHub repository are merely wrappers for open source tools that must be obtained separately, such as LibreOffice (which is used for conversion of office formats to RTF) and the OCR programs Cuneiform and Tesseract.

The TT can be installed on a personal computer, for example in the Windows Subsystem for Linux. Such an instance of the TT even runs while the computer is cut off from the internet and can then be used for handling very sensitive input, e.g. non-anonymized juridical documents.

9 Conclusion

We have in a few words described the current state of the Text Tonsorium, a Workflow Manager for NLP tools that has been re-launched to the CLARIN community. The Text Tonsorium automatically computes workflow designs that fulfil the user's goal and executes the ones that the user selects. To illustrate that the final, sometimes hard, decision which workflow design to use, has to be taken by the user and cannot be automated away, we have devoted a section to how a recent project's aims were backed by the Text Tonsorium, suggesting several workflow designs, each with their advantages and disadvantages. Finally, we discuss how we hope to improve the user experience of the Text Tonsorium. We would be pleased to receive more suggestions from users and testers in the CLARIN community.

References

- Tomaž Erjavec, Štefan Bruda, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Peter Holozan, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Igor Shevchenko, Kiril Simov, Lydia Sinapova, Han Steenwijk, Laszlo Tihanyi, Dan Tufiş, and Jean Véronis. 2010a. MULTEXT-east free lexicons 4.0. Slovenian language resource repository CLARIN.SI.
- Tomaž Erjavec, Ivan Derzhanski, Dagmar Divjak, Anna Feldman, Mikhail Kopotev, Natalia Kotsyba, Cvetana Krstev, Aleksandar Petrovski, Behrang QasemiZadeh, Adam Radziszewski, Serge Sharoff, Paul Sokolovsky, Duško Vitas, and Katerina Zdravkova. 2010b. MULTEXT-east non-commercial lexicons 4.0. Slovenian language resource repository CLARIN.SI.
- Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. Weblicht: Web-based LRT services for German. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pages 25–29. The Association for Computer Linguistics.
- Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 145–153. Association for Computational Linguistics.
- Bart Jongejan. 2013. Workflow management in CLARIN-DK. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT*

⁶<https://github.com/kuhumcst/texton>, <https://github.com/kuhumcst/texton-bin>, <https://github.com/kuhumcst/texton-linguistic-resources> and <https://github.com/kuhumcst/DK-ClarinTools>

Proceedings Series 20, number 89, pages 11–20. Linköping University Electronic Press; Linköpings universitet.

Bart Jongejan. 2016. Implementation of a workflow management system for non-expert users. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 101–108, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Lene Offersgaard, Bart Jongejan, and Bente Maegaard. 2011. How Danish users tried to answer the unaskable during implementation of clarin.dk. In *Proceedings of SDH*, November. SDH 2011- Supporting Digital Humanities.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udxpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

When Size Matters. Legal Perspective(s) on N-grams

Paweł Kamocki

Leibniz-Institut für Deutsche Sprache

Mannheim, Germany

kamocki@ids-mannheim.de

Abstract

N-grams are of utmost importance for modern linguistics and language technology. The legal status of n-grams, however, raises many practical questions. Traditionally, text snippets are considered copyrightable if they meet the originality criterion, but no clear indicators as to the minimum length of original snippets exist; moreover, the solutions adopted in some EU Member States (the paper cites German and French law as examples) are considerably different. Furthermore, recent developments in EU law (the CJEU's *Pelham* decision and the new right of press publishers) also provide interesting arguments in this debate. The paper presents the existing approaches to the legal protection of n-grams and tries to formulate some clear guidelines as to the length of n-grams that can be freely used and shared.

1 Introduction

N-grams are generally defined as sequences of n items from a sample of text. In the field of linguistics, these items can be letters, phonemes or syllables, but perhaps most importantly: words. In this paper, we will discuss n-grams in a narrowed-down sense, i.e. as sequences of n words.

The use of n-grams in language research can be traced back to Shannon (1948), or even further back to Markov (1913). Today, n-grams are fundamental for computational linguistics and language technology, and lists of n-grams are a valuable resource used especially, but not exclusively, in developing language models for Machine Translation purposes. The importance of n-grams is best illustrated by the popularity of Google N-gram Viewer (<https://books.google.com/ngrams>), launched in 2009, and by the fact that its use by researchers has become commonplace, despite its questionable quality and lack of metadata (Koplenig, 2017).

Many linguists attempt to compile their own re-usable lists of n-grams. In doing so, they are confronted with the question of legality. Well aware of the fact that copying and sharing of text in principle requires permission from the copyright holder, they are wondering if, and to what extent, this also applies to very short excerpts of text. Indeed, the question about the number of words below which an excerpt becomes copyright-free is among the questions that legal experts are asked the most by language researchers.

The only possible *in abstracto* answer to this question is unfortunately disappointing: while in general very short n-grams can be used and shared without consequences (at least from the copyright perspective), there are no clear rules as to the length of copyright-free n-grams. The decision on where to draw the line should be made for every project on a case-by-case basis, and depend on several parameters such as the language of source texts, their genre, the primary purpose for which the list n-grams will be re-used and - last but not least - the 'risk appetite' of the research team. This paper will hopefully provide some guidance on this issue.

Before our analysis can start, it should be noted that it only concerns situations where the use of source material is not based on a licence¹ (the licence may, e.g., allow sharing of long excerpts, or

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹For cases where the source material is, like this paper, available under a Creative Commons license, see Eckart de Castilho et al. (2018) for guidance.

prohibit any sharing whatsoever) — in such cases, the license should generally prevail.² Our analysis is also context-independent — in situations covered by statutory exceptions, such as citation or research exception, the rules will differ. Without entering into the intricacies of statutory exceptions in the many national laws of CLARIN members, it can generally be said that the citation exception allows even for long passages to be quoted, as long as the citation is included in an independent work (e.g. a research paper) and justified by its content. For example, it is justified to quote a set of sentences from a corpus to illustrate or disprove a hypothesis. When it comes to the research exception, it may allow compilation of lists of long n-grams, but these lists can subsequently only be shared within a research team, and they cannot be re-used for commercial purposes.

The guidelines in this paper are for all those who, rather than relying on a context-specific statutory exception, want to compile a list of n-grams that can be shared as an independent resource (not as part of a paper or an annex to a book) and in Open Access conditions, i.e. re-usable by anyone and for any purpose (including for commercial purposes).

2 The Traditional Approach, or ‘Originality, You Fool!’

In copyright law, the traditional approach is that parts of works (including literary works) are protected as long as they are themselves original. In this approach, a snippet is regarded as a work in its own right: if it is original, then its reproduction and communication to the public require authorisation of the rightholder, unless they are allowed by a statutory exception.

This approach was adopted *inter alia* by the Court of Justice of the European Union in the 2009 *Infopaq* case. In this seminal decision, the Court ruled: ‘it should be borne in mind that there is nothing in [EU law] indicating that [parts of works] are to be treated any differently from the work as a whole. It follows that they are protected by copyright since, as such, they share the originality of the whole work. (...) the various parts of a work thus enjoy protection under [copyright], provided that they contain elements which are the expression of the intellectual creation of the author of the work [i.e.: which are original (see below)] — PKJ’.³

Further analysis of the question requires a brief presentation of the notion of originality. Originality (sometimes also called ‘individuality’) is a fundamental concept in copyright law — only original works can be protected by copyright. However, since the scope of copyright is extremely broad, encompassing all sorts of human creations from an opera to a piece of software to a cartoon drawing, the notion of originality is necessarily very vague, and its application to a specific case often uncertain. It also used to differ considerably between jurisdictions and time periods: ‘original’ was interpreted as ‘originating from the author’, i.e. not copied (in the US), but also as showing a degree of ‘labour, skill and judgement’ (in England), or even ‘an imprint of the personality of the author’ (in France).

The above-mentioned judgement of the Court of Justice of the European Union in the *Infopaq* case was also the Court’s first attempt to harmonise the notion of originality across EU Member States (Rosati, 2013). The Court ruled that a work is original if it constitutes ‘the author’s own intellectual creation’: this definition existed in some EU Directives,⁴ but incidentally it is also the traditional definition of originality (*Individualität*) in German copyright law (*persönliche geistige Schöpfung*).⁵ The Court further elaborated on this definition in the *Painer* case,⁶ where it ruled that a work is an intellectual creation of the author if it reflects his personality and expresses his free and creative choices.

The possibility to exercise choice in the creative process is therefore a necessary (if not sufficient) condition of originality. One could then hypothesise that the choice of a single word can be sufficient to meet this requirement. The 2008 French Court of Cassation decision⁷ declaring a work consisting of a single word (*Paradis*) original could be quoted to support this statement. However, it should not be

²The question remains whether violation of a license that would prohibit sharing of short excerpts would constitute copyright infringement, or just breach of contract (which are quite different from legal perspective, with different consequences). The answer may vary between jurisdictions, however see: CJEU, judgement of 18 December 2019, case C-666/18 (IT Development vs. Free Mobile) stating that the breach of an IP clause of a software licence constitutes copyright infringement.

³CJEU, 16 July 2009, C-5/08 (*Infopaq*)

⁴Art. 6 of the Term Directive (2006/116/EC); Art. 1.3 of the Software Directive (2009/24/EC); Art. 1.3 of the Database Directive (96/9/EC)

⁵Section 2(2) of the German Copyright Act.

⁶CJEU, 1 December 2011, C-145/10 (*Painer*)

⁷Cour de cassation, 1 Civ., 13 novembre 2008, no. 06-19.021

forgotten that the word was not only written in a very specific font, but also - more importantly - placed in a very specific context: above the toilet of a mental hospital. In other words, the protection was not granted to the word 'Paradis' as a literary work, but rather to the whole setting, which constituted an artistic (and not literary) work.

In 2009, once again in the *Infopaq* ruling, the Court of Justice of the European Union seems to have definitively denied copyright protection of single words, stating that: 'considered in isolation, [words] are not as such an intellectual creation of the author who employs them. It is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation'.⁸

In this part of the ruling the Court referred to another approach to originality, which states that originality manifests itself in 'selection and arrangement' of various elements constituting a work. This definition formally concerns only compilations, but it can be applied more generally (after all, literary works for linguists are but compilations of words). Interestingly, this approach stems from a canonical US copyright case *Sarony* concerning a lithography of Oscar Wilde;⁹ today, it appears not only in the Berne Convention (Article 2.5), but also, as an alternative (selection OR arrangement), in the TRIPS Agreement (Article 10.2), the EU Database Directive (Article 3.1) and numerous national laws (e.g. Article L112-3 of the French Intellectual Property Code or Section 4 of the German Copyright Act). In its original form (as a conjunction), known from the Berne Convention or from the *Sarony* case, this criterion would imply that mere choice (selection) is not enough to constitute originality, and that another aspect - arrangement - is also necessary. In other words, the constitutive elements of a work not only have to be chosen by the author, but also placed in a particular order, which in the context of snippets would imply that two words are the absolute minimum for a snippet to be original.

It is indeed our opinion that in 'pure' international copyright (i.e., without taking into account national laws and case law) two words can be enough - still, only in extremely limited cases - to constitute an original work, or an original snippet. The position according to which two-word snippets can (in very rare cases) be original is also supported by Article 2.1 of the Berne Convention, according to which copyright protection is independent from 'the mode or form' in which the work is expressed, which arguably includes also very short forms. This same rule can also be found in Article L112-1 of the French Intellectual Property Code.

An example of a potentially original two-word sequence is 'krwista krówka' ('bloody (rare) little cow' in Polish). There are several reasons to consider this sequence original: first of all, both words bear some phonetic similarities (the /'kr-/ at the beginning, the /v/ in the middle, and the /a/ at the end). Secondly, the use of the adjective 'krwisty' (bloody or rare), usually associated with a cooked steak, to describe a little cow can be perceived as shocking or humorous, and is highly unusual. On top of this, 'krówka' is also the name of a traditional Polish fudge candy, which makes the association with the adjective 'krwisty' even more unusual. Moreover, this two-word sequence has 0 occurrences in the National Corpus of Polish (<http://www.nkjp.uni.lodz.pl>), which makes it highly probable that it was indeed created by its author, and not merely copied from another source.

Obviously, only a court can authoritatively decide whether 'krwista krówka' is indeed original or not, and admittedly a lot would depend on the talents of the attorneys representing both parties in the procedure. Nevertheless, it is the author's opinion that it could potentially qualify for protection. This leads to the conclusion that original 2-grams can exist, albeit they are extremely rare.

This, however, does not seem to be the position of French judges who relatively often find two-words combinations to be original. For example: *du rififi*¹⁰, *Charlie Hebdo*¹¹, *Bourreau d'enfants*¹² or

⁸CJEU, 16 July 2009, C-5/08 (*Infopaq*), no. 45.

⁹U.S. Supreme Court, 17 March 1884, *Burrow-Giles Lithographic Company v. Sarony*, 111 U.S. 53; in short, the Court ruled that although the lithographer did not create Oscar Wilde, his attire and the scenography of the famous lithograph, he did select these elements and arranged them in a particular way, thereby creating an original work.

¹⁰Cour d'appel de Paris, 4e ch., 24 janvier 1970, RTD com. 1971, p. 94, obs. H. Desbois

¹¹Cour d'appel de Paris, 4e ch., 25 octobre 1995, JurisData n° 1995-024506

¹²Tribunal de Grande Instance de Seine, 3e ch., 2 février 1960, RTD com. 1960, p. 844, obs. H. Desbois

*Paris Canaille*¹³ were declared protected by copyright as original titles.¹⁴ These decisions, however, are rather old. More recently, French courts denied copyright protection to such slogans as *Le marketing du désir* (desire marketing),¹⁵ or *Le permis libre* (Free/open permit),¹⁶ but admitted copyright protection of *À fond la forme*¹⁷ (roughly translatable as ‘Full fitness’, a play on words *fond* (content) and *forme* (form, physical shape)) or *Un nom pour un oui* (a name for a ‘yes’).¹⁸

German courts seem to be more demanding when it comes to originality of very short literary works. The shortest work declared copyright-protected by a German court that we have been able to identify is a four-word slogan *Ein Himmelbett im Handgepäck*¹⁹ (a canopy bed in hand luggage); however, the decision dates back to 1964. Modern German case law seems to generally refuse copyright protection of slogans and titles, which are deemed too short to constitute original works.²⁰ This is also the position in the United States, where the Copyright Office states that ‘*short phrases, such as names, titles, and slogans, are uncopyrightable because they contain an insufficient amount of authorship*’.²¹

So when does a literary work become ‘long enough’ to be considered for copyright protection according to modern standards? As stated above, there is no definitive answer, but some guidance has been provided by the CJEU in the above-mentioned *Infopaq* case. The Court ruled that snippets of 11 consecutive words can be original (although the evaluation of their actual originality was, of course, left to the national court).²² The 11-word *limes* resulted simply from the facts of the case (this was the length of snippets used by *Infopaq*, an early news aggregator service), and the decision should not be interpreted as meaning that 10-word or shorter snippets are free from copyright; it does, however, provide an argument in the discussion. The 10-word limit for snippets should, in our opinion, be considered as very liberal. The truth lies therefore somewhere between 2 and 10.

3 A Few Words About Trademarks, or ‘Some Rights Reserved’

A careful reader might be wondering — if an 1-gram turns out to be a registered trademark (like *Mercedes* or *CLARINS*), can it still be lawfully included in a list of n-grams and shared with the community?

This time the answer is a clear yes. It is so because the exclusive rights conferred by trademark law are in fact much more limited than those conferred by copyright. The copyright holder can in principle prevent others from accomplishing any acts of reproduction (copying) and communication to the public (sharing) of his work (unless they are expressly authorised by a statutory exception); the trademark holder, however, can only prevent others from using his trademark in the course of trade for the purposes of distinguishing goods or services.²³ The inclusion of an n-gram in a list, even if the list is then used for commercial purposes (e.g., developing a language model for a commercial Machine Translation service) is irrelevant from the point of view of trademark law.

In conclusion, there is no need to remove trademarks, or what appears to be trademarks, from lists of n-grams.

¹³Cour d’appel de Paris, 1er ch., 30 mai 1956, *Léo Ferré c/ Sté Océan Films et a.*, JCP G 1956, II, 9354

¹⁴Article L112-4 of the French Intellectual Property Code states that ‘*Le titre d’une oeuvre de l’esprit, dès lors qu’il présente un caractère original, est protégé comme l’oeuvre elle-même*’; however, it can be argued that titles are protected by *sui generis* copyright, which only restricts the use of an original title as a title of another work.

¹⁵Cour d’Appel de Paris, 7 novembre 2017

¹⁶TGI de Paris, 7 juillet 2016

¹⁷TGI Paris, 8 janvier 2002.

¹⁸CA Paris, Ch. 2, 17 juin 2011, RG n°10/12092.

¹⁹Oberlandesgericht Düsseldorf, 28 Februar 1964 – 2 U 76/63.

²⁰E.g. both *DEA – hier tanken Sie auf* (OLG Hamburg, Urteil vom 09.11.2000, Az. 3 U 79/99) and *Für das aufregendste Ereignis des Jahres* (OLG Frankfurt, Beschluss vom 04.08.1986, Az. 6 W 134/8) were denied copyright protection.

²¹US Copyright Office, Circular 33: Works Not Protected by Copyright.

²²CJEU, 16 July 2009, C-5/08 (*Infopaq*), no. 48.

²³Recital 18 of the Trademark Directive of 16 December 2015 (2015/2436); for more details, see also Article 10 of the same Directive.

4 The New Approach, or 'Name that Tune'

A new approach to 'reproduction in part' was adopted by the CJEU in a recent case *Pelham*.²⁴ The facts involved not a literary work, but a sound recording (phonogram) by Kraftwerk, a short part of which ('approximately two seconds rhythm sequence') was used as a sample in another recording. The relevant aspect of the case was decided not on the grounds of copyright, but on the grounds of a related right of phonogram producers, a legal framework that is independent from originality.

The CJEU ruled that the use of a short excerpt of a sound recording constitutes 'reproduction in part' (and therefore an act that in principle requires authorisation of the rightholder) if the excerpt is 'recognisable to the ear'.

Can this approach be applied to text snippets? Arguably, it would mean that n-grams can be freely reused as long as they are not *hapax legomena* (i.e., as long as they occurred independently in more than one text in the language, to the extent that this can be established), and therefore their exact source cannot be identified with certainty. This approach can be viewed as stricter than the one based on originality — a purely descriptive, banal paragraph (e.g., a relation from a football match) will at some point become a *hapax legomenon* if it is long enough, but it will still lack originality. However, the 'recognisability' approach has the advantage of being more objective, and therefore perhaps easier to apply *in abstracto*.

It also presents a practical advantage, as it is quite commonsensical: if the excerpt is not recognisable to the rightholder, then he will not sue for copyright infringement, and if it is not recognisable to the judge, its use will not be qualified as copyright infringement.

This approach may also be particularly appealing in compiling lists of n-grams to be used for training language models (e.g., for Machine Translation purposes). *Hapax legomena* should not be used for such purposes, as the resulting language model would lack the necessary context (Kamocki et al., 2016). Therefore, eliminating *hapax legomena* from such compilations of n-grams is not perceived as a constraint.

On the other hand, the approach also carries some risk of overgeneralisation, as it may lead to deleting n-grams that are certainly not copyright-protected. To illustrate, this claim, let us come back to our example from the previous section. The 2-gram discussed above as potentially original - '*krwista krówka*' - would also be a *hapax legomenon* in the National Corpus of Polish. As such, it should be deemed as non-reusable both in the 'originality' approach, and in the 'recognisability' approach. However, '*szczęśliwy Zenobiusz*' (*happy Zenobiusz* — Zenobiusz being a highly uncommon first name in Polish) is definitely not original, as there is nothing creative about the use of a basic adjective and a known (albeit uncommon) first name; and yet, this 2-gram also has no occurrences in the National Corpus of Polish. Therefore, one can hypothesise that it would be reusable under the 'originality' approach (as it is non-original), but not under the 'recognisability' theory (as its source would be easy to establish).

Another difficulty lies in the fact that it is indeed difficult to establish with enough certainty whether an n-gram is a *hapax legomenon* at the level of the entire language. It is possible (albeit rather unlikely) that the expression '*krwista krówka*' has never been uttered in Polish before (due to its creative, playful and humorous nature). It is, however, virtually impossible that the words '*szczęśliwy Zenobiusz*' have never been uttered, as there have been people named Zenobiusz, and at least some of them must have been described as happy at some point in their lives. Still, both expressions are absent from the National Corpus of Polish, which is the largest corpus of the language. Therefore, the fact that an expression does not appear in a corpus, even very large, can hardly be regarded as proof of its truly unique nature.

5 New Related Right of Press Publishers: A Trench War Has Begun

Article 15 of the new Directive 2019/790 on Copyright in the Digital Single Market (DSM Directive) introduced and harmonised a new related right of press publishers. The right protects said publishers against parasitic use of press articles by commercial news aggregators (Papadopoulou, Moustaka, 2020), and as such is not directly relevant for language research. However, this new right is only triggered when an online service uses more than 'individual words or very short extracts of a press publication'. It will therefore be necessary to define precisely, via case law or via a collective agreement, the maximum length of a 'very short extract'. In Germany, where this right was first introduced in

²⁴CJEU, 29 July 2019, C-476/17 (*Pelham*)

2013 (before the DSM Directive), the Patent and Trade Mark Office initially (2015) recommended 7 consecutive words as a freely reusable ‘very short extract’ of a press publication. However, this recommendation no longer appears on the Office’s website.²⁵

In Germany, the debate on this question has recently been revived by the publication of a series of government bills on the implementation of the DSM Directive.²⁶ The first bill did not define the maximum length of a ‘very short extract’, but instead stated that headlines (*Überschriften*), which can sometimes be quite long (even longer than 7 words), should be considered as such. In the consultation process, three German associations of press publishers (BDZV, VDZ and VDL) issued a joint statement starkly disapproving the inclusion of headlines in the definition of ‘very short extracts’, and arguing that the freely-reusable extracts should be limited to three consecutive words (BDZV et al., 2020).

Although made in a specific context, very different from language research, the statement may be interpreted as indicating that German press publishers will generally not oppose the use of 3-grams extracted from their articles, as they consider them of little value, at least from the commercial standpoint. If press publishers are ready to let news aggregators use three-word snippets, *a fortiori* they should let language researchers do at least as much.

It should be noted here that in German a 3-gram (not to mention a 7-gram) can carry quite a lot of information, as this language uses compound nouns and is known for its unique ability to convey complex meanings in single words (e.g. *Schadenfreude*²⁷ or *Futterneid*,²⁸ to quote the most common examples). It is therefore imaginable that e.g. English- or French-language publishers would be ready to accept free use of slightly longer extracts.

Unfortunately, the most recent developments seem to indicate that there will be no clear answer to the question of the exact length of ‘very short extracts’ under the right of press publishers. The German government has recently published another bill on the same issue, which regrettably does not contain any definition of ‘very short extracts’. In France, Google and an association of press publishers signed an agreement on the application of this new right;²⁹ while it is likely that it defines the minimum length of a freely re-usable snippet, its content has not yet been made public (as of 7 February 2021).

6 Conclusion

As promised in the introduction, we will try to formulate some guidelines concerning the use of short snippets of text without permission from the rightholders:

- 3-grams should be regarded as generally free from copyright.³⁰ Only very exceptionally and only in some jurisdictions expressions of 3 words or shorter can be found original — even then, such very short original expressions seem to occur almost exclusively in very specific contexts, as titles, slogans or in poetry. Even in very large corpora, original 3-grams are likely to be *hapax legomena*; therefore, eliminating *hapax legomena* from the list of n-grams (which

²⁵But cf. the 2015 news report at: <https://www.internet-law.de/2015/09/leistungsschutz-recht-dpma-schlaegt-einigung-vor.html> (retrieved 4 September 2020).

²⁶*Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz — Entwurf eines Ersten Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts, 15 Januar 2020*, available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/DiskE_Anpassung%20Urheberrecht%20digitaler%20Binnenmarkt.pdf?blob=publicationFile&v=1 (retrieved 7 February 2021), followed by *Gesetzentwurf der Bundesregierung — Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, 3 Februar 2021*, available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Gesetz_Anpassung_Urheberrecht_digitaler_Binnenmarkt.pdf?blob=publicationFile&v=5 (retrieved 7 February 2021).

²⁷‘Enjoyment obtained from the troubles of others’ (definition by Merriam-Webster Online Dictionary)

²⁸Literally ‘food envy’, used to describe a feeling of jealousy towards someone who, while eating at the same restaurant, ordered something that looks more appetising than our meal.

²⁹L’Alliance de la Presse d’Information Générale et Google France signent un accord relatif à l’utilisation des publications de presse en ligne, *Le blog officiel de Google France*, 21 janvier 2021, <https://france.-googleblog.com/2021/01/APIG-Google.html> (retrieved 7 February 2021).

³⁰Interestingly, Linden (2014) seems to have reached the exact same conclusion.

is sometimes desirable for linguistic purposes) substantially increases legal certainty about the possibility of its re-use;

- the use of 7-grams, while not risk-free, may be seen by many as a reasonable compromise. In order to mitigate the associated risk, one might attempt to reduce the ‘recognisability’ of text snippets by removing the elements that are ‘highly identifying’, such as *hapax legomena*, proper names and other named entities, or very unusual syntactic structures (e.g., ‘Yoda-Speak’);
- the use of 10-grams may be justified under a very liberal interpretation of the *Infopaq* case; in our opinion, however, it would carry significant risk of copyright infringement.

The guidelines presented above only apply if the use of the data is not based on a license (in which case the license should take precedence), or is not covered by a statutory exception (e.g. citation or research exception).

References

- BDZV (Bundesverband Deutscher Zeitungsverleger Verband), Deutscher Zeitschriftenverleger Verband, Deutscher Lokalzeitungen. 2020. *Stellungnahme zum Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz für ein Erstes Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts vom 16. Januar 2020*.
- Eckart de Castilho, Richard, Giulia Dore, Thomas Margoni, Penny Labropoulou and Iryna Gurevych. 2018. A Legal Perspective on Training Models for Natural Language Processing. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Kamocki, Paweł, Jim O’Regan and Marc Stauch. 2016. All Your Data Are Belong to us. European Perspectives on Privacy Issues in ‘Free’ Online Machine Translation Services. In: David Aspinall, Jan Camenisch, Marit Hansen, Simone Fischer-Hübner, Charles Raab [Eds.]. *Privacy and Identity Management. Time for a Revolution?* Springer International Publishing.
- Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169-188.
- Linden, Krister. 2014. Update from language resources. Oral presentation at the Legal Issues in Language Resources and Infrastructures Workshop. LREC 2014.
- Markov, A.A. 1913. Essai d’une recherche statistique sur le texte du roman “Eugène Onëgin”, illustrant la liaison des épreuves en chaîne. *Bulletin de l’Académie Impériale des Sciences de St.-Pétersbourg. VI série*, 7 (3): 153–162.
- Papadopoulou, Maria-Daphne and Evanthia-Maria Moustaka. 2020. Copyright and the Press Publishers Right on the Internet: Evolutions and Perspectives. In: Tatiana-Eleni Synodinou, Philippe Jougoux, Christiana Markou, Thalia Prastitou [Eds.]. *EU Internet Law in the Digital Era*. Springer: Cham.
- Rosati, Eleonora. 2013. *Originality in EU Copyright: Full Harmonization Through Case Law*. Edward Elgar Publishing: Cheltenham, Northampton.
- Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal*, 27 (3): 379–423.

Sharing is Caring: a Legal Perspective on Sharing Language Data Containing Personal Data and the Division of Liability between Researchers and Research Organisations

Aleksei Kelli
University of Tartu,
Estonia
aleksei.kelli@ut.ee

Pawel Kamocki
IDS Mannheim,
Germany
pawel.kamocki@gmail.com

Gaabriel Tavits
University of Tartu,
Estonia
gaabriel.tavits@ut.ee

Irene Kull
University of Tartu,
Estonia
irene.kull@ut.ee

Andres Vutt
University of Tartu,
Estonia
andres.vutt@ut.ee

Krister Lindén
University of Helsinki,
Finland
krister.linden@
helsinki.fi

Arvi Tavast
Institute of the
Estonian Language,
Estonia
arvi@tavast.ee

Mari Keskküla
University of Tartu,
Estonia
mari.keskkula@
gmail.com

Age Värvi
University of Tartu,
Estonia
age.varv@ut.ee

Silvia Calamai
University of Siena
Italy
silvia.calamai@
unisi.it

Kadri Vider
University of Tartu,
Estonia
kadri.vider@ut.ee

Ramūnas Birštonas
Vilnius University,
Lithuania
ramunas.birstonas@
tf.vu.lt

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Merle Erikson
University of Tartu,
Estonia
merle.erikson@ut.ee

Abstract

The article focuses on determining responsible parties and the division of potential liability arising from sharing language data (LD) containing personal data (PD). A key issue here is to identify who has to make sure and guarantee the GDPR compliance. The authors aim to answer 1) whether an individual researcher is a controller and 2) whether sharing LD results in joint controllership or separate controllership (whether the data's transferee becomes the controller, the joint controller or the processor). The article also analyses the legal relations of parties involved in data sharing and potential liability. The final section outlines data sharing in the CLARIN context. The analysis serves as a preliminary analytical background for redesigning the CLARIN contractual framework for sharing data.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. License details: <http://creativecommons.org/licenses/by/4.0/>

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Arvi Tavast, Ramūnas Birštonas, Gaabriel Tavits, Mari Keskküla, Penny Labropoulou, Irene Kull, Age Värvi, Merle Erikson, Andres Vutt and Silvia Calamai 2021. Sharing is Caring: a Legal Perspective on Sharing Language Data Containing Personal Data and the Division of Liability between Researchers and Research Organisations. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 129–147.

1 Introduction

The article focuses on determining responsible parties and the division of potential liability arising from sharing language data (LD) containing personal data (PD). The General Data Protection Regulation (GDPR) defines personal data as "*any information relating to an identified or identifiable natural person ('data subject')*" (Art. 4 (1)). Sharing personal data constitutes its processing¹ which has to follow the GDPR.²

A key issue here is identifying who has to guarantee the GDPR compliance and assumes liability for a GDPR violation. The controller and the processor are named as responsible parties (GDPR Art. 4).³ The GDPR defines the controller as "*the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data*" (Art. 4 (7)). The processor is defined as "*a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller*" (Art. 4 (8)).

The authors place this question in the context of language research. The following practical questions are addressed: 1) how to differentiate between obligations of the research organisation and individual researchers (e.g. whether an individual researcher is a controller) and; 2) whether sharing LD results in joint controllership or separate controllership (whether the transferee of the data becomes the controller, the joint controller or the processor).

In the first part after the introduction, the authors outline the general principles explaining the relationship between the organisation (e.g. the university) and its employees (researchers). Short case-studies serve as practical examples of how the university-researcher relationship is addressed in European countries. The authors' profile and expertise determined the choice of countries.

In the second part of the article, the authors focus on the legal relations of parties involved in sharing language data. It is essential to distinguish a situation when both parties are controllers (incl. joint controllers) and a situation when one party is the controller, and the other party is the processor (assists the controller in a limited way).

The third part is dedicated to liability and aims to describe what happens if the GDPR requirements are violated. In the last part, sharing language data within the CLARIN framework is briefly analysed.

The article continues and draws on previous research exploring the intersection of language research and personal data protection. The authors started their research by exploring the concept of personal data and analysing legal bases for its processing in the field of language technology (see Kelli et al. 2019; Klavan et al. 2018; Lindén et al. 2020). In this article, the authors with diverse backgrounds address the issue of responsible parties and liability arising from a violation of personal data laws. The paper serves as a preliminary conceptual analysis behind redrafting the CLARIN contractual framework for data sharing (for previous discussions on the contractual framework, see Kelli et al. 2015; Kelli et al. 2018).

Due to the specific focus and limited space, the article does not address the transfer of PD outside the EU, which concerns special provisions of the GDPR and EU case law (e.g. C-311/18).

2 Duty-bearers for the GDPR Compliance within Research Settings

The key feature of the controller is the **determination** of the **purposes** and **means** of the processing of PD. According to WP29 "*Determination of the 'means' therefore includes both technical and organisational questions where the decision can be well delegated to processors (as, e.g. 'which hardware or software shall be used?') and essential elements which are traditionally and inherently reserved to the determination of the controller, such as 'which data shall be processed?', 'for how long shall they be processed?', 'who shall have access to them?', and so on*" (2010: 14).

¹ Article 4 (2) of the General Data Protection Regulation (GDPR) defines processing extensively so that it covers any operation which is performed on personal data (e.g. collection, storage, alteration and sharing).

² For the sake of clarity, not all language data (LD) contains personal data (PD). Furthermore, even if LD has PD, it is recommendable to make it anonymous (if possible) since the GDPR does not apply to anonymous data (GDPR Rec. 26).

Anonymous data can be freely shared as long as PD protection is concerned. However, the authors focus on the cases when language data contains personal data. Language data or data within the context of this paper refers to language data that has personal data. For further discussion on anonymization see WP29 2014; Guidance 2019; ICO 2012; Sartor 2018.

³ When it comes to liability, then as a general rule, any person who has suffered damage as a result of a GDPR violation has the right to receive compensation from the controller or processor (GDPR Art. 82 (1)).

A relevant issue for the research community is to analyse whether an individual researcher and/or the university is the controller. The research setting is a unique context since academic freedom is defined as a fundamental right.⁴ This often gives researchers more freedom compared with other employees.

Considering relations between companies and their employees WP29 (2010: 15) indicates that *"preference should be given to consider as a controller the company or body as such rather than a specific person within the company or the body. It is the company or the body which shall be considered ultimately responsible for data processing and the obligations stemming from data protection legislation"*. According to the European Commission (EC) *"if your company/organisation decides 'why' and 'how' the personal data should be processed, it is the data controller. Employees processing personal data within your organisation do so to fulfil your tasks as a data controller"*. DP Handbook (2018: 102) similarly asserts that the legal entity (not its employees) is the controller.

WP29 (2010: 6) further explains that *"Controller and processor and their staff are therefore considered as the 'inner circle of data processing' and are not covered by special provisions on third parties"*. This is compatible with employment law. The employment law literature and case law also set forth that one of the characteristics of an employee is the fact that the employee is merged with the employer's team of other employees. The employee is employed and acts within the employer's economic activities (Risak & Dullinger 2018; C-22/98 para 26; C-66/85; C-256/01).

Although individual researchers conduct research with some freedom, their work is coordinated by the university. The university is responsible for the GDPR violations, and it has to implement appropriate technical and organisational measures (incl. data protection policies) to ensure the protection of PD (GDPR Art. 24) and to maintain a record of processing activities (Art. 30 GDPR).⁵ The question is whether academics could be considered the controller as well. The issue is not settled yet. However, European Parliamentary Research Service in its study concerning research (EPRS study 2019: 34) suggests that *"researchers and universities should assume that when processing personal data, their activities render them data controllers"*.

Furthermore, it is worth noting that controllership is an element of fact. Any contractual or other arrangements made by the interested parties (e.g. mere designation of X as the controller or a contract assigning controllership to B, whereas A determines the means and purposes of processing) is not binding on the data subject or data protection authorities.⁶

The approaches of different countries are analysed in alphabetical order to outline a possible model and to understand whether there is a common ground for a unified approach within CLARIN.

2.1 Estonia

According to the Organisation of Research and Development Act (ORDA), a research institution is a legal person or an institution in the case of which the principal activity is carrying out fundamental research, applied research or development, or several of the aforementioned activities (§ 3 (1) clause 1). The activities of Estonian public universities are regulated by special acts explicitly established for them. In this subsection, the authors rely on the example of the University of Tartu (UT) in explaining the responsibilities of a researcher and a research institution.

Under the University of Tartu Act (UTA), UT is a legal person in public law (§ 2 (2)). Despite academic freedom, academic staff do not have a special status. Under the Higher Education Act (HEA), the employment relationships of academic staff are regulated by an employment contract (§ 34). Thus, the division of responsibilities between a research institution and a researcher concerning the processing of personal data for research purposes can be based on the general approach to protecting personal data, according to which the employer is the responsible person (i.e. the employer is the controller).

⁴ According to the EU Charter of Fundamental Rights *"The arts and scientific research shall be free of constraint. Academic freedom shall be respected"* (Art. 13).

⁵ There might be a practical problem when an individual researcher collects data himself and the host university is unwilling to be the controller. One approach could be that if the host university of a researcher is unwilling to become the controller, the CLARIN Centre needs to do a due diligence to make sure that the data has been properly collected. If the due diligence is done correctly, the host of the CLARIN Centre may as well become co-controller of the data set.

⁶ WP29 (2010: 9): *"...even though the designation of a party as data controller or processor in a contract may reveal relevant information regarding the legal status of this party, such contractual designation is nonetheless not decisive in determining its actual status, which must be based on concrete circumstances"*.

As the controller, the research institution must implement appropriate technical and organisational measures (including a data protection policy) to ensure the protection of personal data (GDPR Art. 24 (1)) and to keep records of processing operations (GDPR Art. 30). UT has followed this approach. To ensure the GDPR compliance, UT has adopted the Data Protection Policy that explains the processing of PD and information concerning the privacy of individuals and several internal guidelines on the processing of personal data, including in research.⁷ This shows that the university is the controller.⁸ The university must introduce the instructions on data processing to the employee processing PD and demand that they are complied with.

If the employee (incl. researcher, professor and other academic employees) violates personal data rights, the university becomes liable. The data subject can claim damages, termination of the violation and so forth. Simultaneously, the employee (researcher) has breached his/her duties and can be held liable by the university.⁹ This means that if an employee does not comply with the rules on the protection of PD in force at the university, it is a breach of duty, for which the employee may be warned (that his employment contract is terminated if the violation is repeated) or his employment contract may be terminated. However, if the data subject has filed a claim for damages against the university due to non-compliance with the instructions given regarding the processing of personal data, the university may recover damages from the employee by way of recourse.

There could be cases when the researcher does not have an employment relationship with the research institution. Research work is carried out based on a contract of mandate or contract for services, and a doctoral student (as well as a master's student) who has the status of a student may also participate in the research. Unlike an employee, neither a mandatary and contractor nor a doctoral student is integrated in the research institution through subordination. However, in these cases, the research institution is also the research coordinator and the place of research. If the researcher works on the basis of a contract of mandate or contract for services or is a student and the research institution performs functions that are specific to the controller, the research institution is responsible for the processing of personal data. Considering the particular nature of the research and the academic freedom of the researcher, sometimes a researcher acting on a contract of mandate or contract for services or a student and a research institution may also be co-controllers who must enter into the corresponding agreement and inform the data subject about that agreement (GDPR Art. 26).

2.2 Finland

Universities in Finland are corporations under public law since 2009, at which time all employees, including professors, became regular employees. A professor and related teaching staff can be laid off if the university decides that there is no need for a particular discipline, e.g. no students are applying, or the university wishes to profile the university in a specific direction by deselecting a particular discipline.

The universities operate on a mandate to do teaching and research in the public interest.¹⁰ They are funded based on annual negotiations with the government and by external funding for research. Despite freedom of research being granted in the Universities Act¹¹ and the Constitution of Finland, researchers

⁷ Control over the processing of personal data is also subordinated to UT through the system of the ethics committee. The requirement of the ethics committee arises from the Personal Data Protection Act (PDPA), according to which the prior control of the ethics committee is required for the processing of special types of personal data for scientific purposes (PDPA §6(4)).

⁸ UT also has a data protection specialist, which refers to UT's status as the controller.

⁹ In exceptional cases, a person working in an enterprise or institution on the basis of an employment contract may be liable to the data subject as the controller pursuant to Art. 82 (1) of the GDPR. The Data Protection Working Party considers that *"the one liable for a data protection breach is always the controller, i.e. the legal person (company or public body) or the natural person as formally identified according to the criteria of the Directive. If a natural person working within a company or public body uses data for his or her own purposes, outside the activities of the company, this person shall be considered as a de facto controller and will be liable as such"* (WP29 2010: 17).

¹⁰ Universities Act of Finland (§2 (1)): *"The mission of the universities is to promote independent academic research as well as academic and artistic education, to provide research-based higher education and to educate students to serve their country and humanity at large"*.

¹¹ Universities Act of Finland (§ 6 (1)): *"While universities enjoy freedom of research, art and teaching, teachers must comply with the statutes and regulations issued on education and teaching arrangements"*.

as employees of an established research organisation acting on behalf of the organisation are not personally responsible for research activities sanctioned by the organisation as long as they adhere to organisational rules and regulations.

Even if a university relies on the researchers to specify the purpose of their research activities and determine the means for their data processing activities, the university still controls the activity with internal regulations, e.g. concerning data protection. The university usually becomes the controller of personal data involved in the research, although the researcher may remain a co-controller.

The problematic cases are grant-funded researchers who do not have formal employment at a research institution. Such individuals may enter into an agreement with a university for access to research facilities at a favourable cost. In this case, if personal data is collected for research purposes, and the university agrees to assist in the process, the university becomes a co-controller. However, an independent grant-funded researcher may opt to remain the sole controller.

In Finland, many universities help their researchers with guidelines and templates on how to formulate a privacy notice and a record of processing when collecting research material containing personal data (e.g. University of Helsinki, University of Jyväskylä, Aalto University).

2.3 France and Germany

The analysis is likely to be different in countries (like Germany or France) where university professors are not employees of the university, but public servants (*fonctionnaires*, *Beamten*) appointed for life and independent in the exercise of their missions (not unlike, e.g. judges). In both countries, this status is seen as a fundamental guarantee of freedom of academic research. It is due to this independence of researchers that in both Germany and France, e.g. copyright in the works created by academics in principle belongs to them and not to their university or institution.¹² In light of this rule (referred to as '*professors' privilege*'), it may be hard to argue that despite the fact that professors can reap benefits of their work (precisely because they are free to decide how to do it), it is the university that should bear the responsibility for how researchers process personal data.

Along these lines, the French National Centre for Academic Research (CNRS), in its guide on data processing for research purposes (CNRS guide 2019: 12), defines the director of a unit (an individual, not an institution) as the data controller. Then, the director of a unit designates the CNRS' Data Protection Officer (DPO) as 'his' DPO. This practice is in line with Article 37(2) of the GDPR, according to which "[a] group of undertakings may appoint a single data protection officer provided that a data protection officer is easily accessible from each establishment". Once designated, the DPO should be involved "properly and in a timely manner, in all issues which relate to the protection of personal data" (GDPR Art. 38 (1)). Having a common DPO for the whole institution is, therefore, an interesting way of providing for 'bottom-up standardisation' of data processing practice. However, the responsibility *stricto sensu* remains decentralised, as the DPO is not personally responsible for data processing. According to the principle of accountability, the responsibility always remains with the controllers.

This interesting approach in the CNRS is not necessarily shared by all French research institutions. Guidelines issued jointly by several institutions from the Paris region clearly state that in a research project, the researcher is a data controller, as long as he or she determines the means and purposes of processing (French University Guidelines 2019: 9).

In Germany, the situation is no less complex. The University of Cologne openly admits on its website that there are two confronting views regarding the data controllership, attributing the controllership either to the university, or to the researcher and that the university subscribes to the latter (University of Cologne, point 3). In an article published in the *Frankfurter Allgemeine Zeitung*, a leading German scholar on data protection suggests that – precisely for the reasons explained above, i.e. the freedom of research – it would be excessive to assign the liability for data processing to the university, and that it should be shared between the university and the researcher (as joint controllers) (Schwartzmann, 2019).

¹² The solution is likely to differ when it comes to patents – e.g. in Germany, the professors' privilege in patent law was abolished in 2002, and the rights to a patentable invention developed by an academic now belong to his or her institution. Since in the field of language technologies university patents remain rather exceptional, we believe that in the context of this paper an analogy with copyright is more accurate.

This approach would imply that universities should conclude agreements with their researchers to determine their respective responsibilities for compliance with the obligations under the GDPR (Art. 26). There is no requirement that these responsibilities should be shared equally.

2.4 Greece

Research in Greece is performed at universities and specific research institutes. They are legal entities governed by public or private law and supervised by the General Secretariat of Research and Technology (Ministry of Education). The employment of the respective personnel (academic personnel/professors at universities and research personnel at research institutes) is stipulated in two different legal acts (SOQI and RTDI). There are three ranks at which professors and researchers can be hired. They are elected for three years for the lower rank, while the two upper ranks are associated with permanent tenure. Universities and research institutes can also hire scientific and technical personnel to conduct research with private contracts of a restricted or indefinite term. The conditions under which research can be conducted are stipulated in the law on "Research, Technological Development and Innovation" (RTDI).

Chapter E ("Committees for Research Ethics") of the Act 4521/2018 (Act UWA) implements the GDPR in the context of research. The Act stipulates the constitution of a committee for research ethics at all universities and research institutes. According to Sec. 2, one of the objectives of the committees for research ethics is to *"control whether a research project is conducted respecting the value of human beings, [...] their private life and personal data..."*. Before the beginning of a funded project that includes research related to human beings, scientific coordinators must submit before the committee an application that *"includes a questionnaire and a short report on the adequacy and compliance of the project with the current law"*. Further specifications on the application and evaluation procedure and required documents are included in the Regulation of Principles and Operation of the Committee of each institute.

A survey of the regulations of research ethics committees of various academic and research institutes¹³ indicates the adoption of the same policy across them: the institute is acknowledged as the *"entity responsible for the processing of PD"* (the data controller), but the scientific coordinators and all researchers involved in the processing of PD are also held accountable for the processing (joint controllers). They must make sure that they comply with the GDPR and take appropriate measures to safeguard PD throughout the whole procedure (e.g. obtain the data subjects' consent, use (pseudo-) anonymisation for the published data). Failure to comply with this obligation may result in administrative measures such as the project's termination, the reparation of damages, remunerations of affected subjects, and even the discharge from their positions.

In general, professors, researchers, and employees at the institute are bound by these regulations as a result of their professional affiliation. Also, the scientific coordinators must sign a form stating that they are aware of the institute's Code of Ethics, that they will conform to it, and that no changes will be made to the project as described in the application. If any changes are required, an application must be resubmitted. Individuals engaged specifically for the project with a special contract (contract of services) must sign an additional contract of terms and conditions for the processing of PD.

Simultaneously, the committees for research ethics provide their assistance and guidance to the researchers whenever required. Ready-to-use forms and templates are available to researchers.

2.5 Italy

Italian universities can be either public or private bodies. According to the Constitution of the Italian Republic (Art. 33), every university acts on the principles of independence and responsibility. State-run universities of Italy are under the supervision of the Italian's Ministry of University. Independence means that every university establishes self-government through its competent bodies (its faculty). Due

¹³ The survey could not be extensive due to the fact that the Committees of some institutes are still in the process of writing up or updating these regulations. We note here the following: Aristotle University of Thessaloniki [AUTH], Athens University of Economics and Business [AUEB], Foundation for Research and Technology-Hellas [FORTH], University of Crete [UoC], University of Macedonia [UoM], University of Peloponnese [UoP], University of Patras [UoPa], University of Thrace [UoT], University of West Attica [UWA], National Center of Social Research [NCSR].

to the autonomy of the university, researchers and professors working in a public university are considered employees and not public servants/officials.

When the university employee (researcher, professor, temporary research staff) collects research material containing PD, the status of the controller is assumed by his/her university (as a legal entity). The individual scholar is designated as authorised for the treatment, as required by the law in force.

The Personal Data Protection Code (PDPC) sets forward the following principles (Section 2-o):

1. The controller or processor may provide under their responsibility and within the framework of the respective organisation that specific tasks and functions relating to the processing of personal data be allocated to expressly designated natural persons acting under the controller's or processor's authority.

2. The controller or processor shall set out the most appropriate arrangements to authorise the persons acting under their authority to process personal data.

However, the individual researcher cannot be designated as responsible for the processing since this role can only be covered by external subjects who under a contractual arrangement 'act' on behalf of the controller (Art 28 GDPR).

In 2017, the Association of general managers of Italian universities (*Convegno dei Direttori generali delle Amministrazioni Universitarie – CoDAU*) introduced guidelines for personal data processing (CoDAU guidelines). At the same time, the Italian association of Italian University Rectors (CRUI) prepared an internal draft regulation for Italian universities (CRUI regulations), which was last updated in January 2019.

Several universities, starting from CRUI regulations, published their own rules of procedures for processing personal data (e.g. Turin University, Modena and Reggio Emilia University, Bari University).

2.6 Lithuania

In Lithuania, the situation regarding the status of an individual researcher is not self-evident. For example, Vilnius University, the biggest and leading university in Lithuania, has adopted (2018) the Description of the procedure for PD processing at the University of Vilnius (the Description). The Description is silent on the issue if and how the status of the controller is divided between the university and the researcher. This question is not settled in the legal practice nor the legal doctrine.

In such a situation, the general norms and their interpretations should apply. From the outset, it should be noted that the position of professors and other researchers in Lithuanian universities is quite different compared with their counterparts in Germany. Lithuanian researchers have no special status and are regarded as regular employees. The general approach is that employees are not normally deemed the data controllers (DP Handbook 2018: 102).

Paragraph 4 of the Description also indicates that the university is the controller of all data collected during the university activities and internal administration processes as well as the controller of personal data transferred by data subjects and third parties.¹⁴ This provision is further elaborated by paragraph 2 stating that the Description is applicable and compulsory to the data controller, *i.e.* the University of Vilnius and all University employees who are processing personal data in the course of work.

Thus, if professors and other researches are acting in the course of their academic duties (which are specified in their labour contracts, descriptions of positions, other general or local regulations, universities' programs and projects in which researchers participate and similar documents), they should not be regarded as data controllers, but just the employees of the controller (*i.e.* the university). While it is true that researchers have certain discretion while conducting their research, still the general rule is that the overarching goals of the research activities are set by the university and, by signing the employment contract and taking up his/her position, the researcher simply acts on behalf of the university.

Researchers are deemed controllers when they are not acting on behalf of the university. In other words, they act outside their employment duties and set their research purposes and means independently. However, in this scenario, they can no longer be considered university researchers.

¹⁴ The same rule could be found in the earlier document (see Rules for processing PD 2015).

3 Legal Relations of Parties Involved in Sharing Language Data

The determination of a legal basis for processing (incl. sharing) language data is a key issue for language research. The suitable legal grounds could be the data subject's consent, public interest research or legitimate interest (GDPR Art. 6 (1) (a), (e), (f)).

The consent of the data subject should guarantee high-level protection of the data subject's rights and freedoms. However, this is not always possible (e.g. data was collected a long time ago, and there are no contact details). At the same time, the GDPR does not say that one legal ground is to be preferred over the others. Therefore, all suitable legal grounds (consent, public and/or legitimate interest) are equally applicable.

Since the issue has been previously studied (see, Lindén et al. 2019; Kelli et al. 2019) and due to the specific focus of the article, legal grounds are not further analysed here, and attention is given to legal relations of parties involved in data sharing.

3.1 Legal Relations between the Data Controllers

A key issue here is how much freedom parties have in determining who has which obligations under the GDPR. The European Data Protection Board (EDPB) has indicated that the concepts of the controller, joint controller and processor play a crucial role in the application of the GDPR since they determine who shall be responsible for compliance with different data protection rules and how data subjects can exercise their rights in practice (EDPB 2020: 3). According to WP29, *"Being a controller is primarily the consequence of the factual circumstance that an entity has chosen to process personal data for its purposes"* (2010: 8). WP29 clarifies further that the control could originate from the factual influence and the assessment of contractual relations is helpful since relevant actors often see themselves as facilitators rather than controllers. The contractual terms, however, are not decisive (2010: 11). EDPB (2020: 3) adds that the controller is a body that decides certain key elements of the processing, controllership may be defined by law or may stem from an analysis of the factual elements or circumstances of the case, and certain processing activities can be seen as naturally attached to the role of an entity (an employer to employees, a publisher to subscribers or an association to its members).

WP29 (10/2006) found that an entity (SWIFT) was a controller despite presenting itself as a processor based on a functional influence test. This demonstrates that any designation of controller/processor which does not correspond to the facts is void, and the actual situation is decisive, not the contract. The controller is who factually determines the purposes for which the personal data are processed.

When sharing language data for scientific purposes, it can be assumed that both parties (the party sharing data and the recipient) are acting as the controllers. *"Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers"* (GDPR Art. 26 (1)). EDPB (2020: 3) explains that joint participation can take the form of a common decision taken by two or more entities or result from converging decisions by two or more entities, where the decisions complement each other, and they have a tangible impact on the determination of the purposes and means of the processing. EDPB (2020: 3) indicates, *"An important criterion is that the processing would not be possible without both parties' participation in the sense that the processing by each party is inseparable, i.e. inextricably linked"*. EDPB (2020: 21) has issued an example of joint controllership, which allows drawing parallels to assess the relationship between institutions sharing of language data: *"Several research institutes decide to participate in a specific joint research project and to use to that end the existing platform of one of the institutes involved in the project. Each institute feeds personal data it holds into the platform for the purpose of the joint research and uses the data provided by others through the platform for carrying out the research. In this case, all institutes qualify as joint controllers for the personal data processing that is done by storing and disclosing information from this platform since they have decided together the purpose of the processing and the means to be used (the existing platform). Each of the institutes, however, is a separate controller for any other processing that may be carried out outside the platform for their respective purposes"*.

EDPB (2020: 20) also notes that *"the use of a common data processing system or infrastructure will not in all cases lead to qualify the parties involved as joint controllers, in particular where the processing they carry out is separable and could be performed by one party without intervention from the other"*.

In the case of a joint controllership, a transparent arrangement between the joint controllers must be agreed upon to comply with the GDPR (Art. 26 (1)). The essence of this arrangement should be made available to the data subjects (Art. 26 (2)). WP29 (2010: 24) explains it as follows *"Parties acting jointly have a certain degree of flexibility in distributing and allocating obligations and responsibilities among them, as long as they ensure full compliance"*. The controller's responsibilities must be clearly defined in accordance with actual data processing. The arrangement must reflect the respective roles and relationships of the joint controllers *vis-à-vis* the data subjects (GDPR Art. 26 (2)). EDPB (2020: 41) adds, *"It should be made clear here that all responsibilities have to be allocated according to the factual circumstances in order to achieve an operative agreement"*. Joint controllers need to define who is in charge of answering requests of data subjects, providing needed information and fulfilling lawfully the requests of data subjects ("right to be forgotten", etc.). They need to ensure that the whole processing fully complies with the GDPR (EDPB 2020: 41). Otherwise, as indicated by WP29 (2010: 24), the processing is considered *"unlawful due to a lack of transparency and violates the principle of fair processing"*.

Although the GDPR does not specify the legal form of arrangement between joint controllers and therefore, parties are free to agree on the arrangement. The EDPB (2020:43) recommends that such arrangement be made in the form of a binding document such as a contract or other legally binding act under EU or Member State law to which the controllers are subject. EDPB (2020: 43) adds: *"the use of a contract or other legal act will allow joint controllers to demonstrate that they comply with the obligations imposed upon them by the GDPR"*.

3.2 Legal Relations between the Controller and the Processor

The controller (language data owner) can use limited and clearly defined assistance in processing data (e.g. structuring data, making it available). If a CLARIN consortium member makes the language data available itself, it is the controller. The person/entity assisting the controller is the processor.¹⁵ The described situation is also applicable to the case when someone deposits language data with a CLARIN consortium member (see Sec 5). Any processing of PD by the processor must be governed by a contract (GDPR Art. 28).

The GDPR requires that the controller uses only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing meets the requirements of the GDPR and to ensure the protection of the rights of the data subject (Art. 28 (1)). The processor can process the personal data only on documented instructions from the controller (Art. 28 (3) a). Therefore, special attention should be given to the content of this agreement. The controller and the processor may choose to compile their contract including all the compulsory elements or to rely upon, in whole or in part, on standard contractual clauses (SCCs) adopted by the European Commission (Art. 28 (6)).¹⁶ The contract may be concluded under terms negotiated separately between the parties or may be based on contract terms drafted in advance for use in standard contracts or which the parties have not negotiated individually for some other reason (standard terms and conditions). The use of standard terms in data processing contracts is the most common way in practice due to a large number of parties. In the sharing of language data, it is also the most reasonable to use a drafted in advance standard data processing agreement by standardising the requirements applicable to data sharing procedures.

The GDPR requires the contract to regulate the following: *"the subject-matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects and the obligations and rights of the controller"* (Art. 28 (3)).

The parties must sufficiently describe the nature of responsibilities, taking into account the real risks and activities carried out under the specific research project to ensure that all appropriate safeguards are provided (see Art. 89 (1)). EDPB explains it as follows: *"clauses which merely restate the provisions of Article 28(3) and (4) are inadequate to constitute standard contractual clauses. /.../ a contract under*

¹⁵ WP29 (2010: 1) explains that the processor has to meet two conditions: 1) a separate legal entity with respect to the controller; 2) it processes PD on behalf of the controller.

¹⁶ On the 14. January 2021 the EDPB and European Data Protection Supervisor (EDPS) have adopted joint opinions on two new updated sets of contractual clauses (SCCs): one opinion on the SCCs for contracts between controllers and processors and one on the SCCs for the transfer of personal data to third countries. Available at https://edpb.europa.eu/our-work-tools/consistency-findings/edpb-edps-joint-opinions_en (27.1.2021).

Article 28 GDPR should further stipulate and clarify how the provisions of Article 28(3) and (4) will be fulfilled" (2019: 5).

In addition to the list of obligations and rights in Article 28, the GDPR also contains other parties' obligations that should be agreed upon in the contract. For instance, one such obligation is the processor's obligation to notify the controller of data breaches (Article 33 (2)). The parties could agree on the reasonable deadlines of notifying breaches or refer to written documented instructions governing more specific data sharing procedures in such cases. Additionally, it is necessary to agree on other conditions of the parties' rights and obligations not mentioned in the GDPR, including reimbursement of expenses and remuneration, procedures for amendments, and the contract's termination.

The agreement between the controller and the processor must be concluded in written form, including electronic form (Art. 28 (9)). EDPB (2020: 30) has pointed, that non-written agreements (regardless of how thorough or effective they are) cannot be regarded as sufficient laid down in Article 28. Therefore, to avoid any difficulties in demonstrating that the contract is actually in force, the EDPB recommends ensuring that the necessary signatures are included. Otherwise, the competent supervisory authority will be able to direct an administrative fine against both parties (2020: 31).¹⁷ The electronic form has been clarified by the European Parliament (2018): "*However, the rules for entering into contracts or other legal acts, including in electronic form, are not set forth in the GDPR but in other EU and/or national legislation. The e-commerce Directive (Directive 2000/31/EC) provides for the removal of legal obstacles to the use of electronic contracts. It does not harmonise the form electronic contracts can take. In principle, automated contract processes are lawful. It is not necessary to append an electronic signature to contracts for them to have legal effects. E-signatures are one of several means to prove their conclusion and terms*".

The GDPR does not determine whose responsibility is to ensure that the contract for personal data processing is concluded correctly. It is important to note that the requirements of the GDPR are infringed where data processing starts without a binding written contract, and the processor cannot demonstrate the existence of documented instructions (Art. 28 (3)). In such a case, the processor may be considered as the controller in respect of such processing and is subject to the obligations and increased liability of the controller (Art. 28(10)).¹⁸ The obligation to use only processors who are providing sufficient guarantees does not end when the contract is concluded. It is a continuous obligation, and the controller should regularly verify the processor's guarantees through audits and inspections to ensure that the actual data processing is correspondent to the contract and properly and lawfully executed (EDPB 2020: 30).

4 Legal Remedies in Case of Non-compliance with the GDPR

The GDPR provides severe administrative penalties for failure to comply, and the personal data protection authority has been given broad powers. In addition to that, a claim for compensation may be submitted by the person (data subject) who has suffered damage as a result of a breach of data protection requirements (Art. 82 (1)). In some cases, claims for damages may also be filed by family members or heirs (Wybitul, Haß & Albrecht:113-118; Cordeiro 2019: 492-499). A claim can be filed against the controller or the processor. For example, suppose the requirements of the GDPR have been violated in the processing of data for research purposes. In that case, the data subject may file a claim for damages with the university (but not against specific employees who process the data).

A precondition for satisfying a claim for damages is a breach of obligations provided for in the GDPR and in the Member State laws specifying rules of the regulation (Rec. 146).¹⁹ Controllers have obligations that can be described as obligations to *achieve a specific result*. For example, personal data must be collected for legitimate purposes and processed in a way compatible with those purposes (GDPR Art.

¹⁷ Infringements of the obligations of the controller and the processor pursuant to Article 28, be subject to administrative fines up to 10 000 000 EUR, or in the case of an undertaking, up to 2 % of the total worldwide annual turnover of the preceding financial year, whichever is higher (GDPR Art. 83 (4) a)).

¹⁸ Otherwise, the processor is liable to the data subject for the damage caused by the processing only if processor has failed to comply with the requirements of the GDPR specifically addressed to the processors or processor has not complied with the legal instructions of the controller or has acted against them (Art. 82 (2)).

¹⁹ See also Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications). Brussels 10.1.2017. COM (2017) 10 final. 2017/0003(COD). Awaiting Parliament's position in 1st reading.

24 (1)). Most of the controller's obligations can be characterised as obligations to make *reasonable efforts* to do something. For example, Article 6 (1) (d) of the GDPR provides that the controller must take "*every reasonable step*" to ensure that data that is inaccurate or incomplete shall be erased or rectified. To ensure and be able to prove the GDPR compliance, the controller must implement appropriate technical and organisational measures as provided by Art 25 (1) and Art. 32 (1) (for further discussion, see Van Alsenoy 2016). When data is processed on a contractual basis, it is necessary to consider which measures are appropriate for the fulfilment of specific instructions and obligations and to describe them as precisely as possible to assess whether the data controller has fulfilled its obligations. The contract should consist of the technical and organisational measures that are used to demonstrate the GDPR compliance (Art. 24 (1)). When processing is based on consent, the controller must be able to present evidence that the data subject has consented to the processing of his personal data (GDPR Art. 7 (1)).

The processor involved in the processing is liable for breaches of its obligations if it has not complied with the controller's legal instructions or has acted against them. However, the controller is not released from liability if the processor has violated the data processing requirements. The controller may, in turn, claim from the processor compensation for the damage caused by the breach of obligations by the processor and compensated the data subject by the controller (GDPR Art. 82 (5)). The parties may agree in the contract how the compensated damage will be shared (Alsenoy 2016: 285).

It is presumed that the controller and processor are liable for events that result in damage suffered by data subjects. This means that in court proceedings, the controller is required to prove that it is in no way responsible for the damage (GDPR Art. 82 (3)). Here the compliance certificates (GDPR Art. 43) can be of major practical importance. However, the certificate is not a safeguard against civil liability but can be used as a piece of evidence that the GDPR standards were implemented and that a party is not responsible²⁰. The processor also has some options to defence, especially if damage occurred is due to actions outside or contrary to the lawful instructions of the controller. In conclusion, the controller's liability is strict, i.e., the processor is released from liability only if he can prove that the data processing requirements have not been breached (Van Alsenoy 2016: 276, Strugala 2020: 77).

A personal data breach may result in physical, material or non-material damage to natural persons such as loss of control over their PD or limitation of their rights, discrimination, identity theft or fraud, financial loss, unauthorised reversal of pseudonymisation, damage to reputation, loss of confidentiality of personal data protected by professional secrecy or any other significant economic or social disadvantage to the natural person concerned (GDPR Rec. 85). On such occasions, the data subjects may claim full and effective compensation for the damage suffered, which means that both material and non-material damage is subject to compensation (Truli 2018).

The data subject's various costs (e.g. legal costs) could be considered as material damage. In addition, material damage may occur in the form of loss of income. For example, the data subject may lose income in the case of termination of an employment contract due to the disclosure of certain data or if stricter contractual conditions were imposed on him or her by financial institutions (Cordeiro 2019: 495). As the GDPR does not give any guidelines about the assessment of damages, national case law on non-material damage varies widely. In some Member States, the mere processing of data contrary to data protection legislation is not a sufficient violation to justify an award of non-material damage (Sein et al. 2018: 112). The person must have suffered noticeable disadvantage, and it had to be an objectively comprehensible impairment of personality-related interests with a certain weight (Oberlandesgericht Dresden 2019; Landesgericht Karlsruhe 2019)²¹. In some Member States, the mere fact of misuse of data can be sufficient to justify non-material damages (van Alsenoy 2016: 271; Rechtbank 2020). There is a significant risk that the GDPR infringement cases will give rise to a very differentiated court practice regarding the assessment of damages for non-pecuniary harm (Strugala 2020: 68).

If the controller or the processor has paid full compensation for the damage suffered in a situation where more than one controller or processor are involved in the same processing of personal data, the

²⁰ According to the Recital 81 of the GDPR the controller should use only processors who can provide sufficient guarantees in particular in terms of expert knowledge, reliability and resources, to implement technical and organisational measures, including for the security of processing. So, the adherence of the processor to an approved code of conduct or an approved certification mechanism may be used as an element to demonstrate compliance with the obligations of the controller.

²¹ For example, the German Higher Regional Court of Dresden held that minor loss did not give rise to any claim for non-material damages pursuant to Article 82.

controller or processor is entitled to claim back from the responsible parties the compensation corresponding to their part of the responsibility for the damage ((Art. 82 (5); see also Van Alsenoy 2016). This rule is based on the general principle provided for in the Art. 28 (4) of the GDPR that the initial processor shall remain fully liable to the controller for the performance of that other processor's obligations.

In addition to filing a claim for damages, the data subject can also demand the termination of the activity, causing the damage and refrain from doing so in the future.²²

5 Sharing Language Data within the CLARIN Framework

When it comes to data sharing inside the CLARIN community, we can distinguish between two different situations: 1) an external individual or entity deposits LD with a national CLARIN consortium member; 2) a national CLARIN consortium member itself makes LD available.

The first case involves the conclusion of a deposition agreement between the depositor and the CLARIN consortium member. The depositor determines the access and use conditions which makes the depositor the controller under the GDPR, and the CLARIN consortium member acts as the processor since it processes personal data on behalf of the depositor.

In the second scenario, the CLARIN consortium member shares LD on its behalf and is the controller.

The main question in both scenarios is whether the sharing of LD leads to joint controllership between the party who shares and the party who receives the data. According to the GDPR joint controllers jointly determine the purposes and means of processing.²³ They need to determine their respective duties (Art. 26). The European Court of Justice (ECJ) has explained that *"a broad definition of the concept of 'controller', the effective and comprehensive protection of the persons concerned, the existence of joint liability does not necessarily imply equal responsibility of the various operators engaged in the processing of personal data. On the contrary, those operators may be involved at different stages of that processing of personal data and to different degrees"* (C-40/17 para 70). This means that processing at different stages can result in joint controllership. The court has also maintained that *"a religious community is a controller, jointly with its members who engage in preaching, of the processing of personal data carried out by the latter in the context of door-to-door preaching organised, coordinated and encouraged by that community, without it being necessary that the community has access to those data, or to establish that that community has given its members written guidelines or instructions in relation to the data processing"* (C-25/17 para 75). It says that it is possible to be a joint controller even without having access to PD. This could apply to data sharing situation as well.

From the CLARIN perspective, the proposed agreement structure for the transfer of personal data aims to establish a CLARIN Centre as a data processor serving the national CLARIN consortium with each of the consortium members, or an external party, as a controller of its data sets. To this end, the Finnish CLARIN consortium proposes a CLARIN Framework Deposition Agreement (FADA).

The CLARIN FADA is intended to establish a framework of standard deposition rules for data sets that can be communicated by a CLARIN Centre. Individual data sets are added as attachments to the CLARIN FADA, which thereby reduces to a 1-page main document for each data set referring to the general conditions and four data set specific appendixes:

- 1) the data identification, description and citation texts,
- 2) the deposition license conditions with an end-user license agreement template,
- 3) a list of third-party copyrights or database rights,
- 4) the personal data description and the purpose of use of the data set.

Appendixes 3 or 4 may explicitly be left empty if there are no third-party rights or no personal data in the data set.

²² For example, Estonian case law has satisfied the requirement to submit an application to the information search systems Google, AltaVista and Yahoo to stop disclosing defamatory personal data in order to end a situation that damages a person's reputation. See Riigikohus 2010, p 11.

²³ In practice, 'purposes' are much more important than 'means' for determining the controller, cf. WP29 opinion: *"while determining the purpose of the processing would in any case trigger the qualification as controller, determining the means would imply control only when the determination concerns the essential elements of the means"* (2010: 14). It can be argued that in the CLARIN context, where the 'essential elements of the means' are generally similar and known to everyone (computational analysis), only purposes matter.

In the CLARIN infrastructure, there are three main licensing categories dividing language resources into three groups: 1) Publicly available (PUB); 2) For academic use (ACA) and; 3) For restricted use (RES) (for further discussions, see Oksanen et al. (2010) and Kelli et al. (2018)). The CLARIN RES licensing category (for restricted use) is suitable for sharing data sets with PD.

In the suggestions for how to implement the ethical intent of the GDPR in a research setting, Pormeister (2020) recommends that the original controller stays informed about all further use of a personal data set to inform the data subjects about such further use when necessary. The CLARIN RES license requires that data sets not be communicated to a third party by the end-user because a new legitimate end-user can always obtain a copy directly from CLARIN. As the CLARIN Centre, in most cases, remains a mere processor of personal data with the task to communicate such data to research organisations. The original controller stays informed about all requests for further use of a data set.

If there is a request for using a data set for a research purpose that is not sufficiently compatible with the original purpose of use, the data subjects need to be informed. From the CLARIN perspective, it is a practical question whether the CLARIN Centre as a processor is commissioned to inform the data subjects or the original controller notifies them, and how one goes about informing them in practice, i.e. will personal communication be possible or is a public announcement sufficient.²⁴

6 Conclusion

The language research community is aware of personal data protection. At the same time, it is not clear who has to guarantee the GDPR compliance, which contractual arrangements are needed and what the legal consequences and remedies are in case of non-compliance. The following graph summarises the main analysed aspects concerning the sharing of the data:

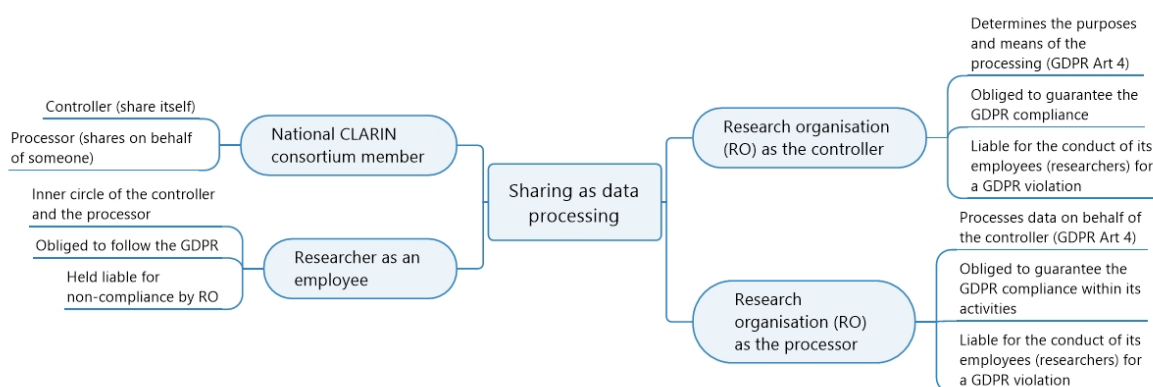


Figure 1. Sharing as data processing

According to the GDPR, the controller is the main responsible party since the controller "*determines the purposes and means of the processing of personal data*" (GDPR Art. 4 (7)). In academic settings, it is not always clear who the controller is. It should also be emphasised that a key feature of academic settings is academic freedom. Therefore, the relevant question is whether the university or an individual researcher is the controller? To answer the question, the authors analysed regulations, policy documents and case law to determine a general framework. Additionally, the authors conducted small case studies in several countries. Generally speaking, the issue is not settled yet. However, it can be assumed that the tendencies are that the university is considered the controller. There are practical reasons for this approach. Firstly, despite academic freedom, the university still controls and directs its employees. Secondly, the university has more resources to compensate for possible damage to the data subject and to take measures ensuring that the damage does not occur.

Researchers are not usually considered controllers. This does not mean that they are not responsible for their actions. In the case of a GDPR violation, they have to answer to their host universities.

²⁴ It seems to depend on whether the data were collected from the data subject (Art. 13 has only one exception to the obligation to provide information) or not (Art. 14 considers impossibility or disproportionate effort).

The authors also explored the relations between controllers and joint controllers. This may be relevant when two or more universities jointly conduct research (collect data, share it, etc.). The main conclusion is that controllership is not anything that can be contractually determined. It is a factual question. It is crucial that processing PD is transparent for the data subject.

In some case, the controller could use assistance in processing PD. The assisting party is called the processor when the assistance is clearly defined and limited. For example, this could be the case when another party is asked to help to structure or share data. To protect the data subject, the GDPR has several mandatory requirements, which are discussed in the article.

There is always a possibility that something goes wrong and someone's rights are violated. Therefore, the article also covers legal remedies in case of non-compliance with the GDPR. Since remedies and violations are two sides of the same coin, it was necessary to cover some obligations of the controller as well. One of the conclusions is that the controller has to demonstrate the GDPR compliance and take all possible measures to avoid harm to the data subject. Acting in good faith is a starting point.

The last section aimed to preliminarily place the previous analysis in the CLARIN context. Sharing language data within CLARIN requires specific arrangements depending on the nature of the relationship between the contractual parties. If a third party deposits data with a CLARIN consortium member, then the third party is the controller and the CLARIN consortium member is the processor. If a CLARIN consortium member shares data on its own behalf, it is the controller.

Several contractual arrangements are needed for sharing language data. However, the main reason for a CLARIN consortium member to be the controller of the PD is that researchers are often very mobile. Sometimes they no longer stay in academia, and sometimes researchers pass away. If someone needs access to the data later, a CLARIN Centre can still carry on with that role. Even if the researcher no longer is in a position to grant access to the data, the data is accessible also with a CLARIN Centre in a joint controller position where both a CLARIN Centre and the researcher separately have control of the data.

References

- [Aalto University] Aalto University. How to handle personal data in research? Available at <https://www.aalto.fi/en/services/how-to-handle-personal-data-in-research> (18.1.2021).
- [Act UWA] Foundation of the University of West Attica and other provisions. Act 4521/2018. Available at <https://www.kodiko.gr/nomothesia/document/345491/nomos-4521-2018> (27.1.2021).
- [AUEB] Regulation of Principles and Operation of the Research Ethics Committee of the Athens University of Economics and Business. Available at <http://rc.aueb.gr/el/static/home> (13.3.2021).
- [AUTH] Regulation of Principles and Operation of the Research Ethics Committee of the Aristotle University of Thessaloniki. Available at <https://www.rc.auth.gr/Documents/Uploaded/b4498638-1d32-45c9-b380-ec88ec6d143d.PDF> (27.1.2021).
- [Bari University] Regolamento in materia di protezione dei dati personali in attuazione del Regolamento UE 2016/679 del Decreto Legislativo 30 giugno 2003, n. 196 Codice in materia di Protezione dei Dati personali. Available at [https://manageweb.ict.uniba.it/ateneo/bollettino-ufficiale/Regolamento protezione dati dr1587 13032019.pdf](https://manageweb.ict.uniba.it/ateneo/bollettino-ufficiale/Regolamento%20protezione%20dati%20dr1587%2013032019.pdf) (27.1.2021).
- [C-311/18] Case C-311/18. Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems (16 July 2020). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1598506855221&uri=CELEX:62018CJ0311> (17.3.2021).
- [C-40/17] Case C-40/17. Fashion ID GmbH & Co. KG vs. Verbraucherzentrale NRW eV, interveners: Facebook Ireland Ltd, Landesbeauftragte für Datenschutz und Informationsfreiheit Nordrhein-Westfalen (29 July 2019). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587057502926&uri=CELEX:62017CJ0040> (17.3.2021).
- [C-25/17] Case C-25/17. Tietosuojavaltuutettu, intervening parties: Jehovan todistajat (10 July 2018). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587066001018&uri=CELEX:62017CJ0025> (17.3.2021).

- [C-22/98] Case C-22/98. Criminal proceedings against Jean Claude Becu, Annie Verweire, Smeg NV and Adia Interim NV (16 September 1999). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999262601&uri=CELEX:61998CJ0022> (17.3.2021).
- [C-66/85] Case C-66/85. Deborah Lawrie-Blum vs. Land Baden-Württemberg (3 July 1986). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999733855&uri=CELEX:61985CJ0066> (17.3.2021).
- [C-256/01] Case C-256/01. Debra Allonby v Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of State for Education and Employment (13 January 2004). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999966374&uri=CELEX:62001CJ0256> (17.3.2021).
- [CNRS guide 2019] CNRS Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte. GUIDE POUR LA RECHERCHE. Available at https://inshs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf (17.3.2021).
- [Constitution of the Italian Republic] Costituzione della Repubblica Italiana. Available at https://www.cortecostituzionale.it/documenti/download/pdf/Costituzione_della_Repubblica_italiana.pdf (16.3.2021).
- [Cordeiro 2019] Cordeiro, António Menezes. Civil liability for processing of personal data in the GDPR. European Data Protection Law Review 2019/5(4), 492-499.
- [CoDAU guidelines] Convegno dei Direttori generali delle Amministrazioni Universitarie (CoDAU). Linee guida in materia di privacy e protezione dei dati personali in ambito universitario. Versione 1.1 – novembre 2017. Available at http://www.codau.it/sites/default/files/verbali/all_3_linee-guida_privacy_gdpr_ravera.pdf (27.1.2021).
- [CRUI regulations] Regolamento in materia di protezione dei dati personali in attuazione del Regolamento UE 2016/679 del Parlamento Europeo e del Consiglio e del decreto legislativo 30 giugno 2016, n. 196 Codice in materia di protezione dei dati personali. Available at https://www.fondazionecru.it/wp-content/uploads/2019/02/bozza_schema_regolamento_privacy.pdf (27.1.2021).
- [DP Handbook 2018] European Union Agency for Fundamental Rights and Council of Europe (2018). Handbook on European data protection law 2018 edition. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-coe-edps-2018-handbook-data-protection_en.pdf (17.3.2021).
- [Data Protection Policy] Data Protection Policy of the University of Tartu. Available at <https://www.ut.ee/en/data-protection-policy> (17.3.2021).
- [Description] Order No. R-316 of the Rector of the University of Vilnius of 25 May 2018 concerning the approval of the description of the procedure of personal data processing at the University of Vilnius. Available at <https://www.vu.lt/en/privacy-policy#general-provisions> (18.1.2021).
- [EC] European Commission. What is a data controller or a data processor? Available at https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controller-processor/what-data-controller-or-data-processor_en (17.3.2021).
- [EDPB 2020] European Data Protection Board. Guidelines 07/2020 on the concepts of controller and processor in the GDPR. Version 1.0. Adopted on 02 September 2020. https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_202007_controller-processor_en.pdf (16.03.2021).
- [EPRS study 2019] European Parliamentary Research Service. How the General Data Protection Regulation changes the rules for scientific research. July 2019. Available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU\(2019\)634447_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU(2019)634447_EN.pdf) (17.3.2021).
- [European Parliament 2018] European Parliament (2018). Parliamentary questions. Available at https://www.europarl.europa.eu/doceo/document/E-8-2018-003163-ASW_EN.html (17.3.2021).

- [EU Charter of Fundamental Rights] Charter of Fundamental Rights of the European Union. 2012/C 326/02. OJ C 326, 26.10.2012, p. 391-407. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT> (17.3.2021).
- [FORTH] Research Ethics committee of the Foundation for Research and Technology-Hellas. Available at https://www.forth.gr/index_main.php?c=46&l=g&s=&p=1 (27.1.2021).
- [French Universities Guidelines 2019] Université Paris Lumières, Université Paris Nanterre, Université Paris 8, Règlement général pour la protection des données. Fiches pratiques à destination des chercheurs. Available at: http://triangle.ens-lyon.fr/IMG/pdf/guide_rgpd_2019_web.pdf (17.3.2021).
- [GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679> (17.3.2021).
- [Guidance 2019] Data Protection Commission (2019). Guidance on Anonymisation and Pseudonymisation. Available at <https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf> (23.3.2021).
- [HEA] Higher Education Act (Kõrgharidusseadus). Entry into force 1.09.2019. Available at <https://www.riigiteataja.ee/en/eli/525062020001/consolide> (27.1.2021)
- [ICO 2012] Information Commissioner's Office (2012). Anonymisation: managing data protection risk code of practice. Available at <https://ico.org.uk/media/1061/anonymisation-code.pdf> (23.3.2021).
- [Kelli et al. 2019] Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labrpolou, Maria Gavriliidou, Pavel Straňák (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 72-82. Available at <http://www.ep.liu.se/ecp/159/008/ecp18159008.pdf> (17.3.2021).
- [Kelli et al. 2018] Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki, Pavel Straňák (2018). Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? In: Maciej Piasecki (Ed.). Selected papers from the CLARIN Annual Conference 2017. Linköping University Electronic Press, 102-111. Available at <http://www.ep.liu.se/ecp/147/009/ecp17147009.pdf> (17.3.2021).
- [Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. In: Koenraad De Smedt (Ed.). Selected Papers from the CLARIN Annual Conference 2015. Linköping University Electronic Press, 13-24. Available at <http://www.ep.liu.se/ecp/article.asp?issue=123&article=002> (17.3.2021).
- [Klavan et al. 2018] Jane Klavan, Arvi Tavast, Aleksei Kelli (2018). The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. *Frontiers in Artificial Intelligence and Applications*, 307, 71–78. Available at <http://ebooks.iospress.nl/volumearticle/50306> (28.1.2021).
- [Landesgericht Karlsruhe 2019] LG Karlsruhe, Urteil vom 02.08.2019-8 O 26/19. *OpenJur* 2020, 69001. Available at <https://openjur.de/u/2293311.html> (17.3.2021).
- [Lindén et al. 2020] Krister Lindén; Aleksei Kelli, Alexandros Nousias (2020). A CLARIN Contractual Framework for Sharing Personal Data for Scientific Research. In: Kiril Simov, Maria Eskevich (Ed.). Selected Papers from the CLARIN Annual Conference 2019 (75–84). Linköping University Electronic Press. Available at <https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=172&ArticleNo=10> (28.1.2021).
- [Lindén et al. 2019] Krister Lindén, Aleksei Kelli, Alexandros Nousias, (2019). To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest. *Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019*. Ed. K. Simov and M. Eskevich. CLARIN, 56-60. Available at <https://office.clarin.eu/v/CE-2019-1512-CLARIN2019-ConferenceProceedings.pdf> (17.3.2021).
- [Modena and Reggio Emilia University] Modena e Reggio Emilia University Regolamento in materia di protezione dei dati personali in attuazione del Regolamento UE 2016/679 del Parlamento Europeo e

- del Consiglio e del Decreto Legislativo 30 giugno 2003, n. 196 Codice in materia di Protezione dei Dati personali. Available at <https://www.unimore.it/hreg/RegolamentoPrivacy.pdf> (27.1.2021).
- [NCSR] Regulation of Principles and Operation of the Research Ethics Committee of the National Centre of Social Research. Available at https://www.ekke.gr/uploads/announcements/privacy_policy/fek_ekke_privacy_policy.pdf (27.1.2021).
- [Oksanen et al. 2010] Ville Oksanen, Krister Lindén, Hanna Westerlund (2010). Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN' in Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Available at <https://helda.helsinki.fi/handle/10138/29359> (17.3.2021).
- [Oberlandesgericht Dresden 2019] Oberlandesgericht Dresden 2019, Beschluss v. 11.06.2019 - Az.: 4 U 760/19. Available at <https://www.datenschutz.eu/urteile/Bei-bloßen-Bagatellverstößen-ohne-ernsthafte-Beeinträchtigung-fuer-das-Selbstbild-oder-Ansehen-einer-Person-besteht-kein-Schadensersatzanspruch-nach-Art-82-DSGVO-Dresden-Oberlandesgericht-2019061> (27.01.2021).
- [ORDA] Organisation of Research and Development Act (Teadus- ja arendustegevuse korralduse seadus). Entry into force 02.05.1997. Available at <https://www.riigiteataja.ee/en/eli/503062019008/consolide> (17.3.2021).
- [PDPA] Personal Data Protection Act (Isikuandmete kaitse seadus). Entry into force 15.01.2019. Available at <https://www.riigiteataja.ee/en/eli/523012019001/consolide> (17.3.2021).
- [PDPC] Personal Data Protection Code containing provisions to adapt the national legislation to Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Available at <https://www.garanteproperty.it/documents/10160/0/Data+Protection+Code.pdf/7f4dc718-98e4-1af5-fb44-16a313f4e70f?version=1.3> (17.3.2021).
- [Pormeister 2020] Kärt Pormeister (2020). Transparency in Relation to the Data Subject in Genetic Research - an Analysis on the Example of Estonia. Doctoral dissertation. Irene Kull; Jaak Vilo; Katrin Õunap; Barbara Evans (sup). University of Tartu. Available at <https://dspace.ut.ee/handle/10062/66697> (17.3.2021).
- [Rechtbank 2020] Rechtbank Noord-Nederland, 15-01-2020, C / 18 / 189406 / HA ZA 19-6. Available at <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBNNE:2020:247> (17.3.2021).
- [Riigikohus 2010] Riigikohtu tsiviilkolleegium, 9. detsember 2010, otsus nr 3-2-1-127-10. Available at <https://www.riigikohus.ee/et/lahendid?asjaNr=3-2-1-127-10> (17.3.2021).
- [RTDI] Research, Technological Development and Innovation Act and other provisions. Act 4310/2014. Available at <https://www.kodiko.gr/nomothesia/document/100926/nomos-4310-2014> (17.3.2021).
- [Risak & Dullinger 2018] Martin Risak, Thomas Dullinger. The Concept of 'Worker' in EU Law: Status Quo and Potential for Change. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3190912 (17.3.2021).
- [Rules for processing PD 2015] The Rules concerning the processing of personal data for scientific purposes at Vilnius University of 23 November 2015. Available at <https://www.vu.lt/teises-aktai> (18.1.2021).
- [Sartor 2018] Nicolas Sartor (2018). Data Compliance in the GDPR – How anonymization allows you to stay compliant in your data analysis. Available at <https://aircloak.com/data-compliance-in-the-gdpr/> (23.3.2021).
- [Schwartzmann 2019] Rolf Schwartzmann. Wer schützt die Forschungsdaten? Frankfurter Allgemeine Zeitung, 23 September 2019. Available at: <https://www.faz.net/-hfh-9ri6m> (3.2.2021).
- [Sein et al. 2018] Karin Sein, Monika Mikiver, Paloma Krööt Tupay (2018). Pilguheit andmesubjekti õiguskaitselahenditele uues isikuandmete kaitse üldmääruses. Juridica 2, 94-115.

- [SOQI] Structure, operation, quality assurance of studies and internationalisation of higher education institutes. Act 4009/ 2011. Available at <https://www.kodiko.gr/nomothesia/document/120922> (27.1.2021)
- [Strugala 2020] Strugala, Radoslaw. Art. 82 GDPR: Strict Liability or Liability Based on Fault? *European Journal of Privacy and Law&Technologies (EJPLT)*. Special issue, 2020, 71-79.
- [Truli 2018] Emmanuela Truli. The General Data Protection and Civil Liability, Chapter 12 In: Mohr Backum et al., *Personal Data in Competition, Consumer Protection and Intellectual Property: Towards a Holistic Approach?* Springer Verlag 2018, 303-329.
- [Turin University] Turin University Regolamento in materia di protezione dei dati personali in attuazione del Regolamento UE 27 aprile 2016, n. 679 del Parlamento Europeo e del Consiglio e del Decreto Legislativo 30 giugno 2003, n. 196 Codice in materia di Protezione dei Dati personali. Available at https://www.unito.it/sites/default/files/reg_protezione_dati_personali_870_2019.pdf (27.1.2021).
- [University of Cologne] Universität zu Köln, Stabsstelle 02.3 - Datenschutz und IT-Sicherheit, Forschungsdatenschutz. Available at: https://verwaltung.uni-koeln.de/stabsstelle02.3/content/forschungsdatenschutz/index_ger.html (17.3.2021).
- [Universities Act of Finland]. Universities Act of Finland. 558/2009. English translation available at https://www.finlex.fi/en/laki/kaannokset/2009/en20090558_20160644.pdf (18.1.2021).
- [University of Helsinki] University of Helsinki. Research data management. Available at <https://www.helsinki.fi/en/research/services-for-researchers-and-research-policy/research-data-management> (18.1.2021)
- [University of Jyväskylä] University of Jyväskylä. Instructions for researchers. Available at <https://www.jyu.fi/en/university/data-privacy/tietosuojaohjeet/researchers> (18.1.2021).
- [UoC] Regulation of Principles and Operation of the Research Ethics Committee of the University of Crete. Available at <https://www.ehde.uoc.gr/index.php/el/157-category-leitourgia-epitrophs/384-kwdikas-deontologias-gr-2> (27.1.2021).
- [UoM] Code of the Research Ethics Committee of the University of Macedonia. Available at <https://www.uom.gr/ethics/kodikas-hthikhs-kai-deontologias-ths-episthmon-ikhs-ereynas> (27.1.2021).
- [UoP] Code of the Research Ethics Committee of the University of Peloponnese. Available at https://elke.uop.gr/?page_id=2194 (27.1.2021).
- [UoPa] Code of the Ethics Committee for Scientific Research of the University of Patras. Available at https://ehde.upatras.gr/wp-content/uploads/2020/11/kodikas_hthikhs_kai_deontologias_pp-1.pdf (27.1.2021).
- [UoT] Regulation of Principles and Operation of the Research Ethics Committee of the University of Thrace. Available at <https://ethics.duth.gr/> (13.3.2021).
- [UT Data Protection Policy] University of Tartu. Data Protection Policy. Available at <https://www.ut.ee/en/data-protection-policy> (27.1.2021).
- [UTA] University of Tartu Act (Tartu Ülikooli seadus). Entry into force 21.03.1995. Available at <https://www.riigiteataja.ee/en/eli/527122019004/consolide> (27.1.2021).
- [UWA] Code of Research Ethics and Conduct Committee. Available at <https://research-ethics-committee.uniwa.gr/kodikas-deontologias/> (27.1.2021).
- [Van Alsenoy 2016] Brendan Van Alsenoy (2016). Liability under EU Data Protection Law. From Directive 95/46 to the General Data Protection Regulation. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 7, 271-288. Available at <https://www.jipitec.eu/issues/jipitec-7-3-2016/4506> (3.02.2021).

- [WP29 2014] WP29. Opinion 05/2014 on Anonymisation Techniques Adopted on 10 April 2014. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (23.4.2021).
- [WP29 2010] Article 29 Working Party. Opinion 1/2010 on the concepts of "controller" and "processor". Adopted on 16 February 2010. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf (17.3.2021).
- [WP29 2006] Article 29 Working Party. Opinion 10/2006 on the processing of personal data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT). Adopted on 22 November 2006. Available at <https://www.dataprotection.ro/servlet/ViewDocument?id=234> (27.4.2021).
- [Wybitul, Haß & Albrecht 2018] Wybitul, Tim, Haß, Detlef, Albrecht, Jan Philipp. Abwehr von Schadensersatzansprüchen nach der Datenschutz-Grundverordnung. NJW 2018/3, 113-117.

The Literary Irony in the Works of Juliusz Słowacki

Anna Medrzecka

The Institute of Literary Research
The Polish Academy of Sciences
anna.medrzecka@gmail.com

Abstract

The purpose of this text to present the research project on the presence of various types of irony in the work of Juliusz Słowacki, one of the most important and influential poets of Polish Romanticism. This poet is known for his masterful use of various types of irony, which are unique in Polish literature. The nature and manner of its usage changes with the development of the poet's work. The first element of said project is to show the development of the use of irony and the artistic means that are used to achieve an ironic effect. I present different types of irony, my main interest being *literary irony* (Hammon, 1997).

When examining irony, I use not only recognized methods of traditional literary studies, but also methods introduced by digital humanities, with particular use of tools provided by the CLARIN consortium. In my paper I present the process of creating a digital corpus of texts by Juliusz Słowacki and its use in research. The first important step was to use the LEM tool to generate statistics of lemmas and tags within the entire corpus, individual texts and groups of texts. Next, I created a list of stylistic irony markers that can be compiled in such a way that they can be detected using CLARIN tools. In some cases, the tools need to be adjusted, but most often it is possible to set the currently available tools so that the expected results can help in search of the described irony indicators.

1 Introduction

The purpose of this text is to present research carried out as part of the doctoral project *Irony in the works of Juliusz Słowacki in the light of digital humanities research*. The research is aimed both at verifying whether the definitions and theories of irony established in the literature on the subject like e.g. (Szturc, 1992) (Hammon, 1997), can be supported by computer analysis, and at checking whether the computer analysis can provide a new point of view on the issue of irony and romantic irony. The project has a unique character due to its subject matter, which are poetic and literary texts. This makes it impossible to use tools typical for rhetorical irony analysis, such as the analysis of the tone of voice or the context of the statement. In addition, it raises the need to distinguish between elements of the ironic style – which is a subject of said research – and those elements that are characteristic of a literary text in general, which sometimes can be mistaken with the irony determinants.

The research is carried out with the use of presently available CLARIN-PL infrastructure, in the co-operation with Wrocław University of Science and Technology. Given its particular requirements, it constitutes a major contribution to the development of tools that can be used in other projects.

The research is carried out in several stages, described later in this article.

First, it was necessary to use philological analysis and textological knowledge to compile the corpus of Słowacki's works, including the creation of a list of available materials and the selection of appropriate versions. This stage has already been completed.

Then, selected texts were digitized and saved in files adapted for computer analysis.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

The next step was to compile a list of determinants of the ironic style, especially those characteristic for Słowacki's work. The list has been preliminarily prepared, but is still being developed, because the conducted research often points out to unforeseen phenomena.

The current stage of the project is the selection, adaptation, testing, and development of digital tools that will best allow to explore the sought determinants of the ironic style in the research corpus. This part of work is still ongoing, so solutions proposed in this paper are likely to be developed still.

2 Juliusz Słowacki and his works – preliminary difficulties

The project is focused on the analysis of the body of work of one author, Juliusz Słowacki – one of the greatest Polish poets of the Romantic period. The preparation of the corpus and its adjustment to the analyses planned was the first important task that presented significant difficulties arising from the peculiar character of the author's artistic legacy.

Even as much as 75 percent of known and published works by Juliusz Słowacki are texts which remained unpublished until after the author's death. Afterwards, the poet's legacy ended in the hands of his mother, Salomea Bécu and his uncle, Teofil Januszewski. The latter handed over his late nephew's autographs to Antoni Małecki, a classical philologist, who in his few treatises on Polish literature mentioned Słowacki in a rather unfavourable way. Still, it is Małecki to whom we owe the first publication of many works left by the poet in autograph. Sadly, Januszewski did not prepare a catalogue of manuscripts delivered to the future publisher, so we are unable to determine the number of works Małecki decided to leave unedited for any reason. The unpublished works include not only letters or private notes of their author, including journal entries or a 'glossary' of Arab terms written down during his travels to the East, but also the overwhelming majority of poems, dramas, and narrative poems. Some of them present few textual problems – the versions conveyed in the manuscripts are uniform and coherent. However, a large portion of the legacy is made up of fragments of works or the so-called 'portions' later merged by the editors into the presently well-known works like the later parts of *Król-Duch* 'King-Spirit' or the so-called later songs of *Beniowski*. In case of the later works, separating individual portions would require massive editorial effort. It was therefore necessary to select the basis for texts put into the body of corpus. In most cases, I decided to base on the imperfect but still canonical Juliusz Kleiner's edition, both in terms of designating the beginning and end of individual texts, and the lection of the autographs. In case of texts having a newer critical edition, like e.g. Samuel Zborowski, published from manuscript by Marek Troszyński (Troszynski, 2017) numerous corrections were taken into account.

2.1 Corpus characteristic

In terms of corpus research, the varied status of texts (and the resulting different status of a word depending on its location) poses a significant problem. The works can be divided into:

1. **Texts published by the poet in his lifetime.** This group caused the fewest problems, because it can be assumed that the first printing reflects the author's intention, the status of a word is strong and unquestionable. Apart from obvious editing mistakes, some doubt can be posed in places by spelling and punctuation, as the comparison with manuscripts often shows that they may be a result of editorial intervention. For the purpose of the analyses described, spelling was standardized in all researched works in order to limit the number of incorrectly recognizes forms ('ign.') to the minimum and to obtain the possibly most reliable statistics. In researching irony spelling and punctuation are not without meaning. However, given the diversity of the material and the fact that it was impossible to obtain similarly reliable records in case of all texts, it was decided not to include this element in the analysis.
2. **Texts published after the poet's death, manuscripts missing.** In case of these texts there is a significant problem with assessing the reliability of the content (this group also includes the vast majority of the letters). Unfortunately, because it was impossible to verify the edited version against the manuscript original, it was decided for the sake of this analysis to accept critical editions.

3. **Texts published after the poet's death, surviving autographs.** Most often, these texts come in numerous versions, the relations of which is impossible to be determined today. In some cases, each version differs much from the others, in other the differences are minor. In classic editions these texts are often published with 'text variants'. In order to avoid falsifying the analysis results, it was decided to include text variants in the corpora, but without repeating fragments identical in several versions.

2.2 Text types and genres

Having determined the composition of the corpus formation, it became necessary to solve the problem arising from the diversity of genres of the researched texts. The corpus includes:

- 24 dramatic works and fragments
- 9 philosophical works
- 261 letters
- 30 narrative poems
- 25 prosaic works and fragments
- 209 poems

Here, it's worth adding that the texts differ much in terms of size. This results in a significant statistical imbalance. In order to avoid an extreme unbalance of texts in the corpus, the texts were divided into smaller files. This did not eliminate the disproportion completely, but still allowed to limit the falsification of data resulting from the use of the instances per million (i.p.m.) measurement (e.g. by the Kontext tool). I.p.m. shows the number of occurrences of a given word per million, related to the whole corpus or subcorpus (depending on the kind of analysis; corpus can consist of only one text). This measurement can be used to compare frequencies between corpora of different sizes. However, it should be remembered that although theoretically the relative frequency of the occurrence of a given word determined with the use of this tool will have a similar value for texts differing in size, in practice the lexical resource of a longer text is by definition much richer, so the results should be interpreted with great caution..

The base version of the corpus, containing all available texts in all existing versions, has been uploaded to the Kontext tool (Machálek, 2020) tool) which allows searching and creating frequency lists of words and tags. Kontext statistics show, that the tekst includes 1 180 127 word forms. The corpus is divided into subcorpora according to the generic characteristics described above.

2.3 Establishing corpus and subcorpora for analysis

In face of the apparent lack of balance in the corpus, a researcher has to solve numerous problems before attempting a proper analysis:

1. **How to select texts for analysis?** It was decided that wherever possible, all available texts should be included. It means including literary works as well as personal notes and letters.
2. **Is it methodologically justified to compare texts of different size and genre?** The problem of size was solved by dividing longer texts into shorter parts, and then using the i.p.m. measurement. It was also necessary because some of the used tools (e.g. WebSty) can only process files no larger than 20 Kb. The problem of genre was solved by dividing the corpus into subcorpora of texts with similar genre features. This allowed to single out features of irony characteristic to a given genre, not appearing in other subcorpora. For example, the use of verbs in 1st person singular in narrative poems may suggest its autothematic character, whereas in drama it is a genre feature.
3. **How to include changes related to the chronology of works in the analysis?** Texts from different periods differ in terms of subject and have certain specific stylistic features. The word *Duch* 'Spirit'

may serve as an example here. It is quite rare in works prior to 1843, whereas in the last years of the poet's life it becomes one of the most important lexemes, so much that it occupies the first place on the frequency list of the entire corpus. At the same time, given the nature of the legacy, it is not possible to precisely determine the order in which some of the later works were written. That is why while determining the analysis parameters it was necessary to balance them in a way that would make them possibly insensitive to elements related to the subject of a given work.

4. **How to treat erasures?** A word left in the manuscript is susceptible to removal, crossing out, being replaced with a new version. From today's point of view, it is impossible to describe relations of texts and versions. A text left in manuscript is truly an open work. So would it be responsible to treat manuscript words and published words as equal? Unfortunately, the only available solution would be to completely abandon analysing unpublished texts, and that would drastically reduce the research material. For this reason, it was decided to accept this imperfect decision and treat manuscripts and published works as equal sources, with relevant annotations.

3 In search for irony – current stage of research

The main subject of the research is the search for irony in the works of Juliusz Słowacki, with particular focus on romantic irony understood according to the theory by Friedrich Schlegel (Schlegel, 2019) and the classic publications for Polish irony studies like works of Włodzimierz Szturc and Zofia Mitosek. While elaborating on the concept of irony the researcher referred to literary studies, like the French *Ironie littéraire* (Hammon, 1997). The current state of machine text processing research on irony turned out relatively uninteresting, as it mostly focuses on small utility texts, like Twitter posts, or product reviews on online stores.

For this research it is important to determine the degree to which literary irony, discernible by a human, is visible in results of computer analysis. Definitely, this is not about creating a simple tool for irony recognition – it would seem that such a tool would not be possible at this stage, nor needed. In this project, computer analyses serve mainly as support for classical literary analyses. It is also a purpose of this research to find features of literary irony only visible via machine text processing. Already some interesting phenomena have been discovered as part of the work carried out so far, e.g. different types of temporal phrases used in texts recognised as ironical. The catalogue of features considered characteristic to an ironic text is currently being prepared for this purpose based on subject literature. The catalogue is still open. What's more, analyses conducted by a research team allow us to detect unexpected occurrences, and that is the greatest advantage of using digital tools.

Each occurrence is then described in a way allowing for its verification with the use of digital tools, thus making it possible for the researchers to verify characteristics of ironic style in the first place, and second to detect other stylistic elements pointing to an ironic reading that would be impossible to discern in traditional reading.

At this stage the catalogue of features is based mostly on the text material of the narrative poems. Detected ironic style features include:

1. **Repetitions**, broadly understood. Firstly, these are of course repetitions of one lexeme. Secondly, some fragments contain groups of different words related on a formative level, such as:
 - (1) Znałem... lecz, szczęściem ('fortune'), uleczona z żalu
Saffone, bardzo podobna do greckiej.
Ta sie, nieszczęściem ('misfortune'), kochała w Moskalu
A Moskał zginał na wojnie tureckiej
Ta poszła zabrać na warneńskim polu
Zwłoki, a uszy - w Konstantynopolu.
Smutna! Ubrana w kwiaty sympatyczne
Poszła nieszczesna ('unfortunate') na brzegi Marmora

2. **Proper nouns and foreign words on rhyme positions.** The rhyme position is naturally salient; emphasizing it by the use of a foreign word, one that naturally draws attention, provides yet another element of parabasis.
3. **Stylistic excess**, namely numerous appearance of words similar in meaning, emotional character, or register, in a limited fragment of a text. Unlike the simple repetitions, this may include various word forms and lexemes, and of very different meaning, only building up the ambience to the point when it becomes impossible to take seriously.

Bo i tu — i tam — za morzem — i wszędzie,	'Because here - and there - and everywhere
Gdzie tylko poszle przed sobą myśl biedna,	Wherever I put my poor thoughts
Zawsze mi smutno, i wszędzie mi jedno;	I'm always sad and always I feel the same
I wszędzie mi źle — i wiem, że źle będzie.	And I always feel bad - and I know I will feel bad.
Wiec już nie myślę teraz tylko o tém,	So I have no thoughts but this one
Gdzie wybrać miejsce na smutek łaskawe,	Where I can find a good place for my sadness
Miejsce, gdzie żaden duch nie traci lotem	A place no spirit would touch
O moje serce rozdarte i krwawe;	My broken and bloody heart
Miejsce, gdzie księżyc przyjdzie aż pod ławę	A place where a moon will come
Idac po fale...	On the waves...'

4. **Juxtaposed fragments differing in register**, such as comical fragments and fragments full of pathos. This feature of irony is explicitly called by the poet in several places, it is also characteristic for romantic irony, in which the elimination of language transparency is of particular importance.
5. **Autothematism, “writing about writing”, parabasis** – classic elements related to irony. This one is a higher-rank phenomena than those described above; here expressed by the accumulation of words related to writing ('I write', 'I describe', 'I say', 'I listen', 'I see', 'I sing', 'poet', 'poetry', 'song', etc.). In case of verbs, the ones appearing in 1st person singular (especially in narrative poems) are particularly important, their use in an author-narrated text has no other meaning than parabasis.

Naturally, the phenomena described above, are not invariably signs of irony; the purpose of this research is to verify the degree to which their appearance and co-appearance may point to an ironic reading. The examples above do not form a complete list of phenomena suggesting the use of irony in a literary text.

4 CLARIN tools

In order to verify the aforementioned features in texts from the corpus tools provided by the CLARIN consortium are being used.

4.1 Inforex

The tool to create annotated text corpora was used to gather all works and provide basic annotations with information about e.g. possible text variants (Marcińczuk, Oleksy, 2019).

4.2 LEM (Literary Exploration Machine)

LEM (Maryl, Piasecki, Walkowiak, 2017) is a versatile tool for text exploration, which in one place combines functionalities of other tools adding to their options. The main features of LEM include: lemmatization, part-of-speech tagging, generating custom wordlists and lemmatized texts. In the initial phase the tool was used to create the frequency list for the entire corpus, individual subcorpora, and singular texts. This allowed the researcher to perform preliminary stylometric analyses involving the frequency of words in individual subcorpora. The new “own categories” function allowed the text to be searched with reference to the above mentioned stylistic elements.

4.3 WebSty

The WebSty (Piasecki, Walkowiak, Eder, 2018) tool is used for stylometric analysis. As part of the project, the researcher is planning to use tools both for standard stylometric analysis, and with parameters precisely selected in terms of features described in the first part of this text. The planned analyses include the comparison of texts based on word lists pointing to autotelism – the list includes verbs like 'to write', 'to sing', 'to create', and nouns like 'poet', 'poetry', 'muse', 'song', etc.

4.4 NER

The NER tool turned out particularly useful while searching for autotelic fragments, as it allows users to analyse the occurrence of proper nouns and temporal words in a given text. The Liner2 (Marciniczuk, Kocon, Oleksy, 2017) tool made it possible to detect an additional feature of ironic texts, namely the occurrence of verbs differing in character on a limited space.

4.5 Słowosieć

plWordNet (Słowosieć) proved a particularly important tool in this research. The tool shows words grouped into sets of cognitive synonyms (synsets). It facilitates all analyses of tone, register, and search for synonyms and repetitions. It was also tested in terms of untypical word combinations or diminutive forms.

5 “Okno” ('Window') – an idea for the future

The characteristic feature of many stylistic elements related to irony is not just their occurrence (most of the features may carry other tropes or phenomena), but their frequency, or combination of several stylistic measures in one place. Hence the need for a tool that would not only check the occurrence of specified features in a whole text, but also detect their occurrence in one specific fragment. Such a tool is still in the concept stage, but its function could be based on numerous identical analyses for subsequent fragments, and later combining them in one statistic.

6 Current project state and perspectives for further research

Up to this point, all project stages focused on corpus forming tasks and allowed for a preliminary preparation of a catalogue of phenomena whose occurrence in text will be analyzed. The catalogue remains open also because new features appear as a result of examination research with the use of available tools. For it to be finished, it will be necessary to prepare additional functionalities for some of the elements of the CLARIN infrastructure; the current cooperation with the team at the Wrocław University of Science and Technology allows us to continuously introduce necessary improvements and functionalities.

The project is still underway so no final conclusions can be made. It seems, however, that the results obtained so far are promising and justify an optimistic outlook on future research in the designated direction, and the scientific methods prepared may be used in research on other corpora.

References

- Piasecki, M., Walkowiak, T., Eder, M. (2018, May). Open Stylometric System WebSty: Towards Multilingual and Multipurpose Workbench. In: *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017* (No. 147, pp. 145-158). Linköping University Electronic Press.
- Maryl, M., Piasecki, M., Walkowiak, T. (2018, May). *Literary Exploration Machine A Web-Based Application for Textual Scholars*. In: *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017* (No. 147, pp. 128-144). Linköping University Electronic Press.
- Marciniczuk, M. Oleksy, M. (2019). Inforex — a Collaborative System for Text Corpora Annotation and Analysis Goes Open. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, pages 711–719. Varna, Bulgaria. INCOMA Ltd.*
- Szturc, W. 1992. *Ironia romantyczna*, Wydawnictwo Naukowe PWN, Warszawa 1992.

- Troszyński, M. 2017. *Alchemia Rekopisu. Samuel Zborowski Juliusza Słowackiego* Wydawnictwo IBL, Warszawa 2017.
- Machálek, T. (2020). *KonText: Advanced and Flexible Corpus Query Interface*. In: Proceedings of LREC 2020, s. 7005–7010.
- Marcińczuk, M., Kocoń, J., Oleksy, M. (2017, April). Liner2—a generic framework for named entity recognition. In: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (pp. 86-91)
- Schlegel, F. 2019. *Fragmenty*, trans. Bartl, C. Wydawnictwo Uniwersytetu Jagiellońskiego.
- Hammon, Ph. 1997. *L'Ironie littéraire. Essai sur les formes de l'écriture oblique* Hachette, Paris.

Digitizing University Libraries – Evolving from Full-Text Providers to CLARIN Contact Points on Campuses

Manfred Nölte
State and University Library
Bremen, Germany
noelte@suub.uni-
bremen.de

Martin Mehlberg
State and University Library
Bremen, Germany
martin.mehlberg@suub.uni-
bremen.de

Abstract

Based on the example of the State and University Library Bremen (SuUB) we will outline in this paper, how academic libraries with digitization activities (hereinafter referred to as *digitizing libraries*) could establish even closer ties to CLARIN in the future. After describing SuUB's past and current CLARIN-related activities (especially full-text transfers to a CLARIN-D centre) we suggest that this collaboration could be expanded by providing advice and training for researchers of the Digital Humanities as potential CLARIN users. Equally important from our point of view is the discussion about future structural options on the level of research infrastructures. We suggest a collaboration between digitizing libraries to jointly agree upon standards of data quality, file formats, interfaces and web services. We discuss the foundation of local CLARIN contact points to pass scholars and researchers on to the respective contact or service of CLARIN. The relevance to CLARIN activities, resources, tools or services is described at the end of each respective section. From the conclusions, the reader will notice: It is the right time for change.

1 Digitizing University Libraries as Full-Text Providers for CLARIN

The State and University Library Bremen is one of many libraries dedicated to the digitization of its historical collections. Digitization and especially the generation of full text is an important instrument for improving the accessibility of valuable information contained in fragile historical documents. It facilitates academic research and teaching and is indispensable to the Digital Humanities. By doing so, these libraries play a very important role as full-text providers or creators of data.

Usually, university libraries undertaking digitization projects produce digital images, metadata for cataloguing and web-navigation purposes, and optical character recognition (OCR) full text for searching. These resources are made available through the library's web portal for digital collections. However, digital humanists need rather high-quality full texts enriched with metadata in the appropriate format in order to process and analyze them with powerful software tools (like regular expression search, part-of-speech tagging, named-entity recognition or topic modeling). To satisfy this specific demand, the SuUB has actively transferred full texts created through digitization projects (funded by the German Research Foundation, DFG) to the Berlin CLARIN-D centre and thus via metadata harvesting to the CLARIN research infrastructure. All these CLARIN tools, concentrated, documented and ready to use, have been a great motivation for this activity. We would like to outline our approach adopted so far, the results and the dissemination achieved within the scientific community. Later on, in section 3 we discuss the underlying structure and concept and how we might intensify operations like this.

The historical journal *Die Grenzboten* was the first full text transferred from the SuUB to CLARIN (Geyken et al., 2018). *Die Grenzboten* is a long running serial publication (1841–1922) which can be classified as a literary journal that also covers politics and arts. It was founded by Ignaz Kuranda (1811-1884) in Brussels in 1841 and later on published in Leipzig and Berlin. We demonstrate that good OCR quality and a page-by-page structuring are prerequisites for the creation of a high-quality

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Text Encoding Initiative (TEI) version of a full text. The TEI version was created in cooperation with the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) (Nölte et al., 2016).

We digitized more than 185,000 single pages in 270 volumes. Almost 33,000 articles were digitized via optical character recognition (OCR) and the titles of the articles were manually captured. The resulting OCR full text was processed by the OCR software ABBYY Finereader 9 and consists of approximately 500 million characters and 65 million tokens. As a second aspect of text quality, we enhanced the level of document structure according to an agreed standard format together with our partners, the Deutsches Textarchiv (DTA; Haaf Geyken and Wiegand, 2014/15). Figure 1 shows manually corrected and tagged “zoning information” based on coordinates provided by the ABBYY Finereader XML files. Using this structure information, we converted the OCR output format to an interoperable TEI format. The metadata of the 33,000 articles also contain information about the publication dates, so that it is possible to analyze the full texts over time.

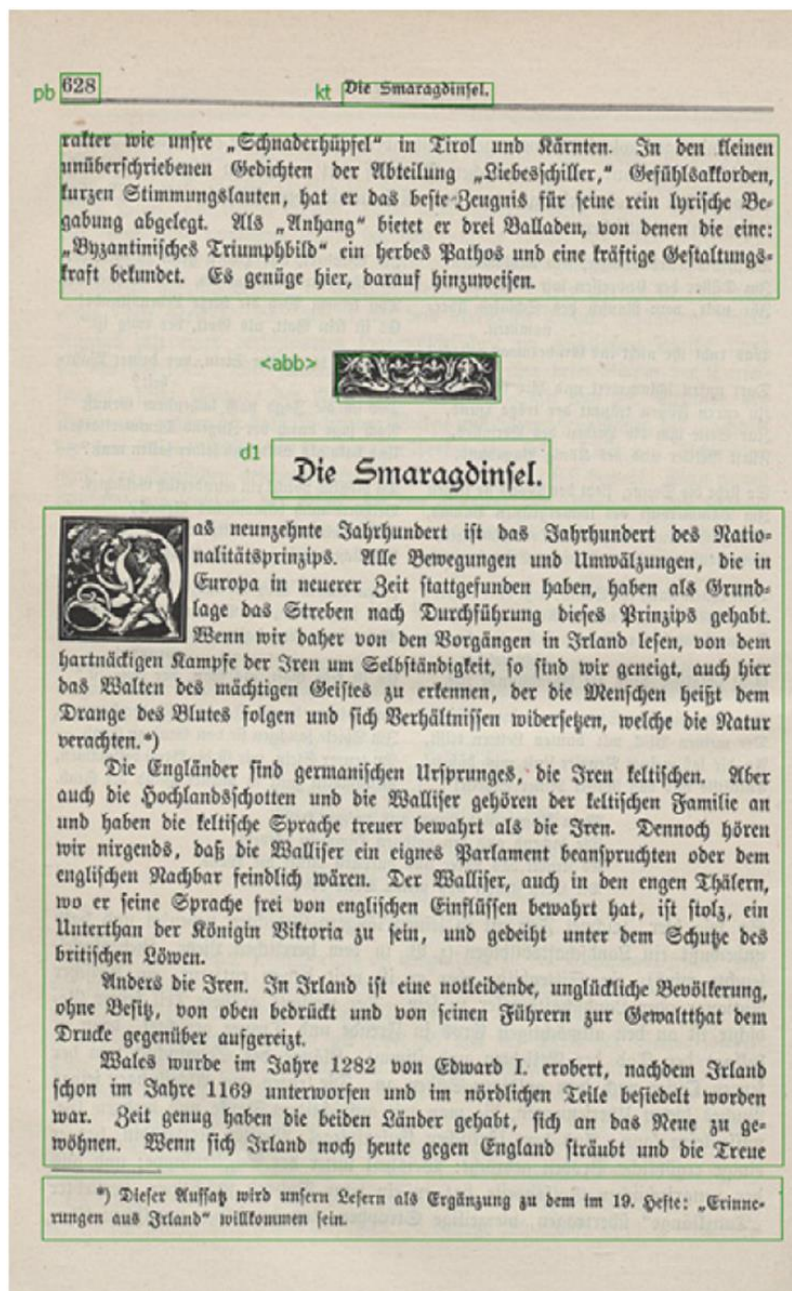


Figure 1: “Zoning” – adding structural annotation to the pages

We were active to disseminate our digitized historical journal enhancing its outreach. It has been used as textual material by computer linguists, digital humanists and philologists, as well as being part of

academic teaching at universities like Ghent, Würzburg and Göttingen¹. Together with ground truth full text data, it has been used by OCR post correction system providers like PoCoTo (Vobl et al., 2014) and ProjectComputing (Evershed and Fitch, 2014) and big projects like OCR-D (Neudecker et al., 2019). The journal *Die Grenzboten* was subject to diachronic collocation analyses² (Jurish and Nieländer, 2020) as well as to analyses like topic modeling (Jannidis, 2016; Fechner and Weiß, 2017; Graham et al., 2012). The project website at SuUB lists five other research projects in the context of the humanities.

It is clear that this usage is a result of actively providing the journal full text to the scholarly community and to research infrastructures as described in this paper. Up to now this journal is just an example of a serial source available at this level of quality and accessibility. A lot more have been digitized or are in the process of being digitized. Later on for example, the SuUB and the University Library Frankfurt digitized over 1000 book titles with about 245,000 pages within the project “Digitale Sammlung Deutscher Kolonialismus” (a digital collection of texts from the period of German colonialism). This project also generated full texts that have been transferred to CLARIN. Both full-text transfers are showcases for a collaboration/teamwork between a digitizing library and CLARIN. The question now is: What has to be done to intensify the transfer of huge amounts of digitized full texts in a reasonable and cost-effective routine manner? We will revisit this question in more detail in section 3.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. As described above, the SuUB has actively transferred full texts to a CLARIN-D centre in an early phase directly after the digitization process. Doing so, we helped to increase the amount of language resources provided by CLARIN and together we generated a bigger perception of this full text within the scientific community. CLARINs “ingest service” and the possibility to host the full text in the CLARIN repositories also clearly helped to increase the dissemination of our full texts within the scientific community.

Digitization activities have the potential to create huge amounts of digitized full texts. This in turn stimulates inter- and cross-disciplinary research. This first collaboration between CLARIN and the SuUB Bremen can be seen as a showcase scenario of how content-providing libraries and CLARIN can mutually benefit from these kinds of digitization projects.

Like CLARIN, we seek to take the requirements of the user community into consideration. For a library it is business as usual to have a lot of contact with our “users” (patrons), especially if they are interested in the [digitized] material we hold. This way we know the demands and expertise of our library patrons.

With our digitized full-text resources we have been serving a lot of different user communities: German philology, linguists, Digital Humanities, political science, history, etc.

2 Counselling and Training Activities of the SuUB Bremen

Academic libraries are close to scholars and researchers not only in terms of physical closeness but also in terms of subject proximity (providing information and services). The main task of libraries has been (and still is) the supply of literature/scientific information. All libraries have contact points and recently they have even gained importance on many campuses, because they provide a “learning space” for students and scientists.

Libraries of course know “their own material”, i.e. texts digitized by themselves, best. They know the subject, the context and the quality of the digitized material. The latter manifests itself in the

¹ The journal *Die Grenzboten* as research data on GitHub; ‘Expertenworkshop: Topic Modelling’ at the University of Göttingen (May 2018); a workshop by Bryan Jurish, Thomas Wernecke, and Maret Nieländer on ‘Diacollo and *Die Grenzboten*: Exploring Diachronic Collocations in a Historical German Newspaper Corpus’ at the Genealogies of Knowledge I — Translating Political and Scientific Thought across Time and Space conference at the University of Manchester (December 2017); ‘CIS OCR Workshop v1.0: OCR and Post Correction of Early Printings for Digital Humanities’ at the Ludwig Maximilian University of Munich (LMU) (September 2015); and a module, also at LMU, by Florian Fink on *PoCoTo: Practice* (2015) [all accessed 8 October 2018].

² Jurish, Bryan, M. Nieländer, and T. Wernecke. 2017. “DiaCollo and *die Grenzboten*.” Talk presented at the conference Genealogies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December, 2017. Jurish, B., M. Nieländer, and T. Wernecke. “DiaCollo and *die Grenzboten*.” Talk presented at the conference Genealogies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December, 2017.

condition of the original material, pixel images, error rates of the full texts as well as the scope and the quality of the metadata. Technical issues, for example, are the interfaces to get access to the data and the formats delivered by the respective systems. In summary, this enables libraries to help with quality issues, access related issues, even content-related issues or options to use web services like IIF³ (Snydman et al., 2015). For example, the quality of the section titles within the full text of *Die Grenzboten* is poor, due to the usage of a special character type, whereas the same information contained in the METS-XML files has perfect quality, because it was captured manually.

In the past, the SuUB has gathered some experience with the personal counselling of scholars from across the humanities disciplines, including linguistics and political science. In general, the questions were of a technical nature (relating for example to formats or system interfaces), but sometimes also questions of a more theoretical nature were discussed (e.g., kinds of quantitative analyses like topic modeling, diachronic collocations, etc.).

The aim of the technical advice was primarily to provide researchers with the knowledge necessary to make full texts available in an interoperable format that meets the requirements of specific software tools. Especially with structured full texts (TEI or in general all sorts of XML), format issues have to be considered. Some quantitative tools, like *mallet* (topic modelling; Graham et al., 2012) only need plain text. But the pre-processing or the whole tool chain (e.g., including a graphical presentation for the analysis findings) nearly always requires the above-mentioned features: structured pages (i.e., semantically tagged full text) and metadata (year of publication, authors, etc.).

Another point of view is to consider the several target groups (like bachelor, master and Ph.D. students, postdocs, researchers, lecturers and citizen scientists) within the library patrons appropriately. What target group is supposed to be passed on to CLARIN? What level of counselling are libraries able to fulfil? We need to find and define a certain transfer point.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. As shown, digitizing libraries are in a good position to start researcher training activities with respect to their full-text resources. Furthermore, they can help access web services or metadata offered by the digital collections software systems, like IIF and OAI-PMH.

Actively supporting the above-mentioned full-text transfers and the mentioned counselling activities will result in considerably better outcomes in all fields of automated and computer-aided research across disciplines working with digitized material. It will enable the employment of quantitative methods and approaches such as authorship attribution studies, clustering techniques (i.e., for literary genre analysis), topic modeling etc.

3 Prospects for Future Collaboration Between CLARIN and Academic Libraries

As shown above digitizing libraries already play a role in the context of CLARIN and the group of CLARIN users. The next step should be to intensify the collaboration between CLARIN and those libraries in order to harmonize the provisioning or transfer of digital textual material, to jointly agree upon common activities or to even establish CLARIN contact points on university campuses. The most appropriate place for these contact points is a library as we will demonstrate in the next section.

³ An IIF-manifest enables an IIF-viewer to establish a direct metadata link to the respective digitized resource. Here a viewer on universalviewer.io links to the SuUB Bremen:
<http://universalviewer.io/uv.html?manifest=https://brema.suub.uni-bremen.de/i3f/v20/1702471/manifest#?c=0&m=0&s=0&cv=6&xywh=-981%2C-124%2C4978%2C2444>

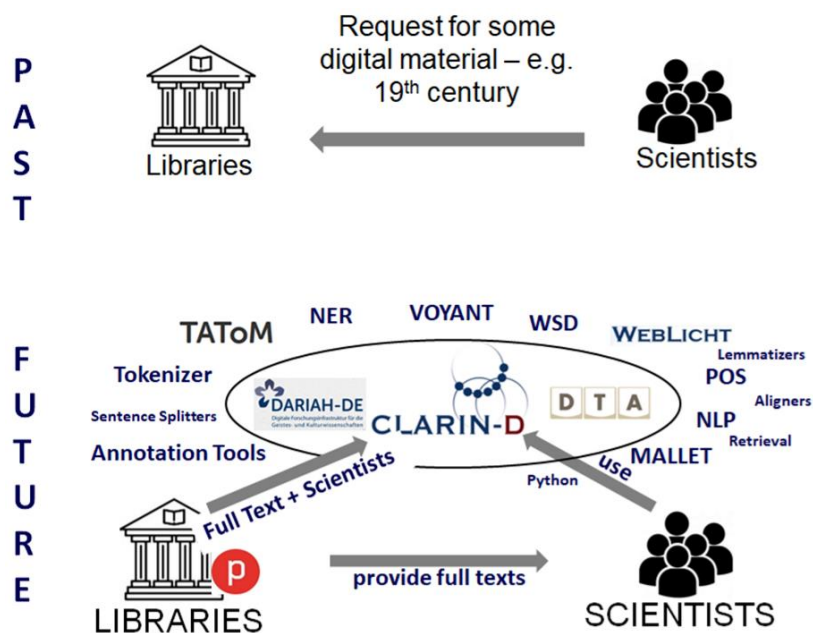


Figure 2: Prospects for Future Collaboration between CLARIN and Academic Libraries

Having done full-text transfers, mentioned in section 1, a few times we list some criteria that might be in need of a more precise specification together with potential requirements. Here we give only short explanations, see (Nölte and Blenkle, 2019) for more examples and details.

- *Full-text quality:* For example, a maximum error rate of characters or for other elements of the metadata, like the structuring of the text. These criteria may vary for different centuries or decades or different software tools or scientific approaches.
- *File formats and metadata:* Transferring plain text is not an option, nor is the output of OCR engines (such as ABBYY-XML). There has to be a decision for ALTO, PAGE, TEI or other file formats, possibly together with ‘annotation guidelines’. (Haaf et al., 2014/15)
 Supplemental note: The use of format converters should be considered with care. There will always be a loss of information converting from a format to another. A good and simple test might be to convert from format *A* to format *B* and back to *A*, and to compare the emerging differences.
- *Persistent back links:* Within the full text there should be back links (page-by-page or at least by sections) to the scanned images in the digital collections of the respective library or archive. If possible, these back links should be persistent at a page “URN granular” level (Sommer, 2010)⁴. Researchers appreciate having the possibility to check the original image quality or to have access to supplemental material such as graphics, images, advertisements, or vignettes.
- *Line breaks:* There should be a guideline whether to transcribe line breaks as is. We have cooperated with partners with varying opinions on this question. Some institutions wanted line breaks as is, whilst the transcription for Wikisource had to be without wrapped words.⁵
- *Strictness of character transcription:* Within historical full texts the spelling, of course, should be transcribed as is; for instance, ‘Säugethiere’ with ‘th’ and ‘Entwicklung’ instead of the modern form ‘Entwicklung’. The same should apply to the transcription of single historical characters using UTF8 codes, like ligatures or special historical glyphs. However, some tools (especially quantitative tools) or workflows may well require transliterated standard versions of the special characters.

⁴ There has been a pilot project for the persistent identification for individual pages, but up to now the registration of massive numbers of URNs is not common practice within digitization activities at libraries.

⁵ https://de.wikisource.org/wiki/Die_Grenzboten

The image shows two examples of historical characters. On the left, the word 'ältlicher' is displayed, where the 'ä' is a historical version of an umlaut-a. On the right, the word 'ift ist' is displayed, where the 'i' is a long-s.

Figure 3: Two examples of historical characters; left: a historical version of an umlaut-a; right: the long-s

Reflecting on these criteria also helps to remind the researcher of the original source of the material. With an OCRed full text being the ‘model’ of a paper original, Piotrowski (2019) mentioned a ‘mapping property’ and a ‘reduction property’ of models that researchers should be aware of. For example, the analyses of serial sources should also consider basic textual properties. An example is the distribution of text quality over time (i.e. year of publication), which might have an impact on methods used in the Digital Humanities. It is therefore necessary to develop methods with a stability⁶ towards varying OCR error rates. Methods and models requiring a constant error rate should be adapted or it should be possible to critically assess and interpret the results. Similarly, some of the mentioned criteria (such as ‘line breaks’ and ‘strictness of character transcription’) will also have an impact on the need for adaptation or interpretation of the used methods or an adapted pre-processing of the full texts.

Ideally, there should be documentation listing all the above-mentioned information: the level of the ‘full-text and metadata’ quality, whether there are further file formats available, the availability of back links, and the status of line breaks and character transcriptions. If this ‘full-text metadata’ is realized with computer readable XML formats, pre-processing scripts might automatically decide what pre-processing remains to be done, and what analysis tools or scientific approaches might be applicable. Licensing and intellectual property would be a further major issue to address.

Jointly discussing and agreeing upon criteria and requirements like this will lead to a best practice approach for future transfers of full texts to CLARIN. Together we might even go for a fully automated full-text transfer or harvesting as a long-term goal. And the above mentioned ‘full-text metadata’ might partially automate the preselection of texts to process.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. Here we have also addressed the issue of data quality. OCRed full-text resources have a certain error rate, and metadata may be rich and good or sparse and of poor quality. Setting up the described collaboration will standardize and harmonize all future full-text transfers and training activities in the context of CLARIN and digitizing libraries, i.e., together we will create best practice approaches that provide scholars and researchers with the best possible quality and interoperability of language resources and services.

4 Prospects for Future CLARIN Contact Points on Campuses

As proposed in section 3, another future activity is the establishment of CLARIN contact points for scholars and scientists at academic libraries. These already have a proficiency in counselling and offering services in the respective domains. They also function as learning spaces, aiming at creating the best atmosphere for the exchange of knowledge. Other outstanding advantages of libraries are: Libraries constitute sustainable structures in the scientific world, they are research infrastructures themselves and they are local on the campus. Every university has a library and universities are places where students become scientists who might be passed on to CLARIN as potential users.

Currently there are a lot of activities to establish event formats at libraries as scholar or researcher training activities with names like “digital lab”, “hands-on lab”, “GLAM⁷ lab”, “innovation lab”, “data lab”⁸, “HackyHours”⁹, “Digital Learning Lab”¹⁰, “Digital Humanities lab”¹¹, “Library Labs”¹²,

⁶ The ‘stability’ of an algorithm or method refers to the quality of the results of a stable method that does not degrade (that much) whilst the given input has a reduced or degraded quality.

⁷ GLAM is an acronym for “galleries, libraries, archives, and museums”.

⁸ DataLab, <https://www.uni-goettingen.de/de/daten+lesen+lernen/592287.html>

⁹ HackyHours, <https://librarycarpentry.org/blog/2019/06/hackyhours-zbmed/>

¹⁰ Digital Learning Lab, <https://www.uni-marburg.de/de/ub/lernen/kurse-beratung/wissen-organisieren/dll>

¹¹ Digital Humanities Lab, <https://ub.fau.de/forschen/digital-humanities-lab/>

¹² Library Labs, <https://www.bl.uk/projects/british-library-labs#>

“scholarly makerspaces”, “Digital Literacy”¹³ and more combinations of these words. CLARIN might play a role to harmonize this multitude of activities, to combine these with CLARIN’s experience, services and tools. The above-mentioned start of collaboration between libraries and CLARIN might be a good first step.

To establish CLARIN contact points on Campuses in the most appropriate manner, the needs/demands/requirements of the scholars and the current situation of academic libraries have to be taken into account. What is on the scholars’ wish lists? What concrete services would researchers like to see? What are libraries able to fulfil? The VDB¹⁴ annual report of the commission for research-oriented services (Leiß, 2020) states: Scientists want sample solutions and best-practice solutions for typical use cases in order to identify and omit problems early. And there is already a professional qualification and knowledge of typical difficulties and pitfalls together with their respective solutions. Scientists want central services and contact persons, which libraries offer.

But of course, something is new. The digital change has come into play. As we described above, services, book titles and pages have already partially turned into web services, metadata and files of different formats. This, for sure, will have an impact on new job profiles within academic libraries. A new strategy for recruiting and professional training will be necessary. Libraries have historically found and will in the future find a level to cope with the multitude of requirements, topics and user communities to accomplish a good start regarding the proposed activities. Another helpful approach is the collaboration between libraries. We have participated within YUFE¹⁵ network activities for an exchange of experience. From the library of Maastricht we have learned that a specific service needs time for being accepted and for being recognized and well used. This means that libraries need to increase outreach efforts to maximize the dissemination of their services. Last but not least, an integration of these new library services into the curricula of the respective departments of the university will help to stabilize these services and the “flow” of participants right from the lectures to the libraries and finally to CLARIN.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. Relevance to the CLARIN activities: As shown, the SuUB together with CLARIN has a big potential to establish user assistance, a help desk or contact point. While documenting the digital collections software system with user manuals, we might support scholars and researchers within the domain of digital language resources. An example is information about our systems’ persistent identifiers and citation mechanisms (see the above criterion for “Persistent back links”).

Contact points on Campuses also might be a useful activity of an academic library to increase the awareness of useful tools and resources of CLARIN.

5 Conclusions

Here we describe conclusions for scholars of the Digital Humanities and for data providing institutions, i.e., for CLARIN, libraries and the whole GLAM sector. As the authors come from the professional context of a digitizing library, they refer to this type of institution in the list of conclusions. Some or all of the conclusions may also be applicable to the entire GLAM sector or to all data-providing institutions, as indicated or in a modified form. The following list is a summary of statements discussed in the previous sections. Furthermore, invitations are given to start or intensify joint activities (in italics).

5.1 Conclusions for CLARIN and digitizing libraries

- *Get together and coordinate.*

Together we should setup a network to streamline the collection and propagation of the requirements of the user community back to the data creation institutions and to address the issues of data quality and data completeness.

We should standardize and harmonize all future full-text transfers and training activities in the context of CLARIN and digitizing libraries. Libraries are in a good position to start researcher

¹³ Digital Literacy, <https://www.tu-braunschweig.de/lehre/konzepte-tools-und-projekte/future-skills#c647807>

¹⁴ VDB: Association of German Librarians (Verein Deutscher Bibliothekarinnen und Bibliothekare, <https://www.vdb-online.org/>)

¹⁵ YUFE – Young Universities for the Future of Europe, <https://yufe.eu/>

training activities with respect to their full-text resources. They are used to address diverse scientific communities; they are local on the campus and are mostly already well equipped with learning spaces. This way we will create best practice approaches that provide scholars and researchers with the best possible quality and interoperability of language resources and services.

- *Help with establishing CLARIN contact points on campuses*
Scholars and researchers might be passed on to CLARIN or directly to the respective contact or service of CLARIN. There are a few questions to consider. What target group is supposed to be directed to CLARIN (Potential target groups: students (bachelor, master, Ph.D.), postdocs, researchers, lecturers and citizen scientists)? What level of counselling are libraries able to fulfil? We need to find and define a certain transfer point.
- *Consider each other as partners*
Libraries help to increase the amount of language resources provided by CLARIN and might also help to increase the awareness of useful tools and resources of CLARIN. CLARIN helps with the possibility to host the full text in the repositories with the “ingest service”. It also increases the dissemination of full texts within the scientific community.
- *Integrate the whole GLAM sector*
Together with data providing institutions inter- and cross-disciplinary research will be stimulated in the best possible manner.

Finally, there are some specific conclusions for libraries.

- *Continue to develop*
As we have learned from contacts to the above mentioned YUFE network, there are already libraries with new positions and job titles like data steward, data specialist, information specialist [digital] humanities, research data manager or education and research technician. Libraries should establish further professional training and should create those new positions, to meet new requirements. A further strategy might be to adapt the recruiting policy.
- *Pay attention to CLARIN activities, services and tools*
It should be clearly seen what CLARIN has achieved so far and how fast services and tools around digital resources have changed. Libraries should keep up or even better engage with existing and upcoming activities in the domain of digital research infrastructures.
- *Collaborate with other libraries or institutions from the GLAM sector*
The best way for an exchange of experience is by collaboration. With regard to approaches of digitization of research and education, locality and regionality are playing an increasingly minor role anyway.
- *Streamline IPR issues*
Intellectual property rights (IPR) should be kept as simple, open and transparent as possible.

5.2 Conclusions for scholars of the Digital Humanities

- *Help and engage yourself*
Help us to actively shape and build the above-mentioned training activities, and in the long run CLARIN contact points.
- *Consider academic libraries as helpful places*
Academic libraries will continue to provide huge amounts of digital content, respectively digital full texts. Libraries have valuable corpora for historical research. We suggest that researchers should regard libraries (and research infrastructures) even more intensively as partners to get access to historical full-text materials. Everybody might agree that digitizing libraries contribute significantly to the amount of the available digitized material. Still, a novel approach is to actively transfer (or provide via metadata harvesting) full texts to the researchers and to research infrastructures, which is one of the central issues of this paper.
Libraries have always been helpful with knowledge of “their” library stock (collections, literary remains, manuscripts, etc.). More and more they also help to access web services or metadata offered by the digital collections software systems, like IIRF (Snydman et al., 2015) and OAI-PMH.
- *Use “digitization on demand” services*
Help with increasing the amount of digitized full-text resources.

- *Know about the resources you use*
E.g., be aware of the original material and the whole process it has undergone (see the list of criteria in section 3).
- *Please give feedback*¹⁶
Tell us about your work with regards to the material you have been using. Contact the respective library or CLARIN (these both should intercommunicate anyway). Tell whether the material has matched your requirements with respect to the criteria discussed in section 3.

It is the right time for change, for new collaborations, new structures and powerful digital research infrastructures.

Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments and suggestions. This work has benefited a lot from the results of several earlier projects funded by the German Research Foundation (DFG).¹⁷

References

- Evershed, John and Kent Fitch. 2014. ‘Correcting Noisy OCR: Context Beats Confusion’, in: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 45–51 (New York: ACM). <http://dx.doi.org/10.1145/2595188.2595200>.
- Fechner, Martin and Andreas Weiß. 2017. ‘Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts’, in: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2017_005.
- Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Henning Lobin, Roman Schneider, Andreas Witt (Hgg.): *Digitale Infrastrukturen für die germanistische Forschung* (= Germanistische Sprachwissenschaft um 2020, Bd. 6). Berlin/Boston, 219–248. Online-Version, DOI: 10.1515/9783110538663-011.
- Graham, Shawn, Scott Weingart, and Ian Milligan. 2012. ‘Getting Started with Topic Modeling and MALLET’. DOI: 10.46430/phen0017
- Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014/15. ‘The DTA “Base Format”: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources’, *Journal of the Text Encoding Initiative*, no. 8, no page.
- Jannidis, Fotis. 2016. ‘Quantitative Analyse literarischer Texte am Beispiel des Topic Modeling’, *Der Deutschunterricht*, 68.5, 24–35.
- Jurish, Bryan and Maret Nieländer. 2020. Using DiaCollo for Historical Research. Selected papers from the CLARIN Annual Conference 2019. *Linköping Electronic Conference Proceedings 172*: 172 33–40.
- Leiß, Caroline. 2020. Bericht der Kommission für forschungsnahen Dienste 2019. O-Bib. *Das Offene Bibliotheksjournal / Herausgeber VDB*, 7(4), 1-3, DOI: 10.5282/o-bib/5628 .
- Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Kay-Michael Würzner, Matthias Boenig, Elisa Herrmann, and Volker Hartmann. 2019. OCR-D: An end-to-end open-source OCR framework for historical documents, in: *Proceedings of the 3rd International Conference on Digital Access to Textual*

¹⁶ The following question has been discussed within the session “Data Curation, Archives and Libraries” at the virtual CLARIN 2020 conference: “How to get feedback on whether the work is **deployed** by the community and **if it actually is meaningful**?” First, **every imaginable type** of feedback would definitely be appreciated by all data providing institutions. A **late feedback** would be having a look at publications on the respective work. We had such positive feedback. In the past, we had direct contact to the users, but on a large scale that is definitely not possible. With respect to the question “**if it actually is meaningful**”: We suggest the collection and forwarding of whether the material was matching the requirements or the type of **further** requirements, to establish some kind of backpropagation through this network of research infrastructures.

¹⁷ DFG funded projects: <https://gepris.dfg.de/gepris/projekt/196492153?language=en> (two projects), <https://gepris.dfg.de/gepris/projekt/324473798>

- Cultural Heritage*, Brüssel 09.05.2019, 53–58.
<https://dl.acm.org/doi/10.1145/3322905.3322917> [accessed 27 April 2020].
- Nölte, Manfred and Martin Blenke. 2019. ‘Die Grenzboten on its Way to Virtual Research Environments and Infrastructures’, *Journal of European Periodical Studies*, 4.1, 19-35.
- Nölte, Manfred, Jan-Paul Bultmann, Maik Schünemann, and Martin Blenke. 2016. ‘Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*’, *o-bib*, 3.1, 32–55 (p. 32) [accessed 27 April 2020].
- Piotrowski, Michael. 2019. ‘Historical Models and Serial Sources’, *Journal of European Periodical Studies*, 4.1
- Snydman, Stuart, Robert Sanderson, and Tom Cramer. 2015. ‘The international image interoperability framework (IIIF): a community & technology approach for web-based images’, in: *Archiving Conference*, vol. 2015, 16–21. Society for Imaging Science and Technology.
- Sommer, Dorothea. 2010. ‘Persistent Identifiers: the ‘URN Granular’ Project of the German National Library and the University and State Library Halle’, *LIBER Quarterly*, 19.3-4, 259–274. DOI:
<http://doi.org/10.18352/lq.7965> [accessed 26 January 2021].
- Vobl, Thorsten, Annetee Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2014. PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61. DATeCH '14. New York, NY, USA: ACM. DOI:10.1145/2595188.2595197.

Towards Semi-Automatic Analysis of Spontaneous Language for Dutch

Jan Odijk

UiL-OTS

Utrecht University, the Netherlands

j.odijk@uu.nl

Abstract

This paper presents results of an application (*Sasta*) derived from the CLARIN-developed tool *GrETEL* for the automatic assessment of transcripts of spontaneous Dutch language. The techniques described here, if successful, (1) have important societal impact, (2) are interesting from a scientific point of view, and (3) may benefit the CLARIN infrastructure itself since they enable a derivative program that can improve the quality of the annotations of Dutch data in CHAT-format.

1 Introduction

This paper presents results of an application (*Sasta*) derived from the CLARIN-developed tool *GrETEL* for the automatic assessment of transcripts of spontaneous Dutch language. The techniques described here, if successful, (1) have important societal impact, since they enable semi-automatic analysis of spontaneous language in a clinical setting, which is an important ingredient of assessments but requires specialised linguistic expertise and takes a lot of effort; (2) are interesting from a scientific point of view (various phenomena get a linguistically interesting treatment), and (3) may benefit the CLARIN infrastructure itself since they enable a derivative program that can improve the quality of the annotations of Dutch data in CHAT-format (CHILDES data, (MacWhinney, 2000)).

Section 2 introduces methods for the analysis of spontaneous language. Section 3 introduces the CLARIN-developed application *GrETEL* that *Sasta* has been derived from. Section 4 briefly describes other work that has been done on automating the analysis of spontaneous language. Section 5 describes the initial experiment that we carried out to assess the potential of the envisaged method. The results were so promising that a small project, called the *SASTA* project, was started up. It is described in section 6. Section 7 describes the most important problems we encountered, and section 8 describes how we addressed a first set of these problems. In section 9 we report on recent results obtained. We end with our conclusions and plans to address the remaining problems (section 10).

2 Analysis of Spontaneous Language

The analysis of spontaneous language is considered an important method for determining the level of language development and for identifying potential language disorders. Crystal et al. (1976) and Crystal et al. (1989) developed the LARSP method for language assessment, remediation and screening.¹ Many researchers developed variants of LARSP for other languages, see e.g. (Ball et al., 2012). Also for the Dutch language various methods have been developed for the analysis of spontaneous language, both for assessment of language development, e.g. GRAMAT (Bol and Kuiken, 1989), TARSP², a variant of LARSP for the Dutch language (Schlichting, 2005; Schlichting, 2017), and STAP³ (van Ierland et al., 2008; Verbeek et al., 2007) as well as for assessment of aphasia, e.g. ASTA⁴ (Boxum et al., 2013).

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

¹LARSP= Language Assessment, Remediation and Screening Procedure.

²Acronym for the Dutch variant of the expansion for LARSP.

³Acronym for *Spontane Taal Analyse Tool* ‘Spontaneous Language Analysis Tool’.

⁴Acronym for *Analyse voor Spontane Taal bij Afasie* ‘Analysis for Spontaneous Language in case of Aphasia’.

Though analysis of spontaneous language is important, it requires specialist linguistic knowledge and expertise, is very time consuming and requires full concentration, so that there is a clear need to investigate whether the process can be automated in full or partially.

The whole process involves several steps. We distinguish here four major stages.⁵ An assessment procedure starts with a session with the child or a patient to elicit spontaneous speech. This session is recorded. In the second stage, the recording is transcribed. This is a very intense and time-consuming process. Fully or partially automating this stage is highly desirable, and we are investigating it but it is not the focus of this paper. In a third stage the transcript is annotated grammatically, and these annotations are used to make an assessment of the patient's language development or language disorder. Finally, remediation goals and procedures are defined.

All assessment methods define so-called *language measures*. Each language measure defines a particular linguistic phenomenon, e.g. a grammatical construction or a morphological property of a word. Each occurrence of this phenomenon in the spontaneous language transcript is marked by a code, and the number of occurrences of this code determined, and used in a comparison with reference data to assess whether there are any deviations, development disorders or what the nature of the aphasia is. A concrete example from TARSP is the language measure indicated with the code *H_{ww} i*, i.e. auxiliary with infinitive. The number of occurrences of the grammatical construction in which an auxiliary co-occurs with an infinitive as its complement is determined and compared to reference data. In the appendix of (Schlichting, 2005) five examples of this construction occur, as in (1). I added the utterance identifier and bolded the auxiliary and the infinitive:

- (1) Examples of *H_{ww} i* in the (Schlichting, 2005) appendix
 - a. **blijft** die wel **staan** (10)
stays that indeed stand
'Will it continue to stand up'
 - b. **kan** je **vastmaken** (11)
can one fasten
'one can fasten it'
 - c. ik **wil** blauwe ogen **tekenen** (13)
I want blue eyes draw
'I want to draw blue eyes'
 - d. wat **moeten** we met die stukjes **doen** ? (22)
what must we with those pieces-DIM do ?
'What should we do with those little pieces?'
 - e. weet je wat ik **kan doen** (28)
know you what I can do
'Do you know what I can do'

These sentences must be annotated with the code *H_{ww} i*, and this is done manually.

This paper reports on initial experiments to partially automate the grammatical annotation stage of this process, with the goal to gain efficiency and possibly also to increase the quality of the annotations.

3 GrETEL

GrETEL (Augustinus et al., 2012) is an application to query treebanks. It makes existing manually verified treebanks for Dutch such as *LASSY-Small* for written Dutch (van Noord et al., 2013) and the *Spoken Dutch Corpus* (Oostdijk et al., 2002) available for search. The syntactic structures inside the treebanks are encoded in XML. GrETEL offers XPath to search in these syntactic structures for words, grammatical properties and constructions. In addition, it offers query-by-example facilities.

Version 4 of GrETEL, GrETEL4, (Odijk et al., 2018), enables a researcher to upload a text corpus and associated metadata, and have it automatically parsed by the Alpino parser (Bouma et al., 2001), after

⁵(Crystal et al., 1989) distinguish seven stages. We collapsed some of their stages.

which the resulting parsebank⁶ is made available for search. It also offers various ways of analysing the search results, for data and metadata combined.

The text corpus upload functionality also makes it possible to upload a transcript of a spontaneous language session and to analyse it for grammatical properties. We describe an experiment with this in section 5.

4 Related Work

To our knowledge, Bishop (1984) was the first to propose and partially implement an automation of LARSP (for English). Long et al. (1996 2000) developed a different system for English. For French, F-LARSP was automated by Parisse et al. (2012), though it could deal reliably only with inflectional properties and the lower stages of development (stages I–III but not stages IV and V). To our knowledge, no attempts have been made before to automate any of the spontaneous language assessment methods for Dutch. However, there has been work on automating the determination of readability, e.g. with the tools *T-Scan*⁷ and *Lint* (Pander Maat and Dekker, 2016; Pander Maat, 2017). Though these are different applications applied to a different domain (prepared written texts) and with different purposes (readability assessment), many of the underlying technologies are shared. For example, T-Scan also uses the Alpino parser. van Noord et al. (2020) developed a Syntactic Profiler of Dutch (SPOD⁸), as part of the treebank query application PaQu (Odijk et al., 2017). SPOD also targets prepared written texts.

5 Schlichting Appendix Test

In order to assess the potential of GrETEL for automating the TARSP analysis, we experimented on the appendix of (Schlichting, 2005). This appendix is intended for illustrating the TARSP analysis and contains a number of example sentences together with their analysis in the form of annotations. We use the analysis as our reference material. For reasons that will become clear below, we call this the *Bronze* reference. The utterances themselves together with their utterance identifiers have been entered in a plain text file in a format supported by GrETEL4.⁹ This file has been uploaded into GrETEL4, which results in a parsebank. This parsebank is publicly available in the GrETEL4 application.¹⁰

An example utterance and analysis is provided in table 1.¹¹

Utt	Level	word1	word2	word3	word4	ann	stages
10	Utt	blijft	die	wel	staan		
10	gloss	stays	that	indeed	stand		
10	Zc	W	Ond	B		+Inv	III,III
10	Wg	Hww i					III
10	VVW		AVn				I

Table 1: Example TARSP analysis

The *Utt* column contains the utterance identifier (*10*). The *Level* column contains the label *Utt* for the actual utterance and labels for the levels of analysis: *Zc* (sentence constructions), *Wg* (word groups), and *VVW* (Connectives, pronouns, and word structure). Next, there as many columns as there are word occurrences (word1, ..., word4), followed by an annotation column for annotations that are not aligned to any specific word occurrence. The final column contains the stages (of language development) that the annotations at that level belong to. The annotations at the *Zc* level are aligned to specific words (*W*=verb,

⁶We call a text corpus in which each sentence has been assigned a syntactic structure automatically a *parsebank*; if the syntactic structures have been manually verified we speak of a *treebank*.

⁷<https://webservices-1st.science.ru.nl/tscan>.

⁸<https://paqu.let.rug.nl:8068/spod>

⁹<https://surfdrive.surf.nl/files/index.php/s/Arsz81uZWbD10z8>.

¹⁰<http://gretel.hum.uu.nl/gretel-upload/index.php/treebank/show/tarvb2>.

¹¹The *gloss* row does not belong to the analysis but has been added here for convenience. The translation of this utterance is ‘Does that one really stay standing up’.

Ond=subject, *B*=adverb) except for *+Inv* (=with inversion). The phenomena associated with these annotations belong to developmental stage III. At the *Wg* level the annotation *Hww i* stands for ‘auxiliary verb with an infinitival complement’. It is aligned to the auxiliary verb *blijft* but also (implicitly) annotates the infinitive *staan*. Finally, at the *VVW* level, the word *die* has been annotated as a substantively used demonstrative pronoun (*AVn*), typical for stage I.

Though some annotations are aligned to specific word occurrences, the actual usage of such alignments is rather inconsistent. We have disregarded the alignment in the experiment.

The annotated data have been encoded in TSV-format and made available in an Excel file in the data folder.¹² Queries have been written for the TARSP language measures that cover the annotations. These queries yield a list of matches in the GrETEL4 application. The queries themselves as well as URLs which execute these queries directly on the parsebank in GrETEL4 are included in a file that contains a summary of the whole analysis.¹³ This file also contains, for each match, the utterance identifier for the utterance in which the match was found. Multiple matches can occur in the same utterance, so the queries yield multisets of utterance identifiers. The Bronze reference has also been specified with a multiset of utterance identifiers for each language measure.

The query used for the code *Hww i*, used as an example in section 2, yields the utterances with utterance identifiers 10, 11, 13, 22 and 28, exactly corresponding with the manual analysis.¹⁴

We noticed after doing several experiments that GrETEL finds many matches that are (in our view) correct though they do not occur in the Bronze reference. We therefore created a second, improved reference, which we have called the *Silver* reference, which includes the utterance identifiers found by GrETEL that are not in the Bronze reference and had them judged for correctness by one of the clinical linguists that we cooperate with. We suspect that the omission of these annotations in the Bronze reference is partially due to human oversight, and partially due to the fact that these data were never created as reference data but rather as illustrative analyses. Though a comparison with a Silver reference probably yields a higher score than a comparison with a truly complete reference (a *Gold* reference), it is a useful way to get an impression of what kind of performance is attainable. Having a Silver reference enables us to do three comparisons: (1) GrETEL v. Bronze reference, as a measure of quality; (2) GrETEL v. Silver reference, as an improved measure of quality; (3) Bronze v. Silver reference, as a measure of the quality of purely human annotation.

For this experiment, we wrote initial versions of queries to implement the TARSP method, but we only wrote queries for language measures that occur in the Schlichting appendix.

We use *recall*, *precision* and *F1-score* as defined in (2) as performance measures. Here *O* is the multiset of results and *R* is the reference multiset:

(2) Performance measures:

- a. Recall: $\frac{|O \cap R|}{|R|}$ (undefined when $|R| = 0$)
- b. Precision: $\frac{|O \cap R|}{|O|}$ (undefined when $|O| = 0$)
- c. F1-score: $\frac{2 * Recall * Precision}{Recall + Precision}$

The results of the experiment have been summarised in table 2.

The figures that we observe here are promising, though it must of course be noted that the experiment has not been carried out on an independent test set. Also note that recall of the automatic system when compared to the silver reference (0.89) is slightly higher than the recall of the human annotation (0.88): inspecting the relevant examples shows that this is caused by the fact that human experts easily overlook instances. However, humans clearly remain superior for precision (0.90 for human annotation, 0.86 for annotation by the system).

¹²<https://surfdrive.surf.nl/files/index.php/s/jJvj16TsDprIKXb>

¹³<https://surfdrive.surf.nl/files/index.php/s/P71is33HVDgbsKK>.

¹⁴This link executes this query in GrETEL: <http://shorturl.at/kzEH3>.

Comparison / Measure	R	P	F1
GrETEL v. Bronze	0.88	0.79	0.83
GrETEL v. Silver	0.89	0.86	0.87
Bronze v. Silver	0.88	0.90	0.89

Table 2: Performance of GrETEL versus a human-created Bronze reference, versus an improved reference called Silver, and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1).

6 The SASTA Project

The results described in section 5 were considered promising by ourselves and the Dutch Association of Clinical Linguistics (VKL). For this reason we decided to extend the development, in a project called SASTA (acronym for a Dutch expansion meaning Semi-Automatic Assessment of Spontaneous Language).

In the project we have developed a research prototype application called *Sasta* aimed at clinical linguists that takes as input (1) a transcript to be analysed; and (2) an assessment method to be applied. The application yields as output (1) a standard profiling form in accordance with the assessment method, plus an assessment of the language development stage or the language disorder of the patient; (2) the transcript enriched with annotations. The automatically annotated transcript can be manually adapted and then offered to *Sasta* again for generating a revised profiling form. We support three different assessment methods (TARSP, STAP and ASTA). Each method is defined as a set of queries, special modules that are needed, measures to deal with deviating input, etc. associated to language measures of the method.

In order to develop *Sasta* we have developed *Sastadev*, a piece of software intended for developers that enables input of multiple reference data in multiple formats and compares the output of *Sasta* with the references and provides a detailed analysis of the differences. Many data provided by VKL members and other clinical linguists have been used for developing the system.

Sasta and *Sastadev* reuse components of GrETEL (the Alpino parser, the upload functionality, and the query functionality) but apply them differently: GrETEL is optimally suited to apply a single query to a large treebank, while *Sasta* and *Sastadev* are more suited to apply multiple queries to a small treebank.

Automating TARSP requires formalising certain aspects of the methods. For example, Schlichting herself uses codes in her examples that are not defined in the definition of the method, though they resemble them.¹⁵ Data that we received from clinical linguists sometimes use yet other variants of the codes. Multiple annotations on a single word are separated by a hyphen or space though hyphens and spaces also occur inside codes (e.g. in *aan-uitloop* and *hww i*). It also appears that the coding scheme does not use fixed codes but presupposes a productive syntax. However, the coding scheme has not been formalised, and uses natural language words, which gives rise to all the horrors of natural language.¹⁶ We have formalised the annotations while at the same time allowing as much flexibility as possible to accommodate actual practice.

7 Problems to Be Addressed

There are many problems that the data and the technology pose and that have to be addressed.

First, the transcripts of the spontaneous language sessions contain a large amount of deviations of normal language use. These are partially due to annotation conventions, and partially due to the fact that the children who are still learning the language and patients with aphasia make imperfect utterances.

¹⁵For example *+inv* instead of *inv*; *v.u.soc.divers* instead of *v.u. sociale uitdrukkingen*; *neg* instead of *xneg*, etc. And in the form provided, sometimes yet other codes are used (e.g. *V.U.Soc.Ster* v. *v.u. sociale uitdrukkingen: stereotiepe uitdrukkingen*. Most of them are easily interpretable by humans but not by software.

¹⁶Successful communication is seriously hampered by natural language, even in as simple a domain as words or terms: natural language words have associations, have a (common sense) meaning, are often ambiguous, are specific to one language, and have variations (abbreviations, acronyms etc.). These properties make successful communication difficult if not impossible, surely between humans and machines but often also between humans. It is much better practice to use arbitrary labels that at best resemble existing words for mnemonic reasons but that are no natural language words.

Conventions for annotating the data had to be made more formal and more detailed. For example, TARSP has the convention that the actual utterance can be accompanied by additional remarks by the annotator between round brackets. However, such round brackets contain two different types of annotations: (1) indication for non-existing words: which word was intended by the patient, according to the annotator; (2) other remarks by the annotator. We want to make optimal use of these annotations, but then these two different uses must be formally distinguished. We therefore require annotations based on the CHAT-format, which formally encodes these two different cases differently.

All kinds of deviations occur in the transcripts. Here is a list of the most common deviations:

- Often, a string is a non-existing word because the transcript also describes how the word was pronounced, e.g. *mouwe* instead of *mouwen* ‘sleeves’ with the *n* unpronounced; *isse* instead of *is een* ‘is a’, *zie-ken-huis* with hyphens to indicate the separated pronunciation of the syllables of the word *ziekenhuis* ‘hospital’.
- overregularisation of word forms (e.g. *gevald* instead of *gevallen* ‘fallen’), and even misspellings of such overregularisations (*gevalt*).
- wrong inflected forms, e.g. *gekeekt* instead of *gekeken* ‘watched’).
- filled pauses.
- dialectal or sociolectal form variants, e.g. *-ie*-diminutives. (*boekie*) instead of *(t)je*-diminutives (*boekje* ‘booklet’).
- repetitions of (sequences of) word occurrences.
- partial repetitions of repeated words.
- false starts.
- other often-occurring grammatical errors, e.g. use of the wrong article, or of the wrong auxiliary for perfect tenses, agreement errors, etc.

For some of these, we are quite confident that we can address them in a sufficiently reliable manner to improve the analysis, and we have made some initial steps towards this. For others, however, we are less confident but we will nevertheless investigate how far we can get.

Second, the Alpino parser has limitations. It cannot analyse all compounds as compounds, it provides insufficient information on verbless utterances, it provides insufficient information on verb-first sentences, it sometimes parses an utterance incorrectly, it sometimes analyses an utterance in a way that differs from the reference (but is not incorrect). Alpino does not consider the context, can do very little when semantic restrictions apply, and cannot deal with intonation .

Third, certain items require queries that cannot be expressed in XPath or only with great difficulty, e.g., the TARSP item *6+* which requires 6 or more constituents in a clause, or the STAP query for adverbs other than locative and temporal adverbs (this query takes up 315 lines in XPath!).

8 Towards Solutions

Many of the problems identified in section 7 can be addressed and several have already been addressed.

For example, by writing the right queries we can analyse certain adverbs inside phrases as if they occur at a sentential level. For queries that cannot be easily formulated in XPath we enable functions in a full programming language (we use Python). In addition, we allow macros inside XPath queries to make the queries shorter and easier to read and to facilitate reuse. For example, the definition of ‘auxiliary verb’ in Tarsp requires a long enumeration of lemmas, and this exact same enumeration must be used in two different queries (*H_{ww} i* and *H_{ww}Z*). With macros the enumeration has to be stated only once.

We developed new modules for normalising orthography, for analysing compounds, for dealing with regional spoken language diminutives ending in *-ie(s)*,¹⁷ for overgeneralised inflectional forms of verbs (even misspelled ones), and for automatically detecting filled pauses and repetitions. We use these to adapt each utterance that Alpino cannot deal with to a variant of this utterance that Alpino can deal with. Some examples have been given in table 3.

Original utterance	Corrected utterance	Gloss
mama mouwe hoog	mama mouwen hoog	mum sleeves high
niet goed uitgekijken	niet goed uitgekeken	not well looked-out
die stukjes	die stukjes	those pieces-DIM
zie-ken-huis	ziekenhuis	hospital

Table 3: Some examples of automatic corrections to improve the performance of Alpino.

We are using data provided by the VKL and by several clinical linguists, all example sentences of (Schlichting, 2005) and Dutch CHILDES data during development.

We use these modules to generate ‘corrected variants’ of deviant utterances, so that Alpino can parse the utterance correctly. The system annotates each utterance for the errors encountered and the corrections applied, so that also an error analysis results. After parsing the corrected utterance the system replaces the corrected words by the original words on the basis of the metadata.¹⁸

We have also developed a module to automatically detect filled pauses and repetitions, and are experimenting with a module for automatically detecting false starts.

9 Recent results

Schlichting	%		O v B			O v S			B v S		
Eval Meth	Corr	Exts	R	P	F1	R	P	F1	R	P	F1
Sastadev	No	No	86.5	80.8	83.6	88.4	89.3	88.8	89.8	97.0	93.3
Sastadev	Yes	No	88.5	81.2	84.7						
Sastadev	No	Yes	88.0	70.1	78.0	89.0	77.1	82.6	86.8	94.4	90.4
Sastadev	Yes	Yes	91.2	70.8	79.7						

Table 4: Performance of Sastadev (version of early 2020) for the Schlichting Appendix (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1). Results are given for the original version of TARSP, i.e. the version also used in the initial experiment described in section 5 (*Exts=No, Corr=No*), for the original version of TARSP with corrections (*Exts=No, Corr=Yes*), and for an extended version of TARSP without (*Exts=Yes, Corr=No*) and with (*Exts=Yes, Corr=Yes*) corrections.

Table 4 shows the performance of Sastadev in the version of early 2020 for the Schlichting Appendix (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1).¹⁹ Results are given for the original version of TARSP (*Exts=No, Corr=No*), for the original version of TARSP with corrections (*Exts=No, Corr=Yes*), and for an extended version of TARSP without (*Exts=Yes,*

¹⁷This entails more than just replacing an *i* by a *j*, e.g. *bekkie* corresponds to *bekje* ‘beak’, *bekie* to *beekje* ‘brook’, *cluppie* to *clubje* ‘club’, etc.

¹⁸Alpino actually provides some facilities for this, by so-called bracketed input, but we decided to use our own implementation in SASTA.

¹⁹The scores of the automated comparison differ from the manual comparison (as in table 2) because some codes were wrongly counted in the manual comparison, e.g. the word *stukkies* ‘small pieces’ was wrongly analysed as a singular compound (*stuk* ‘broken’ + *kies* ‘tooth’) instead of as a plural diminutive, and this was counted as single error; but it should have been counted as three errors.

Corr=No) and with (*Exts =Yes, Corr=Yes*) corrections. The corrections applied are certain orthographic normalisations (addition of *n* after a word ending in *e*, separation of incorrectly concatenated words, dehyphenation), normalisation of regional diminutives, normalisation of overgeneralisations for verbs, and a compound identification module). These corrections are currently only applied for words that are not contained in a Dutch lexicon or in a name list, and at most one variant is considered.

We note a significant drop in precision for the version with the extensions (from 80.8 to 70.1 for the Bronze reference, and from 89.3 to 77.1 for the Silver reference). This is caused by the fact that the extensions, which involve improved versions of existing queries as well as new queries that were not defined before, cause SASTA to identify even more hits that were not marked in the manual annotations, and, for the new queries, even had no occurrence at all in the reference. For this reason we created a new Silver reference for the Schlichting appendix. Table 5 contains the scores for the current version of SASTA (early 2021). Compared to this new silver reference, SASTA scores higher than 90% for recall, precision and f1-score for the Schlichting Appendix. The corrections improve the recall by more than 2 percent points also here, and improve precision marginally.²⁰

Schlichting	%	O v B			O v S		
Eval Meth	Corr	R	P	F1	R	P	F1
Sastadev	No	90.1	72.4	80.3	92.0	91.4	91.7
Sastadev	Yes	93.1	73.1	81.9	94.4	91.6	93.0

Table 5: Performance of Sastadev (version of January 29, 2021) for the Schlichting Appendix (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), in terms of recall (R), precision (P) and F1-score (F1). Results are given for the most recent extended version of TARSP without corrections (*Corr=No*), and for the most recent extended version of TARSP with corrections (*Corr=Yes*). Note that the Silver reference used here differs from the one used in table 4.

We are still developing our implementation of TARSP and the SASTA modules. In the current implementation the corrections that we experimented with for the Schlichting Appendix have not been integrated yet. The results of the most recent system²¹ are given in table 6 (for the TARSP data), in table 7 (for the STAP data), and in table 8 (for the ASTA data).

We have worked with too little data to be able to keep a subset of the data separate, so here we can only report on results on data that have been used during the development of the system. However, we still do have independent data, and hope to report on results on these data in the near future when they have been converted to a format usable by Sastadev. We observe that recall for TARSP compared to the Bronze reference is of reasonable quality, but precision is rather low. However, compared to the Silver reference, precision increases dramatically (for Sample_08 more than 23 percent points!) and recall is also higher when compared to the Silver reference. We observe also here that recall of Sastadev is sometimes higher than recall in the purely human annotation (e.g. in samples 4, 7, 8, 9, 10) though also here precision by human annotators remains superior to Sastadev.

For STAP, we observe that recall and precision are already pretty good when compared to the Bronze reference, and also here they both increase when compared to the Silver reference.²²

For ASTA, we do not have Silver reference data for all samples yet. However, for the Silver reference data that are available, we see again that precision increases dramatically, and also recall increases. The ASTA scores overall are lower than the scores for TARSP and STAP, and this is very likely caused by the fact that the ASTA data contain no annotations for repetitions, false starts, filled pauses or for incorrect words at all, though such phenomena are mostly (though not always) explicitly marked in the TARSP and STAP data. The queries for detecting filled pauses, repetitions, false starts, and incomplete sentences are very complicated and currently score relatively low. This also affects the results for certain other queries (e.g. for nouns and main verbs), since words must be annotated for part of speech differently

²⁰We did not make the Bronze v. Silver comparison for these data yet.

²¹Measurement done on 2021-01-22.

²²For STAP we never received the reference annotations for sample_01, so we report only on the results for 9 samples.

TARSP	O v. B			O v. S			B v. S		
Sample	R	P	F1	R	P	F1	R	P	F1
Sample_01	85.1	72.2	78.1	87.2	86.1	86.6	85.9	100.0	92.4
Sample_02	87.1	64.3	74.0	87.9	69.0	77.3	92.9	98.9	95.8
Sample_03	77.3	63.6	69.8	81.4	81.4	81.4	82.2	100.0	90.2
Sample_04	93.2	81.2	86.8	93.9	90.0	91.9	90.8	100.0	95.2
Sample_05	83.2	69.1	75.5	85.6	83.0	84.3	85.0	99.3	91.6
Sample_06	75.0	56.8	64.7	79.8	75.0	77.3	79.8	99.0	88.4
Sample_07	86.2	66.7	75.2	88.7	80.0	84.1	82.3	96.0	88.6
Sample_08	77.7	63.2	69.7	82.7	86.5	84.6	77.7	100.0	87.4
Sample_09	89.4	69.0	77.9	91.4	87.0	89.1	80.6	99.3	89.0
Sample_10	80.3	69.1	74.3	84.1	89.7	86.8	79.5	98.6	88.1

Table 6: Results of Sastadev for the TARSP data (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1).

STAP	O v. B			O v. S			B v. S		
Sample	R	P	F1	R	P	F1	R	P	F1
Sample_02	78.6	79.0	78.8	81.1	92.2	86.3	86.8	98.2	92.2
Sample_03	91.7	85.7	88.6	92.0	89.8	90.9	92.0	96.1	94.0
Sample_04	92.5	93.4	93.0	92.7	95.3	94.0	97.7	99.5	98.6
Sample_05	92.6	85.5	88.9	93.2	92.8	93.0	92.3	99.5	95.8
Sample_06	92.3	88.5	90.4	93.0	97.9	95.4	91.0	100.0	95.3
Sample_07	91.6	92.4	92.0	91.9	96.0	93.9	95.3	98.7	97.0
Sample_08	94.4	79.7	86.4	95.0	90.5	92.7	87.8	99.0	93.0
Sample_09	95.6	84.5	89.7	96.1	95.3	95.7	88.3	99.1	93.4
Sample_10	91.5	80.8	85.8	92.5	92.5	92.5	88.4	100.0	93.8

Table 7: Results of Sastadev for the STAP data (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1).

when they are in a repetition or a false start. For example in a sequence *beits afbeits* ‘stain stain-off’ *beits* is analysed as a noun though it should have been analysed as a verb since it is a false start of the following word *afbeits*, which is a verb.

10 Concluding Remarks and Future Work

We have presented *Sasta*, an application to analyse transcripts of spontaneous language. Though the *Sasta* application applies only to Dutch, the techniques described here can be applied to any language provided there is a parser for that language and a query system for querying the syntactic structures resulting from the parser. We observe that SASTA scores pretty well on the grammatical analysis of transcripts of spontaneous language sessions. We also found that corrections of deviant language not only improve the deviant parts but also the overall analysis. SASTA also often finds more examples for a grammatical phenomenon than human annotators (who often overlook instances), but the human annotators remain superior in precision. Whether the quality of the grammatical analysis is good enough to make the whole process more efficient remains to be seen. With the VKL we will carry out experiments (starting in January 2021) in which *Sasta* will actually be used in the clinical setting so that we can assess this and optimally integrate *Sasta* into the normal workflow procedures of the hospitals and clinics. In addition, we have secured funding for a small successor project (SASTA+) in which we will investigate more advanced methods for the detection and correction of deviations, including cases in which all

ASTA	O v. B			O v. S			B v. S		
Sample	R	P	F1	R	P	F1	R	P	F1
Sample_01	77.0	78.3	77.6						
Sample_02	77.5	78.0	77.7	79.5	85.3	82.3	90.5	96.6	93.5
Sample_03	75.5	78.7	77.1						
Sample_04	79.3	83.4	81.3	81.7	91.0	86.1	92.3	97.7	95.0
Sample_05	73.8	76.4	75.1	78.2	91.8	84.4	86.9	98.5	92.3
Sample_06	95.6	83.8	89.3						
Sample_07	82.1	79.2	80.6						
Sample_08	89.4	78.5	83.6						
Sample_09	89.2	84.3	86.7						
Sample_10	78.8	82.3	80.5						

Table 8: Results of Sastadev for the ASTA data (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1). Silver references are currently available only for samples Sample_02, Sample_04 and Sample_05.

words in the utterance are correct and cases where multiple variants should be considered. The automatic corrections developed here can also be used to improve existing CHILDES CHAT annotation files, and we will create a side result of this work, a program to improve and enrich existing CHAT files.

Acknowledgements

I would like to thank the VKL working group members for contributing the data and providing us with their knowledge and expertise on the assessment methods, various other linguists who provided us with data (Mieke Stap, Wim Tops, Jacqueline van Kampen, Liesbeth Schlichting, Eliska Heukels) and my colleagues Jelte van Boheemen and Sjoerd Eilander. This research was funded by CLARIAH-PLUS (an NWO project, Grant 184.034.023), the Dutch Association for Clinical Linguistics (VKL) and the Dutch Language Technology Foundation (Stichting Taaltechnologie).

References

- [Augustinus et al.2012] Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odiijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- [Ball et al.2012] Martin J. Ball, David Crystal, and Paul Fletcher, editors. 2012. *Assessing Grammar: The Languages of LARSP*. Number 7 in Communication Disorders across Languages. Multilingual Matters, Bristol.
- [Bishop1984] D. V. M. Bishop. 1984. Automated LARSP: Computer-assisted grammatical analysis. *British Journal of Disorders of Communication*, 19(1):78–87.
- [Bol and Kuiken1989] Gerard Bol and Folkert Kuiken. 1989. *GRAMAT: Methode voor het diagnosticeren en kwalificeren van taalontwikkelingsstoornissen*. Berkhout, Nijmegen.
- [Bouma et al.2001] Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.
- [Boxum et al.2013] Elsbeth Boxum, Fennetta van der Scheer, and Mariëlle Zwaga. 2013. *Analyse voor Spontane Taal bij Afasie. Standaard in samenwerking met de VKL*. VKL, October. <https://klinischelinguistiek.nl/uploads/201307asta4eversie.pdf>.
- [Crystal et al.1976] D. Crystal, P. Fletcher, and M. Garman. 1976. *The grammatical analysis of language disability*. Edward Arnold, London.

- [Crystal et al.1989] D. Crystal, P. Fletcher, and M. Garman. 1989. *The grammatical analysis of language disability*. Cole and Whurr, London, 2nd edition.
- [Long et al.1996 2000] S.H. Long, M.E. Fey, and R.W. Channell. 1996–2000. Computerized profiling, versions 9.0.3-9.2.7 (ms-dos) [computer program]. Software, Department of Communication Sciences, Case Western Reserve University, Cleveland, OH.
- [MacWhinney2000] Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.
- [Odijk et al.2017] Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: <http://dx.doi.org/10.5334/bbi.23>. License: CC-BY 4.0.
- [Odijk et al.2018] Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. <http://aclweb.org/anthology/W17/W17-7608.pdf>.
- [Oostdijk et al.2002] N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodríguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.
- [Pander Maat and Dekker2016] Henk Pander Maat and Nick Dekker. 2016. Tekstgenres analyseren op lexicale complexiteit met TScan. *Tijdschrift voor Taalbeheersing*, 38(3):263–304.
- [Pander Maat2017] Henk Pander Maat. 2017. Zinslengte en zinscomplexiteit. *Tijdschrift voor Taalbeheersing*, 39(3):297–328.
- [Parijsse et al.2012] Christophe Parijsse, Christelle Maillart, and Jodi Tommerdahl. 2012. F-LARSP: A computerized tool for measuring morphosyntactic abilities in French. In Martin J. Ball, David Crystal, and Paul Fletcher, editors, *Assessing Grammar: The Languages of LARSP*, number 7 in Communication Disorders across Languages, chapter 13.
- [Schlichting2005] Liesbeth Schlichting. 2005. *TARSP: Taal Analyse Remediëring en Screening Procedure. Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar*. Pearson, Amsterdam, 7th edition.
- [Schlichting2017] Liesbeth Schlichting. 2017. *TARSP: Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar met aanvullende structuren tot 6 jaar*. Pearson, Amsterdam, 8th edition.
- [van Ierland et al.2008] Margreet van Ierland, Jeannette Verbeek, and Leen van den Dungen. 2008. *Spontane Taal Analyse Protocol. Handleiding van het STAP-instrument*. UvA, Amsterdam.
- [van Noord et al.2013] Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.
- [van Noord et al.2020] Gertjan van Noord, Jack Hoeksema, Peter Kleiweg, and Gosse Bouma. 2020. SPOD: Syntactic profiler of Dutch. *Computational Linguistics in the Netherlands Journal*, 10:129–145.
- [Verbeek et al.2007] Jeannette Verbeek, Leen van den Dungen, and Anne Baker. 2007. *Spontane Taal Analyse Protocol. Verantwoording van het STAP-instrument, ontwikkeld door Margreet van Ierland*. UvA.

Stimulating Lexicographical Knowledge Exchange via Trans-national Access – the ELEXIS Travel Grants as a Use Case

Sussi Olsen
University of Copenhagen,
Denmark
saolsen@hum.ku.dk

Bolette S. Pedersen
University of Copenhagen
Denmark
bspedersen@hum.ku.dk

Tanja Wissik
Austrian Centre for Digital
Humanities, Austria
Tanja.Wissik@oeaw.ac.at

Anna Woldrich
Austrian Centre for Digital
Humanities, Austria
Anna.Woldrich@oeaw.ac.at

Simon Krek
Jozef Stefan Institute,
Ljubljana, Slovenia
simon.krek@guest.arnes
.si

Abstract

This paper describes the intermediate outcome of one of the initiatives of the ELEXIS project: Transnational Access. The initiative aims at facilitating interaction between lexicographers/researchers from the EU and associated countries and lexicographical communities throughout Europe by giving out travel grants. Several of the grant holders have visited CLARIN centres, have been acquainted with the CLARIN infrastructure and have used CLARIN tools. The paper reports on the scientific outcome of the visits that have taken place so far: the origin of the grant holders, their level of experience, the kind of research projects the grant holders work with and the outcomes of their visits. Every six months ELEXIS releases a call for grants, the fourth call closed January 2020. Since then calls and visits have been suspended due to the COVID-19 situation. So far 23 visits have been granted in total; 13 of these visits have been concluded and the reports of the grant holders are publicly available at the ELEXIS website.

1 Background and Motivation

Even though lexicography has a long history of international research conferences, it has traditionally been a research area with limited knowledge exchange outside of each lexicographical institution, and in many cases lexicographic data has only been accessible to researchers from the institution who created the data and held the copyright. This tradition is partly related to the fact that practical lexicography has a strong commercial basis; lexicographical data used to be good business. But it also relates to the fact that enabling easy access to restricted data requires significant effort into facilitating and controlling this access - which again requires time and money not easily found in the budgets of lexicographic projects.

To this end, an important objective of the ELEXIS project is to stimulate knowledge exchange between lexicographical research facilities, infrastructures and resources throughout Europe, which can consequently mutually benefit from the vast experience and expertise that exist in the community. Inspired by other EU projects such as EHRI¹, RISIS², InGRID³, and sobigdata⁴, ELEXIS offers transnational access activities in the form of visiting grants that enable researchers, research groups and lexicographers to work with lexicographical data, which are not fully accessible online. Furthermore, grants offer access to professional on the spot expertise in order to ensure and optimise mutual knowledge exchange. Finally, grant recipients can gain knowledge and expertise by working with

This work is licenced under a Creative Commons Attribution 4.0 International Licence.
Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹ <https://ehri-project.eu/ehri-fellowship-call-2016-2018>

² <http://datasets.risis.eu/>

³ <http://www.inclusivegrowth.eu/visiting-grants>

⁴ <http://www.sobigdata.eu/access/transnational>

lexicographers and experts in NLP and artificial intelligence. The CLARIN infrastructure is one of the important infrastructures for these travel grant visits.

The trans-national access activities are expected to have a long-term impact specifically but not only for lesser-resourced languages, boost the network and infrastructure of the European lexicographic community, and facilitate future collaboration and knowledge exchange.

The objectives of the ELEXIS trans-national activities can be summarised as follows:

- to offer opportunities to researchers or research teams to access research facilities with an excellent combination of advanced technology and expertise
- to support training of new specialists in the field of e-lexicography in order to conduct high-quality research and ensure sustainability of the infrastructure
- to ensure support for excellent scholarly research projects and innovative enterprises and also support the complex multi-disciplinary research
- to encourage the integrative use of technology and methodologies as developed in ELEXIS and in the lexicographical institutions.
- to improve the overall services (lexicographic and technical) available to the research community
- to exchange knowledge and experience and to work towards future common projects and objectives
- to create an interdisciplinary community, collaborating on activities that are fully or partially of relevance to the proposed work of the grant holder
- to create knowledge at the interaction between academia and society

The trans-national activities represent a way of ELEXIS to enable access to restricted data, which has so far not been available outside of the hosting institutions, to researchers from other institutions and countries. As the results of research conducted in trans-national activities become available under open-access licenses, the international lexicographic community will become acquainted with previously inaccessible resources.

2 The Grants

The transnational activities consist of visiting grants of 1 to 3 weeks for researchers to experiment with and work on lexicographical data in a context of mutual knowledge exchange with the hosting institutions. Around five visiting grants are made available twice a year during the entire project period, amounting to 35-40 grants in total. However, since February 2020 calls and visits have been temporarily suspended due to the COVID-19 situation.

The following lexicographic institutions accept transnational visits during the ELEXIS project:

1. ELEXIS-SL: Institut Jozef Stefan (JSI, Slovenia)
2. ELEXIS-NL: Institute for Dutch Language (INT, The Netherlands)
3. ELEXIS-AT: Austrian Academy of Sciences (OEAW, Austria)
4. ELEXIS-RS: Belgrade Center for Digital Humanities (BCDH, Serbia)
5. ELEXIS-BG: Institute of Bulgarian Language Lyubomir Andreychin (IBL, Bulgaria)
6. ELEXIS-HU: Hungarian Academy of Sciences (RILMTA, Hungary)
7. ELEXIS-IL: K-Dictionaries (KD, Israel)
8. ELEXIS-DK: Det Danske Sprog- og Litteraturselskab, University of Copenhagen (DSL/UCPH, Denmark)
9. LEXIS-DE: Trier Center for Digital Humanities (TCDH, Germany)
10. ELEXIS-EE: Institute for Estonian Language (EKI, Estonia)
11. ELEXIS-ES: Real Academia Española (RAE, Spain)

Out of the 11 countries, where the hosting institutions are located, 8 countries participate in CLARIN and five out of the 11 hosting institutions are operating CLARIN B Centres.

2.1 The Calls

Researchers and lexicographers within the EU member states and associated countries are invited to apply for a visit of free access to and support from one of the lexicographical institutions.

The calls for applications include descriptions of the institutions and the lexicographical resources, tools, and expertise that are made available for the visitors. Researchers and lexicographers interested in visiting a particular host institution are encouraged to make motivated applications describing their background, the purpose of the visit etc.

2.2 Dissemination and Reporting

The calls are disseminated through the ELEXIS website⁵, mailing lists, newsletters, as well as through Facebook and Twitter. Additionally, attention to the open call(s) is drawn at presentations and/or booths at various conferences, and flyers are distributed to partners and the audience at relevant events. For the dissemination activities we used CLARIN channels such as mailing lists of national CLARIN consortia, the CLARIN Newsflash and the CLARIN Twitter channel.

Particular effort was invested in disseminating the ELEXIS travel grants via social media campaigns on Facebook and Twitter. For the first call, not only the call was advertised but also each hosting institution was presented in a separate post⁶, as seen in Figure 1. For the second, third and fourth call, only the call itself was advertised.

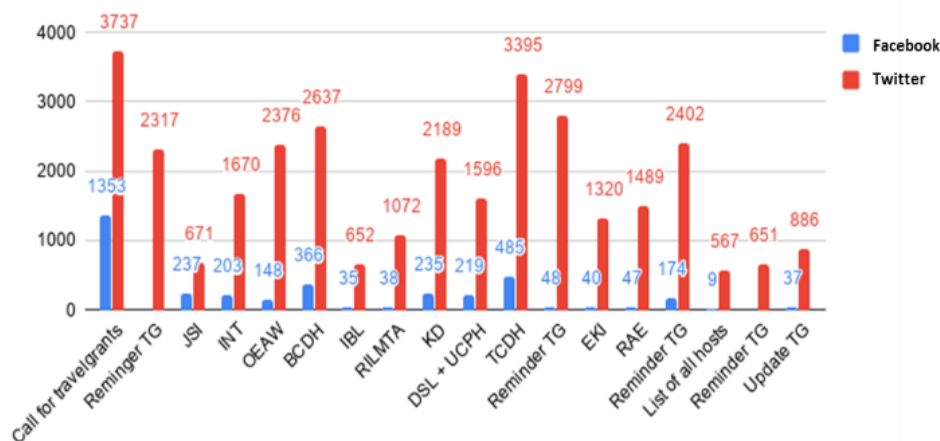


Figure 1. Reach (Facebook)/ Impressions (Twitter) of the dissemination campaign of the 1st call

Furthermore, we accompany the grant holders from the announcement through their travel visits with social media posts as well as website portraits as shown in Figure 2.

⁵ <https://elex.is/grants-for-research-visits/>

⁶ For instance, see: <https://twitter.com/i/events/1025308219126763520>

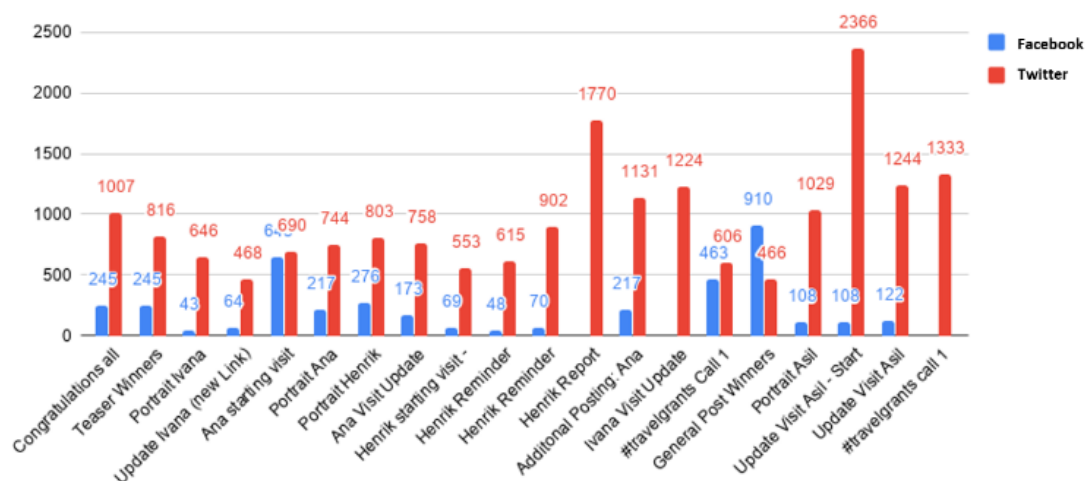


Figure 2. Reach of the dissemination campaign of grant holders of the 1st call.

Besides the website portraits that are published before the research visits in form of a written interview⁷, the grant holders report on their visits afterwards, and the reports are published at the ELEXIS website⁸ and on social media.

The promotion of the open calls via Facebook and Twitter resulted in 88 postings that inform about the open call for research grants (1, 2, 3 and 4) and 106 postings about the grant winners, their projects and journey.

Summing up, in the course of quarter 2 2018 until quarter 1 2020 ELEXIS EU published a grand total of 194 social media postings about the calls for research visit grants on both platforms.

3 Status after four Calls

Four calls for applications were launched and 23 grants were granted, of which 13 were completed, while the other 10 visits have been postponed due the COVID-19 situation. The fifth call has been postponed several times for the same reason and currently it is not to be foreseen when the next call can be launched.

The four calls received applications from Albania, Austria, Bulgaria, Croatia, Denmark, Georgia, Germany, Iceland, Ireland, Israel, Latvia, Poland, Portugal, Republic of North Macedonia, Russia, Slovakia, South Africa⁹, Spain, Turkey, and the UK. The final winners of the first four calls come from institutions located in the countries shown in Figure 3. The fact that the applications received originate from a wide range of countries proves that the transnational access program has indeed reached out to both the strong European lexicographic community as well as to communities that do not have an equivalent strong infrastructure.

⁷ <https://elex.is/category/grants-for-research-visits/>

⁸ <https://elex.is/travel-grant-reports/>

⁹ Since neither Russia nor South Africa are associated countries in the Horizon 2020 program, an application from these countries unfortunately cannot be granted.

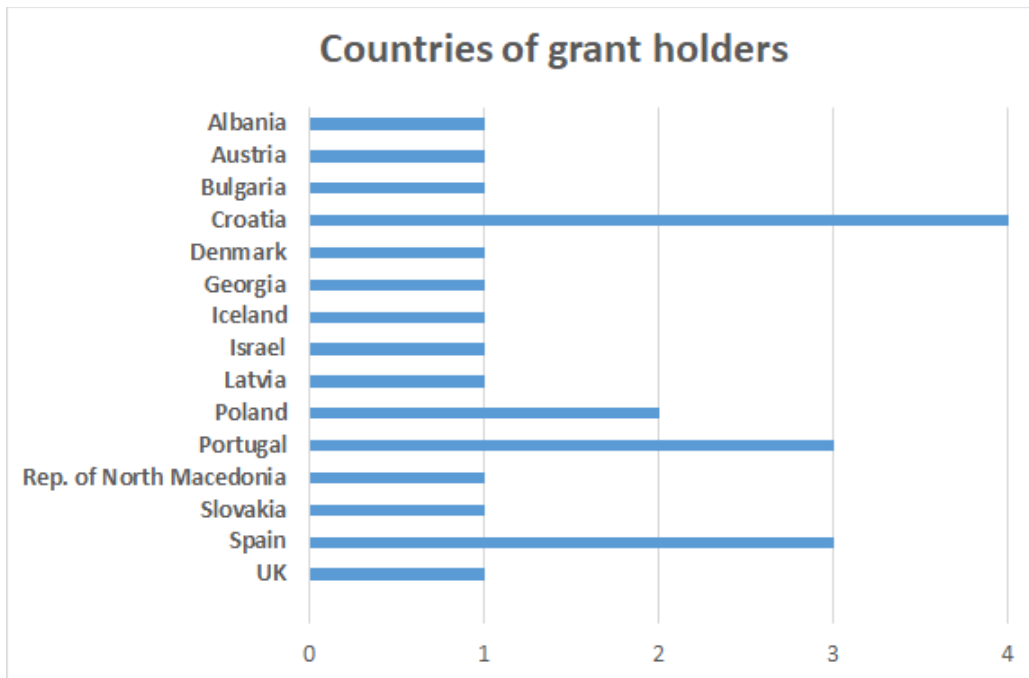


Figure 3. List of countries of the grant winners of the first four calls and number of winners from each country.

After four calls, most of the infrastructures have had one or more visits. Some infrastructures are much more applied to than others. However, since each host has a fixed budget for approximately three visitors, infrastructures that have reached their maximum number of visits and spent their budget are left out of the list of hosting infrastructures for upcoming calls. Hence, in order to make sure that visits are somewhat evenly distributed among infrastructures, there were cases where the Transnational Access Committee, which selects the winners of each call, has given priority to applications addressing less popular infrastructures on the condition that these applications were of sufficient quality.

In order to investigate the research experience of the grant winners, we divided the grant holders into three groups as can be seen in Figure 4.

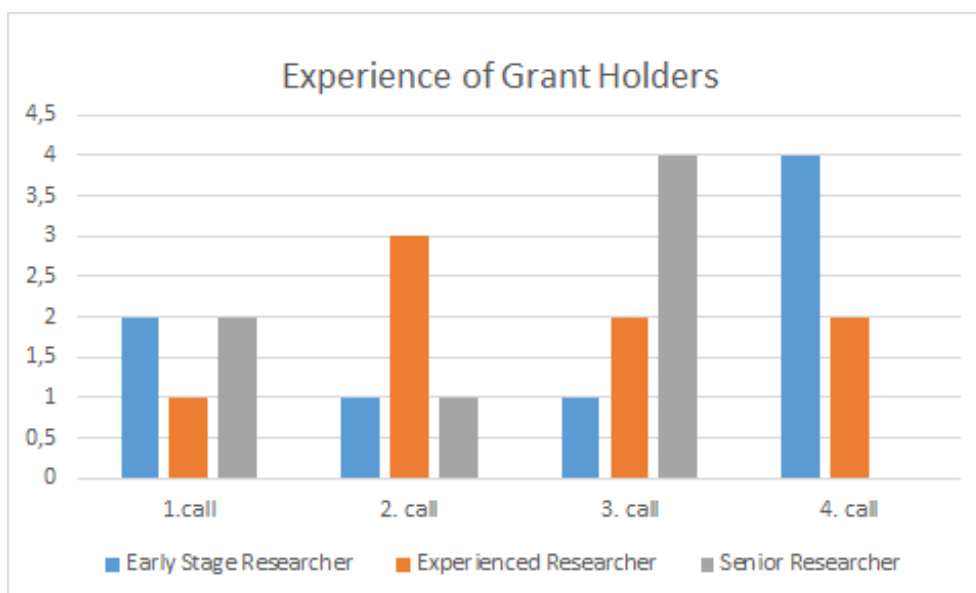


Figure 4. Experience of grant holders divided into three groups.

The early stage researchers span from students doing their master thesis to PhD scholars. In these cases, the visits are reported as having a career boosting effect. Researchers with a couple of years of research experience (after their PhD) belong to the second group, and the last group is defined by senior researchers or lexicographers with years of experience.

As shown in Figure 4, the grant holders represent a good mixture of different levels of experience: Eight early stage researchers, eight experienced researchers and seven senior researchers. At their research visit, most grant holders prove to have limited previous experience in the fields that they work in. Thus, not surprisingly, most of the experienced researchers apply for projects that gratify them with an upgrade of skills in areas that are not previously within their range of experience.

We also did a study of the gender distribution, see Figure 5.

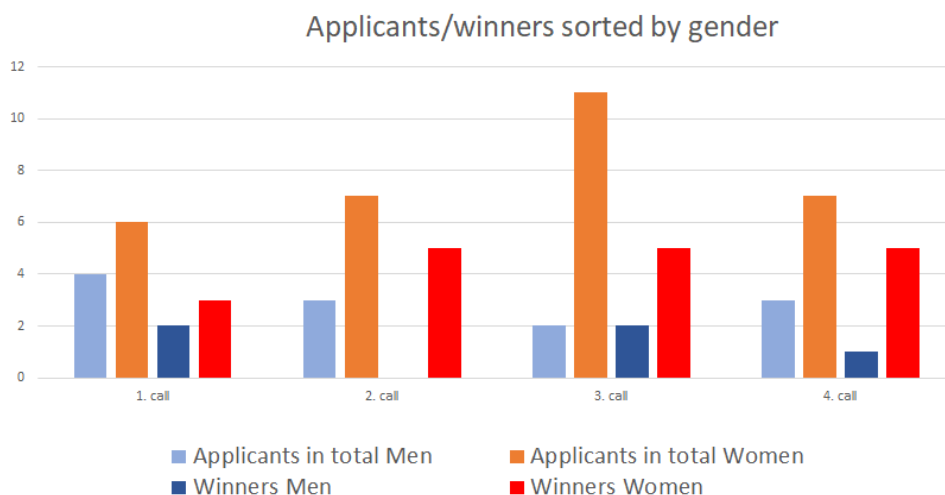


Figure 5. Gender distribution of applicants and grant winners.

We received more than twice as many female applications as male, 12/31. Of the 23 grants given in total, 18 winners were female, 5 male. The success rate has been 53 % on average; for female applications, the success rate is 58 % while only 41 % for the male applicants. It is a fact that lexicography is a field with a strong female representation but it is still a bit unclear what the reason can be for such a large discrepancy.

4 Scientific outcome of the visits

4.1 Research visit projects

Just as the grant holders come from a great variety of countries and communities, the topics of the grant holder projects are quite diverse. Most of the projects focus on the compilation of dictionaries of different kinds and the primary objective of most visits is to be acquainted with the hosts' dictionary writing systems, corpus tools, the methodology behind the dictionary and corpus compilation, standards such as TEI and ISO, and to discuss their own project with experienced lexicographers and terminologists.

From the reports of the visits carried out, the following summary can be made: two projects are about retro-digitalization of older dictionaries, a Latin-Bulgarian dictionary¹⁰ and a Croatian dictionary of literary language¹¹. The grant holders visited different hosting institutions but both gained knowledge

¹⁰ https://elex.is/wp-content/uploads/2019/10/Elexis_report_Elina_Boeva.pdf

¹¹ https://elex.is/wp-content/uploads/2019/03/Elexis-report-on-travel-grant_Ivana-Filipovic-Petrovic.pdf

of corpus digitalization, tools and methods of similar lexical projects as well as hands on experience with dictionary encoding in XML and introduction to the TEI guidelines for dictionaries.

Two projects deal with the creation of dictionaries in combination with terminological content: one proposes a multilingual and multimodal online dictionary combining resources of various kinds (Ramos, 2019), the other an English-Georgian bilingual dictionary of maritime terminology (Tenieshvili, forthcoming). The two grant holders were introduced to relevant terminological as well as lexical projects and to tools and systems that can facilitate their future work, including introduction to international standards such as the ISO standard for term definition.

Two other projects are of a more technical kind: One focuses on comparing the macro/microstructure of a new Portuguese dictionary with the structure of the Diccionario de la Lengua Española (RAE), to which inside access as well as communication with the lexicographers was necessary (Salgado, 2019). In this project, the focus was on the treatment of terminological content in general language dictionaries. The other project dealt with the optimization of methods for automatic extraction of data from a corpus and import into a dictionary writing system¹². Both researchers gained invaluable knowledge for their project thanks to the hosts' introductions to the tools and the various technical solutions.

Another project investigated Nordic e-dictionaries compared to the Croatian Web Dictionary, with a focus on the Danish Dictionary DDO during the visit¹³. The grant holder gained knowledge of the structure of the dictionary as well as the methodology behind, e.g. the lemma selection. The grant holder also participated in workshops about ethical dilemmas in dictionary writing e.g. potentially offensive content, a topic of current interest.

A rather different lexicographic project was a map based data visualization application that analyses the variants of the Spanish language¹⁴. This application could not have been carried out without access to the Dictionary of the Real Academia Española and assistance from its staff. Furthermore this project was also part of a Master's thesis, submitted in 2019¹⁵.

Yet another project dealt with business models of lexicography¹⁶: Based on earlier studies, the grant holder developed a general and more broadly founded business model by arranging several workshops and conducting research interviews at the hosting institution K-Dictionaries.

Common to all the visits are the introduction to dictionaries at the host institution and to the tools, systems and methodology behind. Several visitors received valuable introduction into dictionary writing and structuring including both the linguistic and the technical aspects. Another interesting topic was how to collect a well-balanced corpus for lemma selection.

In their reports all grant holders emphasize the importance of the communication and discussions with the experts of the hosting institution. Several participated in workshops and other events and everybody reports of an instructive and rewarding visit.

4.2 Further scientific dissemination and output related to the research visit projects

Besides the research visit projects and the travel grant reports that are available on the ELEXIS website¹⁷, the travel grant holders of the first call had the opportunity to present their projects during the poster session at the ELEXIS Observer Event in Vienna 2019. Furthermore, several of the research visit projects have led to scientific presentations or publications or were part of a Master thesis. Examples of this kind of output for the first two calls can be found in Woldrich and Wissik (2019).

4.3 Collaboration with and benefits for the CLARIN Network

Some of the ELEXIS travel grant recipients come from countries that are not yet part of CLARIN, hence, they were introduced to the infrastructure during their visit:

¹² https://elex.is/wp-content/uploads/2019/11/final-report_ELEXIS_TanaraZKuhn.pdf

¹³ <https://elex.is/wp-content/uploads/2019/08/Elexis-report-Daria-Lazic.pdf>

¹⁴ https://elex.is/wp-content/uploads/2019/05/Final_Report_ELEXIS_Grant_CETIN_Asil-1.pdf

¹⁵ Asil Çetin (2019). Multi-Faceted Visual Data Analysis for Corpus Research. Master's Thesis. University of Vienna. <http://othes.univie.ac.at/60378/1/64562.pdf>

¹⁶ https://elex.is/wp-content/uploads/2019/01/Report_Transnational_Research_Grant_Henrik_K%C3%B8hler_Simonsen_Research_Visit_Report_final_final.pdf

¹⁷ <https://elex.is/travel-grant-reports/>

The travel grant holder from Spain, who was visiting the Austrian Academy of Sciences in December 2019, was introduced to CLARIN AT and learned about all the facilities that CLARIN offers to researchers. Consequently a visit to the CLARIN K-Centre for Terminology Resources and Translation Corpora at the University of Vienna was organized. This was a perfect opportunity to exchange knowledge in the field of terminology with researchers involved and to learn more about the CLARIN K-Centre infrastructure.

A Croatian grant holder that visited the Society for Danish Language and Literature and University of Copenhagen was introduced to the CLARIN-DK infrastructure and CLARIN EU. Especially the CLARIN-DK NLP tools were of interest to her and with the help of CLARIN-DK staff, some tools from the CLARIN-DK toolbox, i.e. lemmatizer and pos-tagger, were trained for Croatian for the grant holder's future benefit.

During his visit in Ljubljana, a grant holder from the Republic of North Macedonia will be working with CLASSLA, the CLARIN knowledge centre for South Slavic languages. The objective of the visit is to obtain knowledge about corpora management software in order to be able to learn how to train POS taggers for Macedonian, and to create a corpus, ultimately to be used for corpus-based lexicographic work at the Macedonian Academy of Sciences and Arts. Thus it represents an ideal match between CLASSLA's offer of expertise on language resources and technologies for South Slavic languages, and the ELEXIS objectives of bridging the gap between more advanced and lesser-resourced lexicographic communities.

Thanks to the call for ELEXIS Travel Grants, various ELEXIS hosting institutions introduced and will continue to introduce CLARIN to a community that would not have approached a CLARIN centre due to a lack of knowledge. Hence we expect to observe a snowball effect in the future, where ELEXIS grant winners introduce CLARIN within their (national) research communities and approach a CLARIN centre thanks to ELEXIS acting as an intermediate. Through the ELEXIS travel grants, the usage of certain CLARIN tools and services, introduced during the ELEXIS research visits, might increase due to additional users and user scenarios.

In addition, ELEXIS aims to provide interoperability with CLARIN by forming an ELEXIS-CLARIN subgroup¹⁸ to prepare a strategy of integration of ELEXIS services into CLARIN at the end of the project (Summer 2022). We foresee that the subgroup will be formed by members of national CLARIN infrastructures, especially those with CLARIN B centres. Thus, ELEXIS sustainability will be enabled via national consortia to guarantee an afterlife for the efforts made.

5 Conclusion

We have presented the mid-term outcomes of the transnational access initiative of the ELEXIS project where 13 visits to 9 different lexicographical infrastructures in Europe have been completed to date.

The grant holders - be they early stage researchers or senior staff - sought to tailor their research visit in a manner that enabled them to gain new knowledge by physically visiting lexicographical milieus with specific expertise in certain topics and technologies that are highly relevant to their research.

The reports of the grant holders clearly show that the travel grants serve several purposes: the above-mentioned gaining of new knowledge and the network building and knowledge exchange are obvious results. For the individual visiting researchers, however, the visits also serve as a career boost either by helping the early stage researchers establish themselves in the field or by leading the more experienced ones towards new fields. The fact that many experienced researchers apply for a grant to deepen their knowledge and gain new expertise shows that the travel grants meet an existing need not covered by other initiatives.

ELEXIS has a network of 52 observer institutions¹⁹ that benefit from early access to newly developed tools and services, as well as to activities aimed at improving and enriching their own lexicographic data. In the upcoming calls, we expect to receive more applications from these observer institutions. Furthermore, the observing institutions that have lexicographic data, will be invited to join in the network of ELEXIS infrastructures who host the travel grants. They will not receive compensation for work at the institution but visitors will be compensated in the same manner as when visiting existing infrastructures.

¹⁸https://elex.is/wp-content/uploads/2019/04/ELEXIS_03_Observer_Session5_ELEXIS_CLARIN-DARIAH.pdf

¹⁹ <https://elex.is/observers/>

At the time of writing, the second half of the ELEXIS grant activities has been put on hold due to the covid-19 pandemic. Consequently, new calls are being postponed, and already granted and planned visits have not yet been completed. Hopefully, in autumn 2021 it will again be possible to travel around Europe to complete fruitful exchange of lexicographical knowledge. 9 out of 10 of the grant holders who had their visits postponed, claim that they are still very interested in visiting the host institutions when it will once more be possible.

Most presumably, the future grant visits will follow the same line as the previous ones. However, we expect to see an increased interest in the integrative use of the lexicographical tools, methodologies and resources that are just currently being developed and made available through ELEXIS, e.g. the automated linking tool NAISC (McCrae, 2018) and Lexonomy, an open-source platform for writing and publishing dictionaries (Měchura, 2017), which several grant holders already report having worked with.

In conclusion, the ELEXIS travel grants are an opportunity for the CLARIN infrastructure to get known in a community that without ELEXIS would not approach a CLARIN centre. When the project ends we aim at establishing full interoperability with existing CLARIN centres, enabled via national consortia.

References

- McCrae, J. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*. 18. 109-123. 10.2478/cait-2018-0010.
- Měchura, M. B. (2017). 'Introducing Lexonomy: an open-source dictionary writing and publishing system'. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, 19-21 September 2017, Leiden, The Netherlands.
- Olsen, S. & Pedersen, B.S. (2020). D.9.2 Report on trans-national access – year 2. Available at: https://elex.is/wp-content/uploads/2020/07/ELEXIS_D9_2_Report-on-TNA-2.pdf
- Ramos, M., Costa, R. & Roche, C. 2019. Dealing with specialized co-text in text mining: The terminological verbal collocations. In C. Roche, ed. Terminology and Text Mining, TOTh 2019 Proceedings. (https://www.academia.edu/39887099/Dealing_with_specialized_co-text_in_text_mining_The_terminological_verbal_collocations).
- Salgado, A. and Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. In *III Jornadas internacionales sobre investigaciones lexicográficas y lexicológicas (inLÉXICO2019)*, 4–5 abril, 2019, Universidad de Jaén, Spain.
- Tenieshvili, A. (forthcoming). Why It Is Necessary to Create and Adopt Georgian Maritime Terminology?. In *Journal of International Terminology: Translation and Standardization*. Batumi State University, Georgia.
- Woldrich, A. and Wissik, T. (2019). D7.5 First Year Communication and Dissemination Report and Updated Communication Plan. Available at https://elex.is/wp-content/uploads/2019/03/ELEXIS_D7_5_First_year_dissemination_and_communication_report_and_updated_communication_plan.pdf
- Woldrich, A. and Wissik, T. (2020). D7.6 Second Year Communication and Dissemination Report. Available at https://elex.is/wp-content/uploads/2020/05/ELEXIS_D7_6_Second_Year_Dissemination_and_Communication_Report.pdf

An Internationally Fair Mediated Digital Discourse Corpus: Improving Knowledge on Reuse

Rachel Panckhurst

Dipralang EA 739
Université Paul-Valéry Montpellier 3
Montpellier, France

rachel.panckhurst@univ-montp3.fr

Francesca Frontini

Istituto di Linguistica Computazionale
“A. Zampolli” - ILC - CNR
Pisa, Italy

& CLARIN ERIC

francesca.frontini@ilc.cnr.it

Abstract

In this paper, the authors present a French Mediated Digital Discourse corpus, (*88milSMS* <http://88milSMS.huma-num.fr> <https://hdl.handle.net/11403/comere/cmr-88milSMS>). Efforts were undertaken over the years to ensure its publication according to the best practices and standards of the community, thus guaranteeing compliance with FAIR principles and CLARIN recommendations with pertinent scientific and pedagogical reuse. Since knowledge on how resources are reused is sometimes difficult to obtain, ways of improving this are also envisaged.

1 Introduction

The adoption of Open Data and Open Science principles is producing important effects in SSH disciplines and has been enhanced by widespread awareness of the internationally ratified FAIR principles, aiming at ensuring that research data should be Findable, Accessible, Interoperable and Reusable¹. In Linguistics and Natural Language Processing (NLP) the importance of curating Language Resources (LRs) for replicability and reuse has been particularly recognized, with relevant initiatives dating back several decades. Even before the formalisation of the FAIR principles as such, various initiatives have promoted good data management practices in the domain of Language Resources, starting with the FLareNet² action and culminating with the creation of the CLARIN infrastructure. With its network of consortia and centres, CLARIN is making it easier for researchers to adhere to the requirements of the FAIR principles (de Jong et al., 2018), something which is increasingly required by evaluation and funding agencies. In France, over and above its role as CLARIN observer, various national centres are now active under the leadership of the national Huma-Num infrastructure, offering services and promoting the sharing of textual data (among other types) which meet the FAIR principles and the CLARIN best practices.

European Computer-Mediated Communication (CMC) and Mediated Digital Discourse (MDD) corpora initiatives are becoming more visible: Belgian *sms4science*, *Vos Pouces*, (Fairon et al., 2006; Cougnon, 2015; Cougnon and Fairon, 2014; Cougnon et al., 2017); Dutch SoNaR, (Oostdijk et al., 2008); French CoMeRe, (Chanier et al., 2014); German DeRik, (Beißwenger et al., 2013); Swiss *What’s up Switzerland?*, (Ueberwasser and Stark, 2017; Frey et al., 2016). These data types are often difficult to process, standardize, analyze, owing to their complex nature, including ‘noisy’ content (Frey et al., 2019; Poudat et al., 2020).

The objective of this paper is to present *88milSMS*, a French CMC/MDD corpus with a focus on scientific and pedagogical reuse. Over the years the authors of the corpus have undertaken several efforts to ensure its publication according to the best practices and standards of the community, which in turn guarantee compliance with FAIR principles and CLARIN recommendations. However assessing how this translates into an increased usability is a cumbersome task. Knowledge on how resources are reused

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<http://datafairport.org/fair-principles-living-document-menu>

²<http://www.flarenet.eu/>

is sometimes difficult to obtain. After having provided surveys on reuse and analysed data emanating from them, the authors provide ideas on how to improve access to information about what people are doing with corpora when they reuse them.

2 Project & Corpus

The *sud4science* project³ was part of a vast international initiative, entitled *sms4science*⁴, which aimed at building a worldwide database and analysing authentic text messages in different languages — mainly French, but also Creole, German (written in Switzerland and Germany), Italian, Romansh (Dürscheid and Stark, 2011), and English (Drouin and Guilbault, 2016). Many scientific projects analyse authentic data, but ensuing corpora are not always made available for the scientific community and the general public, sometimes owing to legal requirements and commercial issues. However, there is a crucial need for researchers from a wide range of disciplines to have easy access to authentic data, in order to conduct analyses pertaining to their particular research fields. From the onset of the *sud4science* project, the possibility of easy access and reuse of authentic data was of utmost importance to the scientific team.

In 2011, over 88,000 authentic French text messages were collected during a 13-week period from the general public in Montpellier, France (Panckhurst et al., 2014; Panckhurst et al., 2016b) and SMS ‘donors’ were also invited to fill out a sociolinguistic questionnaire (Panckhurst and Moïse, 2014). An anonymization phase was conducted (Patel et al., 2013), owing to legal requirements for data-protection of private data (Ghliss and André, 2017). This involved anonymizing names, telephone numbers, places, brand names, addresses, codes, URLs⁵. In 2014, the finalised largest digital resource of 88,000 ‘raw’ anonymized French text messages, the specific *88milSMS* corpus, two samples (1,000 transcoded SMS, 100 annotated SMS), and the sociolinguistic questionnaire data were made available for all⁶ to download. The researchers chose the Huma-Num web service⁷, then in 2016, they made a TEI/XML version of *88milSMS* available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence on the ‘Ortolang’ platform, which provided the corpus with a citable persistent identifier⁸. Contributions to DARIAH and ELRA were also made in 2015, and *88milSMS* therefore has an ISLRN. These initiatives preceded the FAIR principles but are in strict alignment with them.

3 Towards Being FAIR

Findability refers to initiatives aimed at ensuring long-term preservation of LRs by depositing in a specialized data centre (Ortolang repository, France), documented by a rich and standardized set of metadata, which in turn can be harvested by international meta-catalogues (CLARIN Virtual Language Observatory), allowing international visibility. Thanks to the deposit on the Ortolang repository, the *88milSMS* corpus is **Findable** from the VLO⁹, where it will appear by performing free text searches such as “cmc corpus” or “SMS corpus” and filtering by language. In addition to the visibility within the VLO, the corpus is indexed on Google, searchable on the ELRA catalogue¹⁰, as well as on Isidore¹¹.

Accessibility is gained by adopting a clear set of licences and promoting open access within copyright limits and data protection regulations; single-sign-on technologies allow researchers to gain access to resources based on their institutional identifier. *88milSMS* is fully **accessible**, despite its sensitive content, thanks to the thorough anonymization and verification work (Ghliss and André, 2017) carried out with the help of the University’s legal advisors; a short mandatory form needs to be completed to download from

³<http://sud4science.org>; (Panckhurst et al., 2016b).

⁴<http://www.sms4science.org>; (Fairon et al., 2006; Coughon and Fairon, 2014; Coughon, 2015).

⁵By default, first names, surnames and any data which enable identifying information are anonymized. It is of course frustrating for linguists and other scientists to feel that anonymization causes loss of information, which will not be able to be retrieved at a later stage, but it is a stringent legal requirement.

⁶Both the scientific community and the general public.

⁷<http://88milSMS.huma-num.fr> (Panckhurst et al., 2014), .

⁸<https://hdl.handle.net/11403/comere/cmr-88milSMS> (Panckhurst et al., 2016a).

⁹<https://vlo.clarin.eu>

¹⁰<https://catalogue.elra.info>

¹¹Isidore is a French search engine for documents and resources in SSH <https://isidore.science/document/http://hdl.handle.net/11403/COMERE/V3.3/CMR-88MILSMS>

the bilingual (French/English) Huma-Num web interface, with a user free-of-charge licence. However, no authentication or form completion are needed to download the corpus from Ortolang, where it is available via the Creative Commons CC BY 4.0 licence.

Interoperability and **re-usability** for LR are particularly important, and crucially enabled by the use of standard annotation formats and common best practices, allowing researchers to exploit data from different projects. At the time of the data collection, similar initiatives took place at the international level (*cf.* § 1) thus making the overall philosophy of the corpus attuned to that of these other datasets. Indeed, other authentic data collections projects followed on in more recent years (Ueberwasser & Stark 2017; Coughon *et al.* 2017). From the point of view of encoding, initial formats of the corpus were .ods spreadsheets and *ad hoc* .xml. The use of utf-8 was crucial at the time (2011), in particular to ensure the preservation of a subset of SMS containing the first instances of emoji (Panckhurst and Frontini, 2020). The work to make *88milSMS* fully **interoperable** was carried out later, with the inclusion within the CoMeRe initiative, where the project adopted a common TEI format. Thanks to the aforementioned efforts, *88milSMS* has been **reused** beyond the initial scope of the project, boasting 1,067 downloads from 52 countries (as of 01/01/2021), and with a broad spectrum of multidisciplinary applications (to name a few: language sciences, computational linguistics and text-mining processing initiatives, geographical place name identification, psychology case studies).

4 Towards Scientific and Pedagogical Reuse

4.1 Initial 2017 Survey

Three years after providing *88milSMS* for public download and dissemination, a survey on scientific usage of the corpus was conducted¹². Results have shown a strong disciplinary tendency towards language sciences and computing including NLP, text mining and corpus linguistics research, mainly from higher education establishments. In terms of dissemination, 50% of the research cited was successfully circulated in Master's theses, PhDs, habilitations, books, articles, proceedings, etc. (Panckhurst *et al.*, 2020).

4.2 Follow-up 2019 Survey

In 2019, an update survey was conducted in order to find out if colleagues had cited/used *88milSMS*. Responses were unfortunately minimal but they do indicate that the corpus is being used in language sciences, as is to be expected, but also in other disciplines:

1. Language Sciences and NLP:
 - university courses for 2nd-year students; identifying and improving spelling mistakes (Poitiers University); discourse genres (Lorraine University);
 - recent PhDs: French as a foreign language and how to include SMS-writing in didactic situations; linguistic analysis of French SMS-writing; SMS communication: NLP and information extraction;
 - qualitative comparative analysis between differing corpora, related to morphosyntactic French question-form usage and interactional aspects comparing SMS and oral language.
2. Geography: identification of place names and interpretation of variations (Master's 2 internship subject, 2019, IGN-Paris & Paris-Est Marne-la-Vallée University).
3. Psychology: digital communication and teenagers (relational, emotional romantic aspects, 12-16 year-olds, Master's 1 thesis 2019, Toulouse Jean-Jaurès University).

¹²Researchers and the general public who had signed up to an optional scientific newsletter were contacted.

4.3 Final Questionnaire

A further (third) survey was proposed to the scientific community in November-December 2020 (cf. Appendix for questions) during a three-week period, via three channels:

1. Optional scientific newsletter which had been used for the previous 2019 survey;
2. French natural language list, LN¹³;
3. North American Linguist¹⁴ list.

The respondents' countries were varied across the continents, although only a handful (8 countries out of 52 cited in the download forms) were mentioned: Australia, China, France, Germany, India, Serbia, Switzerland, Uruguay, with an understandably high percentage from France (47%). Students make up 38% of the respondents (undergraduate and graduate 23%, doctoral 15%); another 54% are teachers/researchers from a public institution, and 8% are engineers. The *88milSMS* corpus was used for the following research fields¹⁵:

- Linguistics (75%);
- NLP (25%);
- Psychology (8%);
- Other¹⁶ (8%).

Within research fields, six types of tasks were mentioned¹⁷:

- Linguistic analyses (60%);
- Psychological analyses (10%),
- Statistical studies (10%);
- Corpus linguistics studies (40%);
- Resource building (30%);
- Data mining (20%).

In 50% of cases, the research has been disseminated¹⁸ in internship reports (20%), PhD manuscripts (10%), scientific publications, including articles, chapters, books (20%). In another 40%, the research has not been disseminated, and finally, 30% mention 'other' without specifying any further information. When respondents indicate that they have cited the *88milSMS* corpus, it is in pedagogical situations and/or directly in the bibliography of some of their publications, via the official citation format either mentioning (Panckhurst et al., 2014) or (Panckhurst et al., 2016a). Interestingly enough, a very high percentage (78%) of the respondents did not know that a second version of the corpus (XML/TEI) had been released on Orlolang¹⁹ in 2016! This seems to indicate that there is a lack of information on new versions of LRs being published, which would be important to address. Among those who knew about the latter version, reasons for downloading it were stipulated: "I want to conduct the study of linguistics

¹³https://www.atala.org/liste_ln

¹⁴<https://linguistlist.org/>

¹⁵ Respondents could cite more than one field, hence percentages indicate more than 100% overall.

¹⁶ Respondents did not necessarily mention which other research field was pertinent.

¹⁷ Again, several boxes could be ticked, hence a total of more than 100% is obtained. One item indicated in the survey, "Sociological analysis", was not used for any tasks.

¹⁸ The percentages also go beyond 100% owing to checking several boxes, and two items were not chosen at all: "Yes, through software" and "Yes, with the publication of a resource or dataset".

¹⁹ (<https://hdl.handle.net/11403/comere/cmr-88milSMS>)

area, and I need more up-to-date information”, “I’m going to download the xml-tei version to be able to use it with TXM”²⁰. The final survey question was added for any further information respondents wished to submit. Answers were given by 24% and mainly stipulated pedagogical usage by teachers (“I also use it for pedagogical purposes as an example of data constitution. I hope to have more opportunities to use it also in my publications.”) and students (“Thank you very much for the linguistic corpus, it was very useful for my presentation on the French language.”).

4.4 What Are the Difficulties Linked to Obtaining Reuse Information?

Unfortunately, as in 2019, only a minimal number of questionnaires were completed (corresponding, in terms of figures, to just under 5% of those having signed up for the scientific *88milSMS* newsletter). Firstly, there is no way of knowing which of the three channels was preferred for accessing the survey. Secondly, the worldwide COVID pandemic combined with end-of-the-year obligations may well have negatively affected the number of responses. Finally, it is difficult to assess whether respondents are, on the one hand, actually just not interested in providing this information to resource producers, or, on the other hand, if they really are too overloaded with many more urgent matters. In the case of the *88milSMS* corpus, there are two more factors which may contribute to low numbers of questionnaire responses:

1. Only 37% of those researchers having downloaded the first version of the corpus from the HumaNum platform have signed up to the optional scientific newsletter, which is the main contact method²¹;
2. Downloading the second version of the corpus from the Ortolang platform does not require authentication or mandatory form completion.

4.5 How Do We Gain More Accurate Knowledge on Reuse?

Knowledge on resource reuse is, at times, difficult to obtain, as indicated above. Yet it is essential for resource producers to have access to this, in order to tailor future productions and provide useful person-machine interfaces for users. So the authors definitely need to improve the way in which corpus reuse information is obtained. But how can this be done? If the corpus is reused on another online platform, it might be possible to collaborate with software engineers to gain information (all the while respecting legal aspects) on how often the resource is used and what are the types of formulated queries.

Currently, tools such as the LRE (Language Resources and Evaluation) map have tried to systematically document resource citations in papers²², and proposal for a Language Resource Impact Factor was made some years ago (Mariani and Francopoulo, 2015). However such approaches are incomplete, since they measure the impact of papers only on a limited number of sources. For instance the LRE map only gathers data from a set of conferences (COLING, IJCNLP, Interspeech, LTC, ACLHT, O-COCOSDA, RANLP) and one journal (the LRE Journal) and mentions of resources outside this scope are not referenced. As a consequence, for instance no reference to *88milSMS* can be found on the LREmap, which is surprising, given that the corpus has existed on the ELRA website since February 2015²³.

In the long term, infrastructures such as CLARIN could provide systematic information to researchers concerning the reuse and citation of their data, for instance by promoting data citation practices and systematically collecting information from publications and other sources; of course, practices should also be harmonised across disciplines, within the SSH and beyond, so that reuse beyond disciplinary boundaries can be detected.

In the current panorama, various initiatives are on-going to promote and systematise data citation²⁴.

²⁰TXM a corpus exploration tool widely used in France in the field of SSH <http://textometrie.ens-lyon.fr/>

²¹Download numbers of the *88milSMS* 2014 corpus from the HumaNum platform, as of 1/1/2021: 1,067 total downloads: 675 persons (63%) having not enrolled to receive the newsletter; 392 persons (37%), on the contrary, having enrolled to receive it.

²²<https://lremap.elra.info/>

²³<https://catalog.elra.info/en-us/repository/browse/ELRA-W0082/> ISLRN: 024-713-187-947-8 ID: ELRA-W0082

²⁴See for instance the Declaration on Data Citation principles <https://www.forcell.org/datacitationprinciples>

Within the SSHOC project, which aims at building the Social Sciences and Humanities Open Cloud, work is under way on how to identify, describe and collect data citations (Larrousse et al., 2019).

Given that standards for data citation are still under construction, efforts by individual researchers or researcher teams to manually collect evidence of data reuse can provide ground truth (such as the surveys mentioned above), indicating what type of information researchers are more keen to obtain.

5 Conclusion and Future Work

In this paper, the authors indicated how the French Mediated Digital Discourse *88milSMS* corpus complies with the four FAIR principles (findability, accessibility, interoperability and reusability), also taking into account CLARIN recommendations. In particular, since reusability is an important goal of open and collaborative research, we concentrated on addressing the reusability factor (§4) by analysing various survey results and providing more in-depth discussion about difficulties and knowledge linked to reuse. FAIR principles, which were considered in §3, could be indeed adapted after future research is conducted on other CMC corpora.

Even though the corpus is fairly widely consulted, downloaded and used across the scientific community and beyond, it remains difficult to have sufficient access to other researchers' scientific and pedagogical reuse, despite implementation of an optional scientific newsletter and surveys.

The authors' own pedagogical usage at Université Paul-Valéry Montpellier 3 (under-graduate and post-graduate levels) for Language Science students includes studying discourse analysis and NLP techniques with contemporary instant messaging authentic data such as the *88milSMS* corpus, which has recently been incorporated on the widely consulted Sketch Engine²⁵ platform, thus allowing online analysis via a user-friendly interface without mandatory downloading²⁶. The advantage of this sort of integration is to provide ever-increasing interdisciplinary scientific and pedagogical reuse possibilities.

In this sense, the collaboration on CMC corpora which has already started within CLARIN is crucial²⁷ for the harmonization of formats across international projects, for the identification of common technical solutions for browsing interfaces, and finally for the implementation of a Federated Content Search. Moreover, the centralised and curated access to different projects on similar themes which the CMC 'Resource Family' provides, allows to easily find comparable data from other projects. For instance, it would be interesting to compare *88milSMS/sms4science/CoMeRe*, (Panckhurst et al., 2014; Panckhurst et al., 2016b; Fairon et al., 2006; Cougnon, 2015; Cougnon and Fairon, 2014; Chanier et al., 2014) with the French sub-corpus of WhatsApp messages from *What's up Switzerland?* (Ueberwasser and Stark, 2017; Ueberwasser, 2017)²⁸, or with the FaceBook, Viber, WhatsApp messages from *vos pouces* (Cougnon et al., 2017), or even with daily writing during WW1 in *Corpus14* (Praxiling - UMR 5267, 2019), in order to see how communication has evolved over time and medium.

Finally, an important step for the preservation of *88milSMS* will be its long-term archiving, together with other FAIR corpora from French CLARIN repositories, at the National Computing Center for Higher Education (CINES)²⁹, thus providing insight on digital textuality usage for future generations.

²⁵<http://www.sketchengine.eu/>

²⁶One of the authors used Sketch Engine in an introductory NLP undergraduate course during the autumn 2020 semester, and the students were very enthusiastic about the easy-to-use interface, which they immediately adopted and preferred, compared to other corpus linguistics tools. In actual fact, she presented the platform rather hastily during an online synchronous class during COVID lockdown, so she was pleasantly surprised by the quality of the data analysis the students provided in their end-of-term assignments on *88milSMS* queries and regular expression usage.

²⁷See for instance the past thematic event <https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources> as well as the CMC 'Resource Family' entry <https://www.clarin.eu/resource-families/cmc-corpora>

²⁸The latter project has recently provided open access to the corpus. See here for a general overview of the Swiss CMC-corpora initiative: <https://cmc-corpora.ch/>

²⁹<https://www.cines.fr/en/>

References

- Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537. Publisher: Oxford Academic, <https://academic.oup.com/dsh/article/28/4/531/1077484>.
- Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J., and Seddah, D. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *JLCL - Journal for Language Technology and Computational Linguistics*, 29(2):1–30. <https://halshs.archives-ouvertes.fr/halshs-00953507>.
- Cougnon, L.-A. and Fairon, C., editors. 2014. *SMS Communication*. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Cougnon, L.-A., Maskens, L., Roekhaut, S., and Fairon, C. 2017. Social media, spontaneous writing and dictation. Spelling variation. *Journal of French Language Studies*, 27(3):309–327. <https://www.cambridge.org/core/journals/journal-of-french-language-studies/article/div-classtitle-social-media-spontaneous-writing-and-dictation-spelling-variation/div/9574CD6BF604BD8F866A270E1EC909A3>.
- Cougnon, L.-A. 2015. *Langage et sms: Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.
- de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1515>.
- Drouin, P. and Guilbault, C. 2016. De ‘Viens regarder la partie avec moi’ à ‘Come regarder the game with me’. In *Abstracts, PLIN 2016*, Louvain-la-Neuve, Belgium.
- Dürscheid, C. and Stark, E. 2011. sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin, T. and Mroczek, K., editors, *Digital Discourse. Language in the New Media*. Oxford University Press. ISBN: 9780199795437, <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199795437.001.0001/acprof-9780199795437-chapter-14>.
- Fairon, C., Klein, J. R., and Paumier, S. 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*. Presses universitaires de Louvain. Manuel.CD-Rom., Louvain-la-Neuve.
- Frey, J.-C., Glaznieks, A., and Stemle, E. W. 2016. The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Corazza, A., Montemagni, S., and Semeraro, G., editors, *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016. 5-6 December 2016 Napoli*, pages 157–161, Torino. Academia University Press. <https://bia.unibz.it/handle/10863/8949>.
- Frey, J.-C., König, A., and Stemle, E. W. 2019. How FAIR are CMC corpora? In Longhi, J. and Marinica, C., editors, *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora19)*, Cergy-Pontoise University, France, 9-10 September 2019, pages 26–31. <https://bia.unibz.it/handle/10863/11294>.
- Ghliiss, Y. and André, F. 2017. Après la collecte, l’anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88milSMS. In Ciara R. Wigham, G. L., editor, *Corpus de Communication Médinée par les Réseaux*, pages 71–84. L’Harmattan, Paris. <https://hal.archives-ouvertes.fr/hal-01722169>.
- Larrousse, N., Broeder, D., Brase, J., Concordia, C., and Kalaitzi, V. 2019. SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning, December. Final version - Approved by the European Commission.
- Mariani, J. and Francopoulo, G. 2015. Language Matrices and a Language Resource Impact Factor. In Gala, N., Rapp, R., and Bel-Enguix, G., editors, *Language Production, Cognition, and the Lexicon*, pages 441–471. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-08043-7_25.
- Oostdijk, N., Reynaert, M., Monachesi, P., Noord, G. V., Ordelman, R., Schuurman, I., and Vandeghinste, V. 2008. From D-Coï to SoNaR: a reference corpus for Dutch. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/pdf/365_paper.pdf.
- Panckhurst, R. and Frontini, F. 2020. Evolving interactional practices of emoji in text messages. In Thurlow, C., Dürscheid, C., and Diémoz, F., editors, *Visualizing Digital Discourse. Interactional, Institutional and Ideological Perspectives*, pages 81–103. De Gruyter Mouton.

- Panckhurst, R. and Moïse, C. 2014. French text messages. From SMS data collection to preliminary analysis. In Cougnon, L.-A. and Fairon, C., editors, *SMS Communication. A Linguistic Approach*, pages 141–168. John Benjamins. <https://hal.archives-ouvertes.fr/hal-01485595>.
- Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. 2014. 88milSMS. A corpus of authentic text messages in French, produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8, <https://hal.archives-ouvertes.fr/hal-01485560>.
- Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. 2016a. 88milSMS. A corpus of authentic text messages in French. In Chanier, T., editor, *Banque de corpus CoMeRe*. Nancy, France. Ortolang, <https://hdl.handle.net/11403/comere/cmr-88milSMS>.
- Panckhurst, R., Roche, M., Lopez, C., Verine, B., Détrie, C., and Moïse, C. 2016b. De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire. *Histoire Epistémologie Langage*, 38(2):63–82. <https://hal.archives-ouvertes.fr/hal-01485577>.
- Panckhurst, R., Lopez, C., and Roche, M. 2020. A French text-message corpus: 88milSMS. Synthesis and usage. *Corpus [online]*, (20). <http://journals.openedition.org/corpus/4852>.
- Patel, N., Accorsi, P., Inkpen, D., Lopez, C., and Roche, M. 2013. Approaches of anonymisation of an SMS corpus. In *Proceedings of CICLING 2013, LNCS*, pages 77–88, March 24–30, 2013, University of the Aegean, Samos, Greece. Springer-Verlag. <https://hal-lirmm.ccsd.cnrs.fr/lirmm-00816285>.
- Poudat, C., Wigham, C. R., and Liégeois, L. 2020. *Corpus complexes. Traitements, standardisation et analyse des corpus de communication médiée par les réseaux*. Corpus (20).
- Praxiling - UMR 5267. 2019. Corpus 14. ORTOLANG (Open Resources and TOols for LANGUAGE). <https://hdl.handle.net/11403/corpus14/v2>.
- Ueberwasser, S. and Stark, E. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5). <https://bop.unibe.ch/linguistik-online/article/view/3849>.
- Ueberwasser, S. 2017. What's up, switzerland?: Challenges of a large, multilingual cmc corpus. CLARIN-PLUS workshop "Creation and Use of Social Media Resources", Kaunas, Lithuania. <https://www.clarin.eu/sites/default/files/SimoneUeberwasser.pdf>.

Appendix

Bilingual French-English 2020 Survey on *88milSMS* Reuse

Retour d'expérience sur l'utilisation du corpus *88milSMS*. Pour pouvoir approfondir les données concernant les utilisations scientifiques et pédagogiques de *88milSMS*, Rachel Panckhurst et Francesca Frontini vous seraient très reconnaissantes de prendre 5 minutes pour répondre à quelques questions. Nous vous remercions pour votre aide.

Feedback on usage of the *88milSMS* corpus. In order to gain further insight into scientific and pedagogical usage of *88milSMS*, Rachel Panckhurst Francesca Frontini would be very grateful if you could take 5 minutes to answer a few questions. We are very grateful for your help.

1. Nom de famille / Last name + Prénom / First name
2. Affiliation, institution, autre/other
3. Pays / Country
4. Quel est votre profil ? / I am a:
5. Dans quel(s) domaine(s) d'étude avez-vous utilisé le corpus *88milSMS* ? / I have used the *88milSMS* corpus for the following research fields:
6. Pour quelle(s) tâche(s) avez-vous utilisé *88milSMS*? / I have used *88milSMS* for the following task(s):
7. Les travaux ont-ils été valorisés ? / Has the research been disseminated?
8. Avez-vous cité le corpus *88milSMS* dans la bibliographie de vos travaux ? Si oui, pourriez-vous indiquer ci-dessous la citation utilisée ?
Have you cited the *88milSMS* corpus in the bibliography of some of your publications? If so, can you insert the citation format you used below?
9. Le corpus *88milSMS* a été téléchargé plus de 1050 fois depuis sa mise à disposition en 2014 (<http://88milSMS.huma-num.fr/>) Saviez-vous qu'une deuxième version du corpus (XML/TEI) a été mise sur Ortolang (<https://hdl.handle.net/11403/comere/cmr-88milSMS>) en 2016 ? The *88milSMS* corpus has been downloaded more than 1050 times since its release in 2014 (<http://88milSMS.huma-num.fr/>). Did you know that a second version of the corpus (XML/TEI) was released on Ortolang (<https://hdl.handle.net/11403/comere/cmr-88milSMS>) in 2016?
10. Si vous avez répondu oui à la question précédente, avez-vous également téléchargé *88milSMS* depuis Ortolang? Pour quelle(s) raison(s) ? Si vous avez répondu non à la question précédente, pourriez-vous nous dire si vous comptez télécharger la version de 2016 à l'avenir ? Pourquoi ?
If you answered yes to the previous question, have you also downloaded *88milSMS* from Ortolang? For what reason(s)? If you answered no to the previous question, could you tell us if you plan to download the 2016 version in the future? Why?
11. Complément d'information que vous souhaiteriez apporter, notamment si vous avez répondu "autre" à l'une des questions précédentes.
Further information you wish to submit, including if you answered "other" in any previous question.

Complementing Static Scholarly Editions with Dynamic Research Platforms: Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) for the Wittgenstein Nachlass

Alois Pichler

Wittgenstein Archives at the University of Bergen

Bergen, Norway

alois.pichler@uib.no

Abstract

In 2000 the Wittgenstein Archives at the University of Bergen (WAB) published the CD-ROM edition of *Wittgenstein's Nachlass: The Bergen Electronic Edition* (BEE). However, since the publication, it has become increasingly obvious that this edition does not meet all demands of the community that uses Wittgenstein's manuscripts (his "Nachlass") for research and learning. WAB has, therefore, for more than a decade now worked towards complementing the static CD-ROM edition with an interactive web platform that allows a more comprehensive, yet also a more tailored and user-specific utilization, of WAB's Nachlass resources. The paper describes and discusses two specific web service tools of this platform: Interactive Dynamic Presentation (IDP) of the Wittgenstein Nachlass, and Semantic Faceted Search and Browsing (SFB) of Wittgenstein metadata. The paper argues that it is only when these two tools are fully implemented and functional that WAB can adequately serve the scholarly needs of the Wittgenstein Nachlass user community. The paper describes some selected features and functionalities of these two tools in detail. While pilot versions of both tools are already in use on the platform, they need substantial extension and optimization. This upgrade is being implemented within the Norwegian CLARINO+ project.

1 Data and Metadata for Use of Wittgenstein in Research and Education

During his lifetime, the Austrian-British philosopher Ludwig Wittgenstein (1889–1951) published only one philosophical book, the *Logisch-philosophische Abhandlung / Tractatus logico-philosophicus* (1st ed. 1921/22), and a *Dictionary for Elementary Schools* (1st ed. 1926). However, on his death in 1951, he left behind a significant 20,000 page corpus of unpublished philosophical notebooks, manuscripts, typescripts and dictations. This oeuvre, called "the Wittgenstein papers" or "Wittgenstein's Nachlass" (von Wright, 1969), was early recognized to be one of the most important philosophy archives of all times. It was subsequently brought to the wider public through posthumous book publications such as *Philosophical Investigations* (1st ed. 1953) and *Culture and Value* (1st ed. 1977).

The practice of bringing the Nachlass to *digital* users reached its first milestone in 1998 with Vol. 1 of the Bergen CD-ROM edition *Wittgenstein's Nachlass: The Bergen Electronic Edition* (BEE; Wittgenstein, 2000), edited by the Wittgenstein Archives at the University of Bergen (WAB, <http://wab.uib.no/>).¹ The edition became notable for creating unprecedented new access and research possibilities (Meschini, 2020, ch. 4). Since its establishment in 1990, WAB has worked towards providing digital data and metadata for using the Wittgenstein Nachlass in research and education (Huitfeldt, 2006). This includes the creation of machine-readable transcriptions with specialized markup. The transcriptions were originally produced in MECS-WIT (Huitfeldt, 1994). But at present they are maintained in XML TEI format (Pichler, 2010). XML TEI transcription samples of 5000 Nachlass pages were made available CC BY-NC 4.0 on WAB's website within the

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

¹ WAB continues cooperation with Oxford University Press towards producing a new edition of the BEE.

frameworks of COST Action A32 (2006–10) and the Discovery project (2006–09). Most importantly, since 2015 transcriptions of the *entire* Nachlass, along with high quality Nachlass facsimiles, are made available open access in HTML format (Wittgenstein, 2015– and Wittgenstein, 2016–). In addition, WAB is working on the implementation of semantic web methods and technology. WAB offers free download of a continuously growing Wittgenstein ontology in OWL (RDF) format from its website (http://wab.uib.no/wab_philospace.page).

In order to provide for a common and persistent system of reference for its Nachlass resources, WAB has, within the framework of the Discovery project, assigned a unique identifier to the following: (i) each single one of the (about 150) Nachlass manuscript and typescript items, (ii) each single one of the (about 20,000) Nachlass pages, and (iii) each single one of the (about 55,000) Nachlass “Bemerkungen” (remarks).² A Wittgensteinian *Bemerkung* is typically no longer than half a page and separated from other *Bemerkungen* by one or more blank lines. Moreover, WAB’s Nachlass transcriptions and facsimiles are not only published in the BEE but are also available in a open access edition online (Wittgenstein, 2015–). Thus, with the aforementioned reference system in place and the open access availability of content, metadata and ontology, it may seem that WAB has achieved its goal of sufficiently equipping the user community. But, unfortunately, this is not the case. A *static* scholarly edition of Wittgenstein’s Nachlass, even if it is regularly updated for content, style and technical formats, will, by its very nature, always be inadequate in meeting the ever-evolving, dynamic user needs. A static scholarly edition of Wittgenstein’s Nachlass certainly remains indispensable as a source of stable, authoritative and easily citable text. But for the community to make the most of the resources in research and learning, much more than such an edition is required.

Any static edition is necessarily the result of selection and decision processes. The user needs for such an edition can in the end only be satisfied by complementing the edition with a platform that offers (i) access to datasets and aspects of the source not available through the edition and (ii) specific user need driven and tailored access and use. In the following, I want to argue for the implementation of two web service tools for responding to the increasing and potentially unlimited number and kinds of the needs of the Wittgenstein Nachlass community. I shall describe some of their functionalities along selected features. The two tools are Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) of the Wittgenstein Nachlass and Wittgenstein metadata. I contend that the user community of the Wittgenstein Nachlass continues to have needs and expectations that are not met even after all requirements of a static scholarly edition are fulfilled. I further claim that it is only when the two tools, IDP and SFB, are implemented and fully functional that we begin to adequately address the needs and expectations of the users of the Wittgenstein Nachlass. The paper goes on to describe some selected features of some of the functionalities of these two tools in detail. In my description I shall give most room to IDP, since this service, although apparently less standard and generally known, is of primary importance to WAB and its users.

WAB has for more than a decade now worked towards complementing the static scholarly edition with an online (distributed) research infrastructure that includes pilot versions of the IDP and SFB tools. Both the infrastructure platform and the tools themselves need substantial extension and optimization towards the strengthening of interactive functionalities and matching of user needs. These tasks of extending and better adapting the two tools are being implemented within the Norwegian CLARINO+ project (2020–26).

2 Open-ended User Needs

The BEE brings together three sub-editions – a facsimile, a normalized transcription and a diplomatic transcription – and can be called a “combined edition” (Pichler & Haugen, 2005). The diplomatic and the normalized transcription represent different levels of intervention on WAB’s source transcriptions of the originals. While the diplomatic, for example, retains deleted words,

² For details on the reference system see Pichler, 2010; for *Bemerkung* identifier examples see Figures 2 and 3 below.

deleted characters, marks insertions as insertions and does not intervene in spelling and grammar, the normalized is directed towards providing a standardized, easy to read, and finally also easy to cite, stable authoritative text. To put it more theoretically, one could say that the normalized version is strongly *text*-focused while the diplomatic version primarily attends to the *document* or even to the *document-carrier*. Thus, one could call the two formats two limiting cases of scholarly editing. The diplomatic version is an extremely helpful aid if one wants to start one's Nachlass research by reading the facsimile, but from time to time it needs deciphering help that is then supplied by the diplomatic version. For diplomatic and linear transcription samples see Figures 1–4.³

The triple structure of facsimile, diplomatic and normalized version enabled the BEE to respond in one and the same publication to a spectrum of research needs, rather than simply fulfilling, for example, the request for only an easy to read final version of a text. The BEE demonstrated the significant advantages that digital editions have over print editions in that the former, for example, allow more user-flexible access to the edited material. But at the same time, the edition was still not dynamic enough for adequately responding to the full spectrum of research needs and interests that Nachlass users have and, furthermore, can legitimately expect *digital* editing to provide. It was not dynamic enough by its very nature of being a *static scholarly edition* the purpose of which is to provide a *stable* and citable authoritative text. No static edition alone will ever be dynamic enough to meet the challenge of accommodating the diverse and evolving needs of the research community. Therefore, while a static scholarly edition will always be required, it must at the same time be complemented by a *dynamic research platform* that not only offers additional resources, but also additional and interactively available and toggleable analysis tools, additional presentation and filtering options.⁴ Let me illustrate the claim with a few examples.

Need for the possibility of chronological sorting: During his military service in WW1, Wittgenstein kept diaries – MSS 101–103 (1914–17) – where he not only wrote his philosophical reflections that eventually resulted in his first philosophical book, the *Tractatus logico-philosophicus* (1921/22), but also noted down deeply personal and private remarks. As a rule, he used for the personal and private entries the verso pages and a code, while for the logical and philosophical remarks the recto pages (and no code; see Figure 1).⁵ When editing the material for readers primarily interested in Wittgenstein's *philosophy*, the two Wittgenstein trustees G.E.M. Anscombe and G.H. von Wright selected from the notebooks what they considered the philosophically relevant portions only and turned these into a normalized book edition called *Tagebücher / Notebooks* in 1960. Many years later, in an unauthorized and sensational edition called *Geheime Tagebücher* (1985), Wilhelm Baum published the coded personal and private remarks. To date there is no German or English book edition that contains both kinds of remarks in one and the same book. While the BEE contains both, it contains them as separate blocks. For each of the three notebooks, the BEE first presents the sequence of the personal and private remarks and then presents the sequence of the philosophical remarks. It makes sense to separate the two types of remarks since they each belong to their own specific discourse. However, as a consequence of such editing practices, the Wittgenstein community has learned to receive the remarks as two separate strings. The division of the content of Wittgenstein's writings into the philosophical and the personal may seem appropriate and satisfactory from a certain rigid scholarly perspective, but upon reflection the practice clearly reveals its disadvantages. One disadvantage is that it not only splits the text sequence, but also splits the chronological sequence of the remarks into two and thus, for example, makes it cumbersome to put in context all remarks that Wittgenstein wrote on a specific day. Researchers are increasingly voicing precisely this need of connecting Wittgenstein's personal and private remarks with his simultaneous reflections on philosophy and logic (and vice versa) for a better understanding of Wittgenstein's works (see Figure 2).

Need for the possibility of including / suppressing revision layers: After Wittgenstein's return to Cambridge in 1929, an event that is often regarded as simultaneous with his return to philosophy, Wittgenstein wanted to publish a second philosophical book. He had different ideas of

³ For a short introduction to “diplomatic”, “linear”, “normalized” and other scholarly edition and transcription types see, for example, Pierazzo, 2009. For the distinction between document carrier, document and text see Pichler, 2021.

⁴ See also Gabler, 2013. For an early discussion and promotion of “dynamic edition” see Rehbein, 1998.

⁵ Wittgenstein's cypher consisted in, roughly speaking, reversing the alphabet such that, e.g., “ich” becomes “rxs”; for details see Wittgenstein, 1998, xve.

the book's contents and form at different times. But the so-called Big Typescript, TS 213 (1933), is widely regarded as a definite and substantial steppingstone in this project, if not as an actual candidate for the envisaged book. In this typescript, Wittgenstein collected between two and three thousand remarks selected from the manuscript volumes he had written since 1929 and organized them into chapters and subchapters. He introduced each chapter with a philosophical topic heading (e.g., "Meaning") and each subchapter with a philosophical statement (e.g., "The concept of meaning originates in a primitive conception of language"). However, Wittgenstein soon felt uncomfortable with the arrangement in the Big Typescript and started to not only reorganize the ordering but also to revise the text itself. Moreover, this reworking took place not only in the typescript itself but also in a number of new notebooks and manuscript volumes such that the resulting new text(s) were spread over several items. When producing a book edition from this project, *Philosophical Grammar* (Wittgenstein, 1969), the third Wittgenstein trustee Rush Rhees tried to come as close as possible to Wittgenstein's final intended revision of the text for the Big Typescript. However, his edition was criticized by some for blurring the distinction between the actual and the virtual text of the corpus, and it was remarked that Rhees should have edited the Big Typescript "as it stood", i.e. without Wittgenstein's later handwritten revisions (Kenny, 1984, 37; as a response see Rhees, 1996). The BEE wisely returned – as a documentary edition – to the actual documents and offered diplomatic and normalized transcriptions thereof. It included hyperlinks where Rhees actually carried out Wittgenstein's instructions for arranging the text in a different order or replacing part of it altogether. Although this was a required step, at the same time it deprived readers of an easy way to follow and cite the text in the sequence that resulted from Wittgenstein's revision. This text was often only virtually given in the Nachlass, but had been offered by Rhees' edition. It is clear that one should be able to have it *both* ways, and that it should precisely be the digital edition that *gives* it to you both ways: the text before and after the revision – actually, any text before, with, and after *all* revisions. This is especially relevant for work with heavily revised typescripts, such as TS 213 or TS 226 (see Figure 3). However, this is not something that the BEE could achieve and currently it is still not fully achieved by WAB's Nachlass editions on the web.

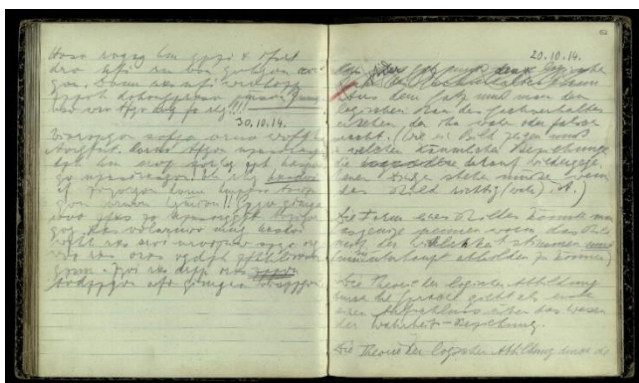
Need for the possibility of filtering according to different parameters: From his earliest to his latest writings, whenever Wittgenstein revisited his remarks with an eye for further editing and processing he marked them with symbols in the margins of the page. At WAB these symbols are called "section marks". A section mark could for example be a slash, an asterisk, a circle, a letter like a capital S, or a letter like lower case x. Similarly, when considering changing the arrangement of his remarks, Wittgenstein would add numbers or combinations of numbers and letters in the margins such that through these symbols the remarks were collected into groups.⁶ The meaning of the single symbols, especially when it comes to the section marks, is to date only partially known. Whenever they have been included in print or digital editions, the editors simply tried to reproduce them in their graphical appearance. This was also the case with (the diplomatic transcription of) the BEE. Against this practice one could object that the entire point of these symbols is to signalize that Wittgenstein wanted to *do* things with the remarks thus marked: dictate them, omit them, discard them, revise them, rearrange them, group them etc. Accordingly, it is to these *action intentions* that the reader should be directed to rather than simply receiving a visual representation of the symbol only. For example, about the remarks marked with a slash in MSS 105–108 (1929–30) we know that most of them were dictated to a typist (Pichler, 1994), resulting in what in Nachlass research is called TS 208. About the number and letter combinations with which Wittgenstein marked several thousand remarks cut out from his typescripts of 1930–31, we know that they constituted the reference system according to which the "Zettel" collection of TS 212 (1932–33), which subsequently became TS 213, was to be organized (Rothhaupt, 2016). However, there is a great deal of such editorial symbols left in the Nachlass and we do not have sufficient knowledge about their functions. Some of them will contain an instruction for how to proceed with, or from them; others will serve to express an evaluation of the remark tagged with the specific symbol. Now, users who

⁶ These editorial numbers should not be confused with the sequential paragraph numbers used by Wittgenstein in his more finished works, and with which the reader will already be familiar from books such as *Philosophical Investigations* (Wittgenstein, 2009) containing remarks ("paragraphs") §§ 1–693.

want to study the meaning of these symbols further or even to convert their meanings to resulting text selections, groupings and arrangements, have a non-negotiable requirement. The requirement is that one has an edition that not only renders the remarks in their original sequence with these symbols included, but also permits filtering and arranging the texts *according* to these symbols while retaining the possibility of including or omitting the symbols themselves in the resulting output. It should thus, for example, be possible to extract all remarks and only the remarks which in the Nachlass are marked by Wittgenstein with a slash, or an asterisk, or a backslash, etc., or a specific combination of them. It is only then that these users' needs will be adequately addressed and the scope of the scholarly utilization of WAB's Nachlass resources can be greatly widened. So, for example, users may become equipped to study specific genetic processes in the Nachlass or recognize thematic groups of which the symbols are often the "indices". Or they may become enabled to perform more basic tasks such as learning about the function and meaning of the symbol itself. Recently, Rothhaupt (2013) has argued that the remarks which Wittgenstein marked with a circle / circle-like symbol (a "Kringel") contain Wittgenstein's attempts at a philosophy of culture. Again, it is only if the user has access to a filtering tool such as the one described here, permitting easy extraction of all *Bemerkungen* and only the *Bemerkungen* which are marked by Wittgenstein with a "Kringel"-symbol, that she is in the position to efficiently and reliably investigate that this hypothesis is sound, or to discover other elements in these remarks that led Wittgenstein to mark them all with the "Kringel"-symbol. While WAB's website (Wittgenstein, 2016–) today already offers filtering of Nachlass documents according to section mark parameters, this feature still stands in need of improvement and does not yet fully meet all user requirements.

Need for the possibility of conducting metadata search / combined text and metadata search: The BEE already offered some semantic search functionalities – e.g. search for references to persons, taxonomies for mathematical and logical notation as well as for graphics, possibility to focus on the coded passages or other groupings only. At the same time, many more valuable metadata had been recorded in the transcriptions or via stand-off markup that users could greatly benefit from if only they had processing access to them. Some users would, for example, not only want to search and browse the Nachlass by references to persons or works (for a sample see Figure 4), but specifically all references to persons or works that have come about or, alternatively, got discarded by later revision (e.g., the revision of the Big Typescript). Or a reader, who has an interest in influences on Wittgenstein but is most acquainted with the book publications from the Nachlass only, may want to search for Wittgenstein's references to persons and works in all and only the remarks which hitherto were *not* included in any of the book publications from the Nachlass. Or a reader most interested in Wittgenstein's writing in code may want to search for any passage in code that hitherto was not published in print. Or one may want to check whether there is a correlation between the remarks marked with a slash "/" and eventual publication of the remarks by the trustees in one of the book publications. One may want to do so either in order to better understand Wittgenstein's use of the section marks or to find out to what extent the Nachlass editors let themselves be guided by the section marks for their selection of materials to be published. Moreover, one may want to find out whether there is a correlation between the sequence of the remarks in a specific work by Wittgenstein, e.g., the *Philosophical Investigations* (Wittgenstein, 2009), and their chronological origin. Other user needs relate to a remark's genetic path(s), its place of origin in the Nachlass corpus, references to places, events and other named entities, similarity to other remarks (Ullrich, 2019, and Huitfeldt, 2020), adherence to text type and genre (philosophical remark, preface, motto, dedication, instruction, aphorism, diary entry, autobiographical remark, personal and private remark, coded remark, mathematical-logical notation, graphic etc.), adherence to Nachlass group (notebook, loose sheet, "Zettel", ledger, typescript, dictation etc.), work status (first draft, elaborated version, final work etc.), script type (shorthand, code etc.), the language the remark is written in (German, English etc.), writing and revision instrument (different kinds of pencil, black ink, red ink, blue ink etc.), research literature referring to it, and so on. Finally, one may frequently also need to conduct searches with both the text *and* the metadata as one's research base. To be able to *combine* text and metadata search becomes, for example, pertinent where one needs to find all

and only those documents that contain both a specific word used by Wittgenstein *and* a specific reference to a person or work. Or one may remember only one or two words from a passage in the *Philosophical Investigations* and try to find the passage with the help of these words, restricting one's search to precisely the *Philosophical Investigations* corpus only. These are all relevant and legitimate needs that can turn out to be pressing in either research or learning. The list in fact seems endless. To meet these and similar needs, efficient and selective access to metadata, and iteratively faceted processing of metadata becomes essential. Unfortunately, neither the BEE nor WAB's Wittgenstein web services currently fully meet this challenge.



Sehe jetzt so klar & ruhig wie nur in den besten Zeiten. Wenn ich nur diesmal alles wesentliche lösen könnte ehe die gute Zeit um ist!!! —.

30.10.14.

Erhielten heute eine deutsche Zeitung. Keine guten Nachrichten was so viel heißt als schlechte Nachrichten! Es ist schwer zu arbeiten wenn solche Gedanken einen stören!! Habe trotzdem auch am Nachmittag gearbeitet. Ich empfinde oft schwer dass ich hier niemand habe mit dem ich mich etwas aussprechen kann. Aber ich will mich allen Gewalten zum Trotz erhalten.

Figure 1: Facsimile of MS 101, facing pages 51v (dated 29.–30.10.1914) and 52r (dated 20.10.1914), with diplomatic transcription of p. 51v (see http://www.wittgensteinsource.org/Ms-101,51v_f). Italics indicates code writing.

30.10.14.

Erhielten heute eine deutsche Zeitung. Keine guten Nachrichten was so viel heißt als schlechte Nachrichten! Es ist schwer zu arbeiten wenn solche Gedanken einen stören!! Habe trotzdem auch am Nachmittag gearbeitet. Ich empfinde oft schwer daß ich hier niemand habe mit dem ich mich etwas aussprechen kann. Aber ich will mich allen Gewalten zum Trotz erhalten.

30.10.14.

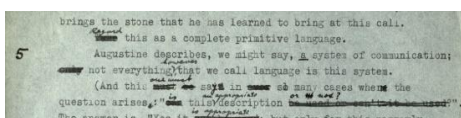
Könnte man sagen: „in „ $\sim\phi(x)$ “ stellt „ $\phi(x)$ “ vor wie es sich nicht verhält“?

Man könnte auch auf einem Bild eine negative Tatsache darstellen indem man darstellt was nicht der Fall ist.

Figure 2: Linear transcription of remark (Bemerkung) Ms-101,51v[2] “Erhielten heute ...” from MS 101, p. 51v, chronologically co-arranged with remarks Ms-101,69r[4] and Ms-101,69r[5], equally dated 30.10.1914, but from p. 69r (see http://www.wittgensteinsource.org/Ms-101,69r_f). Italics indicates code writing.

Diplomatic transcription of “typescript only”:

Augustine describes, we might say, a system of communication; only not everything that we call language is this system.



Diplomatic transcription, including revisions in Wittgenstein's hand:

5 Augustine describes, we might say, a system of communication; only not everything, ^{however}, that we call language is this system.

Figure 3: Part of TS 226, p. 2 (see http://www.wittgensteinsource.org/Ts-226,2_f) with diplomatic transcription of remark Ts-226,2[2] “Augustine describes ...”; for a diplomatic transcription of the entire remark see [http://www.wittgensteinsource.org/Ts-226,2\[2\]_d](http://www.wittgensteinsource.org/Ts-226,2[2]_d).

3 Interactive Dynamic Presentation (IDP)

In the preceding section I have given some examples of the great variety of needs encountered in the user community. If we try to respond to these needs by providing the requested functionalities and services in *one and the same static edition*, it is likely to fail either technically or in terms of usability (or both). Rather, what one needs is toggleable services that in dynamic and interactive ways produce highly adjustable outputs serving the wide spectrum of resources needed: from heavy apparatus to readers' editions, from original to standardized orthography and grammar text, from physical to chronologically arranged document sequences, etc. etc.⁷ Even users who in the beginning are satisfied with having at their disposal the standard presentation formats of the diplomatic (recording all deletions, insertions, overwritings etc.) and the normalized (giving the resulting text) transcription, are likely to discover that these represent editorial intervention, or a lack thereof, which does not satisfy all research needs one can rightly expect scholarly digital resources to fulfill. Some may find out, for example, that a minor thing such as representation of the original's line breaks can make a big difference for purposes of their research. Edited text outputs that follow the original line order have indeed many advantages, one of them being that they easily permit comparison of edited text version and facsimile. In other contexts, however, indication of the original's line-breaks is completely inessential to one's research interests, or even distracting. IDP relieves the editors from having to statically freeze a representation that comes with or without the line breaks; this decision can simply be left to the user who can have it both ways depending on what they judge to be best in a particular context.

One might want to entertain the belief that user needs that fall outside of what is already covered by the diplomatic and the normalized formats can be dealt with by simply increasing the number of formats pre-produced and offered. In this spirit one could suggest adding to the diplomatic and normalized versions, if they come without representation of original line breaks, diplomatic and normalized versions *with* representation of original line breaks. Moreover, one might suggest adding a linear version that occupies a middle ground between the diplomatic and normalized version, and a "typescript only" version that omits handwritten revisions in the typescript (for an example of the latter see Figure 3). But such amendments would not yet help with gaining access to, for example, chronologically sorted outputs of either a single Nachlass item or an entire Nachlass group. Such sorting is relevant for the whole of the Nachlass, for example, on the first *Bände* series MSS 105–122 where there are many chronological jumps from the midst of one manuscript to the other. Furthermore, the above-mentioned amendments would not suffice if one requires versions that are still between, on the one hand, *the diplomatic and the linear*, and on the other hand, between *the linear and the normalized*. A version that is between the diplomatic and the linear would be, for example, one that includes deleted portions of entire words or even entire sentences – in distinction to the diplomatic version where all deleted parts are included, and in distinction to the current linear version where no deleted parts are included. A version that is between the linear and the normalized would be, for example, a version that marks every editorial intervention into orthography and grammar – in distinction to the current linear version where only interventions on word and punctuation level are marked, and in distinction to a normalized version where no such intervention may be marked as such. Moreover, even for the "typescript only" version one would again have to distinguish between at least its diplomatic and normalized variants: the former rendering the typescript "as it stood"; the latter permitting text search across unified orthography and grammar and including replacement symbols for logical and mathematical notation as well as other characters not available on the typewriter: "ss" for "ß", "ae" for "ä", "Ae" for "Ä" etc. etc. If we continue along this line, we will sooner or later encounter the question: In the end, even if we were to meet only the most important user requests, how many additional versions must be added? Attempting to answer the question will make one realize that the number one is likely to end up with is already too big to

⁷ Saying this, one is not even touching upon the challenges created by the fact that different users and publication channels have different preferences regarding formatting styles, e.g. how emphasis is to be conveyed (italics rather than underlining, etc.), or platform (mobile rather than stationary platform, etc.). Neither are issues addressed that relate to individual handicap (sight, colour blindness, etc.) or other conditions that can make use of the services difficult.

feasibly continue along this line of offering pre-made static editions. Rather, one needs to provide something like an interactive and dynamic “laboratory” setting that lets the user create the variety of outputs needed on the fly. Thus, all add-ons will eventually lead us exactly to the point where IDP already *is*. Thus, instead of trying to add any extra editions as *static* add-ons that are pre-made by the editors in advance, it is a much better strategy to provide for the possibility that the user generates the edition(s) from the underlying text archive on the fly and as required in the hic et nunc-situation. All that is needed for this strategy to work is, in addition to software, something like a text archive in the form of encoded transcriptions, stylesheets for their conversion to the specific output needed, and an interface for running on the transcriptions precisely those parameters of the stylesheets that are required for achieving the needed output.

IDP can fulfil its task thanks to its compliance with three principles (Pichler and Bruvik, 2014): (1) Separation of matters of transcription (encoding, markup) from matters of presentation; (2) Empowerment of users to let them interactively co-produce editions rather than being passive receivers of expert-editor produced editions only; (3) A dynamic and multi-relational view of the relation between the source document and potential presentations of it. These three principles are central ingredients of all text encoding based digital editing at WAB. WAB’s IDP site <http://www.wittgensteinonline.no> (Wittgenstein, 2016–) allows the user to interactively produce from the ever latest version of the XML TEI (P5) source transcriptions text archive, more tailored and a greater variety of outputs than those already available from the pre-made static editions offered by WAB on e.g. <http://www.wittgensteinsource.org> (Wittgenstein, 2015–). These outputs are created in HTML through XSLT transformation from the one XML TEI (P5) guided transcription source in the text archive which contains for each Nachlass item, page, *Bemerkung*, sentence, formula, drawing, word, letter and character detailed philological / structural / semantic information.⁸ This makes the output resource for the users a dynamic, adjustable, revisable and continuously updatable entity. The users have access to the ever *newest and improved* version of the source transcription and are given the possibility to *interactively* process the source transcription to the *presentation* that is best tailored to their specific research needs.

Naturally, the more detailed and explicit an encoding the source transcription contains, the more powerful and adjustable the IDP can become. With regard to WAB, chronological sorting of the Wittgenstein Nachlass, for example, can be implemented thanks to two features of WAB’s transcriptions. First, for each single Nachlass remark there exists a self-contained complete XML transcription such that the entire Nachlass can be constructed out of the transcriptions of its single remark. Second, for each of the single Nachlass remarks there also exists WAB metadata providing a, albeit often alleged only, dating for the remark. The two taken together provide for the possibility of a complete sorting of the entire Nachlass according to chronological parameters. Omission of handwritten revision in typescripts can be achieved thanks to explicit encoding of handwriting in typescripts. Filtering and sorting of the Nachlass texts according to Wittgenstein’s editorial marks or numbers can be put to practice thanks to the specific encoding WAB uses for them. It cannot be stressed often enough that it is a principle of the IDP model precisely to not provide a different source transcription for each of the different desired outputs, or at least for the most important of them. Rather, IDP works on the basis that one explicitly marks everything that is to be subsequently processed in the *one and only* master source transcription, possibly combined with additional standoff markup. Also, that one then leaves questions of presentation, filtering, sorting etc. to the stylesheet and user interface. E.g., with regard to subsequent IDP manipulation of handwritten revisions in typescript, every hand-produced writing act is encoded in such a way that it can be filtered and processed *independently* of everything that is typewritten (and vice versa).⁹ While

⁸ The same principle of *One base source – Many outputs / apps on top of it* applies to WAB work generally. Wittgenstein, 2015– (BNE), Wittgenstein, 2016– (IDP), WiTTFind, SFB as also the upcoming new BEE edition at Oxford, while each offering different entrance points and each coming with their specific strengths and foci, are all produced from one and the same set of XML source transcriptions and metadata sources.

⁹ This functionality has been crucial for researching Wittgenstein’s handwritten revisions to Rhees’ translation of the early version of Wittgenstein’s *Philosophical Investigations* in TS 226 (Pichler, 2020a).

WAB's encoding is still far from complete enough to be sufficiently prepared for all legitimate IDP requests, these examples should suffice to show that the possibilities and capacities that IDP offers for responding to user needs are simply enormous. Sorting functionalities (sorting according to chronology; sorting according to physical sequence; sorting according to discourse sequence; sorting in order to more easily relate personal diary entries and philosophical remarks; sorting to better see the chronological sequence of Wittgenstein's philosophical work, etc.), multiple presentation functionalities (inclusion and omission of handwritten revisions in a typescript such as insertions, deletions, overwritings, underlinings etc.; presentation of typed text only, in order to study the vocabulary before the typescript was revised and investigate genetic processes, etc.), filtering functionalities (filtering of the Nachlass according to the marks and numbers that Wittgenstein assigns to his remarks; filtering in order to identify thematic groups; filtering in order to separate (according to Wittgenstein's own judgment) 'good' from 'not so good' remarks; filtering in order to identify genetic processes, etc.) as well as their combinations (sorting, filtering and inclusion/omission according to text revision stage etc.) offer benefits that Nachlass scholars, prior to the introduction of IDP, could only ask for but could not expect the requests to be fulfilled.

One of the most important effects of IDP is that it creates in the user an awareness that what she is dealing with is not something that simply falls from the sky: that even *premade* editions do not fall from the sky, but result from selection and decision processes. Where the user previously, maybe with a sort of innocent and uncritical attitude, simply received and accepted what the editors – be it Wittgenstein's heirs (Erbacher, 2020) or others – gave her (Pichler a.o., 2011), with IDP she suddenly recognizes that any edition is a product of human action deserving scrutiny and verification. The user may just as well take an active role herself in the making of the edition, and through IDP she can indeed become a co-agent, taking on some of the editorial responsibilities herself. But equipping the user with the tools required to do this and thus to make the best out of WAB's resources, will involve much more than cutting edge digital editorial philology methods and tools. It also requires giving the user access to at least the most basic contemporary semantic web technology and methods.

4 Semantic Faceted Search and Browsing (SFB)

WAB's Nachlass reference system provides a URL for each single Nachlass component. It goes without saying that this is crucially important for working with IDP user-generated content: with the reference system, the component researched or cited not only becomes easy to refer to, but also exactly describable, traceable and tractable throughout all filterings, sortings and rearrangements, as well as throughout all research articles and annotation or semantic web environments where it enters into. The same system of reference is also applied to the facsimiles which contributes significantly to user-friendliness and the quality of research in terms of its coherence and consistency. Last but not least, the reference system also makes up the backbone of the Wittgenstein ontology and is thus also essential for working with "semantic Wittgenstein".

In section 2 we identified a need for conducting metadata search and search *combining* text and metadata. This is precisely what semantic faceted search and browsing (SFB) is about. SFB can be briefly explained as follows: First, the source is the data and metadata in their semantic, classificatory and, ideally, also taxonomic aspects. Second, SFB applies digital semantic technologies; it is thus about organizing and investigating a domain's semantics, rather than a field of editorial philology. Finally, the search and browsing works with *facets* as vehicle. Facets are properties / dimensions / relations of the domain's objects according to which these objects can be classified, and thus include metadata. The term "faceted" stands for the metadata and data filtering through incremental *faceting*, and thereby the filtering of data and metadata down to the desired result (the "hit"). Naturally, the domain's data and metadata can also be accessed via a direct string search. There can be a great many and a great variety of relations between the objects of a domain, and at different points there will be a need to focus on different objects and relations. SFB permits to do precisely that and to identify the object(s) which match the focus one has at a particular place and time. Furthermore, each object can have a great number of relations to other objects which one may be aware of but for which one lacks an overview, as well as many additional relations which

one might not yet be aware of. SFB helps to see and make explicit the relations which, though straight on the surface, previously remained unseen, and thus helps the user achieve a synoptic view of the objects and their relations. SFB is also the tool for simply exploring and working with the relations that one already *is* aware of, but maybe still needs highway type of routes for in order to easily move from one known to another known node in the semantic landscape.

The data model behind WAB's SFB pilot on <http://wab.uib.no/sfb/> is an ontology in RDF OWL format that aspires to organize not only the Nachlass, but also the entire Wittgenstein domain under three top classes: *Source*, *Person* and *Subject* (Pichler and Zöllner-Weber, 2012). The *Source* class houses primary and secondary sources relevant for Wittgenstein research; the subclass *Primary Source* further divides into Wittgenstein sources, such as the *Tractatus*, and external sources, such as Augustine's *Confessions*. The lowest subclass of Wittgenstein primary sources is the remark, the *Bemerkung*. The *Person* class contains historical persons such as authors Wittgenstein refers to (Biesenbach, 2014). The *Subject* class contains subclasses such as *Concept* (topic) and *Point* (claim); *Concept* refers to subjects dealt with by Wittgenstein himself or in Wittgenstein research, such as 'elementary proposition', 'picture', 'state of affairs', 'essence', 'logical analysis' 'logical independence', 'philosophy', 'proof'. Instances of *Bemerkung* can be interlinked with instances of *Concept* via the property *discusses*. *Point* refers to the point made or discussed by an individual Wittgenstein remark and may contain an entire statement such as 'The elementary proposition is a picture of a state of affairs'. Instances of *Bemerkung* can again be interlinked with instances of *Point* via the property *discusses*.

In the preceding discussions, I have presented only a brief glimpse of how the instances of the rich and comprehensive semantic Wittgenstein domain can be intertwined and their complex relations modeled, so that both the instances and the relations between them can subsequently be made available for SFB. As of December 2020, the SFB site offers search and browsing of more than 56,000 instances of *Source* and more than 1000 instances of *Person* from the Wittgenstein domain. This goes far beyond only Nachlass-related sources and persons.

5 IDP, SFB and the Wittgenstein Archives in CLARINO+

At present, WAB already runs IDP and SFB pilots on its website (Wittgenstein, 2016– and <http://wab.uib.no/sfb/>; for an independent assessment of the two tools, see Meschini, 2020, ch. 4.2 and 4.3). But while it has come far in producing and maintaining excellent transcriptions of the Wittgenstein Nachlass and excellent metadata and metadata organization for the Wittgenstein domain more generally, this is only one half of what the community needs. The stylesheets, interface and entire infrastructure and setup upon which IDP is built (XML technologies such as Xalan, Saxon, XSLT, HTML and PHP programming etc.) also need to be in top shape. Even to put all already existent encodings to work and offer them for IDP toggling demands a substantial programming investment. While some of the technical features and functionalities required and mentioned above already work, many do not yet work flawlessly or on all required levels. Chronological sorting, for example, while it has worked in the pilot for single items already for a long time, is needed the most at higher levels such as the chronological arrangement of Nachlass item *groups* or even the *entire* Nachlass corpus. With regard to the above-mentioned WW1 diary group MSS 101–103 (1914–17) where the chronological sequence is dispersed across different page sequences, this functionality will finally enable users to much more easily connect Wittgenstein's personal and private remarks with his simultaneous reflections in philosophy and logic. Moreover, the feature of toggling on and toggling off handwritten revisions so that one can view one and the same typescript page, e.g. from the Big Typescript, with and without handwritten corrections and additions, does not yet work as it should. One example where this becomes highly relevant is, as discussed above, the study of the sources for *Philosophical Grammar* (Wittgenstein, 1969). Another important case is Wittgenstein's handwritten revisions to Rush Rhees' translation of the early version of the *Investigations*, contained in TS 226 (Pichler, 2020a). Generally, the current insufficiencies can be briefly summarized as follows: The IDP tool currently manages to offer access to (1) only a fraction of the encoding, (2) only a fraction of combinatorial possibilities of the encoding, (3) only a fraction of the presentation, sorting and filtering possibilities and needs, and (4)

in all these three fields it is susceptible to errors due to undesired interference. It is in fact a major challenge to provide for combined filtering, sorting and presentation modes that work in tandem and do not negatively interfere with each other. This challenge results in the following limitations: with regard to (1), users cannot, for example, yet filter the transcriptions for insertions of a specific subtype; with regard to (2), users are not yet able to, for example, combine filtering of insertions with filtering of the encoding of text alternatives; and with regard to (3), it is not yet possible to render, for example, the type of insertions selected in ways other than what is set by WAB as the default for the IDP site. Moreover, with regard to (2), currently it is, for example, not possible for the user to combine a marking of Wittgenstein’s text alternatives with a diplomatic rendering, or with the inclusion / exclusion of his own markers for text alternative, or with a toggling of including / excluding the alternatives discarded by him.¹⁰

WAB’s SFB pilot “Wittgenstein Ontology Explorer”, however, at present already permits powerful search and browsing of the Nachlass along a number of facets, incl. reference to a person, reference to a work, a remark, its dating and its relation to “published works”. It displays any resulting remark hit along with a link to the corresponding facsimile in the *Bergen Nachlass Edition* on Wittgenstein Source (Wittgenstein, 2015–). The tool is built on the University of Bergen Library’s search infrastructure (<http://marcus.uib.no>), which involves technologies such as Elasticsearch (<https://www.elastic.co>), Apache Jena (<https://jena.apache.org/>) and Angular framework (<https://angular.io/>) and is fed with metadata found either in the XML transcriptions or in standoff markup. But before the metadata can become available on the SFB front end, they must be ingested into the tool; and before being ingested, their relations to each other must be modeled precisely in the above described ontology. Luckily, at present WAB’s reference system for the Nachlass is already implemented in the ontology and the SFB site such that the semantic faceted search and browsing metadata branch can fully communicate with WAB’s editorial philology branch.

Linear transcription:

- 5 Augustine describes, we might say, a system of communication; not everything, however, that we call language is this system.
 (And this one must say in so many cases when the question arises: “is this an appropriate description or not?”. The answer is, “Yes, it is appropriate; but only for this narrowly restricted field, not for everything that you professed to describe by it.” Think of the theories of economists.)

Screenshot of SFB-representation with metadata *Document type*, *Refers to work*, *Refers to person* and *Date*:

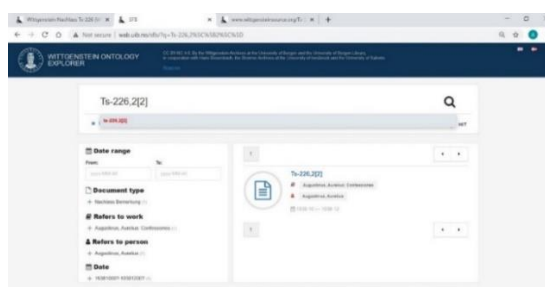


Figure 4: Linear transcription (see [http://www.wittgensteinsource.org/Ts-226,2\[2\]_n](http://www.wittgensteinsource.org/Ts-226,2[2]_n)) and SFB-representation (<http://wab.uib.no/sfb/?q=ts-226,2%5C%5B2%5C%5D>) of remark Ts-226,2[2].

One substantial and important addendum to the current SFB pilot will be the WiTTFind lemmatized Wittgenstein Nachlass *lexicon* that is modeled as a subclass of the *Primary Source* class. The WiTTFind Wittgenstein lexicon (Röhler, 2019) contains a lemmatized index for all occurrences of words in the Nachlass. The lexicon is the outcome of a long-standing cooperation between WAB

¹⁰ For a detailed study of Wittgenstein’s writing and marking of text alternatives see Pichler, 2020b.

and the Centrum für Informations- und Sprachverarbeitung (CIS) at the Ludwig Maximilians Universität München on the search tool WiTTFind (<http://wittfind.cis.lmu.de/>). In this joint project, WAB contributed its facsimiles and encoded XML transcriptions of the Wittgenstein Nachlass as well as XSLT stylesheets for their processing, while CIS provided programming and computational linguistics personnel resources as well as a grammatically encoded digital lexicon of the German language (Hadersbeck, Pichler et al., 2016). WiTTFind offers lemmatized online text search access to the entire Nachlass, displays each sentence containing any grammatical form of the word searched for within the context of the larger remark, and additionally highlights the segment of the facsimile corresponding to the remark hit. WiTTFind continues WAB's reference system all the way down to sentence level. WiTTFind is a fine example of the added value created by making one's data available for research and reuse by others. Implementing the WiTTFind Wittgenstein lexicon in the SFB tool will permit it to adequately respond to one of the needs identified above: simultaneous and combined SFB of *both metadata and text data*. Researchers interested in the genesis of Wittgenstein's philosophy, for example, may want to know when Wittgenstein started to use the expression "game" in places where earlier he had written "calculus", and whether this development can be linked to any other development, e.g. reference to particular works of others, other changes in vocabulary, developments in letter correspondence, meetings and discussions with friends and colleagues, etc. Doing this kind of research becomes possible through an integration of the WiTTFind lexicon into SFB.

In addition to integration of the WiTTFind lexicon, tasks of improving the SFB tool include correcting errors and deficiencies in the overall browsing and combinatorial setup, as well as adding and organizing missing facets. This includes facets for the above-mentioned personal and private coded remarks, mathematical and logical notation and graphics, or for Wittgenstein's writing tools. Another functionality to be added is chronological sorting of a remark's variants, which currently are only displayable in alphanumeric order. Finally, yet another need is the possibility to view the remark hit resulting from one's searching and browsing along with a linear or diplomatic *transcription* presentation of the remark. Currently only the remark's identifier along with a link to the corresponding *facsimile* is displayed. The extension and optimization of the IDP and SFB pilots, incl. the integration of the WiTTaFind lexicon, towards the web services that users need, are all being implemented within the Norwegian CLARINO+ (2020–26) project. Already within the Norwegian CLARINO project (2012–16) the current IDP *pilot* on Wittgenstein, 2016– was developed. Also within CLARINO+, WAB's XML TEI transcription corpus is additionally deposited in the CLARINO Bergen Repository so that users with XML programming competence can by *themselves* respond to their research needs by directly processing and querying the XML transcription files. This task includes generation of CLARIN CMDI-conformant metadata as well as designing licenses for the use of both the transcriptions and the metadata made available. But it is only a fraction of Wittgenstein Nachlass users who master XML technology and most will need an interface that offers the resources in a digital language and style that they understand. For this, IDP and SFB will be central tools. While at present WAB's web services incl. the IDP and SFB pilots already enjoy a large number of international users¹¹, it is only when these and other additional services are upgraded and optimized, that researchers will be able to take full advantage of WAB's resources. Only then will users be equipped to fully exploit the multifaceted interrelations between and within Wittgenstein data and metadata provided by WAB for research and learning about Wittgenstein. At the same time, it is also only then that the deep issues about the relation between on the one hand the contents and forms of Wittgenstein's philosophy and work, and on the other hand their interpretation and application, can properly begin to play out in sufficiently complex formats via interactive digital media.

¹¹ Google Analytics lists for <http://wittgensteinonline.no/> more than 5300 and for <http://wab.uib.no/sfb/> more than 1500 users for the period 2017-20. – In writing this paper I have benefitted from comments by K. De Smedt, N. Gangopadhyay, Ø. Gjesdal, J. Hendrickson, C. Huitfeldt, H. Al Ruwehy and two anonymous reviewers, as well as financial support by the Norwegian Research Council funded Clarino+ project.

References

- Biesenbach, H. 2014. *Anspielungen und Zitate im Werk Ludwig Wittgensteins*, Sofia.
- Erbacher, C. 2020. *Wittgenstein's Heirs and Editors*, Cambridge.
- Gabler, H. W. 2013. *Wittgenstein's Nachlass: The Bergen Electronic Edition*, in: P. Henrikson and Chr. Janss (eds.), *Geschichte der Edition in Skandinavien*, 167–176, Berlin.
- Hadersbeck, M., Pichler, A., Bruder, D. and Schweter, S. 2016. *New (Re)Search Possibilities for Wittgenstein's Nachlass II: Advanced Search, Navigation and Feedback with the FinderApp WiTTFind*, in: *Contributions of the Austrian Ludwig Wittgenstein Society*, 90–93, Kirchberg a. W.
- Huitfeldt, C. 1994. *Toward a Machine-Readable Version of Wittgenstein's Nachlaß*, in: H. G. Senger (ed.), *Philosophische Editionen. Erwartungen an sie – Wirkungen an sie*, Beihefte zu editio 6, 37–43.
- Huitfeldt, C. 2006. *Philosophy Case Study*, in: *Electronic Textual Editing*, Modern Language Association of America, 181–196.
- Huitfeldt, C. and Sperberg-McQueen, C.M. 2020. *Document similarity: Transcription, edit distances, vocabulary overlap, and the metaphysics of documents*, in: *Proceedings of Balisage: The Markup Conference 2020*, Balisage Series on Markup Technologies, 25, <https://doi.org/10.4242/BalisageVol25.Huitfeldt01>.
- Kenny, A. 1976. *From the Big Typescript to the Philosophical Grammar*, in: J. Hintikka (ed.), *Essays on Wittgenstein in Honour of G. H. Von Wright*, *Acta Philosophica Fennica*, 28, 41–53.
- Meschini, F. 2020. *Oltre il libro: forme di testualità e digital humanities*, Milan.
- Pichler, A. 1994. *Untersuchungen zu Wittgensteins Nachlaß*. Working Papers from the Wittgenstein Archives at the University of Bergen 8, Bergen.
- Pichler, A. 2010. *Towards the New Bergen Electronic Edition*, in: N. Venturinha (ed.), *Wittgenstein After His Nachlass*, 157–172, Houndmills.
- Pichler, A., Biggs, M. A. R. and Uffelmann, S. A. 2011. *Bibliographie der deutsch- und englischsprachigen Wittgenstein-Ausgaben*, in: *Wittgenstein-Studien*, 2, 249–286. [Updated January 2019, <http://www.ilwg.eu/files/Bibliographie%20-%202019-11-26.pdf>]
- Pichler, A. and Zöllner-Weber, A. 2013. *Sharing and debating Wittgenstein by using an ontology*, *Literary and Linguistic Computing*, 28 (4), 700–707.
- Pichler, A. and Bruvik, T. M. 2014. *Digital Critical Editing: Separating Encoding from Presentation*, in: D. Apollon, C. Bélisle, Ph. Régner (eds.), *Digital Critical Editions*, 179–199, Urbana Champaign.
- Pichler, A. 2019. *A brief update on editions offered by the Wittgenstein Archives at the University of Bergen and licences for their use (as of June 2018)*, in: *Wittgenstein-Studien*, 10(1), 139–146.
- Pichler, A. 2020a. *Wittgenstein Nachlass Ts-226: A case of Wittgensteinian (Self-)Translation*, in: P. Oliveira, A. Moreno, A. Pichler (eds.), *Wittgenstein in /on translation*, *Colecao CLE* 86, 153–188, Campinas.
- Pichler, A. 2020b. *A Typology of the Philosopher Ludwig Wittgenstein's Writing of Text Alternatives*, *Aisthesis. Pratiche, linguaggi e saperi dell'estetico*, 13(2), 107–116.
- Pichler, A. 2021, forthcoming. *Hierarchical or Non-hierarchical? A Philosophical Approach to a Debate in Text Encoding*, in: *DHQ: Digital Humanities Quarterly*.
- Pierazzo, E. 2009. *Digital genetic editions: the encoding of time in manuscript transcription*, in: M. Deegan, K. Sutherland (eds.), *Text Editing, Print and the Digital World*, 169–186, Farnham.
- Rehbein, M. 1998. *Die dynamische digitale Textedition: Ein Modell*, in: Ed. Hans-Heinrich Ebeling, Hans-Reinhard Fricke, Peter Hoheisel, Malte Rehbein & Manfred Thaller (eds.), *Vom digitalen Archiv zur digitalen Edition*, 5–21. Göttingen.
- Röhler, I. 2019. *Lexikon, Syntax und Semantik – computerlinguistische Untersuchungen zum Nachlass Ludwig Wittgensteins*, Master's thesis at LMU München, Munich.

- Rothhaupt, J. and Vossenkuhl, W. (eds.). 2013. *Kulturen und Werte: Wittgensteins KRINGEL-BUCH als Initialtext*. Berlin.
- Rothhaupt, J. 2016. *Wittgensteins Zettelsammlung TS212 – The Proto-Big Typescript – Ein Paradebeispiel für innovative Möglichkeiten der digitalen und virtuellen Nachlassforschung (BEE, BNE, Wittgenstein Source, HyperWittgenstein, WITTFind)*, in: Contributions of the Austrian Ludwig Wittgenstein Society, 206-210, Kirchberg a. W.
- Suber, P. 2003. *Removing the Barriers to Research: An Introduction to Open Access for Librarians*, in: College & Research Libraries News, 64, 92–94, 113 [unabridged online version at <http://legacy.earlham.edu/~peters/writing/acrl.htm>].
- Ullrich, S. 2019. *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*, Master's thesis at LMU München, Munich.
- Wittgenstein, L. 1998. *Vermischte Bemerkungen. Eine Auswahl aus dem Nachlaß / Culture and Value. A Selection from the Posthumous Remains*, ed. by G. H. von Wright in collaboration with Heikki Nyman, revised edition of the text by Alois Pichler, transl. by Peter Winch. Oxford.
- Wittgenstein, L. 2000. *Wittgenstein's Nachlass: The Bergen Electronic Edition (BEE)*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Claus Huitfeldt, Oxford.
- Wittgenstein, L. 2009. *Philosophical Investigations / Philosophische Untersuchungen*, ed. by P. M. S. Hacker and Joachim Schulte, trans. G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte, New York.
- Wittgenstein, L. 2015–. *Wittgenstein Source Bergen Nachlass Edition (BNE)*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, in: Wittgenstein Source (2009–) [wittgensteinsource.org], Bergen.
- Wittgenstein, L. 2016–. Interactive Dynamic Presentation (IDP) of Ludwig Wittgenstein's philosophical Nachlass [<http://wittgensteinonline.no/>], ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, Bergen.
- Wright, G.H. von. 1969. *The Wittgenstein papers*, The Philosophical Review 78(4), 483–503.

LABLASS and the BULGARIAN LABLING CORPUS for Teaching Linguistics

Velka Popova
Laboratory of Applied
Linguistics, Konstantin
Preslavsky University of
Shumen, Bulgaria
v.popova@shu.bg

Radostina Iglíkova
Laboratory of Applied
Linguistics, Konstantin
Preslavsky University of
Shumen, Bulgaria
r.iglikova@shu.bg

Krasimir Kordov
Laboratory of Applied
Linguistics, Konstantin
Preslavsky University of
Shumen, Bulgaria
krasimir.kordov@shu.bg

Abstract

The article reviews the first steps in integrating CLARIN into the curriculum at Konstantin Preslavsky University in Shumen, Bulgaria. It discusses the transition from informational seminars for undergraduate and PhD students of different majors regarding the possibilities of this European interdisciplinary network all the way to the specific first steps towards integrating its resources and instruments in the process of education. The focus here is on the approbation of resources and instruments developed within the ClaDA-BG project and, specifically, on the application of two products of the LABLING Laboratory of Applied Linguistics at Shumen University as technological partner to the National Consortium ClaDA-BG - the LABLASS web-based system for researching free speech associations and the BULGARIAN LABLING CORPUS of systematized child speech data. The introduction of their pilot versions emphasizes their importance for achieving higher standards of research work although the accent falls on their application within teaching Linguistics. They concern the updating of the curriculum content and the practical modules of various linguistic disciplines, the creation of new resources, the introduction of interdisciplinary scenarios for classwork with teachers of different academic backgrounds (a linguist and an IT specialist), as well as the transitioning of the teaching process out of the classroom and into the research lab.

1 Introduction

The CLaDA-BG national infrastructure aims at supplying resources and instruments to arts-, humanities-, and social studies researchers with the expectation that the actual applications will exceed the research frame and the results will thus be applicable to the field of education as well (Osenova and Simov, 2020). This idea fits naturally in the new *CLARIN in the Classroom* initiative which has to do with the inclusion of CLARIN resources, instruments and services in university education where the competences paradigm is continuously being recognized as one of the main ways of resolving the crisis between the accumulation of large volumes of knowledge and the inability of the students to use it in practice (for more details see (Popova, 2018)). And since the research process guided by the idea of “learning through exploration” to a large extent determines the competences paradigm in education, the *CLARIN in the Classroom* initiative can be recognized as timely and useful.

The article reviews the first steps in integrating CLARIN into the curriculum at Shumen University. The transition from informational seminars for undergraduate and PhD students of different majors regarding the possibilities of this European interdisciplinary network all the way to the specific first steps towards integrating its resources and instruments in education is discussed. The focus here is on the approbation of resources and instruments developed within the ClaDa-BG project and specifically on the application of two products of the LABLING Laboratory of Applied Linguistics at Shumen University as technological partner to the National Consortium ClaDa-BG – the LABLASS web-based system for researching free speech

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

associations and the BULGARIAN LABLING CORPUS of systematized child speech data. The introduction of their pilot versions emphasizes their importance for achieving higher standards of research work, although the accent falls on their application within the teaching of Linguistics.

2 The Role of the General Information and Specialized Seminars for the Professional Development of the Student-Participants in the CLaDA-BG Project

The *CLARIN in the Classroom* initiative would not have found fertile ground for development in the education process at Shumen University if it had not been preceded by informational events which provided a high level of awareness among the academic staff regarding their wide applicability in the social sciences and the humanities. Therefore, immediately after Bulgaria signed the Memorandum of Understanding in 2012, thus becoming one of the 9 founding members of CLARIN ERIC, LABLING began to organize and hold annual seminars for undergraduate and PhD students from different majors to popularize the already available resources and functionalities of the network for interdisciplinary interaction, having discussions regarding the problems of corpus linguistics and the creation of digital linguistic models.

The seminars were organized in the context of excellent cooperation and mutual assistance between the separate partner associations within CLaDA-BG. This ensured that undergraduate and PhD students had access to information regarding the resources, instruments and services contributed not only by researchers within LABLING but also by other teams within the Bulgarian consortium and the European CLARIN. In order to achieve this the national CLaDA-BG coordinator, prof. Kiril Simov of the Institute of information and communication technologies at the Bulgarian Academy of Sciences and prof. Petya Osenova of Sofia University were regularly invited as guest lecturers. A staple of the programme at each of these events was the discussion on the problems featured in the lectures.

With the development of the CLaDA-BG infrastructure and the establishing of LABLING as a technological partner came inner-circle, more specialized seminars in a narrower format. They were related to the preparation of the participants in the project for work on various research activities as well as for working with various platforms. The knowledge and skills acquired in the process further broadened the professional skills of the students and helped their personal and language development. In this sense, this extracurricular activity can definitely be interpreted as a specific element of the university education process which has transitioned out of the classroom and into the research lab.

The specialized LABLING seminars held so far are related to the two priority research areas for the LABLING team, namely the creation of collections of associative and child speech databases. For this purpose the undergraduate and PhD students were introduced to the resources and instruments of the respective platforms (LABLASS and CHILDES) and the necessary skills for working in this specific environment developed on this basis. In the context of this training of sorts the young people were divided into small groups to ensure excellent work feedback and the ability to independently fulfill research tasks. Along with the specific skills necessary for every linguist such as the ability to collect and systematize data, transcribe speech data in specific formats, annotate and code corpora etc., the young people were able to gather enough experience with working as a team, as well as to develop in themselves the associated personal qualities.

In conclusion it can be stated that before CLARIN entered the academic classroom with its resources, instruments and services the classroom itself had become a workshop of sorts for CLARIN where in the process of creating child speech corpora and associative collections the trainees acquired research competences and skills, as well as the self-esteem that their future products will return to them and their colleagues in the university auditorium.

2.1 The LABLASS Web-Based System and the Bulgarian LabLing Corpus in Teaching Linguistics at University

The pilot versions of the LABLASS web-based system and the Bulgarian LabLing Corpus are already a fact after two-years of work (2019-2020). They have been immediately included in the curricula of linguistics disciplines and the newly published textbook by one of the authors of the present article (Popova, 2020). Within the context of the importance of these two CLaDA-BG products, the following section will attempt to present the specific projections of their useful application in teaching Linguistics in some of the philological and pedagogical majors at Shumen University. The fall of 2020 saw the addition of the first Bulgarian child speech corpus (Bulgarian LabLing Corpus) to the Slavic languages section of the CHILDES database platform which includes data about the acquisition of numerous languages from various language families (see Fig. 1).


CHILDES				Slavic Corpora	
Corpus	Age Range	N	Media	Comments	
<i>Bulgarian</i>					
LabLing	1-5	5, 50	some audio	5 longitudinal, 50 narrative	
<i>Croatian</i>					
Kovacevic	1;3-2;8 1;10-2;11 0;10-3;2	3	audio	Two girls and a boy learning Croatian in Zagreb	
MAIN	5-63	143	audio	MAIN protocol	
<i>Czech</i>					
Chromá	1;7-3;9	6	audio	Two boys and four girls learning Czech in Prague	
<i>Polish</i>					
CDS	1-8	many	-	Frequency list for child-directed speech from eight corpora	
Szuman	1;5-7;9	10	-	Diary data collected by Szuman and his students and computerized by Magdalena Smoczynska	
WeistJarosz	1;7-2;6	3	audio	in PhonBank	
<i>Russian</i>					
Protassova	1;6-2;10	1	-	Longitudinal study of a child learning Russian	
Tanja	2;5-2;11	1	-	A child learning Russian in a monolingual environment in the United States	
<i>Serbian</i>					
SCECL	1;6-4;0	8	audio	Recordings in homes with many people included	
<i>Slovenian</i>					
Zagar	5;0	20	-	Arguments patterns in kindergarten children	

Figure 1. Bulgarian LabLing Corpus in the CHILDES Slavic Collection.

The published first corpus of children's speech (Bulgarian LabLing Corpus) is freely accessible for researchers at <https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html>. Its broad applicability is due to the fact that each of the transcripts includes data for identification of the research subjects (demographic and language parameters) and regarding

the type of the corpus itself (longitudinal or cross-sectional). Meanwhile, with its addition to the common database of CHILDES, the abilities of the system for crosslinguistic research is enriched with another Slavic language. Additionally, the Bulgarian linguistic tradition is enriched with another universal applicable standard for researching language ontogenesis thanks to which researchers can make fast, exact and reliable juxtapositions with a large number of languages and build solid typologies and modern theories.

We also need to point out the unquestionable benefits of broadening the students' corpus competence by including knowledge about CHILDES as part of TalkBank – one of the most successful contemporary platforms for studying human speech behaviour. Thus the curricula of the disciplines Psycholinguistics, Foundations of language acquisition, Linguistics, Child Linguistics have been thematically extended which in turn improved the standard and the quality of independent student research in the form of course assignments and theses. In addition, the data from the Bulgarian LabLing Corpus are often used in the teaching process as their multimodality makes them useful and applicable in different demonstrations (see Fig. 2).

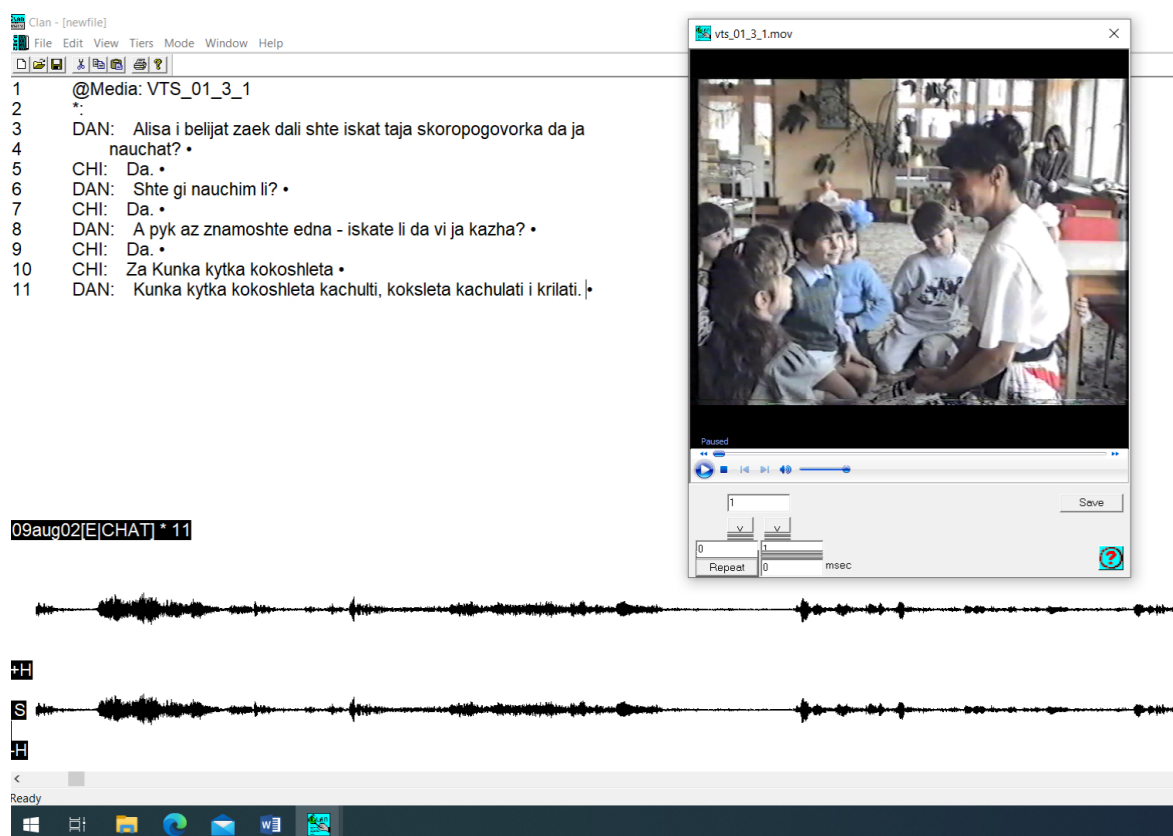


Figure 2. Multimodal data presentation.

The LABLASS web-based system developed within the CLaDA-BG project does not at this point provide free access to users although this is projected to happen by the end of 2021. Its usefulness for researchers is not limited to the availability of the associative collections of included lexicographic sources but is instead much broader and concerns providing the option for the user to create their own new dictionaries and to visualize and compare data from various sources. This web-based system has its place in the practical units of the disciplines dealing with topics such as the mental lexicon and language ontogenesis since word associations are an extremely important source of information in that respect.

In their classes, students can test their working hypotheses by comparing and analyzing published data, creating their own dictionaries. On this basis they develop their own research within the specific course projects. A very important positive result in this respect is the successful de-

fense in 2019 of an MA thesis entitled “Specificities of the Vocabulary of the Bulgarian Native Speaker Nowadays. A Psycholinguistic Study”.

The success of independent student research is supported not only by the resources and instruments of the web-based system LABLASS and the CHILDES platform, but also by the additional unit in the curriculum presented in the textbook, *Psycholinguistics as Experimental Linguistics* (Popova, 2020). It is a workshop of sorts aimed at developing students’ skills for planning and conducting associative experiments and for gathering empirical data and structuring it as a corpus. Also undeniable is the importance of the updated curriculum content of the practical units as well as the academic guidance on the part of the professor and the inclusion of interdisciplinary scenarios for classes with the participation of specialists from different academic fields (a linguist and an IT specialist).

In order to illustrate the ideas mentioned above we can use as an example a pilot model for curricular work in Psycholinguistics with Special Pedagogy students at Shumen University during the winter semester of the 2020–2021 academic year. We shall observe the scenario which includes a cycle of classes supplying the students with knowledge and skills for planning and completing an independent study.

The first class was entitled “Theory and practice of the psycholinguistic experiment”. It was held on the 3rd of November, 2020. The interdisciplinary scenario with the participation of specialists from different academic fields – the psycholinguistics leader and the IT specialist guest lecturer (in fact one of the creators of the LABLASS web-based system) – first was successfully tested (see Fig. 3).

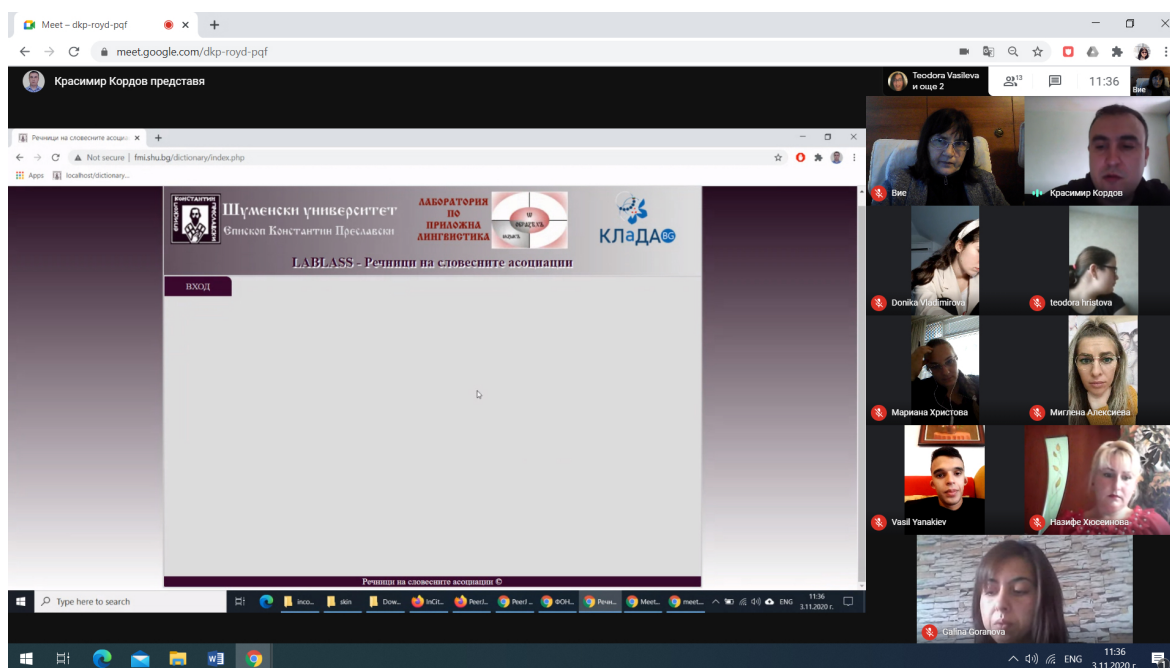


Figure 3. An interdisciplinary scenario class - 3.11.2020, Shumen University.

At the beginning of the class the professor with the leading role in the class, the psycholinguist, introduced the topic in accordance with the compulsory curriculum. After that the guest-lecturer (as illustrated above) was given the opportunity to present the functionalities of the newly-developed web-based system to the students. The discussion following the demonstration of the platform included the contributions of the student-participants in the CLaDA-BG project as well.

A similar scenario (albeit without a guest-lecturer) was employed during the next class dedicated to the importance of corpora and Corpus Linguistics for the professional capacity of the special pedagogue. The TALKBANK and CHILDES platforms were used as general illus-

tration after which, within the context of introducing their functionalities and resources, the functionalities of the system were demonstrated.

After the first two classes came a discussion of the research intentions of each of the trainees with a view to the planning and implementing of an independent study.

This pilot cycle integrated within the Psycholinguistics course is aimed mainly at research techniques and work methods since they are the ones which create the opportunity for achieving the optimal balance between theoretical knowledge and practical skills in the process of education, as well as for the development and broadening of the competences of the trainees. In addition, the updating of the respective curriculum content with regard to the introduction to the pilot versions of LABLASS and the Bulgarian LabLing corpus enables students to be more active in the process, i.e. they are not handed down knowledge but instead co-discover it along with the teacher as academic advisor.

Similar was the scenario for the classes in Child Linguistics and Foundations of language acquisition for students of philology where the practical unit also stood out. In the Introduction to general linguistics course, however, the work on improving the corpus competences of the students was done only in the form of an addition of sorts to the curriculum content, and done only for some topics. Moreover, during the course of the specific classes, while working on particular linguistic case studies and in the completion of individual and group tasks for learning and research, the students were given the opportunity to test some of the functionalities of the web-based system LABLASS as well as those of the CHILDES platform.

3 Conclusion

The article attempted to demonstrate some of the positive effects of the *CLARIN in the Classroom* initiative, of the example of the pilot testing of the web-based system LABLASS and the first Bulgarian child speech corpus (Bulgarian LabLing Corpus) developed within the CLaDA-BG project for teaching linguistics to students of philology and special pedagogy at Shumen University. To summarize, the specific applications of these products in the university teaching practice concern mainly the updating of the curriculum content and the practical units of the specific linguistic disciplines, the creation of teaching resources, the introduction of interdisciplinary scenarios for classes with the participation of specialists from different academic fields, as well as the transitioning of the teaching process out of the class-room and into the research lab. The preliminary results include the improved theoretical and practical competences, language development, research curiosity and professional self-esteem directly reflected in the independent articles, presentations, thesis projects of students where the young people's interest in other CLARIN resources, instruments and services is already observable.

Acknowledgements

This research was partially funded by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO01-272/16.12.2019.

References

- Simov K., Osenova P. 2020. Integrated Language and Knowledge Resources for CLaDA-BG. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings: 137–144.
- Popova V. 2018. *IzledovatelSKIyat podhod v obuchenieto po balgarski ezik mezhdu tradicionnoto i inovativnoto* - Bulgarian Language and Literature. 62(3) (in Bulgarian).
- Popova V. 2020. Psiholingvistikata kato experimentalno ezikoznanie. *Shumen: Univ. izdatelstvo "Episkop Konstantin Preslavsky"* (in Bulgarian).

A Pipeline for Manual Annotations of Risk Factor Mentions in the COVID-19 Open Research Dataset

Maria Skeppstedt, Magnus Ahltop, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, the Institute for Language and Folklore, Sweden
firstname.lastname@isof.se

Abstract

We here demonstrate how a set of tools that are being maintained and further developed within the Språkbanken Sam and SWE-CLARIN infrastructures can be employed for creating manually labelled training data in a low-resource setting. As example text, we used the “COVID-19 Open Research Dataset”, and created manually annotated training data for its associated Kaggle task, “What do we know about COVID-19 risk factors?”. We first used our *topic modelling tool* to i) select a text set for manual annotation, ii) classify the texts into preliminary classification categories, and iii) analyse the texts in search for potential refinements of the annotation categories. We then annotated the text set on a more granular level by labelling the token sequences that indicated the existence of the refined categories in the text. Finally, we used the granularly annotated text set as a seed set, and applied our *active learning tool* for actively selecting additional texts for annotation. For the token-sequence annotations, we used our *text annotation tool*, which includes support for incorporating automatic pre-annotations.

1 Introduction

The COVID-19 Open Research Dataset (CORD-19) is a free resource with scholarly articles on viruses from the coronavirus family, and on related topics (Wang et al., 2020). Associated with the dataset is the Kaggle COVID-19 Open Research Dataset Challenge (Allen Institute For AI, 2020), which consists of nine different tasks, all with the aim of extracting from the data what has been published regarding different COVID-19-related research questions.

In order to demonstrate how a set of tools that are being maintained and further developed within the Språkbanken Sam and SWE-CLARIN infrastructures can be combined into a pipeline and employed for creating a manually labelled text corpus, we used the CORD-19 dataset as an example text set. In particular, the aim was to demonstrate how the tools can be useful in a low-resource setting, i.e. with no existing previously annotated data, and with only limited time available for performing manual annotations.

As the example task, for which we aimed to create manually labelled data, we selected one of the Kaggle tasks associated with the dataset, the task “What do we know about COVID-19 risk factors?”. For our tool demonstration, we used the version of the CORD-19 dataset that was made available in spring 2020, which contains around 40,000 full text articles.

The first tool used in the pipeline was our topic modelling tool, *Topics2Themes* (Skeppstedt et al., 2018). The tool was used for i) *selecting* data for manual annotation, ii) *classifying* the data into preliminary classification categories, and iii) *analysing* the text material in search for potential refinements of the preliminary annotation categories.

The second tool used was our Språkbanken Sam tool for manual text annotation of token sequences. We employed this tool to manually annotate the texts selected using *Topics2Themes* according to the refined annotation categories established in the previous step. The annotations were in the form of token sequences that indicated the existence of one of the refined categories in the text.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

The third tool employed in the pipeline was our tool for active learning, *PAL*. We used this tool for actively selecting additional texts to manually annotate. These further annotations were also carried out using our Språkbanken Sam tool for manual text annotation. *PAL* not only actively selects what texts to annotate, but also provides these texts with pre-annotations, using the same machine learning model it uses for text selection. Our manual text annotation tool allows the annotator to correct these pre-annotated labels, as well as to provide the text with new labels. *PAL*, together with our text annotation tool, thus combines parts of the functionality that is made available by i) a tool such as WebAnno (Yimam et al., 2014), in which annotation suggestions are provided, with ii) the functionality of tools such as JANE (Tomanek et al., 2007) and WordFreak (Morton and LaCivita, 2003), in which data selection through active learning is carried out.

Qualitative text analyses, similar to our approach in the first step of the pipeline, have been conducted in previous research as a preparation for creating annotation categories in the medical domain (Mowery et al., 2012). However, we are not aware of any previous studies that have used a topic modelling tool like Topics2Themes for this task. We are also not aware of previous studies in which this approach has been combined with a pipeline that includes active learning and pre-annotation.

We will here present how we used the three tools in a pipeline for producing annotated texts. The annotation tool will be presented and discussed twice, first when it is used for annotating the texts selected with the topic modelling tool, and thereafter when it is used with pre-annotation of texts selected in the active learning process. We will, however, start by briefly presenting the Kaggle task and our approach for using a set of risk factor keywords for compiling a “Risk factor sub-corpus” from *CORD-19*.

2 The Kaggle Task and Our “Risk Factor Sub-Corpus”

For the “What do we know about COVID-19 risk factors?”-task, the participants are asked to find out what epidemiological studies report about potential risk factors for COVID-19. Our suggestion for how to approach this task is to train a model to recognise language expressions that are used for describing risk factors for diseases in general, i.e. expressions that we could call “risk factor triggers”. To be able to train such a model, a text corpus where such expressions have been manually labelled is needed.

As the first step for creating such a corpus, we compiled a “Risk factor sub-corpus” by extracting 30,000 paragraphs from *CORD-19* that contained a risk factor seed word. We used a list of 104 seed words¹ which we had compiled in the following manner: (i) We first constructed a list of the words (or sometimes bi- and tri-grams) that occurred in the call for the “What do we know about COVID-19 risk factors?”-task, and that we estimated would be good seed words for more general text about risk factors. These included, for instance, “risk factors”, “factors”, “co-infections”, “co-morbidities”, “high-risk”, “pre-existing”, and “susceptibility of”. (ii) We, thereafter, expanded the list by adding synonyms from the Gavagai living lexicon (Sahlgren et al., 2016). (iii) Words in the list that occurred very often in the *CORD-19* corpus were then removed from the seed list as unigrams and specified further with bi-grams. E.g., “factors” was removed and replaced by bi-grams such as “socio-economic factors” and “environmental factors”. (iv) We finally read some of the paragraphs containing the seed words, in search for new words to add, and found words such as “more common among” and “more likely”.

3 Selecting, Classifying and Analysing the Data with Our Topic Modelling Tool

3.1 Method

As our time available for manual annotation was limited, we decided to focus our effort on text paragraphs with a content typical for the 30,000 paragraphs in the risk factor sub-corpus. With limited annotation resources, we are not likely to be able to catch outliers, or even moderately infrequent content, but we might be able to gather data for training a model that catches typical expressions. Finding topics that represent re-occurring content in a text collection, and creating automatic classes in this content, can be done in an unsupervised fashion, for instance by using topic modelling. For this task, we therefore used the topic modelling tool, Topics2Themes. We used the tool’s ability to automatically find synonym

¹<https://www.kaggle.com/mariaskoppstedt/trigger-words>

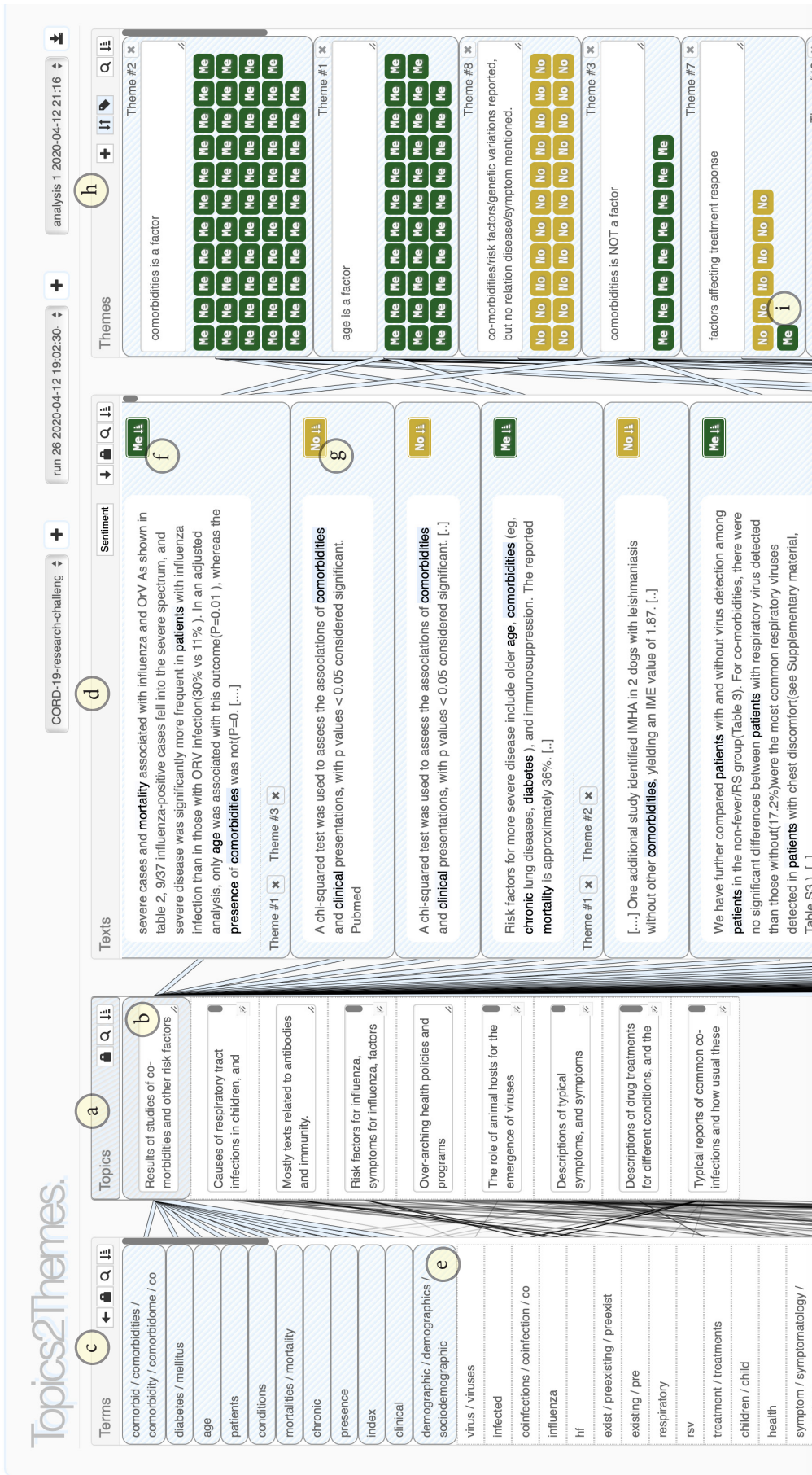


Figure 1. (a) The *Topics* panel contains the nine automatically extracted topics. (b) The user has selected the first topic “Results of studies of co-morbidities and other risk factors” by double-clicking on the topic element in the panel. This has resulted in that the terms associated with this topic, in (c) the *Terms* panel – as well as the texts associated with this topic, in (d) the *Texts* panel – have been sorted as the top-ranked elements in their panels. (e) The *Terms* panel also shows the results of the automatic synonym clustering, based on word embeddings. (f) Texts with a green label (Me) have been manually classified to contain mentions of risk factors, (g) whereas texts with a yellow label (No) have been classified as not mentioning risk factors. (h) The *Themes* panel contains themes that the user has manually identified when analysing the texts. (i) The theme element labels show how many of their associated texts that have been classified as mentioning risk factors.

clusters in the text with the help of word embeddings. We used embeddings pre-trained on biomedical text (McDonald et al., 2018; Brokos, 2018).

We configured Topics2Themes to use NMF topic modelling to try to find a maximum of 20 topics in the dataset, since our limited annotation resources would not allow us to analyse more than 20 topics. Since NMF is a randomised algorithm, which might produce slightly different results each time it is run, the algorithm was run 50 times, and only topics stable enough to occur in all re-runs were retained. This resulted in nine stable topics being detected, which indicates that it is not likely that the NMF algorithm would have been able to find more than 20 stable topics in this text collection, even if we had configured it to search for more topics. For each topic detected by the topic modelling algorithm, the terms and the texts that are closely associated with the topic are presented by the Topics2Themes tool. For each of the topics extracted, we used the annotation functionality of the tool to manually classify its 15 most closely associated texts as to whether they contained a mention of a risk factor for a disease or not. For the topics that had more than one text among these 15 texts that contained mentions of risk factors, we manually classified additional associated texts in the same manner (up to 70 additional texts for each of these topics).

In addition to classifying the texts for risk factor mentions, the functionality in Topics2Themes for documenting re-occurring themes that are identified when manually analysing the texts was used. The name of the tool, Topics2Themes, reflects this functionality of the tool, i.e., (i) that the automatically created categories – the *topics* detected by the topic modelling algorithm – are presented to the user, and (ii) that the user then manually analyses the output of the algorithm and creates user-defined categories – *themes* – which correspond to the content that the user identifies as important in this output. Typically, such manually identified themes represent re-occurring information in the text collection on a more detailed level than the automatically extracted topics. The aim of the manual analysis was in this case to identify whether there were annotation categories in addition to “describing a disease risk factor” that would be relevant, as well as to determine whether this category should be refined.

3.2 Results and Reflections

Topics2Themes produced nine stable topics for our corpus. Figure 1 shows the user interface of the Topics2Themes tool, and the nine topics detected are shown in its *second* panel. The topic descriptions were given by us, after we had analysed the texts most closely associated with the topics. The figure shows when the topic “Results of studies of co-morbidities and other risk factors” has been selected by the user (as indicated by the blue background). Blue lines connect this selected element with terms associated with the topic (to the left), and texts associated with the topic (to the right).

In the *first* (and left-most) panel, which thus contains the terms associated with the topics extracted, examples of the embedding-based synonym clustering can be seen.

In the *third* panel, which thus contains the texts associated with the topics, each text has an assigned label that is a result from our manual classification. A green label (Me) assigned to the text shows that the text has been classified as mentioning a risk factor for a disease, whereas a yellow label (No) shows that the text has been classified as not containing any information on risk factors. We performed a manual classification for a total of 418 texts, and 150 texts among them were classified as describing risk factors for diseases.

Most of the manually classified texts were associated with five of the topics detected. For these five topics, more than one of the top 15 most closely associated texts described a risk factor for a disease. As described above, we classified additional texts for these topics, in addition to the top 15 most closely associated texts. The five topics were: “Results of studies of co-morbidities and other risk factors”, “Causes of respiratory tract infections in children, and whether such previous infections influence the development of asthma”, “Mostly texts related to antibodies and immunity”, “Risk factors for influenza, symptoms for influenza, factors influencing whether people vaccinate or not”, and “Typical reports of common co-infections and how usual these are, and sometimes, also studies of whether they effect severity”.

The *fourth* (and right-most) panel contains elements added by the user, in the form of themes identified when analysing the texts. Among the themes identified, 18 re-occurring themes were found. Examples

include “genetics (and family history) is a factor”, “co-infection is a factor” and “age is a factor”. Five additional examples are shown in the right-most panel in Figure 1. In 24 of the texts analysed, it is described that something could *not* be shown to be a risk factor. Five of the re-occurring themes identified described information of this type. Studies, in which it has *not* been possible to show that something is a risk factor, should be important to identify when mining for risk factors, since the information mined otherwise would be biased towards positive research results. We therefore decided to also include this information in the data to annotate, and consequently classified each such text with the green label that signifies that the text describes risk factors. We also used this output of the analysis – i.e., the frequent occurrences of themes describing that something could *not* be shown to be a risk factor – for refining our preliminary annotation categories, as will be described in the next section².

4 Annotation of Token Sequences Using Our Manual Annotation Tool

4.1 Method

We then decided to provide the texts with more granular annotations, which could be useful for training a machine learning model to detect the language that is used for expressing that something is a risk factor for a disease, i.e., what could be called “risk factor triggers”. We therefore extended the 150 paragraphs in which disease risk factors were described by labelling the token sequences that the authors had used for expressing that something is a risk factor. That is, we did not annotate the token sequence that describes the disease, nor the token sequence that describes the risk factor, but the language expressions used for indicating that something is a risk factor. For instance the expression “carry a heightened risk of”, as shown in Figure 2.

For this more granular annotation, we also – based on the analysis in the previous step – decided to refine the annotation categories by differentiating between if the text described that something was shown to be a risk factor, or if it described that something could *not* be shown to be a risk factor. We consequently created two annotation categories to differentiate between if the token sequences functioned as (i) a risk factor trigger, or (ii) a trigger indicating that something could *not* be shown to be a risk factor. For instance, in the text: “*2019nCoV was of clustering onset, is more likely to infect older men with comorbidities [...]*”, the underlined text was annotated as a risk factor trigger. In contrast, the underlined text in “*The incubation periods did not significantly differ according to age, sex, or the presence of comorbidities [...]*” was annotated as a trigger describing that something could *not* be shown to be a risk factor.

For the token sequence annotations, we used our Språkbanken Sam tool for manual annotation of sequences of text. The user interface of the tool is shown in Figure 2. The tool provides support for two different types of annotations, i) one-token annotations where the annotatable tokens are pre-defined, and ii) annotations of token sequences that are to be IOB-coded. We here used the tool setting that provides support for the latter type.³ The tool is optimised for annotation speed, both when adding new labels and when changing existing labels. The annotator can either use the mouse to select a single token or to select a sequence of tokens. This results in a pop-up (shown in Figure 2) which allows the annotator to quickly choose which label to use for the token(s) selected. If a B-tag is chosen for a selected sequence of tokens, the first token in this sequence will be given the B-tag, and the subsequent tokens in the selected region will be given I-tags.

4.2 Results and Reflections

The total size of the text set annotated was around 50,000 tokens. In this set, there were a total of 282 token sequences indicating that something is a risk factor, and 44 token sequences indicating that something could *not* be shown to be a risk factor. The types of expressions that we target thus occur very rarely in our corpus, despite the fact that texts with a higher probability of containing mentions of risk factors had been selected for the text set. That is, the 50,000-token text set was not a randomly selected

²The full configuration and analysis is available at: <https://www.kaggle.com/mariaskeppstedt/cord19clarin2020analysis>

³The IOB format for the entities in the examples sentence were thus coded as:
[... (is, **B-RISK**) (more, **I**) (likely, **I**) (to, **I**) ...] and [... (did, **B-NO**) (not, **I**) (significantly, **I**) (differ, **I**) (according, **I**) (to, **I**) ...]

Worthy of note is that 11 % of all infants are born premature, and this population thus represents some 12.9 million infants per year worldwide. 5 Preterm infants carry a heightened risk of infectious ailments of both bacterial and viral cause and undernourishment, aggravating this susceptibility during the first years of life, with rhinovirus being a major cause of mortality and morbidity worldwide, particularly in low-income countries. Although strict hygiene measures have been shown to reduce transmission and thus diminish the incidence of rhinovirus infections, no definitive preventive measures have been discovered for the effective control of this entity. On the basis of our results, gut microbiota modulation through the use of specific prebiotics, probiotics, or both could offer a cost-effective tool in the fight against RTIs, hopefully also in the developing world.

Figure 2. The pre-annotation functionality has labelled the token “carry” with the category *B-RISK*. The user has then selected the subsequent four tokens for annotation. This has resulted in a pop-up with the four possible annotation categories: (i) *B-RISK* is the first token – in a sequence of tokens – in an expression which indicates that something is a risk factor, (ii) correspondingly, *B-NO* is the first token in an expression indicating that something could not be shown to be a risk factor, (iii) *I* is the subsequent tokens in the expression (i.e., the category of the first token in an expression determines its type in the annotation tool), and finally (iv) *O* signifies that a token is not included in an expression (this is also the default category for tokens not annotated). The user has here chosen the *I* category, which has the effect that the sequence of tokens “carry a heightened risk of” will be labelled as belonging to the *RISK* category when the annotated data is exported from the tool.

text set, but one with texts that had first been selected based on that they contained keywords for risk factors, and many of them were associated with one of the five topic-modelling-topics that were related to risk factors. Thereby, in our low-resource setting, it would probably not have been fruitful to select a random set of texts for annotation, since it is likely that very few risk factor mentions would have been found.

5 Selecting More Texts for Token Sequence Annotation, Using Our Active Learning Tool

5.1 Method

The third component in our pipeline for creating annotated training data while still only employing limited manual annotation resources was to use our active learning tool, PAL (Skeppstedt et al., 2016). Active learning is a machine learning/data selection technique, where data for manual annotation is actively selected by a machine learning model. Thereby, the machine learning model has the possibility to select the data points – from a large pool of unannotated data – which are most useful for improving the model. For instance, the machine learning model could use uncertainty sampling (Schein and Ungar, 2007; Settles, 2009), i.e., select those data points for which the model is most uncertain regarding how the data should be classified. In a successful active learning set-up, a machine learning model trained on relatively few data points would yield the same performance as a model trained on a larger dataset. It is thereby possible to limit the number of training data points that need to be manually labelled. We have previously conducted experiments with PAL, evaluating the active learning functionality through simulations on labelled datasets (Skeppstedt et al., 2019). For the task of training a model to recognise three different kinds of named entities in tweets, active learning was shown to be more efficient in the use of manually labelled training data than a random selection of manually annotated tweets. That is, models trained on actively selected tweets performed better than models trained on the same amount of randomly selected tweets.

PAL is targeted towards small training datasets, and therefore uses an active learning approach that is more likely to function on small datasets, in the form of uncertainty sampling using a token-level logistic regression classifier. The tool can incorporate features in the form of word embedding vectors

when training the logistic regression classifier. When previously evaluating PAL's performance for named entity recognition in tweets, the incorporation of embeddings was useful for two of the three named entity categories evaluated.

We constructed a pool of 5,000 unlabelled data points, in the form of 5,000 paragraphs from the CORD-19 dataset that had not yet been annotated. That is, we derived the paragraphs for the unlabelled pool from the entire CORD-19 corpus, and not only from our keyword-based "Risk factor sub-corpus". We then used the annotated corpus constructed in the previous steps as the seed set, in order to let PAL train a model to use for active selection of paragraphs from the pool of unlabelled data. We configured the tool to select the 35 most uncertain data points in each active selection/annotation round, but to prioritise texts which the model predicted to contain at least one pre-labelled token. As features, the model was configured to use a concatenated vector consisting of the one-hot encoding of the token to be classified, the one-hot encodings for its four preceding and four following tokens, as well as the embedding vectors for all these nine tokens. We used the same biomedical embeddings (McDonald et al., 2018; Brokos, 2018) that we used for the Topics2Themes tool.

We ran the active learning process in nine iterations. For each iteration, 35 new paragraphs (those containing the 35 tokens for which the machine learning model was most uncertain) were actively selected for manual annotation. After having annotated these 35 paragraphs, they were added as new data samples to the training dataset, i.e., to be used for training the machine learning model for the next active learning iteration.

For each iteration in the active learning process, PAL generates a plot, in order to provide the user with an understanding of how the uncertainty estimations for the unlabelled data pool changes during the process.

5.2 Results and Reflections

Plots generated by PAL for two of the nine iterations are shown in Figure 3. The plots shown are those generated when i) the active learning process is first run (with 418 labelled samples available for training the machine learning model), and ii) the training dataset contains a total of 698 manually labelled samples. The plot included here is shown from the point of view of a model which detects expressions that indicate that something is a risk factor. Thereby, the colour red is used for tokens that the model classifies as risk factor indicators, and blue is used for all other tokens. Corresponding plots (not included here) are generated from the point of view of a model which detects expressions that indicate that something could *not* be shown to be a risk factor.

To the left in the plot, the content of the pool of unlabelled data is visualised, through a t-SNE plot (van der Maaten and Hinton, 2008) of the most frequently occurring words in the data pool. Again, the biomedical embeddings (McDonald et al., 2018; Brokos, 2018) were used. The plot thus shows the semantic distribution of the words in the corpus, where semantically similar words are represented by dots that are positioned close to each other in the plot. The hue of the dot is determined by the token instance of the word for which the model is most uncertain, i.e., the larger the uncertainty for the classification of this token, the darker is the colour with which it is displayed.

To the right of the plot, the 35 tokens for which the model is most uncertain are shown, i.e., the tokens on which the decision for which 35 paragraphs to select for manual annotation is based. The actual token is shown in the center, and to its left and right, its textual context is shown. The bars show the level of uncertainty with which the model has classified the tokens, and the bar colour is determined by the class of the token (as classified by the model). When the model was too uncertain to be able to make a decision for how to classify the token, the bar is shown in black.

Despite the few iterations in which the active learning process was run, the plots generated by PAL show how the state of the pool of unlabelled data changes. After nine iterations, there seems to be less uncertainty left in the data pool. This is most evident by the lengths and colours of the bars representing the tokens, but there is also a small indication through lighter colours in the t-SNE plot and through the bar representing the mean uncertainty in the pool of unlabelled data.

The aim of providing visualisations for the uncertainty left in the pool of unlabelled data is to make the


Expressions indicating that something is a risk factor

418 training samples

Classification uncertainty for the most uncertain tokens in data pool:

RISK model trained on 418 samples

Red: Tokens classified as RISK
Blue: Other tokens.

Data pool: 
2% mean uncertainty left

1: 100%	studies in mice did
2: 100%	.. hospitalizations for stroke
3: 100%	This suggests that proteases
4: 100%	ACE 2 was identified
5: 100%	S. pneumoniae was
6: 100%	The reproduction number was
7: 100%	..ourable and fatal infections
8: 100%	to change the hydrogen
9: 100%	..ificantly increased compared
10: 100%	and humans have been
11: 100%	.. The prevalence was
12: 100%	1 concentration has been
13: 100%	proteins in particular,
14: 100%	are each strongly and
15: 100%	does predict the difference
16: 100%	no more bacteria were
17: 100%	We have shown that
18: 99%	the youngest children are
19: 99%	days from symptom onset
20: 99%	these levels decreased as
21: 99%	did not show any
22: 99%) and to the
23: 99%	..monocyte differentiation can
24: 98%	6.4 days)
25: 98%	has advanced in industries
26: 98%	outbreaks of NEC have
27: 98%	We noted an increase
28: 98%	contrast, those regions
29: 98%	the infection was the
30: 98%	therefore both are likely
31: 97%	plasma protein systems are
32: 97%	This increase may
33: 97%	K d is the
34: 97%	the cell clones showed
35: 97%	study demonstrated a greater

not
by
play
as
significantly
estimated
may
bond
to
reported
calculated
show
may
independently
in
present
there
at
was
expected
significant
deep
have
long
where
been
in
exhibit
interaction
to
selectively
reflect
apparent
different
than

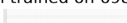
reveal such immune reactions
3.588 amongst the elderly
a role in the
the functional receptor for
reduced in mice immunized
to be as high
be seen. 13
donor / acceptor pro
children with non-LRTI.
among MERS-CoV patients ()
as the number of
to correlate with BALF
produce higher quality models
associated (P <
increment between a C-G
within lymphoid tissues ()
has been a close
highest risk for hospitaliza..
significantly associated wit..
due to their relatively
changes (Fig.
cervical lymph nodes into
both protective and immunopa..
time between onset of
there is a risk
linked to viral infections
BALF cells in both
a high degree of
of the F-glycoprotein of
traffic through lymph nodes
activated ? d Plasma
enhanced production of chemo
dissociation constant of the
degrees of CPE appeared
two-fold increase in potency

698 training samples

Classification uncertainty for the most uncertain tokens in data pool:

RISK model trained on 698 samples

Red: Tokens classified as RISK
Blue: Other tokens.

Data pool: 
1% mean uncertainty left

1: 100%	.. (40 ; 11-38)
2: 100%	ACE 2 was also
3: 100%	..sease incidence patterns are
4: 100%	- 2 levels were
5: 100%	there was no
6: 100%	also could explain the
7: 100%	was associated with a
8: 100%	..ual risk factors potentially
9: 100%	and RANTES were significantly
10: 100%	have been identified as
11: 100%	tract infection and may
12: 100%	we found no significant
13: 100%	to identify patients more
14: 100%	could increase the risk
15: 100%	..ly white participants showed
16: 99%	case fatality ratio and
17: 99%	..ommunity-acquired pneumonia,
18: 99%	consequence, there is
19: 99%	.. 25 which might
20: 99%	No association
21: 99%	viruses, which are
22: 99%	psychiatric consultation had a
23: 98%	life histories are predicted
24: 98%	MBL plays a critical
25: 97%	ratio R is a
26: 96%	they could have greater
27: 96%	an antibody titer increase
28: 95%	the HCAI event rather
29: 95%	..virus infections become more
30: 94%	or IFN γ may
31: 89%	..ated connectivity classes of
32: 88%	ACE 2 plays a
33: 85%	PCV 7 has not
34: 83%	terminated at day 21
35: 82%	.. and is nearly

for
found
prone
compared
significant
increased
significantly
contributing
increased
risk
of
correlation
likely
a
the
an
be
was
found
higher
to
role
crucial
impact
fourfold
than
likely
play
linked
critical
reduced
due
independent

those younger than 50
to play a significant
to underreporting owing to
only with c.
correlation between virus pe..
risk of vRTI in
higher rate of previous
to incidences of stroke
in the patients with
factors in the pathogenesis
individuals with the
between MS patients with
to suffer disease which
mumps, whereas the
stronger relationship between..
difference was marginally st..
increased proportion of pati..
increased proportion of adults
an additional factor affecting
found between responders and
commonly among military cons
risk for psychological distr..
have a greater number
in mediating development of
factor in our risk
. Point of entry
or greater between acute
true risk factors for
with increasing age.
a contributory role in
individuals.
role in SARS pathogenesis
nasopharyngeal carriage of S
to subsequent growth of
of the airflow patterns

Figure 3. Two of the plots generated by PAL during the active learning process: (i) When the active learning process is first run (with 418 annotated training samples), and (ii) after a total of 698 samples have been annotated and added to the training data. To the left, the total uncertainty in the pool of unlabelled data is shown through a t-SNE plot (the darker the colours of the dots, the more uncertainty is left). To the right, the classifier uncertainty for the 35 most uncertain tokens are shown, i.e., the tokens on which the choice of which paragraphs to select for annotation was based.

B-NO

Our study did not identify strong **associations** with underlying chronic illnesses, most likely because the prevalence of such conditions was low (< 10 %) in this population. HCPs with a history of smoking had a **risk** for infection almost 3 times that of nonsmokers. We found **no** association between MERS-CoV infection and sex. Most case series to date have demonstrated a male **predominance** among casepatients (15, 23, 24), but our study suggests this association might be explained by social and behavioral factors that increase exposure to MERS-CoV, rather than a sex-specific difference in biological susceptibility.

Figure 4. Automatic pre-annotations produced by PAL and imported into the Språkbanken Sam tool for manual annotation.

B-NO I I I I I

Our study **did** not identify strong **associations** with underlying chronic illnesses, most likely because the prevalence of such conditions was low (< 10 %) in this population. HCPs with a history of smoking **had** a **risk** for infection **almost** 3 times that of nonsmokers. We found **no** association between MERS-CoV infection and sex. Most case series to date **have** demonstrated a male **predominance** among casepatients (15, 23, 24), **but** our study suggests this association might be explained by social and behavioral factors that increase exposure to MERS-CoV, rather than a sex-specific difference in biological susceptibility.

Figure 5. Manual annotations carried out in the Språkbanken Sam annotation tool.

active learning and annotation process more interesting. Thereby, there is a possibility to also increase the annotator’s intrinsic motivation for the annotation task. That there was a change in visualised uncertainty levels already after nine iterations shows that there is a potential for using these kinds of visualisations for increasing the interest in the annotation task, also very early in the active learning process.

6 Manual Annotation of Token Sequences with Pre-Annotation

6.1 Method

The same logistic regression model, which is used for actively selecting training data samples, is also used by PAL for providing the selected samples with pre-annotated labels.

The pre-annotations from PAL were in previous versions of the tool only provided in the format of the annotation tool BRAT (Stenetorp et al., 2012). However, BRAT provides a rather extensive set of functions for text annotation, which also has the effect that the procedure for annotating token sequences is more time consuming than when using an annotation tool specifically adapted for this task, e.g., the annotation tool which we have developed at Språkbanken Sam. We have therefore made it possible to also import pre-annotations from PAL into the Språkbanken Sam annotation tool.

Figure 4 shows an example of a pre-annotated text, and Figure 5 shows how manual annotations have been provided to the same text.

6.2 Results and Reflections

Also this second part of the manually annotated corpus contains around 50,000 tokens. For this text set, we found a total of 224 token sequences indicating that something is a risk factor, and 39 token sequences describing that something could *not* be shown to be a risk factor. That is, slightly fewer token sequences were detected in the actively selected sub-corpus, than in the one compiled through topic modelling.

While annotating the texts, we could observe that the quality of the pre-annotations was not very high. This is exemplified by the text paragraph in figures 4 and 5. In fact, this paragraph is chosen as an example paragraph here because it contained many instances of annotated token sequences, not because it was a

paragraph representative for the performance of the pre-annotation. It contains one pre-annotated token sequence that was not at all altered by the annotator, which was quite rare.

We had expected that low-quality pre-annotations would disturb the flow of the manual annotations, and therefore not be perceived as useful by the annotator. However, the opposite was experienced. That is, many consecutive sentences without pre-annotated content were subjectively perceived as boring to annotate, while sentences with one or several pre-annotations were found interesting. That the lower-quality pre-annotations were not found disturbing might be explained by the fact that the annotation tool is optimised for annotation speed, also for altering pre-annotated content. When pre-annotated tokens are selected by the annotator and given a different annotation category than the one provided by the pre-annotations, the pre-annotated content is automatically removed by the annotation tool. It is also easy to change the annotation category of a sequence, but to keep the annotated token span, by just changing the annotation category of the first token. The simplicity with which pre-annotations can be altered comes with a trade-off, since – unlike for the BRAT tool – a token cannot be assigned to several categories with this set-up. However, with an annotation task that only allows one category per token, we believe it is better to choose a tool that is optimised for annotation speed.

The sentiment towards pre-annotations with a lower quality that we present here was a subjective assessment made by the annotator in this study. The attitude towards lower-quality pre-annotations might vary between annotators, and not everyone might find that the presence of pre-annotations makes the annotation task more interesting. However, one of the lessons learnt here is that it might be worth to at least give the annotator a choice to include pre-annotations, also pre-annotations with a lower quality. That said, pre-annotations with a very low quality are probably not found useful by any annotator.

7 Discussion and Tool/Data Availability

Our resources for manual annotation were scarce, not only in terms of the number of man-hours that we were able to spend on creating the annotated corpus, but also in terms of competence within the medical domain. One of the authors had previous experience in annotation guideline creation and medical text annotation in collaboration with physicians, and was also the one who carried out the manual annotations. However, without a medical education, some text content is difficult to understand, e.g., to distinguish between risk factors for a disease, and causes, signs and symptoms associated with the disease. Even more difficult is the development of comprehensive annotation guidelines, without access to medical knowledge, e.g., guidelines regarding which annotation categories to include and regarding exactly what should be counted as a risk factor for a disease. We, therefore, did not construct any detailed annotation guidelines, apart from the short description of the two annotation categories given above.

While the fact that we lack medical competence might decrease the value of the annotated corpus created, it also highlights the importance of annotation pipelines, similar to the one we have demonstrated here. That is, annotation pipelines with the potential of supporting annotation guideline creation, facilitating annotation, and minimising the amount of annotated data required. While it is possible to obtain laymen annotations for English texts at a low cost, annotations carried out by annotators with extensive medical knowledge tend to be more expensive. Thereby, it is important to use the resource of medical expertise wisely. A topic modelling tool might give the medical expert an overview of typical categories in the texts, which might help in determining annotation categories and creating guidelines. An annotation tool with a high usability, and through which the annotator is able to track the status of the active learning process, might make it faster to annotate and increase the expert's intrinsic motivation for the annotation task. Finally, an active selection of training data samples that are useful for a machine learning model might make it possible to train useful models without having to manually annotate very large corpora.

Both Topics2Themes⁴ and PAL⁵ are freely available on GitHub. We plan to continue the development of our tool for manual text annotation, and to also make this tool freely available. We will also continue the development of Topics2Themes. After this study was conducted, we have added the functionality of allowing the user to provide a manually constructed list of multiword expressions, i.e., expressions that

⁴<https://github.com/mariask2/topics2themes>

⁵<https://github.com/mariask2/PAL-A-tool-for-Pre-annotation-and-Active-Learning>

are then treated as any other word by the topic modelling algorithm. This functionality could, however, be extended by also providing an automatic detection of multiword expressions.

We have also made our two annotated datasets, i.e., the set selected through topic modelling and the set selected through active learning, freely available at Kaggle.⁶ The two datasets consist of around 100,000 tokens, with a total of 506 token sequences annotated as expressions used for describing that something is a risk factor, and 83 token sequences annotated for descriptions of when something could *not* be shown to be a risk factor. Although the small size of these annotated datasets might not be sufficient for training high-performing classification models, we welcome anyone to use these annotations for classifier experiments, or to use them as seed sets in active learning and/or pre-annotation approaches for further expanding the training dataset. We would also appreciate efforts from others – in particular annotators with a medical background – to annotate the same dataset, to be able to compute inter-annotator agreement, or to use the annotations as a support for developing a set of detailed annotation guidelines.

Our next step will consist of making use of data contributed by others at Kaggle. For the risk factor task, there is structured data collected regarding studies of COVID-19 risk factors, together with links to relevant articles. These articles might be used for collecting and annotating a text set that can be employed as an independent gold standard, against which our approach for creating a training dataset for risk factor mentions can be evaluated.

Although the main purpose of this study has been to demonstrate the use of different types of tools for the creation of an annotated dataset – rather than the resulting dataset – we still consider this type of data, and its associated Kaggle task, as important. Natural language processing tools that can help researchers to access the content of scientific papers regarding risk factors for COVID-19 are useful, for instance when criteria for COVID-19 vaccine prioritisation must be established. Such tools can be developed using the kinds of annotated datasets that we have created here, and methods for efficiently creating these datasets are therefore important.

Acknowledgements

This work was supported by the Swedish Research Council (2017-00626).

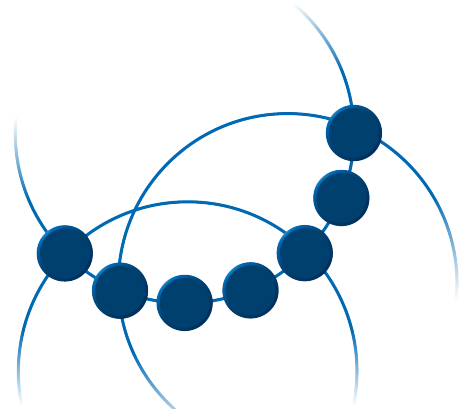
References

- Allen Institute For AI. 2020. COVID-19 open research dataset challenge (CORD-19). <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks>.
- Georgios-Ioannis Brokos. 2018. Biomedical pre-trained word embeddings. <https://github.com/RaRe-Technologies/gensim-data/issues/28>.
- Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Thomas Morton and Jeremy LaCivita. 2003. Wordfreak: An open tool for linguistic annotation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4 (NAACL HLT)*, pages 17–18, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Danielle L Mowery, Sumithra Velupillai, and Wendy W Chapman. 2012. Medical diagnosis lost in translation – analysis of uncertainty and negation expressions in English and Swedish clinical texts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 56–64, Montréal, Canada, June. Association for Computational Linguistics.
- Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The Gavagai living lexicon. In *Proceedings of the Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).
- Andrew I. Schein and Lyle H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Mach. Learn.*, 68(3):235–265, October.

⁶<https://www.kaggle.com/mariaskepstedt/manually-annotated-risk-factor-expressions>

- Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report #1648, University of Wisconsin–Madison. <http://research.cs.wisc.edu/techreports/2009/TR1648.pdf>.
- Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2016. PAL, a tool for Pre-annotation and Active Learning. *JLCL*, 31(1):91–110.
- Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.
- Maria Skeppstedt, Rafal Rzepka, Kenji Araki, and Andreas Kerren. 2019. Visualising and evaluating the effects of combining active learning with word embedding features. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 91–100. German Society for Computational Linguistics and Language Technology (GSCL).
- Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the Linguistic Annotation Workshop*, pages 9–16, Stroudsburg, PA, USA, June. Association for Computational Linguistics.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.
- Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Muhie Seid Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

CLARIN



Common Language Resources and Technology Infrastructure

Linköping Electronic Conference Proceedings
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)
ISBN 978-91-7929-609-4

180
2021

Front Cover Illustration:
Picture Composition by CLARIN ERIC